UNIVERSIDAD DEL BÍO-BÍO
Facultad de Ciencias Empresariales
Escuela de Ingeniería Civil Informática

ANÁLISIS AUTOMÁTICO DE SENTIMIENTOS SOBRE OPINIONES Y/O COMENTARIOS DE NOVELAS EN ESPAÑOL

PROYECTO DE TÍTULO PRESENTADO EN CONFORMIDAD A LOS REQUISITOS PARA OBTENER EL TÍTULO DE INGENIERO CIVIL EN INFORMÁTICA

Alumna: Carol Alejandra Oyarzún Delgado.

Profesor Guía: Clemente Rubio Manzano.

Profesor Co-Guía: Alejandra Segura Navarrete.

16 de Octubre de 2014

El siguiente proyecto de título se la dedico a Manuel Vásquez, por su apoyo incondicional, aunque suene trillado, además de ser un motivador constante e impulsarme siempre emprender nuevos desafíos.

AGRADECIMIENTOS

Me gustaría agradecer en primer lugar a mi familia, por acompañarme en todo este proceso. A Geraldin Senoceain, por su infinita compresión. A mis profesores guías Clemente Rubio y Alejandra Segura, por su orientación en este proyecto de título. Al grupo de investigación SoMoS (Software, Modelling & Science) por su apoyo. A las chicas y chicos del foro Bookzinga, por tomarse el tiempo de contestar la encuesta. Finalmente, a mis compañeros de MCC por siempre inyectar buena onda.

RESUMEN

En este trabajo se enfoca en el análisis de sentimientos, abarcando el proceso para calcular la polaridad de un corpus de opiniones de críticas literarias. Dicho procedimiento, se compone de una primera etapa donde se realiza la extracción de datos, seguido de una limpieza de datos, para posteriormente realizar un pre-procesamiento de los comentarios.

En la etapa de limpieza, tiene como objetivo suprimir, todos los caracteres que puedan dificultar el posterior cálculo de polaridad, como es el caso de las faltas de ortografía, las palabras ajenas al idioma del lexicón. Luego, en la etapa de pre-procesamiento, donde se utiliza la herramienta GATE, que proporciona la posibilidad de implementar una serie de plugins.

Paralelamente, se genera un lexicón, herramienta con la cual se compara cada palabra, para determinar la polaridad. Los lexicones que componen nuestro diccionario son el WordNet-Affect, *fullStrengthLexicon, mediumStrengthLexicon,* una adaptación de *ANEW.* Y finalmente se presenta el diseño e implementación de las pruebas que validarán los resultados.

ABSTRACT

This paper focuses on sentiment analysis, covering the process to calculate the polarity of a corpus of reviews from book reviews. This procedure consists of a first stage, data extraction, followed by a data cleansing, is later does a pre-processing of the comments made.

In the cleaning step, is to remove all characters that may hinder the subsequent calculation of polarity, such as misspellings, the words foreign to the language of the lexicon. Then in the step of pre-processing, where the GATE tool that provides the ability to deploy a number of plugins is used.

Simultaneously, a lexicon, tool which each word is compared, to determine the polarity is generated. The Lexicons that comprise our dictionary are WordNet-Affect, fullStrengthLexicon, mediumStrengthLexicon, an adaptation of ANEW. Finally the design and implementation of tests that will validate the results is presented.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

1 ESTRUCTURA Y ORGANIZACIÓN DEL DOCUMENTO

En este capítulo, se presentan las partes que componen este proyecto de título, describiendo brevemente el contenido de estos.

Capítulo 1 Estructura y Organización del Documento: En este capítulo se presenta la disposición de los capítulos dentro del informe.

Capítulo 2 Introducción Y Motivación: En este capítulo se exhibe los principales motivos impulsado la confección de este proyecto de título, sus objetivos, contribuciones y restricciones.

Capítulo 3 Resumen del Proyecto: En este capítulo se describe la problemática que rodea a este proyecto y se definen los conceptos relacionados.

Capítulo 4 Estado del Arte: En este capítulo se presenta en detalle el estado actual de la disciplina. Señalando las técnicas más utilizadas, y los distintos enfoques existentes.

Capítulo 5 Marco Teórico: En este capítulo se profundizan los conceptos de red social, los diferentes tipos de opinión, y lo enfoques que existen para clasificar su polaridad.

Capítulo 6 Herramientas y Recursos: En este capítulo se describe las herramientas utilizadas en este proyecto de título, y que no han sido desarrolladas por el autor de este informe.

Capítulo 7 Caso de Estudio: En este capítulo se exhibe el procedimiento de selección de la muestra, el método de recolección de información, construcción del corpus, la creación del lexicón, la formulación de la detección automática.

Capítulo 8 Experimentación: En este capítulo se entrega el diseño de las pruebas que se realizaran en este proyecto.

Capítulo 9 Conclusiones y Trabajo Futuro: En este capítulo se recopila las conclusiones extraídas en este trabajo y presenta las posibles proyecciones de este proyecto.

2 INTRODUCCIÓN Y MOTIVACIÓN

En este capítulo se presenta una breve introducción al tema, junto a lo que motiva la realización de este trabajo.

2.1 INTRODUCCIÓN

Como se ha planteado en el contexto de este trabajo, el auge de los medios sociales ha eliminado prácticamente la barrera que permite a los usuarios encontrar y compartir sus opiniones con otros. De hecho, en general se reconoce que las opiniones publicadas por los usuarios son una fuente importante de información para la toma de decisiones (SAKUNKOO & SAKUNKOO, 2009). Por lo tanto, las empresas, las organizaciones, los gobiernos y los diversos grupos en general, también han expresado interés en conocer esa información (FERRAN & HURTADO, 2013), e incluso, la difusión de las aplicaciones comerciales han fomentado dicha situación. Lo anterior, proporciona una fuerte motivación para la investigación, si añadimos que, por primera vez en la historia humana, tenemos una gran cantidad de datos de asesoramiento en la web. (LIU, Bing, 2012)

Dado que la información entregada por el usuario en la web, se presenta como norma general, como un texto poco estructurado, es necesario para realizar un análisis de opinión recurrir a técnicas avanzadas de procesamiento del lenguaje natural.

En los últimos años, ha surgido una nueva disciplina para el problema del análisis computacional de las opiniones, los sentimientos y la subjetividad en el texto, llamado *análisis sentimental*, que a veces se hace referencia como *minería de opinión*. El análisis sentimental incluye varias tareas como la detección de la subjetividad, la polaridad y la intensidad basado en tópicos. (CUADRADO, 2011)

2.2 MOTIVACIÓN

Nos encontramos inmersos en un mundo en el que millones de personas expresan sus opiniones sobre productos comerciales en blogs, wikis, foros, chats y redes sociales, la enorme cantidad de información no estructurada puede ser un factor clave para las empresas que desean crear una imagen o identidad, en la mente de sus clientes (CAMBRIA & HUSSAIN, 2012). Es en este escenario donde la minería de opiniones ha adquirido verdadera importancia con el objetivo de hacer uso de esta gran cantidad de información.

El análisis de sentimientos consiste en identificar opiniones, emociones y valoraciones, tanto negativas como positivas (WILSON, WIEBE, & HOFFMANN, 2005) , utilizando el poder de concretar, especificar para formalizar y polarizar este contenido, permitiendo así obtener información oportuna, acotada y contextualizada sobre un objeto de opinión, logrando la toma decisiones correcta.

A pesar de que existen publicaciones entre las décadas de 1980 y 1990, los avances reales empezaron con el realce de la web 2.0, y específicamente con la consolidación de las redes sociales, lo que hace un tema novedoso e importante. (VILARES, ALONSO, & GÓMEZ-RODRÍGUEZ, 2013) Cabe destacar que el grueso de las publicaciones es acerca de habla inglesa, la cual difiere bastante del habla hispana. (MARTÍNEZ, MARTÍN, & UREÑA, 2011)

Este proyecto se enfoca en el análisis sobre los comentarios de novelas en español mediante un lexicón, qué como caso de estudio es útil, debido a que estas muchas veces se tiende a relacionar que una buena novela es aquella que vende una gran cantidad de ejemplares, y no se toma en cuenta el valor de las opiniones de los lectores de estas, más aún no siempre todo lo que leemos nos suscita una buena opinión.

2.3 OBJETIVOS

En este capítulo, se presenta los objetivos generales y específicos que espera cumplir este proyecto de título, junto a sus aportes y limitaciones.

2.3.1 OBJETIVOS GENERALES

Analizar la polaridad sentimental de los comentarios realizados sobre las novelas de un sitio web.

2.3.2 OBJETIVOS ESPECÍFICOS

- Realizar una investigación del estado del arte concerniente a las técnicas de análisis de sentimiento.
- Investigar las herramientas o API’s para realizar un análisis de opinión.
- Generar un corpus de opiniones literarias por medio de la extracción de los datos de un determinado sitio.
- Realizar un pre-procesamiento de los datos, para el uso de una herramienta o API’s.

- Realizar un análisis sentimental
- Evaluar la validez de los resultados.

2.4 METODOLOGÍA

La metodología utilizada en este proyecto consta de tres partes: *Investigación, Experimentación y Validación*.

- **Investigación:** En esta etapa se realizó un estudio preliminar del tema, donde se recopiló los antecedentes para la bibliografía. En esta etapa se generó el estado del arte acerca del problema, se definió el marco conceptual. A su vez se redactó el marco teórico como resultado final de esta etapa, profundizando en los temas más relevantes del análisis de sentimientos.
- **Experimentación:** Se inició con el estudio de las herramientas existentes en la actualidad en el área de estudio. Además se planificó, diseñó y codificó los experimentos, para su posterior análisis e implementación.
- **Validación:** En esta etapa se validaron los resultados obtenidos en la etapa de experimentación, permitiendo generar conclusiones acordes a los objetivos planteados al inicio del proyecto, además de determinar los aportes reales que entrega esta investigación y el posterior trabajo futuro.

2.5 APORTES

Los principales aportes de este proyecto son:

- Generar un corpus compuesto sólo de opiniones literarias en español.
- Generación de un lexicón a partir de cuatro ya existentes.
- Realizar un análisis de sentimental.
- Generar una lista de las palabras desconocidas por el lexicón.

2.6 LÍMITES

Este proyecto busca la utilización de herramientas ya existentes, en ningún momento tiene la finalidad de innovar en algún algoritmo que optimice el análisis sentimental, y tampoco solucionar el tema de la ironía y el sarcasmo presente en ellos. Solamente se utilizará como caso de estudio opiniones referentes a novelas literarias, de un sitio

seleccionado y la validez del análisis será evaluado sobre una muestra de los resultados obtenidos.

3 RESUMEN DEL PROYECTO

En este capítulo se presenta una síntesis del trabajo, definiendo el problema y presentando los conceptos previos para comprender el siguiente proyecto.

3.1 DEFINICIÓN DEL PROBLEMA

El objetivo principal de este proyecto consiste en analizar la polaridad sentimental de los comentarios realizados sobre las novelas de un sitio, por lo cual, es imperante indagar en la o las herramientas para realizar dichos análisis con el fin de determinar la correlación entre la apreciación que tienen los usuarios de una novela y el ranking asociado del sitio estudiado. Por lo mismo, en el siguiente proyecto de título se enfreta a los siguientes desafíos:

- La faltas de ortografía, jergas, abreviaturas (q, pk), o alargamiento innecesario de palabras (“buuueno”).
- La escasez de herramientas PLN basadas en el idioma español.

3.2 MARCO CONCEPTUAL

Para una mejor comprensión del siguiente trabajo, se presenta la definición de los términos fundamentales que lo componen:

- ***Procesamiento del Lenguaje Natural (PNL)***: Consiste en el estudio y análisis de los aspectos lingüísticos de un texto a través de programas informáticos. Además de investigar mecanismos computacionalmente eficaces para la comunicación entre personas o entre personas y máquinas. (LEIVA, 1996).

- ***Análisis de Sentimientos:*** Es la tarea de identificar opiniones, emociones y valoraciones, tanto negativas como positivas, utilizando el poder de procesamiento informático, para formalizar y polarizar este contenido. (WILSON, WIEBE, & HOFFMANN, 2005).

- ***Corpus:*** Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación. (RAE)

- ***Lexicón:*** Es un conjunto de palabras previamente polarizadas (WILSON, WIEBE, & HOFFMANN, 2005), es decir, donde cada palabra está etiquetada de acuerdo a su orientación semántica. (CUADRADO, 2011)

- ***Lexicón de Affect:*** Corresponde a un tipo de lexicón, donde las palabras polarizadas, se encuentran categorizadas de acuerdo al afecto que transmite el comentario, por ejemplo: *alegría, entusiasmo, ira, tristeza, disgusto, sorpresa, miedo*. (VALITUTTI & STRAPPARAVA, 2004)

- ***Lexicón de Sentic:*** Corresponde a un tipo de lexicón, donde las palabras polarizadas, se encuentran categorizadas de acuerdo al sentimiento que transmite el comentario, por ejemplo: *positivo*, *negativo* (PÉREZ-ROSAS, BANEA, & MIHALCEA). O en algunos casos de acuerdo a una dimensión, por ejemplo: *valencia*, (que oscila entre lo agradable y desagradable) *excitación* (que oscila desde la calma a la agitación), *dominio o control* (que oscila desde el control a la salida de control). (Redondo, 2007)

- ***Orientación Semántica:*** Consiste en predecir la orientación de una opinión, a través de los adjetivos o adverbios que contiene. Una frase tiene una orientación semántica positiva cuando tiene asociaciones buenas (ejemplo “matiz sutil”) y una orientación negativa cuando tiene malas asociaciones (ej. “muy arrogante”). (TURNEY, 2002) En Turney la orientación semántica se calcula como la información mutua (PMI) entre la frase dada y la palabra "*excelente*" menos la información mutua entre la frase dada y la palabra "*pobre*".

- ***Opinión:*** Se define como la exposición de los sentimientos y vivencias de los usuarios respecto a un objeto de opinión. (CUADRADO, 2011) . Las opiniones se encuentran compuestas por: un objetivo *(g)* y un sentimiento *(s)* sobre el objetivo, es decir, *(g, s)*, donde *g* puede ser cualquier entidad o aspecto de la entidad sobre la ha expresado la opinión, y *s* es un sentimiento positivo, negativo o neutral, o una puntuación numérica que expresa la fuerza e intensidad del sentimiento (por ejemplo, 1-5 estrellas*).* (LIU, Bing, 2012)

- ***Hechos:*** Se define como fragmentos de texto donde el usuario expresa sentencias objetivas sobre el producto o servicio. Un ejemplo de esto, es la siguiente frase “*Compramos este teléfono para reemplazar un teléfono inalámbrico que tuve por cerca de nueve años.*” (CUADRADO, 2011)

- ***Sentimiento:*** Se refiere a tanto a un estado de ánimo, como a una emoción conceptualizada que determina el estado de ánimo (RAE). En el contexto de la PNL, el término sentimiento consiste en las emociones o la polaridad que transmite el texto. (CAMBRIA & HUSSAIN, 2012)

- ***Emoción:*** Las emociones son nuestros sentimientos y pensamientos. Las cuales han sido estudiadas en varios campos, por ejemplo, la psicología, la filosofía y la sociología. Y están estrechamente relacionadas con los sentimientos. La fuerza de un sentimiento u opinión está típicamente ligada a la intensidad de ciertas emociones. (LIU, Bing, 2012)

- ***Clasificación de polaridad:*** Tarea que pretende, como última finalidad, clasificar fragmentos de texto, que pueden ser desde documentos hasta sintagmas, en positivo o negativo dependiendo de su significado emocional. (CUADRADO, 2011)

- ***Minería de opiniones:*** Hace referencia a las técnicas computacionales para extraer, clasificar, comprender y valorar las opiniones expresadas en diversas fuentes de noticias en línea, comentarios de medios sociales, y otros contenidos generados por los usuarios. (CAMBRIA & HUSSAIN, 2012)

- ***Synset:*** Es un conjunto de relaciones conceptuales, que se consideran semánticamente equivalentes, y que al menos en parte, consideramos independiente de su idioma. (VALITUTTI & STRAPPARAVA, 2004)

4 ESTADO DEL ARTE

En el siguiente capítulo se presenta los trabajos más relevantes en ámbito del análisis de sentimientos.

Para comenzar tenemos que el término análisis de sentimientos fue usado por primera vez en la literatura en la obra de (SANJIV & CHEN, 2001) y (TONG, 2001) en la predicción de juicios para analizar el comportamiento de mercado.

Unos años más tarde, Pang, Lee, Vaithyanathan y Turney, abordarón el problema de la clasificación de la polaridad. Donde en (PANG, LEE, & VAITHYANATHAN, 2002) presentó el uso de técnicas de aprendizaje de máquina a través de tres *algoritmos Naive Bayes, Maximum Entropy y Support Vector Machines*. Por otra parte, (TURNEY, 2002) lo realizó a través de un clasificador no supervisado. Este clasificador determinaba la naturaleza positiva o negativa de un documento basado en la orientación semántica de términos que pueden estar representados por el algoritmo de *PMI-IR (Pointwise Mutual Information- Information Retrieval)*, que se basa en la frecuencia de co-ocurrencia de los términos. De estos trabajos, se desprendieron los dos enfoques existentes, el *semántico* (TURNEY, 2002) y el *supervisado* (PANG, LEE, & VAITHYANATHAN, 2002)

Años más tarde, (WILSON, WIEBE, & HOFFMANN, 2005) propuso determinar preliminarmente la *polaridad* de una *opinión*, para luego eliminar las que presentaban una *polaridad* ambigua. En (FERMÍN, A., ENRIQUEZ, & ORTEGA, 2008) se realizó una clasificación de un *corpus* de opiniones sobre críticas de cine utilizando el algoritmo PMI-IR. Posteriormente (FERNÁNDEZ, GÓMEZ, BOLDRINI, & MARTÍNEZ-BARCO, septiembre de 2011) utilizó los *corpus* EmotiBlog Kyoto, Emotiblog Phones, JRC, para determinar el beneficio que implicaba su utilización en la determinación de intensidad y la emoción de las opiniones. Luego, (Saralegi Urizar, 2012) planteo una solución supervisada que comprendía el tratamiento de emoticones, la negación, y tareas de lematización y etiquetado. Paralelamente (Trilla, 2012) presentó una clasificación de texto basado en Multinomial Naive Bayes para procesar mensajes de twitter. En (Martínez Cámara, 2012) se utilizó el algoritmo *SVM* (Máquina de Vectores de Soporte) para determinar la *polaridad* de una serie de tweet en español. Ese mismo año, en (Fernández Anta, 2012) se compararon los rendimientos de varios clasificadores supervisados. Posteriormente, (FERRAN & HURTADO, 2013) se aplicó el algoritmo *SVM*, a través de la *libreriaSVM* que se integraba a *WEKA*, que determinaba un *análisis de sentimientos* a nivel global y de entidad de los tweet, clasificarlos por tópicos, y finalmente la tendencia política de los usuarios. Por último,

en (VILARES, ALONSO, & GÓMEZ-RODRÍGUEZ, 2013) se planteó una aproximación híbrida, que combina conocimiento lingüístico obtenido con técnicas de *aprendizaje automático*, que posteriormente entrenaba a un *clasificador supervisado*.

5 MARCO TEÓRICO

En este capítulo, se presenta una breve explicación de qué son las redes sociales, qué es una opinión, su clasificación y finalmente cómo se clasifica su polaridad.

5.1 REDES SOCIALES

Un sitio web puede considerarse como una red social cuando primero, permite que un individuo construya un perfil público o parcialmente público, dentro del sistema. Segundo, puede construir una lista de usuarios con el objetivo de compartir una conexión o información. Y tercero permite ver y recorrer la lista de conexiones que él y sus contactos hayan creado. De acuerdo a la definición anterior, la primera red social fue creada 1997, llamada *SixDegrees.com*, que permitía a los usuarios crear perfiles, lista de sus amigos y, a partir de 1998 navegar por las listas de amigos. (ELLISON, 2007)

Una característica fundamental, es que entrega la posibilidad de generar contenido de opinión, pero medio de comentarios realizados por el usuario, que los demás integrantes de la red social, pueden rebatir o complementar. (FERNÁNDEZ, 2013)

5.2 OPINIÓN

Tal como se definió en la **sección 3.3,** una *opinión* se define como la exposición de los sentimientos y vivencias de los usuarios respecto a un producto o servicio. (CUADRADO, 2011) A continuación se definen los distintos tipos de opinión existentes, según (LIU, Bing, 2012)

5.2.1 TIPOS DE OPINIONES

5.2.1.1 Opinión Regular

Corresponden simplemente a una opinión en la literatura y tiene dos principales subtipos (BING, Liu, 2006 y 2011):

- ***Opinión Directa*:** Se refiere a una opinión expresada directamente en una entidad o un aspecto de la entidad, por ejemplo, "*La calidad de imagen es muy grande.*" (LIU, Bing, 2012)
- ***Opinión Indirecta:*** Una opinión indirecta es una opinión que se expresa indirectamente a una entidad o aspecto de una entidad sobre la base de sus efectos sobre algunas otras entidades. Por ejemplo "*Después de la inyección de la droga, mis articulaciones se sintieron peor"*, describe un efecto no

deseado de la droga en "*mis articulaciones*", que indirectamente da una opinión negativa. (LIU, Bing, 2012)

5.2.1.2 Opinión Comparativa

Corresponde una opinión expresa una relación entre dos o más objetos, ya sea, sus similitudes o diferencias y/o una preferencia superior basados en algunos aspectos comunes entre las entidades. Por ejemplo "*Marta es más inteligente que Sofía"* y *"Este es el peor plato de comida que existe".* (JINDAL, Nitin; BING, Liu, 2006a.) (JINDAL & BING, Mining comparative sentences and relations, 2006b.)

5.2.1.3 Opinión Implícita

Una opinión implícita es una declaración subjetiva que da una opinión regular o comparativa. Por ejemplo, *"Coca-Cola tiene un gran sabor",* y *"Coca-Cola sabe mejor que Pepsi."* (LIU, Bing, 2012)

5.2.1.4 Opinión Explicita

Una opinión implícita es una declaración objetiva que implica una opinión regular o comparativa. Tal declaración de objetivos por lo general expresa un hecho deseable o indeseable. Por ejemplo, "*Compré el colchón hace una semana, y un valle se ha formado*", y *"La vida de la batería de los teléfonos Nokia es más largo que los teléfonos Samsung.*" (LIU, Bing, 2012)

5.3 CLASIFICACIÓN DE POLARIDAD DE LAS OPINIONES

Con el fin de calcular la polaridad de las opiniones existen dos enfoques, que abordan dicha problemática, que se explican a continuación.

5.3.1 ENFOQUE AUTOMÁTICO

El enfoque automático se caracteriza, por utilizar algoritmos de aprendizaje automático, esto implica, que a medida que estos son utilizados se va adquiriendo conocimiento de los patrones existentes en el texto, que luego serán utilizados para clasificar otros documentos (FERNÁNDEZ, 2013). Entre las principales herramientas se encuentran: el clustering, entropía máxima, redes bayesianas, SVM (Máquina de Vectores de Soporte), redes neuronales, árboles de decisión y regresión lineal. (MANNING, RAGHAVAN, & SCHAJTZE, 2008) (WESTERSKI, 2007)

El principal problema de este enfoque, reside en su dependencia del dominio en el que esté inmerso y el costo de crear conjuntos de entrenamientos. En estas herramientas el texto estudiado se representa como una bolsa de palabras, las cuales la aprende

correctamente en un ambiente específico (VILARES, ALONSO, & GÓMEZ-RODRÍGUEZ, 2013). Es por lo mismo, que su rendimiento cae considerablemente al clasificarlos en un dominio distinto. (TABOADA, BROOKE, TO, & STEDE, 2011)

5.3.2 ENFOQUE SEMÁNTICO

El enfoque semántico se caracteriza por emplear un diccionario genérico (VILARES, ALONSO, & GÓMEZ-RODRÍGUEZ, 2013), es decir, no requieren de ningún algoritmo de entrenamiento. (MARTÍNEZ, GARCÍA, A., MARTÍN, GARCÍA, & UREÑA, 2012) La principal herramienta que utiliza este enfoque es el lexicón. (En la **sección 6.2**, se detalla su funcionamiento)

Por otro lado, a pesar que este enfoque ha demostrado ser útil en distintos ambientes, su rendimiento en Twitter disminuye, debido que en dicho medio existe una elevada frecuencia de abreviaturas, emoticones, expresiones o jergas que no se encuentran en un lexicón genérico. (ZHANG, MOHAMED, & BING, 2011)

6 HERRAMIENTAS Y RECURSOS

En este capítulo se describen las distintas herramientas utilizadas en el desarrollo de este proyecto de título.

6.1 HERRAMIENTAS LINGÜÍSTICAS

Las herramientas lingüísticas se pueden definir como un software que implementa técnicas y métodos para realizar análisis lingüísticos a distintos niveles de un texto de entrada. (CUADRADO, 2011), en nuestro caso se utilizó la herramienta *GATE*, debido a la variedad de complementos que posee.

6.1.1 GATE

La herramienta GATE, del inglés *General Architecture for Text Engineering* (CUNNINGHAM, MAYNARD, BONTCHEVA, & TABLAN, 2002) es una arquitectura para el desarrollo y despliegue de componentes de software que procesan el lenguaje humano. GATE es un software gratuito de código abierto, (CUNNINGHAM H. , y otros, 2012) desarrollado por la Universidad de Sheffield (Reino Unido), en respuesta a la necesidad de un proceso de estandarización para el desarrollo de herramientas de análisis lingüístico. El proyecto comenzó en 1995 y todavía está activo con muchas actualizaciones y mejoras. (CUADRADO, 2011).

Por regla general, un sistema de ingeniería lingüística hace uso de tres componentes básicos; algoritmos, datos de entrada y una interfaz gráfica para representar los datos. Es por lo anterior que los recursos en GATE están clasificados en las siguientes tres categorías (CUADRADO, 2011):

- *Recursos Lingüísticos*: Representan entidades como léxicos, corpus u ontologías. (CUNNINGHAM H. , y otros, 2012)
- Recursos de Procesamiento: Representan entidades que son principalmente algorítmica, tales como analizadores, generadores o modeladores. (CUNNINGHAM H. , y otros, 2012)
- Recursos de Visualización: Representan componentes de visualización y edición que participan en interfaces gráficas de usuario. (CUNNINGHAM H. , y otros, 2012)

El conjunto de recursos integrados con GATE se conoce como CREOLE, una colección de objetos reutilizables para ingeniería lingüística. Todos los recursos se empaquetan en un archivo Java (o 'JAR'), además de algunos datos de configuración XML (CUNNINGHAM H. , y otros, 2012). GATE admite datos de entrada en múltiples

formatos, incluyendo XML, RTF, email, HTML, SGML y texto plano (CUNNINGHAM H. , y otros, 2012).

Además GATE dispone de una gran variedad de plugins en los distintos idiomas, por ejemplo *Spanish Plugin*, ANNIE, *Readability Tools*, que se explicarán en detalle en la **sección 7.3.2.4**. Además de su amplio uso dentro de la comunidad científica para el procesamiento del lenguaje natural, (CUADRADO, 2011) son unas de las razones por la que se utilizó esta herramienta para el pre-procesamiento del corpus.

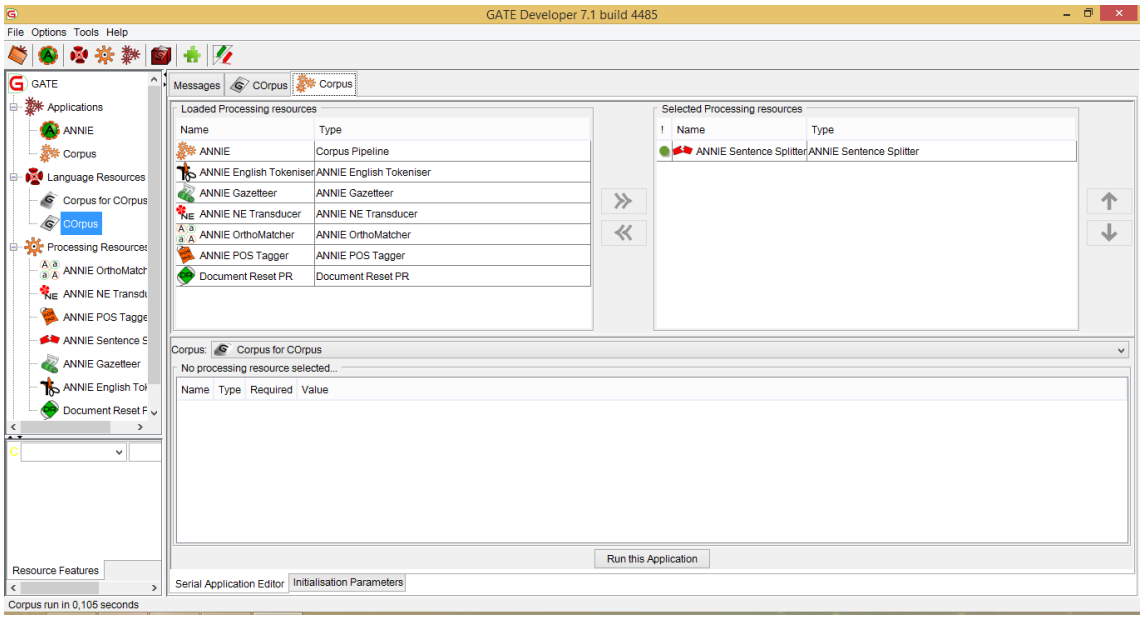


Figura 1: Interfaz de la Herramienta GATE

6.2 LEXICONES

Un lexicón, corresponde a un conjunto de palabras, previamente polarizadas (WILSON, WIEBE, & HOFFMANN, 2005). Donde su principal ventaja, es la posibilidad de encontrar fácilmente y rápidamente un gran número de palabras. A pesar que se pueden encontrar errores, estas pueden ser resueltas fácilmente con una simple comprobación manual. La principal desventaja, es que no considera el contexto o dominio de las palabras, es decir, que una determinada palabra puede estar etiquetada previamente como negativa, pero su verdadera polaridad considerando otro dominio sería positiva. Por ejemplo, para el altavoz de un teléfono, está *callado*, por lo general es una característica negativa. Sin embargo, para un auto, está *callado,* es una apreciación positiva. (CUADRADO, 2011). En este proyecto, se trabajó con un lexicón conformado por cuatro de estos, desarrollados en los paper´s que se describirán a continuación.

6.2.1 THE SPANISH ADAPTION OF ANEW

Corresponde a un lexicón de sentic, realizado por el grupo de Jaime Redondo et al. , basándose en el *Affective Norms for English Words,* (ANEW) compuesta de 1034 palabras en inglés, que posteriormente fueron traducidas.

Para crear este lexicón se reunió a 720 personas, quienes evaluaron estas palabras, y las clasificaron en tres dimensiones, donde cada una presentaba una valoración en una escala de 1 a 9. Por un lado, tenemos la dimensión valencia (oscila entre lo agradable y desagradable), la dimensión excitación (oscila desde la calma a la agitación) y finalmente la tercera dimensión dominio o control (oscila desde el control a la salida de control). Un ejemplo de esto sería, la palabra *funeral*, que tiene una *valencia* de 1.48 y una *excitación* de 5.06, es decir, representa una palabra *agradable* para el usuario y además produce una *excitación* intermedia. Otro ejemplo, es la palabra *madre*, con una *valencia* 8.19, y *excitación* 5.19, que representa una palabra *agradable* con una *excitación* neutra. (Redondo, 2007)

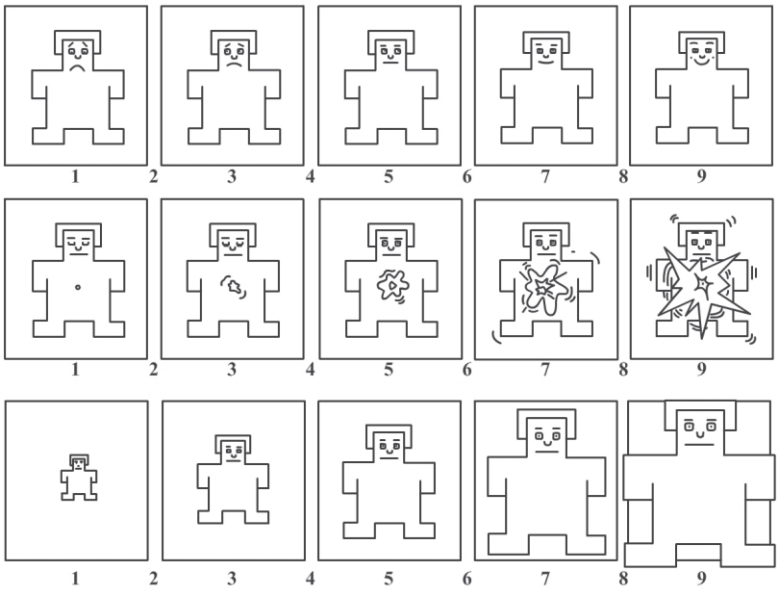


Figura 2: Escala the spanish adaptation of ANEW

En la **figura 2**, se entrega la escala utilizada por este lexicón, donde la primera fila representa la dimensión *valencia*, donde el extremo izquierdo representa algo *desagradable* y el extremo derecho algo *agradable*. La segunda la dimensión de *excitación*, el lado izquierdo representa la *calma* y el derecho *agitación*. Y la tercera el *dominio*, que varía de *control* a *descontrol* de izquierda a derecha, respetivamente.

6.2.2 LEARNING SENTIMENT LEXICONS IN SPANISH

Corresponde a un lexicón de sentic, realizado por Verónica Pérez-Rosas et al., quienes elaboraron dos lexicones ambos en español, el primero *fullStrengthLexicon* contiene un léxico proveniente del *OpinionFinder lexicón* (WIEBE & RILOFF, 2005), conformado por 1347 palabras que fueron clasificadas de manera manual, consultando las distintas ponderaciones que estaban asociadas a una palabra en particular, y tomando su valor más alto. El segundo lexicón, *mediumStrengthLexicon*, conformado por 2496 palabras, que fue generado de manera automática, basándose en el *SentiWordNet* (ESULI & SEBASTIANI, 2006), seleccionando el synset (conjunto de relaciones conceptuales) que contiene una puntuación superior a 0.5. Ambos lexicones fueron etiquetados como positivo (pos) y negativo (neg). (PÉREZ-ROSAS, BANEA, & MIHALCEA)

6.2.3 DEVELOPING AFFECTIVE LEXICAL RESOURCES

Consiste en un lexicón de affect, realizado por Alessadro Valitutti et al., quienes elaboraron el *WordNet-Affect* a partir del *WordNe*t, por medio de la selección y etiquetado del syntec (relaciones semánticamente equivalentes), representando afectivamente el concepto. El syntec proporciona una correlación entre un concepto y las palabras correspondientes. Paralelamente se le añadió una etiqueta de dominio, que representa el concepto afectivo que personifica un estado emocional, entre los que se encuentran: *alegría, sorpresa, ira, tristeza, disgusto y miedo*. (VALITUTTI & STRAPPARAVA, 2004) Por ejemplo, podemos tener un conocimiento afirmativo de que la gente le tiene *miedo* a los terremotos. Basándonos en algunas declaraciones y analizándolas, nos entrega relaciones semánticas que las conectan entre sí. Así que, cada vez que la palabra *terremoto* aparezca, el sistema podría saber que produce *miedo,* y por lo tanto clasificar el texto con esta dimensión adicional. (VALITUTTI & STRAPPARAVA, 2004)

7 CASO DE ESTUDIO

A continuación, se describe el proceso que se efectúa en este proyecto.

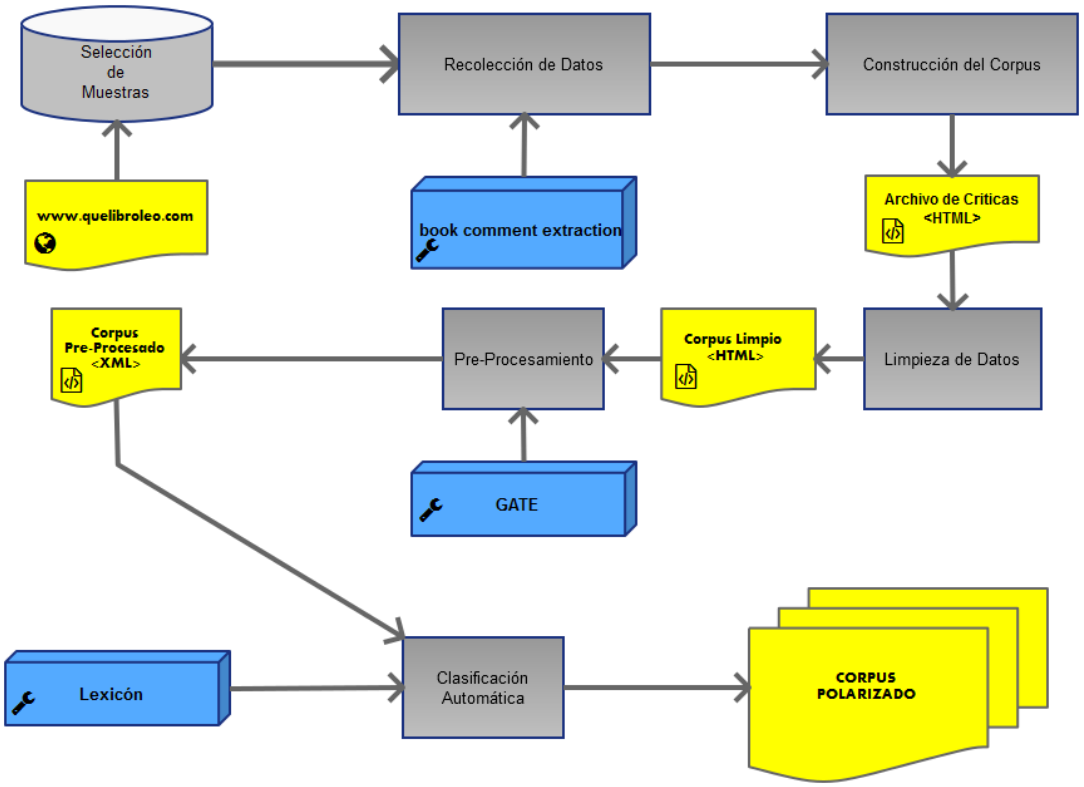


Figura 3: Diagrama de Caso de Estudio

En la **figura 3**, se presentan las etapas que inician con la selección de la muestras, es decir, la elección del sitio donde extrajeron los datos. Posteriormente, se procedió a la recolección de los datos a través de un archivo en java. Con los datos obtenidos, se construyó el corpus. Dicho corpus se encontraba con mucho ruido, es decir, con caracteres que dificultaban el cálculo de la polaridad. Una vez, que se obtuvo el corpus limpio en un formato HTML, se procedió a realizar el pre-procesamiento, por medio de la herramienta GATE, en conjunto con una serie de plugins. Paralelamente se generó, el Lexicón, a partir de la unión de cuatro lexicones, existentes. Finalmente se realizó la clasificación automática, es decir, se calculó la polaridad de cada comentario, culminando con un corpus polarizado. El detalle de cada parte de este proceso, se describe a continuación.

7.1 SELECCIÓN DE MUESTRA

Para la creación de nuestro corpus, se buscó una página web en español, que se dedicará reseñar libros de distintos géneros, y permitiera a los usuarios opinar y/o comentar sobre estos.

Las páginas encontradas fueron las siguientes:

Lecturalia[1]*:* Fundada en 2006, es una de las redes sociales de literatura más extendida en España. Está página cualquiera puedes obtener valoraciones y referencias a todo tipo de novelas y libros, con la peculiaridad de que también es posible adquirir directamente su versión digital pagando.

Figura 4: Lecturalia

Quelibroleo[2]*:* Considerada el primer 'facebook literario'. Lanzada en 2008 por José Luis y Alberto Ramírez. El registro en esta página es gratuito y permite al usuario poner nota a sus lecturas favoritas, así como conocer qué libros le recomiendan otros usuarios.

[1] www.lecturalia.com
[2] www.quelibroleo.com

Figura 5: Quelibroleo

Wattpad[3]*:* Es una comunidad en línea entorno a temas de escritura y la narración. Los usuarios pueden publicar artículos, relatos y poemas sobre cualquier cosa, ya sea en línea o a través de la aplicación Wattpad. El contenido incluye obras de autores desconocidos y publicados. Los usuarios pueden comentar y pueden votar por las historias o unirse a grupos asociados con el sitio web.

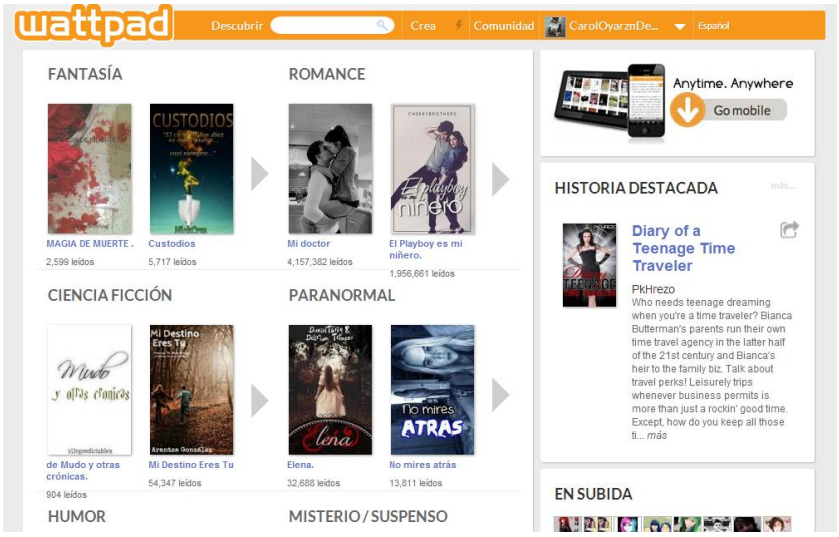


Figura 6: Wattpad

Librote[4]*:* Red social que ayuda a registrar libros y obtener recomendaciones. Lanzada en mayo de 2013. Entre sus funcionalidades se encuentran redactar críticas, comentarios, añadir amigos, editar la información de los libros, enviar libros, etc.

[3] http://www.wattpad.com/
[4] https://www.librote.com/

Figura 7: Librote

Para seleccionar el sitio en donde se extrajó nuestro corpus, este debía cumplir con las siguientes características (FERMÍN, A., ENRIQUEZ, & ORTEGA, 2008):

1. Un alto número de críticas disponibles, se considera alto a partir 2000 opiniones registradas en el sitio.
2. Cada crítica debe llevar asociada la puntuación que el usuario le da a la novela en cuestión. Que permitirá distinguir si una crítica contiene una opinión favorable o desfavorable.

En la siguiente tabla se compara los cuatro sitios descritos anteriormente.

	Criterio 1	**Criterio 2**
Lecturalia	Cumple. (El libro más criticado consta de 150 opiniones)	Cumple
Quelibroleo	Cumple. (El libro más criticado consta de 229 opiniones)	Cumple
Wattpad	Cumple. (El libro más criticado consta de 246 opiniones)	No Cumple
Librote	No cumple. (El libro más criticado consta de 26 opiniones)	Cumple

Tabla 1: Tabla Comparativa

De acuerdo a los datos presentados los sitios que cumplen con los criterios señalados son *Lecturalia* y *Quelibroleo.* Pero se trabajó con *Quelibroleo,* ya que este presentaba una mayor cantidad de críticas por libro.

7.2 RECOLECCIÓN DE DATOS

La extracción de los datos, se realizó a través de un programa en java, que funciona de la siguiente manera.

```
INICIO
        Archivo_url = leer; // el programa lee la url, que se encuentra previamente
        almacenado en un archivo txt.
        Para cada url del archivo
                Página= obtener_HTML (URL);
                Para cada substring de la página
                        Si substring es un comentarios
                                Comentarios -> agregar(substring);
                        FIN SI
                FIN Para
                Para cada comentario es comentarios
                        Comentario = etiquetar (comentario) // se le agrega div al inicio
                        de cada comentario
                        Corpus_libro-> concatenar (comentario)
                FIN Para
                        Guardar Corpus_libro;
        Fin Para
FIN
```

Tabla 2: Recolección de Datos

A grandes rasgos, el programa identifica la ubicación de los comentarios de acuerdo a las estructura HTML del sitio. El enlace del programa utilizado, las instrucciones de uso, junto con los comentarios extraídos, la lista de URL de cada libro se encuentran en los enlaces presentes en los siguientes **anexos 11.14.1, 11.14.4, 11.14.2, 11.14.5.**

7.3 CONSTRUCCIÓN DEL CORPUS

La extracción de los datos se realizó el día 31 de Mayo de 2014, se recopilaron un total de 2545 críticas provenientes de 69 libros con una nota promedio de 6,5.

Cada crítica tiene asociada una nota colocada por el usuario. El detalle de la escala se detalla a continuación[5]:

Nota	Categoría
1-2	Pésimo
3-4	Malo
5	Regular
6	Bueno
7-8	Muy Bueno
9-10	Excelente

Tabla 3: Escala quelibroleo

De los 69 libros, 6 pertenecían a la categoría de *Excelente*, 18 a *Muy Bueno*, 9 a la categoría de *Bueno*, 28 a *Regular*, 8 a *Malo*.

Lo anterior se representa en el siguiente gráfico.

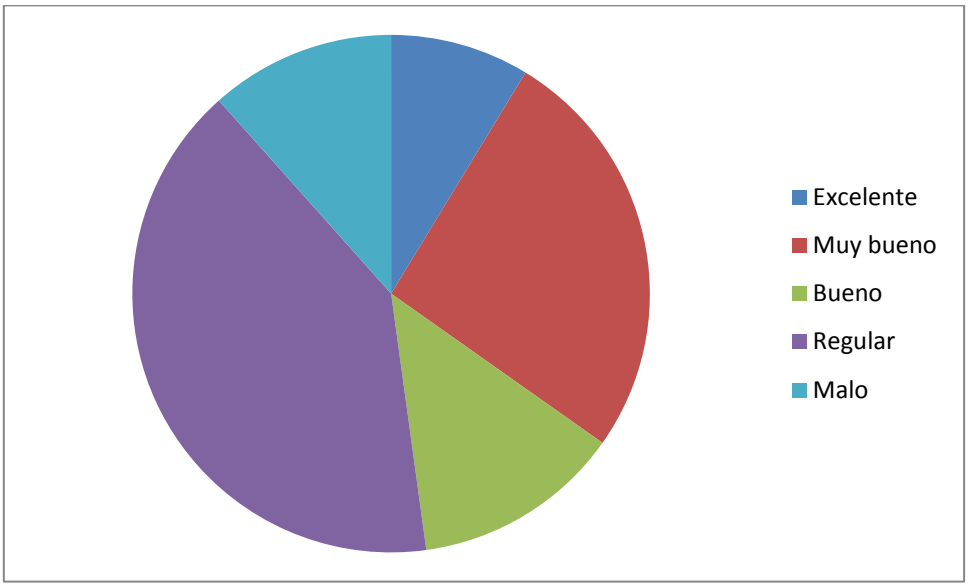


Figura 8: Distribución de Categorías

La lista de cada uno de los libros seleccionados, se encuentra en el enlace del **anexo 11.14.3.**

[5] http://www.quelibroleo.com/

7.3.1 LIMPIEZA DE DATOS

El Corpus empleado en esta fase se encuentra, en el enlace presente en el **anexo 11.14.6.** Dichos comentarios fueron sometidos a una serie de procedimientos, con el fin de reducir la mayor cantidad de ruido posible, que corresponde a las expresiones, signos que interfieren en la extracción de la polaridad. A continuación se presenta una tabla, donde se presenta los principales criterios empleados.

Criterio	Acción
Si aparecen palabras abreviadas.	Se sustituye la palabra condensada por su forma completa.
Si se presentan risas.	Se normaliza la risa, por una única expresión.
Si aparecen emoticones.	Se suprimen dichas expresiones.
Si se presentan expresiones sin significado aparente.	En el caso de la expresión no represente un aporte lingüístico se elimina. En caso contrario se reemplaza por un sinónimo.
Si aparecen palabras de origen extranjero.	Se sustituye por su homólogo en español.
Si presentan palabras con letras repetidas.	Se suprimen las letras repetidas de la palabra.
Si aparece una palabra repetida.	Se elimina la palabra duplicada.
Si se presenta una URL.	Se suprime la URL.
Si aparece un Nick de un usuario dentro de un comentario.	Se elimina el nombre del usuario en cuestión.
Si se presentan faltas ortográficas y de tipeo	Se corrige el error en cuestión sin alterar la semántica de la oración.
Si aparece una serie de asteriscos.	En este caso se elimina dichos símbolos que son indicadores de palabras grosera dentro de los comentarios.
Si se presentan deletreo de palabras.	Se recompone la palabra en su expresión conocida.
Si aparecen comentarios spam.	Se eliminan dichos comentarios.
Si se presenta el puntaje del comentario dentro de este.	En el caso de que el comentario solo contenga el puntaje asignado al libro se suprime

	completamente. En cambio si el comentario contiene una opinión sobre el libro, y además el puntaje asociado, sólo se elimina la calificación.

Tabla 4: Resumen de Criterios Empleados

El detalle de cada de cada criterio se describe a continuación:

- **Corrección de abreviaturas más frecuentes:** (VILARES, ALONSO, & GÓMEZ-RODRÍGUEZ, 2013) Se sustituyó algunos de las abreviaturas por su forma conocida. Entre las que destacan: *q, desd, est, m, xq, aunk, aunq, peli, KK, tod, muxo, x, pq, tb, +, =, k, d, enganxa, niñ@, ppio, qu, tod@, convenci@, int.*
  Del total de comentarios, el 2% contenía palabras abreviadas, donde, la más frecuente fue *q.* A continuación, se presentan algunos ejemplos, junto con su respectiva corrección.

  "es un clásico q se debe leer, es fantástico una gran obra."
  es un clásico que se debe leer, es fantástico una gran obra.

  "lo tuve muchos años en mi estanteria,cuando lo lei m reproche no haberlo echo antes,excelente"
  lo tuve muchos años en mi estantería, cuando lo leí me reproche no haberlo hecho antes, excelente

  "Una trama perfectamente hilada+ unos personajes totalmente creibles y profundos+un mundo rico y complejo= Un libro increible!empiezo ahora el segundo libro,pero tengo claro k se trata de una historia k merece ser leida."
  Una trama perfectamente hilada más unos personajes totalmente creíbles y profundos más un mundo rico y complejo igual Un libro increíble! empiezo ahora el segundo libro, pero tengo claro k se trata de una historia k merece ser leída.

  "Me parecio un libro super original en la forma de contarlo y la historia me engancho desde el principio, es un poco duro pero eso tb hace que la historia sea muy real. Le doy un 10 de lo mejor que he leido últimamente"

Me pareció un libro súper original en la forma de contarlo y la historia me engancho desde el principio, es un poco duro pero eso también hace que la historia sea muy real. Le doy un 10 de lo mejor que he leído últimamente

"Leible pq es de Gabo, pero de lo mas flojo que he leído de el".
Leíble porque es de Gabo, pero de lo más flojo que he leído de él.

- **Normalización de las risas:** (MARTÍNEZ, GARCÍA, A., MARTÍN, GARCÍA, & UREÑA, 2012) Dado que la risa se puede expresar de múltiples formas, se sustituye por una misma expresión "*jajá*".
  Del total de comentarios, el 0,43% contenía expresiones de risas, donde, la más frecuente fue *jejeje.* A continuación, se presentan un ejemplo, junto con su respectiva corrección.

  "Mucho relleno , descipciones largas y agotadoras . Si vas a por lo sustacial del libro , te lo puedes leer en 30 minutos y el caso es que lo tiene pero no te deja mucho lugar a imaginación. Eso si , la sustancia es MUY buena. Es un clásico de juventudes cuarentonas , jijiji"
  Mucho relleno, descripciones largas y agotadoras. Si vas a por lo sustancial del libro, te lo puedes leer en 30 minutos y el caso es que lo tiene pero no te deja mucho lugar a imaginación. Eso sí, la sustancia es MUY buena. Es un clásico de juventudes cuarentonas, jajá.

- **Eliminación de Emoticones:** Se eliminaron la expresiones tales como: :) , //\\, ;) , :S , :(, =D , O_o, XD, xd, XDDD, :D, :P, ;P. Dado que en nuestro trabajo no profundiza en la etiquetación de emoticones.
  Del total de comentarios, el 0,90% contenía palabras contenía emoticones, donde, la más frecuente fue *:) .* A continuación, se presentan algunos ejemplos, junto con su respectiva corrección.

  "El primero libro de Gabo que leí, me gusto mucho, sin duda me encanta este auto y sus historias :)"

  El primero libro de Gabo que leí, me gustó mucho, sin duda me encanta este auto y sus historias

"muy buen libro ya me leeré el segundo!! es una lastima que el autor haya muerto :(. "

muy buen libro ya me leeré el segundo!! es una lástima que el autor haya muerto.

"Me encantó, muy muy bueno, vi la película y definitivamente tenia que leer el libro y TODOS los libros... me encantó sobre todo porque caí enamorada.. =D!"

Me encantó, muy muy bueno, vi la película y definitivamente tenía que leer el libro y TODOS los libros... me encantó sobre todo porque caí enamorada…

- **Normalización de Expresiones:** Se eliminaron expresiones tales como: *ufff, ejem, bluff, bla bla bla, uy!, buaa, puff, buff,* ya que no presentan un aporte lingüístico. Paralelamente se encontraron frases como *ni fu, ni fa*, que fue reemplazado por *ni bien, ni mal*. Y *plis plas* fue reemplazado por *un momento*
  Del total de comentarios, el 0,63% contenía estas expresiones. A continuación, se presentan algunos ejemplos, junto con su respectiva corrección.

"De incomprensible éxito... Aunque su éxito retrata a la sociedad europea del siglo XXI perfectamente: no por su contenido (simple a más no poder) sino porque hace falta que un texto tenga solo treintaipico hojas para que la gente lea algo de ensayo... Ni fu, ni fa".

De incomprensible éxito... Aunque su éxito retrata a la sociedad europea del siglo XXI perfectamente: no por su contenido (simple a más no poder) sino porque hace falta que un texto tenga solo treinta pico hojas para que la gente lea algo de ensayo... Ni bien, ni mal.

"Entretenida pero ni mucho menos me parece una novela como para haberle dado tanta bombo, sinceramente me parecio un blufff. Novelas como esta o mejores hay muchas para mi gusto".

Entretenida pero ni mucho menos me parece una novela como para haberle dado tanta bombo, sinceramente me pareció un engaño. Novelas como esta o mejores hay muchas para mi gusto.

"El libro se lee en un plis plas ....porque la historia engancha"

El libro se lee en un momento....porque la historia engancha

"Dios Mío... yo nunca he podido con él. Sé que es un clásico, que debería leerlo, bla, bla, bla, pero no puedo".

Dios Mío... yo nunca he podido con él. Sé que es un clásico, que debería leerlo, y todo lo demás, pero no puedo.

- **Palabras Extranjeras:** Se sustituyó las palabras de origen extranjero, por su homólogo en español. Entre las que destacan: *yankees, WTF?, podés, let me in, light, WOW, leit motin, volao, liado, releche, naïf, copy, not bad, to be continued?, woman, queres.*
  Del total de comentarios, el 0,66% contiene palabras extranjeras. A continuación, se presentan algunos ejemplos, junto con su respectiva corrección.

  "es lobro es maravilloso porque describe los yankees se introducen alos naciones"

  es libro es maravilloso porque describe los norteamericano se introducen a los naciones

  "El ritmo vertiginoso del libro es la única razón por la cual lo terminé, se lee en un "volao". lo que me da que pensar que su éxito se debe únicamente a conspiranoicos y periodistas literarios trasnochados ávidos de morbo anticlerical."

  El ritmo vertiginoso del libro es la única razón por la cual lo terminé, se lee en un "segundo". lo que me da que pensar que su éxito se debe únicamente a

conspiranoicos y periodistas literarios trasnochados ávidos de morbo anticlerical.

"Naïf, dulce, lectura rápida"

Ingenua, dulce, lectura rápida

"No es tan malo pero a veces se hace aburrido leer de tantos aromas Not bad La mejor parte es el capítulo penúltimo; qué narrativa!"

No es tan malo pero a veces se hace aburrido leer de tantos aromas no está mal La mejor parte es el capítulo penúltimo; qué narrativa!

"Un libro muy entretenido, pero el final es totalmente ?to be continued?."

Un libro muy entretenido, pero el final es totalmente ?Continuará?.

- **Normalización de palabras con letras repetidas:** El proceso consistió en transformar a la palabra a su forma conocida. Entre las que destacan: *graan, laaaargo, muuchaaa, taaan, perfectooo, muchooooo, essss miiiiiaaaaa, leeeeeto, grandeeee.*
  Del total de comentarios, el 0,39% contiene palabras con letras repetidas. A continuación, se presentan algunos ejemplos, junto con su respectiva corrección.

"Repetitivo, aburrido y laaaargo, esa es mi humilde opinión sobre este libro. Me costó horrores acabarlo y cuando lo hice no entendí el por qué de su buena fama".

Repetitivo, aburrido y largo, esa es mi humilde opinión sobre este libro. Me costó horrores acabarlo y cuando lo hice no entendí el por qué de su buena fama.

"Diferente, fresco, entretenido y rápido. Bueno para leerlo, pero la pelicula deja muchooooooooo q desear"

Diferente, fresco, entretenido y rápido. Bueno para leerlo, pero la película deja mucho que desear

"Grandeeeeeeeeee".

Grande

- **Eliminación de palabras repetidas:** El proceso consistió en eliminar estas palabras, ya sea que fueran escritas intencionalmente para hacer hincapié a una determinada característica o inconscientemente y tratarse de un simple error de tipeo. De lo anterior, las más destacadas fueron: *muchos, tan, gran, muy, malo, flojito, mucha, mucho, no es en, se una, un, los, la de.*
  Como en nuestro trabajo no se abordará el tema de la intensidad, y por lo tanto la eliminación de estas palabras, no afecta la extracción de la polaridad.
  Del total de comentarios el 0,55% contiene expresiones que amplifican un comentario, y el 0,43% errores de tipeo. Por lo tanto, sumando ambos porcentajes solo 0,98% contiene palabras repetidas. A continuación, se presentan algunos ejemplos, juntos con su respectiva corrección.

"Una historia tan tan bizarra y tan gélida como el ambiente en el que se mueven los personajes sórdida en algunos momentos. No creo que lea más de esta autora",

Una historia tan bizarra y tan gélida como el ambiente en el que se mueven los personajes sórdida en algunos momentos. No creo que lea más de esta autora,

"Durante muchos muchos años fue mi libro favorito. Supongo que su lectura dejó una huella imborrable en la mente del muchacho que yo era cuando lo leí, y que contribuyó a cimentar, desde entonces, mi pasión por los libros con aquella frase rotunda con la que acaba "... porque las

estirpes condenadas a cien años de soledad no tenían una segunda oportunidad sobre la tierra"".

Durante muchos años fue mi libro favorito. Supongo que su lectura dejó una huella imborrable en la mente del muchacho que yo era cuando lo leí, y que contribuyó a cimentar, desde entonces, mi pasión por los libros con aquella frase rotunda con la que acaba "... porque las estirpes condenadas a cien años de soledad no tenían una segunda oportunidad sobre la tierra".

"Nunca un libro me habia parecido tan sencillo y bueno a la vez. Gabo es es un genio escribiendo y cuando más lo demuestra es en libros como este. Su lectura te atrapa y ya ves tú, el argumento realmente es de risa. Un auténtico mago de la narrativa".

Nunca un libro me había parecido tan sencillo y bueno a la vez. Gabo es un genio escribiendo y cuando más lo demuestra es en libros como este. Su lectura te atrapa y ya ves tú, el argumento realmente es de risa. Un auténtico mago de la narrativa.

"Se lee bien, pero es flojito flojito".

Se lee bien, pero es flojito.

"Mala mala mala con avaricia. Nada que ver con el Ocho, yo me releí ésta antes de leer su segunda parte y, aunque es necesario hacerlo para enterarte de algo, no hace que El Fuego te guste más. Algunas partes son hasta ridículas".

Mala con avaricia. Nada que ver con el Ocho, yo me releí ésta antes de leer su segunda parte y, aunque es necesario hacerlo para enterarte de algo, no hace que El Fuego te guste más. Algunas partes son hasta ridículas.

"Lo leí hace alrededor de de 6 años y fue mi primera experiencia de enganche con un libro, en mi opinión vale la pena leerlo! es una historia hermosa".

Lo leí hace alrededor de 6 años y fue mi primera experiencia de enganche con un libro, en mi opinión vale la pena leerlo! es una historia hermosa.

- **Eliminación de URL:** El proceso consistió en eliminar las URL del comentario, las cuales no se encuentran presente en el lexicón. Por otra parte, el estudio de las URL no forma parte de nuestra línea de trabajo, por lo que su eliminación no afectará el cálculo de la polaridad. Del total de comentarios, el 2,31% las contenía. A continuación se presentan algunos ejemplos, junto a su respectiva corrección.

  "El libro que siempre quise leer.
  http://juanmanuelpr.blogspot.com.es/p/al-otro-lado-del-cristal.html"
  El libro que siempre quise leer.

  "http://dimequeleer.blogspot.com.es/2012/10/cincuenta-sombras-james-el.html Aquí"
  ~~*http://dimequeleer.blogspot.com.es/2012/10/cincuenta-sombras-james-el.html Aquí*~~

  "Mi reseña: http://letrasyfotogramas.blogspot.com.es/2012/05/un-espectaculo-con-la-muerte.html"
  ~~*Mi reseña: http://letrasyfotogramas.blogspot.com.es/2012/05/un-espectaculo-con-la-muerte.html*~~

- **Eliminación de Nick de Usuarios:** El proceso consistió en eliminar la alusión a Nick de otros usuarios, los cuales no se encuentran presentes en lexicón y además al tratarse de nombre de usuario carecen de polaridad, por otra parte, el estudio de los Nick no forma parte de nuestra línea de trabajo, debido a esto su eliminación no afectará el cálculo de la polaridad. Del total de comentarios, el 1,96% las contenía. A continuación se presentan algunos ejemplos, junto a su respectiva corrección.

"Opino lo mismo que claudia y alberto. Añadiría que el libro tiene momentos durísimos que dejan a uno con la boca abierta, pero también hay pasajes muy tiernos y emotivos, como la vida misma."

Añadiría que el libro tiene momentos durísimos que dejan a uno con la boca abierta, pero también hay pasajes muy tiernos y emotivos, como la vida misma.

"Coincido con janca_lb cumple todo lo que se ha dicho de él."

cumple todo lo que se ha dicho de él.

- **Corrección de faltas de ortografía y errores de tipeo:** El proceso consistío en modificar los errores ortográficos y de tecleo propio del lenguaje natural. Del total de comentarios el 40,86% contenía dichas faltas. A continuación se presentan algunos ejemplos, junto con su respectiva corrección.

"La historia es predecible...me pareció un libro frío. En este genero hay otros mejoes."

La historia es predecible...me pareció un libro frío. En este género hay otros mejores.

"Muy regularcito, sobre todo después de haber leido los de Camila Lackberg."

Muy regularcito, sobre todo después de haber leído los de Camila Lackberg.

- **Especiales:** Son aquellos comentarios, que no pertenecían a ninguna de las categorías anteriores, pero que de igual forma se debieron modificar. Y por otra parte, se encontraron dos comentarios repetidos, a los cuales no se le realizó cambio, fuera de lo ya descrito en este documento. Del total de comentarios, el 0,66% pertenece a esta categoría. A continuación se detallan los criterios utilizados.

    - Los comentarios donde se observaron una serie de asteriscos, los cuales son agregados cuando la palabra es reconocida como grosera por la

página, se decidieron eliminar los símbolos, ya que era imposible determinar la palabra en sí.

> **“Creo que el primer libro es muy bueno. Para mí, la historia podría acabarse aquí, sin el ******* de los siguientes. Siempre puedes inventar el final a tu gusto :s.”**
>
> *Creo que el primer libro es muy bueno. Para mí, la historia podría acabarse aquí, sin el de los siguientes. Siempre puedes inventar el final a tu gusto*

- Los comentarios que, tal como se aprecia en la palabra *“perfecta”* citado en el siguiente comentario, fue descompuesta en sílabas, con el fin de darle más énfasis. Y tal como se explicó, la intensidad no forma parte de los tópicos que abordan en este trabajo, por lo que dichas palabras se modificaron.

> ***“Una novela per-fec-ta. García Márquez es tan bueno - y lo sabe - que, en la primera frase de la novela, se permite la chulería de decirte lo que va a pasar al final. Y desde ahí comienza un pasmoso viaje en el tiempo (hacia atrás, hacia adelante, incluso muchos años), entre los personajes (a los que vemos abocados a hacer algo que en realidad no quieren hacer, como si el destino lo hiciera inevitable), para volver a acabar prácticamente en el punto donde comenzó. Y no se pierde ni un momento de tensión ni de interés. Chapó, maestro. Chema.”***
>
> *Una novela perfecta. García Márquez es tan bueno - y lo sabe - que, en la primera frase de la novela, se permite la chulería de decirte lo que va a pasar al final. Y desde ahí comienza un pasmoso viaje en el tiempo (hacia atrás, hacia adelante, incluso muchos años), entre los personajes (a los que vemos abocados a hacer algo que en realidad no quieren hacer, como si el destino lo hiciera inevitable), para volver a acabar prácticamente en el punto donde comenzó. Y no se pierde ni un momento de tensión ni de interés. Chapó, maestro. Chema.*

- Los comentarios que no guardan relación con la opinión de un libro, fueron considerados como un spam y posteriormente eliminados.

  "MARKETING PURO!!! Recomiendo que lean 11.99 Euros (no recuerdo el nombre del autor en este momento) y entenderán lo que digo."

  ~~*MARKETING PURO!!! Recomiendo que lean 11.99 Euros (no recuerdo el nombre del autor en este momento) y entenderán lo que digo.*~~

- Los comentarios que guardan relación con el puntaje entregado a la novela, pero no contengan opinión sobre este fueron eliminados.

  "Nota: 9"

  ~~*Nota: 9*~~

  "Para mí, Intriga: Intriga (3,5)/Histórica (4'5): 4 Descripciones: 3´5 Acción: 3 Personajes: 3 Tema: 6 Trama: 4 Nota media : 3'9 = 4"

  ~~*Para mí, Intriga: Intriga (3,5)/Histórica (4'5): 4 Descripciones: 3´5 Acción: 3 Personajes: 3 Tema: 6 Trama: 4 Nota media : 3'9 = 4*~~

- Los comentarios que contenían el puntaje puesto por el usuario al libro, pero una opinión sobre este, se le eliminó la ponderación, para evitar cual tipo de influencia o predisposición a la hora de calcular la polaridad.

  "Es un libro entretenido y después de la primera parte "técnica" se convierte en un libro que es imposible abandonar. Me pareció muy itneresante el concepto desde la óptica de un periodista investigador. Le pondría un 8"

Es un libro entretenido y después de la primera parte "técnica" se convierte en un libro que es imposible abandonar. Me pareció muy interesante el concepto desde la óptica de un periodista investigador.

Finalmente de los 2545 comentarios, 1287 fueron limpiados, lo cual equivale a un 50,56%. Por otra parte, y tal como se explicó algunos comentarios fueron eliminados al ser considerados spam, quedando así un total de 2534 comentarios en nuestro corpus.

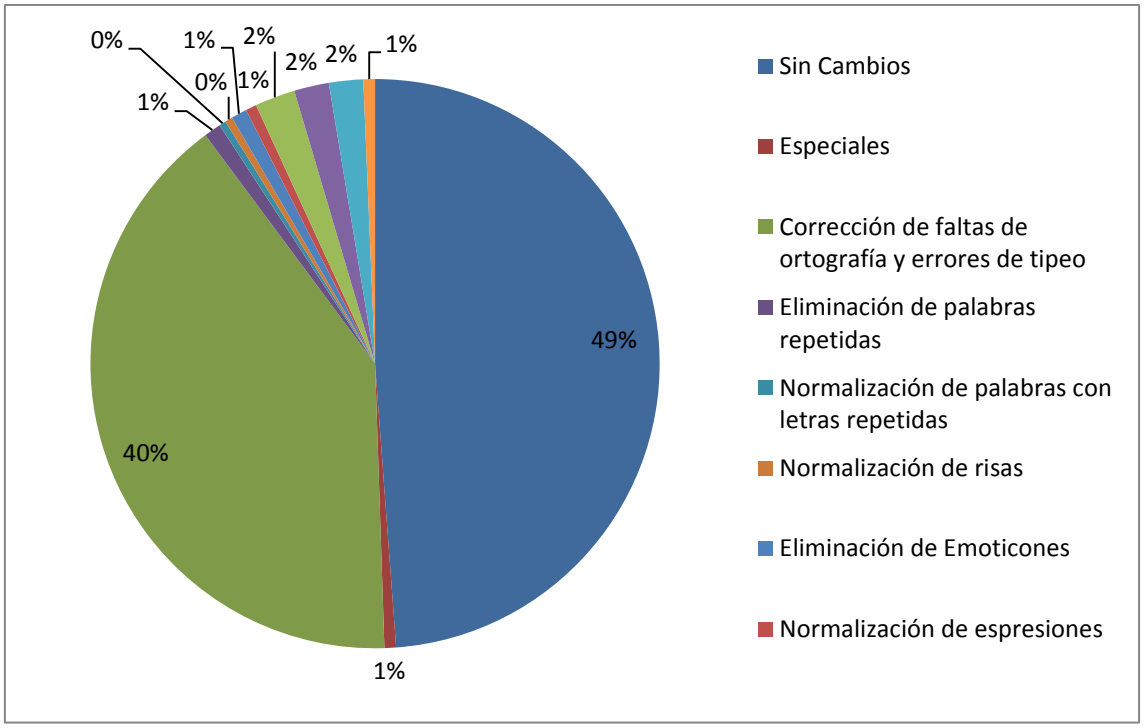


Figura 9: Gráfico de Limpieza de datos

El gráfico anterior resume el porcentaje de cambios realizados al corpus, paralelamente muestra que las mayores modificaciones realizadas fueron en el ámbito ortográfico. El detalle de las modificaciones aplicadas se encuentran en el enlace presente en **anexo 11.14.7,** junto con el corpus limpiado (ver **anexo 11.14.8**)

7.3.2 PRE PROCESAMIENTO

En esta etapa, la información lingüística, que será indispensable para llevar a cabo las siguientes etapas, por lo cual, como ya se explicó en la **sección 6.1.1**, la herramienta

GATE permite implementar una serie de recursos lingüísticos. A continuación se detalla paso a paso el desarrollo de esta etapa y los módulos utilizados.

7.3.2.1 Creación del Corpus

Primero iniciamos el programa y creamos un nuevo documento, donde se alojará el corpus. Para hacerlo, sé hace clic con el botón derecho en *Language Resources,* luego en *New*, y finalmente en *GATE Document*.

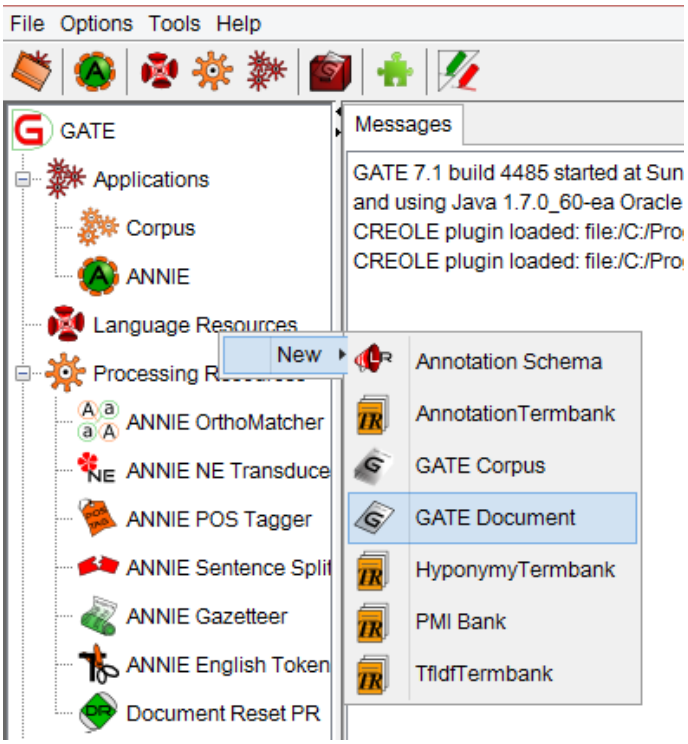


Figura 10: Crear Documento

Luego, aparecerá una ventana como la siguiente (ver **figura 11**). Y se completa los campos *Name* (Corpus Criticas de Libros), *enconding* (utf8), URL (se selecciona el archivo, donde se encuentra en corpus)

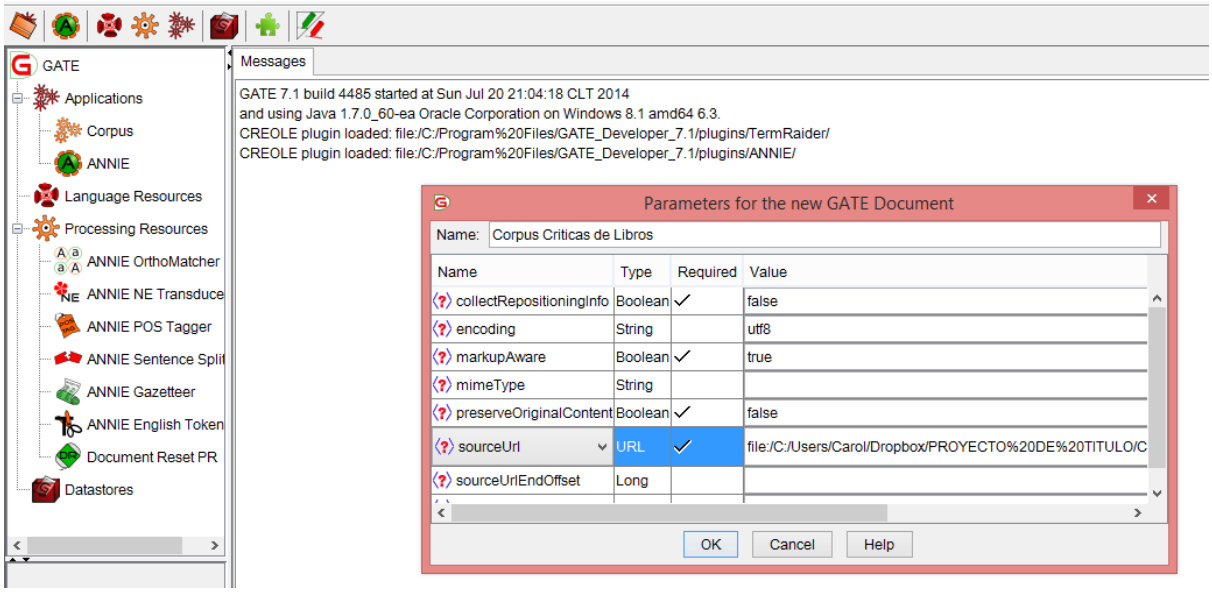


Figura 11: Creación de Documento Gate

Una vez creado el documento, se procede a crear el Corpus (ver **figura 12**), haciendo clic con el botón derecho sobre *Corpus Criticas de Libros*, para luego hacer clic en *New Corpus with this Document*, donde sobre él trabajarán las herramientas, que se relatarán más adelante.

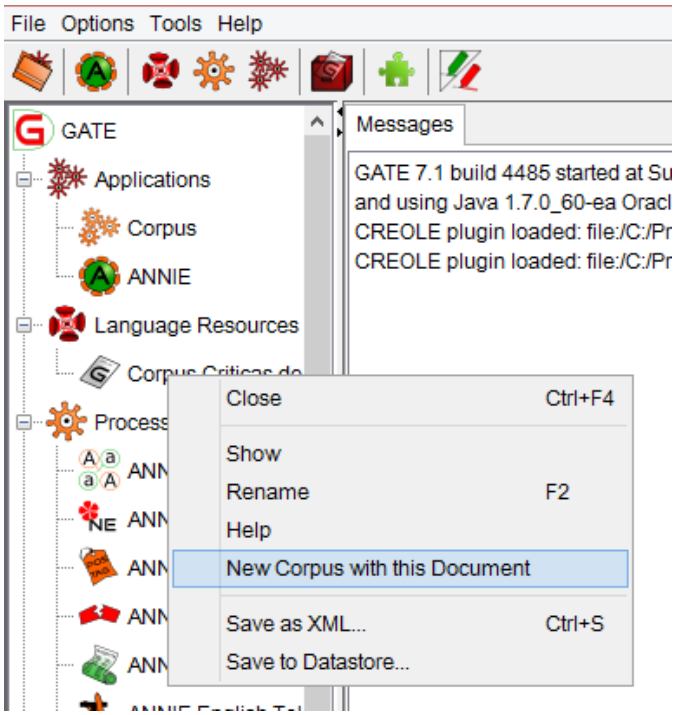


Figura 12: Crear Corpus

7.3.2.2 Creación de una nueva Aplicación

Para crear nuestra propia aplicación se hace clic primero en *Applications*, *luego Create New Application*, y finalmente en *Corpus Pipeline,* el cual permite ejecutar una aplicación en todo un corpus, arrastrando los recursos en él (Ver **figura 13**).

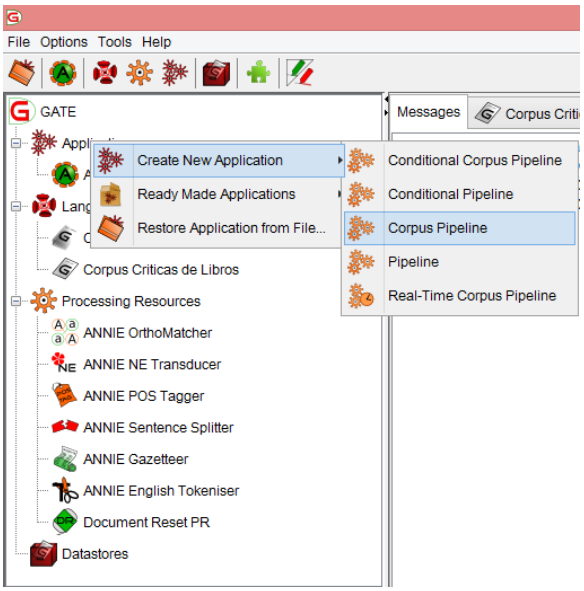


Figura 13: Creación de Corpus Pipeline

Al seleccionar nuestra aplicación, que denominamos *Pre-procesamiento*, se abre la siguiente ventana, que se presenta a continuación, donde al lado izquierdo aparecen los recursos disponibles, los cuales arrastramos al lado derecho, seleccionamos el corpus, y ejecutamos con *Run this Application*.

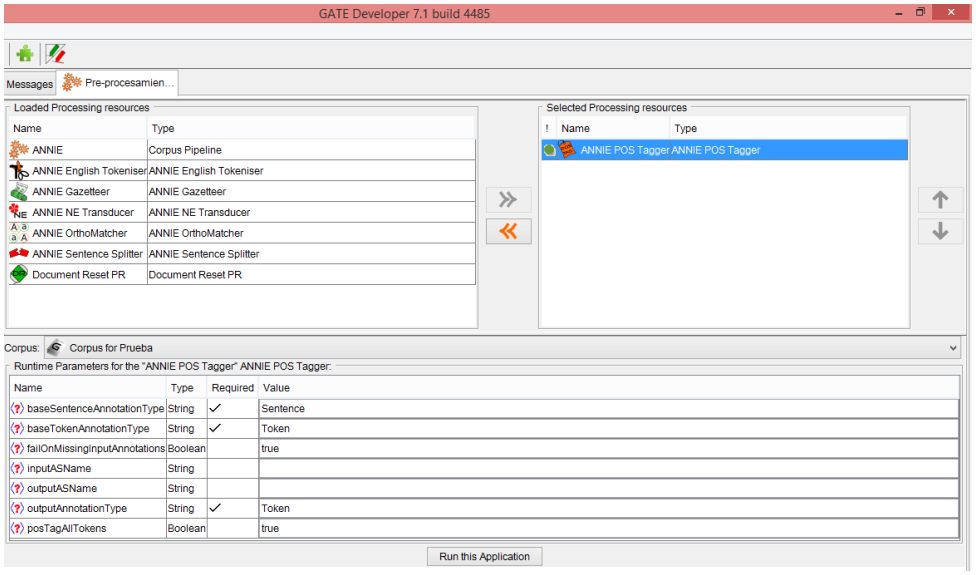


Figura 14: Pre-Procesamiento

7.3.2.3 Instalación de Plugin

Por otro lado, y tal como se mencionó en la **sección 6.1.1.** GATE, dispone de una variedad de plugins, los cuales al hacer clic en el icono +, permite instalar herramientas, que servirán en nuestro trabajo.

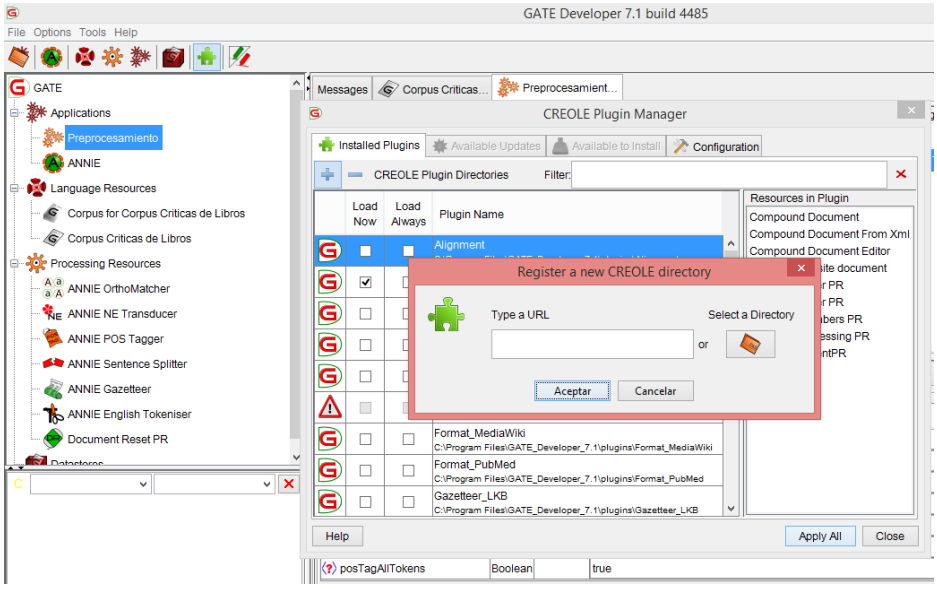


Figura 15: Instalación de Plugins

Una vez realizado el paso anterior, y para finalizar la instalación, se hace clic en *Processing Resource,* con el botón derecho, y se busca la herramienta a utilizar, en este caso seleccionaremos *Spanish POS Tagger* (ver **figura 16**), luego se le da un nombre clave (ver **figura 17**), y con ello, ya es posible la implementación como se explicó en la **sección 7.3.3.2.**

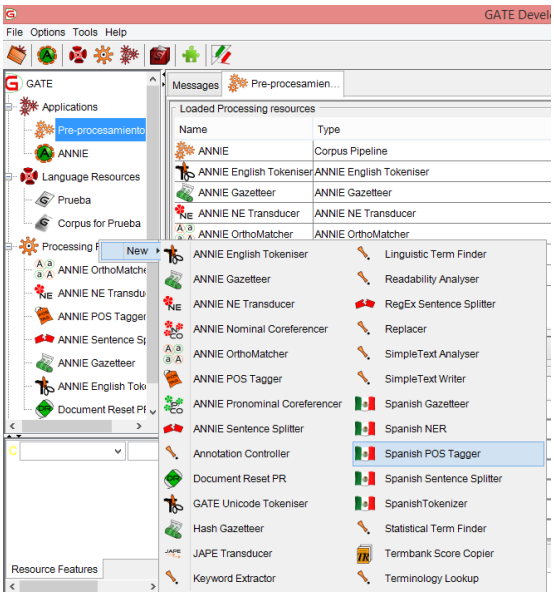


Figura 16: Instalación de herramienta

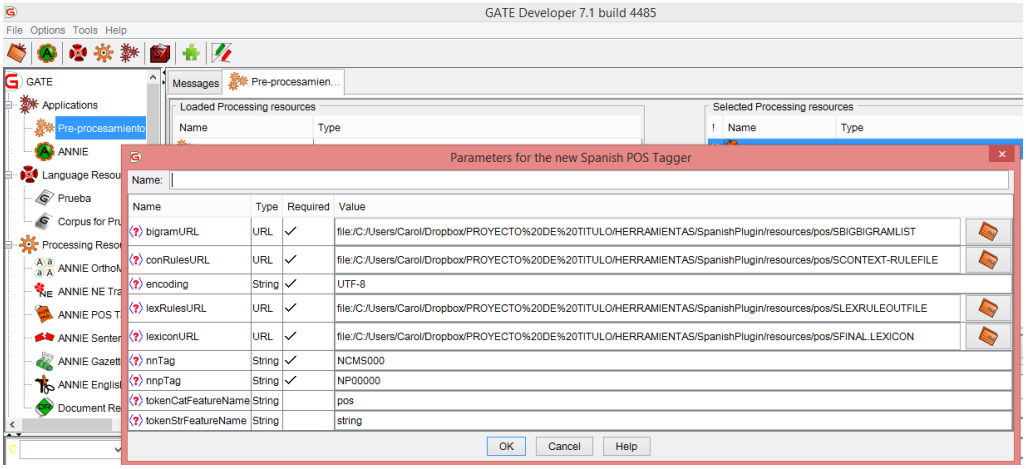


Figura 17: Nombre herramienta

7.3.2.4 Plugin Implementados.

A continuación, se presentan los plugins utilizados en la etapa de pre-procesamiento.

- *Spanish Plugin*[6]*:* Corresponde a un conjunto de herramientas para procesar textos en español. (PORRAS, 2008)
    - *Spanish Tokenizer:* Toma en cuenta una lista de abreviaturas típicas del español para poder delimitar los tokens. El resultado de aplicar este recurso a un corpus (o documento de GATE) es un conjunto de anotaciones del tipo Space Token, que identifica los espacios existentes en el texto y de tipo Token con los siguientes atributos:

[6] Se puede descargar de http://sourceforge.net/projects/nlptools-es/

- *Kind*: El tipo de token, puede ser uno de los siguientes: word, punctuation, number, url, email.
- *Length:* El número de caracteres del token.
- *Orth:* La ortografía del token, puede ser uno de los siguientes: lowercase, upperInitial, allCaps, mixedCaps
- *String:* El string que determina ese token. (PORRAS, 2008)

A continuación, en la **figura 18**, se visualiza la implementación del recurso. En el **anexo 11.1** y **11.2**, se entrega una muestra de las tablas que entrega GATE. Lo anterior, se extrajo en un XML, que se encuentra disponible en el enlace presente en el **anexo 11.14.16**.

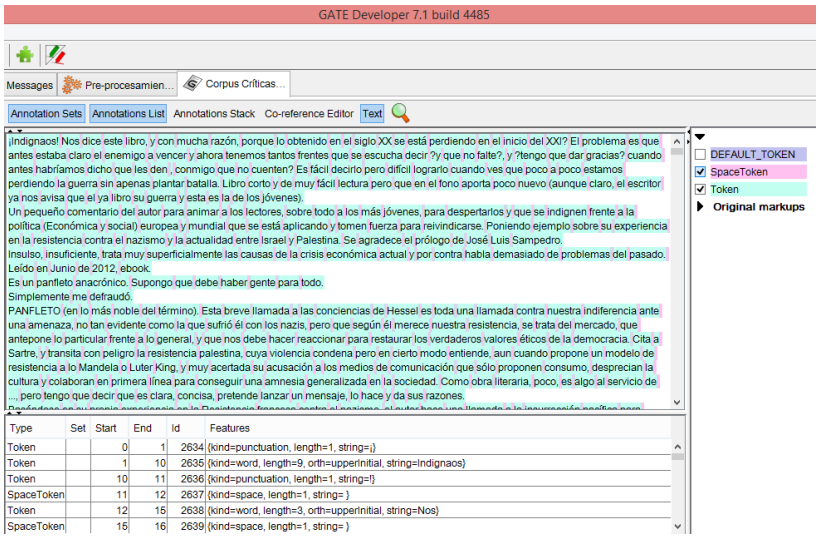


Figura 18: Ejecución del Tokenizer

En la siguiente tabla se visualiza los datos entregados al seleccionar la notación *Token*, donde la primera columna, hace referencia al *tipo*, en este caso *Token*, seguido de *Set*. Luego *Start* y *End*, indican la posición de la palabra. Le sigue el *id*, y finalmente *features*, en esta columna se entrega la clase (*kind,* puede adquirir los valores w*ord* (palabra), *punctuation* (signos puntuación)), el largo (*length*), *orth* (que puede adquirir los valores *upperInitial* (inicial mayúscula), *lowercase* (minúscula)) y la cadena analizada (*string*).

Type	Set	Start	End	Id	Features
Token		1	10	2635	{kind=word, length=9, orth=upperInitial, string=Indignaos}

Tabla 5: Salida Notación Token

En la siguiente tabla se visualiza los datos entregados al seleccionar la notación *SpaceToken*, que determina los espacios entre las palabras, donde la primera columna, hace referencia al *tipo*, en este caso *SpaceToken*, seguido de *Set*. Luego *Start* y *End*, indican la posición del espacio. Le sigue el *id*, y finalmente *features*, esta columna se entrega la clase *(kind*), el largo (*length*), y la cadena analizada (*string*) que en el ejemplo corresponde a un espacio.

Type	Set	Start	End	Id	Features
SpaceToken		11	12	2637	{kind=space, length=1, string= }

Tabla 6: Salida Notación SpaceToken

- *Spanish Sentence Splitter:* Corresponde a un segmentador de oraciones. El resultado al aplicar este recurso a un corpus, es la generación de las anotaciones *sentence, split* (PORRAS, 2008)

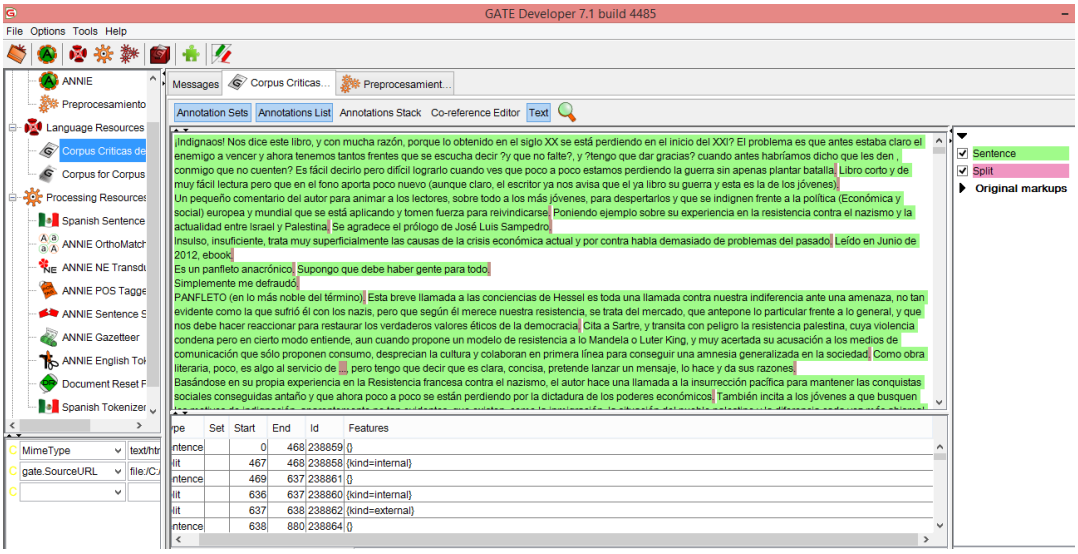


Figura 19: Ejecución de Spanish Sentence Splitter

En la **figura 19**, se visualiza la implementación del recurso, donde se observa las notaciones *Sentence* y *Split*. En el **anexo 11.3** y **11.4**, se entrega una muestra de las tablas que entrega GATE. Lo anterior, se extrajo en un XML, que se encuentra disponible en el siguiente enlace presente en el **anexo 11.14.15.**

En la siguiente tabla se visualiza los datos entregados al seleccionar la notación *Sentence*, que reconoce las oraciones en el texto, donde la primera columna, hace referencia al *tipo*, en este caso *Sentence*, seguido de *Set*. Luego *Start* y *End*, indican la posición de la oración. Le sigue el *id*, y finalmente *features*.

Type	Set	Start	End	Id	Features
Sentence		0	468	238859	{}

Tabla 7: Salida Notación Sentence

En la siguiente tabla se visualiza los datos entregados al seleccionar la notación *Split*, que reconoce los puntos presentes en el texto, donde la primera columna, hace referencia al *tipo*, en este caso *Split*, seguido de *Set*. Luego *Start* y *End*, indican la posición del punto. Le sigue el *id*, y finalmente *features*, esta columna se entrega la clase *(kind*) que adquirir los valores *internal* (para los puntos seguidos, puntos suspensivos. e.d. los puntos inmersos dentro de un párrafos) o *external* (para los puntos que dividen un párrafo)

Type	Set	Start	End	Id	Features
Split		467	468	238858	{kind=internal}

Tabla 8: Salida de Notación Split

- *Spanish POS Tagger:* Corresponde a un etiquetador morfosintáctico, (PORRAS, 2008) que recibe un corpus previamente tokenizado, segmentado por oraciones y regresa como resultado el texto etiquetado con categoría gramatical, es decir, es necesario haber implementado el *Spanish Tokenizer* y *Spanish Sentence Splitter*, en el corpus. Como resultado al aplicar este recurso a un corpus, es la generación de la una anotación de tipo Token en el atributo *pos*.

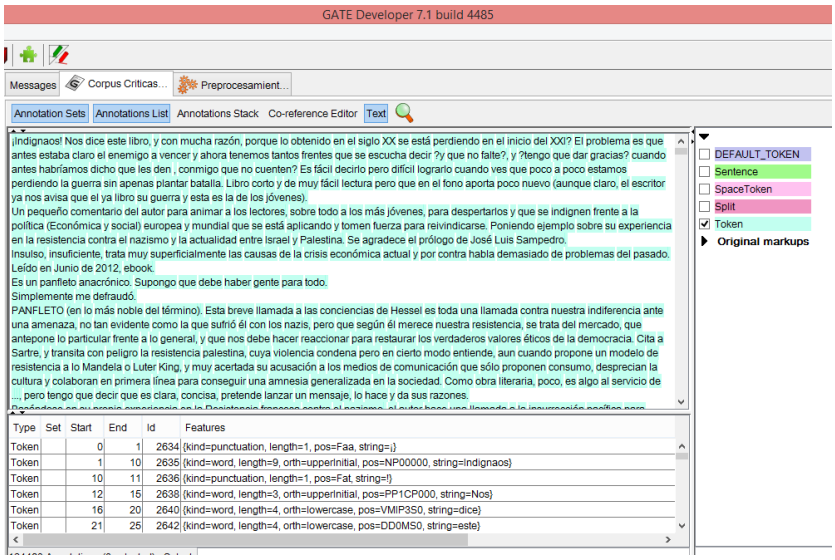


Figura 20: Ejecución de Spanish POS Tagger

En la **figura 20**, se visualiza la implementación del recurso, donde se observa las notaciones Token. En el **anexo 11.5**, se entrega una muestra de las tablas que

entrega GATE. Y el **anexo 11.6** la lista de etiquetas utilizadas. Lo anterior, se extrajo en un XML, que se encuentra disponible en el siguiente enlace presente en el **anexo 11.14.14.**

En la siguiente tabla se visualiza los datos entregados al seleccionar la notación *Token*, donde la primera columna, hace referencia al tipo, en este caso *Token*, seguido de *Set*. Luego *Start* y *End*, indican la posición de la palabra. Le sigue el *id*, y finalmente *features*, esta columna se entrega la clase (*kind,* puede adquirir los valores w*ord* (palabra), *punctuation* (signos puntuación)), el largo (l*ength*), *orth* (que puede adquirir los valores *upperInitial* (inicial mayúscula), *lowercase* (minúscula)), *pos*, que corresponde a la etiqueta que introduce el Spanish POS tagger, en el ejemplo presentado se entrega el valor de *NP00000*, que de acuerdo a la tabla presentada en el **anexo 11.6**, la primera letra representa la categoría, en este caso N hace referencia a los *nombres*, la segunda letra P, corresponde al tipo *propio*, el conjunto de ceros implica que no se utilizó una clasificación semántica (clasificación por significado) y finalmente la cadena analizada (*string*).

Type	Set	Start	End	Id	Features
Token		1	10	2635	{kind=word, length=9, orth=upperInitial, pos=NP00000, string=Indignaos}

Tabla 9: Salida de la Notación Token con el Atributo POS

- ANNIE[7]: Es un sistema de extracción de información (A Nearly-New Information Extraction System), que comprenden un conjunto de módulos integrados por un *tokenizer*, un *gazetteer*, una *sentence splitter*, un *part of speech tagger*, un *transductor de entidades con nombre* y un *etiquetador correferencia*. (CUNNINGHAM, MAYNARD, BONTCHEVA, & TABLAN, 2002)

    - Document Reset: Permite restablecer el documento a su estado original, mediante la eliminación de todos los conjuntos de anotaciones. (CUNNINGHAM H. , y otros, 2012) El documento reseteado se encuentra en el siguiente enlace presente en el **anexo 11.14.10**.

    - ANNIE Part of Speech Tagger: Produce una etiqueta del texto como una anotación en cada palabra o símbolo. La lista de etiquetas utilizadas se presenta en **anexo 11.7.** Para su implementación, requiere la ejecución previa de *Spanish Tokenizer* y *Spanish Sentence Splitter*.

[7] Se encuentra por defecto en el GATE

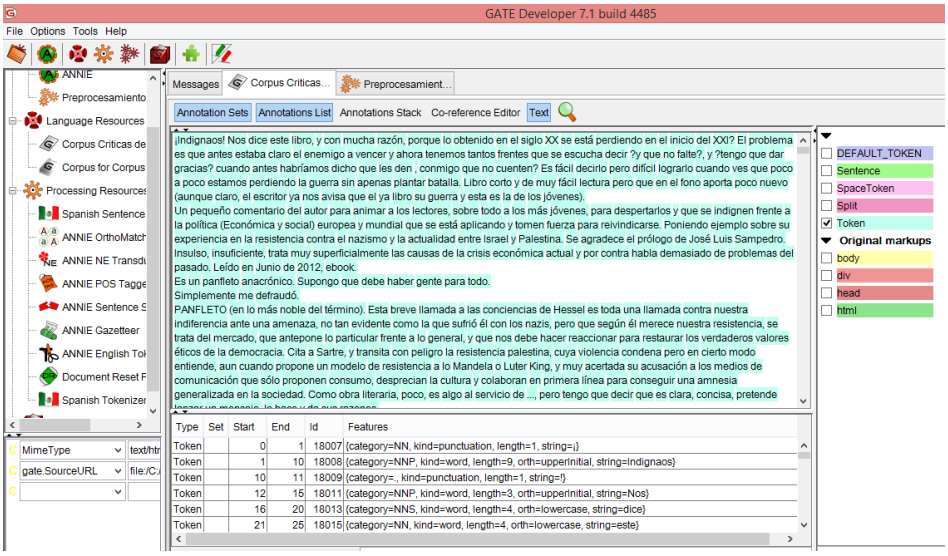


Figura 21: ANNIE Pos Tagger

En la figura 19, se visualiza la implementación del recurso, donde se observa las notaciones *Token*. En el **anexo 11.8**, se entrega una muestra de las tablas que entrega GATE. De lo anterior, se extrajo en un XML, que se encuentra disponible en el siguiente enlace presente en el **anexo 11.14.9**.

En la siguiente tabla se visualiza los datos entregados al seleccionar la notación Token, que reconocer y etiqueta las palabras, donde la primera columna, hace referencia al tipo, en este caso Token, seguido de *Set*. Luego *Start* y *End*, indican la posición de la palabra. Le sigue el *id*, y finalmente *features*, esta columna se entrega la categoría, en este caso corresponde NNP, que si verificamos en el **anexo 11.7** hace referencia a *Nombre propio – singular*, luego tenemos la clase (*kind,* puede adquirir los valores w*ord* (palabra), *punctuation* (signos puntuación)), el largo (*length*), *orth* (que puede adquirir los valores *upperInitial* (inicial mayúscula), *lowercase* (minúscula)) y la cadena analizada (*string*).

Type	Set	Start	End	Id	Features
Token		1	10	18008	{category=NNP, kind=word, length=9, orth=upperInitial, string=Indignaos}

Tabla 10: Salida de la Notación Token al Aplicar ANNIE Pos Tagger

- *Readability Tools*[8]*:* Corresponde a una herramienta, utilizada para obtener datos estadísticos, del corpus, por ejemplo, la frecuencia de determinadas palabras. El sistema *utiliza Plain English thesaurus*, también incluye la búsqueda de la terminología por medio *ISO TC 3*, y proporciona un sistema de gestión de terminología (TMS) a través de la norma *ISO 16642*. (READABILITY TOOLS) Por otro lado, se integra perfectamente a GATE, y a ANNIE, por lo que requiere de la ejecución *ANNIE POS tagger*, para su normal funcionamiento.

  - *Readability Analyser:* Este plugin guarda el número de palabras, sílabas, frases, personajes y palabras polisílabas contenidas dentro de un documento. Estos valores se utilizan para el cálculo de fórmulas de legibilidad, como la facilidad de lectura de *Flesch, Fórmula Kincaid, SMOG, ARI y Fog Index*. (READABILITY TOOLS)

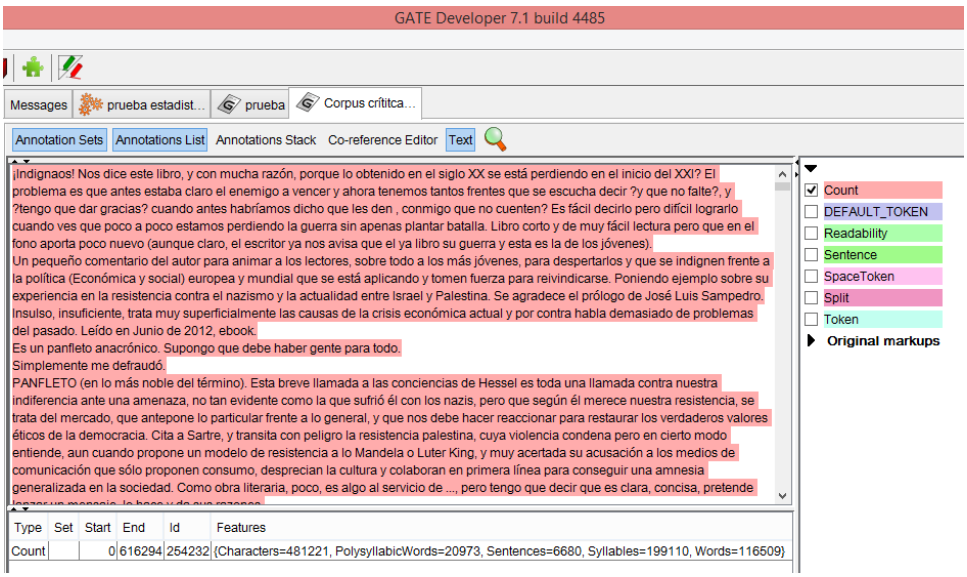


Figura 22: Notación Count

Type	Set	Start	End	Id	Features
Count		0	616294	254232	{Characters=481221, PolysyllabicWords=20973, Sentences=6680, Syllables=199110, Words=116509}

Tabla 11: Salida de la Notación Count del Statical Term Finder

En la **figura 22**, se visualiza la implementación del recurso, donde se observa las notaciones *count*, quien estrega la siguiente información: el corpus se encuentra conformado por un total de 481.221 caracteres, 20.973 palabras

[8] Se puede descargar de http://www.cs.surrey.ac.uk/BIMA/Projects/LIRICS/liricsSoftware.html

polisílabas, 6.680 sentencias, 199.110 sílabas, 116.509 palabras, resumido en la **tabla 11**. En el **anexo 11.9**, se entrega un la tabla que muestra GATE.

Type	Set	Start	End	Id	Features
Readability		0	616294	254231	{ARI=6.74460327690241, Execution=1, FOG=14.177060609488525, Flesch=44.55332901601679, Kincaid=11.377979089058577, SMOG=13.25158129617029}

Tabla 12: Salida de la Notación Readability del Statical Term Finder

Por otro lado, al seleccionar la notación *Readability* (ver **figura 23, tabla 12**), obtenemos la siguiente información: el *Flesch* (corresponde a la facilidad de comprensión de un documento) es de 44.55332901601679. De acuerdo a la escala presentada en el **anexo 11.10**, el corpus es fácilmente entendido por un estudiante de 13 años en adelante. Luego tenemos la *fórmula de Kincaid*, (traduce la puntuación de Flesch, a una puntuación de 0-100, la calificación estadunidense en las instituciones educativas) de 11.377979089058577.

Luego, tenemos el *SMOG* (mide de la legibilidad que estima los años de educación necesarios para comprender una pieza de escritura) de 13.25158129617029. Seguido de *ARI* (mide la compresibilidad de un texto, se basa en un factor de caracteres por palabra) es de 6.74460327690241. Y finalmente *FOG* (El índice calcula los años de educación formal necesarios para comprender el texto en una primera lectura.) es de 14.177060609488525. Las ecuaciones de cada una de los indicadores se presentan en el **anexo 11.10**, junto a su respectiva tabla.

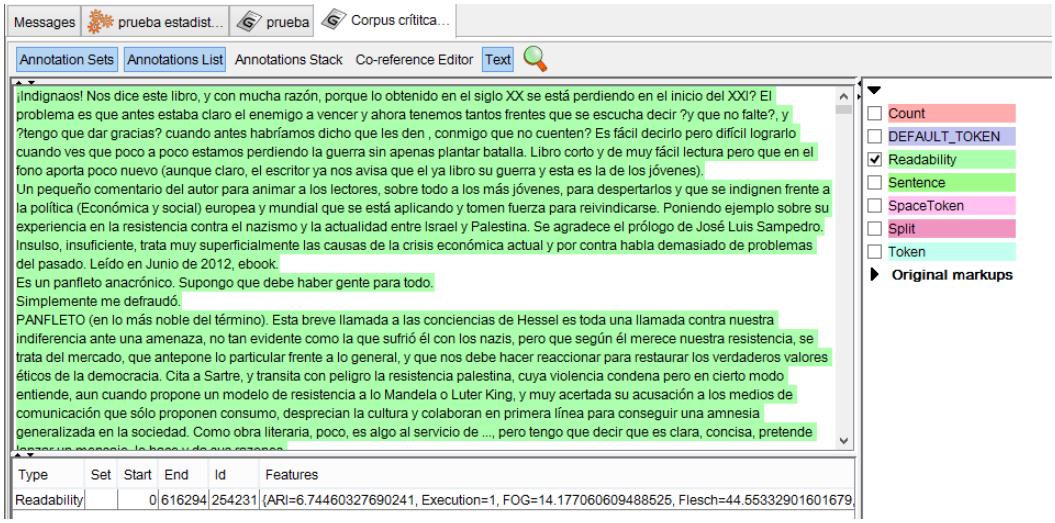


Figura 23: Notación Readability

Por último, se extrajo la información anterior en un XML, que se encuentra disponible en el siguiente enlace presente en el **anexo 11.14.13**

- *Linguistic Term Finder:* Determina todos los nombres compuestos en el documento con la parte de las anotaciones de voz creadas por el etiquetador ANNIE POS. (READABILITY TOOLS)

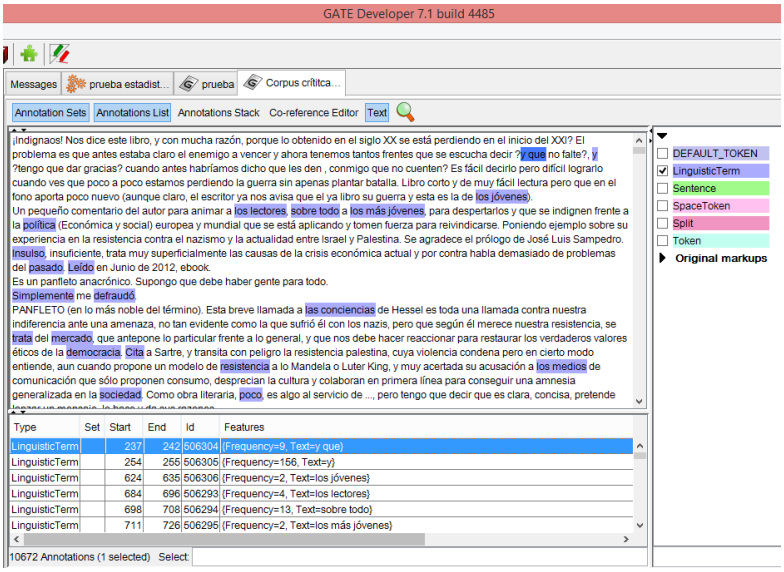


Figura 24: Linguistic Term Finder

En la **figura 24**, se visualiza la implementación del recurso, donde se observa la notación *LinguisticTerm,* el cual entrega que la palabra más frecuente corresponde a "*que"* con 191 apariciones. En el **anexo 11.11** entrega una muestra de las tablas que entrega GATE. Lo anterior, se extrajo en un XML,

que se encuentra disponible en el siguiente enlace presente en el **anexo 11.14.12.**

En la siguiente tabla se visualiza los datos entregados al seleccionar la notación *LinguisticTerm*, donde la primera columna, hace referencia al tipo, en este caso *LinguisticTerm* seguido de *Set*. Luego *Start* y *End*, indican la posición de la palabra. Le sigue el *id*, y finalmente *features*, esta columna entrega frecuencia de aparición (*frequency*) y la cadena analizada (*Text*).

Type	Set	Start	End	Id	Features
LinguisticTerm		549335	549342	507063	{Frequency=10, Text=acuerdo}

Tabla 13: Salida de la Notación Linguisticterm

El XML resultante luego de aplicar *Spanish plugin, Readability tools*, y al finalizar el pre-procesamiento, se encuentran almacenada en los siguientes enlaces presentes en los **anexos 11.14.17, 11.14.11, 11.14.18.**

En la **figura 25** se visualiza requisitos de cada plugin para su correcto funcionamiento, por ejemplo *Readability tools* necesita que previamente se haya ejecutado *Annie part of speech tagger*.

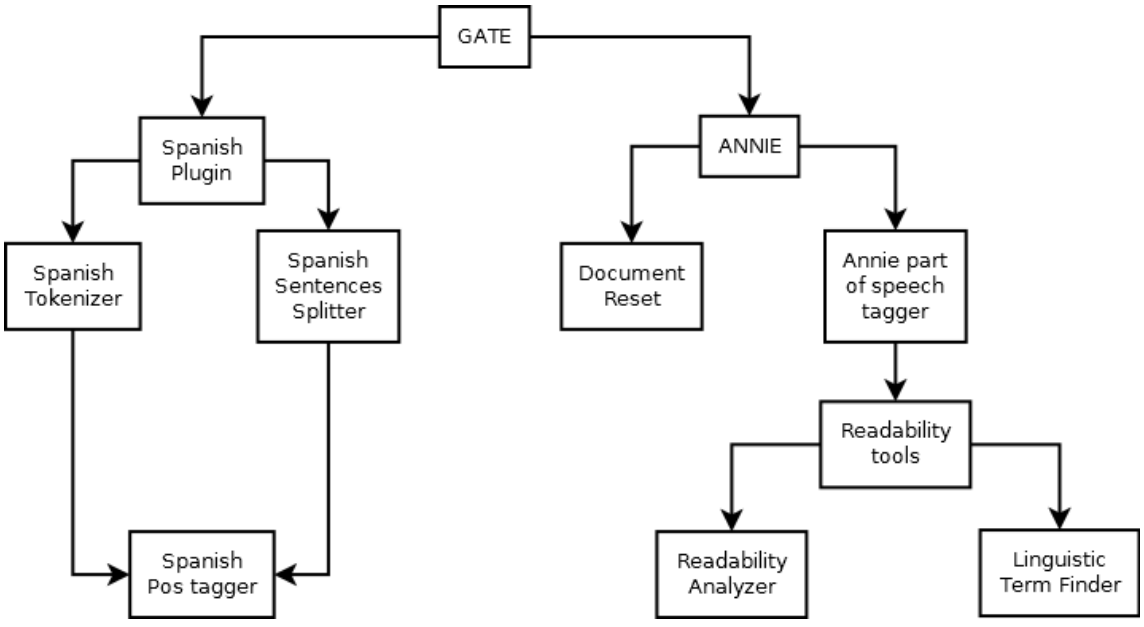


Figura 25: Plugin Implementados

7.4 LEXICÓN

A continuación se detalla el procedimiento para la creación del lexicón, el cual se basó en la fusión de cuatro ya existentes, que se observan en la **figura 26**. Para generarlo se realizaron las siguientes etapas: nivelación de métricas, eliminación de palabras repetidas, compresión de palabras. Cada una de estas fases se explica a continuación.

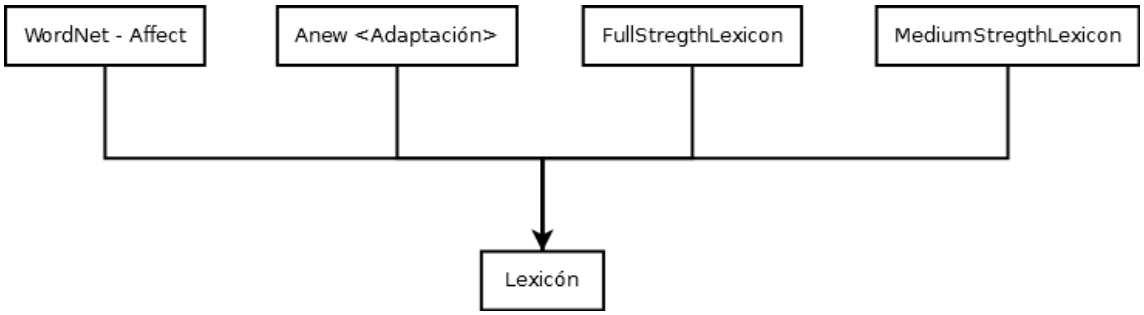


Figura 26: Generación de Lexicón

7.4.1 GENERACIÓN DE LEXICÓN

Para la creación del lexicón que se utilizó en este proyecto, primero se unificó las métricas que utilizaba cada uno, de acuerdo a nuestro caso de estudio, la polaridad se calculó bajo las siguientes categorías, positiva (pos), ambigua (amb) y negativa (neg).

Por lo tanto, para unificar las métricas fue necesario realizar las siguientes conversiones:

- En el lexicón de Redondo et al. (Redondo, 2007) descrito en el **capítulo 6** de este proyecto, se consideró solamente la dimensión *valencia*, donde si el índice es mayor igual a 1 y menor a 4 se categoriza como *neg*, en cambio si el índice es mayor igual a 4 y menor a 6 se categoriza como *amb*. Finalmente si es mayor 6 y menor igual a 9 como *pos.*
- En los lexicones *fullStrengthLexicon, mediumStrengthLexicon* (PÉREZ-ROSAS, BANEA, & MIHALCEA), no se le realizó ningún cambio, debido a que ya se encontraban categorizados con la nomenclatura correspondiente. Cabe señalar, que estos lexicones no presentaban palabras etiquetadas como ambiguas, además algunas palabras, poseían una segunda polaridad, que fue obtenida a través de un sondeo con expertos.
- Finalmente en el lexicón de Valitutti et al. (VALITUTTI & STRAPPARAVA, 2004), se basó en lo entregado por Strapparada y Valitutti (STRAPPARAVA &

VALITUTTIi), quienes presentaban la siguiente conversión, para emociones que se caracterizan por transmitir placer, como ejemplo *alegría* o *entusiasmo*, se etiquetaron como positivas (*pos*). En cambio, las emociones que transmiten dolor, como es el caso de la *tristeza, el disgusto* o la *ira,* se etiquetaron como negativas (*neg*). Finalmente como ambiguas (*amb*) aquellas que dependen del contexto semántico como *sorpresa* o *miedo*.

Las conversiones realizadas, se resumen en la siguiente tabla.

Lexicón	**Positivo**	**Negativo**	**Ambiguo**
WordNet-Affect	Alegría	Ira	Sorpresa
	Entusiasmo	Tristeza	Miedo
		Disgusto	
Adaptación ANEW	[6,9]	[1,4[	[4,6[

Tabla 14: Conversiones Lexicón

El lexicón inicial y el generado, se presentan en los siguientes enlaces presentes en los **anexos 11.14.19, 11.14.20**, respectivamente.

7.4.2 ELIMINACIÓN DE PALABRAS REPETIDAS

Como el lexicón a utilizar corresponde a la fusión de lexicones ya existentes, es natural encontrar palabras repetidas entre los lexicones o dentro del mismo, debido a que el lexicón *WordNet-Affect,* proviene de un lexicón en inglés que se tradujo al español, un ejemplo de esto, es el caso de la palabra "*admirable*", "*commendable*" que en español ambos significan "*admiración*".

Con el objetivo de suprimir las palabras repetidas se consideraron los siguientes criterios:

1. Si las palabras repetidas poseen la misma polaridad se procede a mantener solamente una.
2. En cambio, si poseen polaridades distintas, se procede a conservar una, etiquetada como *ambigua.*
3. En el caso de la palabra posea una polaridad entregada por expertos, se procede a considerar esta como la polaridad de la palabra dentro del lexicón que se encuentra inversa. Y se realiza la comparación explicada en los puntos anteriores.

El resultado de este procedimiento genera un lexicón de 5.054 palabras polarizadas, donde 2.034 corresponden a palabras negativas, etiquetadas con *-1*. Por otro lado, 1746 a palabras positivas, etiquetas con *1*. Finalmente 473 palabras ambiguas, etiquetadas con *0*.

El software utilizado junto con el lexicón generado está almacenado en los enlaces presentes en los **anexos 11.14.21, 11.14.22** respectivamente.

7.4.3 COMPRESIÓN DE PALABRAS

Como se mencionó en la sección anterior el lexicón *WordNet-Affect* está formado por palabras originarias del idioma ingles traducidas al español, por lo que existen varias de estas expresiones que son inusuales de encontrar en un comentario, como es el caso de "*de forma vomitiva*". Una vez identificadas estas palabras, se procedió a comprimirlas, es decir, a transformar dicha expresión en una sola palabra, tal que no se comprometiera su significado original, por ejemplo "*a carcajadas*" se redujo por "*carajadas*".

Finalmente del total de las 5.054, 199 formaban parte de este grupo, donde solamente 76 pudieron ser comprimidas. En las restantes 122 no se logró llevar dicho cometido, como fue el caso de "*amor adolecente*", ya que no se consiguió encontrar una palabra que englobara dicho significado sin alterar su acepción original, por lo que se procedió extraerlas del lexicón, con el objetivo de conservar solo las palabras conformadas por un solo término, dejando un total de 4.862 palabras polarizadas, donde 1.673 palabras eran positivas correspondiente al 34,41%; 2.750 negativas equivalente al 56,56% y finalmente 439 palabras ambiguas correspondiente al 9,03%.

Lo anterior se ilustra en el siguiente gráfico.

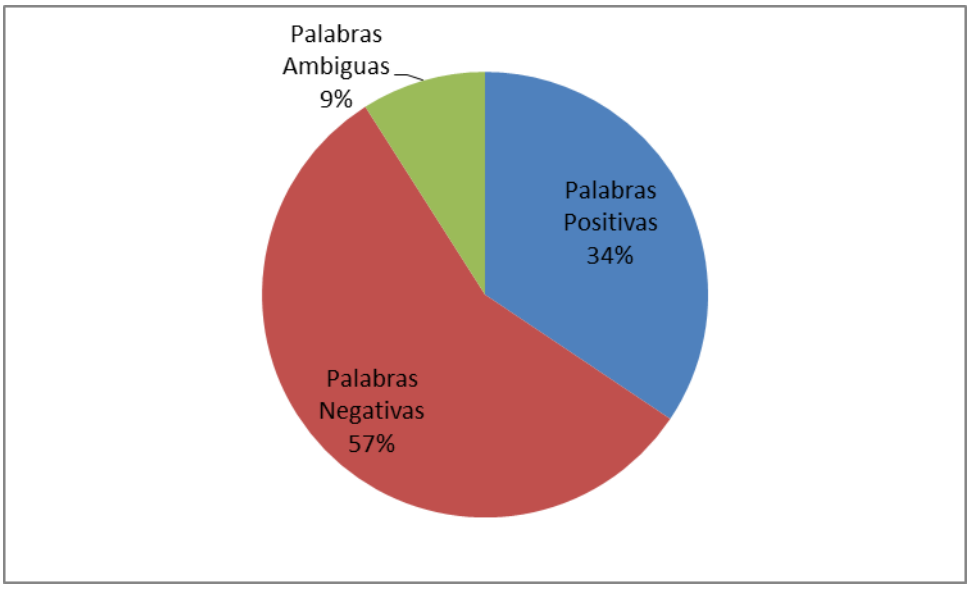


Figura 27: Distribución de Lexicón

El lexicón generado, junto al conjunto de palabras identificadas y comprimidas se encuentra en un archivo Excel almacenado en el siguiente enlace presente en el **anexo 11.14.23** En el **anexo 11.15** se encuentra la lista con la palabras modificadas en esta sección.

7.4.4 INCORPORACIÓN STEMMING

Una vez constituido el total de palabras que conforman el lexicón, se procede a calcular el stemming de cada una de ellas, que consistío en buscar la raíz (stem) de la palabra, que perteneciera a la misma familia semántica a través de la remoción de sufijos. (FERNÁNDEZ, 2013) Por ejemplo "*abandonadamente*" su stemming corresponde a "*abandon*". El objetivo de este procedimiento, es si se diera el caso, en que una palabra no se encuentre directamente en el lexicón, se procede a compararla con el stemming. Y así calcular su polaridad.

El cálculo del stemming se realizó con una librería llamada libstemmer[9] programada en java, basando en el algoritmo de Porter (PORTER, 1980), que consiste en la definición de una serie de reglas dependiendo del idioma, donde si la palabra estudiada tiene por sufijo S_1 y la raíz que lo precede satisface una condición previamente definida, entonces el sufijo S es reemplazado por S_2.

Matemáticamente se define así:

[9] Se puede descargar en http://snowball.tartarus.org/download.php

$$(condición)\ S_1 \rightarrow S_2$$

Ecuación 1: Algoritmo de Porter

El lexicón generado, se encuentra en un archivo Excel almacenado en el siguiente enlace presente en el **anexo 11.14.24.**

7.5 CLASIFICACIÓN AUTOMÁTICA

En esta etapa se determinó la polaridad de las opiniones del corpus.

En primer lugar, se procedió a identificar y eliminar los signos de puntuación y stop-words presentes en el comentario. Donde, un stop-words, corresponde a una palabra que su frecuencia de uso es por sobre la media, o cuyo significado es neutro en relación al ambiente en el que está inmerso, como lo son: los artículos, pronombres, preposiciones, etc. (FERNÁNDEZ, 2013) (RAJARAMAN & ULLMAN). Para identificar los stop-words, se utilizó la librería *Lucene*[10]*,* una API de código abierto, el cual se implementó en JAVA.

A continuación se inicia el proceso de calcular la polaridad de los comentarios, a través de las siguientes etapas:

1- Primero se busca la palabra en cuestión directamente en el lexicón, de encontrarse se etiqueta con su respectiva polaridad.
2- En caso contrario, se procede a calcular el stemming de la palabra y se compara con los stemming presentes en el lexicón. De encontrase, se etiqueta con su respectiva polaridad. En caso de que en el lexicón esté presente el mismo stemming para dos palabras con distintas polaridad, se selecciona el primero en la lista.
3- Finalmente si el stemming de la palabra no se encuentra en el lexicón, se etiqueta como *desconocida*.

La determinación final de la polaridad de un cometario, se calculó de la siguiente manera, se contabilizó el total de palabras positivas con *1*, las negativas con *-1*, y las ambiguas y desconocidas con *0*. Finalmente se realiza una sumatoria total del peso de las polaridades.

Como se ilustra en el siguiente comentario del libro “Si tú me dices ven lo dejo todo…pero dime ven”.

[10] Se puede descargar en http://lucene.apache.org/

De fácil lectura, ameno, penetrante,...ESPECIAL. Te hace pensar.

Luego de quitar los signos de puntuación y stop-words, queda de la siguiente manera.

fácil lectura ameno penetrante especial hace pensar

Posteriormente se identificó la polaridad de la palabra presente en el lexicón, donde las palabras *fácil, ameno, especial* son positivas. Por otro lado, *penetrante* corresponde a una palabra negativa. Ambiguas no se identificó ninguna. Y finalmente como palabras desconocidas quedaron *lectura, hace, pensar.*

Como ya se mencionó a las palabras positivas se les asigna el valor *1*, negativas *-1* y a ambiguas y desconocidas *0*. Por lo que al sumar, las tres palabras positivas, la palabra negativa y las tres palabras desconocidas. Esto es igual, a decir 1+1+1-1+0+0+0=2. Por lo tanto, al ser mayor a 1 indica que el comentario es *positivo*.

Finalmente de los 2.534 comentarios polarizados, 1.939 corresponden a comentarios positivos equivalente a 76,52%. Por otro lado, se identificaron 320 comentarios negativos, equivalente al 12,63. Finalmente se hallaron 275 comentarios ambiguos que equivalen al 10,85%

Lo anterior, se ilustra en la siguiente gráfica.

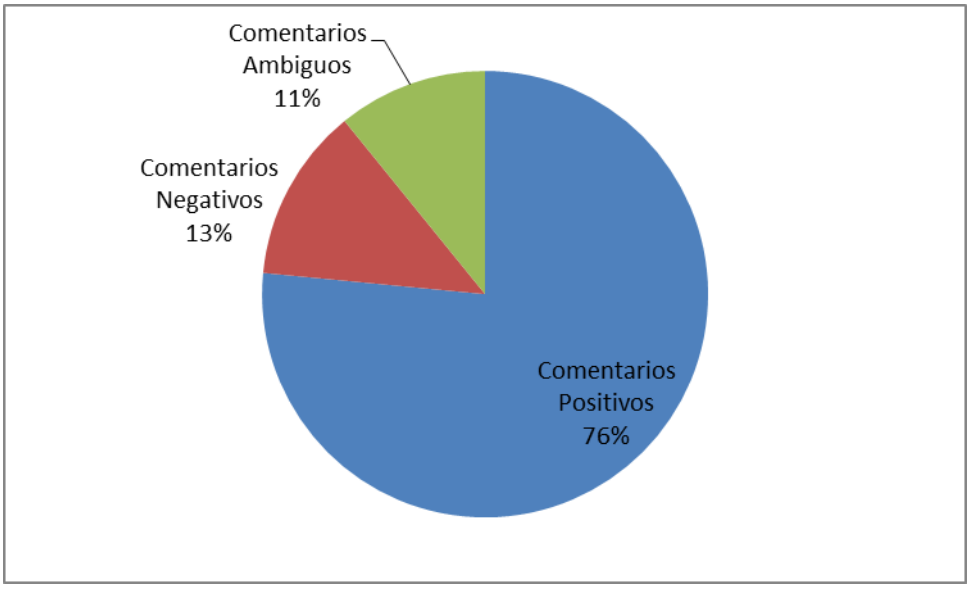


Figura 28: Distribución de Comentarios Polarizados

Recapitulando, tenemos que las palabras que no se pudieron encontrar en el lexicón, se etiquetaron como *desconocidas* que corresponden a un total de 5.513 palabras.

Por otro lado, el software aplicado en el procedimiento, junto con los comentarios polarizados, se encuentran almacenado en los siguientes enlaces, presentes en el **anexo 11.15.25, 11.15.26** respectivamente. Para calcular la polaridad de un comentario en particular y observar el detalle de este, se utilizó un programa almacenado en el enlace presente en el **anexo 11.15.27**

Como complemento al experimento las palabras desconocidas, se les aplicó el *Spanish Gazetter,* que corresponde a un diccionario que contiene nombres de entidades tales como ciudades, organizaciones, días de la semana, etc. (CUNNINGHAM H. , y otros, 2012) Con el objetivo de identificarlas, ya que al tratarse de sustantivos, podrían ser etiquetadas como palabras neutras, es decir, que carecen de polaridad. El cual se encuentra presente en el *Spanish Plugin (***ver sección 7.3.2.4***)*. Para la utilización del Spanish Gazetter, previamente se debe haber ejecutado el *Spanish Tokenizer, Spanish Sentence Splitter y Spanish POS Tagger.*

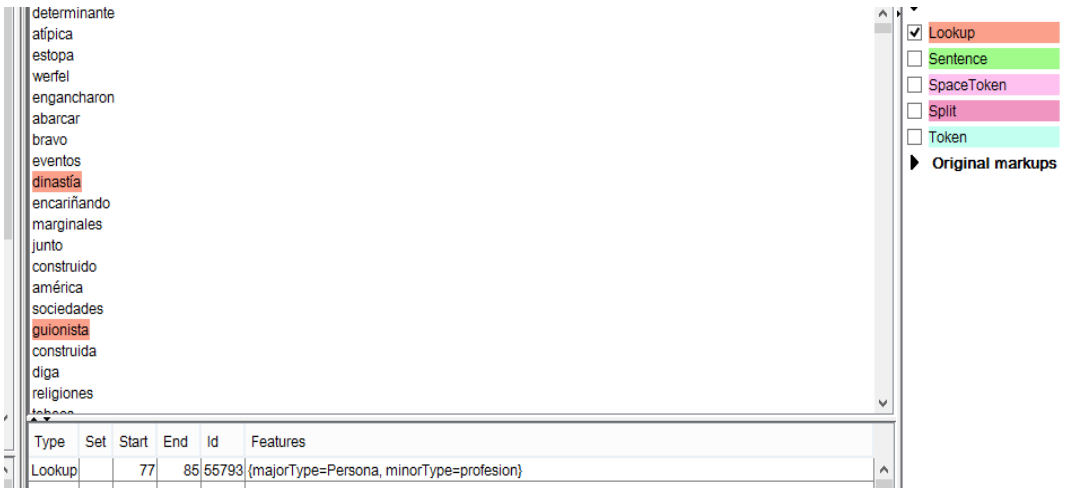


Figura 29: Notación Lookup

En la **figura 29**, se observa la notación *Lookup,* el cual entrega la clasificación que obtiene dicha palabra. Tal como se puede observar, las palabras *dinastía* y *guionista,* destacadas en la imagen, fueron clasificadas de la siguiente manera, *tipo principal* (*majorType*) *persona*, y *sub-tipo* (*miniorType*) *profesión.*

Lo anterior, se resume en la siguiente tabla, donde la primera columna, hace referencia al tipo, en este caso *Lookup* seguido de *Set*. Luego *Start* y *End*, indican la posición de la palabra. Le sigue el *id*, y finalmente *features*. En el **anexo 11.16** se entrega las tablas que proporcionadas por GATE.

Type	Set	Start	End	Id	Features
Lookup		77	85	55793	{majorType=Persona, minorType=profesion}

Tabla 15: Salida de la Notación Lookup

De la aplicación del Gazette, se lograron identificar 238 palabras, equivalente al 4,38%. Donde, 44% de estas pertenecían a la categoría *persona,* seguidas del 11% *a organización.* Lo anterior, se ilustra en la siguiente **figura 30**.

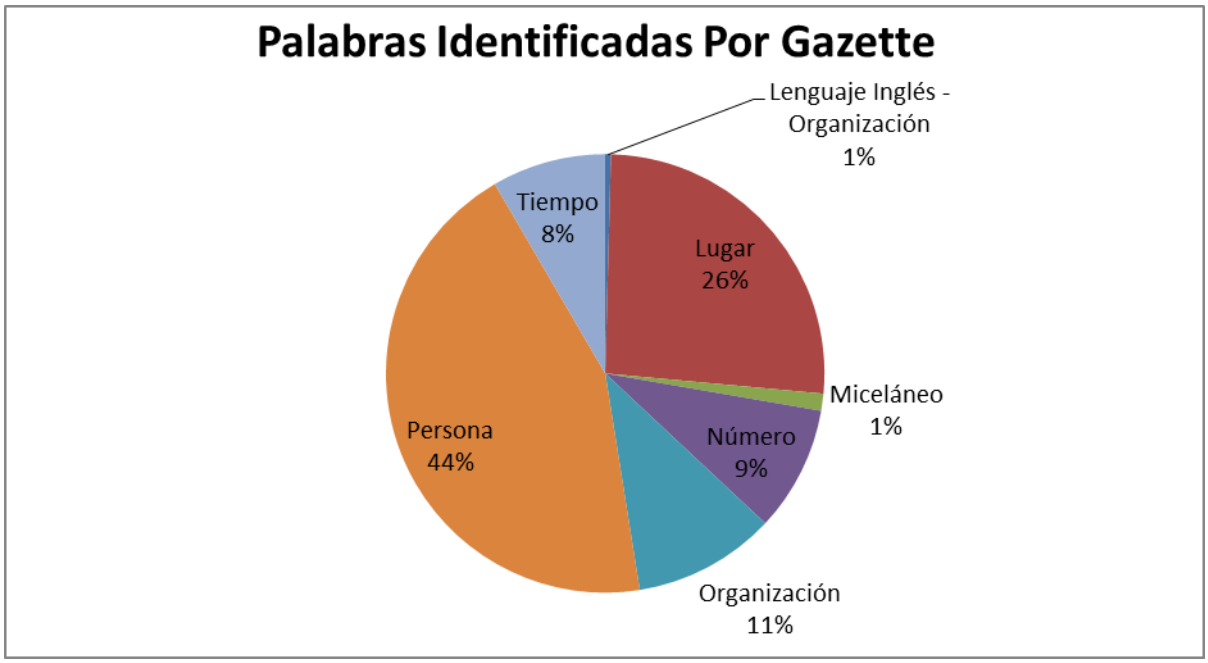


Figura 30: Palabras Identificadas por Gazette

La lista de las palabras desconocidas, junto a las palabras identificadas por *Gazette*, se encuentran en un archivo Excel almacenado en el siguiente enlace, presente en el **anexo 11.14.28.**

8 EXPERIMENTACIÓN

En el siguiente capítulo se presenta, las métricas con la que se evaluaron los resultados, junto con los criterios que se utilizaron para determinar la muestra representativa con la que se compararon los resultados obtenidos.

8.1 MÉTRICAS DE EVALUACIÓN

Las métricas que evaluaron nuestra clasificación serán la *exactitud*, *precisión*, *cobertura* y la *medida-F, e*n conjunto, con una tabla de verdad, representando todas las posibles combinaciones de acierto/error del clasificador, donde *TP* (*true positive*, *verdadero positivo*), representa al documento que ha sido etiquetado como positivo, tanto por el sistema automático, como el sistema manual (encuesta, sitio web). En cambio, *FN*, corresponde cuando el sistema automático lo clasifica como *negativo* y el sistema manual como positivo, por lo que es considerado un *falso negativo*. Algo similar ocurre si el sistema automático lo clasifica como *positivo* y el sistema manual como *negativo,* a esto se le denomina *falso positivo* (*FP*). Finalmente, si el sistema automático lo clasifica como *negativo* y el sistema manual también, adquiere el nombre de *verdadero negativo* (TN). Lo anterior, se resume en la siguiente tabla. (CUADRADO, 2011)

	Automático	
Manual	Positivo	Negativo
Positivo	*TP*	*FN*
Negativo	*FP*	*TN*

Tabla 16: Tabla de Verdad

A continuación, se detallan cada una de las métricas

Exactitud: Mide el porcentaje de aciertos en relación al total de documentos para el clasificador en ambas clases. La exactitud es una medida global de acierto del clasificador en la tarea. (CUADRADO, 2011)

$$exactitud = \frac{TP + TN}{TP + FN + FP + TN}$$

Ecuación 2: Exactitud

Precisión: Mide el número de aciertos del clasificador para una clase entre el total de clasificados en dicha clase. Por ejemplo, un clasificador tiene una buena precisión si posee un bajo número de falsos positivos. (CUADRADO, 2011)

$$precisión_P = \frac{TP}{TP + FP}$$

Ecuación 3: Precisión

Cobertura: Mide el número de aciertos para una clase en relación al número de documentos que deberían haber sido clasificados en esa clase. Es decir, la cobertura se puede ver como la sensibilidad del clasificador para dicha clase, ya que a mayor cobertura menor número de *falsos negativos.* (CUADRADO, 2011)

$$cobertura_P = \frac{TP}{TP + FN}$$

Ecuación 4: Cobertura

Medida-F: Corresponde a la combinación de las medidas anteriores, *precisión* y *cobertura*, que representa la media armónica de la *precisión* y la *cobertura*. (CUADRADO, 2011)

$$medida - F = \frac{2 \times presición \times cobertura}{precisión + cobertura}$$

Ecuación 5: medida-F

8.2 MUESTRA REPRESENTATIVA

A continuación se detalla las muestras extraídas, con el objetivo de validar los resultados obtenidos por el clasificador automático.

8.2.1 SITIO QUELIBROLEO

Como se mencionó, en la **sección 7.1**, una de las razones por la que seleccionó, el sitio *quelibroleo*, fue porque cada comentario escrito tiene asociado una nota, que oscila de 1 a 10 (ver **tabla 3**), por lo que se tuvo que realizar una conversión para conseguir comparar ambos resultados.

	Positivo	**Negativo**	**Ambiguo**
Categoría	Excelente	Pésimo	Regular
	Muy Bueno	Malo	
	Bueno		

Tabla 17: Conversión quelibroleo

Por otra parte, se seleccionó un total de 188 comentarios al azar de un total de 2.534, que se compararon con los resultados entregados al clasificar la polaridad por comentarios. Además, se extrajo un total de 20 comentarios, con una extensión máxima de tres líneas para ser evaluados por el grupo de expertos.

Finalmente se seleccionaron 52 libros al azar, donde se compararon los resultados entregados, al clasificar la polaridad a nivel de libro. En el anexo **11.12** se presenta la lista de libros seleccionados.

8.3 GRUPO DE EXPERTOS

Con el fin de comparar los resultados entregados tanto por el cálculo de la polaridad como con la métrica presente en el sitio. Se realizó un encuesta online, sin restricción en su grupo etario, pertenecientes al foro Bookzinga[11]. En el **anexo 11.13**, se presenta el contenido de la encuesta.

Figura 31: Sitio Bookzinga

La encuesta constó de un total de 24 preguntas, de las cuales 4 hacían referencia al país de origen, edad, cantidad que libros que leen al mes y sexo. Por tratarse de una encuesta no presencial, se determinó que los comentarios que ellos debían categorizar en positivo, negativo o ambiguo, no tuvieran una extensión superior a las tres líneas.

[11] http://www.bookzingaforo.com/

8.4 EXPERIMENTOS

Los siguientes experimentos tienen como objetivo responder a las siguientes preguntas:

¿La clasificación de polaridad del comentario es coherente con los resultados entregados por el sitio?

¿La clasificación de polaridad del comentario es coherente con los resultados entregados por el grupo de expertos?

¿La clasificación de polaridad del libro es coherente con los resultados entregados por el sitio?

8.5 CLASIFICACIÓN DE LA POLARIDAD

8.5.1 CLASIFICACIÓN DE LA POLARIDAD DEL COMENTARIO

En este apartado se clasificó la polaridad de cada comentario, independiente del libro al que pertenecieran. Del total de 2534 comentarios, se seleccionó una muestra representativa de 188 al azar. El cálculo en detalle de este valor se detalla en el **anexo 11.17**

8.5.1.1 Clasificación Muestra Representativa

A continuación se presentan los resultados de la clasificación de los comentarios de una muestra significativa.

De los 188 comentarios, 143 corresponden a comentarios clasificados como *positivos* equivalente al 76,52%. Por otra parte 26 corresponden a comentarios *negativos* equivalente al 13,83% Y finalmente se identificaron un total de 19 comentarios *ambiguos*, equivalente al 10,11%. Como se puede observar la distribución porcentual de los tipos comentarios, es muy similar a la presentada en la **sección 7.5**, con lo cual corroboramos que se trata de una muestra representativa, del universo de 2.534 comentarios.

A continuación se presenta un gráfico, donde se ilustra la distribución porcentual de los comentarios.

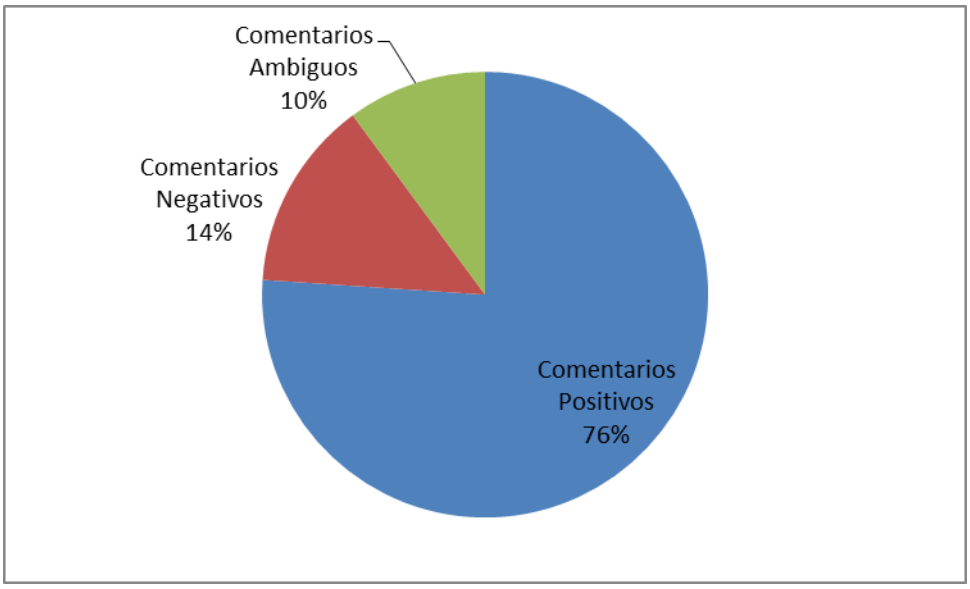


Figura 32: Distribución de la Muestra de Comentarios

Una vez identificados los comentarios, se procedió a identificar la puntuación entregada por cada usuario. Tal como se observa en la siguiente figura.

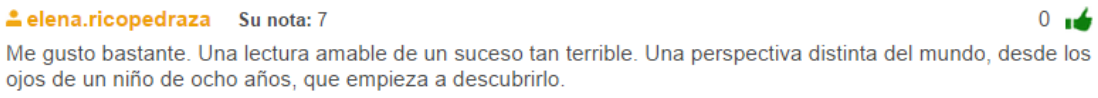


Figura 33: Puntuación de Cometario

Donde el usuario *elena.ricopedraza,* le entrega una nota 7 al libro *el niño con el pijama a rayas,* al realizar la conversión tenemos que el comentario primero se clasifica como *Muy Bueno* de acuerdo a la escala del sitio (ver **tabla 3**), lo cual nosotros lo consideraremos como Positivo (ver **tabla 17**).

Una vez finalizado este proceso, se realiza lo siguiente:

1- Si la clasificación automática coincide con la catalogación del usuario en positivo o ambiguo se representa como *TP* (*true positive*, *verdadero positivo*).
2- En cambio, la clasificación automática etiqueta el comentario como negativo o ambiguo, y el usuario como positivo, se representa como *FN (false negative, falso negativo)*
3- Por otro lado, si la clasificación automática los etiqueta como positivo o ambiguo, y el usuario como negativo, se representa como FP (*false positive, falso positivo*)

4- Si la clasificación automática los etiqueta como positivo o negativo, y el usuario como ambiguo, se representa como *FN (false negative, falso negativo) o* FP (*false positive, falso positivo*), respectivamente.
5- Finalmente si la clasificación automática lo clasifica como negativo y esto coincide con lo catalogado por el usuario, se representa como TN (*true negative, verdadero negativo*)

Con estos datos, se obtuvo la siguiente tabla:

TP	FP	TN	FN
127	18	8	35

Tabla 18: Resultado De Muestra de Comentarios

La cual, permite calcular las métricas de *exactitud, precisión, cobertura, medida-F (*en la **sección *8.1.1*** se detallan las ecuaciones de las métricas*),* los resultados fueron los siguientes:

Exactitud: 71,80%. Indica que existe un gran porcentaje de aciertos en relación al total de comentarios.

Precisión: 87,59%. Esta métrica entrega el número de aciertos para una clase, en este caso se aplicó para las polaridades positivas, ya que un clasificador tiene una buena precisión si posee un bajo número de falsos positivos (CUADRADO, 2011). Enfocándonos en nuestro caso de estudio, implica existe un 79,35% de probabilidad de clasificar correctamente un comentario positivo.

Cobertura: 78,4%. De esto se desprende que la sensibilidad del clasificador para los comentarios positivos en relación a los comentario que debieron ser clasificados (*TP + FN*), que corresponde al 79,3%. Por lo tanto entre más cercano a 100%, menor será el número de falsos negativos

Medida-F: 82,7%. Corresponde la combinación de las medidas *precisión* y *cobertura,* que representa la media armónica (o inverso de la media aritmética) de la *precisión* y la *cobertura.*

La clasificación realizada, se encuentra en un archivo Excel almacenado en el siguiente enlace, presente en el **anexo 11.14.29**

8.5.1.2 Clasificación por Polaridad

En este apartado, y con la finalidad de determinar la precisión de cada clase de polaridad, para determinar cómo se comporta el clasificador con cada una, se tomó una muestra representativa de cada clase.

8.5.1.2.1 Polaridad Positiva

Del total de 1.949 comentarios positivos, se extrajo 184, los cuales todos poseían polaridad positiva, previa clasificación automática. Y se procedió a realizar el proceso descrito en la **sección 8.5.1.1**, que corresponde a identificar el puntaje de cada comentario y realizar su respectiva conversión.

A continuación se procedió a determinar la métrica de *exactitud*, por lo que en primer lugar se calculó los valores de *TP, FN, TN, FP.* Lo cual se resume en la siguiente tabla.

TP	FP	TN	FN
149	0	0	35

Tabla 19: Resultado de Comentarios Positivos.

Tal como se puede observar *TN* y *FP*, son cero, ya que como es natural, al concentrarse los comentarios en solo con polaridad positiva, por lo tanto es congruente que no aparezcan verdaderos negativo o falsos positivos. Por lo anterior, es que la exactitud de aciertos en polaridades positivas es de 80% que corresponde al porcentaje de identificar correctamente un comentario positivo.

8.5.1.2.2 Polaridad Negativa

Del total de 320 comentarios negativos, se extrajo 124, los cuales todos poseían polaridad negativa, previa clasificación automática. Y se procedió a realizar el proceso descrito en la **sección 8.5.1.1**, que corresponde a identificar el puntaje de cada comentario y realizar su respectiva conversión.

A continuación se procedió a determinar la métrica de *exactitud*, en primer lugar se calculó los valores de *TP, FN, TN, FP.* Lo cual se resume en la siguiente tabla.

TP	FP	TN	FN
0	73	51	0

Tabla 20: Resultado de Comentarios Negativos

Por lo anterior, la exactitud de aciertos en polaridades negativas corresponde a un 41,1%, que equivale al porcentaje de identificar correctamente un comentario negativo,

la explicación más detallada de por qué se obtiene este porcentaje se abordará al final de este capítulo.

8.5.1.2.3 Polaridad Ambigua

Del total de 275 comentarios ambiguos, se extrajo 117, los cuales todos poseían polaridad ambigua, previa clasificación automática. Y se procedió a realizar el proceso descrito en la **sección 8.5.1.1**, que corresponde a identificar el puntaje de cada comentario y realizar su respectiva conversión.

A continuación se procedió a determinar la métrica de *exactitud*, en primer lugar se calculó los valores de *TP, FN, TN, FP.* Lo cual se resume en la siguiente tabla.

TP	FP	TN	FN
5	31	0	81

Tabla 21: Resultado de Comentarios Ambiguos

Por lo anterior, la exactitud de aciertos en polaridades ambiguas corresponde a un 4%, que equivale al porcentaje de identificar correctamente un comentario ambiguo, la explicación más detallada de por qué se obtiene este porcentaje se entrega al final de este capítulo.

El detalle de la clasificación por polaridad se encuentra en un archivo Excel almacenado en el siguiente enlace, presente en el **anexo 11.14.30**

8.5.2 CLASIFICACIÓN DE LA POLARIDAD DEL LIBRO

En este apartado se clasificó la polaridad del libro, y dado que éste se encuentra compuesto por un número determinado de comentarios, se procedió primero a calcular la polaridad de cada comentario, para luego determinar la polaridad del libro, la cual se comparó, con la presentada en el sitio.

De los 69 libros, se seleccionaron al azar 52. Y a continuación se procedió a identificar la puntuación general del libro, el cual como se observa en la siguiente imagen, viene dado por los votos entregados por los usuarios, los cuales no están obligados a redactar una crítica de este.

Figura 34: Puntuación de Libro

Donde *Cien Años de Soledad*, tiene un nota de *8,26* que se traduce según la escala del sitio a *Muy Bueno* (ver **tabla 3**). Y de acuerdo a nuestra escala corresponde a *Positivo* (ver **tabla 17**).

Para calcular la polaridad de los libros, se precede a sumar la polaridad de cada comentario que lo compone. Si la sumatoria es mayor o igual a *1*, se considera *positivo*, igual a *0 ambiguo*, y menor o igual a *-1 negativo.*

A continuación, se realiza el siguiente proceso:

1- Si la clasificación automática coincide con lo catalogado por el sitio en positivo se representa como *TP* (*true positive*, *verdadero positivo*).
2- En cambio, la clasificación automática etiqueta el libro como negativo o ambiguo, y el sitio como positivo, se representa como *FN (false negative, falso negativo)*
3- Por otro lado, si la clasificación automática los etiqueta como positivo o ambiguo, y el sitio como negativo, se representa como FP (*false positive, falso positivo*)
4- Si la clasificación automática los etiqueta como positivo o negativo, y el sitio como ambiguo, se representa como *FN (false negative, falso negativo) o* FP (*false positive, falso positivo*) respectivamente.
5- Finalmente si la clasificación automática lo clasifica como negativo o ambiguo y esto coincide con lo catalogado por el sitio, se representa como TN (*true negative, verdadero negativo*)

Para ilustrar de mejor manera este procedimiento, se ejemplificará de la siguiente manera. Si deseamos comparar la polaridad del libro “Circo Máximo. La ira de Trajano” con la entregada por el sitio. Primero debemos calcular la polaridad de cada uno de los

comentarios presentes en la página por medio el clasificador automático. Donde, al finalizar tenemos que todas son de polaridad *positiva*.

Figura 35: Ejemplo de Polaridad Libro

Paralelamente, observamos que el puntaje que entrega el sitio es de 8,87, que corresponde a *Muy Bueno*, y que según nuestra escala el libro tendía una polaridad positiva, es decir, es bien catalogado por los usuarios. Y al coincidir la polaridad entregada por el sitio y la calculada automáticamente, se representa con *TP* (*verdadero positivo*).

Ahora bien de los 52 libros seleccionados, calculando las métricas de *exactitud, precisión, cobertura, medida-F (*en la **sección *8.1.1*** se detallan las ecuaciones de las métricas*),* los resultados fueron los siguientes:

TP	FP	TN	FN
24	7	0	21

Tabla 22: Resultados de Muestra de Libros

Exactitud: 44%. Indica el grado de asertividad al calcular la polaridad de un libro en función de total de este.

Precisión: 77%. Esta métrica entrega el número de aciertos para una clase, el cual se aplicó para los libros con polaridades positivas, esto quiere decir, que existe un 77% de probabilidad de clasificar correctamente un libro como positivo, tomando el total de comentarios que lo componen.

Cobertura: 53%. De esto se desprende que la sensibilidad del clasificador para los determinar la polaridad de lo libros positivos en relación a los libros que debieron ser clasificados (*TP* + *FN*), que corresponde al 53%

Medida-F: 63%. Corresponde la combinación de las medidas *precisión* y *cobertura,* que representa la media armónica (o inverso de la media aritmética) de la *precisión* y la *cobertura.*

La clasificación realizada, se encuentra en un archivo Excel almacenado en el siguiente enlace, presente en el **anexo 11.14.30**

8.5.3 CLASIFICACIÓN DE LA POLARIDAD POR GRUPO DE EXPERTOS

En este apartado se comparó un total de 20 comentarios polarizados provenientes de diversos libros, con un grupo de expertos pertenecientes al foro *Bookzinga*. Encuesta realizada por internet entre los días 8 de Agosto y 16 de Septiembre del presente año. Donde la muestra total correspondió a 101 personas de distintas nacionalidades. El detalle de cada pregunta, junto a su respectivo análisis, se encuentra en el **anexo 11.18**

Por otra parte, considerando que el libro con más comentarios cuenta con un total de 180 críticas, la muestra representativa de este valor, corresponde a un total de 96, y dado que dicha cifra fue superada, consideramos exitosa la encuesta en cuanto a cobertura.

Lo primero que se realizó en este apartado fue identificar el puntaje que poseía cada comentario incorporado a la encuesta, y realizar su posterior conversión. Una vez finalizada esta parte se procedió a etiquetar los resultados, como se explica a continuación:

1- Si la clasificación automática coincide con lo catalogado por el grupo de experto en positivo se representa como *TP* (*true positive*, *verdadero positivo*).
2- En cambio, la clasificación automática etiqueta el libro como negativo o ambiguo, y el grupo de expertos como positivo, se representa como *FN (false negative, falso negativo)*
3- Por otro lado, si la clasificación automática los etiqueta como positivo o ambiguo, y el grupo de expertos como negativo, se representa como FP (*false positive, falso positivo*)
4- Si la clasificación automática los etiqueta como positivo o negativo, y el grupo de expertos como ambiguo, se representa como *FN (false negative, falso negativo) o* FP (*false positive, falso positivo*) respectivamente.

5- Finalmente si la clasificación automática lo clasifica como negativo o ambiguo y esto coincide con lo catalogado por el grupo de expertos, se representa como TN (*true negative, verdadero negativo*)

Finalmente con los datos obtenidos, se obtiene la siguiente tabla:

TP	FP	TN	FN
11	3	1	5

Tabla 23: Resultados Muestra Encuesta

La cual, permite calcular las métricas de *exactitud, precisión, cobertura, medida-F (*en la **sección *8.1.1*** se detallan las ecuaciones de las métricas*),* los resultados fueron los siguientes:

Exactitud: 60%. Indica que existe un gran porcentaje de aciertos de parte del clasificador automático en relación al total de comentarios.

Precisión: 79%. Esta métrica entrega el número de aciertos para una clase, se aplicó para las polaridades positivas, ya que un clasificador tiene una buena precisión si posee un bajo número de falsos positivos (Cuadrado, 2011). Enfocándonos a nuestro caso de estudio, implica existe un 79% de probabilidad de clasificar correctamente un comentario positivo, de acuerdo al catalogado por los expertos.

Cobertura: 68%. De esto se desprende que la sensibilidad del clasificador para los comentarios positivos en relación a los comentarios que debieron ser clasificados.

Medida-F: 73%. Corresponde la combinación de las medidas *precisión* y *cobertura,* que representa la media armónica (o inverso de la media aritmética) de la *precisión* y la *cobertura.*

La clasificación realizada, se encuentra en un archivo Excel almacenado en el siguiente enlace, presente en el **anexo 11.14.32**

8.6 CONCLUSIÓN DE LAS PRUEBAS

En esta apartado, se explican las conclusiones que obtuvieron de las pruebas antes descritas.

8.6.1 CLASIFICACIÓN DE POLARIDAD DEL COMENTARIO

- En primer lugar, durante la revisión de los distintos comentarios polarizados, se observó varias deficiencias del lexicón. Debido, a que varios comentarios, presentaban un alto número de palabras desconocidas, las cuales podían cambiar la polaridad actual calculada.

  Un ejemplo de esto es:

  "Hacía tiempo que no me llevaba una novela de intriga a una lectura tan continuada, apetece seguir leyendo y no te defrauda con el paso de las hojas."

  Luego, se haber eliminado los signos de puntuación y stop-wods. El comentario queda así: *"hacía tiempo llevaba novela intriga lectura tan continuada apetece seguir leyendo defrauda paso hojas"*

  Obtenemos que el comentario es negativo, lo cual se contradice con la puntuación del usuario que corresponde a 10, es decir, positivo. Pero al observar en más detalle, las palabras identificadas como **positiva** es: *leyendo*, como **negativas:** *continuidad, paso*. Como **ambiguas**: *tiempo, intriga*. Y finalmente como **desconocidas**: *hacía, llevaba, novela*, lectura, tan, apetece, *seguir, defrauda, hojas.* Por lo tanto, si sumamos todo, se obtiene que el comentario es negativo.

  Paralelamente se encontraron algunos casos, donde no se encontró ninguna palabra polarizada, esto implicó que por defecto el comentario quedara como ambiguo.

  Un ejemplo de esto es: "*Imprescindible su lectura. Te obliga a la reflexión*."

  Limpiado el comentario queda: *imprescindible lectura obliga reflexión.*

  De este comentario, el clasificador entregó que las palabras: *imprescindible, lectura, obliga, reflexión,* son **desconocidas**, es decir, no se encontraban presentes en el lexicón.

- Otro fenómeno, que se presentó, y como ya se ha mencionado en el transcurso de este trabajo, en ningún momento se pretendió, identificar o calcular la ironía de

manera exitosa, con el objetivo de observar cómo se comporta nuestro clasificador, se presenta el siguiente ejemplo.

Si tomamos la siguiente frase como ejemplo: "*Libro indispensable para ir al baño, si te quedas sin papel, tienes 600 páginas para limpiarte"*

Luego de realizar la etapa de limpieza, queda: "*libro indispensable ir baño si quedas papel páginas limpiarte*".

Donde dicho comentario es considerado positivo, debido a que las palabas **positivas** encontradas son: *libro, baño*. Es con la palabra *libro,* que se produce el fenómeno de ambigüedad, debido a que el lexicón considera *libro*, del verbo *librar* (por ej. "me libro de él") y con el objetivo de no alterar la integridad del lexicón no se modificó su polaridad. Y como **negativa** la palabra*: quedas;* **ambiguas:** *papel.* Finalmente *indispensable, ir, páginas, limpiarte,* se consideran **desconocidas**.

- Otra variable que afectó tanto positiva como negativamente, es que el ranking es designado por los usuarios, por lo que es subjetivo, es decir, un comentario puede contener igual cantidad de palabas *positivas* como *negativas*, en otras palabras, un comentario *ambiguo*, el cual no necesariamente va a tener una **nota 5** por parte del usuario. Un caso de esto es el siguiente ejemplo, "*Demasiado lento y pesado, pero en general a mi García Márquez me provoca esa sensación*". Que a simple vista uno puede considerarlo como un comentario negativo, conclusión que comparte el clasificador automático, etiquetándolo como tal, pero el usuario lo catalogó con un **6**, de una escala de 10, lo cual según el sitio obtiene la etiqueta de *Bueno*. Esto demuestra que no pareciera existir una concordancia entre la clasificación y el comentario. Por lo tanto, surge la duda si se trata de un alago o una ironía por parte del usuario. Interrogantes que afectaron el cálculo de la exactitud en las polaridades ambiguas.
- Finalmente como, se observó al calcular la exactitud entre las distintas clases de polaridad, se obtuvo que existía un mejor desempeño al identificar comentarios positivos que negativos. Esto se debe a que existe una tendencia humana a suavizar las críticas negativas mediante eufemismos o negaciones de términos positivos, (VILARES, ALONSO, & GÓMEZ-RODRÍGUEZ, 2013) lo que dificulta su identificación.

8.6.2 CLASIFICACIÓN DE LA POLARIDAD DEL LIBRO

- Con respecto al cálculo de la polaridad en el libro, se encontró que la cantidad de críticas en algunos casos era muy inferior al número de votos sobre el libro, lo que provocó una cierta variación a la hora de comparar los resultados como fue el caso

del libro “El Peregrino de Compostela (Diario de un Mago)” la cual cuenta con 102 votos versus los 7 comentarios, y al calcular el promedio de los 7 comentarios da que el libro debería estar en la categoría de *Muy Bueno*, y no *Regular* como lo tiene el sitio de acuerdo a los votos.

- Por otro lado se encontraron varios comentarios, que no tenían una puntuación asociada, por lo que el sitio los consideraba como 0, como es el caso de “La Conspiración de Yuste. Hay que matar a Carlos V” que cuenta con la misma cantidad de críticas que votos, que corresponden a 6. Y al calcular el promedio con solo los comentarios que tenían asociada una puntuación, el libro pasaba de la categoría *Malo* a *Bueno*. Hay que aclarar que según la escala del sitio, no se considera en ningún momento el 0 como nota mínima, al contrario, la escala inicia en 1.

8.6.3 CLASIFICACIÓN DE LA POLARIDAD DEL GRUPO DE EXPERTOS

- Con la clasificación de los expertos se corroboró una de las falencias que presenta el clasificador, relacionado a las sutilezas del lenguaje, que para el ser humano es más sencillo identificarlas. Por ejemplo: “*Es mejor que la película...pero sinceramente no es para tanto*”. Donde el clasificador lo etiquetó como positivo, en cambio el 42% de los expertos lo consideró *negativo*, y el otro 50% como *ambiguo*, y solo el 9% como *positivo*. Lo que demuestra nuestro punto.
- Otro hecho que se presentó, es que en algunas ocasiones aplicar el stop-words, provocaba que el comentario adquiera una polaridad diferente a la esperada, debido a que se eliminaba ciertas negaciones. Por ejemplo *"Es mejor que la película...pero sinceramente no es para tanto",* al aplicar stop-words quedaba como *“mejor película sinceramente”,* lo cual repercutía a que el comentario se etiquetara como positivo, esto se debe a que el lexicón no consideró expresiones, solo palabras independientes.

Como conclusión final, se puede rescatar que a pesar de las falencias encontradas, y presentas, el clasificador tiene métricas superior al 60% comparándolos con expertos y un 70% comparándolo con las puntuaciones del sitio. Por tanto, los resultados obtenidos fueron coherentes tanto con la clasificación de comentarios, como la de expertos y el sitio. Y presentó una pequeña falla porcentual al calcular un conjunto de comentarios agrupados por libro.

9 CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se creó un corpus basándonos en críticas literarias, las cuales fueron extraídas, limpiadas, pre-procesadas, clasificadas. Donde, las principales dificultades correspondieron a las transformaciones de aquellas expresiones que no tenían significado por sí solas, pero que transmitían una emoción, como es el caso de *bla bla bla*. Y como se trató de un proceso manual, por medio de un procesador de texto, existieron ciertos caracteres que no pasaron por este proceso, por lo que como trabajo futuro se proyecta realizar este procedimiento de manera automática, y así mejorar los resultados.

Por otro lado, se presentó en el transcurso de este informe variados plugins, que a pesar de no ser creación del autor de este proyecto, fue un desafío encontrarlas, debido a la escasez de herramientas creadas para el idioma español.

Paralelamente, al generar el lexicón, se tuvo que considerar las métricas que utilizaron cada uno de los autores, realizando las pertinentes conversiones. Y a pesar de la gran cantidad de palabras que lo componen, se encontró un total de 5.514 palabras sin repetir que el lexicón no logró identificar. Por lo cual, se proyecta como trabajo futuro, polarizarlas, con el objetivo de mejorar el cálculo de la polaridad. Paralelamente a esto, se encontraron varias palabras que carecían sentido en un contexto escrito, como es el caso de “amor adolecente”, y se pretende como trabajo futuro encontrar una palabra que englobe dicha expresión, sin alterar su sentido. Por otra parte, incorporar en el lexicón expresiones como “no es para tanto”, ya que al ser diseccionadas, se pierde su intencionalidad, es decir, al estar compuesta por stop-words, el clasificador las elimina, debido a esto se pretende buscar dichas expresiones como frase.

Por otro, cuando el stemming de palabra no era único, es decir, una palabra se encontraba asociada a más de un stemming, se consideró la polaridad de la primera en la lista, lo cual es una debilidad de nuestro clasificador. Para contrarrestar dicha situación, se podría calcular la media de la polaridad de los stemming de esta, y asociarla a la palabra.

Además como ya se mencionó, para identificar con mayor certeza los comentarios negativos, se pretende como trabajo futuro implementar en una nueva versión del lexicón que potencie las palabras negativas, y comparar los resultados.

Por último, para abordar el tema de la ironía y/o sarcasmo, se proyecta generar un modelo u algoritmo que permita identificarlas. Así mismo, para determinar el contexto,

es decir, que dependiendo de las palabras que la acompañen, estas tienen distintas acepciones, como fue el caso de *libro.*

Finalmente, como se ha podido observar, en el transcurso de este informe se ha cumplido objetivo general de este estudio, que consistía en analizar la polaridad de los comentarios sobre novelas en un sitio web. Objetivo que se llevó a cabo a través de una investigación del estado del Arte, explorando múltiples plugins. Extrayendo el corpus de un sitio, para luego limpiarlo y pre-procesarlo. Y como punto final se evaluaron los resultados que se encuentran presentes en esta conclusión.

10 BIBLIOGRAFÍA

Cuadrado, J. C. (2011). *Un Modelo Lingüístico-Semántico Basado en emociones para la Clasificación de Textos según su Polaridad e Intensidad.* Tesis para optar al grado de Doctor en Informática, Universidad Complutense de Madrid.

David Vilares, M. A.-R. (2013). *Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento linguistico.* Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 127-134.

Ellison., D. M. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1):article 11.

Erik Cambria, A. H. (2012). Sentic Computing, Techniques, Tools, and Applications. Springer.

Etiquetas Eagles. (s.f.). Recuperado el 21 de Julio de 2014, de http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Eugenio Martínez Cámara, M. M. (2011). *IV Jornadas TIMM Tratamiento de la Información Multilingüe y Multimodal*, (págs. 61-63). Torres, Jaén.

Fermín L. Cruz, J. A. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en españool. *Procesamiento del lenguaje Natural, nº 41*, 73-80.

Fernández Anta, A. P. (2012). Techniques for Sentiment Analysis and Topic Detection of Spanish Tweets: Preliminary Report. Castellón, España.

Fernández, R. A. (2013). *Extracción de Información y Conocimiento de las Opiniones Emitidas por los Usuarios de los Sistemas WEB 2.0.* Santiago: Tesis para optar al grado de Magíster en Gestión de Operaciones.

Ferran Pla, L.-F. H. (2013). ELiRF-UPV en TASS-2013: Análisis de Sentimientos en Twitter. *In proceeding of: TASS workshop at SEPLN 2013. IV Congreso Español de Informática.* Madrid.

Gautam Pant, P. S. (2004). Crawling the Web. *Web Dynamics: Adapting to Change in Content, Size, Topology and Use, Springer-Verlag, Berlin, Germany*.

H. Cunningham, D. M. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. *Proceeding of the 40th Annual Meeting of the ACL*.

Hamish Cunningham, D. M. (29 de Noviembre de 2012). Developing for Language Processing Components with GATE Version 7 (a User Guide). United Kingdom: The University of Sheffield, Department of Computer Science.

Janyce Wiebe, E. R. (2005). Creating Subjective and Objetive Sentence Classifiers fros Unannotated Texts. Ciudad de México, México: In Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics.

Javi Fernández, J. M.-B. (septiembre de 2011). Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog. *Procesamiento del Lenguaje Natural, Revista nº 47*, 179-187.

Jindal, N. a. (2006a.). Identifying comparative sentences in text documents. *In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-2006).*

Jindal, N. a. (2006b.). Mining comparative sentences and relations. *In Proceedings of National Conf. on Artificial Intelligence (AAAI-2006).*

Leiva, I. G. (1996). El procesamiento del lenguaje natural aplicado al análisis del contenido de los documentos. *Revista General De Información Y Documentación 6(2), 205*.

Liu, B. (2006 y 2011). En *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Springer. doi:10.1007/978-3-642-19460-3_12.

LIU, Bing. (2012). *Sentiment Analysis and Opinion Mining.* Toronto: Morgan& Claypool Publishers.

Martí, M. A. (2003). Tecnologías del lenguaje. Barcelona: Editorial UOC.

Martínez Cámara, M. T. (2012). SINAI at TASS 2012. *Procesamiento de Lenguaje Natural*, 50:53-60.

P. Manning, C. D. (2008). An Introduction to Information Retrieval 1st Edition. New York: Cambridge University Press.

Pang, B. L. (2002). Thumbs up? sentimentment classiffication using machine learning techniques. En *En IN PROCEEDINGS OF* (págs. 79-86).

Patty Sakunkoo, N. S. (2009). Analysis of Social Influence in Online Book Reviews. *Proceedings of the Third International ICWSM Conference.*

Porras, V. M. (2008). *Herramientas para la Extracción de Información bajo la arquitectura GATE.*

Porter., M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems 14*, 130-137.

RAE. (s.f.). Recuperado el 14 de Mayo de 2014, de http://lema.rae.es/drae/?val=corpus

Rajaraman, A., & Ullman, J. D. (s.f.). Data Mining. Mining of Massive Datasets.

Readability Tools. (s.f.). Recuperado el 2014 de Julio de 22, de http://www.cs.surrey.ac.uk/BIMA/Projects/LIRICS/liricsSoftware.html

Redondo, J. F. (2007). The Spanish adaptation of ANEW (affective norms for English words). Behavior research methods, 39(3).

Sanjiv R. Das, M. Y. (2001). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *In Proceedings of the Asia Pacific Finance Association Annual Conference*.

Saralegi Urizar, S. V. (2012). TASS: Detecting Sentiments in Spanish Tweets. *En TASS 2012 Working Notes.* Castellón, España.

Sebastiani, A. E. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the 5th Conference on Language Resources and Evaluation.

Strapparava, C. S. (s.f.). WordNet-Affect: an affective extension of WordNet. En *Proceedings of the 4th International Conference on Language Resources and Evaluation* (págs. 1083–1086). 2004.

Taboada, M. J. (2011). Lexicon-based methods for sentiment analysis. (págs. 267-307). Computational Linguistics, 37(2):.

Theresa Wilson, J. ,. (2005). *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.* In Proceedings of the conference on Human Language

Technology and Empirical Methods in Natural Language Processing - HLT ´05. Morristown, NJ, USA: Association for Computational Linguistics, 2005, pp. 347-354.

Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussion. *Proceedings of the Workshop on Operational Text Classification.*

Trilla, A. F. (2012). Sentiment Analysis of Twitter messages based on Multinomial Naive Bayes. *TASS 2012 Wor-king Notes.* Castellón, España.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classiffication of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)* (págs. 417-424). Morristown, NJ, USA.

VALITUTTI, D. S. (2004). Developing affective lexical resources. PsychNology Journal, 2(1).

Verónica Pérez-Rosas, C. B. (s.f.). Learning Sentiment Lexicons in Spanish. LREC 2012: 3077-3081.

Westerski, A. (2007). Sentiment Analysis: Introduction and the State of the Art overview. (págs. 211-218). España: Universidad Politecnica de Madrid.

Wikipedia. (s.f.). Recuperado el 14 de Mayo de 2014, de http://es.wikipedia.org/wiki/Sentimiento#cite_note-1

Wikipedia. (s.f.). Recuperado el 14 de Mayo de 2014, de http://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales

Wikipedia. (s.f.). Recuperado el 2014 de Junio de 22, de http://es.wikipedia.org/wiki/Wattpad

Wikipedia. (s.f.). Recuperado el 22 de Julio de 2014, de http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test#Flesch_Reading_Ease

Wikipedia. (s.f.). Recuperado el 2014 de Julio de 22, de http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test#Flesch.E2.80.93Kincaid_Grade_Level

Wikipedia. (s.f.). Recuperado el 2014 de Julio de 22, de http://en.wikipedia.org/wiki/SMOG_Index

Wikipedia. (s.f.). Recuperado el 2014 de Julio de 22, de http://en.wikipedia.org/wiki/Automated_Readability_Index

Wikipedia. (s.f.). Recuperado el 2014 de Julio de 22, de http://en.wikipedia.org/wiki/Gunning_fog_index

Zhang, L. R. (2011). Combining lexicon-based and learning based methods for Twitter sentiment analysis. Informe Téecnico HPL-2011-89 HP Laboratories, Palo Alto, CA.

11 ANEXOS

11.1 SALIDA DE NOTACIÓN TOKEN

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación Token.

Type	Set	Start	End	Id	Features
Token		0	1	2634	{kind=punctuation, length=1, string=¡}
Token		1	10	2635	{kind=word, length=9, orth=upperInitial, string=Indignaos}
Token		10	11	2636	{kind=punctuation, length=1, string=!}
Token		12	15	2638	{kind=word, length=3, orth=upperInitial, string=Nos}
Token		16	20	2640	{kind=word, length=4, orth=lowercase, string=dice}
Token		21	25	2642	{kind=word, length=4, orth=lowercase, string=este}
Token		26	31	2644	{kind=word, length=5, orth=lowercase, string=libro}
Token		31	32	2645	{kind=punctuation, length=1, string=,}
Token		33	34	2647	{kind=word, length=1, orth=lowercase, string=y}
Token		35	38	2649	{kind=word, length=3, orth=lowercase, string=con}
Token		39	44	2651	{kind=word, length=5, orth=lowercase, string=mucha}
Token		45	50	2653	{kind=word, length=5, orth=lowercase, string=razón}
Token		50	51	2654	{kind=punctuation, length=1, string=,}
Token		52	58	2656	{kind=word, length=6, orth=lowercase, string=porque}
Token		59	61	2658	{kind=word, length=2, orth=lowercase, string=lo}

Tabla 24: Notación Token

11.2 SALIDA DE NOTACIÓN SPACE TOKEN

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación Space Token, que determina los espacios entre las palabras.

Type	Set	Start	End	Id	Features
SpaceToken		11	12	2637	{kind=space, length=1, string= }
SpaceToken		15	16	2639	{kind=space, length=1, string= }
SpaceToken		20	21	2641	{kind=space, length=1, string= }
SpaceToken		25	26	2643	{kind=space, length=1, string= }
SpaceToken		32	33	2646	{kind=space, length=1, string= }
SpaceToken		34	35	2648	{kind=space, length=1, string= }
SpaceToken		38	39	2650	{kind=space, length=1, string= }
SpaceToken		44	45	2652	{kind=space, length=1, string= }
SpaceToken		51	52	2655	{kind=space, length=1, string= }
SpaceToken		58	59	2657	{kind=space, length=1, string= }

Tabla 25: Notación Space Token

11.3 SALIDA DE LA NOTACIÓN SENTENCE

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación Sentence, que reconoce las oraciones en el texto.

Type	Set	Start	End	Id	Features
Sentence		0	468	238859	{}
Sentence		469	637	238861	{}
Sentence		638	880	238864	{}
Sentence		881	996	238866	{}
Sentence		997	1042	238868	{}
Sentence		1043	1187	238871	{}
Sentence		1188	1218	238873	{}
Sentence		1219	1245	238876	{}
Sentence		1246	1285	238878	{}
Sentence		1286	1310	238881	{}
Sentence		1311	1350	238884	{}
Sentence		1351	1727	238886	{}
Sentence		1728	2106	238888	{}
Sentence		2107	2160	238890	{}
Sentence		2160	2259	238892	{}
Sentence		2260	2537	238895	{}
Sentence		2538	2767	238897	{}
Sentence		2768	2981	238899	{}
Sentence		2982	3101	238902	{}

Tabla 26: Notación Sentence

11.4 SALIDA DE LA NOTACIÓN SPLIT

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación Split, que reconoce los puntos presentes en el texto.

Type	Set	Start	End	Id	Features
Split		467	468	238858	{kind=internal}
Split		636	637	238860	{kind=internal}
Split		637	638	238862	{kind=external}
Split		879	880	238863	{kind=internal}
Split		995	996	238865	{kind=internal}
Split		1041	1042	238867	{kind=internal}
Split		1042	1043	238869	{kind=external}
Split		1186	1187	238870	{kind=internal}
Split		1217	1218	238872	{kind=internal}
Split		1218	1219	238874	{kind=external}
Split		1244	1245	238875	{kind=internal}
Split		1284	1285	238877	{kind=internal}
Split		1285	1286	238879	{kind=external}

Tabla 27: Notación Split

11.5 SALIDA DE LA NOTACIÓN TOKEN CON EL ATRIBUTO POS

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación Token, que reconoce y etiqueta las palabras.

Type	Set	Start	End	Id	Features
Token		0	1	2634	{kind=punctuation, length=1, pos=Faa, string=¡}
Token		1	10	2635	{kind=word, length=9, orth=upperInitial, pos=NP00000, string=Indignaos}
Token		10	11	2636	{kind=punctuation, length=1, pos=Fat, string=!}
Token		12	15	2638	{kind=word, length=3, orth=upperInitial, pos=PP1CP000, string=Nos}
Token		16	20	2640	{kind=word, length=4, orth=lowercase, pos=VMIP3S0, string=dice}
Token		21	25	2642	{kind=word, length=4, orth=lowercase, pos=DD0MS0, string=este}
Token		26	31	2644	{kind=word, length=5, orth=lowercase, pos=NCMS000, string=libro}
Token		31	32	2645	{kind=punctuation, length=1, pos=Fc, string=,}
Token		33	34	2647	{kind=word, length=1, orth=lowercase, pos=CC, string=y}
Token		35	38	2649	{kind=word, length=3, orth=lowercase, pos=SPS00, string=con}
Token		39	44	2651	{kind=word, length=5, orth=lowercase, pos=DI0FS0, string=mucha}

Tabla 28: Notación pos

11.6 ETIQUETAS DE SPANISH POS TAGGER

Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. Pero dependiendo de la lengua hay atributos que pueden no especificarse. Si un atributo no se especifica significa que o bien expresa un tipo de información que no existe en la lengua o que la información no se considera relevante. La infra especificación de un atributo se marca con el 0. (ETIQUETAS EAGLES)

En la columna 1 encontramos un número que hace referencia al orden y posición en que aparecen los atributos. La columna 2 hace referencia a los atributos, el número de los cuales varía dependiendo de la categoría. En la columna 3 encontramos los valores que puede tomar cada atributo y, finalmente, la columna 4 representa los códigos que se han establecido para su representación. Las etiquetas en sí sólo son los códigos (columna 4) y se sabe a qué atributo pertenecen por la posición (columna 1) en la que se encuentran. (ETIQUETAS EAGLES)

A continuación, se presentan las tablas de etiquetas agrupadas por categoría.

ADJETIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
3	Grado	Aumentativo	A
		Diminutivo	D
		Comparativo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N

6	Función	-	0
		Participio	P

Tabla 29: Adjetivos

ADVERBIOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adverbio	R
2	Tipo	General	G
		Negativo	N

Tabla 30: Adverbios

DETERMINANTES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Determinante	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
		Artículo	A
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P

		Invariable	N
6	Poseedor	Singular	S
		Plural	P

Tabla 31: Determinantes

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5-6	Clasificación semántica	Persona	SP
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

Tabla 32: Nombres

VERBOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A

		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

Tabla 33: Verbos

PRONOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I

		Interrogativo	T
		Relativo	R
		Exclamativo	E
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Impersonal Invariable	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblicuo	O
7	Poseedor	Singular	S
		Plural	P
8	Politeness	Polite	P

Tabla 34: Pronombres

CONJUNCIONES			
Pos.	**Atributo**	**Valor**	**Código>**
1	Categoría	Conjunción	C
2	Tipo	Coordinada	C
		Subordinada	S

Tabla 35: Conjunciones

INTERJECCIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interjección	I

Tabla 36: Interjecciones

PREPOSICIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída	C
3	Género	Masculino	M
4	Número	Singular	S

Tabla 37: Preposiciones

SIGNOS DE PUNTUACIÓN			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Puntuación	F

Tabla 38: Signos de Puntuación

NUMERALES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Cifra	Z
2	Tipo	partitivo	d
		Moneda	m
		porcentaje	p
		unidad	u

Tabla 39: Numerales

FECHAS Y HORAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Fecha/Hora	W

Tabla 40: Fechas y Horas

11.7 ETIQUETAS DE ANNIE PART OF SPEECH TAGGER

A continuación, se presenta lista de etiquetas, extraídas de (CUNNINGHAM H. , y otros, 2012)

Etiqueta	Significado
CC	Conjunción de coordinación: "y", "pero", "ni", 'o', 'sin embargo', más, menos, menos, los tiempos (multiplicación), durante (división). También 'para' (porque) y 'tan' (es decir, 'de modo que').
CD	Número cardinal
DT	Determinante: los artículos que incluyen 'a', 'un', 'todos', 'no', 'el', 'otro', 'cualquiera', 'algunos', 'los'.
EX	Existencial "ahí": átonas "allí" que provoca la inversión del verbo flexivo y el sujeto lógico; 'Había una fiesta en progreso ".
FW	Palabra extrajera
IN	Preposición o conjunto de subordinación
JJ	JJ - adjetivo: compuestos de con guiones que se utilizan como modificadores; andando- con-felicidad
JJR	Adjetivo - comparativo: Adjetivos con el fin comparativo.
JJS	Adjetivo - superlativos: Adjetivos con el final superlativo.
JJSS	desconocido, pero probablemente una variante de JJS
LRB	Desconocido
LS	Lista de elementos marcadores: Números y letras utilizadas como identificadores de elementos de una lista.
MD	Modal: Por ejemplo "puede", "podría", "se atreven", "puede", "podría", "debe", "debería "," deberá "," debería "," hará "," haría ".
NN	sustantivo - singular o plural
NNP	Nombre propio - singular: Todas las palabras de nombres por lo general son escritas con mayúscula pero los títulos no se consideran.
NNPS	Nombre propio - plural: Todas las palabras de nombres por lo general son escritas con mayúscula pero los títulos no se consideran.
NNS	Sustantivo - plural
NP	Nombre propio - singular
NPS	Nombre propio - plural
PDT	Predeterminante: Determinante como elementos que preceden un artículo o pronombre posesivo.

POS	Final posesivo: Sustantivos acabando " s' o ‘’’ '.
PP	Pronombre personal
PRPR$	Desconocido-, pero pronombre probablemente posesivo
PRP	Desconocido-, pero pronombre probablemente posesivo
PRP$	Pronombre desconocido, pero probablemente posesivo, como 'mi', 'tu', 'su', ‘unos’, 'nuestro', y ‘ellos'.
RB	Adverbio: la mayor parte de palabras que acaban '-mente '.Además de "bastante", "demasiado", "muy", "suficiente", 'de hecho', 'no', y ‘nunca’.
RBR	Adverbio - relativo: los adverbios que terminan en "-er" con un sentido comparativo.
RBS	Adverbio - superlativo
RP	Partículas: Mayormente palabras monosilábicas que también duplican adverbios como direccionales.
STAART	Iniciar marcador de estado (de uso interno)
SYM	Símbolo: símbolos técnicos o expresiones que no son palabras en inglés.
TO	Literal "to"
UH	Interjección: Tales como 'mi', 'oh', 'por favor', 'uh', 'bueno', 'sí'.
VBD	Verbo - tiempo pasado: incluye la forma condicional del verbo "ser"; 'Si yo fuera rico...“.
VBG	Verbo - gerundio o participio presente
VBN	Verbo - participio pasado
VBP	Verbo - sustantivo-tercera persona singular presente
VB	Verbo - forma de la base: imperativos subsumido, infinitivos y subjuntivo.
VBZ	Verbo - 3 ª persona singular del presente
WDT	Determinante ‘wh'
WP$	Posesivo 'wh'-pronombre: incluye' cuya '
WP	Pronombre ‘wh': incluye "qué", "quién" y "quién".
WRB	Adverbio ‘wh’: incluye "cómo", "dónde", "por qué". Incluye "cuándo" cuando se usa en un sentido temporal.
::	Cuando se presenta dos puntos
,	Cuando se presenta una coma.
$	Cuando se presenta signo peso o dólar.
-	Cuando se presenta doble guion.
"	Cuando se presenta comillas dobles.
´	Cuando se presenta una tilde o rayita oblicua que baja de izquierda a derecha de quien escribe o lee (`)

(	Cuando se presenta paréntesis izquierdo.
.	Cuando se presenta un punto.
)	Cuando se presenta paréntesis derecho.
’	Cuando se presenta comilla simple o apóstrofe

Tabla 41: Etiquetas de ANNIE pos tagger

11.8 SALIDA DE LA NOTACIÓN TOKEN AL APLICAR ANNIE POS TAGGER

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación Token, que reconoce y etiqueta las palabras.

Type	Set	Start	End	Id	Features
Token		0	1	18007	{category=NN, kind=punctuation, length=1, string=¡}
Token		1	10	18008	{category=NNP, kind=word, length=9, orth=upperInitial, string=Indignaos}
Token		10	11	18009	{category=., kind=punctuation, length=1, string=!}
Token		12	15	18011	{category=NNP, kind=word, length=3, orth=upperInitial, string=Nos}
Token		16	20	18013	{category=NNS, kind=word, length=4, orth=lowercase, string=dice}
Token		21	25	18015	{category=NN, kind=word, length=4, orth=lowercase, string=este}
Token		26	31	18017	{category=NN, kind=word, length=5, orth=lowercase, string=libro}
Token		31	32	18018	{category=,, kind=punctuation, length=1, string=,}
Token		33	34	18020	{category=NN, kind=word, length=1, orth=lowercase, string=y}
Token		35	38	18022	{category=JJ, kind=word, length=3, orth=lowercase, string=con}
Token		39	44	18024	{category=NN, kind=word, length=5, orth=lowercase, string=mucha}
Token		45	50	18026	{category=NN, kind=word, length=5, orth=lowercase, string=razón}
Token		50	51	18027	{category=,, kind=punctuation, length=1, string=,}
Token		52	58	18029	{category=NN, kind=word, length=6, orth=lowercase, string=porque}

Tabla 42: Notación ANNIE POS tagger

11.9 SALIDA DE LA NOTACIÓN COUNT DEL STATICAL TERM FINDER

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación *Count* esta entrega la cantidad de caracteres, palabras polisílabas, sentencias, silabas y palabras.

Type	Set	Start	End	Id	Features
Count		0	616294	254232	{Characters=481221, PolysyllabicWords=20973, Sentences=6680, Syllables=199110, Words=116509}

Tabla 43: Notación atributo count

11.10 SALIDA DE LA NOTACIÓN READABILITY DEL STATICAL TERM FINDER

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación *Readability* esta entrega Flesch, Kincaid, SMOG, ARI, FOG index.

Type	Set	Start	End	Id	Features
Readability		0	616294	254231	{ARI=6.74460327690241, Execution=1, FOG=14.177060609488525, Flesch=44.55332901601679, Kincaid=11.377979089058577, SMOG=13.25158129617029}

Tabla 44: Notación atributo Readability

11.10.1 ÍNDICE DE FLESCH

Corresponde a la facilidad de comprensión de un documento.

$$206.835 - 1.015\left(\frac{N\,°\,total\ de\ palabras}{N°\ total\ de\ oraciones}\right) - 84.6\left(\frac{N°\ total\ de\ sílabas}{N°\ total\ de\ palabras}\right)$$

Ecuación 6: Flesch

La puntuación puede interpretarse del siguiente modo: (WIKIPEDIA)

Puntuación	**Descripción**
90,0–100,0	Un estudiante medio de 11 años puede entender el texto sin esfuerzo.
60,0–70,0	Un estudiante medio de 13 a 15 años puede entender el texto sin esfuerzo.
0,0–30,0	Universitarios graduados entienden mejor el texto.

11.10.2 ÍNDICE DE KINCAID

Traduce la puntuación de Flesch, a una puntuación de 0-100, la calificación estadunidense en las instituciones educativas. (WIKIPEDIA)

$$0.39\left(\frac{total\ de\ palabras}{total\ de\ sentencias}\right) + 11.8\left(\frac{total\ de\ sílabas}{total\ de\ palabras}\right) - 15.59$$

Ecuación 7: Kincaid

11.10.3 ÍNDICE DE SMOG

Mide de la legibilidad que estima los años de educación necesarios para comprender una pieza de escritura. (WIKIPEDIA)

$$grado = 1.0430\sqrt{número\ de\ palabras\ polosilabas \times \frac{30}{número\ de\ sentencias}} + 3.1291$$

Ecuación 8: SMOG

11.10.4 ÍNDICE DE ARI

Mide la compresibilidad de un texto, se basa en un factor de caracteres por palabra. (WIKIPEDIA)

$$4.71\left(\frac{caracteres}{palabras}\right) + 0.5\left(\frac{palabras}{sentencias}\right) - 21.43$$

Ecuación 9: ARI

11.10.5 ÍNDICE DE SMOG

Calcula los años de educación formal necesarios para comprender el texto en una primera lectura. Por otro lado, las palabras "complejas": son aquellas con tres o más sílabas. No incluye nombres propios, jergas, o palabras compuestas. (WIKIPEDIA)

$$0.4\left[\left(\frac{palabras}{sentencia}\right)+100\left(\frac{palabras\ complejas}{palabras}\right)\right]$$

Ecuación 10: SMOG

11.11 SALIDA DE LA NOTACIÓN LINGUISTICTERM

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación *LinguisticTerm* esta entrega la frecuencia de una palabra.

Type	Set	Start	End	Id	Features
LinguisticTerm		549335	549342	507063	{Frequency=10, Text=acuerdo}
LinguisticTerm		556083	556088	507266	{Frequency=10, Text=ahora}
LinguisticTerm		589085	589097	506018	{Frequency=10, Text=al principio}
LinguisticTerm		481462	481469	510754	{Frequency=10, Text=cultura}
LinguisticTerm		149097	149110	516361	{Frequency=10, Text=decepcionante}
LinguisticTerm		45783	45794	507678	{Frequency=10, Text=descripción}
LinguisticTerm		557971	557976	507310	{Frequency=10, Text=donde}
LinguisticTerm		12555	12563	507101	{Frequency=10, Text=engancha}
LinguisticTerm		163264	163272	516156	{Frequency=10, Text=escribir}
LinguisticTerm		576503	576513	505913	{Frequency=10, Text=gustó nada}
LinguisticTerm		605155	605159	506252	{Frequency=10, Text=hora}
LinguisticTerm		580538	580551	505883	{Frequency=10, Text=impresionante}
LinguisticTerm		248505	248514	511764	{Frequency=10, Text=inocencia}
LinguisticTerm		18335	18337	506626	{Frequency=10, Text=la}
LinguisticTerm		547887	547890	506998	{Frequency=10, Text=lee}
LinguisticTerm		4877	4885	506913	{Frequency=10, Text=los años}
LinguisticTerm		616136	616146	506431	{Frequency=10, Text=los libros}
LinguisticTerm		42927	42932	507421	{Frequency=10, Text=luego}
LinguisticTerm		24963	24971	506884	{Frequency=10, Text=misterio}
LinguisticTerm		604976	604984	506254	{Frequency=10, Text=obstante}

Tabla 45: Notación de LinguisticTerm

11.12 SELECCIÓN DE LIBROS

A continuación, se presenta la lista de libros con los que se comparan los resultados.

N°	Id	Nombre del libro	Clasificación	Polaridad
1	1	¡Indignados!	REGULAR	AMBIGUO
2	3	Alma Inmortal	MALO	NEGATIVO
3	4	Aurora Boreal	REGULAR	AMBIGUO
4	6	Cien Años de Soledad	MUY BUENO	POSITIVO
5	8	Circo Máximo. La ira de Trajano	EXCELENTE	POSITIVO
6	9	Crepúsculo (Saga Crepúsculo I)	BUENO	POSITIVO
7	11	Déjame Entrar	MUY BUENO	POSITIVO
8	12	Desde mi cielo	BUENO	POSITIVO
9	13	Divina Comedia	MUY BUENO	POSITIVO
10	15	El Ángel más Tonto del Mundo	REGULAR	AMBIGUO
11	16	El Ángel Perdido	MALO	NEGATIVO
12	18	El Club Dante	MALO	NEGATIVO
13	19	El Club De Los Viernes Se Reúne de Nuevo	REGULAR	AMBIGUO
14	20	El Código Da Vinci	REGULAR	AMBIGUO
15	21	El Contador de Historias	MALO	NEGATIVO
16	22	El Coronel no tiene quien le Escriba	BUENO	POSITIVO
17	23	El Diario de Bridget Jones	BUENO	POSITIVO
18	24	El Enigma del Cuatro	MALO	NEGATIVO
19	25	El Exorcista	MUY BUENO	POSITIVO
20	26	El Fuego	MALO	NEGATIVO
21	27	El Hobbit	MUY BUENO	POSITIVO
22	29	El Mágico Libro de los Infinitos Cuentos	EXCELENTE	POSITIVO
23	30	El Niño con el Pijama de Rayas	BUENO	POSITIVO
24	31	El Peregrino de Compostela (Diario de un Mago)	REGULAR	AMBIGUO
25	32	El Perfume	MUY BUENO	POSITIVO
26	34	El Principito	MUY BUENO	POSITIVO
27	35	El Señor de los Anillos	MUY BUENO	POSITIVO

28	36	Ella, que todo lo tuvo	REGULAR	AMBIGUO
29	37	Fahrenheit 451	MUY BUENO	POSITIVO
30	40	La Biblia de Barro	REGULAR	AMBIGUO
31	41	La Bodega	REGULAR	AMBIGUO
32	42	La Cena Secreta	REGULAR	AMBIGUO
33	44	LacClave Gaudí	MALO	NEGATIVO
34	45	La Conspiración	REGULAR	AMBIGUO
35	46	La Conspiración de Yuste. Hay que matar a Carlos V	MALO	NEGATIVO
36	47	La Doctora Cole	REGULAR	AMBIGUO
37	48	La Era de los Invidentes	EXCELENTE	POSITIVO
38	52	La Librería	REGULAR	AMBIGUO
39	53	La Tienda de los Suicidas	REGULAR	AMBIGUO
40	54	La Trampa	REGULAR	AMBIGUO
41	55	Las Ardillas de Central Park Están Tristes los Lunes	REGULAR	AMBIGUO
42	57	Los Hombres que no Amaban a las Mujeres (Millennium I)	MUY BUENO	POSITIVO
43	58	Los Juegos del Hambre	MUY BUENO	POSITIVO
44	59	Medianoche	BUENO	POSITIVO
45	60	Memoria de mis Putas Tristes	REGULAR	AMBIGUO
46	61	Memorias de una Geisha	MUY BUENO	POSITIVO
47	62	Muerto Hasta El Anochecer	BUENO	POSITIVO
48	63	Peregrinatio	REGULAR	AMBIGUO
49	66	Si tú me dice ven lo dejo todo…pero dime ven	REGULAR	AMBIGUO
50	67	The Host (La Huésped)	BUENO	POSITIVO
51	68	Un Perfecto Equilibrio	EXCELENTE	POSITIVO
52	69	Una Tienda en París	REGULAR	AMBIGUO

Tabla 46: Selección de Libros

11.13 ENCUESTA A GRUPO DE EXPERTOS

La siguiente encuesta consta de un total de 24 preguntas, donde se le solicita que clasifique las opiniones de la siguiente manera:

Positivo: Si el comentario es agradable. ☺

Ej. *Adrián es muy guapo*[12]

Negativo: Si el comentario es molesto. ☹

Ej. *Los alquimistas son unos torturadores*

Ambiguo: Si no es posible determinar si el comentario es positivo o negativo. :-\

Ej. *Adrián es muy dulce, pero a veces su temperamento se descontrola.*

1) *¿Qué edad tienes?*
   a) *Menor a 15 años*
   b) *Entre 15 y 25 años*
   c) *Entre 26 y 35 años*
   d) *Mayor a 35 años*

2) *¿Cuál es tu país de origen?*
   a) *Alemania*
   b) *Argentina*
   c) *Australia*
   d) *Belice*
   e) *Bolivia*
   f) *Brasil*
   g) *Canadá*
   h) *Chile*
   i) *China*
   j) *Colombia*
   k) *Costa Rica*
   l) *Corea*
   m) *Ecuador*
   n) *España*

[12] Los ejemplos presentados hacen referencia al libro bloodlines, de richelle mead, que es un libro muy popular en el foro bookzinga

o) *Estados Unidos*
p) *El Salvador*
q) *Francia*
r) *Guatemala*
s) *Guyana*
t) *Honduras*
u) *Israel*
v) *Italia*
w) *Jamaica*
x) *Japón*
y) *México*
z) *Nueva Zelanda*
aa) *Nicaragua*
bb) *Panamá*
cc) *Paraguay*
dd) *Perú*
ee) *Portugal*
ff) *Puerto Rico*
gg) *Reino Unido*
hh) *República Dominicana*
ii) *Uruguay*
jj) *Venezuela*

3) *¿Cantidad de libros que lees en promedio al mes?*
   a) *No leo*
   b) *1*
   c) *2*
   d) *3*
   e) *Más de 3*

4) *¿Cuál es tu sexo?*
   a) *Mujer*
   b) *Hombre*

5) *es un clásico que se debe leer, es fantástico una gran obra.*
   a) Positivo
   b) Negativo

c) Ambiguo

6) *Este libro lo leí hace muchos años y lo releí un par de veces más. A quienes les gusta la fantasía mezclado con la magia, léanlo*
   a) Positivo
   b) Negativo
   c) Ambiguo

7) *Lo tuve muchos años en mi estantería, cuando lo leí me reproche no haberlo hecho antes, excelente*
   a) Positivo
   b) Negativo
   c) Ambiguo

8) *Es mejor que la película...pero sinceramente no es para tanto*
   a) Positivo
   b) Negativo
   c) Ambiguo

9) *No sé qué es más malo si el libro o la película*
   a) Positivo
   b) Negativo
   c) Ambiguo

10) *Fue mi primer contacto con el universo de Tolkien y, aunque es más infantil, es totalmente recomendable, sobre todo para continuar leyendo El señor de los anillos y El Silmarillion*
   a) Positivo
   b) Negativo
   c) Ambiguo

11) *El libro no está mal, es entretenido bastante diría yo, pero la verdad me esperaba mucho más.....*
   a) Positivo
   b) Negativo
   c) Ambiguo

12) *Magnífico libro para comprender el horror que se vivió en aquellos días, recomendable al cien por cien*
    a) Positivo
    b) Negativo
    c) Ambiguo

13) *Una historia que entretiene al lector. Sencillo y práctico a la hora de leer. un gran éxito de ventas. Una pequeña gran obra.*
    a) Positivo
    b) Negativo
    c) Ambiguo

14) *A mí me interesó bastante lo leí en una tarde y la verdad me entretuvo, el final muy triste*
    a) Positivo
    b) Negativo
    c) Ambiguo

15) *Tan sencillo y a la vez tan profundo. Libro para leer en un rato recordándolo toda la vida. Un hombre, un niño y una rosa como base para un aprender sin parar.*
    a) Positivo
    b) Negativo
    c) Ambiguo

16) *Puede que no sea tan bueno como el principito, pero igualmente debería ser de lectura obligatoria*
    a) Positivo
    b) Negativo
    c) Ambiguo

17) *Es uno de mis libros preferidos…creo que es el que más veces he leído...recomiendo también, parábola de unas alas*
    a) Positivo
    b) Negativo
    c) Ambiguo

18) *Muy buena novela, buena descripción de los personajes, de los pasajes, te mantiene enganchado, no he leído el primero, pues creía que era independiente uno del otro, no me cansa de leerlo.*
    a) Positivo
    b) Negativo
    c) Ambiguo

19) *Me he leído el código da vinci, ángeles y demonios, y el símbolo perdido y este es el que me pareció más aburrido, me lo acabe a duras penas, no me gustó nada,*
    a) Positivo
    b) Negativo
    c) Ambiguo

20) *Recomendado para quienes gustan del género, sin embargo no me ha gustado lo suficiente para leer las dos obra que me faltan y que completan la trilogía.*
    a) Positivo
    b) Negativo
    c) Ambiguo

21) *Me gustan muchos los libros de temática japonesa, este libro resume la vida de las geishas de una manera que hace que estés allí.*
    a) Positivo
    b) Negativo
    c) Ambiguo

22) *Sin duda una grande obra, excelente libro, me gustó bastante y sobretodo el final que me fascinó y me atrapo totalmente*
    a) Positivo
    b) Negativo
    c) Ambiguo

23) *Esta genial el libro!!tiene acción, suspense, comedia, drama y si, sobre todo romance pero sin llegar a ser empalagoso!!*
    a) Positivo
    b) Negativo
    c) Ambiguo

24) El libro se lee en un momento ....porque la historia engancha

a) Positivo
b) Negativo
c) Ambiguo

11.14 DOCUMENTACIÓN DIGITAL.

A continuación, se dejan los enlaces de los archivos de apoyo que se emplearon durante este proyecto.

11.14.1 PROGRAMA PARA LA EXTRACCIÓN DE COMENTARIOS

http://goo.gl/2OXC49

11.14.2 COMENTARIOS EXTRAÍDOS

http://goo.gl/VMkHhF

11.14.3 LISTA DE LOS LIBROS EXTRAÍDOS

http://goo.gl/sGBNA7

11.14.4 INSTRUCCIONES PARA EL USO DEL PROGRAMA PARA LA EXTRACCIÓN DE COMENTARIOS

http://goo.gl/ctY7FI

11.14.5 URL DE LOS LIBROS EXTRAÍDOS

http://goo.gl/NnVHdm

11.14.6 CORPUS EXTRAÍDO DEL SITIO

http://goo.gl/3ZwFPT

11.14.7 ARCHIVO DONDE SE PRESENTA LOS COMENTARIO EDITADOS

http://goo.gl/9dtlhu

11.14.8 CORPUS LIMPIADO

http://goo.gl/0v1f0C

11.14.9 XML EXTRAÍDO LUEGO DE APLICAR ANNIE POS TAGGER

http://goo.gl/rvGlXI

11.14.10 XML EXTRAÍDO LUEGO DE APLICAR DOCUMENT RESET

http://goo.gl/YOajcw

11.14.11 XLM EXTRAÍDO LUEGO DE APLICAR EL PLUGIN READABILITY TOOLS

http://goo.gl/J53Ifz

11.14.12 XML EXTRAÍDO LUEGO DE APLICAR LINGUISTIC TERMFINDER

http://goo.gl/CUG0qS

11.14.13 XML EXTRAÍDO LUEGO DE APLICAR READABILITY ANALYSER

http://goo.gl/jk5WRJ

11.14.14 XML EXTRAÍDO LUEGO DE APLICAR POS TAGGER

http://goo.gl/gxwIRW

11.14.15 XML EXTRAÍDO LUEGO DE APLICAR SENTENCE SPLITTER

http://goo.gl/yLn1u8

11.14.16 XML EXTRAÍDO LUEGO DE APLICAR TOKEN

http://goo.gl/1WeZgZ

11.14.17 XML EXTRAÍDO LUEGO DE APLICAR SPANISH PLUGIN

http://goo.gl/KrxwBj

11.14.18 XML EXTRAÍDO LUEGO DE FINALIZAR EL PRE-PROCESAMIENTO

http://goo.gl/ILCf88

11.14.19 LEXICÓN INICIALES

http://goo.gl/ARfEAm

11.14.20 LEXICÓN GENERADO AL UNIR LOS CUATRO LEXICÓN

http://goo.gl/OW2YR8

11.14.21 PROGRAMA PARA LA GENERACIÓN DEL LEXICÓN SIN PALABRAS REPETIDAS

http://goo.gl/zeTKNR

11.14.22 LEXICÓN SIN PALABRAS REPETIDAS

http://goo.gl/NF4TC8

11.14.23 LEXICÓN GENERADO AL COMPRIMIR PALABRAS

http://goo.gl/PHnI36

11.14.24 LEXICÓN CON STEMMING

http://goo.gl/mzxiSC

11.14.25 PROGRAMA PARA CALCULAR POLARIDAD

http://goo.gl/51VRRo

11.14.26 COMENTARIOS POLARIZADOS

http://goo.gl/026oXj

11.14.27 PROGRAMA INDIVIDUAL PARA CALCULAR POLARIDAD

http://goo.gl/QLZ14g

11.14.28 Palabras Desconocidas y Palabras Identificadas Por Gazette

http://goo.gl/Bfc1Hy

11.14.29 Muestra Representativa de Comentarios

http://goo.gl/CQ0Kg9

11.14.30 Muestra Representativa por Polaridad

http://goo.gl/zVXljE

11.14.31 Muestra Representativa por Libro

http://goo.gl/jFiESe

11.14.32 Muestra Representativa de Grupos de Expertos

http://goo.gl/4vvrYK

11.14.33 Respuesta de encuesta

http://goo.gl/JEDt39

11.15 PALABRAS COMPRIMIDAS

A continuación se entrega una tabla con las palabras que fueron comprimidas en la **sección 7.4.3**.

Palabra Original	Palabra Modificada
a carcajadas	carcajadas
sentirse abandonado	abandonado
sentir afecto	afecto
muy agradado	agradado
saltar de alegría	alegría
alegría vigorosa	alegría
sentirse animoso	animoso
que muestra aprobación	aprobación
con arrepentimiento	arrepentimiento
que es atroz	atroz
extremadamente bien	bien
cansado de	cansado
los celos	celos
con compasión	compasión
dejar confundido	confundido
sensación conmovedora	conmovedora
muy contento	contento
que causa deleite	deleite
depresivamente oscuro	depresivamente
con desagrado	desagrado
que provoca desprecio	desprecio
que deja embobado	embobado
que enfurece	enfurece
muy entusiasmado	entusiasmado
sentirse espontáneamente entusiasmado	entusiasmado
entusiasmo intenso	entusiasmo
que provoca entusiasmo	entusiasmo
sentir escalofrío	escalofrío
que causa estupefacción	estupefacción
dejar estupefacto	estupefacto
que deja estupefacto	estupefacto

ánimo exultante	exultante
muy graciosamente	graciosamente
muy gracioso	gracioso
que gusta	gusta
gustar de	gustar
con hostilidad	hostilidad
ser implacable	implacable
ira intensa	ira
hacer irritar	irritar
dejar lelo	lelo
que hace llorar	llorar
sentido de logro	logro
que causa miedo	miedo
que da miedo	miedo
tener miedo	miedo
que causa nerviosismo	nerviosismo
con odio	odio
ofender rápidamente	ofender
que da pánico	pánico
en pánico	pánico
expresar pena	pena
sentido de pérdida	pérdida
dejar perplejo	perplejo
sentido de pertenencia	pertenencia
de manera protectora	protectora
erizado de rabia	rabia
rabia acumulada	rabia
rabia criminal	rabia
hacer rabiar	rabiar
que regocija	regocija
que causa regocijo	regocijo
con remordimiento	remordimiento
causar repulsión	repulsión
acumular resentimiento	resentimiento
resentimiento intenso	resentimiento

con satisfacción	satisfacción
que causa satisfacción	satisfacción
que puede ser satisfecho	satisfecho
sentir júbilo	sentir
sentido de suficiencia	suficiencia
hacer sufrir	sufrir
en suspenso	suspenso
que causa suspenso	suspenso
con timidez	timidez
que hace tiritar	tiritar
que causa tristeza	tristeza

Tabla 47: Palabras Comprimidas

11.16 SALIDA DE NOTACIÓN LOOKUP

A continuación entrega una muestra de la tabla que entrega el GATE en el momento de seleccionar la notación Lookup.

Lookup	77	85	55793	{majorType=Persona, minorType=profesion}
Lookup	152	161	55794	{majorType=Persona, minorType=profesion}
Lookup	358	365	55795	{majorType=Persona, minorType=profesion}
Lookup	1200	1204	55797	{majorType=Lugar, minorType=general}
Lookup	1200	1204	55796	{majorType=Persona, minorType=profesion}
Lookup	1200	1204	55535	{majorType=Persona, minorType=Profesion}
Lookup	1262	1270	55798	{majorType=Lugar, minorType=general}
Lookup	1383	1390	55799	{majorType=Persona, minorType=gentilicio}
Lookup	2479	2485	55800	{majorType=Persona, minorType=profesion}
Lookup	2612	2618	55801	{majorType=Persona, minorType=profesion}
Lookup	2646	2654	55802	{majorType=Persona, minorType=profesion}
Lookup	2783	2794	55803	{majorType=Persona, minorType=profesion}
Lookup	2872	2881	55804	{majorType=Persona, minorType=profesion}
Lookup	2914	2925	55805	{majorType=Persona, minorType=profesion}
Lookup	3385	3391	55806	{majorType=Lugar, minorType=general}
Lookup	3888	3899	55807	{majorType=Persona, minorType=profesion}
Lookup	4060	4067	55549	{majorType=Lugar, minorType=Tipo}
Lookup	4092	4097	55550	{majorType=Persona, minorType=Profesion}
Lookup	4471	4486	55551	{majorType=Persona, minorType=Profesion}
Lookup	5393	5404	55554	{majorType=Persona, minorType=Profesion}
Lookup	5616	5621	55555	{majorType=Persona, minorType=Profesion}
Lookup	5927	5935	55556	{majorType=Lugar, minorType=Pais}
Lookup	6765	6773	55559	{majorType=Persona, minorType=Gentilicio}
Lookup	8081	8090	55564	{majorType=Persona, minorType=Profesion}
Lookup	8101	8110	55565	{majorType=Persona, minorType=Profesion}
Lookup	8274	8280	55566	{majorType=Persona, minorType=Profesion}

11.17 CALCULO DE LA MUESTRA REPRESENTATIVA

Para determinar las muestras representativas utilizadas en la etapa de prueba tenemos la siguiente ecuación:

$$\frac{k^2 Npq}{e^2(N-1)+k^2 pq}$$

Ecuación 11: Ecuación para Muestra Representativa

Dónde:

N: es el tamaño de la población o universo.

k: es la constante que depende del nivel de confianza que asignaremos, el cual definimos como un 95%, es decir que nos podemos equivocar en un 5%. Por lo tanto k es igual 1,96.

e: es el error muestral deseado, en tanto por uno. El error muestral es la diferencia que puede haber entre el resultado que obtenemos preguntando a una muestra de la población y el que obtendríamos si preguntáramos al total de ella. En nuestro caso, designaremos el valor 0,03.

p: es la proporción de individuos que poseen en la población la característica de estudios. Como ese dato es desconocido, consideraremos p =0,05.

q: es la proporción de individuos que no poseen esa característica, es decir, 1-p= 0,95

- Para **N=69** libros

  Para calcular la muestra de libro, no entregó 52 libros.

- Para **N=2534** comentarios

  La muestra representativa es de 188 comentarios

- Para **N=1939** comentarios positivos

  La muestra representativa es 184 comentarios positivos.

- Para **N=320** comentarios negativos

  La muestra representativa es de 124 comentarios negativos.

- Para **N= 275** comentarios ambiguos

La muestra representativa es de 117 comentarios ambiguos.

- Para **N=180**, que corresponde al libro con más comentarios en el corpus.

La muestra representativa es de 96 personas que deben contestar la encuesta, para considerarse exitosa.

11.18 DESCRIPCIÓN DE ENCUESTA A EXPERTOS.

La siguiente encuesta fue aplicada un total de 101 personas, entre el 8 de agosto al 16 de Septiembre del año 2014.

La encuesta constaba con un total de 24 preguntas con alternativas, fue completamente anónima y voluntaria.

1) ¿Qué edad tienes?

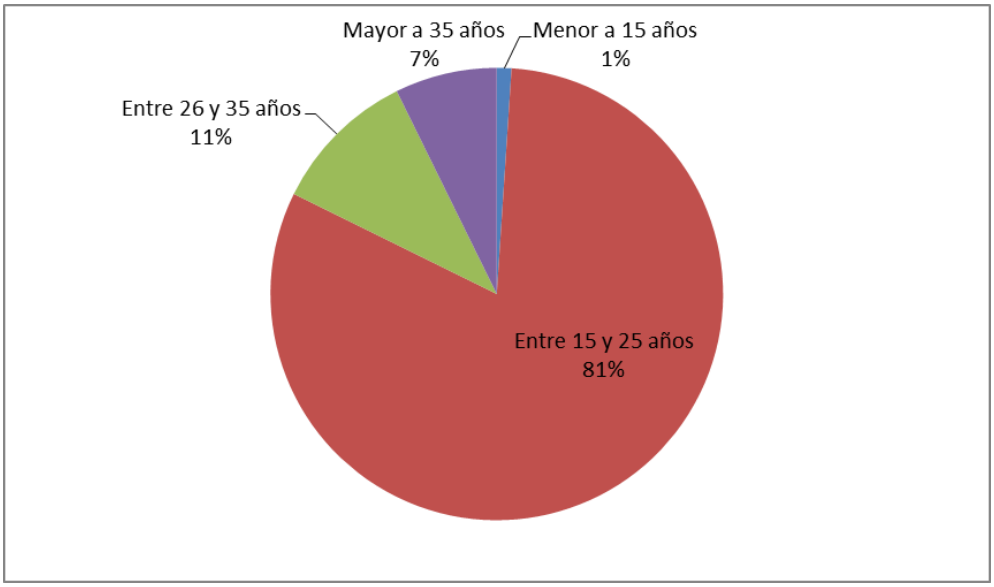


Figura 36: Pregunta 1 Encuesta

Esta primera pregunta, cumple con el objetivo, de determinar el grupo etario que fue encuestado. Donde el 81% se encuentra en los 15 y 25 años. La totalidad de los datos se presentan en la siguiente tabla.

Menor a 15 años	Entre 15 y 25 años	Entre 26 y 35 años	Mayor a 35 años
1	78	10	7

Tabla 48: Pregunta 1

2) ¿Cuál es tu país de origen?

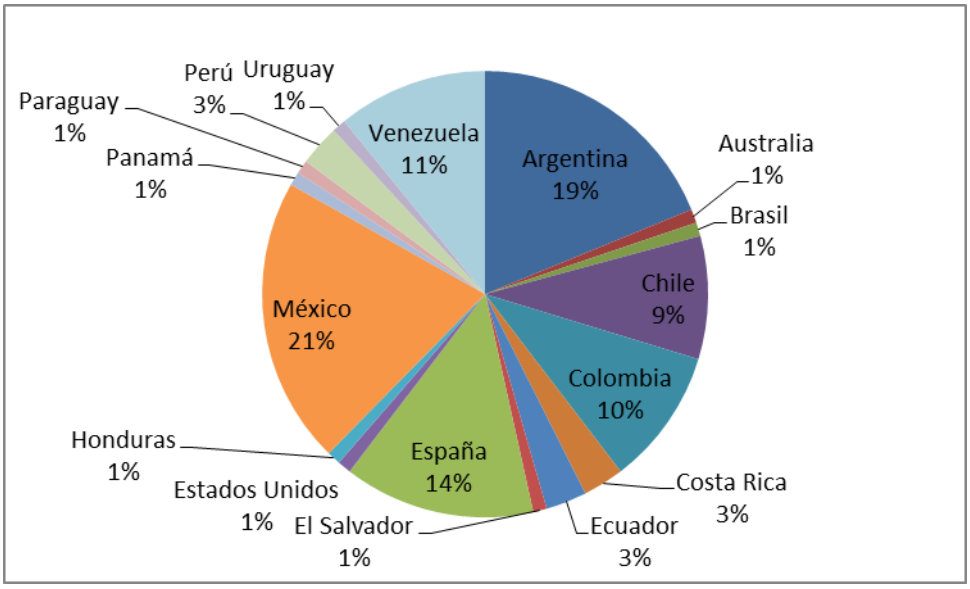


Figura 37: Pregunta 2 Encuesta

Esta siguiente pregunta, cumple con el objetivo, de determinar el lugar de pertenencia de los encuestados. Donde el 21% pertenece a México. La totalidad de los datos se presentan en la siguiente tabla.

País	Total
Argentina	19
Australia	1
Brasil	1
Chile	9
Colombia	10
Costa Rica	3
Ecuador	3
El Salvador	1
España	14
Estados Unidos	1
Honduras	1
México	21
Panamá	1
Paraguay	1

Perú	3
Uruguay	1
Venezuela	11

Tabla 49: Pregunta 2 Encuesta

3) ¿Cantidad de libros que lees en promedio al mes?

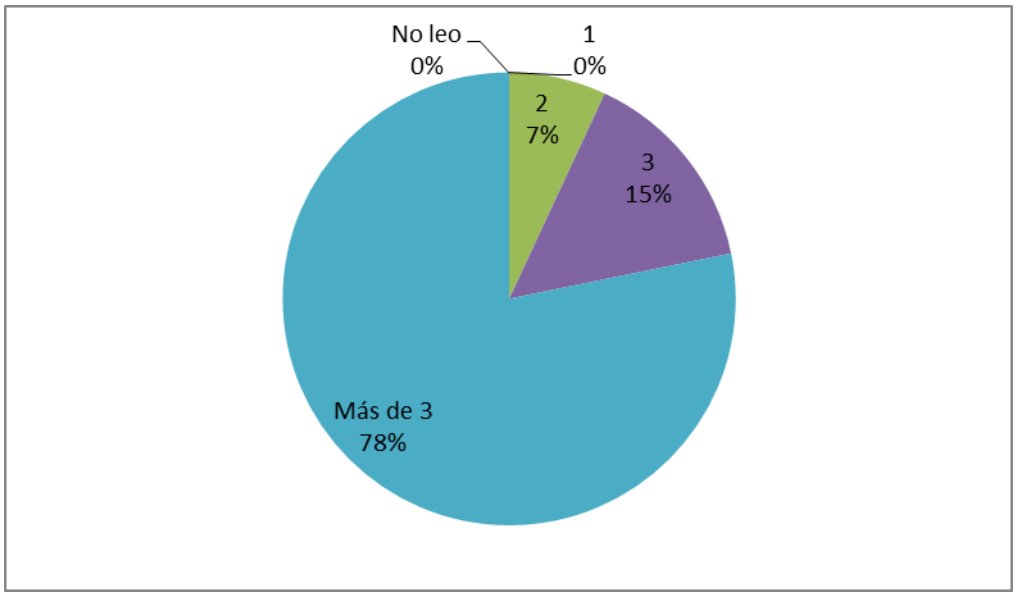


Figura 38: Pregunta 3 Encuesta

La siguiente pregunta, cumple con el objetivo, de comprobar de si se trata certeramente de lectores frecuentes, por lo cual se le consulta la cantidad de libros que leen al mes. Donde el 78% lee más de 3 libros al mes. La totalidad de los datos se presentan en la siguiente tabla.

No leo	1	2	3	Más de 3
0	0	7	15	79

Tabla 50: Pregunta 3 Encuesta

4) ¿Cuál es tu sexo?

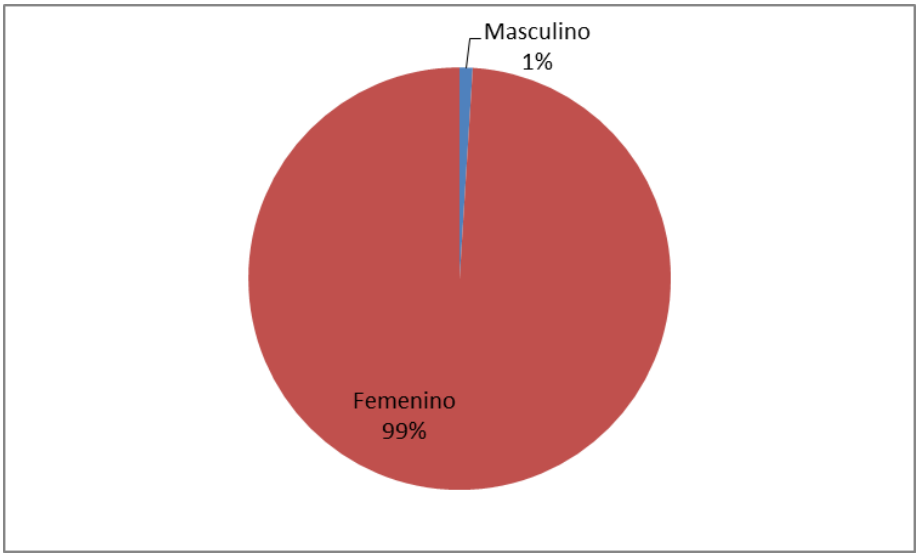


Figura 39: Pregunta 4 Encuesta

La siguiente pregunta, cumple con el objetivo, de determinar el género de los encuestados. Donde el 99% pertenece al género femenino. La totalidad de los datos se presentan en la siguiente tabla.

Masculino	Femenino
1	100

Tabla 51: Pregunta 4 Encuesta

5) "Es un clásico que se debe leer, es fantástico una gran obra."

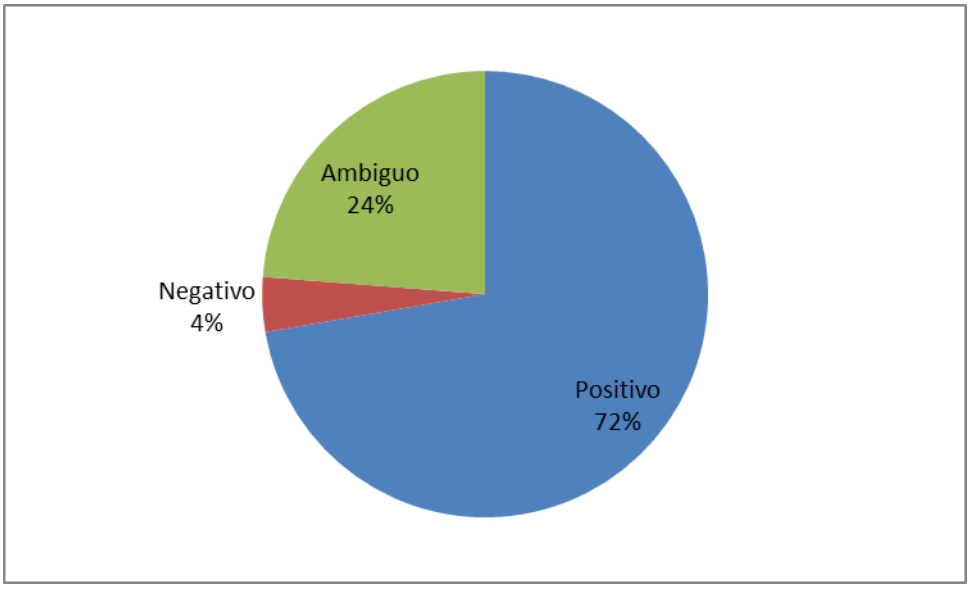


Figura 40: Pregunta 5 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 72%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
73	4	24

Tabla 52: Pregunta 5 Encuesta

6)"Este libro lo leí hace muchos años y lo releí un par de veces más. A quienes les gusta la fantasía mezclado con la magia, léanlo"

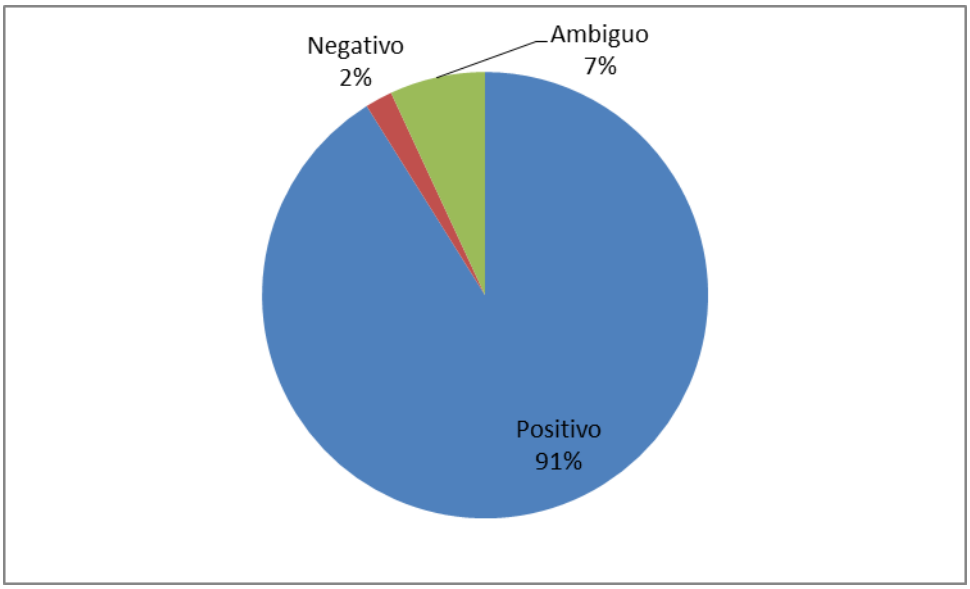


Figura 41: Pregunta 6 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 91%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
92	2	7

Tabla 53: Pregunta 6 Encuesta

7)"Lo tuve muchos años en mi estantería, cuando lo leí me reproche no haberlo hecho antes, excelente"

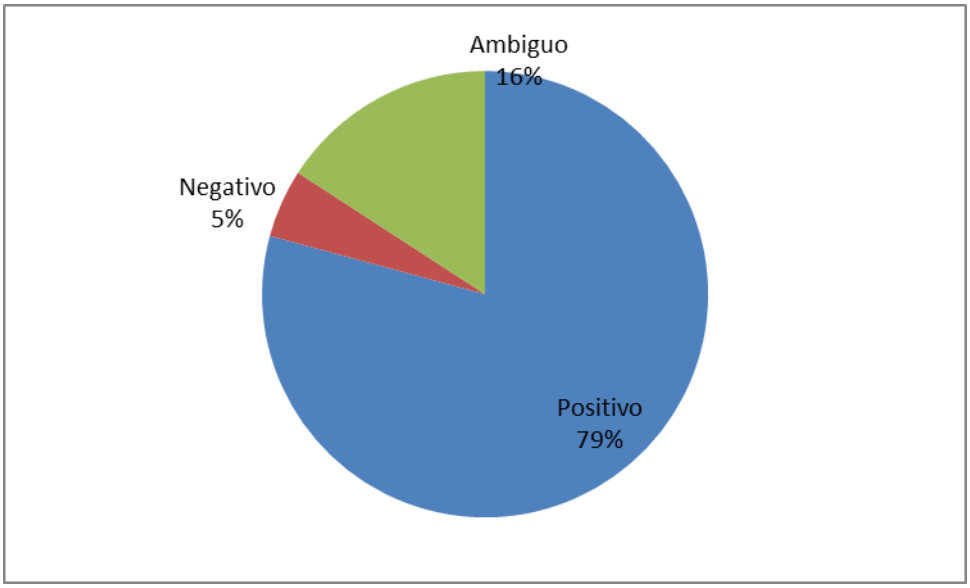


Figura 42: Pregunta 7 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 79%, pero en esta ocasión el clasificador automático lo etiquetó como ambiguo, debido a que solo pudo identificar la palabra *excelente* como positiva y *reproche* como negativa. Dejando un total de 5 palabras desconocidas*: años, estantería, leí, haberlo, hecho*. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
80	5	16

Tabla 54: Pregunta 7 Encuesta

8) "Es mejor que la película...pero sinceramente no es para tanto"

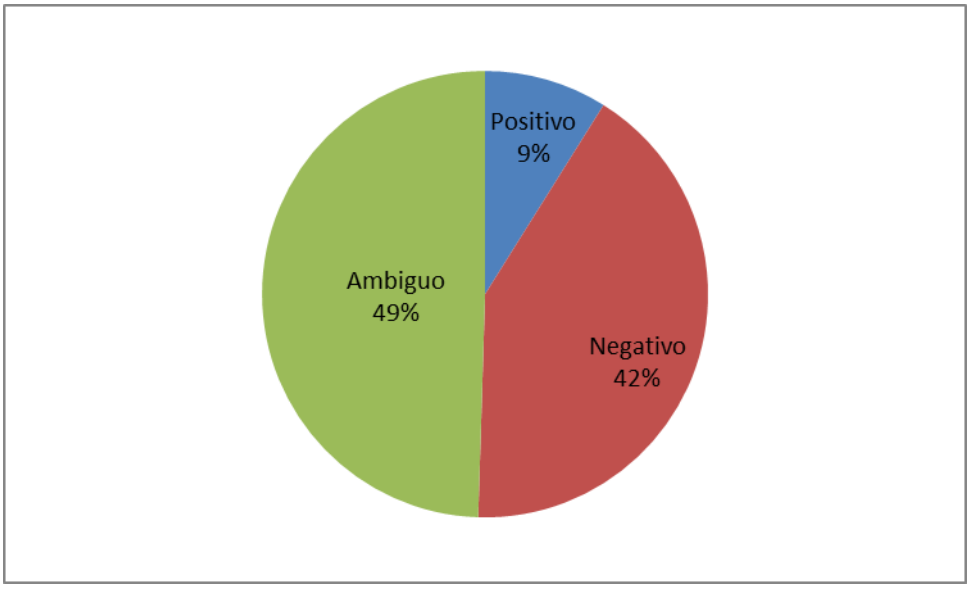


Figura 43: Pregunta 8 Encuesta

En la siguiente pregunta, como se puede observar, existe una diferencia porcentual de un 7%, por lo que se consideraron la polaridad Ambigua y Negativa como válidas. Donde el clasificador automática lo catalogó como positivo, esto se debió que al quitar los stop-words, el comentario quedó como "*mejor película sinceramente*". Las cuales todas palabras positivas. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
9	42	50

Tabla 55: Pregunta 8 Encuesta

9) "No sé qué es más malo si el libro o la película"

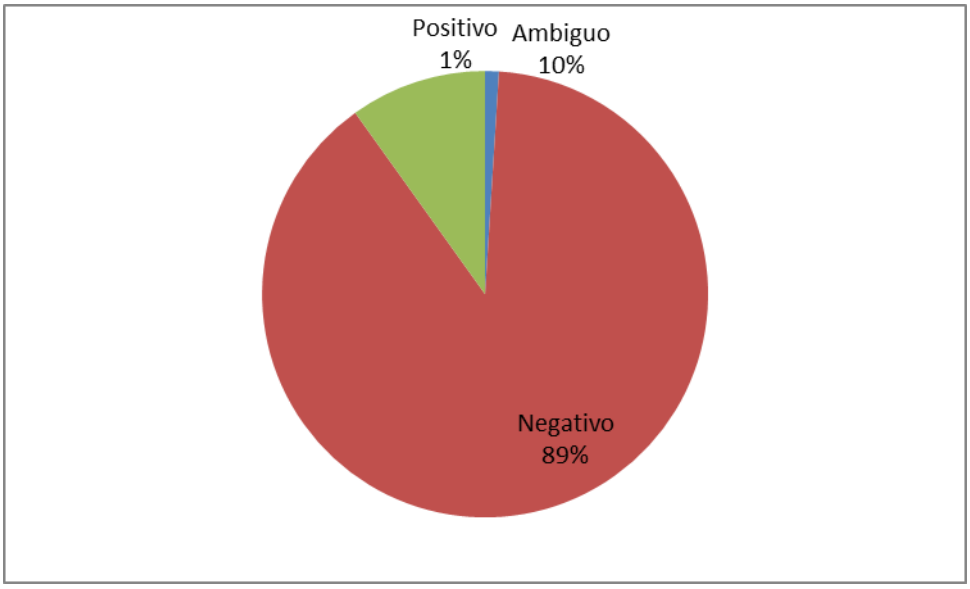


Figura 44: Pregunta 9 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era negativo con un 89%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
1	90	10

Tabla 56: Pregunta 9 Encuesta

10) "Fue mi primer contacto con el universo de Tolkien y, aunque es más infantil, es totalmente recomendable, sobre todo para continuar leyendo El señor de los anillos y El Silmarillion"

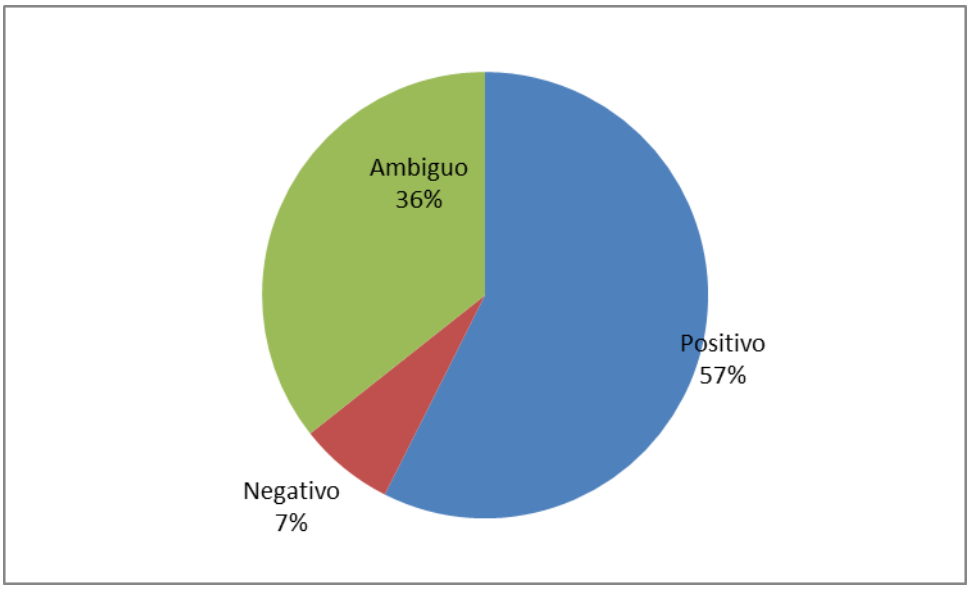


Figura 45: Pregunta 10 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 57%, pero en esta ocasión el clasificador automático lo etiquetó ambiguo, ya que contaba 2 palabras positivas (*totalmente, leyendo*) y 2 negativas (*infantil, continuar*) Por otra parte, se detectaron en este comentario un total de 9 palabras desconocidas (*primer, contacto, universo, Tolkien, aunque, recomendable, señor, anillos, silmarillion*)

La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
58	7	36

Tabla 57: Pregunta 10 Encuesta

11) "El libro no está mal, es entretenido bastante diría yo, pero la verdad me esperaba mucho más....."

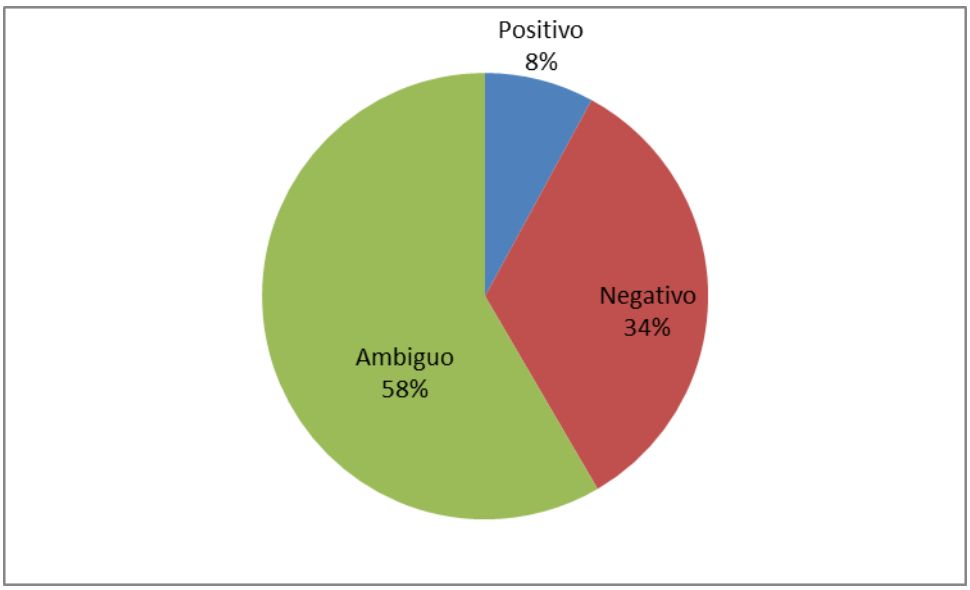


Figura 46: Pregunta 11 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era ambiguo con un 58%, pero en esta ocasión el clasificador automático lo etiquetó como positivo, ya que al quitar los stop-words el comentario quedo de la siguiente forma: *libro mal entretenido bastante diría verdad esperaba.* Donde las palabras positivas corresponden a: *libro, entretenido, verdad, esperaba*. Negativas *mal*. Y finalmente las palabras desconocidas *bastante, diría.*

La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
8	34	59

Tabla 58: Pregunta 11 Encuesta

12) "Magnífico libro para comprender el horror que se vivió en aquellos días, recomendable al cien por cien"

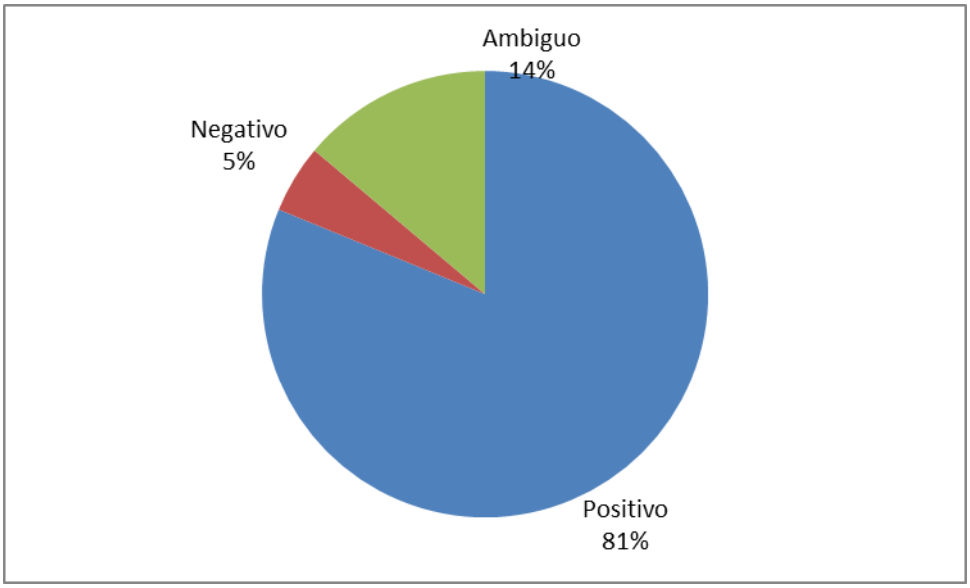


Figura 47: Pregunta 12 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 81%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
82	5	14

Tabla 59: Pregunta 12 Encuesta

13) "Una historia que entretiene al lector. Sencillo y práctico a la hora de leer. un gran éxito de ventas. Una pequeña gran obra."

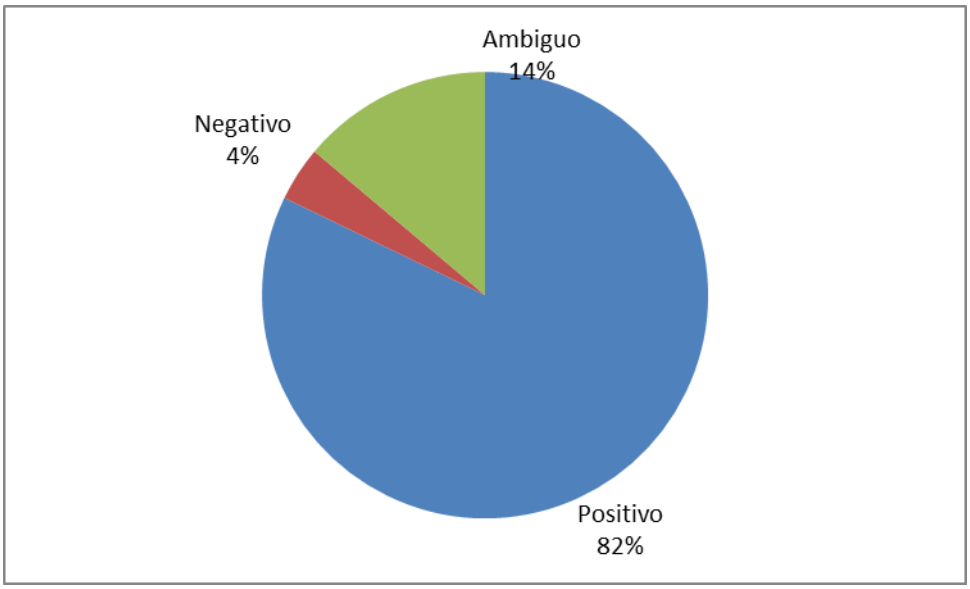


Figura 48: Pregunta 13 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 82%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
83	4	14

Tabla 60: Pregunta 13 Encuesta

14) "A mí me interesó bastante lo leí en una tarde y la verdad me entretuvo, el final muy triste"

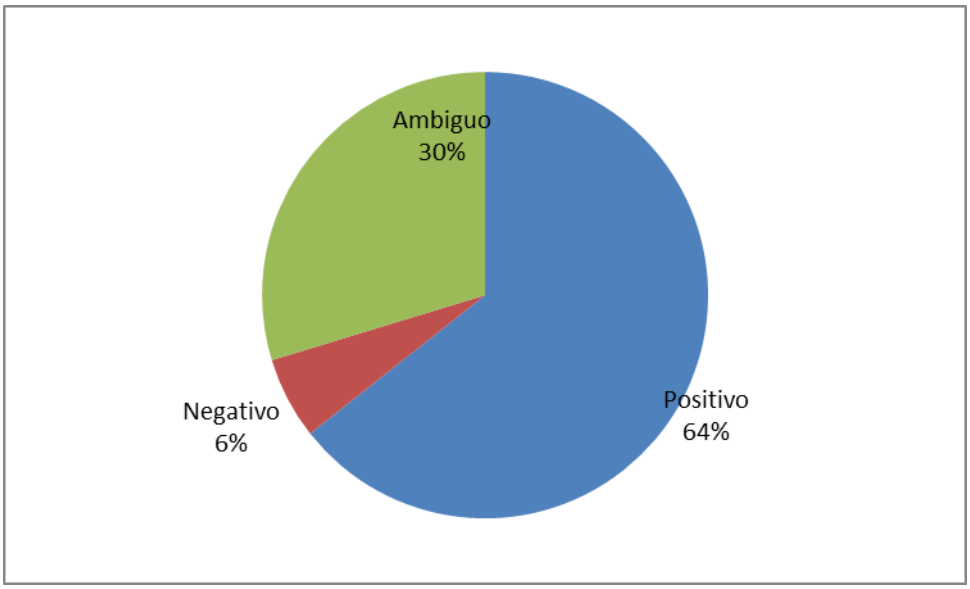


Figura 49: Pregunta 14 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 64%, pero en esta ocasión el clasificador automático lo etiquetó ambiguo, ya que contaba 2 palabras positivas (*interesó, verdad*) y 2 negativas (*final, triste*) Por otra parte, se detectaron en este comentario un total de 4 palabras desconocidas (*bastante, leí, tarde entretuvo*)

La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
65	6	30

Tabla 61: Pregunta 14 Encuesta

15) "Tan sencillo y a la vez tan profundo. Libro para leer en un rato recordándolo toda la vida. Un hombre, un niño y una rosa como base para un aprender sin parar."

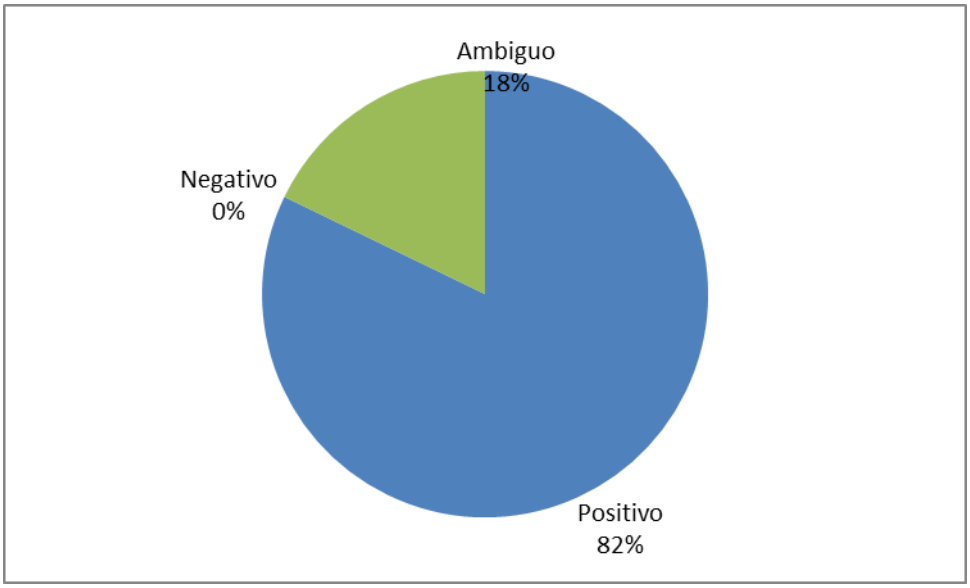


Figura 50: Pregunta 15 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 82%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
83	0	18

Tabla 62: Pregunta 15 Encuesta

16) "Puede que no sea tan bueno como el principito, pero igualmente debería ser de lectura obligatoria"

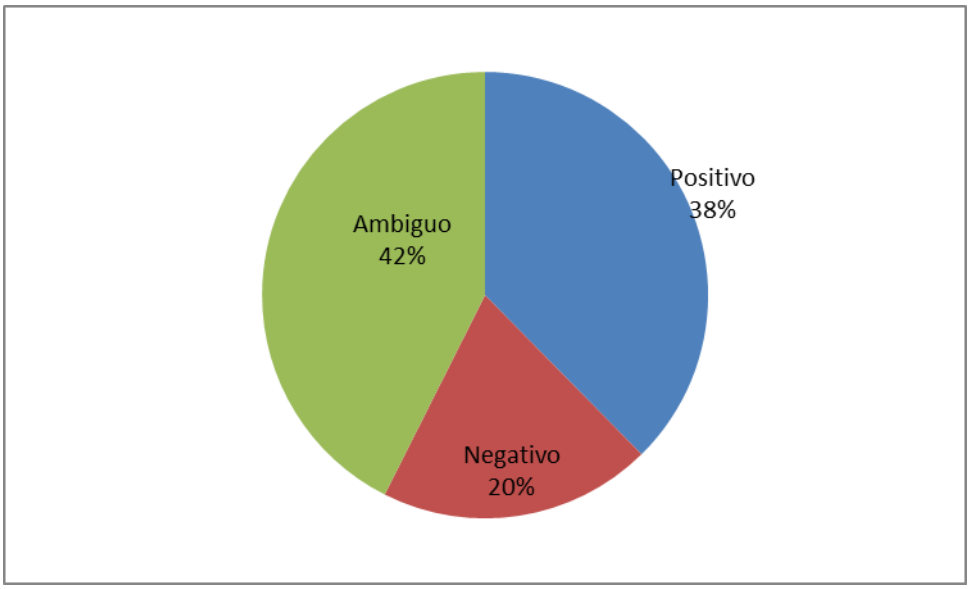


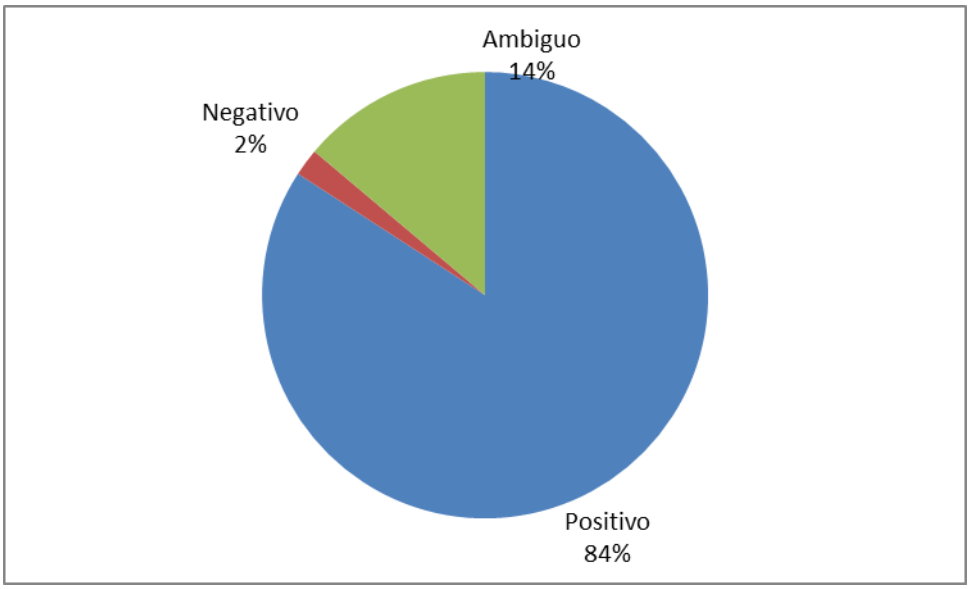
Figura 51: Pregunta 16 Encuesta

En la siguiente pregunta, como se puede observar, existe una diferencia porcentual de un 4%, por lo que se consideraron las polaridades Positivo y Ambiguo como válidas. Donde el clasificador automático lo etiquetó como ambigua. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
38	20	43

Tabla 63: Pregunta 16 Encuesta

17) "Es uno de mis libros preferidos…creo que es el que más veces he leído...recomiendo también, parábola de unas alas"



La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 84%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Figura 52: Pregunta 17 Encuesta

Positivo	Negativo	Ambiguo
85	2	14

Tabla 64: Pregunta 17 Encuesta

18) "Muy buena novela, buena descripción de los personajes, de los pasajes, te mantiene enganchado, no he leído el primero, pues creía que era independiente uno del otro, no me cansa de leerlo."

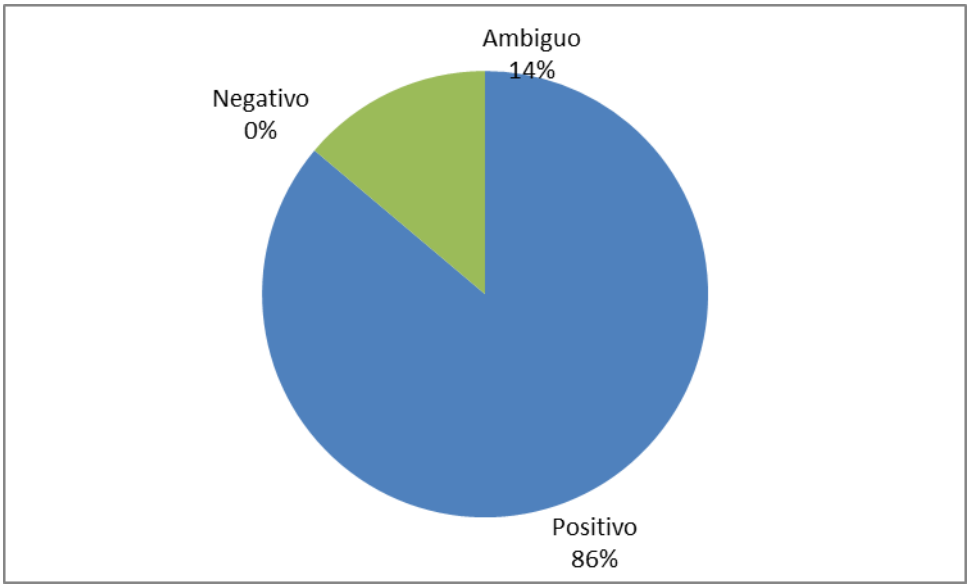


Figura 53: Pregunta 18 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 86%, pero en esta ocasión el clasificador automático lo etiquetó ambiguo, ya que contaba 2 palabras positivas (*buena, buena*) y 2 negativas (*independiente, cansa*). La palabra ambigua (*pasajes*). Y finalmente, se detectaron en este comentario un total de 10 palabras desconocidas (*novela, descripción, personajes, mantiene, enganchado, leído, primero, pues, creía, leerlo*).

La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
87	0	14

Tabla 65: Pregunta 18 Encuesta

19) "Me he leído el código da vinci, ángeles y demonios, y el símbolo perdido y este es el que me pareció más aburrido, me lo acabe a duras penas, no me gustó nada."

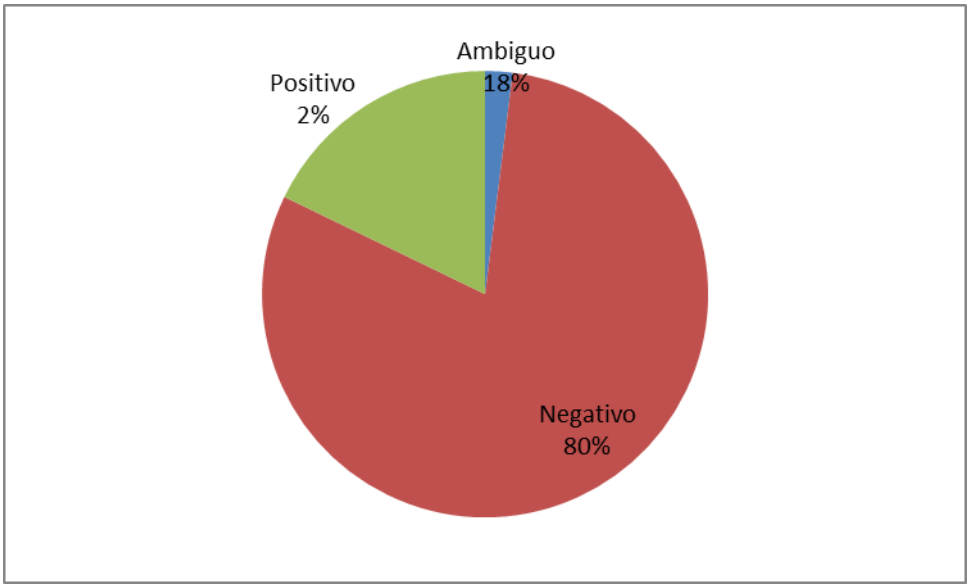


Figura 54: Pregunta 19 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era negativo con un 80%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
2	81	18

Tabla 66: Pregunta 19 Encuesta

20)"Recomendado para quienes gustan del género, sin embargo no me ha gustado lo suficiente para leer las dos obra que me faltan y que completan la trilogía."

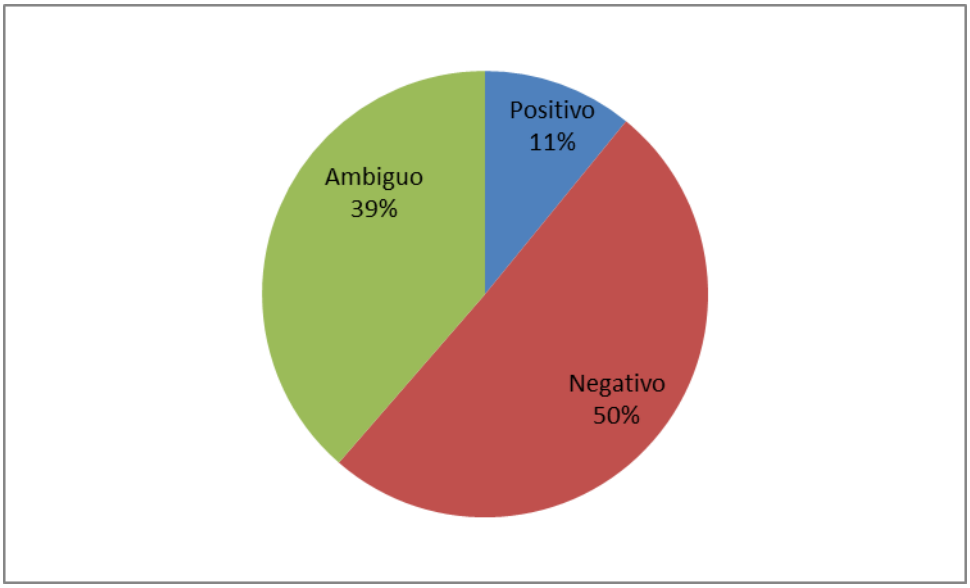


Figura 55: Pregunta 20 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era negativo con un 50%, pero en esta ocasión el clasificador automático lo etiquetó ambiguo, ya que contaba 2 palabras positivas (*gustan, gustado*) y 2 negativas (suficiente*, faltan*). La palabra ambigua (género). Y finalmente, se detectaron en este comentario un total de 7 palabras desconocidas (*recomendado, embargo, leer, dos, obra, completan, trilogía*).

Positivo	Negativo	Ambiguo
11	51	39

Tabla 67: Pregunta 20 Encuesta

21)"Me gustan muchos los libros de temática japonesa, este libro resume la vida de las geishas de una manera que hace que estés allí."

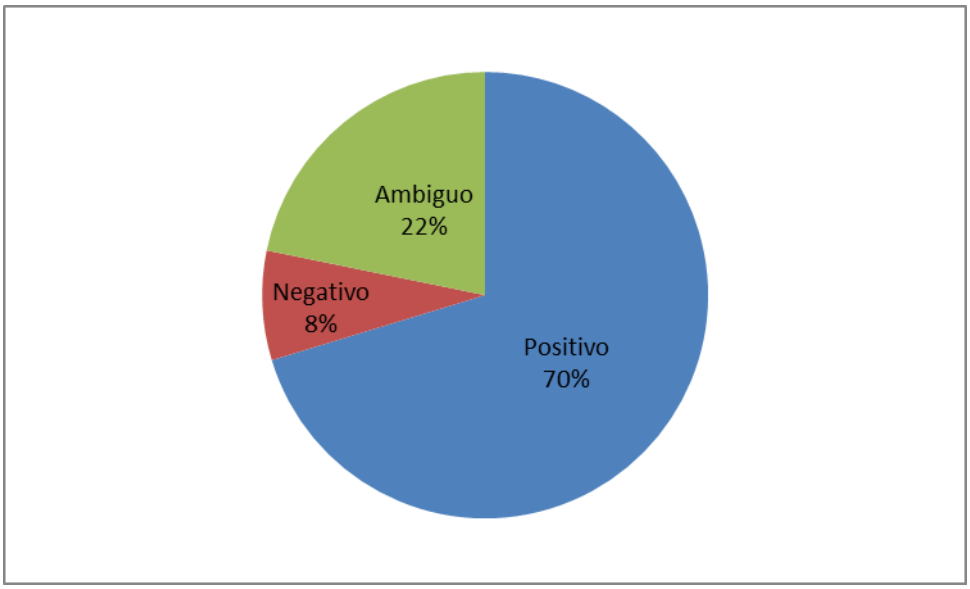


Figura 56: Pregunta 21 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 70%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
71	8	22

Tabla 68: Pregunta 21 Encuesta

22)"Sin duda una grande obra, excelente libro, me gustó bastante y sobretodo el final que me fascinó y me atrapo totalmente "

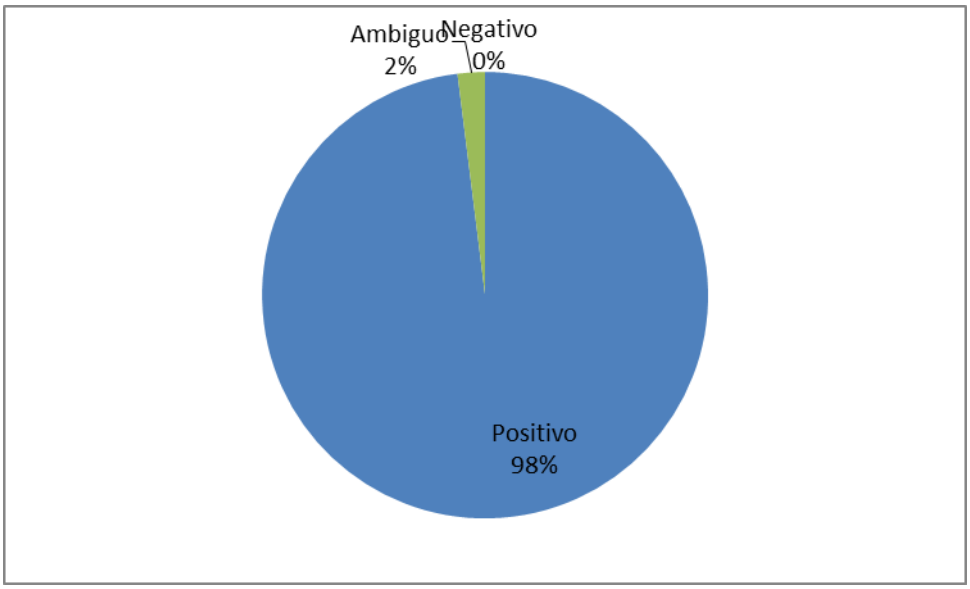


Figura 57: Pregunta 22 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 98%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla.

Positivo	Negativo	Ambiguo
99	0	2

Tabla 69: Pregunta 22 Encuesta

23)"Esta genial el libro!!tiene acción, suspense, comedia, drama y si, sobre todo romance pero sin llegar a ser empalagoso!! "

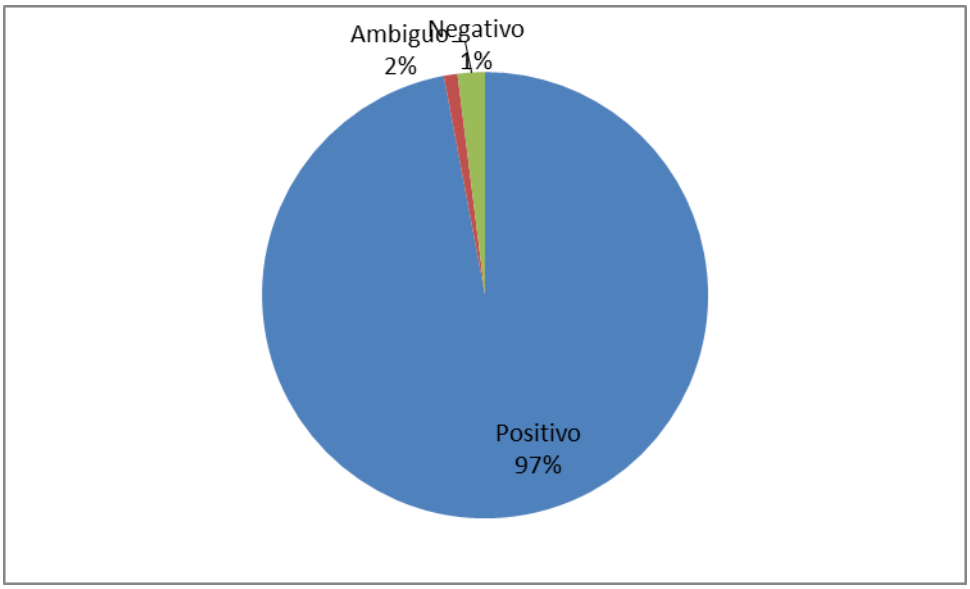


Figura 58: Pregunta 23 Encuesta

Positivo	Negativo	Ambiguo
98	1	2

Tabla 70: Pregunta 23 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 97%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla

24)"El libro se lee en un momento ....porque la historia engancha"

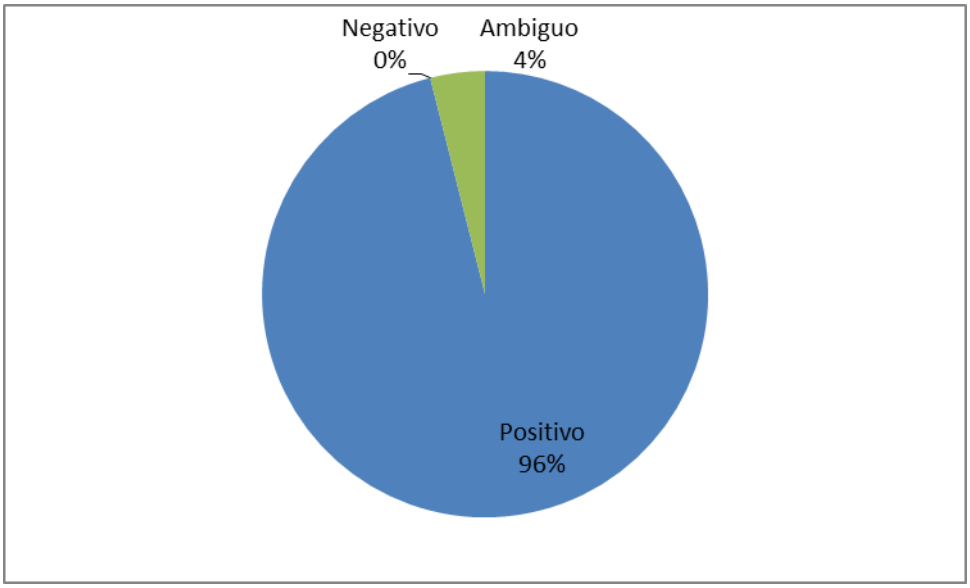


Figura 59: Pregunta 24 Encuesta

La siguiente pregunta el grupo de experto determinó que el comentario era positivo con un 96%, al igual que el clasificador automático. La totalidad de los datos se presentan en la siguiente tabla

Positivo	Negativo	Ambiguo
97	0	4

Tabla 71: Pregunta 24 Encuesta

Al finalizar esta encuesta, podemos concluir que preliminarmente el clasificador automático predijo correctamente la polaridad de 12 comentarios. Lo cual es un resultado favorecedor. Las respuestas de los encuestados se encuentran en un archivo Excel almacenado en el siguiente enlace, presente en el **anexo 11.14.32.**