Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

Unidad Cuajimalpa

Licenciatura en Tecnologías y Sistemas de la Información

División de Ciencias de la Comunicación y Diseño

Proyecto Terminal:

Análisis de Sentimientos y Emociones en Documentos Digitales

José Antonio Hernández Ambrocio

Asesorado por:
M.C. A. Gabriela Ramírez de la Rosa
Dr. Esaú Villatoro Tello

Contenido

1 Introducción

A diario, usuarios activos en la web generan grandes cantidades de información en foros de opinión, blogs, redes sociales, sitios de reseñas, sitios de comercio electrónico. Según algunas estadísticas de la red: Facebook alcanza 1.2 billones de usuarios activos al día, 1.9 billones al mes [14]; Twitter tiene un promedio de 500 millones de tweets al día [15]; Amazon cuenta con 2.4 billones de visitas mensuales [16], sólo por mencionar algunos.

El tráfico de información en la web da oportunidad a figuras públicas, gobierno, empresas, investigadores, entre otros, de conocer la opinión pública sobre un tema cualquiera, de esta manera se pueden tomar acciones para cambiar o mantener la opinión general, por ejemplo, para un producto que está teniendo mala recepción, se pueden hacer cambios en la estrategia publicitaria. El análisis de sentimientos y emociones son dos tareas, del área de procesamiento de lenguaje natural, que permiten explotar dicha información.

Identificar la polaridad, positiva o negativa, de la opinión, ya sea de una persona, tema, producto o cualquier otra entidad, permite detectar defectos, fortalezas y/o oportunidades de dicha entidad, a partir de la percepción general del público.

Para lograr esto, de manera rápida y automática, se requiere de herramientas que implementen el análisis de sentimientos y emociones. Actualmente, existen herramientas que permiten hacer uno o ambos análisis, sin embargo, tienen algunas limitaciones. Entre las limitaciones más comunes se encuentran: el idioma, el costo, la cantidad de texto que puede procesar, el formato del archivo de entrada o la disponibilidad de algún esquema de visualización.

En este proyecto, se propone el desarrollo de una herramienta que sirva para llevar a cabo análisis de sentimientos y emociones, que soporte los idiomas inglés y español, en cualquier corpus textual, es decir, un conjunto de documentos, por ejemplo, notas periodísticas históricas, conjuntos de entrevistas transcritas, conjuntos de reseñas u opiniones; además de contar con esquemas de visualización de datos para los resultados.

El Análisis de Sentimientos, consiste en determinar la polaridad de un texto. La clasificación deriva usualmente de las clases positiva, negativa y, en ocasiones, neutra. Por otra parte el Análisis de Emociones, consiste en la tarea de buscar en un texto la carga afectiva de seis emociones básicas: Alegría, Enojo, Tristeza, Temor, Disgusto y Sorpresa [7]. Ambar tareas pueden ser abordadas desde tres enfoques diferentes: **a)** basados en recursos léxicos, **b)** aprendizaje supervisado e **c)** híbridos, es decir, una composición de los anteriores.

El resto de este documento está formado por los siguientes capítulos:

- Marco Teórico: se explican conceptos y elementos teóricos, fundamentales para comprender el contexto del proyecto, por ejemplo, conceptos como aprendizaje automático, clasificación de textos, modelos de representación

de texto, métricas de evaluación, etc.

- Estado del arte: se muestran algunas herramientas existentes y se hace una comparación entre las características que tiene cada una. Además se explican, de manera breve, algunos trabajos relacionados al tema de análisis de sentimientos y emociones.
- Método propuesto: se explican los métodos para análisis de sentimientos y emociones que se proponen.
- Experimentación: se explican los experimentos realizados a lo largo de la elaboración del proyecto, además de la metodología que se siguió para llevarlos a cabo.
- Integración del sistema: uno de los objetivos era integrar los módulos desarrollados a una plataforma web existente, en esta sección se muestra cómo quedan integrados los módulos en la plataforma.

1.1 Objetivos

El objetivo general del proyecto es:

Implementar una herramienta de visualización de datos, a través del uso de recursos léxicos en español e inglés para el análisis de sentimientos y emociones presentes en colecciones de textos.

1.1.1 bjetivos particulares

- Compilar recursos léxicos para el análisis de sentimientos y emociones.
- Implementar un método basado en recursos léxicos para la identificación de sentimientos y emociones.
- Proponer un esquema de visualización de información que facilite al usuario el análisis de emociones y sentimientos en textos.
- Incorporar los módulos creados en la herramienta web desarrollada previamente.

2 Marco Teórico

El objetivo de esta sección es ofrecer al lector algunos conceptos y elementos básicos que son esenciales para entender este proyecto terminal. En primer lugar, se introduce al lector el tema de aprendizaje automático; después, se trata el tema de clasificación de textos, explicando los modelos de representación que se emplearon; posteriormente, se puntualizan los recursos léxicos y finalmente, se explican las métricas de evaluación.

2.1 Aprendizaje Automático

El aprendizaje automático es una rama de la Inteligencia Artificial que trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos.

Se dice que una computadora es capaz de aprender de una experiencia E con respecto a una tarea T y una medida de desempeño P, si su desempeño en la tarea T, medida por medio de P, mejora con la experiencia E [10].

2.1.1 Clasificación de textos

La clasificación de textos es una tarea de aprendizaje automático que consiste en etiquetar un conjunto de documentos de texto en categorías temáticas de un grupo de clases definidas.

Usualmente, para desarrollar un sistema de clasificación automática de textos es necesario representar la instancia de cierta manera para poder categorizar la información. Algunos modelos de representación de información son:

Bolsa de palabras

En este método de representación, cada documento (ya sea parte del corpus inicial o resultado de fase operativa del sistema) es representado por un vector de n términos ponderados. El peso de los términos están dados generalmente entre 0 y 1 [12].

Vector término-documento

La ocurrencia de un término en un documento establece una relación entre estos. Dicha relación término-documento puede ser cuantificada usando un esquema de pesado, por ejemplo, frecuencia. En forma de matriz puede ser visto como:

$$\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{bmatrix} f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ f_{3,1} & f_{3,2} \end{bmatrix} \end{array}$$

donde cada elemento $f_{i,j}$ se refiere a la frecuencia del término k_i en el documento d_j [2].

El uso de la frecuencia es solamente un ejemplo de pesado. Las formas más comunes para dar un peso a la relación término documento son: **a)** pesado booleano, que consta en asignar peso 1 si el término k_i aparece en el documento d_j y 0 en caso contrario; **b)** pesado por frecuencia de término (TF), se asigna el número de veces que el término k_i aparece en el documento d_j. **c)** pesado TF-IDF[1], este esquema es una combinación entre el valor IDF y la frecuencia del término.

2.1.2 Métricas de evaluación

Por métrica se entiende como una medida que indica el grado que un sistema, proceso o componente posee un atributo, en este caso el grado de fiabilidad.

Una manera fácil y útil para visualizar y reportar los resultados del sistema es la *Matriz de confusión*. Una matriz de confusión es una tabla en la cual las columnas son las predicciones de clases y las filas las clases reales. La figura 1 muestra un ejemplo.

	Predicción clase positiva	**Predicción clase negativa**
Clase real positiva	Verdaderos positivos (TP)	Falsos Positivos (FP)
Clase real negativa	Falsos negativos (FN)	Verdaderos negativos (TN)

Figura 1: Ejemplo matriz de confusión

Algunas de las métricas que existen para = medir la calidad del clasificador son:

- **Precisión:** La cantidad de documentos asignados a la clase c_p,por el clasificador, que realmente pertenecen a la clase c_p.
  $P= \frac{TP}{TP+FP}$
- **Recuerdo:** Es la fracción de los documentos que pertenecen a la clase c_p y fueron asignados correctamente a la clase c_p.
  $R = \frac{TP}{TP+TN}$
- **F-Score:** Es una medida armónica que combina la precisión y recuerdo [2]
  $F = \frac{2PR}{P+R}$

[1] De: Term Frequency (TF) e Inverse Document Frequency (IDF)

2.2 Recursos Léxicos

Un léxico se define, según la RAE[2], como un conjunto de palabras de un idioma que pertenecen al uso de una región, actividad o algún campo semántico determinado.

Dada la temática del proyecto, un recurso léxico es un conjunto de palabras que están asociadas a un sentimiento (positivo o negativo). La asociación entre término-sentimiento puede ser dada explícitamente término-sentimiento, e.g., "Feliz - Positivo", o como la probabilidad de que tal término se incline a tal sentimiento.

Para la realización de este proyecto se usaron los siguientes recursos léxicos: **a)** SentiWordNet [1], cuenta con 117,660 términos aproximadamente y sólo se encuentra en inglés, por cada término tiene una etiqueta *Part Of Speech*, un identificador, un puntaje positivo y uno negativo y una corta definición; **b)** SenticNet[3], cuenta con 50,000 palabras en inglés aproximadamente, por cada término proporciona la polaridad y un valor (entre -1.0 y 1.0) de probabilidad afectiva de su polaridad; **c)** ML-Senticon [6], se trata de varias listas de lemas positivos y negativos para inglés, español, catalán, gallego y vasco. Cada lema viene acompañado de una estimación numérica de su polaridad (entre -1.0 y 1.0) así como de un valor de desviación típica de dicha polaridad. **d)** ANEW[4] [4], el cual cuenta con 1030 términos con un conjunto de valoraciones emocionales (de 0 a 10). Además hay una adaptación al idioma español con 1034 palabras [11].

Estos recursos léxicos han sido usados previamente en otros proyectos que trabajan sobre análisis de sentimientos y/o emociones. En la siguiente sección se hablará de algunos de esos proyectos.

[2]Recuperado de: http://dle.rae.es/?id=ND3Rym3
[3]Disponible en: http://sentic.net
[4]Por sus siglas: Affective Norms for English Words

3 Estado del arte

En esta sección se presentan algunos trabajos de investigación en los que utilizan distintos métodos para analizar emociones y sentimientos.

Como resultado de las grandes cantidades de información que se generan en la Web, ha crecido el interés de contar con herramientas capaces de clasificar información y mostrar tendencias. Debido a esto, existen herramientas que permiten realizar dicha tarea, sin embargo, siguen los esfuerzos por obtener resultados de mayor fiabilidad.

3.1 Análisis de sentimientos

Fermín L. Cruz et al [5]. realizaron un proyecto que consiste en la clasificación de documentos basada en la opinión. Los datos de entrada fue un corpus compuesto por críticas de cine en español. La arquitectura de su método consiste en: extraer bigramas[5] de una crítica; para cada bigrama calcula la orientación semántica (OS), es decir un valor real positivo o negativo, mediante el algoritmo PMI-IR (Pointwise Mutual Information – Information Retrieval); a partir de la suma de las orientaciones semánticas se clasifica la crítica como positiva, si el resultado es mayor a 0, o negativa en caso contrario.

Otro trabajo en el cual hubo una parte fuerte de experimentación fue el de Grigori Sidorov, et al [13]. Este trabajo trata del análisis de sentimientos sobre un corpus formado por *Tweets*. El preprocesamiento del corpus consistió en cuatro procedimientos: corrección de errores, etiquetado de atributos (url, hashtag, etc), etiquetado *Part Of Speech* y procesado de negación. Fueron utilizados tres clasificadores, basados en aprendizaje automático; Naive Bayes, Arboles de decisión C4.5, y Máquinas de vectores de soporte(SVM). En la etapa de experimentación y evaluación fueron cambiando variantes como el tamaño del corpus, número de clases, el uso de diferentes dominos. La métrica de evaluación usada fue la precisión, en general hubo buenos resultados; el rango fue de 60% hasta 85%.

3.2 Análisis de emociones

Carlo Strapparava et al [17]. realizaron un experimento para encontrar emociones en un conjunto de datos, formado por títulos de noticias. Hicieron uso del recurso léxico, SentiWordNet Affect, que es una extensión de SentiWordNet. Desarrollaron cinco sistemas: el primero, el sistema más básico, busca la presencia de palabras en el léxico WordNet Affect; el segundo, tercero y cuarto método son una variación de Latent Semantic Analysis (LSA), usando una bolsa de palabras denotando la emoción, añadiendo los sinonimos en los synsets de

[5]Se tratan los términos de un texto en pares, e.g., hola-amigo

WordNet y todas las palabras emocionales respectivamente; el quinto es bajo un enfoque supervisado, fue implementado con un clasificador Naive Bayes. Los resultados no fueron concluyentes, pues fueron muy variados y debido a que sólo se reporta la medida de precisión es difícil ver un panorama general del desempeño de sus métodos.

En el trabajo de José R. Gálvez-Pérez, et al. [8] fue creado un sistema basado en recursos léxicos para la opinion pública en Twitter. Fue utilizado el léxico *Spanish Emotion Lexicon* (SEL) [13] en el que cada término tiene una probabilidad de uso afectivo (PFA). Su método de clasificación comienza determinando si un *tweet* $\boldsymbol{T}$ es de opinión o informativo, si es informativo es considerado neutro; después se hace una comparación de cada palabra en $\boldsymbol{T}$, en busca del término T_i en SEL. Se hace una sumatoria de los valores del PFA de todos los términos encontrados. El F-Score fue relativamente bajo pues en en la medida $\boldsymbol{F}$ su máximo fue de 64

De los trabajos mencionados, se retomaron algunos de los léxicos que les funcionaron mejor, por ejemplo, Spanish Emotion Lexicon (SEL) para trabajar la parte de emociones y por otro lado SentiWordNet para la parte de sentimientos. También se retomó el uso de algoritmos de aprendizaje, como, Naive Bayes y SVM.

3.3 Herramientas existentes

Actualmente existen herramientas para analizar la polaridad de sentimientos como: Sentiment Analysis with Python NLTK[6] , Alchemy de IBM[7] y Sentiment Analysis de Stanford[8].

En la tabla 1 se pueden observar las características que tiene la propuesta de este proyecto contra otras herramientas. La característica que más denota en la propuesta es que cuenta con un esquema de visualización de los resultados, el soporte de formatos de archivos, que funciona con textos en español.

3.3.1 Sentiment Analysis with Python NLTK

Esta aplicación funciona usando un clasificador que fue entrenado con tweets y reseñas de películas de un conjunto de datos creado por Bo Pang y Lillian Lee. Los resultados son más acertados si el texto de entrada son del mismo dominio que los datos con los que fue entrenado el clasificador. En la figura 2 se muestran los resultados de una prueba que se hizo ingresando un fragmento de una crítica de película.

[6]Disponible en: http://text-processing.com/demo/sentiment/
[7]Disponible en: https://alchemy-language-demo.mybluemix.net
[8]Disponible en: http://nlp.stanford.edu/sentiment/

Tabla 1: Tabla comparativa de algunas aplicaciones existentes y la aplicación propuesta

Criterios	Herramientas			
	Propuesta	Sentiment Analysis with Python NLTK	Alchemy - Watson IBM	Sentiment Analysis by Stanford
Idiomas	ES, EN	EN, FR, EN	Multilenguaje	EN
Visualización gráfica de resultados	✓	✗	✗	✓
Código Abierto	✓	✗	✗	✓
Formatos de archivos	PDF, TXT, DOC	TXT	HTML, TXT, URL	TXT
Limite de longitud de texto	No	50,000 caracteres	Sin dato	200 líneas
Análisis de Sentimientos	✓	✓	✓	✓
Análisis de Emociones	✓	✗	✓	✗
Gratuito	✓	✓	✗	✓

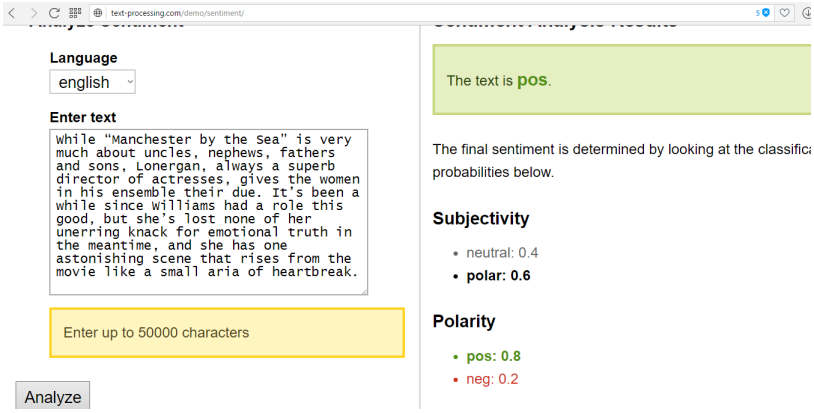

Figura 2: Pantalla de resultados de Sentiment Analysis with Python NLTK. En la parte izquierda se encuentran los datos de entrada; de lado derecho se ven los resultados del análisis, mostrando la polaridad que determinó el clasificador y en la parte inferior se muestra un resumen con más detalle diciendo la probabilidad resultante para cada clase.

3.3.2 Alchemy

El servicio IBM Watson^{TM} AlchemyLanguage [9] es una colección de APIs para análisis de texto que proporcionan información semántica del contenido de entrada. Puede tener como entrada texto, HTML o un URL [4]. Entre los resultados que proporciona del análisis, se encuentra la de análisis de sentimientos y análisis de emociones. En la figura 3 y 4 se muestra la pantalla de resultados de la plataforma Web.

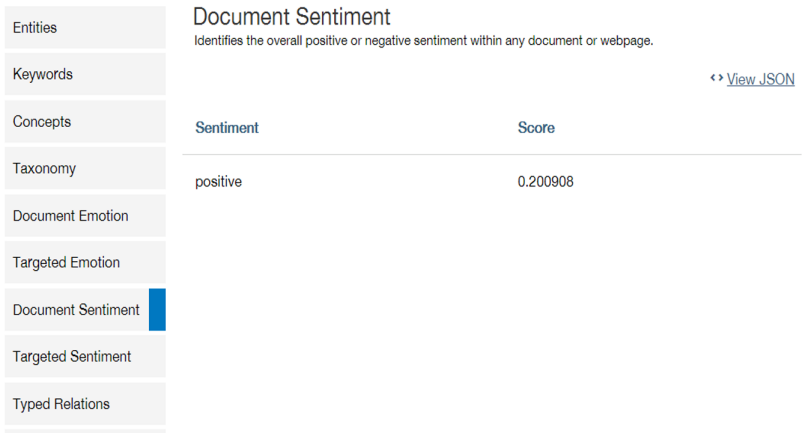

Figura 3: Pantalla de resultados de Alchemy de IBM. De lado derecho está un menú con las opciones de resultados que se pueden visualizar tras completar el análisis, en este caso se encuentra activa la pestaña de "Document Sentiment". De lado derecho se muestra el resultado

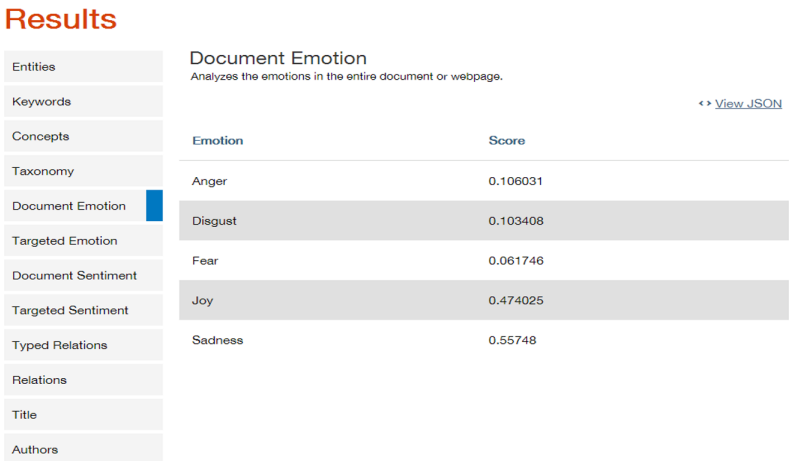

Figura 4: Pantalla de resultados de Alchemy de IBM. De lado derecho está un menú con las opciones de resultados que se pueden visualizar tras completar el análisis, en este caso se encuentra activa la pestaña de "Document Emotion". De lado derecho se muestran una tabla donde cada fila es una emoción acompañada de un puntaje.

3.3.3 Sentiment Analysis by Stanford

Funciona construyendo una representación de oraciones completas basada en la estructura de una oración; de esta manera se considera el orden de las palabras

y evita perdida de información. Por ejemplo, en la oración: “This movie was actually neither that funny, nor super witty” (“En realidad esta película no es divertida ni ingeniosa”) se encuentran las palabras funny y witty, que por sí solas tienen una polaridad positiva, sin embargo, la oración es negativa.

En la figura 5 se muestra un ejemplo de los resultados que arroja la herramienta.

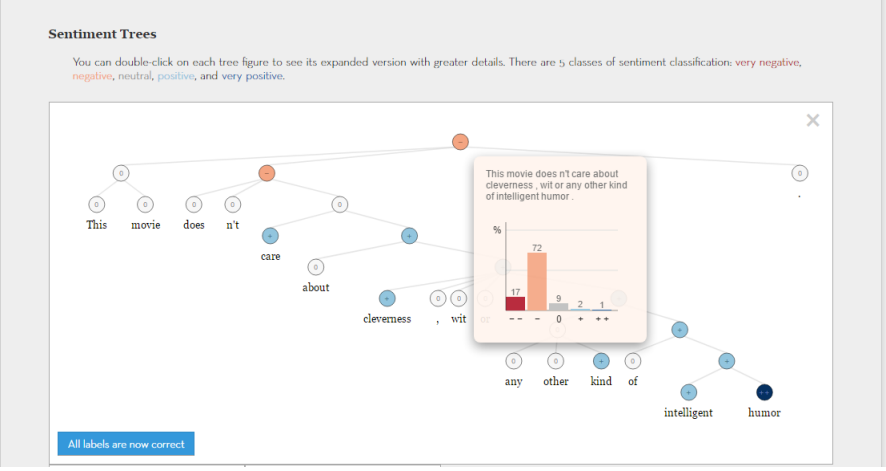

Figura 5: Pantalla de resultados de Sentiment Analysis by Stanford. Los resultados son mostrados en un árbol donde: cada nodo es una oración y el nodo raíz es el documento completo. Para cada nodo se busca darle una polaridad (muy negativa, negativa, neutral, positiva, muy negativa
, a partir de la polaridad de los nodos, se asigna una clase. Si se apunta a un nodo se pueden ver más detalles de ese nodo.

A partir de esto, se puede decir que la plataforma desarrollada tiene como ventaja que: soporta los idiomas español e inglés, cuenta con esquemas de visualización más fáciles de entender, soporta más formatos de archivos de entrada, cuenta con análisis de sentimientos como de emociones y es gratuito

4 Método Propuesto

En esta sección se explican los métodos de análisis de sentimientos y emociones que se proponen para cumplir el objetivo. Se propone usar dos enfoques: basado en recursos léxicos e híbrido que es una combinación de los enfoques basado en recursos léxicos y de aprendizaje automático. Para el análisis de sentimientos se plantea utilizar tanto enfoque en recursos léxicos e híbrido, mientras que para el análisis de emociones solo se cuenta con el enfoque basado en recursos léxicos debido a que no se contó con una colección de textos etiquetada.

4.1 Método base

Se implementó un método base para análisis de sentimientos y uno para análisis de emociones.

4.1.1 Análisis de sentimientos

Para calcular la polaridad del documento, teniendo como entrada una colección de documentos C y un recurso léxico L: se obtiene la lista de tokens de C.

Posteriormente, para cada término, o token, W de C que se encuentre en L, se suma la estimación afectiva de W; teniendo al final, de este proceso, como salida un valor numérico que es la polaridad. Finalmente se asigna la clase en función del valor de POL. Se asigna clase positiva si la polaridad es mayor a cero y, en caso contrario, clase negativa.

En la figura 6 se puede ver el esquema general del funcionamiento del método base.

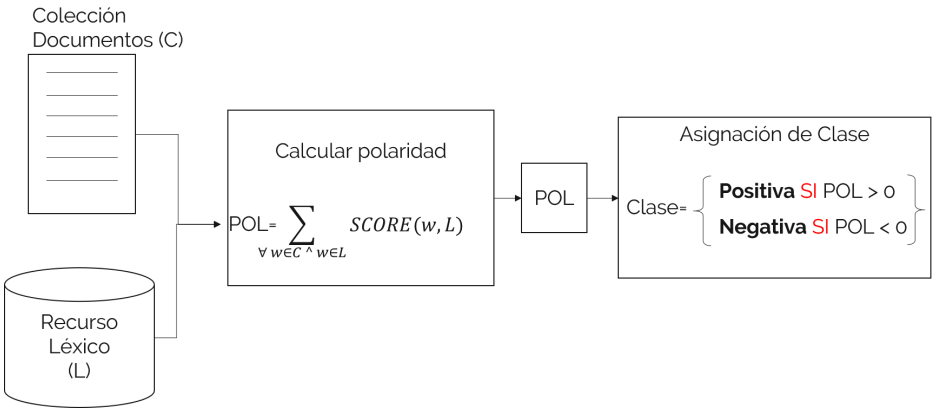

Figura 6: Esquema de método base para determinar la polaridad de una colección de documentos

4.1.2 Análisis de Emociones

Similar al anterior, se tiene una colección de documentos C, un recurso léxico L. Para este método se usa el recurso léxico en español propuesto por Grigori Sidorov [13]. En el cual, cada palabra está asociada a una categoría (alegría, enojo, miedo, disgusto, tristeza y sorpresa) y a un valor denominado Factor de Probabilidad de uso Afectivo (PFA[9]), e.g., "abundancia 0.83 Alegría"".

Para el cálculo de probabilidad afectiva de cada emoción del documento C: por cada término W del documento C que se encuentre en L, se determina a que categoría (emoción) E pertenece W y posteriormente se suma el valor PFA a la categoría E a la categoría E_i. En la figura 7 se muestra un esquema del funcionamiento de este método.

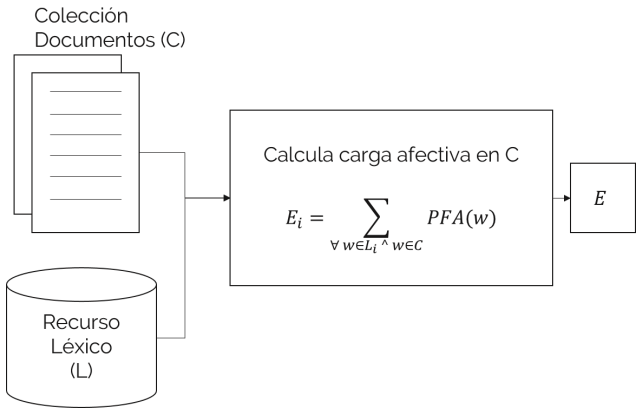

Figura 7: Esquema de método base

4.1.3 Estandarización de recursos léxicos

Aunque la mayoría de los recursos léxicos cuentan con casi la misma información (ver sección 2.2), sin embargo, se pueden presentar en distintos formatos, además, vienen ordenados de diferente manera o algunos tienen más información que otros.

Una parte de este trabajo, que representa una ventaja, fue adaptar todos los recursos léxicos al formato estándar término-polaridad. Esto fue llevado a cabo pues el método base se implementó con la idea de que funcionara con cualquier léxico que tuviese dicho formato, en vez de implementar un algoritmo para cada léxico.

El hecho de tener como entrada un léxico en la forma término-polaridad, da la ventaja de que a futuro se de la opción al usuario de subir un léxico diferente.

[9]Por su acrónimo en inglés: Probability Factor of Affective use

4.2 Método Híbrido

Un método híbrido es una combinación entre los métodos de aprendizaje y basado en recursos léxicos.

Como ya se mencionó, existen métodos de clasificación que funcionan bajo un enfoque supervisado, en el cual las columnas de la matriz término-documento, que es usada para entrenar un modelo[10], están conformadas por el vocabulario que compone el documento. Para un método híbrido, la bolsa de palabras está compuesta por el vocabulario del léxico.

La implementación de este método fue llevada a cabo en dos etapas: la primer etapa consiste en construir un modelo clasificador y la segunda en usar dicho modelo para clasificar una nueva instancia.

Para construir el modelo, se tiene como entrada un léxico y un corpus textual etiquetado, es decir, se conoce a que clase pertenece cada documento; después el corpus pasa por un proceso en el que se descarta toda palabra que no aparezca en el documento, a partir de esto se tiene el vocabulario que compone la bolsa de palabras; posteriormente se construye una matriz término-documento para después pasarlo como entrada al algoritmo de aprendizaje; finalmente, se obtiene como salida un clasificador.

Una vez que se haya construido un modelo clasificador, al llegar un nuevo documento D para clasificar, de entrada, se tiene el corpus textual que se desea clasificar y el vocabulario que se creó durante la etapa de entrenamiento; similar al anterior, se eliminan palabras del corpus que no pertenezcan al vocabulario y se forma la matriz término-documento; finalmente, se obtiene la clase de las instancias usando el clasificador, construido en la primer etapa, pasando como entrada la matriz término-documento. En la figura 8 se presenta un esquema con la arquitectura que sigue el método híbrido y se puede ver el proceso que se explicó.

[10] A partir de éste se hacen las predicciones de nuevas instancias.

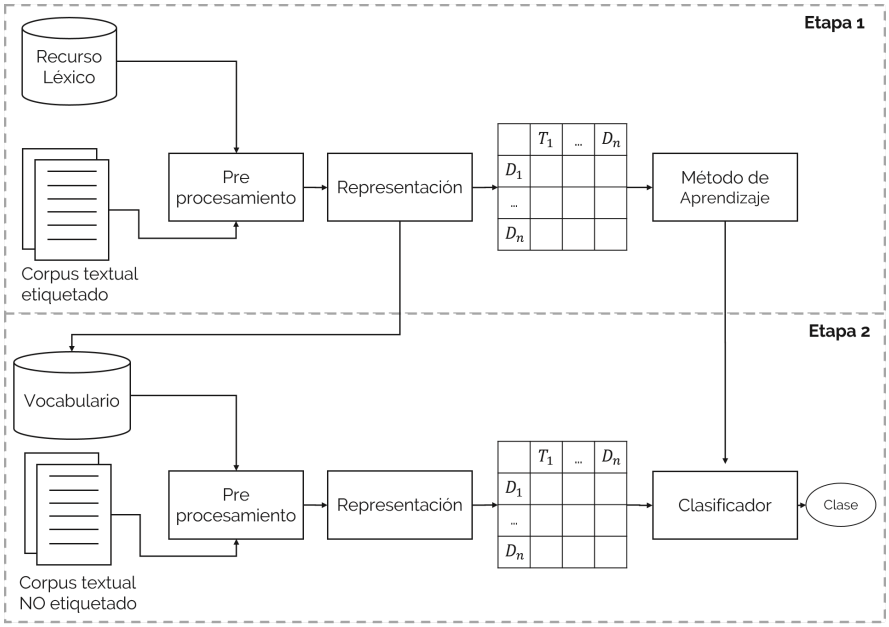

Figura 8: Esquema de método híbrido

5 Experimentación

En esta sección se explicarán los diferentes experimentos que se desarrollaron con el objetivo de encontrar el método con mejores resultados, para cada enfoque que se plantea implementar. Cabe mencionar que la parte experimental se centró en el análisis de sentimientos dado que sólo se contaba con corpus etiquetados para la clasificación de sentimientos.

5.1 Colecciones de datos

Para poder medir la confianza que tiene el sistema, se requiere de una colección de documentos etiquetados, i.e., cuya polaridad sea conocida a priori. Los experimentos de este proyecto se hicieron trabajando con dos corpus: uno en español y el segundo en inglés.

El corpus en inglés fue construido por John Blitzer, et al [3]. Está compuesto por opiniones de cuatro tipos de productos en Amazon: Libros, DVDs, Electrónicos, Artículos de Hogar; contando con 1000 opiniones por dominio. Con las opiniones viene una puntuación de 0 a 5 estrellas al producto. A partir de esa puntuación se etiquetaron las opiniones; a todas las opiniones calificadas de 3 a 5 fueron etiquetadas como positivas y a las opiniones de 0 a 2 se les etiquetó como negativas, las opiniones con 3 estrellas se les ignoró.

Respecto al corpus en español, constaba de un total de 3,553 críticas de cine extraídas del sitio *www.muchocine.net* [5]. Cada crítica cuenta con una calificación que va de 1 a 5. Para separar las críticas se consideraron como positivas a todas aquellas que contaran con una calificación mayor a 3, y en caso contrario a todas las críticas que tuviesen una calificación menor de 3, se les consideró negativas, el resto fueron ignoradas. Tras la separación de las críticas, según su calificación, el corpus fue reducido a 2,304 críticas, de las cuales 1174 resultaron ser positivas, y 1130 negativas.

En ambos casos se requirió un pre procesamiento antes de comenzar a evaluar el sistema. De manera general, para ambos corpus, se eliminaron signos de puntuación, se convirtió todo a minúscula, se hizo un parseo a codificación UTF-8 para evitar perdida de información.

5.2 Configuración experimental

- **Experimento 1: Recursos Léxicos**

  El objetivo de este experimento es encontrar el léxico que se comporte mejor utilizando el método base (sección 4.1) para el análisis de sentimientos. Utilizando como métrica el F-Score para comparar el desempeño entre los léxicos. Los recursos léxicos que se usaron son: SentiCon, SenticNet, SentiWordNet y ANEW.

- **Experimento 2: Ensamble de léxicos**

  El objetivo es crear un ensamble entre un conjunto de léxicos, partiendo de la idea de que funcionaría mejor si se ponderaban los valores de cada término en los léxicos, dando un peso de 0.5 al léxico que funcionó mejor en el experimento 1. También hubo una segunda propuesta para realizar el ensamble, la cual constó en hacer un promedio de los valores. De igual manera, los recursos que se utilizaron fueron: SentiCon, SenticNet, SentiWordNet y ANEW.

- **Experimento 3: Método híbrido**

  Este experimento fue útil para la creación del modelo clasificador, que será utilizado en la plataforma web, para la predicción de un nuevo documento. Se utilizaron dos algoritmos de clasificación: *Naive Bayes* y el algoritmo *Optimización Mínima Secuencial* (SMO[11]) que es una mejora del algoritmo *Máquinas de vectores de soporte* (SVM[12]).

  La primera tarea a realizar en este experimento fue pasar la colección de documentos etiquetada a un vector término-documento. Partiendo de que es un enfoque híbrido; se construye la matriz, donde los atributos son sólo aquellos términos de la colección que aparecen en el recurso léxico. Se creo una matriz por cada sistema de pesado (Booleano, TF, TF-IDF), esto con el objetivo de comparar resultados y encontrar la mejor opción. Posteriormente, se entrena un modelo para cada matriz por cada uno de los dos algoritmos de aprendizaje (Naive Bayes y SMO).

  Para evaluar el modelo se hizo uso de la técnica validación cruzada [2],que consiste en construir k clasificadores: $\psi_1, \psi_2 ... \psi_k$. Para validar el clasificador se dividen los datos de entrenamiento Dt en k partes: $N_{t1}, N_{t2} ... N_{tk}$. El clasificador ψ_i usa la i-ésima parte como datos de prueba y el resto del documento como datos de entrenamiento. Posteriormente, cada clasificador es evaluado independientemente usando las métricas: precisión, recuerdo, F-Score. Finalmente termina con el cálculo de las métricas de los k clasificadores.

5.3 Resultados y análisis

5.3.1 Colección de documentos en inglés

- **Experimento 1: Recursos Léxicos**

  El recurso léxico que se comportó mejor fue SentiWordNet para los cuatro dominios. También se puede notar que el dominio que obtiene mejores resultados es el de *Artículos de hogar*. Una razón de esto es por el tamaño del vocabulario de los léxicos, además del peso que tiene asignado un término en cada léxico.

[11] Por sus siglas en inglés: Sequential Minimal Optimization
[12] Por sus siglas en inglés: Support Vector Machine

En general, SentiCon y SenticNet obtuvieron mejores números al clasificar documentos positivos, 99% y 94% respectivamente, por otro lado tuvieron malos resultados al clasificar documentos que pertenecen a la clase negativa, 5% y 16% respectivamente. Mientras que con SentiWordNet, aunque tiene menor porcentaje de los documentos positivos clasificados correctamente a 82%, se compensa al clasificar documentos negativos pues se obtiene un 35% de recuerdo y aunque igual es un número bajo, es el doble que SenticNet. Esta fue una razón fundamental que determinó el desempeño general de cada léxico. En la tabla 2 se puede ver el F-Score, que da un panorama más concreto del desempeño general de cada léxico.

- **Experimento 2: Ensamble de léxicos** Como se menciono anteriormente, se hicieron dos ensambles: uno en el cual se daría mayor peso al léxico que se desempeñara mejor y en este caso fue SentiWordNet, por lo tanto la ponderación quedó con 0.5 para SentiWordNet, 0.25 a SenticNet y 0.25 para SentiCon; y un segundo ensamble que consiste en promediar los valores. Al hacer este ensamble se esperaba obtener mejores resultados que usar los léxicos de manera individual, sin embargo, los resultados, para esta colección, no fueron los esperados.

  La razón principal de que fallara este experimento es porque los valores de estimación afectiva en los léxicos se contradicen, es decir, mientras en un léxico un término es positivo, en otro es negativo y en algunos casos la diferencia es relativamente grande. Un ejemplo de esto se puede ver en la figura 9.

Término	SentiWordNet	SenticNet	SentiCon
curative	0.875	- 0.59	0.625
clumsiness	-0.375	0.772	-0.5

Figura 9: Ejemplo de dos contradicciones en los léxicos.

En la tabla 2 se reportan los resultados de los experimentos 1 y 2. Se puede ver que el léxico con mejor desempeño es SentiWordNet, mientras que SentiCon tiene un desempeño muy bajo en comparación con los demás.

- **Experimento 3: Método híbrido**

  Se construyeron 24 clasificadores, considerando que son 2 algoritmos de clasificación, 3 tipos de pesado y 4 dominios. En la tabla 3 se puede ver una comparación del F-Score resultante para cada dominio. También se observa que el pesado booleano es el que mejor funciona para ambos algoritmos, siendo 79% el porcentaje más alto para Naive Bayes y 80% para SMO; a partir de estos porcentajes se deduce que la mejor opción es el modelo creado para el dominio de Artículos de Hogar usando el algoritmo SMO con un pesado booleano.

		Libros	DVD	Electrónicos	Artículos hogar	Promedio
	SentiCon	38%	38%	40%	41%	39%
Experimento 1	**SenticNet**	47%	53%	52%	54%	51%
	SentiWordNet	56%	59%	60%	63%	59%
Experimento 2	**Ensamble 1**	48%	49%	48%	50%	49%
	Ensamble 2	47%	48%	48%	50%	48%

Tabla 2: Resultados de los experimentos 1 y 2 para la colección en inglés. En esta tabla se reporta el F-Score, la última columna es el promedio del F-Score. El ensamble 1 es la ponderación: 0.5 a SentiWordNet, 0.25 a SentiCon y 0.25 a SenticNet. El ensamble dos es el promedio entre los tres léxicos.

		Colección en inglés				Colección español
		Libros	DVD	Electrónicos	Artículos hogar	Mucho Cine
	BOOL	71%	79%	78%	78%	63%
Naive	**TF**	68%	70%	71%	71%	59%
Bayes	**TF-IDF**	68%	70%	71%	71%	58%
	BOOL	74%	78%	78%	80%	64%
SMO	**TF**	73%	75%	75%	77%	63%
	TF-IDF	73%	70%	75%	77%	63%

Tabla 3: Resultados para método híbrido. Se muestran los resultados para ambas cada dominio de cada colección, cuatro para inglés y uno en español. Horizontalmente, la tabla está divida en dos, indicando los algoritmos de aprendizaje. Además están descritos los tres esquemas de pesado (Bool, TF, TF-IDF).

5.3.2 Colección de documentos en español

- **Experimento 1: Recursos Léxicos** Los recursos léxicos utilizados para esta colección fueron ANEW y SentiCon, obviamente ambos en su versión en español. En este caso fue ANEW quien tuvo un mejor desempeño.

  Similar al experimento aplicado en la colección en inglés, ambos tuvieron una precisión muy similar, pero fue el recuerdo el que tuvo mayor diferencia y fue este el que afectó el desempeño general. SentiCon clasificó 91% correctamente de los documentos etiquetados cómo positivos mientras que ANEW obtuvo un 65% que es un número bajo si se compara con el otro léxico. Por otro lado, fue ANEW quien clasificó 34% de los documentos negativos correctamente y SentiCon un 13%. En la tabla se muestra el desempeño general de ambos léxicos.

- **Experimento 2: Ensamble de léxicos**

  Para este experimento, sólo se hizo un ensamble que se formó a partir de dar una ponderación de los léxicos. En este caso fue ANEW el que funcionó mejor por lo que se le dio mayor peso (0.6).

  A diferencia de la colección en inglés, aquí si se obtuvieron resultados favorables usando un ensamble de ambos léxicos. En la tabla 4 se muestra el F-Score que se obtuvo para el ensamble.

		F-Score
Experimento 1	Anew	48%
	Senticon	43%
Experimento 2	Ensamble	54%

Tabla 4: F-Score para experimento 1 y 2

- **Experimento 3: Método híbrido** En este experimento se construyeron 6 clasificadores, 3 por cada algoritmo de aprendizaje (Naive Bayes y SMO) usando los tres esquemas de pesado (Booleano, Frecuencia de término y TF-IDF). El clasificador que funcionó mejor fue el que se construyó usando el algoritmo SMO con pesado booleano con un 64$ de puntaje para F-Score, aunque los resultados fueron similares para todos los clasificadores. En la tabla 3 donde se comparan los resultados, reportando el F-Score.

  De los resultados obtenidos en cada uno de los experimentos, se puede observar que la mejor opción, para la tarea de análisis de sentimientos, es usar el método híbrido, bajo el algoritmo SMO, para ambos lenguajes. En la colección de documentos en inglés, este método obtuvo un puntaje de 80% en F-Score contra un 63% si se usa el método basado en recursos léxicos, usando el léxico SentiWordNet. Mientras que para la colección en español, el método hibrido obtuvo un puntaje de 63% contra un 54% de parte del método basado en recursos léxicos usando el ensamble.

6 Integración del Sistema

Actualmente existe una plataforma web [13] para la búsqueda y visualización de concordancias y frecuencias en documentos digitales. Uno de los objetivos de este proyecto es integrar los módulos de análisis de sentimientos y emociones a dicha plataforma web. Se piensa que en un futuro a mediano plazo la plataforma web cuente con diversos módulos para el análisis de textos.

A continuación se explica como se integran los módulos desarrollados en el sistema.

6.1 Esquema General

Durante este proyecto, se agrega a la plataforma las opciones de: (I) realizar análisis de sentimientos y de emociones bajo un enfoque basado en recursos léxicos, (II) análisis de sentimientos usando un algoritmo de aprendizaje.

En la figura 10, se muestra el proceso que se sigue al hacer un análisis usando el método basado en recursos léxicos. Como entrada se tiene un corpus textual que pasa a un modulo de pre procesamiento (quitar signos de puntuación, convertir a minúsculas, etc.) que da como resultado un "documento limpio", posteriormente se pasa a los módulos propuestos, análisis de sentimientos y análisis de emociones, que reciben información de un recurso léxico; finalmente, el resultado es presentado al usuario bajo algún esquema de visualización.

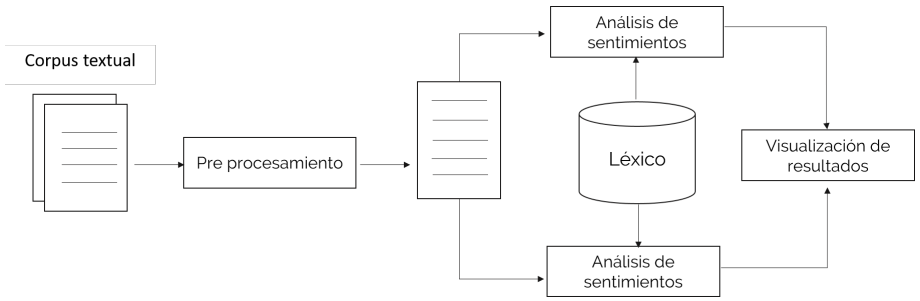

Figura 10: Esquema de los módulo de análisis de sentimientos y emociones bajo un enfoque de clasificación basado en recursos léxicos

En la figura 11 se muestra un diagrama de las clases que componen el sistema. Este diagrama incluye las funciones implementadas antes de este proyecto, además de las integradas durante el desarrollo de este trabajo. Se incluyen métodos a las clases *"Módulo de Visualización"* y *"Operaciones de Análisis de Textos"* y una clase que funciona para gestionar los léxicos.

[13] Disponible en: http://hao.cua.uam.mx:8080/Corpus/pages/index.html

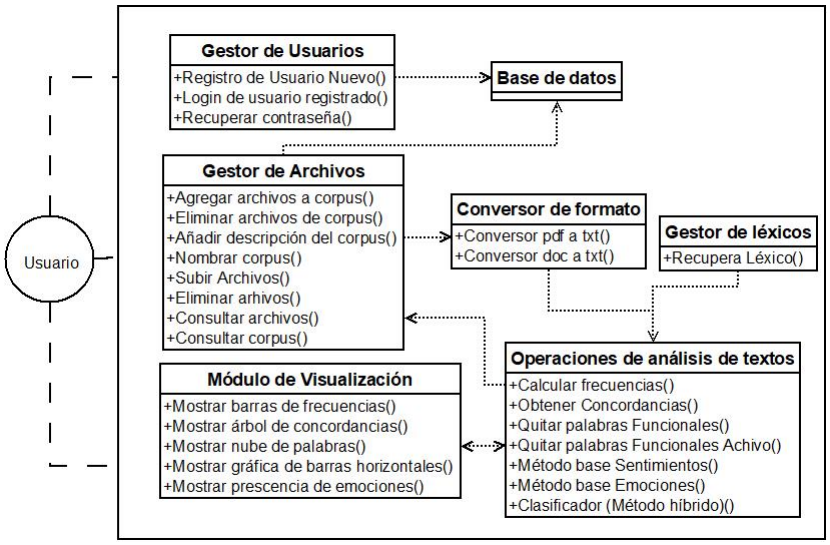

Figura 11: Diagrama de clases del sistema.

6.2 Esquemas de visualización

Una parte importante en este proyecto es el uso de esquemas de visualización para mostrar los resultados del análisis. El sistema cuenta con dos esquemas diferentes: el primero, dedicado para análisis de sentimientos; el segundo, para análisis de emociones. Se hizo uso de la librería *D3.js*[14] de JavaScript que sirve para la manipulación de datos.

El primer esquema es una gráfica de barras horizontales. En el eje de las Y se encuentran cada una de las instancias analizadas y en el eje X se muestra la polaridad. En la figura 12 se muestra un ejemplo.

El segundo esquema hace uso de gráficas de dona y es dedicado al análisis emociones. Por cada instancia en una colección de datos se genera una gráfica que indica que tan presente está cada emoción. En la figura 13 se muestra un ejemplo.

6.3 Vistas del sistema

A continuación se muestran las pestañas que se agregaron a la plataforma web. En las figuras 14 y 15 se describen los elementos principales en el área de trabajo correspondiente a cada sección.

[14]Disponible en: https://d3js.org

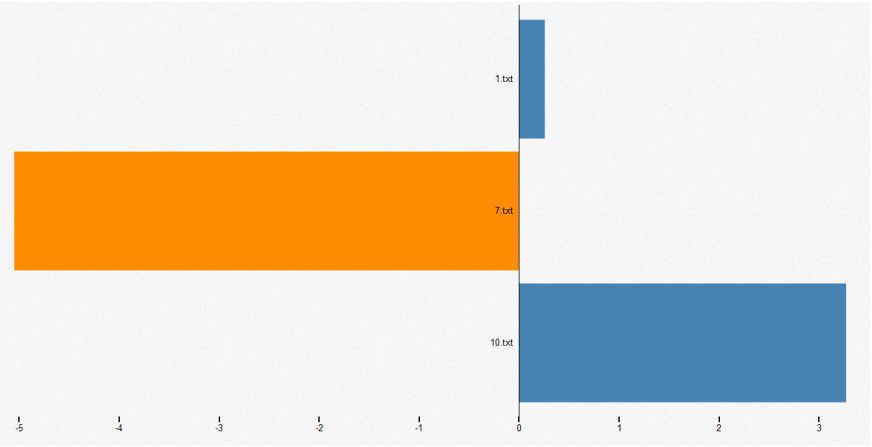

Figura 12: Esquema de visualización: Gráfica de barras horizontal.

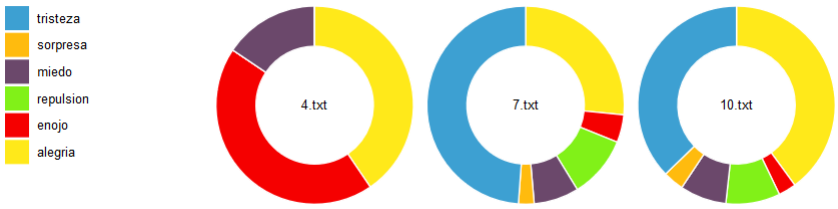

Figura 13: Esquema de visualización: Gráfica de donas.

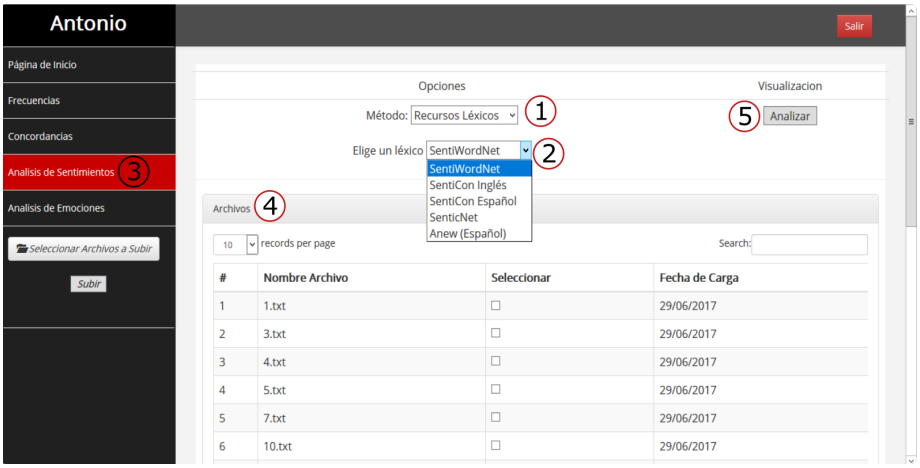

Figura 14: Vista para la sección de Análisis de Sentimientos. En está figura se puede visualizar que la pestaña para análisis de sentimientos está activa(3). En esta área de trabajo se puede seleccionar archivos o un corpus (4), creados con anterioridad, para llevar a cabo un análisis de sentimientos. El usuario puede elegir que método (basado en recursos léxicos o híbrido) desea usar (1), en caso de elegir usar el método basado en recursos léxicos, deberá elegir que léxico utilizar. En la parte derecha superior se encuentra un botón que comienza a hacer el análisis (5).

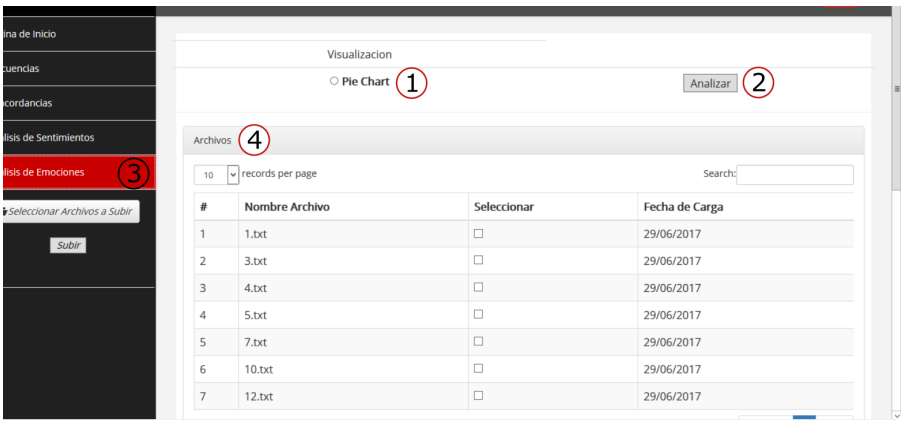

Figura 15: Vista para la sección de Análisis de emociones. En esta ilustración se muestra activa la sección de *"Análisis de Emociones* (3). En está pestaña se muestran las opciones de visualización (1). En la parte inferior se encuentra una tabla con los archivos que ha subido el usuario (4); en la parte derecha superior se encuentra un botón que comienza a hacer el análisis (2).

7 Conclusiones

En esta sección se exponen las conclusiones a las que se llegaron tras terminar este proyecto.

Se puede decir que los objetivos que se plantearon al inicio de este proyecto se han alcanzado al haber desarrollado e integrado, a una plataforma web, un módulo para análisis de sentimientos y un módulo para análisis de emociones.

Una de las motivaciones de este proyecto era la limitación de idiomas en las herramientas existentes, dicha limitación fue afrontada usando recursos en español e inglés y debido a eso la plataforma soporta documentos en español e ingles.

Otro de los objetivos, y motivaciones, que se cumplió de manera aceptable, en este proyecto fue el uso de esquemas de visualización de resultados que permite al usuario poder hacer análisis de textos de manera más fácil.

También se puede asegurar que los resultados que da el sistema tienen un grado de confianza bastante aceptable; esto debido a que hubo una fuerte etapa de experimentación para hallar variables o variaciones en los métodos propuestos para llegar a un mayor grado de confianza.

Se espera que la aplicación desarrollada en este proyecto facilite a los usuarios a procesar, de manera automática, grandes cantidades de información y a partir de los esquemas de visualización poder hacer análisis cualitativos y use esa información como mejor le convenga.

7.1 Trabajo a futuro

- Agregar opciones de visualización a la plataforma para ambos módulos, que ofrezcan más información, con el objetivo de que el usuario pueda hacer un análisis más profundo.
- Implementar un método híbrido aplicado al análisis de emociones.
- Agregar la opción de que el usuario pueda poner una URL para descargar información de dicha página y aplicar análisis de sentimientos y/o emociones a esa información.

Referencias

[1] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition, 2008.

[3] J. Blitzer, M. Dredze, F. Pereira, et al. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007.

[4] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, 1999.

[5] F. Cruz, J. Troyano, F. Enriquez, and J. Ortega. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. In *Procesamiento del lenguaje Natural*, volume 41, pages 73–80, 2008.

[6] F. Cruz Mata, J. A. Troyano Jiménez, B. Pontes Balanza, and F. J. Ortega Rodríguez. Ml-senticon: un lexicón multilingüe de polaridades semánticas a nivel de lemas, 2014-09.

[7] P. Ekman and W. Friesen. *The repertoire of nonverbal behavior: Categories, origins, usage, and encoding*, volume 1. semiotica, 1 edition, 1968.

[8] J. Gálvez-Pérez, B. Gómez-Torrero, R. Ramírez-Chávez, K. Sánchez-Sandoval, V. Castellanos-Cerda, R. García-Madrid, H. Jiménez-Salazar, and E. Villatoro-Tello. Sistema automático para la clasificación de la opinión pública generada en twitter. *Research in Computing Science*, (95), 2015.

[9] IBM. Recuperado de: https://www.ibm.com/watson/developercloud/doc/alchemylanguage/index.html 01/07/2017.

[10] T. Mitchel. *Machine Learning*. McGraw Hill, 1 edition, 1997.

[11] J. Redondo, I. Fraga, I. Padrón, and M. Comesaña. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605, 2007.

[12] F. Sebastiani. A tutorial on automated text categorisation. In *Proceedings of the 1st Argentinian Symposium on Artificial Intelligence (ASAI 1999)*, pages 7–35, 1999.

[13] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, and J. Gordon. Empirical study of machine learning based

approach for opinion mining in tweets. In *Proceedings of the 11th Mexican International Conference on Advances in Artificial Intelligence - Volume Part I*, MICAI'12, pages 1–14, Berlin, Heidelberg, 2013. Springer-Verlag.

[14] SocialBakers. All facebook statistics in one place, 2017. Consultado: 2017-06-19, recuperado de:: https://www.socialbakers.com/statistics/facebook/, author=SocialBakers.

[15] SocialBakers. All twitter statistics in one place, 2017. Consultado: 2017-06-19, recuperado de:: https://www.socialbakers.com/statistics/twitter/.

[16] Statista. Web visitor traffic to amazon.com, 2017. Consultado: 2017-06-19, recuperado de: https://www.statista.com/statistics/623566/web-visits-to-amazoncom/.

[17] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560, New York, NY, USA, 2008. ACM.