



UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior de Linares

Trabajo Fin de Grado

SOFTWARE PARA EXTRACCIÓN DE INFORMACIÓN EN BASES DE CONOCIMIENTO DE OPINIÓN.

Alumno: Laura M^a Miñarro Escribano

Tutor: Molina González, María Dolores

Depto.: Ingeniería de Telecomunicación

Septiembre, 2017

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior de Linares

Trabajo Fin de Grado

**SOFTWARE PARA EXTRACCIÓN DE
INFORMACIÓN EN BASES DE
CONOCIMIENTO DE
OPINIÓN.**

Alumno: Laura María Miñarro Escribano.

Tutor: M^a Dolores Molina González.

Depto.: Ingeniería de Telecomunicación.



Firma del autor /a



Firma del tutor /a

Índice General

1.	Resumen	6
2.	Introducción	8
2.1	Estructura del documento	9
2.2	Motivación	11
3.	Descripción	12
4.	Objetivos	14
5.	Estado del arte	15
5.1	iSOL	21
5.2	MCR	22
5.2.1	EuroWordNet	24
5.3	SentiWordNet	25
5.4	CRiSOL	26
6.	Herramientas para la detección de la polaridad	28
6.1	Weka	28
6.2	NLTK	28
6.3	TextPro	30
7.	Materiales y herramientas de desarrollo	32
7.1	Materiales	32
7.2	Herramientas de desarrollo	32
7.2.1	Lenguaje de programación JAVA	32
7.2.2	Netbeans	34
7.2.3	XML	36
8.	Desarrollo del Software	39
8.1	Contenido de las bases de datos	39
8.1.1	MCR	39
8.1.2	SentiWordNet	42
8.1.3	CRiSOL	45

8.2	Generación de CRiSOL_v2	47
8.2.1	Pruebas de funcionamiento del software	56
8.3	Problemas encontrados en el desarrollo	67
9.	Líneas de futuro	68
10.	Anexos	69
10.1	Manual de instalación del entorno de trabajo Netbeans 8.0.1	69
10.2	Manual de funcionamiento de la interfaz de usuario	80
10.2.1	Ejemplo Práctico	85
11.	Índice de Figuras	87
12.	Índice de Tablas	89
13.	Glosario	90
14.	Referencias bibliográficas	92

1. Resumen

En la actualidad, tras la evolución de la web 2.0 se genera de manera continuada una gran cantidad de datos, dando lugar a una fuente ingente de información que por sí sola es difícil de utilizar para poder sacar alguna conclusión.

Es por esto que se han desarrollado diferentes técnicas que faciliten el procesamiento de datos de una forma lo más automatizada posible para que en las diferentes organizaciones, empresas y clientes potenciales puedan así extraer conclusiones y tomar decisiones a partir de estos datos.

El gran volumen de datos existente online, hace que de ser un producto, este pase a convertirse en una materia prima que necesita ser explotada para poder obtener un activo final, que es el conocimiento.

Hoy día, el conocimiento de la información da poder a cualquier empresa u organización, y es por esto que los gobiernos e instituciones invierten gran cantidad de dinero en las TICs, y esto cada vez se encuentra más relacionado con diversos campos como son la minería de datos, inteligencia empresarial y procesos.

La mayor parte de las investigaciones en el ámbito del Análisis de Opiniones (AO) se ha ejecutado en textos escritos en inglés, debido especialmente a la falta de recursos en otros idiomas. Sin embargo, en la actualidad la presencia de otros idiomas diferentes del inglés en Internet crece cada vez más, por lo que es necesario orientar la investigación a otros idiomas o a sistemas que sean capaces de funcionar independientemente de la lengua utilizada, y no sólo a textos escritos en inglés.

Este Trabajo Fin de Grado tiene como objetivo determinar la orientación (positiva o negativa) de una frase, un documento, una característica o incluso una palabra, para así medir el grado de la polaridad expresada en textos por el autor. Para lograr alcanzar dicho objetivo y así contribuir a la mejora de los sistemas existentes de clasificación de la polaridad, se propone desarrollar e integrar recursos lingüísticos adaptados al español, tercer idioma más usado por los usuarios web, como podemos ver en la Figura 1, para hacer posible la clasificación de opiniones textuales en la web 2.0 basada en Orientación Semántica, tarea englobada dentro del Procesamiento del Lenguaje Natural (PLN).

10 idiomas más usados en Internet

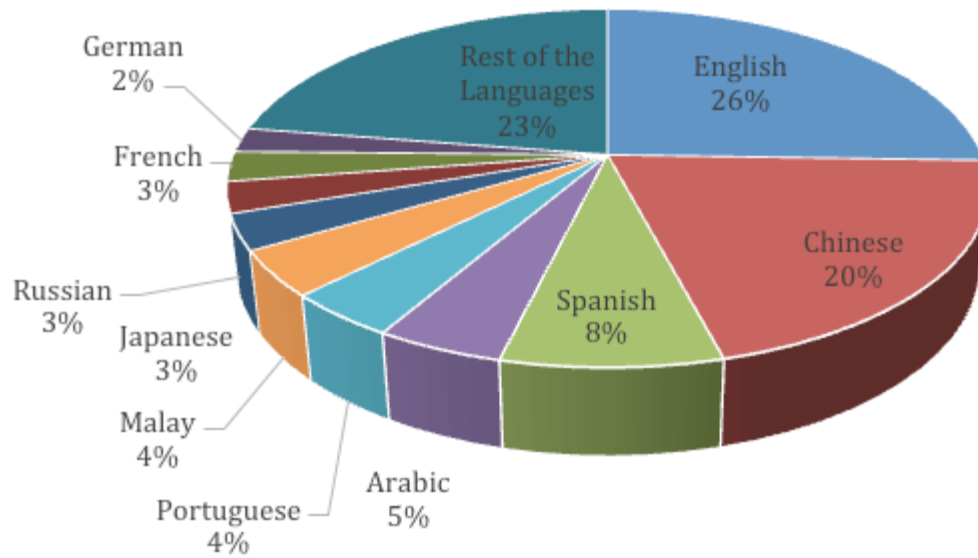


Figura 1. Millones de usuarios en los 10 idiomas más usados en Internet (Mayo 2017)¹

¹ <https://www.eldinero.com.do/41932/casi-300-millones-de-hispanohablantes-poseen-acceso-a-internet/>

2. Introducción

El análisis de sentimientos, conocido también como minería de opiniones, se basa en la utilización de tecnologías de PLN, lingüística computacional y analítica de textos para extraer e identificar información subjetiva de contenido de diferentes tipos.

Mediante este proceso se puede determinar el tono emocional que hay detrás de una serie de palabras, determinando si un acto de habla o una frase contiene una opinión, negativa o positiva, sobre un concepto o sobre una entidad concreta. Esto es utilizado para intentar entender las opiniones, emociones y actitudes expresadas en una mención online.

El uso del análisis de sentimientos es extremadamente eficaz en la monitorización de las revisiones de productos en foros, los medios sociales, y otros sitios web ya que éste nos permite que nos hagamos una idea de la opinión pública general sobre diversos temas.

Tiene importantes y numerosos beneficios, y la habilidad de extraer información de datos es una práctica que diversas organizaciones a nivel mundial ya están adoptando.

Esto no quiere decir que el análisis de sentimientos sea una ciencia perfecta, puesto que el lenguaje humano es complicado y enseñar a una máquina a analizar las diferentes variaciones culturales, matices gramaticales, ironía, faltas ortográficas y jergas de las menciones online es un procedimiento difícil. De hecho, en función del contexto, las cifras de acierto pueden andar decenas de puntos por debajo de la perfección. Un ejemplo de esto es, si pides a seis personas que analicen el sentimiento de un conjunto de frases, es casi imposible que coincidan en la mayoría de estas frases más de cinco de estas personas. Cuando nos referimos a la precisión de una herramienta automatizada, lo hacemos comparándola con la de estos grupos de control humanos, y es por esto por lo que no es posible alcanzar el 100% de eficacia real [1].

Hay sistemas sencillos que están limitados a leer una frase y buscar en ella una palabra o palabras que tienen registradas en su base de datos como negativas, positivas o neutras.

Si por ejemplo aparece una palabra buena como “maravilloso”, esto es una opinión positiva. Si por el contrario aparece una palabra mala como “decepcionado”, es una opinión negativa.



Figura 2. Sentimentanalysis

2.1 Estructura del documento

El documento aquí presente conforma la memoria final de mi trabajo fin de grado para la titulación Grado en Ingeniería Telemática, cursada en la Universidad de Jaén.

En este documento se mostrará la metodología utilizada para desarrollar un software capaz de automatizar la extracción de información de bases de conocimiento de opinión para la creación de un recurso léxico que se utilizará en el análisis de opinión.

La memoria se ha estructurado en 9 apartados que tratan de exponer el estudio que se ha realizado, 2 anexos, un glosario y una bibliografía. Con la intención de que los lectores tengan una visión inicial de cada uno de ellos, se van a exponer los elementos que la componen:

En los cuatro primeros apartados, se realiza un Resumen, una Introducción donde se explica brevemente que es el análisis de sentimientos, su uso y beneficios en la monitorización de los medios sociales, foros y otros sitios web, una Descripción donde se explica el proceso que se llevará a cabo para la realización del trabajo y unos Objetivos donde hacemos mención a las tareas necesarias para cumplirlos.

En el apartado número 5, se encuentra el Estado del arte, donde hacemos un repaso general de la evolución del análisis de sentimientos. Además, explicamos los lexicones utilizados para llevar a cabo la realización del trabajo fin de grado.

En el apartado número 6 se describen algunas de las Herramientas para la detección de la polaridad (WEKA, NLTK y TextPro), profundizando un poco en cada una de ellas.

El apartado número 7 describe los Materiales necesarios para la puesta en marcha del trabajo y las herramientas de desarrollo empleadas.

En el apartado número 8 podemos encontrar el Desarrollo del Software, donde veremos el contenido de las bases de datos y la generación de CRiSOL_v2 con diversas pruebas de funcionamiento del software y algunos de los problemas encontrados durante el desarrollo del mismo.

Por último, en el apartado número 9, describiremos algunas Líneas de futuro,

El primer anexo consta de un manual básico de instalación del entorno de trabajo NetBeans 8.0.1.

El segundo anexo consta del manual de funcionamiento de la interfaz de usuario.

Por último, un glosario donde aparecen todos los términos utilizados en el trabajo fin de grado y una bibliografía donde se muestran todas las referencias utilizadas.

2.2 Motivación

El conocer la opinión de los demás en diferentes áreas de interés ha sido siempre una fuente importante de información. A raíz del auge de Internet, se ha visto como se ha masificado el uso de páginas de opiniones, como son los foros sobre determinadas áreas, blogs e incluso redes sociales, con lo cual se ha abierto una oportunidad para saber lo que las personas opinan con sólo navegar por Internet.

La principal motivación a la hora de elegir este Trabajo Fin de Grado ha sido la de querer realizar un proyecto distinto a lo que he venido desempeñando durante mi trayectoria académica, ya que me ha despertado un gran interés el desarrollar una herramienta que me permita profundizar en el conocimiento del PLN. En definitiva, la motivación radica en el aprendizaje continuado, en dedicar la última oportunidad que brinda la Universidad de trabajo personal para así aprender algo nuevo y descubrir la mejor manera de resolverlo.

3. Descripción

Dentro del campo de investigación de la minería de datos y haciendo uso del análisis de textos, la lingüística computacional y el PLN se encuentra una disciplina a la que se le llama análisis de sentimientos (sentiment analysis), también conocida como análisis de subjetividad (subjectivity analysis) o minería de sentimientos (opinión mining) [2].

El presente trabajo fin de grado se centra en la clasificación de la polaridad de comentarios en Internet y presenta la mejora del recurso léxico CRISOL (Combined Resources in iSOL) [3], llamado CRiSOL_v2.

El objetivo de dicha clasificación de la polaridad es el de obtener una puntuación entre 0 y 1 que indique si un texto expresa una opinión negativa, positiva o neutra.

Este nuevo recurso toma como base la lista de palabras de opinión iSOL (Improved Spanish Opinion Lexicon) [4], a la cual incluye los valores de polaridad de los synsets de SentiWordNet (SWN) [5]. Este es un recurso léxico explícitamente diseñado para apoyar la clasificación de sentimientos y las aplicaciones de minería de opinión en inglés.

En este caso se ha utilizado SWN 3.0, que es una versión mejorada de SWN 1.0, un recurso léxico disponible para motivo de investigación.

Como SWN no es un recurso para español, se ha tenido que usar como pivote la versión española de WordNet incluida en el Repositorio Central Multilingüe (MCR). Tanto SWN 1.0 y 3.0 son el resultado de anotar automáticamente todos los synsets de WordNet según sus grados de positividad, neutralidad y negatividad.

Los nuevos resultados obtenidos con CRiSOL_v2 superan los obtenidos por los lexicones base iSOL y SWN por separado.

El proceso que se llevará a cabo es el siguiente:

- Por un lado tendremos CRiSOL, en el cual buscaremos la palabra que queremos analizar.
- Por otro, tendremos MCR el cual enlazaremos con CRiSOL mediante la palabra anteriormente citada. Aquí junto a ésta, aparecerá un identificador de 8 dígitos y si una palabra se repite 4 veces, habrá 4 identificadores diferentes.
- Una vez encontrados todos los identificadores asociados a una misma palabra, enlazamos MCR con SWN y veremos las puntuaciones (tanto positivas como negativas) de cada una de las palabras.

El resultado final de este Trabajo Fin de Grado será el de un servicio que realice esta labor y que nos permita tener con cada una de las palabras, diversas puntuaciones desde distintos puntos de vista.

4. Objetivos

El objetivo principal de este Trabajo Fin de Grado, es desarrollar un software que permita automatizar la extracción de información de bases de conocimiento de opinión para la creación de un recurso léxico que se utilizará en el análisis de opinión en español.

Como hemos mencionado anteriormente, se pretende obtener una clasificación de la polaridad de comentarios en Internet a través de un recurso mejorado de CRiSOL, al que se le ha llamado CRiSOL_v2.

Para cumplir este objetivo necesitaremos abordar las siguientes tareas:

- Estudiar las diferentes bases de conocimiento de opinión que vamos a emplear en el desarrollo del trabajo.
- Familiarización con el lenguaje de programación escogido para la realización e implementación del software.
- Estudiar el funcionamiento del entorno de desarrollo integrado libre escogido.
- Desarrollo de una interfaz de usuario utilizando el generador del entorno de desarrollo escogido.
- Realización de diversas pruebas de funcionamiento del software diseñado.

5. Estado del arte

El análisis de sentimientos es un área de estudio que analiza las emociones, opiniones, sentimientos, actitudes y evaluaciones de las personas hacia determinados productos, empresas, servicios y/o temas concretos.

Actualmente existen una gran cantidad de nombres para hacer alusión a este ámbito como son: minería de opinión, extracción de opinión, minería de sentimiento, análisis de emociones, etc.

Las primeras apariciones de la terminología de análisis de sentimientos y de minería de opinión aparecieron en los estudios realizados por Kushal et al. en el año 2003 [6].

Las primeras investigaciones relativas a opiniones y a sentimientos aparecieron a partir del año 1999, como es el caso del estudio que elaboró Wiebe [7]. Como ejemplo de algunas tareas típicas dentro de la minería de textos tenemos:

- **Categorización de textos** (Text categorization/Document classification):

Se busca asignar un documento o un texto, de forma particular o numérica, a una o más clases o categorías.

Una de sus diversas aplicaciones podría ser el análisis de sentimientos, etc.

- **Similitud semántica de textos** (Semantic Text similarity):

Se captura el grado de similitud semántica de dos fragmentos de texto.

Su principal aplicación son los buscadores de Internet, dentro del campo de búsqueda y la recuperación de información.

- **Resumen de documentos** (Document summarization):

Se trata encontrar una agrupación de datos representativos del total del documento, identificando sus partes más informativas.

Uno de sus principales servicios son los buscadores de internet.

- Extracción de conceptos o entidades (Named entity recognition):

Se pretende localizar y clasificar los componentes de un texto dentro de unas clases predefinidas, como son nombres de personas, empresas, tantos por ciento, lugares, etc.

La identificación de expresiones importantes para crear vínculos con Wikipedia es una aplicación actual de esta técnica.

- Análisis de sentimientos (Sentiment Analysis/Opinion mining):

Su objetivo principal es determinar la polaridad de un documento o de una frase, tratando de descubrir la actitud del creador en base a las probables emociones o evaluaciones incluidas en el documento. Las etiquetas más frecuentes para la clasificación de la polaridad son: positiva, objetiva o neutra.

Con el crecimiento de los medios de comunicación social (web 2.0) en Internet, estas aplicaciones cobran especial importancia, sobre todo desde el punto de vista empresarial, a la hora de analizar cualquier tipo de expresiones online.

Desde el año 2000, la búsqueda sobre el análisis de opiniones de sentimientos ha crecido considerablemente. Esto es debido principalmente al tremendo número de campos de aplicación que existen en la actualidad, así como al incremento de opiniones recogidas en blogs, foros, redes sociales, etc., que hacen posible disponer de bastantes datos para así avanzar en las investigaciones.

Hoy día este es un campo muy abierto con muchos desafíos por resolver. La simple tarea de catalogar el sentimiento de un texto como positivo o como negativo es a veces tan complejo que diversas personas pueden ponerse o no de acuerdo para otorgar una clasificación definitiva. Esto sucede ya que un mismo texto o frase puede ser interpretado de diferentes formas, en función de factores culturales, personales o incluso el idioma o dialecto [2].

Es por esto que tanto las opiniones como los sentimientos son subjetivos, y para poder extraer la información objetiva lo más importante es poder disponer de un conjunto amplio de opiniones. La clasificación de polaridad de sentimientos posiblemente sea la tarea más estudiada dentro del análisis de sentimientos.

Este problema de clasificación puede abordarse con **aprendizaje automático** supervisado o no supervisado. El **enfoque supervisado** es la técnica de aprendizaje automático que requiere de un conjunto de datos etiquetados para la generación de un modelo con la capacidad de clasificar nuevos documentos. Si lo vemos de una manera más técnica, este precisa de un conjunto de ejemplos de los que se conozca la clase a la que pertenecen. Deben representarse por medio de un conjunto de características, que se emplean para la construcción de un modelo estadístico, el cual modela la distribución que siguen los datos, de manera que cuando se toma un ejemplo sin clase, pero representado de igual forma que el conjunto de entrenamiento, el algoritmo de clasificación es capaz de determinar a qué distribución de datos se ajusta más, y en función de esto le asigna la clase de la distribución a la que más se asemeja [8].

El enfoque supervisado parte de un cúmulo de datos anotado con la polaridad del documento. Esto puede hacerse de forma automática, como por ejemplo cuando los documentos llevan una evaluación numérica, como sucede en los portales de compra de Internet, o de forma manual por medio de anotadores humanos.

La anotación automática tiene una ventaja y es que se puede realizar de forma colectiva y así conseguir un conjunto más grande de datos, pero ocurre que esta anotación pierde precisión frente a la manual. Por el contrario, la anotación manual gana precisión, pero pierde capacidad anotadora, y el conjunto de datos tiende a ser bastante más pequeño [8].

En el trabajo [9], se hizo el primer estudio que adaptó este enfoque para la clasificación de películas en dos clases, positiva y negativa. Aquí los textos eran representados como vectores de palabras que se utilizaban como variables independientes para la clasificación, efectuada con SVM (Support Vector Machine).

Después se han seguido diversas líneas de investigación en cuanto a otros algoritmos de aprendizaje automático y a diferentes maneras de formar un conjunto de variables independientes efectivo, ya que se ha demostrado que esto es fundamental para el éxito de la clasificación.

Como variables independientes suelen usarse:

- Términos del documento junto con su posición y/o su relevancia medida en términos de su valor TF-IDF (Term frequency – Inverse document frequency) o calculada mediante técnicas estadísticas.
- Categorías gramaticales de los vocablos, como por ejemplo los adjetivos, buenos indicadores de subjetividad.
- Expresiones y palabras con sentimiento asociado directamente: feo, bonito, bueno, malo, etc.
- Palabras inversoras de la polaridad: no, nunca, debería, etc. Estas palabras normalmente cambian la polaridad del término que lo sigue.

En cuanto al **enfoque no supervisado**, la idea de éste consistía en comparar las palabras consecutivas (n-gramas) con patrones sintácticos prefijados, por ser los más propensos a ser utilizados para expresar opiniones. Este enfoque no necesita de un entrenamiento previo. Los patrones se formaban en base a la categoría gramatical de las diferentes palabras. A continuación, a estos n-gramas se les asociaba una polaridad en base a su distancia a la palabra positiva “maravilloso” y a su distancia a la palabra negativa “pobre”, y se calculaba la orientación semántica del documento como la media de todas las polaridades conseguidas para los n-gramas.

Otros de los enfoques más actuales están basados en la utilización de lexicones, que consisten en una colección de términos y frases asociados con la polaridad y su intensidad. La clave para desarrollar la clasificación se basa en la identificación de los términos dentro del documento, y es a partir de ahí cuando se completa la técnica con un tratamiento de las palabras que modifican la polaridad o la invierten. Una vez esto, se computa la polaridad general del documento en función de las polaridades de los términos encontrados.

Por sus diversas características, tanto el enfoque supervisado como el no supervisado tienen sus ventajas y sus inconvenientes, si bien con la que mejores resultados se consigue es con la aproximación basada en un enfoque supervisado.

Con la mejora del recurso CRiSOL, se pretende dar un mejor apoyo a un enfoque no supervisado en el análisis de sentimientos en español.

Además del método de aprendizaje automático usado, se pueden considerar tres niveles de profundidad para el análisis de sentimientos [10]:

- Documento: El dilema que se aborda en este nivel es cómo hacer la clasificación general del documento. Se asume que cada documento expresa opiniones sobre una sola entidad, así como que cada opinión es emitida por un solo autor. Un ejemplo de esto son las opiniones sobre un producto en concreto. Es por esto que cuando se comparan diversas entidades este enfoque es insuficiente.
- Frase: Se observa cada frase como una unidad independiente y se asume que cada frase solamente deberá contener una sola opinión. Esto está altamente relacionado con la Clasificación de Subjetividad, que consiste en determinar si una frase es subjetiva o es objetiva.
- Entidad y Aspecto: En este nivel de análisis es en el que se consigue extraer más información de las opiniones. Se centra en la opinión directamente, bajo la idea de que una opinión está compuesta de un sentimiento (positivo o negativo) y un objetivo. Estos objetos se suelen representar por medio de entidades o por medio de sus atributos, que juntos forman el aspecto a analizar. Por ejemplo, en la opinión “por teléfono es muy improbable contactar con el centro, por Internet se contacta muy bien”, esta se podría referir a dos atributos: Contacto telefónico y contacto a través de Internet, teniendo una polaridad negativa el primero y una polaridad positiva el segundo, respectivamente.

Este problema de análisis se divide en varias subtareas:

- Identificar las entidades y los atributos.
- Clasificar sus polaridades.
- Clasificar la polaridad general.

Dichas subtareas siguen siendo problemas abiertos dentro del PLN, y es por esto que este nivel de análisis es el que plantea mayores retos en la actualidad.

A continuación, nombraremos algunos trabajos que contribuyen al avance del Análisis de Opinión en español mediante la generación y creación de lexicones enfocados a dicho idioma.

SO-CAL (Semantic Orientation-CALculator) [11]. Este lexicón es posiblemente el sistema más destacado en AO en español. Este sistema al igual que SO-CAL para textos en inglés (aunque para este idioma no tuvo gran importancia), además de solucionar la orientación semántica guardada a nivel individual en sustantivos, adjetivos, adverbios y verbos, incluye modificadores de la polaridad como son la negación o los intensificadores (muy, poco, bastante...), también es capaz de localizar y eliminar el sentimiento reflejado en el contenido no fáctico del texto, representado, por ejemplo, mediante expresiones condicionales o subjuntivas.

SEL (Spanish Emotion Lexicon) [12] es un lexicón que está formado por 2.036 palabras que fueron anotadas y analizadas manualmente. Estas palabras fueron etiquetadas con las categorías emocionales básicas: alegría, tristeza, miedo, enfado, sorpresa y repulsión y se le asignó un porcentaje de probabilidad de ser usada con un sentido emocional. A este dato se le llamó FPA (Factor de Probabilidad de uso Afectivo). El lexicón está disponible libremente.

ElhPolar [13]. Este es un lexicón que fue creado a partir de diferentes fuentes de datos. Por una parte, se realizó una traducción automática del lexicón de polaridad en inglés OpinionFinder a través de un diccionario bilingüe Inglés-Español. Las traducciones ambiguas se pudieron resolver de forma manual. Como segunda fuente del lexicón, las palabras se extrajeron automáticamente de un corpus generado a partir de tweets. Así mismo, introducen una lista de vocabulario coloquial producido mediante la recopilación de palabras a partir de dos fuentes: “Diccionario de jerga y expresiones coloquiales” y el vocabulario coloquial editado por los diferentes usuarios en una web de crowdsourcing.

ML-SentiCon (Multi-Layered Multilingual Sentiment lexiCon) [14]. Es un método automático que se utiliza para la construcción de lexicones de polaridades semánticas a nivel de lema, para el español, inglés y para catalán, gallego y euskera (otras tres lenguas oficiales en España). Para generar estos lexicones se reprodujo el método original empleado para la construcción de SWN 3.0 añadiéndole unas mejoras, que refleja de manera muy significativa las mejores estimaciones de la positividad y negatividad, de acuerdo con sus evaluaciones.

El método se divide en dos partes notoriamente diferenciadas: un primer cálculo individual de la polaridad, y un segundo cálculo global de la polaridad a partir de los valores conseguidos en la primera parte. Los lexicones fueron estructurados en varias capas, lo que permite seleccionar distintos compromisos a priori entre la cantidad de palabras disponibles y la exactitud de las estimaciones. En total se han creado 8 capas.

Por último realizaremos un breve repaso a cada una de las bases de conocimiento de opinión necesarias para la realización del Trabajo Fin de Grado.

5.1 iSOL

iSOL (Improved Spanish Opinion Lexicon) [4] es un lexicón que fue generado a partir de una traducción automática del inglés al español del lexicón de Bing Liu generando el recurso SOL (Spanish Opinion Lexicon). El recurso iSOL se obtuvo después de llevar a cabo una revisión manual sobre SOL con el fin de mejorar la lista final de las palabras de opinión.

Por un lado, debido a la gramática española, tenemos por ejemplo que mientras un adjetivo inglés no tiene ni género ni número, y por lo general, es representado por un solo término, al adjetivo español le corresponde hasta cuatro posibles palabras traducidas al español, dos para el género (masculino o femenino) y dos para el número (singular o plural).

Por otra parte, siguiendo la filosofía de Bing Liu [10] se introdujo en las listas algunas palabras mal escritas ya que aparecen con mucha frecuencia en el contenido de los medios de comunicación social, como por ejemplo kaput, pillín o coñacete.

Finalmente iSOL se compone de 2.509 palabras positivas y 5.626 palabras negativas. Por lo tanto, este lexicón español contiene 8.135 palabras de opinión.

5.2 MCR

MCR (Multilingual Central Repository) [15] es un recurso lingüístico a gran escala que puede ser usado en procesos semánticos que necesitan gran cantidad de conocimiento lingüístico. MCR integra en el mismo marco de trabajo de EuroWordNet, diversas versiones de WordNet para diferentes lenguas, inglés, español, vasco, catalán y gallego.

Por otro lado, WordNet es una base de datos léxica que proviene del inglés, que agrupa palabras en inglés en conjuntos de synsets, y proporciona definiciones generales y cortas almacenando las relaciones semánticas entre los sinónimos en su conjunto. Su propósito es producir un diccionario cuyo uso sea más intuitivo, y soportar las aplicaciones de ingeniería artificial y los análisis automáticos de texto. Las herramientas del software y la base de datos pueden ser utilizadas y descargadas libremente, además, la base de datos puede ser consultada en línea.

Volviendo al recurso lingüístico MCR, los synsets han sido contruidos siguiendo el modelo propuesto por EuroWordNet, en los cuales los WordNet se enlazan mediante un índice entre lenguas (InterLingual Index-ILI). Por medio de este ILI los lenguajes están conectados, haciendo posible ir desde una palabra de un idioma a otras palabras similares traducidas a otros idiomas gracias a las asignaciones generadas automáticamente entre las versiones de WordNet.

Este hecho es el que nos permite enlazar SWN para el idioma español utilizando el ILI de MCR en el SWN en inglés. La versión final de MCR contiene alrededor de 1,6 millones de relaciones semánticas entre los synsets, siendo la mayoría de ellos adquiridos mediante métodos automáticos. Este recurso está en continuo crecimiento siendo la última versión disponible MCR 3.0.

El MCR 3.0 se distribuye bajo dos licencias diferenciadas [15]:

1. Los datos de sincronización y relación de WordNet en inglés, contenidos en la carpeta engSW, se distribuyen bajo la licencia original de WordNet. Puede encontrarse en <http://wordnet.princeton.edu/wordnet/license>
2. Todos los demás datos de este paquete se distribuyen bajo la licencia de atribución 3.0 Unported (CC BY 3.0). Puede encontrarse en <http://creativecommons.org/licenses/by/3.0>

En la siguiente figura, se muestra un pequeño fragmento de la base de datos del MCR:

a_capella	1	spa-30-00001740-r	r	99	NULL	-----
aspirar	6	spa-30-00001740-v	v	99	NULL	-----
respirar	1	spa-30-00001740-v	v	99	NULL	-----
entidad_fisica	1	spa-30-00001930-n	n	99	NULL	-----
incapaz	1	spa-30-00002098-a	a	99	NULL	-----
abstracción	5	spa-30-00002137-n	n	99	NULL	-----
cosa	11	spa-30-00002452-n	n	99	bier NULL	-----
cosa	4	spa-30-00002684-n	n	99	NULL	-----
objeto	2	spa-30-00002684-n	n	99	NULL	-----
objeto_fisico	1	spa-30-00002684-n	n	99	NULL	-----
objeto_inanimado	1	spa-30-00002684-n	n	99	NULL	-----
apenas	1	spa-30-00003093-r	r	99	NULL	-----
escasamente	1	spa-30-00003093-r	r	99	NULL	-----
aspirar	3	spa-30-00003316-v	v	99	NULL	-----
incipiente	1	spa-30-00003553-a	a	99	NULL	-----
naciente	1	spa-30-00003553-a	a	99	NULL	-----
conjunto	10	spa-30-00003553-n	n	99	NULL	-----
unidad	11	spa-30-00003553-n	n	99	NULL	-----
unidad_completa	1	spa-30-00003553-n	n	99	NULL	-----
suspirar	1	spa-30-00004032-v	v	99	NULL	-----
organismo	2	spa-30-00004258-n	n	99	NULL	-----
ser	2	spa-30-00004258-n	n	99	NULL	-----
ser_vivo	2	spa-30-00004258-n	n	99	NULL	-----
último	7	spa-30-00004296-a	a	99	NULL	-----
organismo	1	spa-30-00004475-n	n	99	NULL	-----

Figura 3. Fragmento MCR

5.2.1 EuroWordNet

EuroWordNet [16] es un proyecto financiado por la C.E. para el Desarrollo de una Base de Datos Multilingüe Wordnet con relaciones semánticas entre palabras. Es un proyecto de investigación y desarrollo para construir un recurso semántico multilingüe. Su objetivo es el desarrollo de redes semánticas y la realización de la investigación necesaria para el desarrollo; el carácter multilingüe de las redes semánticas permite ir de una lengua a otra. Este proyecto no persigue servir para una sola aplicación, sino para todo tipo de aplicaciones dentro del PLN, Inteligencia Artificial, etc.

Las características específicas de la base de datos EuroWordNet son:

- Cada wordnet refleja las relaciones semánticas en un sistema interno a la lengua, manteniendo las diferencias culturales y lingüísticas.
- Contiene relaciones multilingües entre cada wordnet y los significados del inglés.
- Utilizando las relaciones multilingües, es posible ir de una lengua a otra, lo cual hace posible comparar los distintos wordnets para descubrir inconsistencias y diferencias entre las distintas lenguas.
- Cada wordnet está relacionado con una ontología común independiente del lenguaje y con etiquetas de los campos de conocimiento en los que se usa ese significado. Estos pueden ser utilizados por el usuario para adaptar la base de datos multilingüe sin tener que acceder a cada uno de los wordnets de cada lengua.

5.3 SentiWordNet

SentiWordNet (SWN) [5] es uno de los lexicones más utilizados en el Análisis de Opinión y está construido sobre la base de datos léxica WordNet.

WordNet es una base de datos léxica del idioma inglés que agrupa tanto nombres, adjetivos, verbos y otras formas gramaticales en inglés en un conjunto de sinónimos a los que llamamos synsets, proporcionando definiciones generales y cortas y almacenando las relaciones semánticas entre los conjuntos de sinónimos.

Asigna a cada synset en WordNet tres propiedades (negativo, positivo y objetivo), e indica la intensidad de cada uno de estos tres atributos para cada synset. Al tratarse de valores de polaridad, la suma de la tripleta debe ser 1, ya que estos valores están comprendidos entre 0 y 1 y son obtenidos usando un método semi-supervisado cuya suma debe dar 1.

SWN cubre el total de los synsets de WordNet, aproximadamente 117.000. En SWN cada entrada contiene la categoría morfológica (nombre, adjetivo, adverbio, verbo) y un índice, que identifican al synset en WordNet, junto a las tres propiedades que indican la probabilidad de que el synset sea negativo, positivo o neutro. Este recurso se distribuye libremente para su uso no comercial.



Figura 4. SentiWordNet

```
a 00004171 0 0 moribund#2 being on the point of death; breathing your last; "a moribund patient"
a 00004296 0 0 last#5 occurring at the time of death; "his last words"; "the last rites"
a 00004413 0 0 abridged#1 (used of texts) shortened by condensing or rewriting; "an abridged version"
a 00004615 0 0 shortened#4 cut#3 with parts removed; "the drastically cut film"
a 00004723 0 0 half-length#2 abridged to half its original length
a 00004817 0 0 potted#3 (British informal) summarized or abridged; "a potted version of a novel"
a 00004980 0 0 unabridged#1 (used of texts) not shortened; "an unabridged novel"
a 00005107 0.5 0 uncut#7 full-length#2 complete; "the full-length play"
a 00005205 0.5 0 absolute#1 perfect or complete or pure; "absolute loyalty"; "absolute silence"; "absolute truth"; "absolute alcohol"
a 00005473 0.75 0 direct#10 lacking compromising or mitigating elements; exact; "the direct opposite"
a 00005599 0.5 0.5 unquestioning#2 implicit#2 being without doubt or reserve; "implicit trust"
a 00005718 0.125 0 infinite#4 total and all-embracing; "God's infinite wisdom"
```

Figura 5. Fragmento SWN

5.4 CRiSOL

CRiSOL [3] es un nuevo recurso que combina la información de opinión de SWN y de iSOL. Para ello se intenta añadir a iSOL las puntuaciones de polaridad de los conceptos de SWN. iSOL es un recurso formado por palabras, o mejor dicho, por formas, ya que, tanto lemas como algunas de sus derivaciones constituyen iSOL. Por otro lado, SWN es un recurso conformado por conceptos en inglés, de manera que se hace obligatorio el uso de un recurso auxiliar para enlazar las formas de iSOL y SWN, el cual será el ya descrito MCR.

El proceder habitual en el uso de una base de conocimiento léxica basada en la estructura de WordNet, como es el caso de MCR y de SWN, se corresponde con el uso del índice entre lenguas (ILI) para recuperar la información asociada al concepto. En este caso no se cuenta con ILIs, sino con formas lingüísticas de una lista de palabras de opinión. MCR asocia a cada lema un ILI, lo cual identifica inequívocamente uno de los posibles conceptos del lema. Tomando ese ILI, ya si es posible acudir a SWN y obtener las puntuaciones de polaridad asociadas a dicho concepto.

Por tanto, el proceso de generación de CRiSOL comenzó con la obtención de los lemas de las palabras de iSOL. Una vez obtenidos los lemas, el siguiente paso fue encontrar el ILI asociado al lema en MCR. Como es sabido, un lema puede tener asociados varios identificadores, dado que es común que un lema represente a varios conceptos. Para la primera versión de CRiSOL, se siguió como heurística el tomar como ILI el primero de los asociados al lema. El último paso fue el de recuperar de SWN los valores de polaridad asociados al ILI.

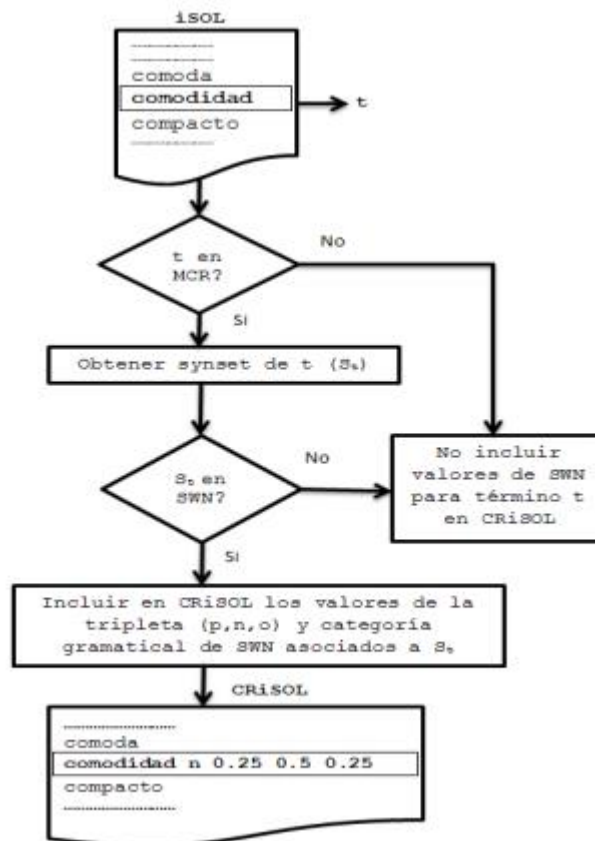


Figura 6. Generación de CRiSOL

```

<is:term xmlns:is="http://sinai.ujaen.es/isol_polar">
  <is:id>ID-23</is:id>
  <is:text>aclamada</is:text>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar" source="SINAI">
    <is:label>positive</is:label>
  </is:polarity>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar" source="SentiWordNet" pos="v">
    <is:positive>0.375</is:positive>
    <is:objective>0.625</is:objective>
    <is:negative>0.</is:negative>
  </is:polarity>
</is:term>

```

Figura 7. Fragmento CRiSOL

6. Herramientas para la detección de la polaridad

6.1 Weka

El Data Mining o la minería de datos, permite realizar patrones de comportamiento y búsquedas de información en grandes bases de datos.

Weka [17] (Waikato Environment for Knowledge Analysis) es un software gratuito de aprendizaje automático y minería de datos. Este tiene un punto a su favor con respecto a otros programas similares, está desarrollado bajo licencia GNU (General Public License). Esta es la licencia de derecho de autor más ampliamente usada en el mundo del software libre y código abierto, y garantiza a los usuarios finales la libertad de utilizar, estudiar, compartir y modificar el software.

Weka es una plataforma desarrollada en Java por la Universidad de Waikato, en Nueva Zelanda. Contiene una gran colección de algoritmos para tareas de minería de datos. Estos pueden aplicarse directamente a un conjunto de datos a través de su interfaz gráfica o bien pueden ser llamados directamente desde su propio código Java.

Incluye herramientas para el pre-procesamiento de datos, clasificación, reglas de asociación y diferentes formas de visualización de los datos, tanto en el inicio del proceso de carga de datos, como después de haber aplicado un algoritmo. También está perfectamente preparado para el desarrollo de nuevos esquemas de aprendizaje sobre máquinas.

6.2 NLTK

NLTK² (Natural Language Toolkit), es un conjunto de programas y bibliotecas para el PLN. Es un software libre, lo que permite que tanto a estudiantes como a personal académico realizar estudios con dicha herramienta sin la necesidad de realizar ninguna inversión económica. Esta herramienta es de código abierto y está disponible en inglés, holandés y francés. NLTK incluye datos de muestra y demostraciones gráficas.

² <http://www.nltk.org/>

NLTK está destinado a apoyar la enseñanza en PLN y la investigación en áreas muy relacionadas, que incluyen las ciencias cognitivas, recuperación de información, lingüística empírica, aprendizaje de la máquina y la inteligencia artificial. NLTK se ha usado exitosamente como una herramienta de estudio individual, como plataforma para los sistemas de investigación de constitución y prototipos y como una herramienta de enseñanza.

Para poder hacer uso de NLTK tenemos que acceder a su web <http://www.nltk.org/> y descargar e instalar Python 2.6 que es la versión recomendada para NLTK.

En la siguiente figura se muestra un ejemplo del funcionamiento de esta herramienta.

The image shows a web interface for sentiment analysis. On the left, under the heading "Analyze Sentiment", there is a "Language" dropdown menu set to "english", a text input field containing the word "Beautiful", and a yellow box indicating a character limit of 50,000. An "Analyze" button is at the bottom left. On the right, under the heading "Sentiment Analysis Results", a green box states "The text is pos.". Below this, a note explains that the final sentiment is determined by classification probabilities. Two sections follow: "Subjectivity" with "neutral: 0.4" and "polar: 0.6", and "Polarity" with "pos: 0.7" (highlighted in green) and "neg: 0.3" (highlighted in red).

Category	Value
Subjectivity - neutral	0.4
Subjectivity - polar	0.6
Polarity - pos	0.7
Polarity - neg	0.3

Figura 8. <http://text-processing.com/demo/sentiment/>

Cómo podemos observar, la palabra Beautiful tiene una polaridad positiva (0.7) y negativa (0.3).

6.3 TextPro

TextPro³ (Text Processing Tools) es un conjunto de herramientas modulares de PLN para el análisis de textos escritos. Ha sido diseñada para integrar y reutilizar el estado de los componentes de la programación Neurolingüística (PNL) desarrollados por los investigadores de la FBK. La versión actual ofrece funciones que van desde la tokenización (seguridad de datos) hasta el análisis y reconocimiento de entidades.

Los diferentes módulos incluidos en TextPro han sido evaluados en el contexto de varias compañías de evaluación como EVALITA (Evaluation of NLP and Speech Tools for Italian) y Semeval 2010 (Evaluation Exercises on Semantic Evaluation).

La arquitectura de TextPro está organizada como una tubería de procesadores donde cada etapa acepta datos de una entrada inicial (o desde la salida de una etapa anterior), ejecuta una tarea específica, y da salida a los datos resultantes (o los envía a la siguiente etapa).

Actualmente distribuye una versión para Linux y Mac, tanto para fines de investigación como para fines comerciales. El código para una versión básica de servicios web también está disponible para su distribución.

En la actualidad TextPro está disponible en español, italiano y alemán. Este software procesa texto según su análisis morfológico, expresiones temporales, etc.

³ <http://hlt-services2.fbk.eu/textpro-demo/textpro.php>

En las siguientes figuras podemos ver el funcionamiento del software.

TextPro online demo

Choose language: English ▼

Please enter your text here:

This house is very big and beautiful

morphological analysis ✕

Morphological Analysis:

1	this+adj+zero	this+pron	this+adv	house+n+sing	be+v+indic+pres+sing3	very+adv	big+adj+zero	and+conj	beautiful+adj+zero
	This			house	is	very	big	and	beautiful

Figura 9. Análisis Morfológico

Choose language: English ▼

Please enter your text here:

This house is very big and beautiful

part-of-speech ✕

Part-of-Speech:

	DT0	NN1	VBZ	AV0	AJ0	CJC	AJ0
1	This	house	is	very	big	and	beautiful

Figura 10. Partes de la oración

7. Materiales y herramientas de desarrollo

7.1 Materiales

Para la realización del trabajo fin de grado se ha recurrido al siguiente material:

- Equipamiento informático.
- Entorno de desarrollo libre.
- Herramientas ofimáticas.
- Referencias bibliográficas.
- Materiales de referencia de fabricantes.
- Acceso a Internet.

7.2 Herramientas de desarrollo

Para el desarrollo de nuestro Trabajo Fin de Grado emplearemos el lenguaje de programación JAVA así como el entorno de desarrollo integrado libre Netbeans 8.0.1.

Las bases de datos con las que estamos trabajando (CRiSOL, MCR y SWN), están implementadas en XML (Extensible Markup Language).

A continuación describiremos todas estas herramientas de forma más detallada:

7.2.1 Lenguaje de programación JAVA

JAVA [18] es un lenguaje de programación concurrente, de propósito general, orientado a objetos que fue diseñado específicamente para tener tan pocas dependencias de implementación como fuera posible.

Su intención es permitir que los desarrolladores de aplicaciones escriban el programa una vez y lo ejecuten en cualquier dispositivo. Esto quiere decir que el código que es ejecutado en una plataforma no tiene que ser recompilado para correr en otra.

Java es, a partir de 2012, uno de los lenguajes de programación más populares en uso, particularmente para aplicaciones de cliente-servidor de web, con millones de usuarios reportados.

Fue desarrollado originalmente por James Gosling de Sun Microsystems (la cual fue adquirida por la compañía Oracle) y publicado en 1995 como un componente fundamental de la plataforma Java de Sun Microsystems con la idea original de usarlo para la creación de páginas web.

Su sintaxis tiene muchas similitudes con el lenguaje C y C++, pero tiene menos utilidades de bajo nivel que cualquiera de ellos. Las aplicaciones de Java son generalmente compiladas a bytecode (clase Java) que puede ejecutarse en cualquier máquina virtual Java (JVM) sin importar la arquitectura de la computadora subyacente.

La compañía Sun desarrolló la implementación de referencia original para las máquinas virtuales, compiladores de Java, y librerías de clases en el año 1991 y las publicó por primera vez en 1995. A partir de mayo del año 2007, en cumplimiento con las especificaciones del Proceso de la Comunidad Java, Sun volvió a licenciar la mayoría de sus tecnologías de Java bajo la Licencia Pública General de GNU.

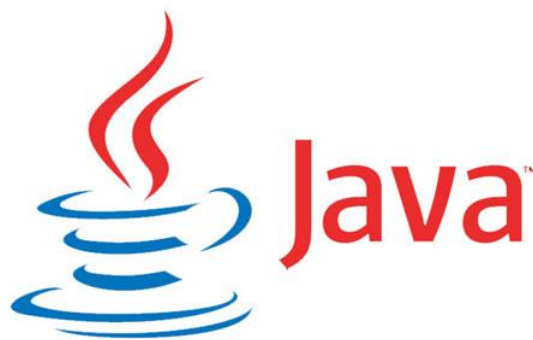


Figura 11. Java

El lenguaje Java se creó con cinco principales objetivos:

- Debería diseñarse para ejecutar código en sistemas remotos de forma segura.
- Debería incluir por defecto soporte para trabajo en red.
- Debería ser fácil de usar y tomar lo mejor de otros lenguajes orientados a objetos, como C++.
- Debería usar el paradigma de la programación orientada a objetos.
- Debería permitir la ejecución de un mismo programa en múltiples sistemas operativos.

En un sentido estricto, Java no es un lenguaje absolutamente orientado a objetos, a diferencia de, por ejemplo, Ruby. Por motivos de eficiencia, Java ha relajado en cierta medida el paradigma de orientación a objetos y así no todos los valores son objetos.

El código Java puede ser a veces redundante en comparación con otros lenguajes. Esto es en parte debido a las frecuentes declaraciones de tipos y conversiones de tipo manual (casting). También se debe a que no se dispone de operadores sobrecargados, y a una sintaxis relativamente simple. Sin embargo, J2SE 5.0 (Java 2 Platform Standard Edition) introduce elementos para tratar de reducir la redundancia, como una nueva construcción para los bucles “foreach”.

A diferencia de C++, Java no dispone de operadores de sobrecarga definidos por el usuario. Los diseñadores de Java tomaron esta decisión puesto que consideraban que, bajo ciertas circunstancias, esta característica podía complicar la lectura y el mantenimiento de los programas.

7.2.2 Netbeans

NetBeans [19] es un entorno de desarrollo integrado libre, donde los programadores pueden escribir, compilar, depurar y ejecutar programas, hecho principalmente para el lenguaje de programación Java, aunque puede servir para cualquier otro lenguaje de programación. Existe además un número importante de módulos para extenderlo. NetBeans IDE es un producto gratuito y libre sin restricciones de uso. Es un proyecto de código abierto de gran éxito con un gran número de usuarios, una comunidad en constante crecimiento, y con cerca de cien socios en todo el mundo. Sun Microsystems fundó el proyecto de código abierto NetBeans en junio del año 2000 y continúa siendo el patrocinador principal de los proyectos.

La plataforma NetBeans permite que las aplicaciones sean desarrolladas a partir de un conjunto de componentes de software llamados módulos. Un módulo es un archivo Java que contiene clases de java escritas para interactuar con las APIs (Application Programming Interface) de NetBeans y un archivo especial que lo identifica como módulo. Las aplicaciones construidas a partir de módulos pueden ser extendidas agregándole nuevos módulos. Debido a que los módulos pueden ser desarrollados independientemente, las aplicaciones basadas en la plataforma NetBeans pueden ser extendidas fácilmente por otros desarrolladores de software.

El NetBeans IDE soporta el desarrollo de todos los tipos de aplicación Java (web, J2SE, aplicaciones móviles y EJB (Enterprise JavaBeans)).

NetBeans IDE 6.5.2, la cual fue publicada el 19 de noviembre de 2008, extiende las características existentes del Java EE. Adicionalmente, el NetBeans Enterprise Pack soporta el desarrollo de Aplicaciones empresariales con Java EE 5, incluyendo herramientas de desarrollo visuales de SOA, herramientas de esquemas XML, orientación a web services, y modelado UML (Unified Modeling Language). El NetBeans C/C++ Pack soporta proyectos de C/C++, mientras el PHP Pack, soporta PHP 5.

Todas las funciones del IDE son provistas por módulos. Cada módulo provee una función bien definida, tales como el soporte de Java, edición, o soporte para el sistema de control de versiones. NetBeans contiene todos los módulos necesarios para el desarrollo de aplicaciones Java en una sola descarga, permitiéndole al usuario comenzar a trabajar inmediatamente.

Tanto Sun Studio, Sun Java Studio Enterprise y Sun Java Studio Creator de Sun Microsystems han sido todos basados en el IDE NetBeans. Desde Julio de 2006, NetBeans IDE es licenciado bajo la Common Development and Distribution License (CDDL), una licencia basada en la Mozilla Public License (MPL). En Octubre de 2007, Sun anunció que NetBeans desde entonces se ofrecerá bajo licenciamiento dual de Licencia CDDL y la GPL versión 2.

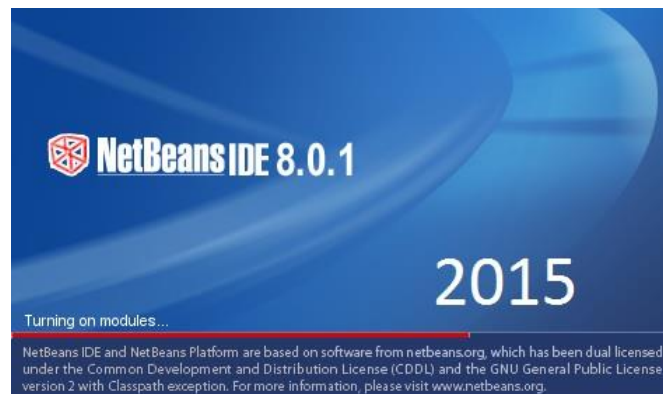


Figura 12. NetBeans IDE 8.0.1

7.2.3 XML

Lenguaje extendido de Marcas [20], del inglés “extensible markup language” es un lenguaje desarrollado por el W3C en febrero del año 1998 aunque está basado en una estandarización anterior la ISO 8879, la última recomendación de la W3C (World Wide Web Consortium) es de 26 Noviembre de 2008. XML es usado ampliamente para almacenar e intercambiar datos en forma legible.

El XML es una adaptación del SGML (Standard Generalized Markup Language), un lenguaje que permite el etiquetado y la organización de los documentos. Esto quiere decir que XML no es un lenguaje en sí mismo, sino un sistema que permite definir lenguajes de acuerdo a las necesidades. El MathML (Mathematical Markup Language), el SVG (Scalable Vector Graphics) y el XHTML (eXtensible HyperText Markup Language), son algunos de los lenguajes que el XML tiene la capacidad de definir.

Los documentos de texto, las bases de datos, las páginas web y las hojas de cálculo son algunos de los campos de aplicación del XML. El metalenguaje aparece como un estándar que estructura el intercambio de información entre las diferentes plataformas.

Los expertos han señalado algunas de las ventajas que derivan de la utilización del XML, como por ejemplo:

- Facilita el análisis y el procesamiento de los documentos XML creados por terceros.
- Es extensible (se pueden añadir nuevas etiquetas tras el diseño del documento)
- Su analizador es estándar (no requiere de cambios para cada versión del metalenguaje)

Entre los lenguajes creados con XML, destacan el XLINK (que intenta trascender las limitaciones de los enlaces de hipertexto en HTML (HyperText Markup Language)) y el XSL (Extensible Stylesheet Language) .

La validez de los documentos (es decir, que su estructura sintáctica se encuentre correctamente desarrollada) depende la relación especificada entre los distintos elementos a partir de un documento externo o una definición.

Según se analice, XML presenta una serie de ventajas para los desarrolladores, especialmente porque permite relacionar aplicaciones de diferentes plataformas y lenguajes. La universalidad que persigue XML puede no llegar nunca si en lugar de aprovecharlo para la resolución de problemas, se generan nuevos sabiendo que tendrán una solución.

Dado que, en gran parte, la utilidad de una herramienta depende de la creatividad de quien la utiliza, resulta imposible resumir todas las aplicaciones de XML. En pocas palabras, se puede decir que ofrece la posibilidad de estructurar y representar datos. En la actualidad, es común que los programas incluyan archivos de configuración en este formato; tal es el caso de las aplicaciones creadas con la tecnología .NET (de Microsoft) y de Apache.

Cuando se desarrolla un programa con interfaz gráfica es necesario organizar todas las imágenes de manera que se vayan cargando a medida que se necesiten, y XML es de gran ayuda en estos casos: permite etiquetarlas agruparlas, relacionarlas con otros datos y especificar su ubicación, según las necesidades de los diseñadores.

Pero además de facilitar la configuración de un programa y la organización de los recursos, XML cumple un papel muy importante que es su punto fuerte: le permite comunicarse con otras aplicaciones, de diferentes plataformas y sin que importe el origen de la información en común. Se pueden tener, por ejemplo, un programa corriendo en Linux con Oracle, y otro en Windows con una base de datos de SQL Server, ambos compartiendo datos gracias a una estructura en XML.

Los servicios web, son componentes de la Red que brindan la posibilidad de realizar una serie variada de operaciones, a través de métodos concretos que aprovechan el metalenguaje XML para sus comunicaciones, gracias a lo cual cualquier plataforma puede hacer uso de sus ventajas.

Para finalizar, XML es una de esas herramientas que a pesar de su poca complejidad esconden un gran potencial, gracias a su innegable utilidad y ser fácil de utilizar.

Un ejemplo de XML:

```
<?xml version="1.0" encoding="UTF-8"?>
<is:isol xmlns:is="http://sinai.ujaen.es/isol_polar">
  <is:term xmlns:is="http://sinai.ujaen.es/isol_polar">
    <is:id>ID-1</is:id>
    <is:text>abiertamente</is:text>
    <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SINAI">
      <is:label>positive</is:label>
    </is:polarity>
    <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SentiWordNet" pos="x">
      <is:positive>0.</is:positive>
      <is:objective>1.</is:objective>
      <is:negative>0.</is:negative>
    </is:polarity>
  </is:term>
</is:isol>
```

8. Desarrollo del Software

A lo largo de este apartado describiremos todos los pasos a seguir que nos han permitido poner en funcionamiento nuestro software.

8.1 Contenido de las bases de datos

En este apartado, explicaremos de forma detallada, el contenido de cada una de las bases de datos que hemos utilizado para poner en funcionamiento nuestro software.

8.1.1 MCR

MCR es un recurso en español que a partir de un identificador de synset es capaz de relacionar una palabra en español con otra en inglés.

Es un archivo de texto plano que contiene un total de 57989 palabras. Cada una de las palabras lleva unos valores asociados. Una misma palabra puede estar repetida varias veces, ya que esta base de datos va asociada a otra llamada SWN, donde cada palabra tiene asignado un significado distinto según su connotación, ya sea positiva o negativa.

Cada línea de MCR consta de:

- Palabra.
- N° de significados distintos.
- Identificador de synsets.
- POS (Part Of Speech): Categoría gramatical (nombre, adjetivo, adverbio...).

PALABRA	Nº SIGNIFICADOS DISTINTOS	IDENTIFICADOR DE SYNSET	POS
Abrazo	2	spa-30-00059728-n	n (nombre)
Abrazo	1	spa-30-00417397-n	n
Abrazo	3	spa-30-00417643-n	n
Abrazo	4	spa-30-00417859-n	n

Tabla 1. MCR Ejemplo nombre

PALABRA	Nº SIGNIFICADOS DISTINTOS	IDENTIFICADOR DE SYNSET	POS
Agitar	7	spa-30-02585050-v	v (verbo)
Agitar	3	spa-30-00014034-v	v
Agitar	12	spa-30-01391946-v	V
Agitar	9	spa-30-01418389-v	v
Agitar	5	spa-30-01419473-v	v
Agitar	2	spa-30-01629958-v	v
Agitar	13	spa-30-01819911-v	v
Agitar	10	spa-30-01864865-v	v
Agitar	4	spa-30-01875295-v	v
Agitar	8	spa-30-01880888-v	v
Agitar	14	spa-30-01885430-v	v
Agitar	1	spa-30-01889610-v	v
Agitar	6	spa-30-01890792-v	v
Agitar	15	spa-30-02093390-v	v
Agitar	11	spa-30-02093610-v	v

Tabla 2. MCR Ejemplo verbo

PALABRA	Nº SIGNIFICADOS DISTINTOS	IDENTIFICADOR DE SYNSET	POS
Apenas	1	spa-30-00003093-r	r (adverbio)
Apenas	2	spa-30-00073763-r	r

Tabla 3. MCR ejemplo adverbio

PALABRA	Nº SIGNIFICADOS DISTINTOS	IDENTIFICADOR DE SYNSET	POS
Abundante	7	spa-30-00013887-a	a (adjetivo)
Abundante	6	spa-30-02558528-a	a
Abundante	4	spa-30-01080900-a	a
Abundante	5	spa-30-00014358-a	a
Abundante	2	spa-30-00014490-a	a
Abundante	1	spa-30-00015247-a	a
Abundante	3	spa-30-00015854-a	a

Tabla 4. MCR ejemplo adjetivo

8.1.2 SentiWordNet

SWN es una base de conocimiento de opinión que está construida a partir de WordNet. Asocia a cada synset tres valores de polaridad: Positivo, Objetivo y Negativo. Estos tres valores son la probabilidad de que el sentido correspondiente a WordNet sea Positivo, Negativo u Objetivo. Al ser valores de probabilidad es evidente que su suma tiene que ser igual a 1 [5].

Cada línea de SWN consta de:

- POS
- ID
- PosScore
- NegScore
- SynsetTerms

POS e ID, identifican de forma única los synsets de WordNet (3.0).

Los valores PosScore y NegScore son la positividad y la negatividad. La objetividad se puede calcular de la siguiente manera:

$$ObjScore = 1 - (PosScore + NegScore)$$

La columna de SynsetTerms informa sobre el significado (en inglés) de la palabra perteneciente al synset.

A continuación pondremos varios ejemplos con diferentes palabras y categorías gramaticales.

En la siguiente tabla se muestra la palabra **abrazo** con diferentes identificadores de synsets y por consiguiente, diferente significado.

POS	ID	POSSCORE	NEGSCORE	SYNSET TERMS
n	00059728	0	0	Clinch
n	00417397	0.125	0	Embracing Embracement Embrace
n	00417643	0.25	0	Snuggle Nestle Cuddle
n	00417859	0.125	0	Squeeze Hug Clinch

Tabla 5. SWN ejemplo nombre

Siguiendo el orden anterior, los diferentes significados para una misma palabra son:

00059728: *(Boxing) The act of one boxer holding onto the other to avoid being hit and to rest momentarily.*

00417397: *The act of clasping another person in the arms (as in greeting or affection).*

00417643: *A close and affectionate (and often prolonged) embrace.*

00417859: *A tight or amorous embrace; "come here and give me a big hug".*

En la siguiente tabla mostramos la palabra abundante:

POS	ID	POSSCORE	NEGSCORE	SYNSET TERMS
a	00013887	0	0.25	Plentiful Bountiful
a	02558528	0.25	0	Substantial Square Solid Satisfying Hearty
a	01080900	0.25	0	Abundant
a	00014358	0	0.25	Galore Abounding
a	00014490	0.125	0	Rich Plentiful Plenteous Copious Ample
a	00015247	0	0.25	Riotous Profuse Luxuriant Lush Exuberant
a	00015854	0	0.25	Plentiful

Tabla 6. STW ejemplo adjetivo

Siguiendo el orden anterior, los diferentes significados para una misma palabra son:

00013887: *Present in great quantity; "An abundant supply of water".*

02558528: *Providing abundant nourishment; "A hearty meal"; "Good solid food"; "Ate a substantial breakfast"; "Four square meals a day".*

01080900: *Producing in abundance; "The bountiful earth"; "A plentiful year"; "Fruitful soil".*

00014358: *Existing in abundance; "Abounding confidence"; "Whiskey galore".*

00014490: *Affording an abundant supply; "Had ample food for the party"; "Copious provisions"; "Food is plentiful"; "A plenteous grape harvest"; "A rich supply".*

00015247: *Produced or growing in extreme abundance; "Their riotous blooming".*

00015854: *Existing in great number or quantity; "Rhinoceroses were once plentiful here".*

8.1.3 CRiSOL

CRiSOL es un recurso que combina la información de opinión de ISOL y SWN. Es un archivo .xml que contiene un total de 8135 lemas.

El proceso de generación de CRiSOL comenzó con la obtención de los lemas de las palabras de iSOL. El siguiente paso fue encontrar el ILI asociado al lema en MCR. Como es sabido, un lema puede tener asociados varios identificadores, dado que es común que un lema represente a varios conceptos. Para la primera versión de CRiSOL, se siguió como heurística el tomar como ILI el primero de los asociados al lema. El último paso fue el de recuperar de SWN los valores de polaridad asociados al ILI [3].

A continuación se muestran algunos ejemplos:

```
<is:term xmlns:is="http://sinai.ujaen.es/isol_polar">
  <is:id>ID-2</is:id>
  <is:text>abrazo</is:text>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SINAI">
    <is:label>positive</is:label>
  </is:polarity>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SentiWordNet" pos="n">
    <is:positive>0.</is:positive>
    <is:objective>1.</is:objective>
    <is:negative>0.</is:negative>
  </is:polarity>
</is:term>
```

```
<is:term xmlns:is="http://sinai.ujaen.es/isol_polar">
  <is:id>ID-13</is:id>
  <is:text>abundante</is:text>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SINAI">
    <is:label>positive</is:label>
  </is:polarity>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SentiWordNet" pos="a">
    <is:positive>0.</is:positive>
    <is:objective>0.75</is:objective>
    <is:negative>0.25</is:negative>
  </is:polarity>
</is:term>
```

Si nos fijamos en la palabra abrazo, vemos que tenemos tres propiedades (positive, objective y negative), cada una de ellas con un valor. Como hemos mencionado anteriormente, se siguió como heurística el tomar como ILI el primero de los asociados al lema, y es por esto que si nos fijamos en SWN, los valores PosScore y NegScore son 0, por lo que en CRiSOL estos valores van a coincidir.

```
<is:positive>0.</is:positive>
<is:objective>1.</is:objective>
<is:negative>0.</is:negative>
```

Si nos fijamos en la palabra abundante en SWN, vemos que los valores PosScore y NegScore del primer ILI asociado al lema son: 0 y 0.25, consecutivamente que coinciden con los obtenidos en CRiSOL.

```
<is:positive>0.</is:positive>  
<is:objective>0.75</is:objective>  
<is:negative>0.25</is:negative>
```

El valor objective se ha calculado de la siguiente manera:

$$\text{Objective} = 1 - (\text{positive} + \text{negative})$$

8.2 Generación de CRiSOL_v2

Para la generación de CRiSOL_v2, se han utilizado las 3 bases de datos mencionadas anteriormente.

En el siguiente esquema se puede ver a groso modo como hemos enlazado estas bases de datos:



Figura 13. Esquema enlace bases de datos.

El proceso de generación de CRiSOL_v2 ha comenzado con la búsqueda de los lemas de las palabras de CRiSOL. MCR asocia a cada lema un ILI, lo cual identifica inequívocamente uno de los posibles conceptos del lema. Una vez obtenidos estos lemas de CRiSOL, el siguiente paso ha sido encontrar el ILI asociado al lema en MCR. Un lema puede tener asociados varios identificadores, debido a que una misma palabra puede representar varios conceptos. Tomando este ILI, acudimos a SWN para así obtener los valores de polaridad asociados a dicho concepto.

Para la primera versión de CRiSOL, se siguió como heurística el tomar como ILI el primero encontrado de los asociados al lema en SWN, y lo que se ha pretendido en este trabajo fin de grado ha sido realizar una mejora de este recurso trabajando con todos los ILIs. De esta forma se han sacado varias clasificaciones que se explicarán con detalle posteriormente.

La siguiente figura representa el proceso de generación de CRiSOL_v2, el cual se trata de una base de conocimiento compuesta por los mismos términos que CRiSOL, que además de contar con la etiqueta de polaridad de iSOL, están complementados por la categoría morfológica y las distintas puntuaciones de polaridad de SWN.

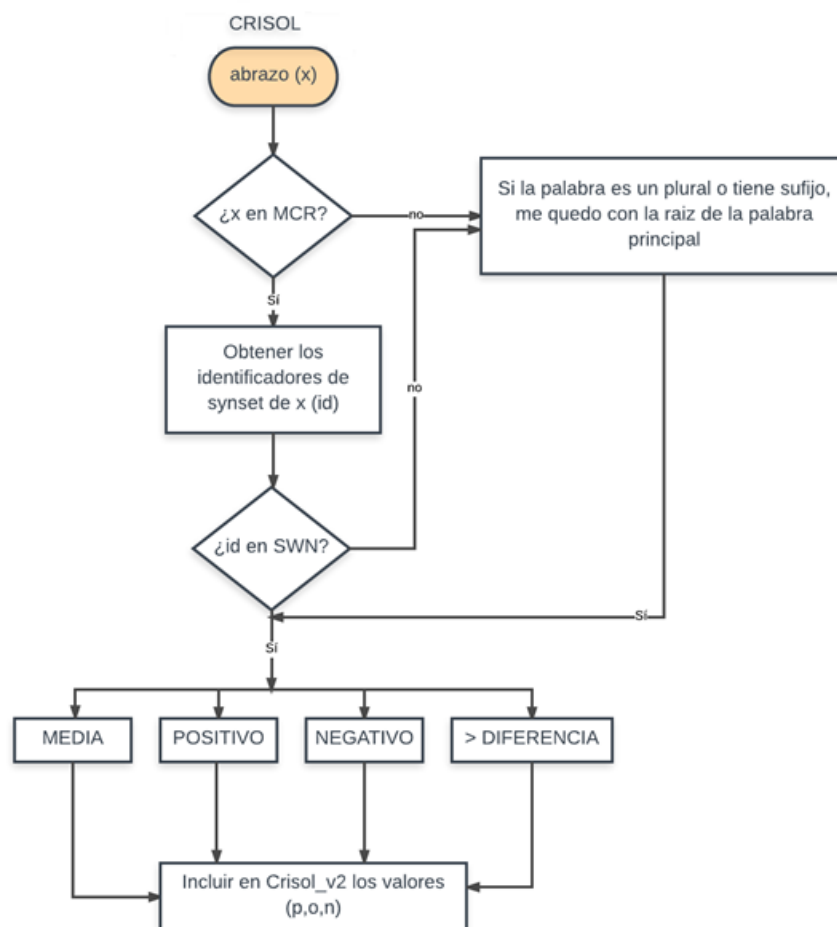


Figura 14. Generación de CRiSOL_v2

Como hemos mencionado anteriormente, al recurso CRiSOL_v2 se le han añadido distintas clasificaciones a partir de todos los valores de polaridad conseguidos de SWN para cada lema. Estas clasificaciones son el cálculo de la media entre palabras homógrafas, es decir, que se escriben igual pero tienen distinto significado, los valores más positivos, quedándonos con la palabra que tenga el valor de positividad más alto, los valores más negativos, quedándonos con la palabra que tenga el valor de negatividad más alto y con las palabras que tengan mayor diferencia entre positivo y negativo de los valores de polaridad.

Para el cálculo de la media se ha utilizado la siguiente fórmula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n a_i = \frac{a_1 + a_2 + a_3 + \dots + a_n}{n}$$

Siendo $(a_1, a_2, a_3, \dots, a_n)$ el conjunto de los valores de polaridad (positiva y negativa) asociados a un lema.

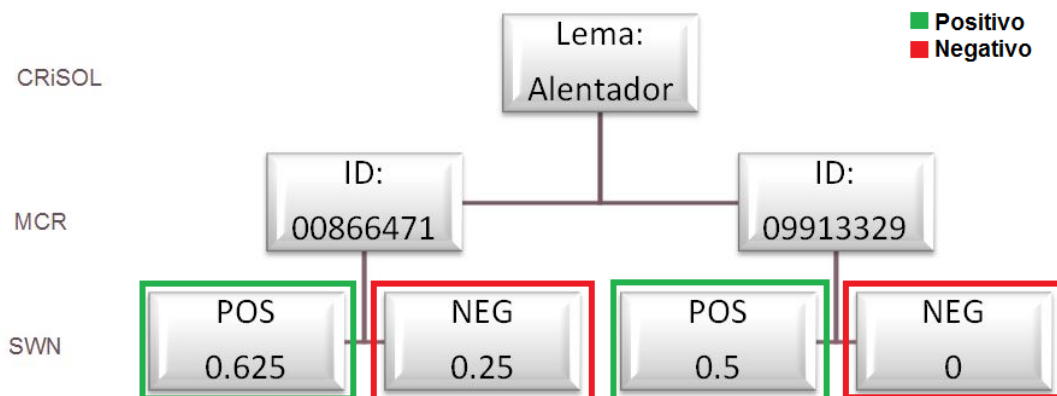


Figura 15. Cálculo de la media en CRiSOL_v2

$$\bar{x}(\text{positive}) = \frac{0.625 + 0.5}{2} = 0.563$$

$$\bar{x}(\text{negative}) = \frac{0.25 + 0}{2} = 0.125$$

$$\text{Objective} = 1 - (\text{positive} + \text{negative}) = 1 - (0.563 + 0.125) = 0.312$$

"SentiWordNet Media"

```
<is:positive>0.563</is:positive>
<is:objective>0.312</is:objective>
<is:negative>0.125</is:negative>
```

El cálculo de los valores más positivos y más negativos se ha realizado de la misma manera. Hemos buscado los diferentes ILIs asociados al lema en MCR y a continuación hemos acudido a SWN para así obtener los valores de polaridad asociados a dicho concepto. Se ha realizado una búsqueda iterativa donde nos quedamos con los valores positivos y negativos del lema, mostrando finalmente el conjunto de valores de polaridad, donde el valor positivo es el más alto.

Esto mismo ocurre con el cálculo de los valores más negativos, pero a la inversa. Aquí nos quedamos con el conjunto de valores de polaridad, donde el valor negativo es el más alto.

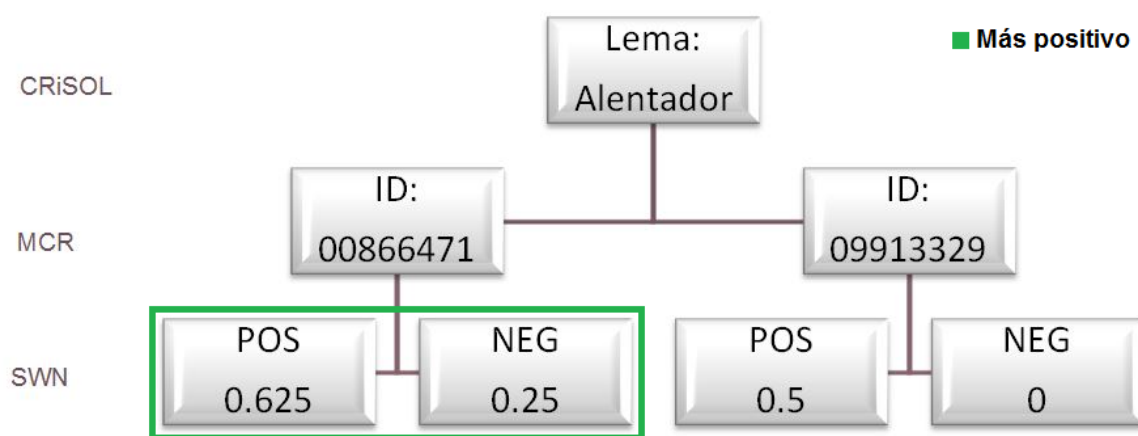


Figura 16. Cálculo valores más positivos CRiSOL_v2

$$Objective = 1 - (positive + negative) = 1 - (0.625 + 0.25) = 0.125$$

"SentiWordNet Positivo"

```

<is:positive>0.625</is:positive>
<is:objective>0.125</is:objective>
<is:negative>0.25</is:negative>
  
```

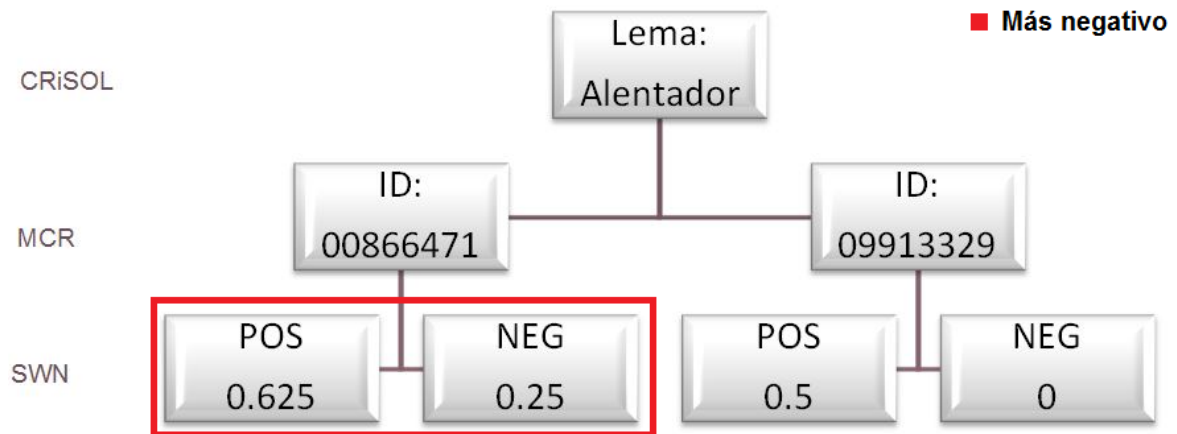


Figura 17. Cálculo valores más negativos CRiSOL_v2

$$Objective = 1 - (positive + negative) = 1 - (0.625 + 0.25) = 0.125$$

"SentiWordNet Negativo"

```

<is:positive>0.625</is:positive>
<is:objective>0.125</is:objective>
<is:negative>0.25</is:negative>

```

Por último, para el cálculo de la mayor diferencia entre Positivo-Negativo lo hacemos de la siguiente manera:

Calculamos el valor absoluto entre $|Positive - Negative|$ y nos quedamos con el valor del resultado con mayor puntuación.

$$Diferencia = \max(|Positive - Negative|)$$

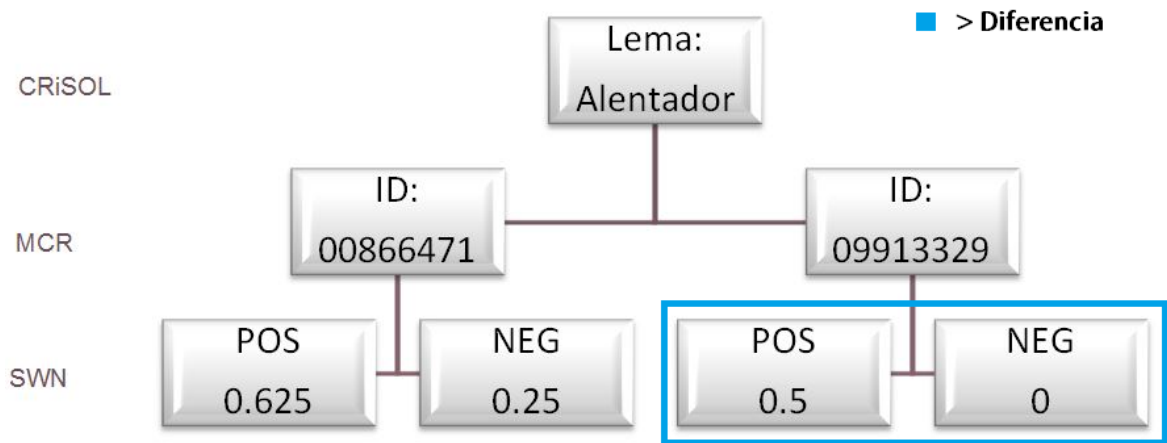


Figura 18. Cálculo de la diferencia CRiSOL_v2

$$Diferencia = |0.625 - 0.25| = 0.375$$

$$Diferencia = |0.5 - 0| = 0.5$$

$$Objetive = 1 - (positive + negative) = 1 - (0.5287 + 0.0283) = 0.443$$

"SentiWordNet Diferencia"

<is:positive>0.5</is:positive>
 <is:objective>0.5</is:objective>
 <is:negative>0.0</is:negative>

A continuación mostraremos un pequeño fragmento del documento .xml generado (CRiSOL_v2), donde pueden verse las distintas clasificaciones obtenidas. Este sigue la misma estructura que CRiSOL, ya que es una versión mejorada del mismo, añadiendo las distintas clasificaciones.

```
<is:term xmlns:is="http://sinai.ujaen.es/isol_polar">
  <is:id>ID-2</is:id>
  <is:text>abrazo</is:text>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SINAI">
    <is:label>positive</is:label>
  </is:polarity>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SentiWordNet" pos="n">
    <is:positive>0.</is:positive>
    <is:objective>1.</is:objective>
    <is:negative>0.</is:negative>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SentiWordNet Media" pos="n">
    <is:positive>0.125</is:positive>
    <is:objective>0.875</is:objective>
    <is:negative>0.0</is:negative>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SentiWordNet Positivo" pos="n">
    <is:positive>0.25</is:positive>
    <is:objective>0.75</is:objective>
    <is:negative>0.0</is:negative>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SentiWordNet Negativo" pos="n">
    <is:positive>0.125</is:positive>
    <is:objective>0.875</is:objective>
    <is:negative>0.0</is:negative>
  <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar"
source="SentiWordNet Diferencia" pos="n">
    <is:positive>0.25</is:positive>
    <is:objective>0.75</is:objective>
    <is:negative>0.0</is:negative>
  </is:polarity>
</is:term>
```

Asimismo, se ha creado una interfaz de usuario, donde además de obtener los valores de polaridad de un solo lema, pueden obtenerse dichos valores en textos completos.

Figura 19. Interfaz de Usuario

En ella se muestran tanto los valores de subjetividad (Positivo y Negativo) como el valor de objetividad (neutro), que se obtienen a raíz de calcular la media, los valores más positivos, valores más negativos y la diferencia entre Positivo-Negativo.

Además se han añadido un buscador de synsets, que hace un conteo de las palabras contenidas en SWN y muestra el número total de ellas y un buscador de palabras, que hace un conteo de las palabras con sentimiento contenidas en CRiSOL_v2.

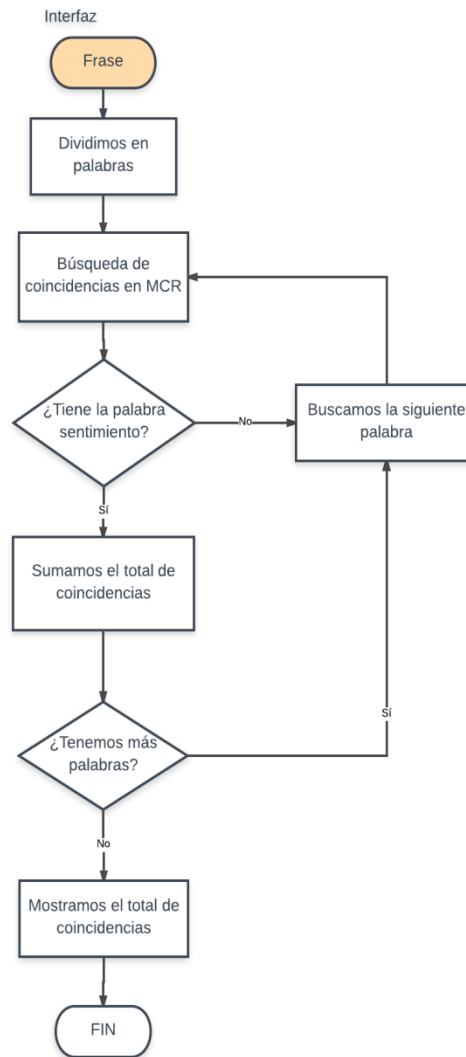


Figura 20. Diagrama de flujo del conteo de synsets

Para mostrar cuales son estas palabras se ha añadido un cuadro de texto llamado Palabras con Sentimiento, que muestra una a una cuales son las diferentes palabras que se están analizando del texto que hemos introducido.

Si queremos generar el .xml de CRiSOL_v2 en su totalidad, pulsaremos el botón Generar Fichero.

A la hora de analizar textos completos, la heurística que se ha seguido ha sido la misma que con un solo lema o palabra. Se ha analizado palabra por palabra (siguiendo los mismos criterios que se han explicado anteriormente) y se ha hecho una media total, dependiendo del número de palabras con sentimiento que se encuentran en el texto.

8.2.1 Pruebas de funcionamiento del software

En este apartado, se muestran distintas pruebas de funcionamiento del software diseñado.

El siguiente texto se ha extraído de <https://www.tripadvisor.es>, sitio web que proporciona reseñas de contenido relacionado con viajes y que incluye foros de viajeros.

PRUEBA 1

“Es uno de los hoteles más estupendos donde me he alojado en mi vida. Todo está planteado con un gusto y una elegancia exquisita. La atención del personal es genial. La propietaria es un encanto. Está correctamente ubicado.”

CLASIFICACIÓN DE POLARIDAD DE SENTIMIENTOS

Introduzca el texto:

Es uno de los hoteles más estupendos donde me he alojado en mi vida. Todo está planteado con un gusto y una elegancia exquisita. La atención del personal es genial. La propietaria es un encanto. Está correctamente ubicado.

Clasificación:

Media (highlighted with a red box)

Más Positivo

Más Negativo

Diferencia

Subjetividad:

POSITIVO 0.5287

NEGATIVO 0.0283

Objetividad:

NEUTRO 0.443

Palabras con Sentimiento:

estupendos
gusto
elegancia
exquisita
genial
encanto
correctamente

Buscador de Synsets: 52

Buscador de Palabras: 7

Generar Fichero

UJa Universidad de Jaén

Laura M^a Miñarro Escribano

Figura 21. Interfaz Media Opinión 1

Como podemos observar, son siete las palabras con sentimiento que se evalúan en el texto. Estas son:

- Estupendos
- Gusto
- Elegancia
- Exquisita
- Genial
- Encanto
- Correctamente

En la siguiente tabla se muestra la evaluación de cada una de estas palabras, obteniendo la media de los valores (positivos, negativos y neutros).

PALABRA	POSITIVIDAD	NEGATIVIDAD	NEUTRALIDAD
Estupendos	0.542	0.042	0.416
Gusto	0.359	0.031	0.61
Elegancia	0.875	0	0.125
Exquisita	0.5	0.125	0.375
Genial	0.85	0	0.15
Encanto	0.45	0	0.55
Correctamente	0.125	0	0.875

Tabla 7. Evaluación Opinión 1

Para el cálculo de la media anterior, se ha buscado en SWN cada una de estas palabras y se han visto los distintos significados para cada una de ellas.

A continuación, mostraremos diversas tablas donde podremos ver en cada una de ellas una palabra con distintos significados y su valoración, ya que según en el contexto en el que se utilice, esta puede tener distintos significados y por consiguiente distinta puntuación.

Palabra **ESTUPENDOS**

	POSITIVO	NEGATIVO
Supreme	0.75	0
Tremendous	0.75	0
Brave/Braw	0.125	0.125

Tabla 8. Estupendos

$$\bar{x}(pos) = \frac{0.75 + 0.75 + 0.125}{3} = 0.542$$

$$\bar{x}(neg) = \frac{0 + 0 + 0.125}{3} = 0.042$$

Palabra **GUSTO**

	POSITIVO	NEGATIVO
Taste	0	0
Sense of taste	0.375	0
Tastiness	0.75	0
Fondness	0	0
Liking	0.5	0.25
Gusto	0.75	0
Flavor	0	0
Appreciation	0.5	0

Tabla 9. Gusto

$$\bar{x}(pos) = \frac{0 + 0.375 + 0.75 + 0 + 0.5 + 0.75 + 0 + 0.5}{8} = 0.359$$

$$\bar{x}(neg) = \frac{0 + 0 + 0 + 0 + 0.25 + 0 + 0 + 0}{8} = 0.031$$

Palabra **ELEGANCIA**

	POSITIVO	NEGATIVO
Elegance	0.875	0
Style	0.375	0
Gentility	0.625	0
Stylishness	0.625	0

Tabla 10. Elegancia

$$\bar{x}(pos) = \frac{0.875 + 0.375 + 0.625 + 0.625}{4} = 0.875 \quad \bar{x}(neg) = \frac{0 + 0 + 0 + 0}{4} = 0$$

Palabra **EXQUISITA**

	POSITIVO	NEGATIVO
Exquisite	0.375	0.125
Refiner	0.5	0
Dainty	0.375	0.25
Of extreme beauty	0.5	0.125
Yummy	0.75	0.25
Tasty	0.625	0.25

Tabla 11. Exquisita

$$\bar{x}(pos) = \frac{0.375 + 0.5 + 0.375 + 0.5 + 0.75 + 0.625}{6} = 0.5$$

$$\bar{x}(neg) = \frac{0.125 + 0 + 0.25 + 0.125 + 0.25 + 0.25}{6} = 0.125$$

Palabra **GENIAL**

	POSITIVO	NEGATIVO
Tremendous	0.75	0
Not bad	0.875	0
Divine	0.875	0
Mean	1	0
Virtuoso	0.75	0

Tabla 12. Genial

$$\bar{x}(pos) = \frac{0.75 + 0.875 + 0.875 + 1 + 0.75}{5} = 0.85 \quad \bar{x}(neg) = \frac{0 + 0 + 0 + 0 + 0}{5} = 0$$

Palabra **ENCANTO**

	POSITIVO	NEGATIVO
Charm	0.5	0
Loveliness	0.875	0
Enchantment	0	0
Elan	0.375	0
Lure	0.5	0

Tabla 13. Encanto

$$\bar{x}(pos) = \frac{0.5 + 0.875 + 0 + 0.375 + 0.5}{5} = 0.45 \quad \bar{x}(neg) = \frac{0 + 0 + 0 + 0 + 0 + 0}{5} = 0$$

Palabra **CORRECTAMENTE**

	POSITIVO	NEGATIVO
Justly	0.125	0

Tabla 14. Correctamente

$$\bar{x}(pos) = \frac{0.125}{1} = 0.125 \quad \bar{x}(neg) = \frac{0}{1} = 0$$

En la siguiente figura, podemos ver la media total de las palabras que tienen sentimiento del texto anterior:



Si hacemos el cálculo, podemos ver que el resultado es correcto:

$$\bar{x}(pos_total) = \frac{0.542 + 0.359 + 0.875 + 0.5 + 0.85 + 0.45 + 0.125}{7} = 0.5287$$

$$\bar{x}(neg_total) = \frac{0.042 + 0.031 + 0 + 0.125 + 0 + 0 + 0}{7} = 0.0283$$

$$Objetive = 1 - (positive + negative) = 1 - (0.5287 + 0.0283) = 0.443$$

PRUEBA 2

“Me ha parecido un hotel vulgar, donde los empleados eran un poco violentos a la hora de hablar. Me ha dado mucho coraje, por lo que no volveré a este hotel.”

En la frase anterior, podemos encontrar tres palabras con sentimiento. Estas son: vulgar, violento y coraje.

A continuación se muestran las distintas valoraciones de cada una de estas palabras, según su significado:

PALABRA	POSITIVIDAD	NEGATIVIDAD
Vulgar	0	0.625
Vulgar	0	0.75
Vulgar	0.25	0.375
Vulgar	0	0
Vulgar	0	0
Vulgar	0.25	0.25
Vulgar	0.222	0.778
Vulgar	0	0.5

Tabla 15. Valoración palabra vulgar

PALABRA	POSITIVIDAD	NEGATIVIDAD
Violento	0	0.625
Violento	0.375	0.5
Violento	0	0
Violento	0.375	0.375
Violento	0	0.25
Violento	0.25	0
Violento	0	0
Violento	0.25	0.125
Violento	0.25	0.25

Tabla 16. Valoración palabra violento

PALABRA	POSITIVIDAD	NEGATIVIDAD
Coraje	0	0
Coraje	0.375	0.5
Coraje	0.25	0.25
Coraje	0.375	0.375
Coraje	0.5	0.5
Coraje	0.25	0.25
Coraje	0	0
Coraje	0.125	0.75
Coraje	0.25	0

Tabla 17. Valoración palabra Coraje

En las siguientes tablas, se muestran la evaluación de cada una de las palabras con sentimiento, obteniendo las distintas clasificaciones (media, más positivo, más negativo y diferencia).

MEDIA

LEMA	POSITIVIDAD	NEGATIVIDAD	NEUTRALIDAD
Vulgar	0.09	0.41	0.5
Violentos	0.167	0.236	0.597
Coraje	0.236	0.319	0.445

Tabla 18. Evaluación Media Opinión 2

$$\bar{x}(\text{pos}_{_total}) = \frac{0.09 + 0.167 + 0.236}{3} = 0.1643$$

$$\bar{x}(\text{neg}_{_total}) = \frac{0.41 + 0.236 + 0.319}{3} = 0.3217$$

$$\text{Objective} = 1 - (\text{positive} + \text{negative}) = 1 - (0.1643 + 0.3217) = 0.514$$

MÁS POSITIVO

LEMA	POSITIVIDAD	NEGATIVIDAD	NEUTRALIDAD
Vulgar	0.25	0.25	0.5
Violentos	0.375	0.375	0.25
Coraje	0.5	0.5	0.0

Tabla 19. Evaluación Más Positivo Opinión 2

$$\bar{x}(pos_total) = \frac{0.25 + 0.375 + 0.5}{3} = 0.375$$

$$\bar{x}(neg_total) = \frac{0.25 + 0.375 + 0.5}{3} = 0.375$$

$$Objetive = 1 - (0.375 + 0.375) = 0.25$$

Clasificación:

Media

Más Positivo

Más Negativo

Diferencia

Subjetividad:

POSITIVO 0.375

NEGATIVO 0.375

Objetividad:

NEUTRO 0.25

Palabras con Sentimiento:

vulgar
violentos
coraje

Figura 22. Prueba 2 Más Positivo

En el caso en el que haya dos palabras con la misma positividad, nos quedaremos con la que tenga menor negatividad.

MAS NEGATIVO

LEMA	POSITIVIDAD	NEGATIVIDAD	NEUTRALIDAD
Vulgar	0.222	0.778	0.0
Violentos	0.0	0.625	0.375
Coraje	0.125	0.75	0.125

Tabla 20. Evaluación Más Negativo Opinión 2

$$\bar{x}(pos_total) = \frac{0.222 + 0 + 0.125}{3} = 0.1157$$

$$\bar{x}(neg_total) = \frac{0.778 + 0.625 + 0.75}{3} = 0.7177$$

$$Objetive = 1 - (0.1157 + 0.7177) = 0.1667$$

Clasificación:

Media

Más Positivo

Más Negativo

Diferencia

Subjetividad:

POSITIVO 0.1157

NEGATIVO 0.7177

Objetividad:

NEUTRO 0.1667

Palabras con Sentimiento:

vulgar
violentos
coraje

Figura 23. Prueba 2 Más Negativo

En el caso en el que haya dos palabras con la misma negatividad, nos quedaremos con la que tenga menor positividad.

MAYOR DIFERENCIA

LEMA	POSITIVIDAD	NEGATIVIDAD	NEUTRALIDAD
Vulgar	0.0	0.75	0.25
Violentos	0.0	0.625	0.375
Coraje	0.125	0.75	0.125

Tabla 21. Evaluación Diferencia Opinión 2

$$\bar{x}(pos_total) = \frac{0 + 0 + 0.125}{3} = 0.0417$$

$$\bar{x}(neg_total) = \frac{0.75 + 0.625 + 0.75}{3} = 0.7083$$

$$Objetive = 1 - (0.0417 + 0.7083) = 0.25$$

Clasificación:

Media

Más Positivo

Más Negativo

Diferencia

Subjetividad:

POSITIVO 0.0417

NEGATIVO 0.7083

Objetividad:

NEUTRO 0.25

Palabras con Sentimiento:

vulgar
violentos
coraje

Figura 24. Prueba 2 Mayor Diferencia

8.3 Problemas encontrados en el desarrollo

Durante el desarrollo de esta aplicación se nos han presentado algunas dificultades que pasaremos a detallar a continuación.

Como hemos mencionado anteriormente, CRiSOL_v2 es una modificación de CRiSOL, y este es un recurso que está compuesto por palabras en español. Por otro lado tenemos SWN es un recurso conformado por conceptos en inglés, de manera que se hace obligatorio el uso del recurso MCR.

En CRiSOL, podemos ver adjetivos a los que se les ha añadido el sufijo –mente para formar un adverbio (ardientemente, amablemente, ampliamente, etc.). Estas palabras no las contempla MCR, por lo que en este caso la solución que hemos tomado ha sido la de quedarnos con la raíz de la palabra (ardiente, amable, etc.). Lo mismo ocurre con los femeninos y plurales (terminados en –s y –es).

Hemos sido conscientes de que estas soluciones han podido afectar a un número mínimo de palabras, como por ejemplo en el caso del adjetivo “demente”, pero la mejora que hemos obtenido de este recurso es muy superior al número de palabras que se han visto perjudicadas.

9. Líneas de futuro

Debido a que nos hemos centrado en la clasificación de la polaridad de comentarios en Internet, mejorando para ello un recurso ya existente, sólo hemos tenido en cuenta las palabras contenidas en dicho recurso. Es por ello que no se ha analizado una frase completa, sino que se han analizado las palabras con sentimiento contenidas en dicha frase.

Sería interesante tener en cuenta el análisis del fenómeno lingüístico de la negación ya que este se utiliza para modificar el valor de verdad de la unidad lingüística a la que se aplica. El tratamiento de este fenómeno implica identificar la presencia de la negación en el texto y, posteriormente, determinar los términos a los que se aplica dentro de este. En los últimos años, el tratamiento de este fenómeno ha despertado un especial interés en la comunidad científica e incluso investigadores de prestigio lo definen como un desafío en PLN debido a su dificultad.

Otro de los aspectos a tener en cuenta sería el uso de intensificadores (muy, nunca, apenas, etc.), ya que estos pueden modificar el valor de la palabra a la que acompañan.

10. Anexos

10.1 Manual de instalación del entorno de trabajo Netbeans 8.0.1

Primero procedemos a descargar NetBeans IDE, para ello nos dirigimos a la siguiente dirección:

https://netbeans.org/downloads/8.0.1/?pagelang=pt_BR

NetBeans IDE 8.2 Download

8.1 | 8.2 | Development | Archive

Email address (optional):

Subscribe to newsletters: ☒ Monthly ☐ Weekly ☒ NetBeans can contact me at this address

IDE Language: Español Platform: Windows

Note: Greyed out technologies are not supported for this platform.

NetBeans IDE Download Bundles in community contributed languages¹

Supported technologies *	Java SE	Java EE	HTML5/JavaScript	PHP	C/C++	All
NetBeans Platform SDK	•	•				•
Java SE	•	•				•
Java FX	•	•				•
Java EE		•				•
Java ME						•
HTML5/JavaScript		•	•	•		•
PHP			•	•		•
C/C++					•	•
Groovy						•
Java Card™ 3 Connected						•
Bundled servers						
GlassFish Server Open Source Edition 4.1.1		•				•
Apache Tomcat 8.0.27		•				•

Download Download Download x86 Download x86 Download x86 Download x86 Download x64 Download x64 Download x64 Download

Free, 100 MB Free, 200 MB Free, 111 - 115 MB Free, 111 - 115 MB Free, 109 - 112 MB Free, 222 MB

Figura 25. Versiones Netbeans

Seleccionamos el idioma que queremos que tenga nuestro entorno de desarrollo. En este caso seleccionamos “Español”.

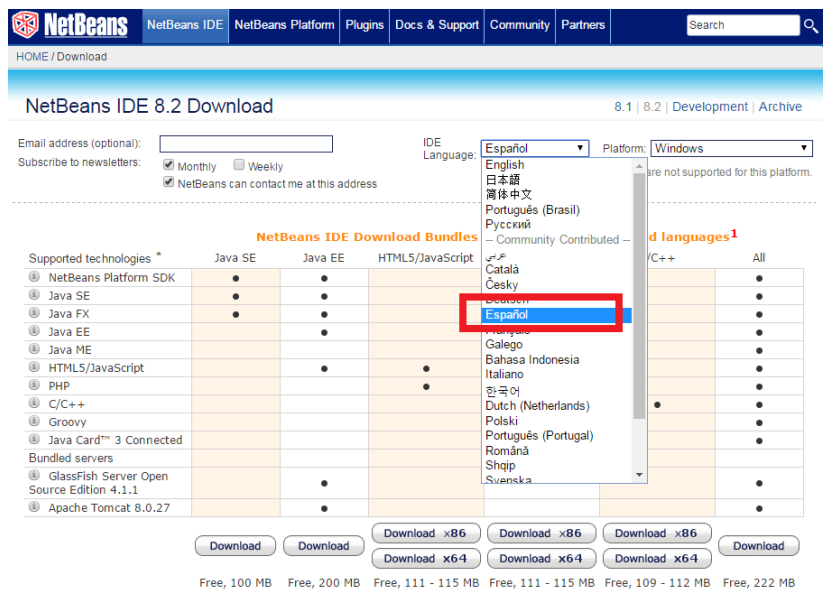


Figura 26. Proceso de descarga

A continuación escogemos el sistema operativo con el que cuenta la máquina donde instalaremos el NetBeans IDE. En este caso escogemos “Windows”.

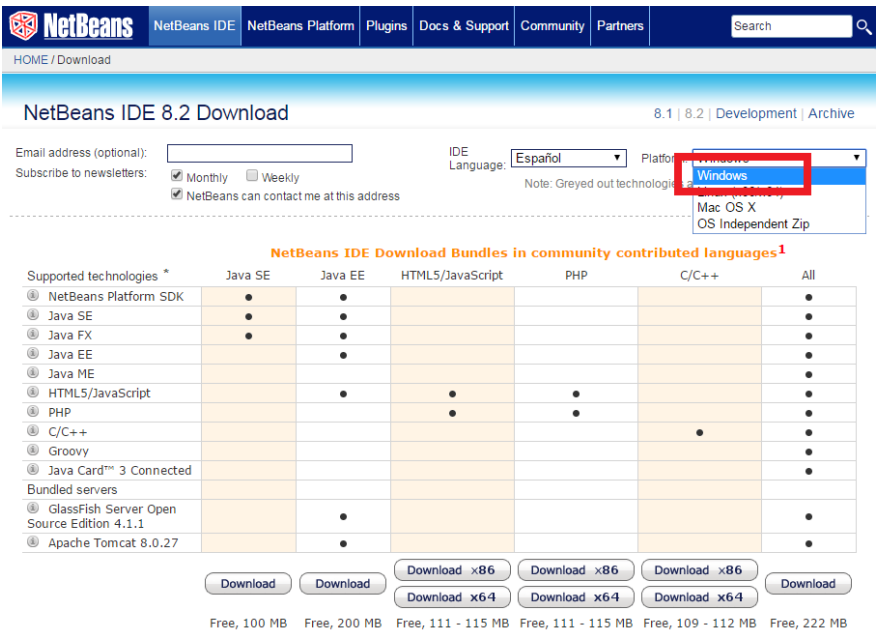


Figura 27. Selección sistema operativo

Por último, seleccionamos el paquete que vamos a descargar. Esto va a depender sobre qué lenguajes vamos a trabajar, ya que algunos paquetes no contienen ciertas tecnologías. Para evitar cualquier tipo de problema, procedemos a instalar el paquete “All”.

NetBeans IDE 8.2 Download

8.1 | 8.2 | Development | Archive

Email address (optional):

Subscribe to newsletters: ☒ Monthly ☐ Weekly

☒ NetBeans can contact me at this address

IDE Language: Español Platform: Windows

Note: Greyed out technologies are not supported for this platform.

NetBeans IDE Download Bundles in community contributed languages

Supported technologies *	Java SE	Java EE	HTML5/JavaScript	PHP	C/C++	All
NetBeans Platform SDK	•	•				•
Java SE	•	•				•
Java FX	•	•				•
Java EE		•				•
Java ME						•
HTML5/JavaScript		•	•	•		•
PHP			•	•		•
C/C++					•	•
Groovy						•
Java Card™ 3 Connected						•
Bundled servers						•
GlassFish Server Open Source Edition 4.1.1		•				•
Apache Tomcat 8.0.27		•				•

Download Download Download x86 Download x86 Download x86 Download x64 Download x64 Download x64 Download

Free, 100 MB Free, 200 MB Free, 111 - 115 MB Free, 111 - 115 MB Free, 109 - 112 MB Free, 222 MB

Figura 28. Descarga paquete seleccionado

A continuación procedemos a descargar el Java Development Kit (JDK), ya que sin este no se podrá instalar NetBeans:

<http://www.oracle.com/technetwork/es/java/javase/downloads/index.html>

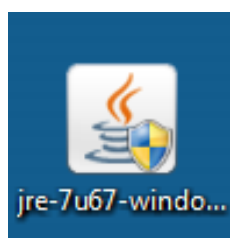


Figura 29. JDK

Se abrirá la ventana de instalación, y daremos a siguiente:



Figura 30. Instalación Netbeans

En la siguiente ventana tenemos la posibilidad de cambiar la ruta de instalación y elegir las características a instalar, pero lo vamos a dejar como está por defecto y daremos clic en siguiente:

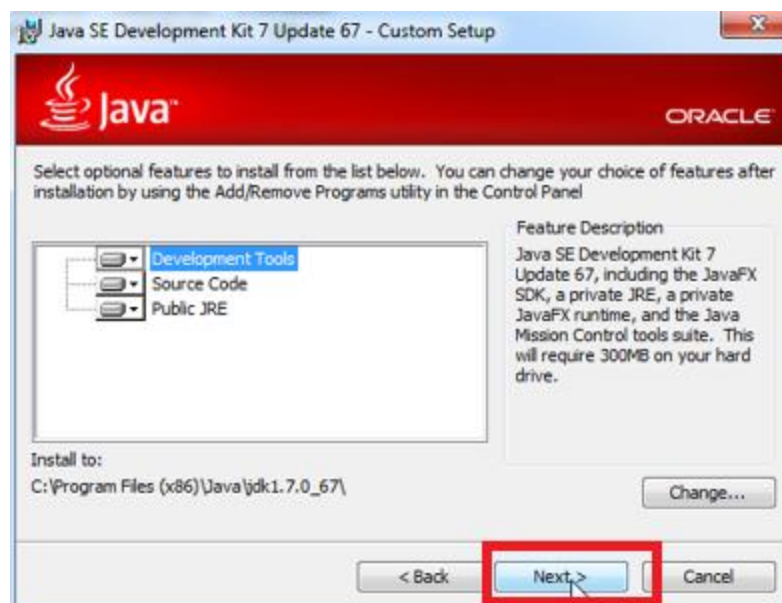


Figura 31. Diferentes paquetes para la instalación

Esperamos mientras se instala el JDK, y se procederá a instalar el Entorno de Ejecución de Java ((JRE), y le damos a siguiente:

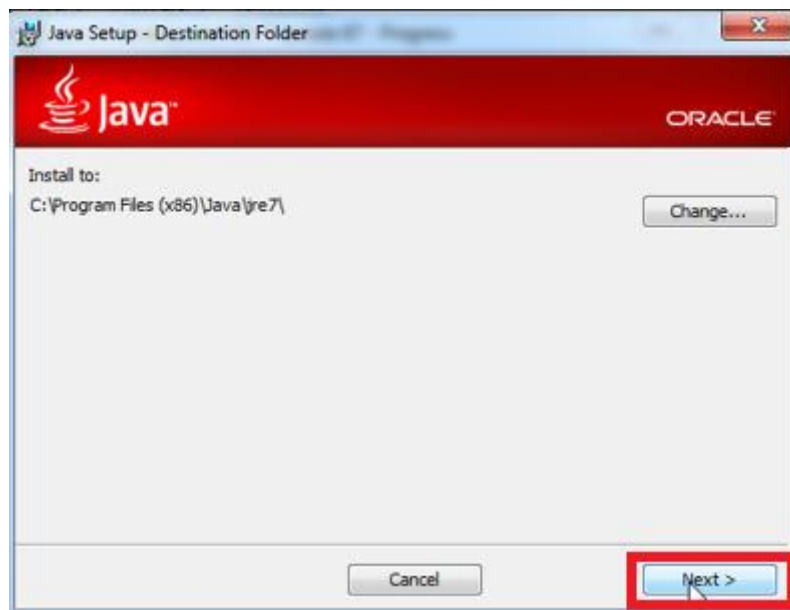


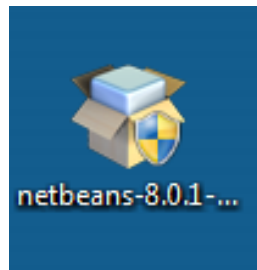
Figura 32. Directorio de la instalación

Una vez instalado, le damos a “cerrar”.



Figura 33. Verificación de la instalación

A continuación hacemos click sobre el ejecutable de NetBeans IDE para proceder con su instalación:



Se nos abrirá una ventana con los componentes y características que serán instaladas. Le damos a siguiente:

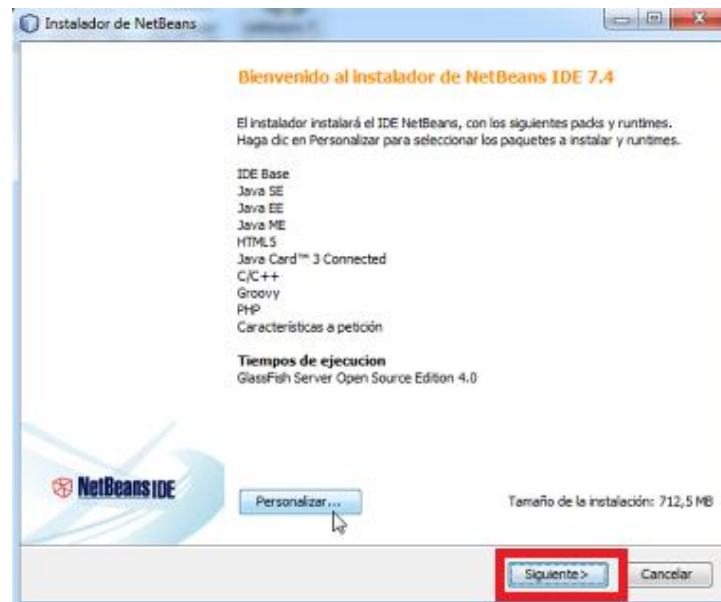


Figura 34. Componentes y características a instalar

Accedemos a otra ventana en la que debemos desactivar cualquier ventana con la característica que no queramos instalar. En el caso que desactivemos algún componente que sea utilizado por otro, aparecerá un mensaje de error en la parte inferior y hasta que no se corrija no se podrá continuar con la instalación.

Tras haber escogido los paquetes y las herramientas que queremos instalar, le damos a “Aceptar” y después a “Siguiente”. El siguiente paso será aceptar el acuerdo de la licencia y continuar con la instalación.

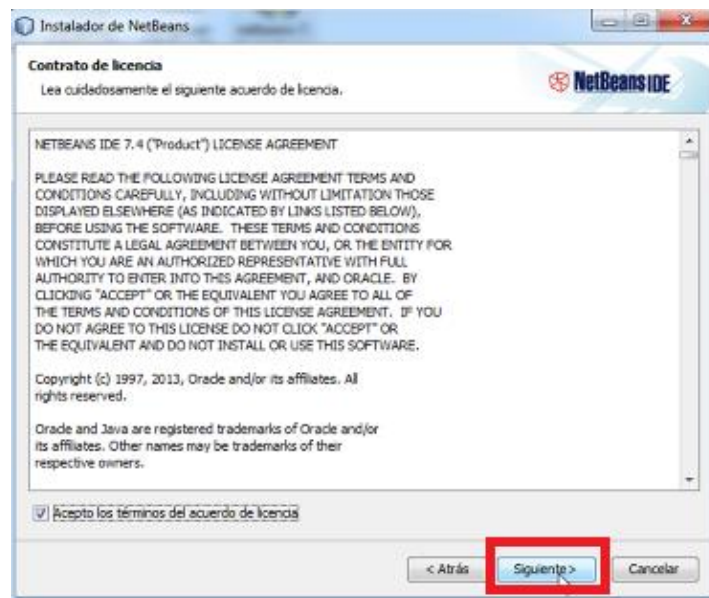


Figura 35. Acuerdo de licencia

En esta ventana nos pregunta si queremos instalar JUnit. Esto es una herramienta para hacer pruebas a nuestras aplicaciones JAVA; pero en esta ocasión no lo necesitamos por lo que le damos a No instalar y hacemos clic en “Siguiente”:

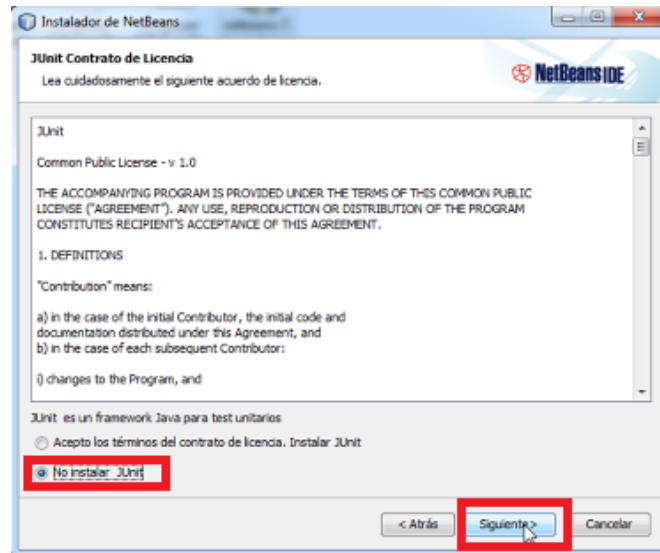


Figura 36. JUnit

En la siguiente ventana nos mostrará la ruta de instalación del NetBeans IDE, que podemos modificar dando a “Examinar” y podemos escoger la carpeta que deseemos. En este caso se dejará la ruta por defecto.

También podremos ver el JDK que usará NetBeans (en nuestro caso el que hemos instalado en los pasos anteriormente citados). Le damos a “Siguiente”:

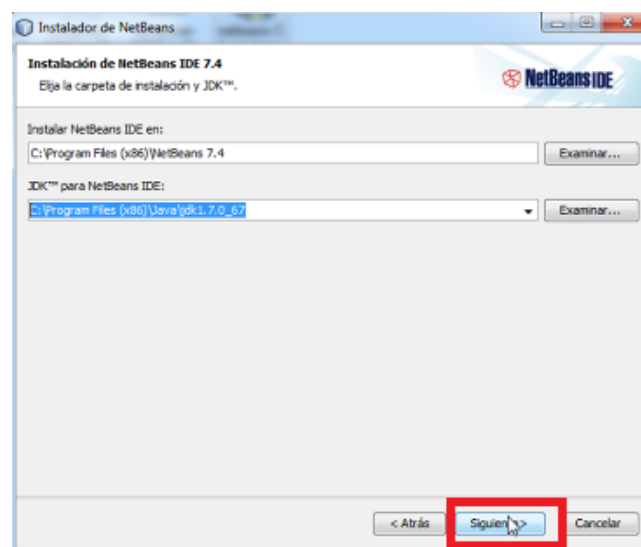


Figura 37. Ruta instalación Netbeans

En esta ventana se nos presentará el mismo caso del punto anterior, sólo que aquí será con el servidor de aplicaciones "GlassFish", damos clic en el botón "Siguiente":

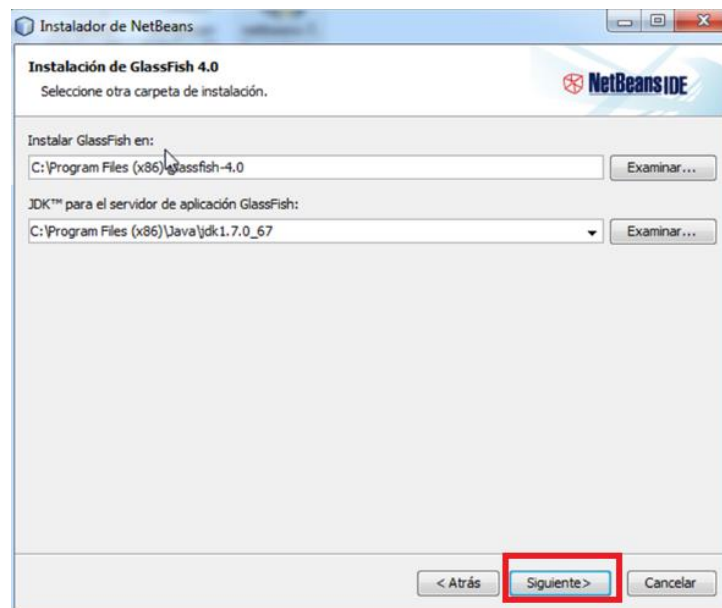


Figura 38. Servidor GlassFish

En este paso se muestra la ruta de instalación para "Apache Tomcat" y dejamos la ruta que aparece por defecto. Le damos a "Siguiente":

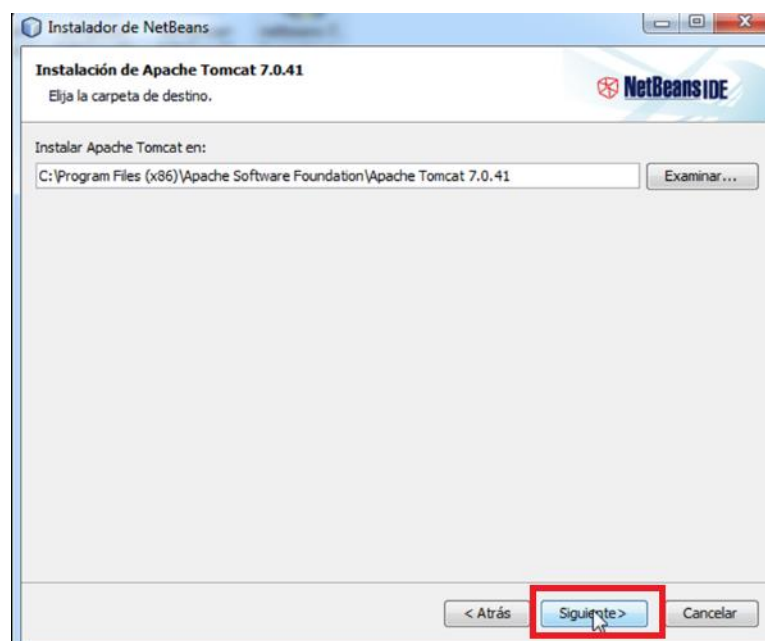


Figura 39. Apache Tomcat

Hacemos click sobre “Instalar” para que se lleve a cabo todo el proceso de instalación de NetBeans IDE en nuestro PC. En el caso de que no queramos que el instalador busque actualizaciones mientras se realice el montaje, desactivaremos la casilla “CheckforUpdates”. En caso contrario la dejaremos activada.

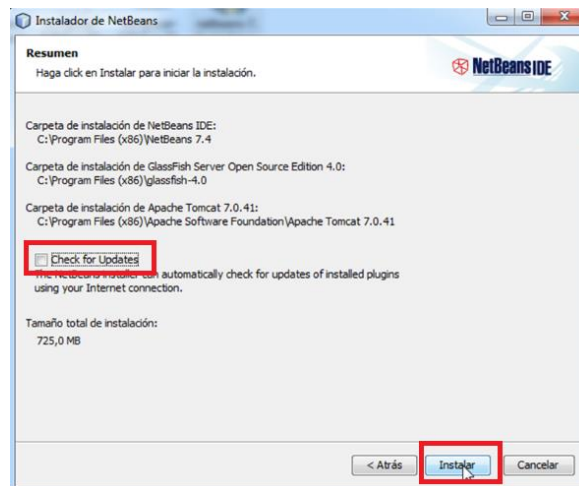


Figura 40. Instalación NetBeans IDE

Por último, nos aparece una ventana en la cual podremos activar/desactivar la casilla “Apoye al proyecto NetBeans proporcionando datos de uso de manera anónima”. En nuestro caso la dejaremos desactivada, puesto que con esta opción se enviarán a la base de datos de estadísticas en el servidor de NetBeans los datos de las herramientas que más utilicemos y no es necesario mantenerla activada para finalizar la instalación. Le damos click en “Terminar”:

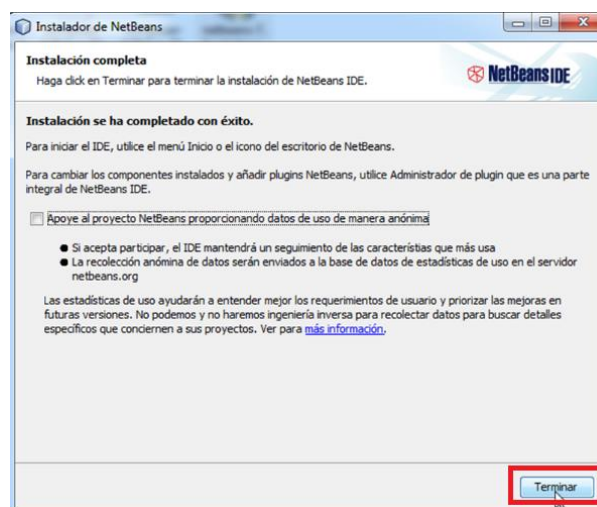
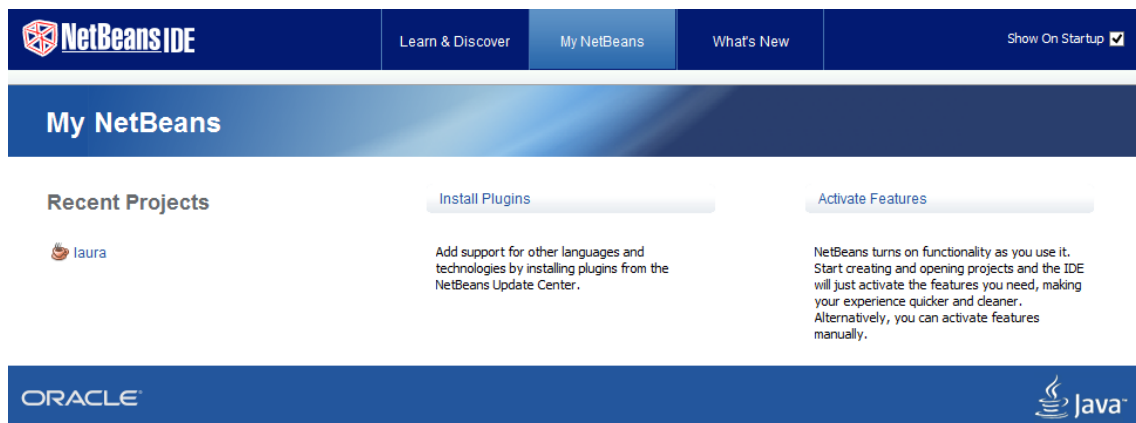


Figura 41. Fin Instalación

Ya hemos terminado la instalación y veremos en nuestro escritorio el acceso directo de NetBeans IDE.



10.2 Manual de funcionamiento de la interfaz de usuario

Presentamos la guía para la utilización de la interfaz de usuario del software para la extracción de información en bases de conocimiento de opinión.

En la interfaz de usuario mostramos las diferentes utilidades del software, diferenciando bien cada una de sus opciones.

Se accede a la interfaz de usuario a través de un ejecutable.

CLASIFICACIÓN DE POLARIDAD DE SENTIMIENTOS

Introduzca el texto:

Clasificación:

- Media
- Más Positivo
- Más Negativo
- Diferencia

Subjetividad:

- POSITIVO
- NEGATIVO
- Objetividad: NEUTRO

Palabras con Sentimiento:

Buscador de Synsets

Buscador de Palabras

Generar Fichero

UJa Universidad de Jaén

Laura M.ª Miñarro Escribano

Figura 42. Interfaz Usuario

Una vez en esta pantalla, se podrá introducir el texto a analizar.

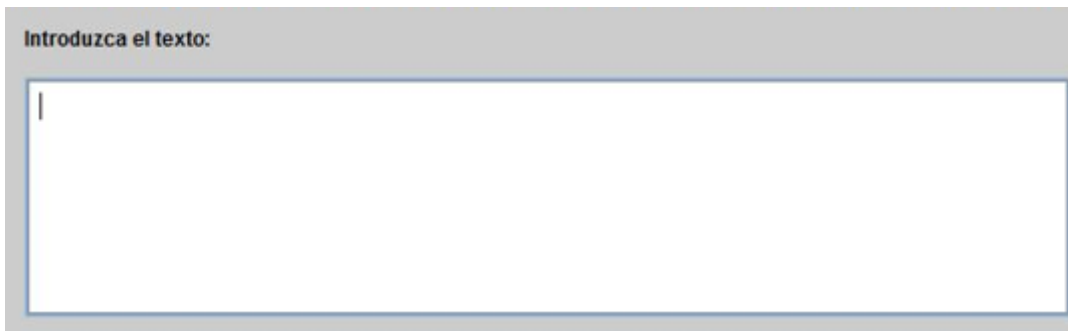
Un formulario con un encabezado que dice "Introduzca el texto:". Debajo de este encabezado hay un cuadro de texto rectangular grande y vacío, con un cursor de texto visible al principio.

Figura 43. Cuadro de texto

A continuación se explicarán las distintas funcionalidades de los botones:

El botón Buscador de Synsets, hace un conteo de las palabras contenidas en SWN. Hay que tener en cuenta que aunque el texto sólo contenga una sola palabra, en SWN puede aparecer varias veces esa misma palabra, ya que según en el contexto en el que se haya utilizado puede tener diferente puntuación.

Un botón rectangular con el texto "Buscador de Synsets" en el centro. A la derecha del botón hay un cuadro de texto rectangular vacío.

Figura 44. Buscador de Synsets

El botón Buscador de Palabras, cuenta las palabras con sentimiento que contiene CRiSOL_v2. Si hay una palabra que se repite varias veces en el texto introducido, se contará el número de veces que dicha palabra aparezca.

Un botón rectangular con el texto "Buscador de Palabras" en el centro. A la derecha del botón hay un cuadro de texto rectangular vacío.

Figura 45. Buscador de palabras

Al pulsar este botón, en “Palabras con Sentimiento”, nos aparecerá una lista con todas las palabras que tengan sentimiento en el texto introducido.



Figura 46. Listado palabras con sentimiento

Los botones MEDIA, MÁS POSITIVO, MÁS NEGATIVO Y DIFERENCIA funcionan de la misma manera.

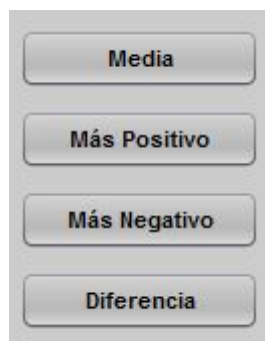


Figura 47. Funciones

El botón media nos calculará la media de las palabras con sentimiento contenidas en SWN.



Figura 48. Media

Si pulsamos en el botón Más Positivo, nos aparecerá la puntuación de la palabra que tenga la valoración más positiva. Si hay varias palabras en el texto, nos calculará la media de las puntuaciones.

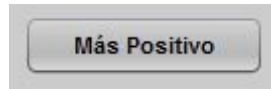


Figura 49. Más positivo

Si pulsamos en el botón Más Negativo, nos aparecerá la puntuación de la palabra que tenga la valoración más negativa. Si hay varias palabras en el texto, nos calculará la media de las puntuaciones.

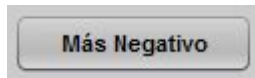


Figura 50. Más negativo

Por último, en el botón Diferencia, se calcula la diferencia entre Positivo-Negativo de cada palabra. Aquí nos quedamos con la palabra que tenga una mayor diferencia. Al igual que en los botones anteriores, si en el texto hay varias palabras, nos calculará la media de las puntuaciones.

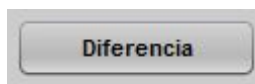


Figura 51. Diferencia

Al pulsar cualquiera de los botones anteriores, nos aparecerán los valores solicitados en el cuadro de resultados, diferenciando los valores de sentimiento (Positivos, Negativos y Neutros) referentes a cada palabra.



Subjetividad:

POSITIVO

NEGATIVO

Objetividad:

NEUTRO

Figura 52. Cuadro de resultados

Por último tenemos el botón Generar Fichero, que nos generará un documento .xml llamado CRiSOL_v2, donde aparecerá calculado todo lo mencionado anteriormente y que hemos extraído de CRiSOL, MCR y SWN. Sólo tendremos que generar este fichero cuando haya modificaciones en CRiSOL.

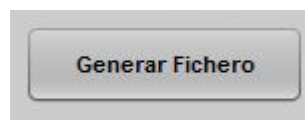


Figura 53. Generación del xml

10.2.1 Ejemplo Práctico

CLASIFICACIÓN DE POLARIDAD DE SENTIMIENTOS

Introduzca el texto:

Personal muy amable y gentil

Clasificación: Media Más Positivo Más Negativo Diferencia	Subjetividad: POSITIVO0.269 NEGATIVO0.0 Objetividad: NEUTRO0.731	Palabras con Sentimiento: amable gentil
---	--	---

Buscador de Synsets Buscador de Palabras	22 2	Generar Fichero
--	--	---

 **UJa** Universidad de Jaén

Laura M^a Miñarro Escribano

Figura 54. Interfaz ejemplo

Con el siguiente texto, podemos ver que en la opción Buscador de Synsets, nos aparece un 22. Esto es porque nos hace una suma total de las veces que cada una de las palabras aparece en SWN:

- Personal: 8
- Muy: 2
- Amable: 10
- Y: 1
- Gentil: 1

En Buscador de Palabras nos aparecen un total de dos palabras con sentimiento, que como podemos ver en la siguiente figura, son **amable** y **gentil**.



Figura 55. Palabras con sentimiento ejemplo

Si pulsamos los botones Media, Más Positivo, Más Negativo y Diferencia nos aparecerán los diferentes valores calculados en cada uno de ellos. Estos valores varían dependiendo de lo que queramos mostrar.

A screenshot of a web application titled "CLASIFICACIÓN DE POLARIDAD DE SENTIMIENTOS". The interface includes a text input field with the text "Personal muy amable y gentil". Below this, there is a section with a red border containing classification buttons and calculated values. The buttons are "Media", "Más Positivo", "Más Negativo", and "Diferencia". The calculated values are: "POSITIVO" (0.269), "NEGATIVO" (0.0), and "NEUTRO" (0.731). To the right of this section is another "Palabras con Sentimiento:" window showing "amable" and "gentil". At the bottom, there are two search fields: "Buscador de Synsets" with value 22 and "Buscador de Palabras" with value 2. A "Generar Fichero" button is also present. The footer includes the logo of the "Universidad de Jaén" and the name "Laura M.ª Miñarro Escribano".

Clasificación:	Subjetividad:	Objetividad:
Media	POSITIVO	0.269
Más Positivo	NEGATIVO	0.0
Más Negativo		
Diferencia	NEUTRO	0.731

Figura 56. Cálculos ejemplo

11. Índice de Figuras

Figura 1. Millones de usuarios en los 10 idiomas más usados en Internet (Mayo 2017) ..	7
Figura 2. Sentimentanalysis.....	9
Figura 3. Fragmento MCR	23
Figura 4. SentiWordNet	25
Figura 5. Fragmento SWN	25
Figura 6. Generación de CRiSOL	27
Figura 7. Fragmento CRiSOL	27
Figura 8. http://text-processing.com/demo/sentiment/	29
Figura 9. Análisis Morfológico	31
Figura 10. Partes de la oración	31
Figura 11. Java.....	33
Figura 12. NetBeans IDE 8.0.1	35
Figura 13. Esquema enlace bases de datos.	47
Figura 14. Generación de CRiSOL_v2.....	48
Figura 15. Cálculo de la media en CRiSOL_v2	49
Figura 16. Cálculo valores más positivos CRiSOL_v2	50
Figura 17. Cálculo valores más negativos CRiSOL_v2	51
Figura 18. Cálculo de la diferencia CRiSOL_v2	52
Figura 19. Interfaz de Usuario.....	54
Figura 20. Diagrama de flujo del conteo de synsets.....	55
Figura 21. Interfaz Media Opinión 1	56
Figura 22. Prueba 2 Más Positivo	64
Figura 23. Prueba 2 Más Negativo.....	65
Figura 24. Prueba 2 Mayor Diferencia	66
Figura 25. Versiones Netbeans.....	69
Figura 26. Proceso de descarga	70
Figura 27. Selección sistema operativo	70
Figura 28. Descarga paquete seleccionado	71
Figura 29. JDK.....	71
Figura 30. Instalación Netbeans	72
Figura 31. Diferentes paquetes para la instalación	72
Figura 32. Directorio de la instalación	73
Figura 33. Verificación de la instalación.....	73
Figura 34. Componentes y características a instalar.....	74

Figura 35. Acuerdo de licencia.....	75
Figura 36. JUnit	76
Figura 37. Ruta instalación Netbeans	76
Figura 38. Servidor GlassFish.....	77
Figura 39. Apache Tomcat.....	77
Figura 40. Instalación NetBeans IDE	78
Figura 41. Fin Instalación.....	78
Figura 42. Interfaz Usuario.....	80
Figura 43. Cuadro de texto	81
Figura 44. Buscador de Synsets	81
Figura 45. Buscador de palabras	81
Figura 46. Listado palabras con sentimiento	82
Figura 47. Funciones	82
Figura 48. Media.....	82
Figura 49. Más positivo.....	83
Figura 50. Más negativo	83
Figura 51. Diferencia.....	83
Figura 52. Cuadro de resultados.....	84
Figura 53. Generación del xml	84
Figura 54. Interfaz ejemplo	85
Figura 55. Palabras con sentimiento ejemplo	86
Figura 56. Cálculos ejemplo.....	86

12. Índice de Tablas

Tabla 1. MCR Ejemplo nombre	40
Tabla 2. MCR Ejemplo verbo	40
Tabla 3. MCR ejemplo adverbio	41
Tabla 4. MCR ejemplo adjetivo	41
Tabla 5. SWN ejemplo nombre	43
Tabla 6. STW ejemplo adjetivo	44
Tabla 7. Evaluación Opinión 1	57
Tabla 8. Estupendos	58
Tabla 9. Gusto	58
Tabla 10. Elegancia	59
Tabla 11. Exquisita	59
Tabla 12. Genial	60
Tabla 13. Encanto	60
Tabla 14. Correctamente	61
Tabla 15. Valoración palabra vulgar	62
Tabla 16. Valoración palabra violento	62
Tabla 17. Valoración palabra Coraje	63
Tabla 18. Evaluación Media Opinión 2	63
Tabla 19. Evaluación Más Positivo Opinión 2	64
Tabla 20. Evaluación Más Negativo Opinión 2	65
Tabla 21. Evaluación Diferencia Opinión 2	66

13. Glosario

- Siglas

AO: Análisis de Opiniones.

API: Application Programming Interface.

CDDL: Common Development and Distribution License.

CRiSOL: Base de Conocimiento de Opiniones para el Español.

EJB: Enterprise JavaBeans.

EVALITA: Evaluation of NLP and Speech Tools for Italian.

GNU: General Public License.

HTML: HyperText Markup Language.

ILI: Índice entre Lenguas.

JDK: Java Development Kit

J2SE: Java 2 Platform Standard Edition.

MathML: Mathematical Markup Language.

MCR: Repositorio Central Multilingüe.

ML-SentiCON: Multi-Layered Multilingual Sentiment Lexicon.

MPL: Mozilla Public License.

NLTK: Natural Language Toolkit.

PLN: Procesamiento del Lenguaje Natural.

POS: Part Of Speech.

SEL: Spanish Emotion Lexicon.

SEMEVAL: Evaluation Exercises on Semantic Evaluation.

SGML: Standard Generalized Markup Language.

SO-CAL: Semantic Orientation-Calculator.

SOL: Spanish Opinion Lexicon.

SQL: Structured Query Language.

SVM: Support Vector Machine.

SVG: Scalable Vector Graphics.

SWN: SentiWordNet.

TextPro: Text Processing Tools.

TF-IDF: Term Frequency-Inverse Document Frequency.

UML: Unified Modeling Language

WEKA: Waikato Environment for knowledge Analysis.

W3C: World Wide Web Consortium.

XHTML: Extensible Hypertext Markup Language.

XML: Extensible Markup Language.

XSL: Extensible Stylesheet Language.

14. Referencias bibliográficas

- [1] Análisis de sentimiento: qué es y para qué se usa, 2016. Brandwatch [online], <https://www.brandwatch.com/es/2015/02/analisis-de-sentimiento/>
- [2] Reconocimiento de la polaridad semántica, 2013. Historia de la Informática [online], <http://histinf.blogs.upv.es/2013/12/17/reconocimiento-de-la-polaridad-semantica/>
- [3] GONZÁLEZ, M. Dolores Molina; CÁMARA, Eugenio Martínez; VALDIVIA, M. Teresa Martín. Crisol: Base de conocimiento de opiniones para el español. Procesamiento del Lenguaje Natural, 2015, vol. 55, p. 143-150.
- [4] MOLINA-GONZÁLEZ, M. Dolores, et al. Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications, 2013, vol. 40, no 18, p. 7250-7257.
- [5] BACCIANELLA, Stefano; ESULI, Andrea; SEBASTIANI, Fabrizio. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. En LREC. 2010. p. 2200-2204.
- [6] DAVE, Kushal; LAWRENCE, Steve; PENNOCK, David M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. En Proceedings of the 12th international conference on World Wide Web. ACM, 2003. p. 519-528.
- [7] WIEBE, Janyce M.; BRUCE, Rebecca F.; O'HARA, Thomas P. Development and use of a gold-standard data set for subjectivity classifications. En Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999. p. 246-253.
- [8] Alejandrocassis, 2015, Aprendizaje Supervisado. Inteligencia Artificial 101 [online]. 21 October 2015. <https://inteligenciaartificial101.wordpress.com/2015/10/20/aprendizaje-supervisado/>

- [9] PANG, Bo, et al. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2008, vol. 2, no 1–2, p. 1-135.
- [10] LIU, Bing. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 2012, vol. 5, no 1, p. 1-167.
- [11] BROOKE, Julian; TOFILOSKI, Milan; TABOADA, Maite. Cross-Linguistic Sentiment Analysis: From English to Spanish. En RANLP. 2009. p. 50-54.
- [12] SIDOROV, Grigori, et al. Empirical study of machine learning based approach for opinion mining in tweets. En Mexican international conference on Artificial intelligence. Springer, Berlin, Heidelberg, 2012. p. 1-14.
- [13] SARALEGI X., & San Vicente I.. (2013). Elhuyar at TASS 2013. In Proceedings of XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural. (pp. 143-150).
- [14] CRUZ, Fermín L., et al. Building layered, multilingual sentiment lexicons at synset and lemma levels. Expert Systems with Applications, 2014, vol. 41, no 13, p. 5984-5994.
- [15] GONZALEZ-AGIRRE, Aitor; LAPARRA, Egoitz; RIGAU, German. Multilingual Central Repository version 3.0. En LREC. 2012. p. 2525-2529.
- [16] D, Pedro Luis, [no date], EuroWordNet. EuroWordNet [online]. <http://elies.rediris.es/elies2/cap334.htm?cv=1>
- [17] WAIKATO, The University of, Weka (3.6.10) - Malavida [online]. <http://www.malavida.com/es/soft/weka/>
- [18] Base de Conocimientos. ¿Qué es el lenguaje de programación JAVA? - Base de Conocimientos - ICTEA [online]. <http://www.ictea.com/cs/knowledgebase.php?action=displayarticle&id=8790>

[19] Bienvenido a NetBeans y www.netbeans.org. Bienvenido a NetBeans y www.netbeans.org, Portal del IDE Java de Código Abierto.

https://netbeans.org/index_es.html

[20] Definición de XML — Definición.de [online],

<http://definicion.de/xml/?cv=1>

