Universidad Internacional de La Rioja (UNIR)

Escuela Superior de Ingeniería y Tecnología

Grado en Ingeniería Informática

Análisis de Sentimientos a través de Twitter

Ubicación del código fuente:
https://github.com/deivit16/TFG

Trabajo Fin de Grado

presentado por: Montoya Ruiz, David

Director/a: Soltero, Francisco

Ciudad: Madrid
Fecha: 18/07/2019

Índice

Resumen

Desde hace varias décadas la manera en la que los usuarios interactúan e intercambian opiniones ha cambiado, principalmente debido a la aparición y el crecimiento experimentado en las redes sociales.

El presente trabajo TFG se encuentra enmarcado dentro de los conceptos de procesamiento del lenguaje natural y analítica del dato cuyo objetivo será desarrollar una aplicación software con interfaz web que permita realizar un análisis de sentimiento del léxico de una muestra de tuits en la red social Twitter.

En función del número de usuarios que interactúen al mismo tiempo sobre esa temática, causará viralidad o tendencia.

Este trabajo fin de grado, realiza un estudio de las opiniones que tienen los usuarios identificando los sentimientos positivos y negativos a través de sus tuits. Además, estudia distintas características relacionadas con el lenguaje natural.

Palabras clave: Procesamiento de lenguaje natural, análisis de sentimiento, estudio de opiniones, Twitter, aplicación software.

Abstract

For several decades, the way in which users interact and exchange opinions has changed, mainly due to the appearance and growth experienced in social networks.

The present work TFG is framed within the concepts of natural language processing and data analytics whose aim will be to develop an application with a web interface that allows to perform a sentiment analysis of the lexicon of a sample of tweets in the social network Twitter.

Depending on the number of users who interact at the same time on that subject, it will cause virality or tendency.

This end-of-grade project, makes a study of the opinions that have users, identifying the positive and negative sentiments through their tweets. In addition, study the different characteristics related to natural language.

Key words: Natural Processing Language, Sentiment Analysis, study of opinions, Twitter, software application.

"Not everything that counts can be counted and not everything that´s counted truly counts" (Edward Bruce Cameron, 1963)

"Stay hungry, stay foolish" (Steve Jobs, 2005)

Agradecimientos

Una vez llegados al punto en el cual, ves cómo se cierra una de las etapas más importantes de mi vida, son muchos los nombres que me vienen a la cabeza.

En primer lugar, quiero agradecer a mi familia y muy especial a mi madre, por la confianza depositada en mí y por ser un pilar importante, haciéndome ver que, todo esfuerzo realizado finalmente tendría su recompensa.

En segundo lugar, quiero agradecer a mi novia Paula, que después de estos 5 años de universidad, siempre ha estado apoyándome, y nadie más que ella ha sufrido en cada examen, cada nota final, cada entrega realizada…

El tercer lugar a mi hermano Israel, por haber sido siempre un espejo en el que mirarme.

Por último, me gustaría agradecer a todos los profesores que he tenido durante toda la carrera ya que de una u otra manera han conseguido trasmitir sus conocimientos y de todos ellos, aprendí algo.

Gracias a todos y cada una de las personas que han formado parte de mi vida y me han apoyado en todo momento.

1 Introducción

1.1 Contexto

Es de naturaleza humana (Naturaleza humana, Sin Fecha), desde el concepto filosófico, que los seres humanos, tiendan a compartir características distintivas e inherentes como pueden ser su manera de pensar, de sentir o de actuar.

Tal y cómo afirma Lozares-Colina en su teoría de las redes sociales (Lozares-Colina, 1996), esta manera de compartir que tienen los seres humanos se realiza en un espacio social o "*red social*", el cual, se compone por un conjunto de personas y el entorno en el que crean un espacio de relaciones.

"Las Redes Sociales pueden definirse como un conjunto bien delimitado de actores - individuos, grupos, organizaciones, comunidades, sociedades globales, etc.- vinculados unos a otros a través de una relación o un conjunto de relaciones sociales" (Lozares-Colina, 1996, pág. 108).

Sin embargo, el auge que ha experimentado Internet en las últimas décadas ha provocado que este espacio social y las relaciones directas que se conformaban haya sido transformado.

El número de personas que utilizan las redes sociales aumenta continuamente. Según el estudio Global Digital Overview (We are Social and Hootsuite´s, 2019), elaborado por la agencia creativa especializada en social media We are Social (We are social, 2019) y la plataforma de gestión de redes sociales Hootsuite (Hootsuite´s, Sin Fecha), el uso en redes sociales creció un 9% en el último año. En este mismo estudio se afirma que la penetración global alcanzó el 45% de la población mundial, el equivalente a 3.484 billones de personas.

Las redes sociales han cambiado nuestra forma de relacionarnos con el mundo, principalmente porque ha proporcionado un espacio en el que los usuarios pueden compartir y expresar libremente tanto sentimientos como opiniones sobre distintos temas.

Los usuarios disponen ahora de nuevos modos de comunicación. Este nuevo modo de comunicación tiene un impacto directo en las opiniones y en la forma en la que se trasmite la información y las noticias. Según el artículo publicado por Criado, M.A. (Criado, 2011), movimientos como el que se surgió en España el 15 de mayo, también conocido como 15M, no hubieran sido posibles si no hubieran existido las redes sociales, pues, el sentir de las emociones colectivas se cargó de emotividad a medida que se divulgaba.

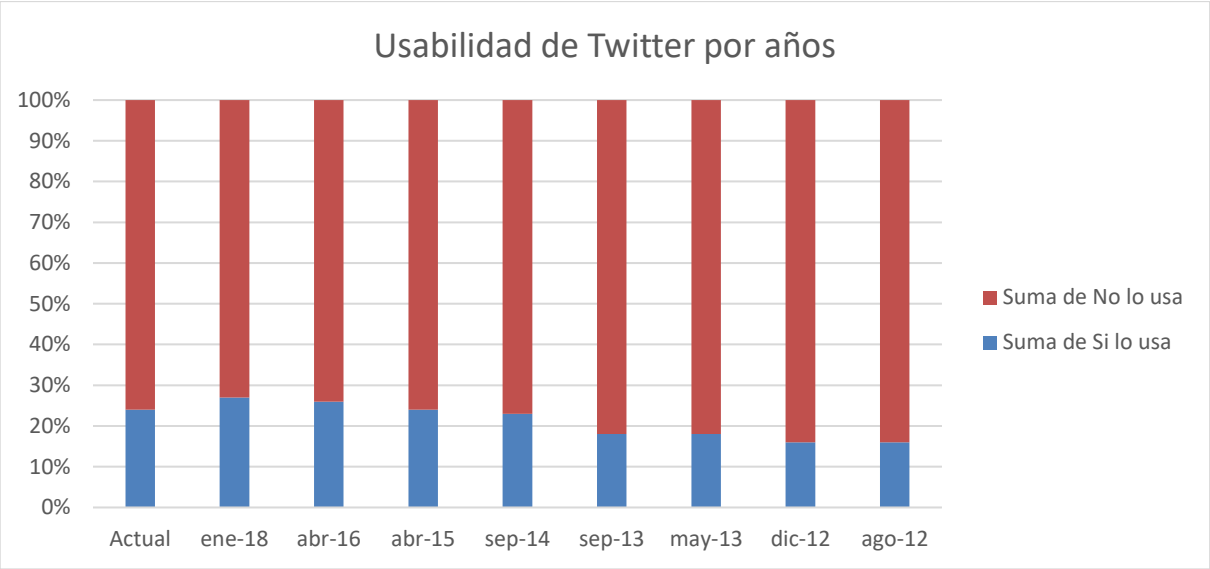
Los consumidores tienen ahora acceso a multitud de opiniones, muchas de las cuales tienen la capacidad de influenciar en las personas que lo leen (Lahuerta Otero & Cordero Gutiérrez, Sin Fecha). Esta capacidad de influenciar sobre los pensamientos, sentimientos y acciones de los usuarios ya fue estudiada por autores como Lozares-Colina (Lozares-Colina, 1996), y al que denominaron influencia social.

El usuario, por tanto, se convierte en un trasmisor de opinión y sentimiento. Los consumidores que toman decisiones sobre productos y marcas están recurriendo cada vez más a la comunicación por ordenador para obtener información en la que basar sus decisiones (Kozinets, 2002) y ese espacio de relaciones ha pasado a ser un espacio de relaciones virtual.

Uno de estos espacios virtuales que permite el intercambio de opiniones y sentimientos es la red social Twitter. Twitter (Twitter, Sin Fecha), está basada en el intercambio de mensajes de manera bidireccional por parte de los usuarios. Es un servicio que permite publicar, compartir e intercambiar cualquier información con un total de 280 caracteres.

Según el último artículo “Social Media Methodology” publicado por la empresa Pew Research Center (Pew Research Center, 2019), encargada de realizar estudios de investigación donde se brinda información sobre problemáticas, actitudes y tendencias en Estados Unidos, (Pew Research Center, 2019), el número de usuarios activos en la red social experimentó un fuerte crecimiento en sus primeros años manteniéndose estable en los tres últimos. El mismo artículo sitúa a Twitter entre las redes sociales más importantes en la actualidad, llegando a alcanzar un 24% de uso entre la población de los Estados Unidos en el año 2019.

En la gráfica 1 de elaboración propia a partir de los datos de (Pew Research Center, 2019), se puede observar el crecimiento en la usabilidad que ha experimentado Twitter entre los años 2012 y 2019.



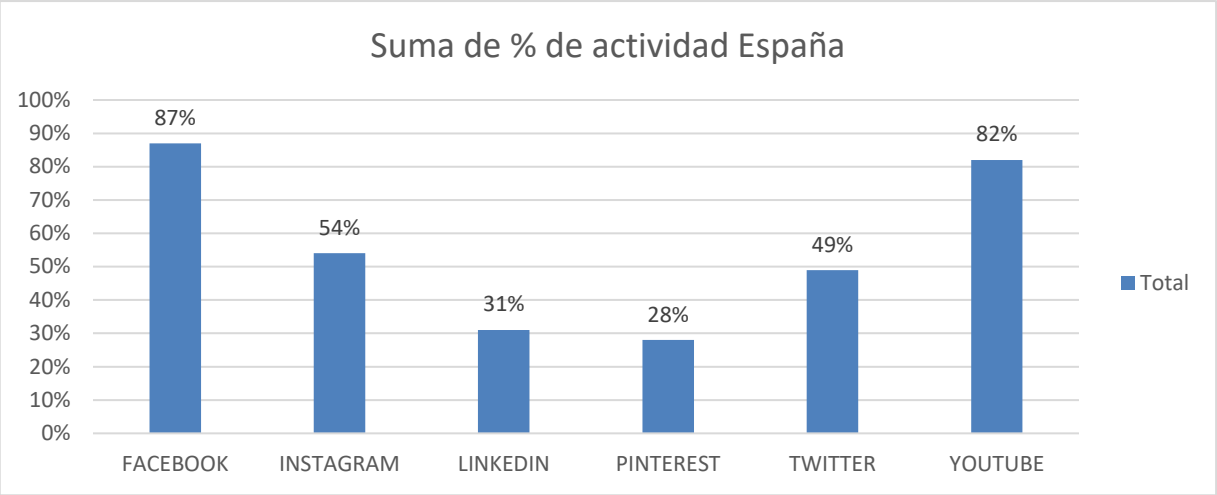
Gráfica 1. Uso de Twitter. Crecimiento anual

La red social Twitter (We are Social and Hootsuite´s, 2019), contaba a principios del año 2019 con aproximadamente 326 millones de usuarios, situándola entre las 20 redes sociales con más número de usuarios. En la gráfica 2 de elaboración propia a partir de los datos de (We are Social and Hootsuite´s, 2019), se detalla esta clasificación basado en usuarios mensuales activos, cuentas de usuarios o visitas únicas a cada plataforma.

Gráfica 2. Redes sociales. Cuentas activas

En España (We are Social and Hootsuite´s, 2019), se sitúa entre las redes sociales más activas. En concreto este estudio la ubica como la quinta red social más activa por detrás de Facebook e Instagram con alrededor de un 49% de aceptación en España tal y cómo se

observa en la gráfica 3 de elaboración propia a partir de datos obtenidos de (We are Social and Hootsuite´s, 2019)



Gráfica 3. Suma de actividad España

En la última década, los trabajos e investigaciones orientados a la minería de sentimientos no han dejado de aumentar, principalmente debido a que actualmente se permite el acceso a una gran cantidad de datos. Los dominios en los cuales, se ha realizado investigaciones orientados al análisis de sentimiento han sido muy variados como se describirá más adelante en el apartado sobre el estado del arte.

Por estas razones, y dado que es una plataforma que permite la expresión de opiniones a través de los mensajes, Twitter es la aplicación idónea para el desarrollo de un proyecto de investigación basado en el análisis de sentimiento.

1.2 Motivación

El grado en ingeniería informática te da un amplio abanico de posibilidades en el que poder enfocar y encarar el trabajo de investigación que permita finalizar los estudios superiores.

Se ha valorado algunas de estas posibilidades sin entrar en detalle de cómo de fácil podría llegar a encarar y enfocar el mismo, si no en aprovechar la oportunidad que se ofrece como autoaprendizaje y desarrollo personal.

La creciente evolución producida en los últimos años en todo lo relacionado con la analítica y transformación de grandes volúmenes de datos en información valiosa, y por tanto en conocimiento de negocio, se ha convertido en una auténtica obsesión de muchas

compañías actuales. Muchas de estas empresas están aprovechando los resultados para crear estrategias que permitan diferenciarse de sus competidores directos.

Según BLS, Bureau Labor Stadistics de EE. UU. (Bureau of Labor Statistics, U.S. Department of Labor, 2019), se espera que los puestos de trabajo relativos a Big Data y científico de datos aumenten un 19% entre los años 2016 y 2026. Esto hace que los empleos que se encuentren englobados en la transformación digital y la analítica de los datos se conviertan en las profesiones del futuro.

Tal y cómo se están desarrollando los acontecimientos, en los próximos años, los sistemas de inteligencia artificial, el procesamiento del lenguaje natural enfocados a robótica, el conocimiento en analítica de datos, formarán parte de la vida diaria de las personas.

Por este motivo, se ha considerado enfocar este trabajo de investigación en aquellas áreas cuyo modelo estuviera cercano a la analítica de datos y a los algoritmos enfocados a la inteligencia artificial, permitiendo tener un crecimiento personal y profesional, aprovechando todas las posibilidades que brinda el Big Data como son: tratamiento y manejo de grandes volúmenes de datos en sistemas no relacionales; tratamiento de datos no estructurados; diseño y analítica de datos mediante interfaces gráficas; conocimiento de las arquitecturas y paradigmas de Big Data; implementación y conocimiento de algoritmos; implementación de tecnologías que permitan resolver problemas de NLP; analítica predictiva; procesamiento en tiempo real; entre otros.

También se ha evaluado el aprendizaje de nuevos lenguajes y dado que, estaba más habituado con programación en .NET, usándose en entorno de desarrollo Visual Studio, se ha considerado realizar este trabajo de fin de grado desarrollando una aplicación software que tuviese una interfaz gráfica web en lenguaje Java.

1.3 Objetivos

En el desarrollo de este trabajo fin de grado, se pretende cubrir y alcanzar los objetivos que se detallan en este apartado.

1.3.1 Objetivo general

Gestión de análisis es un sistema de minería de texto basado en la red social Twitter y cuyo objetivo principal es la de analizar el sentimiento expresado a través de un mensaje o tuit.

Por medio de una aplicación se creará un análisis automatizado, basado en opiniones sobre un cierto tema o tendencia concreto, construyendo así un sistema que recoja los intereses de los usuarios y que pueda obtener datos valiosos sobre ellos para convertirlo en conocimiento.

Gestión de análisis permitirá visualizar a través de una interfaz moderna, las opiniones expresadas en un corpus extraído directamente desde la red social Twitter y estarán segmentadas en cinco grandes categorías: muy positiva, positiva, neutra, negativa y muy negativa.

Adicionalmente, se permitirá tener una geolocalización de los mensajes que dispongan de sus datos precisos de latitud y longitud, dando visibilidad completa en un mapa basado en Google Maps.

1.3.2 Objetivo específico

1. Realizar un estudio del arte que permita conocer los trabajos de investigación más importantes dentro del procesamiento del lenguaje natural y del paradigma del análisis de sentimiento en particular en redes sociales.
2. Implementar un sistema adaptado a las recomendaciones de usabilidad y accesibilidad web a través de interfaces amigables y modernas.
3. Ampliar conocimientos en lenguaje de programación Java y en la construcción de sistemas con interfaz web basados en plataforma Eclipse.
4. Utilizar API´s y librerías de terceros, que permitan la reutilización de software.
5. Usar herramienta de control de versiones para la gestión del código fuente y la gestión de la memoria.
6. Aplicar conocimientos adquiridos en el grado en ingeniería informática enfocados a aplicar metodologías ágiles dentro del marco Scrum para el desarrollo de software.
7. Familiarizarse con el uso de bases de datos no relacionales para la gestión y almacenamiento de información no estructurada.

1.4 Estructura del trabajo

En este apartado, se detalla y define cada uno de los puntos con los que cuenta este trabajo de fin de grado.

Capítulo 1. Introducción: En este primer capítulo de introducción, se asientan las bases del proyecto, se explica y justifica las razones por las cuales se ha decidido realizar este trabajo de fin de grado y se hace una visión general enfocado al trabajo de investigación.

Capítulo 2. Estado del arte: En este capítulo se realiza un estudio general de los trabajos de investigación que han sentado las bases de este trabajo, se detallan los avances realizados en los artículos y trabajos de investigación en el ámbito de estudio del trabajo fin de grado y se describen las ideas generales que harán posible el desarrollo del proyecto.

Capítulo 3. Propuesta software: En este capítulo se recogen los requisitos identificados en el desarrollo del software, se realiza un análisis general de la arquitectura desde los distintos puntos del proyecto, se realiza un plan general que detalle la metodología aplicada para la implementación del software desde su inicio hasta su finalización y, por último, se explican las herramientas utilizadas.

Capítulo 4. Conclusiones: En este capítulo se evalúa el cumplimiento de los requisitos recogidos en el apartado anterior y se exponen las conclusiones generales del trabajo fin de grado.

Capítulo 5. Futuros trabajos: En este capítulo se abordará desde la perspectiva general que detalle aquellas implementaciones que, por falta de tiempo no ha sido viable o no ha tenido cabida desarrollar en este trabajo de fin de grado y que pueden dar una mejora adicional a lo ya existente en el trabajo fin de grado.

2 Estado del arte

Parece interesante que al aproximarse al campo objeto de estudio y dado que es un campo desconocido por una parte de la sociedad, se haga una especial referencia que explique y anteponga en qué consiste este estudio de investigación y cuáles son las aportaciones dentro de la computación, de modo que dé respuestas y sentido a este trabajo de fin de grado.

2.1 Procesamiento del lenguaje natural

Dado que este estudio se centra en el análisis de sentimiento, el procesamiento del lenguaje natural también conocido como NLP toma especial protagonismo.

El procesamiento del lenguaje natural (Procesamiento del lenguaje natural, Sin Fecha) (NLP por sus siglas en inglés, Natural Language Processing), es un campo que se engloba dentro de la computación, la inteligencia artificial y la lingüística y que estudia las interacciones existentes entre los seres humanos y las computadoras a través del lenguaje natural.

Cabe destacar la complejidad del campo de estudio provocado por las propias dificultades existentes en todo lenguaje, a las que se debe añadir la falta de capacidad de las computadoras actuales en su identificación.

Estas dificultades se basan principalmente a propiedades inherentes a cualquier ser humano y que Noan Chomsky, 1998 identificó como recursión e infinitud discreta.

“La recursividad se considera un requisito mínimo y elemental de cualquier teoría sintáctica el formalizar la capacidad humana de generar infinitas oraciones a partir de un número limitado de elementos (Infinitud discreta).” (Chomsky, 2019)

La arquitectura NLP (Procesamiento del lenguaje natural, Sin Fecha), se encuentra englobada en cuatro niveles. Estos niveles parten del análisis de la oración hasta su comprensión. Los cuatro niveles que se proponen son:

- Análisis morfológico: se realiza un análisis de las palabras para extraer raíces, rasgos flexivos, sufijos, prefijos y otros elementos con el objetivo de comprender la manera en la cual se construyen y forman las palabras.
- Análisis sintáctico: se realiza un análisis de la estructura de las oraciones con el objetivo de comprender las reglas con la que se construyen y combinan las palabras.

- Análisis semántico: se extrae el significado de la frase con el objetivo de resolver ambigüedades tanto léxicas como estructurales.
- Análisis pragmático: se realiza un análisis del texto más allá de los límites de la frase, por ejemplo, para determinar los antecedentes referenciales de los pronombres.

Las áreas en las que se encuentran todas las líneas de investigación del procesamiento del lenguaje natural nos muestran un panorama con un crecimiento sorprendente en campos cómo sistemas de búsquedas, reconocimiento del habla, traducción automática o análisis de sentimientos, entre otros.

2.2 Breve historia

El análisis de sentimiento (Análisis de Sentimiento, Sin Fecha), también conocido como minería de opiniones o análisis de opinión, se enmarca en el dominio del procesamiento del lenguaje natural, la analítica de textos y la lingüística computacional, y es el estudio que pretende determinar el tono o sentimiento de un conjunto de palabras, permitiendo una clasificación de un texto escrito en un lenguaje natural.

La primera definición formal que se encuentra en un artículo sobre análisis de sentimiento y minería de opinión fue realizada por (Dave, Lawrence, & Pennock, 2003) en la conferencia celebrada en Budapest, "*Proceedings of the 12th international conference on World Wide Web*":

"Conjunto de resultados de búsqueda para un elemento determinado, generando una lista de atributos del producto y agregando opiniones."

Otros investigadores enfocados a la analítica de sentimientos como Pang y Lee en (Pang & Lee, 2008) lo definen como:

"Tratamiento computacional de opiniones, sentimientos y subjetividad en textos"

La tarea principal en la minería de opinión es la de establecer una clasificación en un texto que indique si tiene un sentimiento positivo o negativo con más o menos intensidad de manera que podamos entender la intención del usuario permitiendo identificar y clasificar la

opinión a la que se está refiriendo. A esta clasificación se la conoce como intensidad de la polaridad.

A pesar de ser un campo relativamente moderno, existen diversos estudios orientados a la minería de opinión y análisis de sentimiento. La mayoría de ellos han abordado el problema siguiendo principalmente dos líneas de investigación: clasificador basado en aprendizaje automático y clasificador basado en reglas.

- **Clasificador basado en aprendizaje automático**: La primera línea de estudio se basa en la utilización de un corpus representativo y en el entrenamiento de los distintos algoritmos de ingeniería artificial, tanto automáticos como supervisados, para clasificar.

  La mayoría de los estudios de aprendizaje automático están basados en tres tipos de algoritmos: clasificador de Naïve Bayes, clasificador de entropía máxima y máquinas vector de soporte.

  Es el caso de las publicaciones realizadas por (Pang, Lee, & Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, 2002) y (Salvetti, Lewis, & Reichenbach, 2006) para identificar la polaridad en las críticas de cine, en los que se expone cómo los algoritmos son entrenados mediante un conjunto de datos utilizando unigramas (vector de palabras con N=1 tokens independientes) y bigramas (vector de palabras con N=2 tokens consecutivos, es decir, conjunto de términos con hasta dos palabras consecutivas) y llegando a la conclusión de que los distintos clasificadores usados de aprendizaje automático mejoran cuando se usan características binarias que indican la presencia de unigramas en el texto, en lugar de una característica numérica que indica el número de apariciones o frecuencia.

  El principal problema en el uso de este tipo de clasificador es que se basan en su comportamiento estadístico, es decir, tienen una fuerte dependencia en el tamaño y en la perfección del conjunto de datos de entrenamiento.

- **Clasificador basado en reglas**: La segunda de estas líneas de estudio está basada en abordar el problema a través de la orientación semántica de las palabras y en base a esto determinar el sentimiento completo de la frase. Este clasificador se basa en la utilización de lexicones: conjuntos de listas de palabras clasificadas con un valor de sentimiento. En este tipo de clasificador

se centraron los estudios de los investigadores (Taboada, Brooke, Tofiloski, & Voll, 2011).

El principal problema en el uso de este tipo de clasificador es que no alcanzan a detectar la semántica asociada al lenguaje natural.

2.3 Primeros estudios

Para el desarrollo de este trabajo de fin de grado, se realizó previamente un trabajo de búsqueda basado en el estado del arte, de manera que se pudiera obtener todas las técnicas y métodos que se usan en la actualidad para identificar los sentimientos basados en el léxico y en el lenguaje natural.

En base a este trabajo previo de investigación, se encuentran los primeros estudios relativos a la utilización de los datos para clasificar las opiniones en positivos o negativos a principios del siglo XXI, donde alcanzó su mayor popularidad, motivado principalmente a dos causas: la primera razón fue el avance en los sistemas de aprendizaje automático; la segunda razón fue la facilidad para obtener un corpus de datos de entrenamiento para entrenar los algoritmos.

Prueba de esta popularidad la encontramos en los trabajos realizados por Peter Turney (Turney, 2002), quien desarrolló un algoritmo de clasificación de opiniones con orientación semántica positiva (pulgares hacia arriba) u orientación semántica negativa (pulgares hacia abajo) y cuya precisión de promedio alcanzó el 74% en las opiniones analizadas sobre los dominios de automóviles, bancos, películas y destinos de viaje.

También es de gran interés las aportaciones que realizaron Bo Pang, (Pang, Lee, & Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, 2002) en el área de la minería de la información y la analítica de los textos, permitiendo tratar la subjetividad, el sentimiento y las opiniones elaboradas sobre películas. Para ello, experimentaron con tres algoritmos distintos que utilizaban técnicas principalmente de machine learning: clasificación de Naive Bayes, clasificación de entropía máxima (MaxEnt o ME) y máquinas de vectores de soporte (SVN´s), probando diferentes enfoques de n-gramas y concluyendo que la incorporación de la frecuencia de n-gramas emparejadas podría ser una característica que podría disminuir la precisión y que estos algoritmos desempeñaron una mayor precisión utilizando unigramas que alcanzaron un 82.9% de promedio en el algoritmo SVN´s.

Sin embargo, la mayoría de estos primeros artículos publicados se centraban en la captación del sentimiento a partir de la identificación de la polaridad positiva o negativa en el texto, siguiendo la conceptualización en dos dimensiones que propone Russell (Russell, 1980), quién afirma que existe sólo dos emociones básicas, positivas y negativas, tal y cómo se muestra en la ilustración 1, sin tener en cuenta la polaridad neutra.

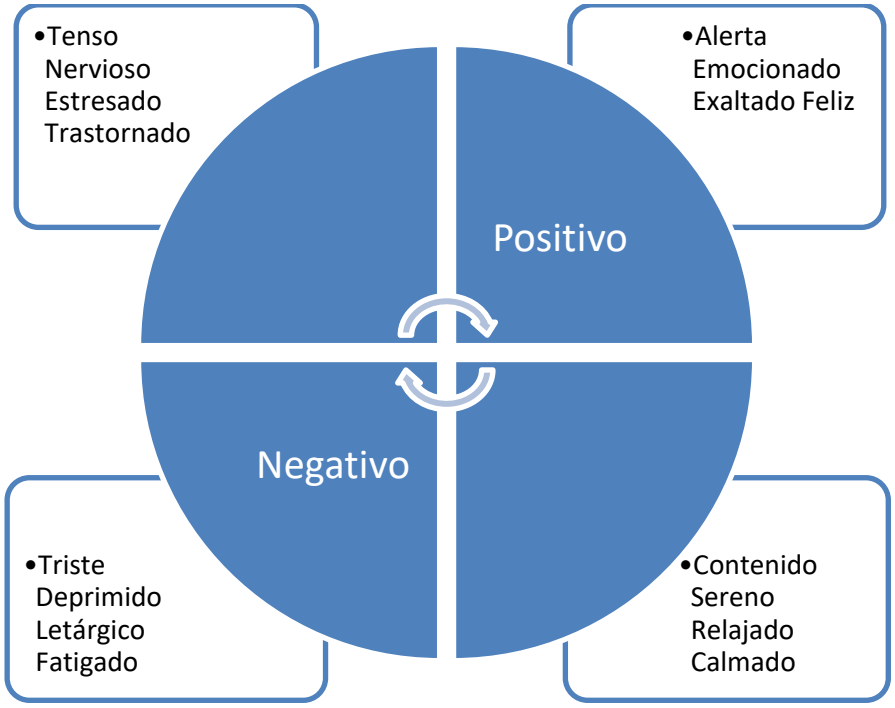


Ilustración 1. Dimensiones de emociones propuesto por Russell

Fuente de elaboración propia a partir de los datos de (Russell, 1980)

Sin embargo, Koppel (Koppel & Schler, 2006) defiende la necesidad de tener una clasificación en tres clases que permita obviar los ejemplos imparciales en el proceso de aprendizaje. De este modo, se consigue una mejor distinción binaria de los documentos positivos y negativos.

2.4 Análisis del sentimiento en redes sociales

El análisis de sentimiento a partir de un conjunto de datos obtenido directamente desde redes sociales no llegó hasta el año 2010 cuando los investigadores Alexander Pak & Patrick Paroubek (Pak & Paroubek, 2010), utilizaron un corpus obtenido directamente desde Twitter para realizar minería de opinión y análisis de sentimiento por medio del clasificador de Navie

Bayes. El clasificador permitía detectar los sentimientos positivos, negativos y neutrales de los tuits.

Desde entonces, la facilidad para obtener un conjunto de datos de entrenamiento para entrenar los distintos algoritmos ha permitido el crecimiento de artículos basados en minería de opinión. Prueba de ello es el aumento en el número de publicaciones sobre el estado del arte obteniendo un conjunto de datos directamente de Internet en muchos y diversos dominios para tratar de conocer la opinión de las personas.

Cabe destacar, por ejemplo, los artículos basados en la captación de la opinión de los votantes para realizar perfiles sociopolíticos que permitieran conocer el comportamiento de las personas en un sistema electoral.

Muchos de los analistas y periodistas políticos han consideran clave el uso de la tecnología y las redes sociales a la hora de atraer votos, especialmente de gente joven. Un ejemplo de esto es el que manifiesta la periodista Sarah Lai Stirland, 2008 (Stirland, 2008), en su artículo "*Las armas secretas de Obama*" en el que revela cómo Barack Obama fue el primer candidato en implantar con éxito técnicas tecnológicas en su campaña electoral, opinión que también es compartida por el periodista Harfoush (Harfoush, 2009), en su libro "Yes We Did! An inside look at how social media built the Obama brand", en el que expone con más detalle como Barak Obama llegó a construir una marca alrededor de sí mismo para su beneficio político en las elecciones generales a la presidencia de los Estados Unidos.

Los primeros trabajos sobre análisis de sentimiento en el campo de la política estuvieron muy presentes en las publicaciones que realizaron Murphy Choy; Michelle L.F. Cheong; Ma Nang Laik; Koo Ping Shung (Choy, Cheong, Laik, & Shung, 2011), y en el que intentaron predecir la magnitud e intención de voto que obtendría cada candidato a la presidencia de Singapur. Para la realización de este ranking, emplearon un corpus con alrededor de 16.500 tuits. Aunque sus estudios no pudieron acertar esta clasificación, si sentaron unas sólidas bases para futuros artículos.

A partir de este momento, el crecimiento en investigaciones en el dominio de la política no paró de sucederse. Es el caso de la publicación que Hao Wang; Dogan Can; Abe Kazemzadeh; François Bar; Shrikanth Narayanan, (Wang, Dogan Can, & François Bar, 2012) realizaron para intentar predecir el número de votos que obtendría cada candidato a la presidencia de Estados Unidos. Para ello, abordaron distintas técnicas que permitían tener en cuenta dos problemas presentes en la captación del análisis de sentimiento en los textos escritos:

1. El primer problema es que el lenguaje utilizado en un tuit difiere mucho del lenguaje natural de las personas.

2. El segundo problema es que los tuits generalmente suelen ser bastante sarcásticos lo que presenta un desafío adicional a los modelos de clasificación.

Sobre esta última premisa se publicaron distintos artículos de relevancia donde destacar el trabajo de los investigadores Roberto González-Ibáñez; Smaranda Muresan; Nina Wacholder, (González-Ibáñez, Muresan, & Wacholder, 2011) en el que exponen esta problemática a través de su artículo *"Identifying Sarcasm in Twitter: A Closer Look"*. En este artículo describen cómo el sarcasmo puede transformar completamente la polaridad de un tuit y concluyendo que existe una gran dificultad para detectar el sarcasmo en un computador. A esta problemática se le añade otras como puede ser la longitud del mensaje o tuit, cuanto mayor es la longitud, más complejo se hace identificarlo.

También son interesantes, los trabajos de investigación realizados en nuestro país en este dominio en particular por los investigadores Vilares Calvo, D., Thelwall, M., & Alonso, M.A., 2015 (Vilares, Thelwall, & Alonso, 2015), donde se estudian las opiniones realizadas en Twitter, monitoreando la política y sociedad española. El trabajo hace distintas contribuciones en este ámbito en las que cabe destacar: el sistema de detección del sentimiento para un tuit político español de manera automatizada; la manera en la cual la negación o los modismos en los caracteres pueden influir en la precisión y en la detección del sentimiento dentro del lenguaje español; clasificación de las opiniones sobre líderes políticos, partidos y personalidades clave. Por tanto, no tratan de obtener un ranking único en un sistema electoral como en otros estudios, sino que el objetivo principalmente se centra en disponer una clasificación de popularidad e intención de voto.

2.5 El poder del análisis del sentimiento en la actualidad

Pero ¿cuál es el motivo por el que existe tanto interés en conocer el sentimiento de los usuarios? ¿Qué beneficios se puede obtener en conocer la opinión de las personas? Estas preguntas se responden con los movimientos surgidos a través de las redes sociales y cuya finalidad está basada en el cambio social.

Los movimientos populares aparecidos recientemente han cambiado la visión que tienen tanto organizaciones como partidos políticos, principalmente por la rápida difusión que tienen y por la carga emocional que desprenden.

Es el caso del movimiento denominado “Primavera Árabe”, movimiento que se organizó en el año 2010. Tal y como se explica en el artículo de Reporteros sin Fronteras (Reporteros sin Fronteras, Sin Fecha), los usuarios usaron las redes sociales como instrumento de movilización y de información, precipitando la caída de dictadores. Fueron revoluciones humanas, principalmente impulsadas a través de Internet y las redes sociales donde se acuñaron los términos de “revolución Twitter” o “revolución Facebook”.

Otros movimientos populares como el acontecido en España el 15 de mayo del 2011, también conocido como 15M, tuvieron un impacto directo en la sociedad (Criado, 2011), y confirmaron que fue un movimiento social espontáneo organizado a través de las redes sociales.

Los primeros artículos de investigación orientados a obtener el sentimiento creado en el movimiento 15M a través de la red de microblogging Twitter fueron el que lideraron los investigadores (Alvarez, Garcia, Moreno, & Schweitzer, 2015). En este artículo detallan como analizaron un corpus con alrededor de 560.000 tuits durante la fase de creación del 15M, publicados por cerca de 85.000 usuarios. El 45% de estos mensajes tenían un sentimiento positivo mientras que el 22% desprendían un sentimiento negativo. Los resultados obtenidos en el estudio verificaron que la difusión de los mensajes con sentimiento negativo fue mucho mayor y que existieron grandes consecuencias derivadas de la propagación viral, pues, en la difusión, los mensajes se cargaron de emotividad.

Por estas razones los partidos políticos han centrado sus esfuerzos en la utilización de Twitter para conocer el pensamiento de los usuarios, creando perfiles sociopolíticos. La clasificación de perfiles sociopolíticos ha tenido su máxima tensión en España a mediados del año 2018 (Varela, 2019), cuando los partidos introdujeron en la ley electoral un artículo en el cuál, daban la posibilidad de recopilar datos personales relativos a opiniones realizadas en redes sociales. De esta manera, daban total libertad a los partidos políticos en crear perfiles ideológicos de las personas, para así, centrar su propaganda y poder realizar actividades políticas durante el periodo electoral. Esto creó mucha controversia e indignación al considerarse una intromisión en la intimidad de los ciudadanos. Finalmente, el tribunal constitucional consideró inconstitucional el artículo en abril del año 2019 al considerar que el mismo no ofrece suficiente protección a los ciudadanos.

Este temor por parte de la sociedad española y su posterior inconstitucionalidad del artículo viene precedido del escándalo vivido con la empresa Cambridge Analytica. Cambridge Analytica (Cambridge Analytica, Sin Fecha), fue una compañía privada que combinaba la minería de datos y el análisis de datos con la comunicación estratégica en un proceso electoral y que, en el año 2018, tuvo que cerrar definitivamente al verse involucrada en prácticas

engañosas para conseguir influenciar en la decisión de las personas, principalmente en opiniones políticas.

Sin embargo, los artículos de investigación que centran su interés en este dominio en particular y en otros dominios en general siguen creciendo a un ritmo vertiginoso.

A este interés, se le añade otros como son la necesidad de disponer de una analítica de datos en tiempo real y la disposición de un corpus que ha pasado de unos cientos de miles de datos a miles de millones de registros. Esto hace inevitable realizar implementaciones basado en un entorno capaz de analizar y gestionar grandes volúmenes de información de distintas fuentes de manera óptima y rápida.

Por esta razón se están desarrollando sistemas en tiempo real que permitan el manejo y tratamiento de esta gran cantidad de información, basados en implementaciones Big Data. A pesar de que el término Big Data está muy extendido en general, no existe una definición exacta y única del término. No obstante, el término Big Data, suele aplicarse en la comunidad científica a todo lo referido con: manejo de grandes volúmenes de datos; variedad en los datos; velocidad en la generación y el procesamiento de los datos; veracidad en los datos. Esto es también conocido como las cuatro V (volumen, variedad, velocidad y veracidad).

Las primeras implementaciones basadas en entornos de Big Data junto con procesamiento del lenguaje natural NLP estuvieron lideradas por los investigadores Erturk, Emre y Hong Shi, (Erturk & Shi, 2016), donde explican la manera de implementar un framework basado en el ecosistema de Hadoop llamado KOSHIK, permitiendo el procesamiento, tratamiento y consulta de una gran cantidad de documentos en lenguaje natural no estructurados.

En los últimos años, se han dado grandes pasos dentro de la analítica de opinión, principalmente debido al avance en los algoritmos de aprendizaje y al acceso a una gran cantidad de fuentes públicas que permiten entrenar a estos algoritmos. Estos pasos han conseguido un porcentaje de éxito bastante aceptables.

Sin embargo, están apareciendo nuevos retos a los que se enfrenta la comunidad científica en el tratamiento de opiniones. Estos nuevos retos, están basados en el tratamiento de imágenes obtenidas desde fuentes de información públicas para clasificar la polaridad de la imagen.

Uno de los primeros artículos de investigación en esta área es "*Emotion Detection and Sentiment Analysis of Images*" realizado por los investigadores Gajarla y Gupta, (Gajarla & Gupta, 2017). En este artículo se detalla la manera en la que, utilizando redes sociales como

Flickr, y a través de SVM´s, clasificaron imágenes en 5 categorías: Amor, Felicidad, Violencia, Miedo y Tristeza. En este artículo obtuvieron una precisión del 67,8, concluyendo que existe diversos problemas para clasificar entre las emociones positivas (Amor y Felicidad) y negativas (Violencia, Miedo y Tristeza), posiblemente debido a que la red está aprendiendo la tonalidad o los colores en las imágenes: los sentimientos positivos generalmente tienen imágenes de colores brillantes y los negativos corresponden a imágenes oscuras.

En España, SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural, Sin Fecha) es la sociedad española que realiza investigaciones en el área del procesamiento del lenguaje natural. Fue fundada en el año 1984 permitiendo crear una comunidad sin ánimo de lucro que realiza investigaciones y publicaciones en el tratamiento automático del lenguaje natural, especialmente con el uso del español como lengua. La agencia realiza dos artículos de investigación que son publicados cada 6 meses.

2.6 Herramientas búsqueda análisis sentimiento

Las herramientas software que ofrece el mercado para ayudar a los usuarios, empresas y gestores de redes sociales en el análisis de sentimiento son varias. Cabe destacar las siguientes:

1. Hootsuite Insights (Hootsuite Insights, Sin Fecha), es una plataforma que permite medir el sentimiento social alrededor de una marca, filtrando los resultados por localización, idioma y género, para una visión multidimensional de los segmentos de mercado.

2. Brandwatch Analytics (Brandwatch Analytics, Sin Fecha), es una herramienta muy útil, que permite monitorizar las redes sociales permitiendo conocer la opinión de los usuarios sobre ciertos temas, facilitando el entendimiento de las actitudes del consumidor y reaccionar adecuadamente

3. Semantria (Semantria, Sin Fecha), es un complemento que funciona en Excel creada por la empresa Lexalytics. Semantria ofrece análisis de texto a través de API y complemento en Excel.

4. Rapidminer (RapidMiner, Sin Fecha), es software especializado que permita trabajar los datos. Es un sistema Open Source para Data Mining. Permite aplicar distintos algoritmos a los datos y realizar una búsqueda de sentimientos en redes sociales.

5. General Architecture for Text Engineering por sus siglas en inglés GATE (General Architecture for Text Engineering, Sin Fecha) es una suite de herramientas de código abierto y gratuito, publicado con licencia LGPL desarrollada por la universidad de Sheffield, Inglaterra. Gate está escrita en código Java y permite realizar tareas para el procesamiento, análisis y extracción del lenguaje natural.

1. Waikato Environment for Knowledge Analysis por sus siglas en inglés Weka (Weka, Sin Fecha) es un software libre distribuido bajo la licencia GNU-GPL. Weka está escrito en código Java y permite el procesamiento de algoritmos de aprendizaje automático y minería de datos.

3 Propuesta Software

3.1 Visión general

En este apartado se realizará una visión general de la propuesta software donde se da un enfoque de alto nivel del desarrollo a implementar en el trabajo de fin de grado. Esta perspectiva, permitirá establecer una discusión abierta sobre los conceptos más importantes que admita asentar las bases de la implementación software.

Los datos serán extraídos de la red social Twitter a través de una librería de código abierto desarrollada en Java para tal efecto, facilitando la extracción de la información de manera limpia y, en caso necesario, facilitando una transformación de los datos para ser adaptados a las necesidades del desarrollo.

Los datos, se categorizarán realizando minería de opinión a través de la librería de código abierto desarrollada para tal efecto llamada Stanford CoreNLP, determinando así la polaridad de los sentimientos del tuit en cinco categorías, que han sido definidas como: muy positivo, positivo, neutro, negativo y muy negativo.

El gestor de base de datos relacional almacenará los datos de acceso a la plataforma desarrollada, tales como usuarios, contraseña y perfiles, estará gestionada en Microsoft SQL Server 2016.

La propuesta del desarrollo del software mantiene una interfaz general basada en lenguaje Java. La interfaz gráfica, dispondrá de una visualización de gráficos utilizando la biblioteca escrita en lenguaje JavaScript llamada D3.js. Además, la geolocalización de los tuits será visible desde la biblioteca llamada Google Maps de código abierto, desarrollada por Alphabet Inc, en lenguaje JavaScript.

La accesibilidad a los datos ubicados en el gestor de base de datos relacional SQL Server, se realizará por medio del controlador JDBC 6.0 de Microsoft para SQL Server, biblioteca escrita en lenguaje Java para facilitar el acceso a la base de datos.

Adicionalmente, en cada búsqueda, se almacenarán los datos de Twitter, así como datos complementarios como la polaridad del tuit ya normalizado, en un sistema basado en Hadoop para futuros análisis.

3.2 Requisitos del sistema

Durante la fase de especificación de requisitos, se ha tenido en cuenta las necesidades con el propósito de definir de manera clara y sencilla el conjunto de funcionalidades y restricciones que contendrá el sistema que se pretende desarrollar.

3.2.1 Alcance del proyecto

El presente trabajo fin de grado tiene como finalidad el análisis, diseño, implementación, pruebas y validación de un sistema basado en tecnología web, que permita la extracción de mensajes o tuits escritos por usuarios en la red social Twitter. Una vez obtenidos, se analiza la información subjetiva que denotan y se clasifica en un "sentimiento" que puede ser, muy positivo, positivo, neutro, negativo y muy negativo.

La aplicación software será un sistema que permitirá la visualización de los datos a través de una interfaz gráfica intuitiva y amigable.

3.2.2 Requisitos funcionales

Sistema

RF-1	• El sistema será un desarrollo software con interfaz web que permitirá extraer la información de Twitter y analizar sus datos.
RF-2	• El sistema será un desarrollo software que permitirá analizar la información subjetiva para representar el sentimiento que desprende un tuit.

Inicio de sesión

RF-3	• Para acceder a la aplicación software, se deberá usar los mismos datos de usuario y contraseña definidos en la base de datos
RF-4	• Cada usuario debe haberse logado previamente en el sistema con su usuario y contraseña para poder acceder e interactuar con el sistema.

Extracción de tuits

RF-5	• El sistema, tendrá un buscador donde los usuarios podrán realizar búsquedas sobre Twitter
RF-6	• El sistema realizará una extracción de los datos de Twitter del texto escrito en el buscador.

3.2.3 Requisitos no funcionales

Documentación

RNF-1	• La aplicación debe poder disponer de un manual de usuario para su usabilidad.
RNF-2	• La aplicación debe disponer de los suficientes comentarios en el código para que su mantenimiento futuro sea lo más rápido y fácil posible.

Compatibilidad

RNF-3	• La aplicación debe poder ser ejecutada en los navegadores comunes (Internet Explorer, Safari, Chrome).
RNF-4	• La aplicación debe poder ser funcional en cualquier entorno de red.

Seguridad

RNF-5	• En la aplicación únicamente tendrán acceso usuarios autentificados.

Usabilidad

RNF-6	• La aplicación debe poder mostrar los datos establecidos en los requisitos funcionales para los cuales ha sido diseñada.
RNF-7	• La aplicación debe respetar los requisitos de usabilidad y responsividad.

Rendimiento

RNF-8	• La aplicación debe ser ágil y disponer de un rendimiento óptimo para su usabilidad.
RNF-9	• La aplicación requerirá un tiempo de respuesta óptimo y rápido, que no debe superar los 10 segundos en la autenticación y los 20 segundos en la búsqueda de mensajes en Twitter.

Estabilidad

RNF-10	• La aplicación debe poder gestionar correctamente a los fallos.

API Externas

RNF-11	• Para obtener la información de la geolocalización de los tuits se utilizará la biblioteca de Google Maps.
RNF-12	• Para obtener la información de los sentimientos de los tuits se utilizará la biblioteca Stanford CoreNLP
RNF-13	• Para obtener la información visualización de los gráficos de los tuits se utilizará la biblioteca D3.JS
RNF-14	• Para obtener la información de la base de datos relacional SQL Server se utilizará la API Microsoft JDBC 6.0

Lenguaje

RNF-15	• Se utilizará para el desarrollo de la aplicación software lenguaje de programación JAVA. Las tecnologías utilizadas se detallan en el apartado de arquitectura en tecnologías utilizadas.
RNF-16	• Se utilizará para la extracción de datos de la base de datos relacional, el lenguaje de programación SQL.

3.3 Metodología

3.3.1 Introducción

Para el desarrollo del proyecto, se utilizará una metodología que permita el desarrollo del software basado en un trabajo incremental e iterativo cuya base fundamental se asiente en el uso de un conjunto de principios ágiles que permita cubrir las necesidades de una manera rápida y sin necesidad de utilizar mucho tiempo en los requisitos del sistema, pero que permita cubrir todos los requisitos funcionales y no funcionales que se pretende.

"Los principios ágiles ponen el énfasis en construir software que funcione y que se pueda usar rápidamente, en vez de pasarse mucho tiempo al principio escribiendo especificaciones." (Deemer, Benefield, Larman, & Vodde, Sin Fecha, pág. 4)

La elección de la metodología está basada en el paradigma Scrum. Scrum (Scrum (desarrollo de software), 2019) es un marco de trabajo que define un conjunto de prácticas y roles, y que puede tomarse como punto de partida para definir el proceso de desarrollo que se ejecutará durante un proyecto.

"Scrum es un marco de trabajo iterativo e incremental para el desarrollo de proyectos, productos y aplicaciones. Estructura el desarrollo en ciclos de trabajo llamados Sprint. Son iteraciones de 1 a 4 semanas, y se van sucediendo una detrás de otra." (Deemer, Benefield, Larman, & Vodde, Sin Fecha, pág. 5)

La razón consistente para la utilización de Scrum en este desarrollo es, que este marco de trabajo permite durante el desarrollo e implementación del software, aprovechar un conjunto de buenas prácticas englobadas dentro de una estructura definida al que denominamos "ciclo de vida del proyecto". Este conjunto de buenas prácticas dará como resultado la finalización satisfactoria del proyecto, cubriendo por tanto todos los objetivos planteados al inicio, e implementando todas las necesidades detectadas previamente durante la fase de ingeniería de requisitos. El ciclo de vida de un proyecto basado en Scrum es el siguiente:

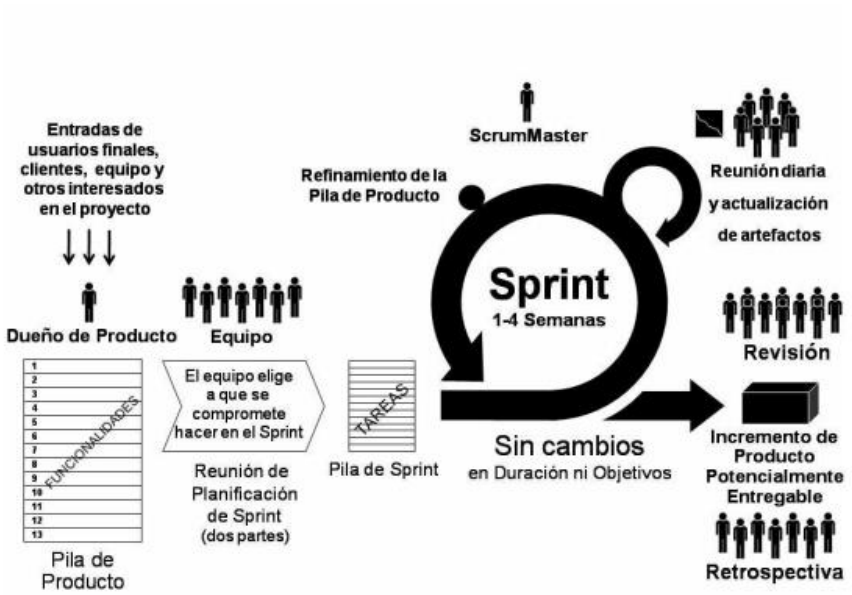


Ilustración 2. Scrum - Ciclo de vida completo

Fuente obtenida de (Deemer, Benefield, Larman, & Vodde, Sin Fecha)

El equipo Scrum, está formado por tres roles distintos, cada uno con su propia funcionalidad ya definida y el conjunto de todos los roles formara el equipo Scrum. Estos roles son:

- **Product Owner**: Es el responsable de maximizar el valor del producto y del trabajo del equipo de desarrollo. Es el único que gestiona la lista de productos también conocido como Product Backlog. (Schwaber & Sutherland, 2013)
- **Scrum Máster**: Es la persona responsable de asegurar que el Scrum es entendido y adoptado por todos los miembros del equipo Scrum. (Schwaber & Sutherland, 2013). Es el evangelizador de Scrum.
- **Development team**: El equipo de desarrollo está formado por un conjunto de profesionales que desempeñan el trabajo de entregar un incremento de un producto "Terminado" y que tras su finalización se pueda poner en producción al final de cada Sprint. (Schwaber & Sutherland, 2013)

El ciclo de vida de un proyecto con metodología Scrum se caracteriza por ciclos temporales cortos.

"Un Sprint es el corazón de Scrum. Es un bloque de tiempo de entre 3-4 semanas durante el cual se crea un incremento de producto" (Schwaber & Sutherland, 2013)

Un sprint es una iteración definida por una fecha de inicio y una fecha de finalización. Se inicia con la reunión de planificación del Sprint (Sprint Planning Meeting) que es un trabajo colaborativo del equipo Scrum al completo, donde se detalla el trabajo a realizar durante la fase de este. Le siguen los Scrum diarios (Daily Scrum), una reunión corta de en torno a 15 minutos para que el equipo de desarrollo sincronice sus actividades y marque un plan de ejecución paras las próximas 24 horas. Continúa con el trabajo de desarrollo en el cual el equipo de desarrollo implementa las funcionalidades definidas en el sprint. Finalmente acaba con la revisión del Sprint (Sprint Review) en el que se inspecciona el sprint y se adapta la lista de productos si fuese necesario con nuevos requisitos y con la retrospectiva del sprint (Sprint Retrospective) en el que se inspecciona el equipo Scrum a sí mismo para crear un plan de mejoras y poder abordarlas en el siguiente sprint. (Schwaber & Sutherland, 2013)

La lista de productos o Product Backlog especifica el conjunto de características, funcionalidades, requisitos, mejoras y correcciones identificadas a ser implementados en el desarrollo del proyecto durante las entregas futuras. Es una lista viva, en la cual, durante la

fase de refinamiento (Refinement) de la lista de productos pueden añadirse nuevas funcionalidades. (Schwaber & Sutherland, 2013).

La lista de pendientes de Sprint o Sprint Backlog es el conjunto de elementos seleccionados de la lista de productos para ser implementados durante el Sprint. Además, llevará consigo un plan de entrega del producto para conseguir la finalización satisfactoria del mismo.

3.3.2 Roles

En la tabla 1, se muestra cada uno de los roles que estarán presentes en el desarrollo del software al completo.

Tabla 1. Roles Scrum - Proyecto TFG

Rol	Persona
Product Owner (Cliente)	David Montoya Ruiz
Scrum Master	David Montoya Ruiz
Development Team	David Montoya Ruiz

3.3.3 Product Backlog

La siguiente tabla, muestra el PB (Product Backlog) del proyecto de desarrollo software con interfaz web. En el PB se identifica la lista de requisitos generales conocido como PBI o Product Backlog Items y la planificación estratégica.

Tabla 2. Product Backlog - Proyecto TFG

Identificador (ID) de la Historia	Enunciado de la Historia	Estado	Dimensión / Esfuerzo	Iteración (Sprint)	Prioridad	Comentarios
HU-0000-0001	Como usuario, quiero que el sistema disponga de un portal de acceso mediante usuario y contraseña, con la finalidad de mantener la seguridad y confidencialidad de la información.		24 horas (3 jornadas)	Sprint 2	3	
HU-0000-0002	Como usuario necesito que el diseño del portal web sea guiado teniendo en cuenta requisitos de usabilidad y responsividad, de manera que su utilización resulte sencilla e intuitiva para mis usuarios		40 horas (5 jornadas)	Sprint 1	2	
HU-0000-0003	Como usuario del sistema, quiero que el sistema pueda ser ejecutado en cualquiera de los navegadores comunes (Internet Explorer, Google Chrome, Safari), con la finalidad de tener un portal accesible desde cualquier lugar sin limitación alguna.		8 horas (1 jornada)	Sprint 1	1	
HU-0000-0004	Como usuario, necesito que la información de los usuarios, contraseñas, perfiles… esté recogida en un repositorio con la finalidad de tener un acceso único al sistema		8 horas (1 jornadas)	Sprint 2	4	
HU-0000-0005	Como usuario, quiero poder consultar los datos de Twitter a través de los indicadores previamente definidos, con la finalidad de tener una visión completa que permita tener una mejora en la toma de decisiones empresariales		24 horas (3 jornadas)	Sprint 5	7	
HU-0000-0006	Como usuario, quiero disponer un buscador que extraiga la información de Twitter y así poder ser analizado.		24 horas (3 jornadas)	Sprint 3	5	
HU-0000-0007	Como usuario, necesito poder clasificar la información por sentimiento. La clasificación estará basada en muy positivo, positivo, neutro, negativo y muy negativo		24 horas (3 jornadas)	Sprint 4	6	
HU-0000-0008	Como usuario, necesito poder visualizar la información en un mapa con la geolocalización de cada tuit siempre que este disponga de la información de latitud y longitud		16 horas (2 jornadas)	Sprint 5	8	

3.3.4 Sprint Backlog

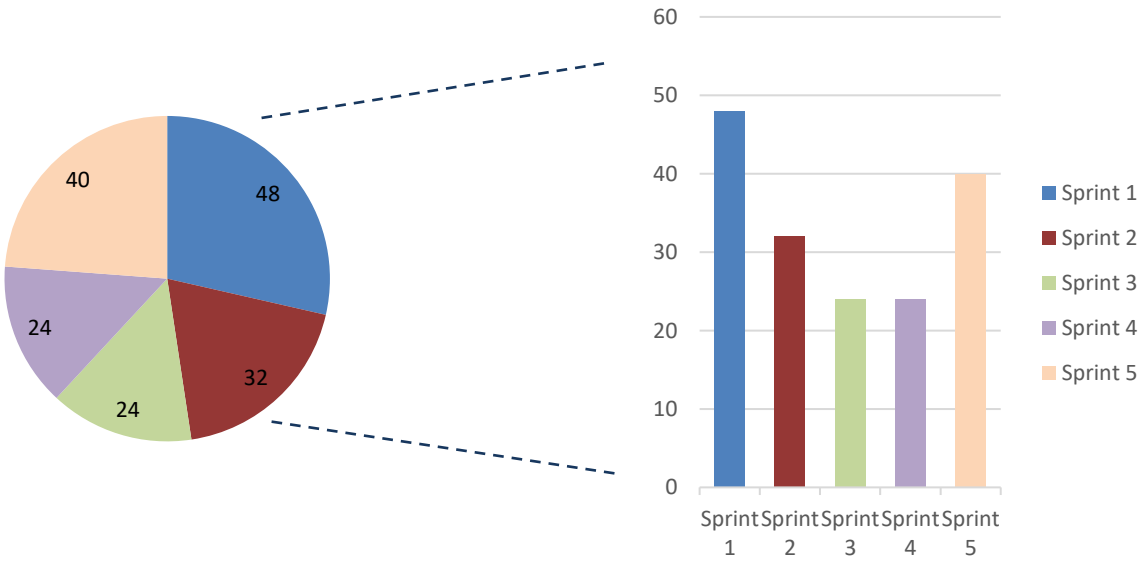
El desarrollo de la aplicación está compuesto por un total de 5 “Sprint” cuya suma de cada uno de ellos, nos mostrará la duración total del proyecto. El tiempo estimado para el desarrollo y consecución del proyecto será de 168 horas. El desglose con la fecha de inicio y fin de cada iteración se mostrará en el apartado plan de trabajo.

En la tabla 3, se detalla el tiempo total de dedicación estimada de cada uno de los sprint que tendrá el desarrollo.

Tabla 3. Número total de horas por Sprint

Sprint	Horas
Sprint 1	48
Sprint 2	32
Sprint 3	24
Sprint 4	24
Sprint 5	40
Total	**168**

Tal y cómo se observa en la gráfica 1 se detalla la distribución del tiempo por cada uno de los sprint.



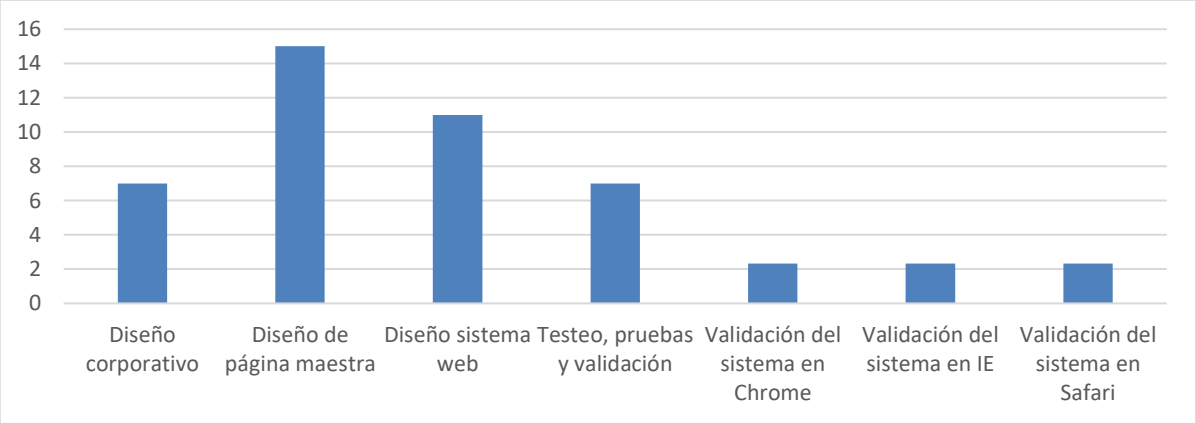
Gráfica 4. Tiempo por Sprint

3.3.4.1 Backlog del primer Sprint

El primer sprint tendrá una duración de 48 horas (6 jornadas de trabajo) estimadas.

El Backlog del primer sprint se desglosa en las siguientes tareas cuyas funcionalidades permitirán el diseño de una aplicación web y la accesibilidad a la misma desde los distintos navegadores convencionales. Funcionalidades por implementar:

La gráfica 5 de elaboración propia, contiene el tiempo de dedicación en cada una de las funcionalidades.



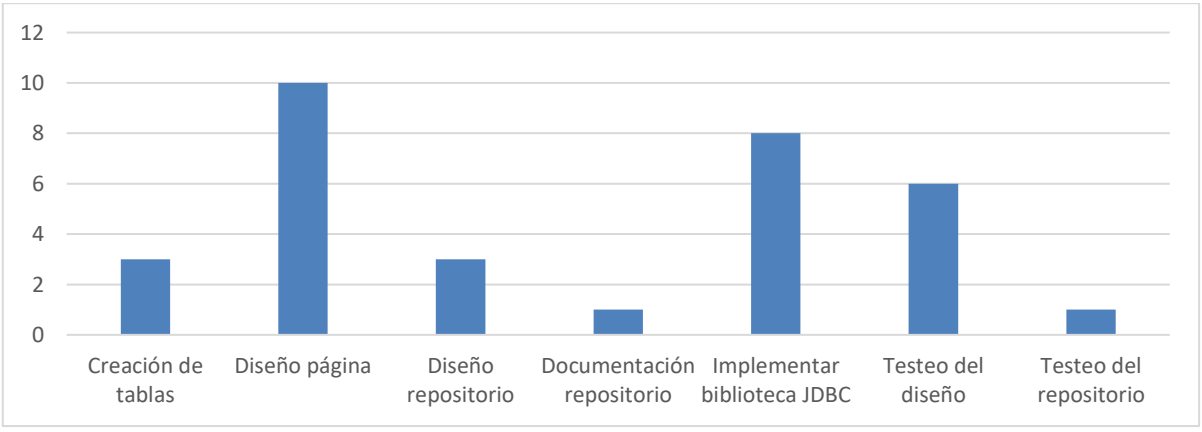
Gráfica 5. Sprint 1 - Horas dedicadas por funcionalidades implementadas

3.3.4.2 Backlog del segundo Sprint

El segundo sprint tendrá una duración de 32 horas (4 jornadas de trabajo) estimadas.

El Backlog del segundo sprint se desglosa en las siguientes tareas cuyas funcionalidades permitirán el acceso al sistema y la creación de un repositorio donde se almacene la información de los usuarios. Funcionalidades por implementar:

La gráfica 6 de elaboración propia, contiene el tiempo de dedicación en cada una de las funcionalidades.

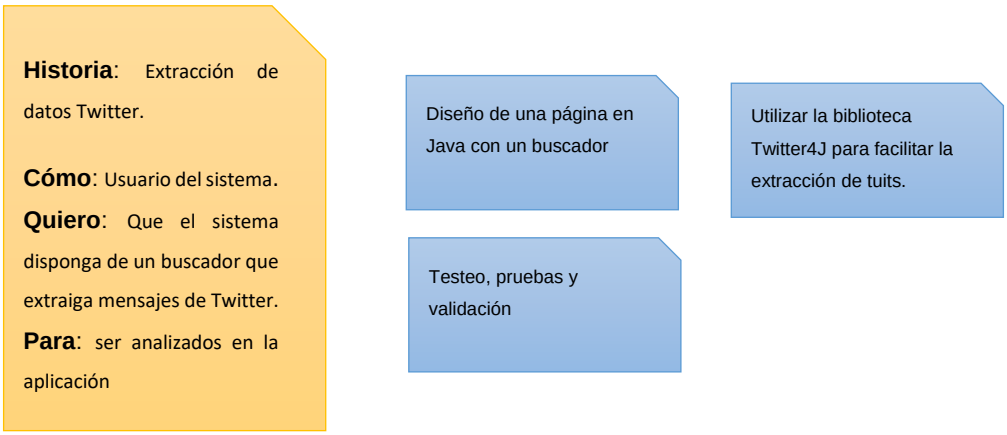


Gráfica 6. Sprint 2 - Horas dedicadas por funcionalidades implementadas

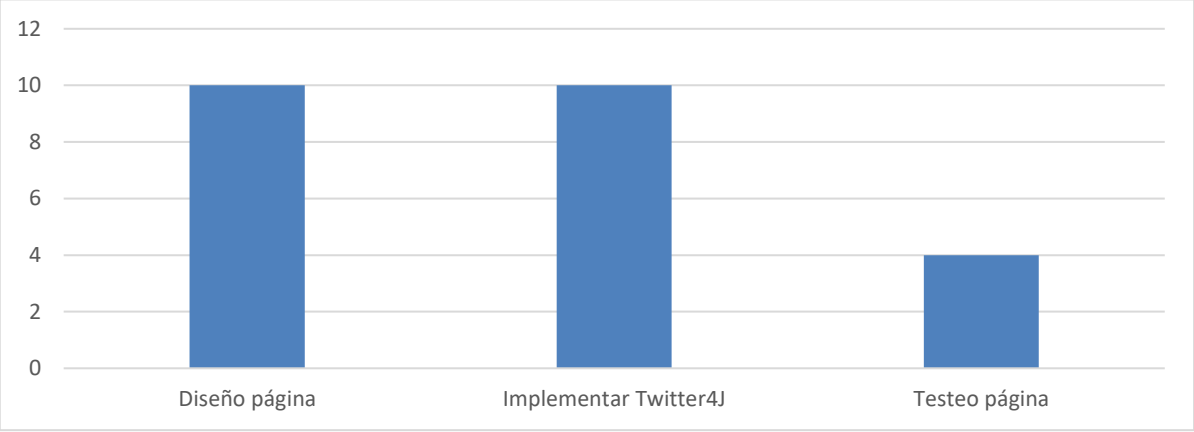
3.3.4.3 Backlog del tercer Sprint

El tercer sprint tendrá una duración de 24 horas (3 jornadas de trabajo) estimadas.

El Backlog del tercer sprint se desglosa en las siguientes tareas cuya funcionalidad permitirá la extracción de datos de Twitter. Funcionalidades por implementar:



La gráfica 7 de elaboración propia, contiene el tiempo de dedicación en cada una de las funcionalidades.



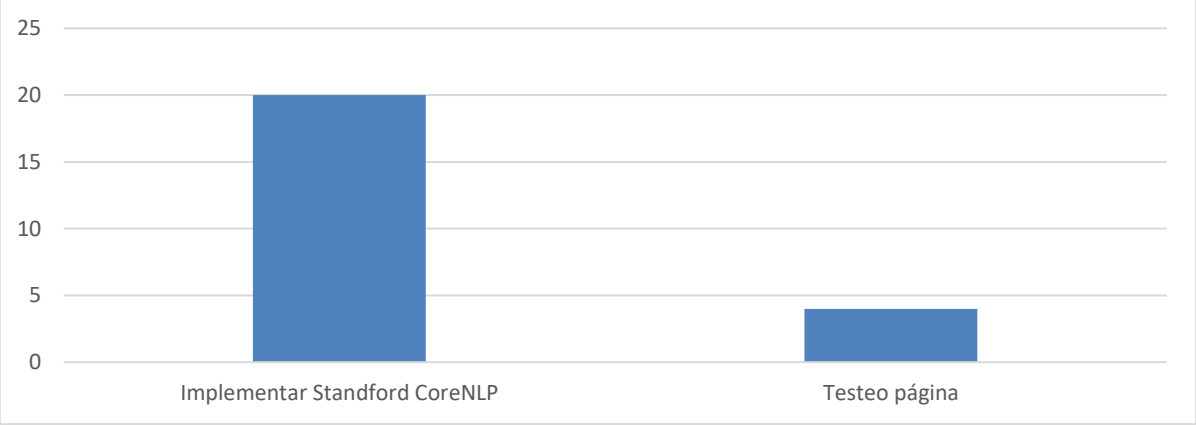
Gráfica 7. Sprint 3 - Horas dedicadas por funcionalidades implementadas

3.3.4.4 Backlog del cuarto Sprint

El cuarto sprint tendrá una duración de 24 horas (3 jornadas de trabajo) estimadas. El Backlog del cuarto sprint se desglosa en las siguientes tareas cuya funcionalidad permite la clasificación del sentimiento de un tuit.

Historia: Análisis de sentimiento **Cómo**: Usuario del sistema. **Quiero**: Que el sistema clasifique la información por sentimientos **Para**: ser analizados en la aplicación	Utilizar la biblioteca Stanford CoreNLP para clasificar el sentimiento.	Testeo, pruebas y validación

La gráfica 8 de elaboración propia, contiene el tiempo de dedicación en cada una de las funcionalidades.



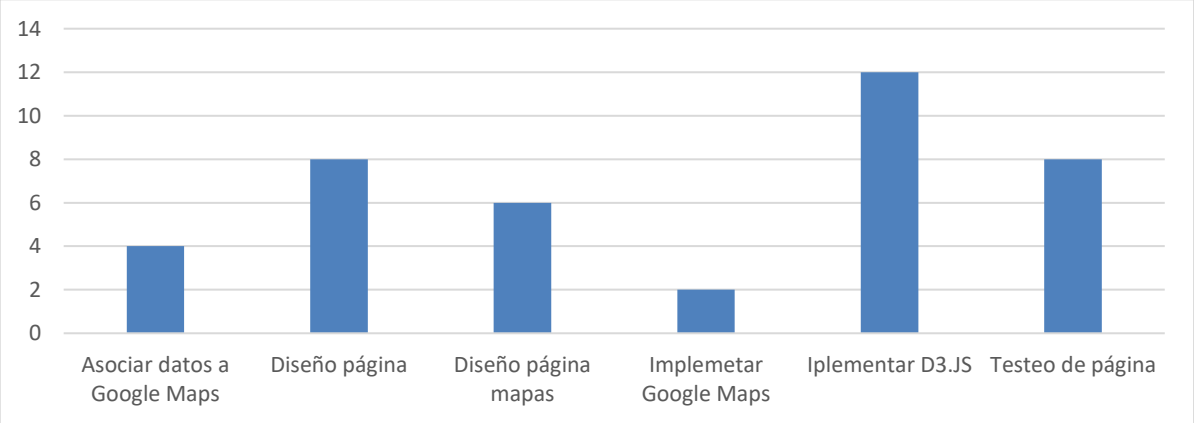
Gráfica 8. Sprint 4 - Horas dedicadas por funcionalidades implementadas

3.3.4.5 Backlog del quinto Sprint

El quinto sprint tendrá una duración de 40 horas (5 jornadas de trabajo) estimadas. El Backlog del quinto sprint se desglosa en las siguientes tareas cuyas funcionalidades permitirán el diseño de la interfaz gráfica de la aplicación.

Historia: Diseño de indicadores **Cómo**: Usuario del sistema. **Quiero**: Consultar los datos de Twitter a través de los indicadores previamente definidos **Para**: Tener visión completa que permita tener una mejora en la toma de decisiones empresariales	Diseño de página web para la visualización de los indicadores.	Utilizar la biblioteca D3.JS para la visualización de datos.	Testeo, pruebas y validación
Historia: Visualización de mapa **Cómo**: Usuario del sistema. **Quiero**: visualizar la información en un mapa. **Para**: Tener la geolocalización de cada tuit	Diseño de página web para la visualización de mapas. Asociar los datos de tuits de latitud y longitud en el mapa.	Utilizar la biblioteca Google Maps para la visualización de datos.	Testeo, pruebas y validación

La gráfica 9 de elaboración propia, contiene el tiempo de dedicación en cada una de las funcionalidades.



Gráfica 9. Sprint 5 - Horas dedicadas por funcionalidades implementadas

3.3.5 Plan de trabajo

En el siguiente apartado, se realiza un desglose del plan de trabajo detallando cada uno de los Sprint identificados.

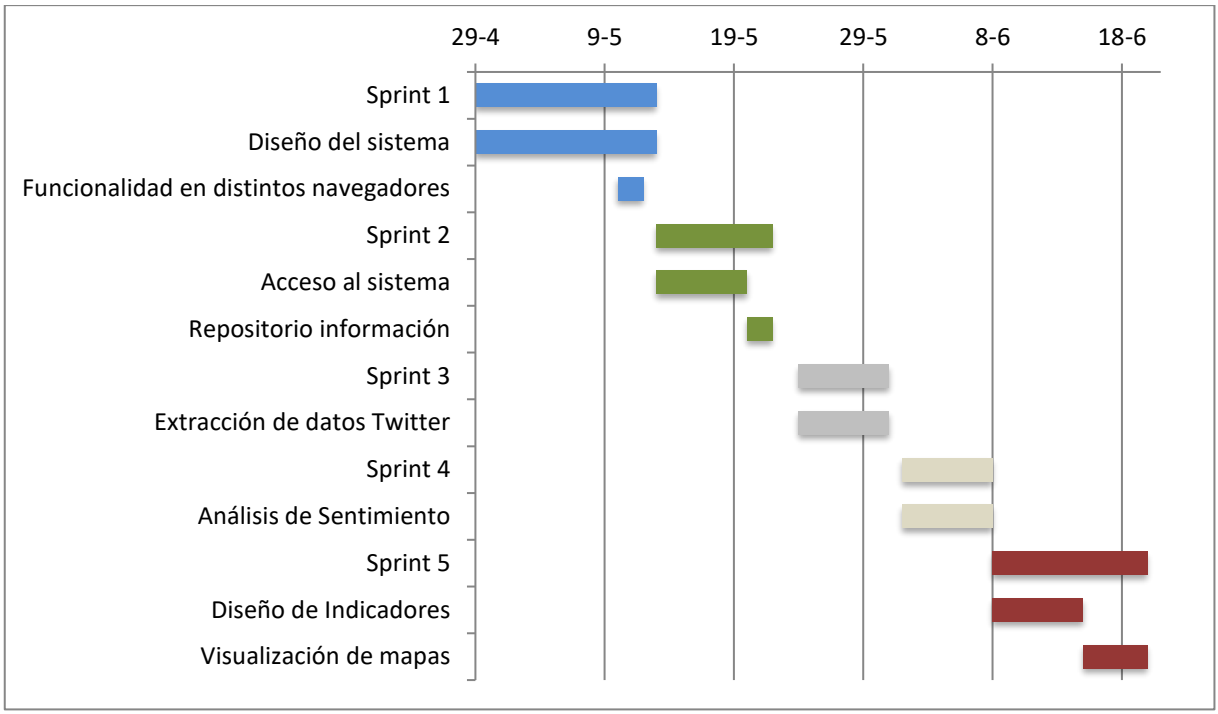
La suma del total de dedicación en la implementación de cada una de las iteraciones que conforman el desarrollo de todas las funcionalidades y restricciones del sistema será de 168 horas.

El tiempo de dedicación aproximado por el equipo de desarrollo será de 20 horas semanales. A continuación, se detalla en la tabla 4, el plan de trabajo para la consecución del trabajo fin de grado en el que se muestra las fechas de inicio y fin, así como el estado de cada uno de los Sprint.

Tabla 4. Plan de trabajo - Scrum

Nombre de la tarea	Responsable	Fecha de inicio	Fecha final	Días	Horas	Estado
Sprint 1	DMR	2019-04-29	2019-05-12	14	48	Finalizado
Diseño del sistema	DMR	2019-04-29	2019-05-12	14	40	Finalizado
Funcionalidad en distintos navegadores	DMR	2019-05-10	2019-05-11	2	8	Finalizado
Sprint 2	DMR	2019-05-13	2019-05-21	9	32	Finalizado
Acceso al sistema	DMR	2019-05-13	2019-05-19	7	24	Finalizado
Repositorio información	DMR	2019-05-20	2019-05-21	2	8	Finalizado
Sprint 3	DMR	2019-05-24	2019-05-30	7	24	Finalizado
Extracción de datos Twitter	DMR	2019-05-24	2019-05-30	7	24	Finalizado
Sprint 4	DMR	2019-06-01	2019-06-07	7	24	Finalizado
Análisis de Sentimiento	DMR	2019-06-01	2019-06-07	7	24	Finalizado
Sprint 5	DMR	2019-06-08	2019-06-19	12	40	Finalizado
Diseño de Indicadores	DMR	2019-06-08	2019-06-14	7	24	Finalizado
Visualización de mapas	DMR	2019-06-15	2019-06-19	5	16	Finalizado

La gráfica 10 de elaboración propia, se detalla el tiempo de dedicación previsto en cada una de las tareas a realizar, derivando en el plan de trabajo total o gráfico Gantt.



Gráfica 10. Gráfico Gantt - Plan de trabajo

3.3.6 Plan de sprint

3.3.6.1 Iteración primer sprint board

En el Sprint 1, se implementará las primeras funcionalidades enfocadas a tener una aplicación robusta cumpliendo los estándares desde el 29 de abril del 2019 hasta el 12 de mayo del 2019.

Backlog ítems comprometidos	Tareas no iniciadas	Tareas en progreso	Tareas Completadas
Historia: Funcionalidad en distintos navegadores **ID: HU-0000-0003**			Validación del sistema en Safari Validación del sistema en Google Chrome Validación del sistema en Internet Explorer
Historia: Diseño sistema. **ID: HU-0000-0002**			Diseño sistema web programado mediante Diseño del corporativo, elección de colores, Diseño de la página maestra. Testeo, pruebas y validación.

3.3.6.2 Iteración segundo sprint board

En el Sprint 2, se implementará las funcionalidades enfocadas al acceso al sistema y a la creación de un repositorio con los usuarios y claves de acceso. Será llevado a cabo desde el 13 de mayo del 2019 hasta el 21 de mayo del 2019.

Backlog ítems comprometidos	Tareas no iniciadas	Tareas en progreso	Tareas Completadas
Historia: Acceso al sistema. **ID: HU-0000-0001** **Historia**: Repositorio información. **ID: HU-0000-0004**			Diseño de la página de validación de usuario y Utilizar la biblioteca JDBC 6.0 Testeo, pruebas y validación Diseño de un repositorio Creación de las tablas maestras de usuario en Documentación del repositorio. Testeo, pruebas y validación

3.3.6.3 Iteración tercer sprint board

En el Sprint 3, se implementará las funcionalidades enfocadas a la extracción de la información de Twitter y a la implementación de una página con un buscador. Será llevado a cabo desde el 24 de mayo del 2019 hasta el 30 de mayo del 2019.

Backlog ítems comprometidos	Tareas no iniciadas	Tareas en progreso	Tareas Completadas
Historia: Extracción de datos Twitter. **ID: HU-0000-0006**			Diseño de una página en Java con un buscador Utilizar la biblioteca Twitter4J para facilitar la Testeo, pruebas y validación

3.3.6.4 Iteración cuarto sprint board

En el Sprint 4, se implementará las funcionalidades enfocadas a la implementación del análisis de sentimiento. Será llevado a cabo desde el 1 de junio del 2019 hasta el 7 de junio del 2019.

Backlog ítems comprometidos	Tareas no iniciadas	Tareas en progreso	Tareas Completadas
Historia: Análisis de sentimiento **ID: HU-0000-0007**			Utilizar la biblioteca Stanford CoreNLP para clasificar el sentimiento. Testeo, pruebas y validación

3.3.6.5 Iteración quinto sprint board

En el Sprint 5, se implementará las funcionalidades enfocadas a la implementación del análisis de sentimiento. Será llevado a cabo desde el 8 de junio del 2019 hasta el 19 de junio del 2019.

Backlog ítems comprometidos	Tareas no iniciadas	Tareas en progreso	Tareas Completadas
Historia: Diseño de indicadores **ID: HU-0000-0005** **Historia**: Visualización de mapa **ID: HU-0000-0008**			Diseño de página web para la visualización de Diseño de página web para la visualización de Utilizar la biblioteca Google Maps para la Asociar los datos de tuits de latitud y longitud en el Utilizar la biblioteca D3.JS para la visualización de Testeo, pruebas y validación Testeo, pruebas y validación

3.4 Arquitectura software

3.4.1 Introducción

En el siguiente apartado, se explicará la arquitectura software empleada para el desarrollo de la aplicación “Gestión del análisis”.

El objetivo principal será mantener una especificación de la arquitectura del sistema, así como un documento actualizado que mantenga organizado todos los componentes internos y externos que serán necesarios, sirviendo de guía para el entendimiento de la arquitectura y para el buen desarrollo de la aplicación software con interfaz web. Ian Sommerville, 2005, define la arquitectura como:

“El diseño arquitectónico es un proceso creativo en el cual se diseña una organización del sistema que cubrirá los requerimientos funcionales y no funcionales de éste”. (Sommerville, 2011, pág. 151)

3.4.2 Arquitectura general del sistema

Uno de los aspectos más importantes en un desarrollo es llegar a convertir y plasmar los requerimientos técnicos en una arquitectura software y es por eso por el cual, debe existir una cohesión entre el proceso de ingeniería de requerimientos y el proceso de ingeniería de arquitectura, tal y cómo afirma Jan Bosch.

“La actividad más compleja durante el desarrollo de aplicaciones es la transformación de una especificación de requisitos en una arquitectura del sistema.” (Bosch, 2000, pág. 24)

Siguiendo estas premisas, el proceso de arquitectura se define como un proceso creativo para detallar los requisitos funcionales y no funcionales. La arquitectura seleccionada para el desarrollo de la aplicación consta de distintas capas que se explicarán con mayor profundidad en este apartado permitiendo visualizar, modelar y documentar el sistema.

La arquitectura general del sistema está basada en una arquitectura web cuya principal ventaja es que es una arquitectura de sistemas distribuidos lo que facilita el acceso de múltiples usuarios a través de una interfaz gráfica amigable e intuitiva basado en el patrón arquitectónico cliente-servidor.

“El modelo arquitectónico cliente-servidor es un modelo de sistema en el que dicho sistema se organiza como un conjunto de servicios y servidores asociados, más unos clientes que acceden y usan estos servicios, (...), y cuya principal ventaja es que es una arquitectura distribuida. Se puede hacer un uso efectivo de los sistemas en red.” (Sommerville, 2011, págs. 226-227).

Dentro del patrón de arquitectura cliente-servidor, el desarrollo del software propuesto se fundamenta en una arquitectura “n-tier” en el cual, cada componente se encuentra englobado dentro de una capa o tier.

Concretamente está basada en una programación en tres capas. Según la Wikipedia (Programación por capas, Sin Fecha), la principal ventaja es la posibilidad de que el desarrollo pueda llevarse a cabo en distintos niveles, de modo que cada capa se abstrae del resto de módulos, permitiendo una arquitectura completamente escalable, que facilita el desarrollo y el futuro mantenimiento del sistema.

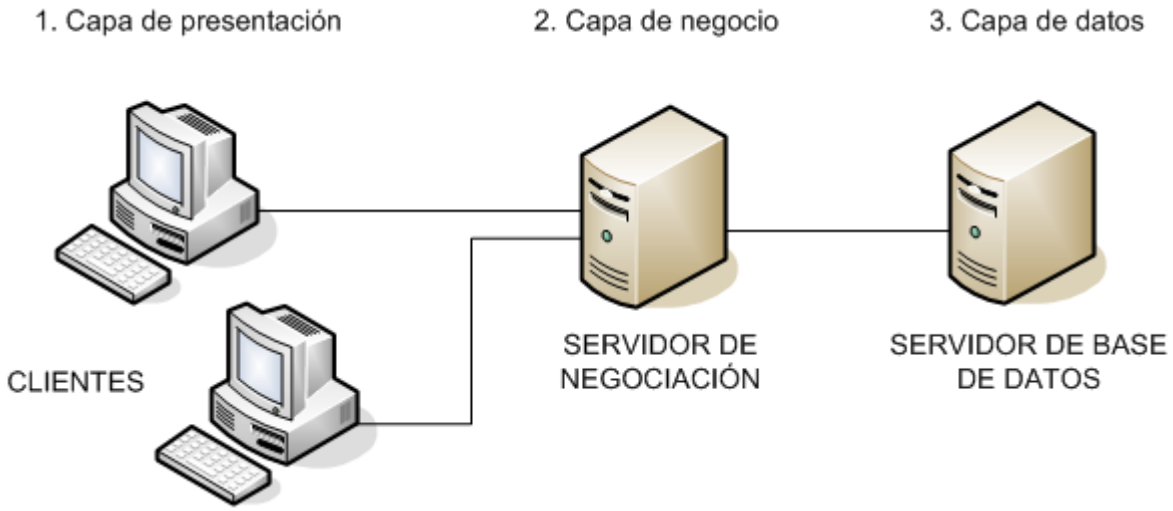


Ilustración 3. Programación por capa - 3 liers

Fuente obtenida de (Programación por capas, Sin Fecha)

Para comprender la arquitectura que mantiene el sistema, es necesario que se conozca los módulos con los que cuenta el software. Cada uno de estos módulos cuenta con su propia funcionalidad completamente independiente.

- Módulo menú: En este módulo se encuentra información sobre el funcionamiento de la aplicación, así como las funcionalidades implementadas y el objetivo del desarrollo.
- Módulo análisis: El módulo análisis, permite realizar una búsqueda de los datos de Twitter y extraer un corpus de los tuits encontrados. Se realizará un análisis de sentimiento del corpus. En base a esto, el módulo análisis, a través de su interfaz gráfica mostrará una serie de indicadores definidos.

- Módulo grafo: Este módulo dibujará un grafo de palabas para que se visualice las iteraciones de cada una de ellas.
- Módulo mapa: En el módulo mapas, se mostrará un mapa con la geolocalización de cada tuit.

Se podrá navegar desde todos los módulos de manera que sea intuitivo para el usuario. Los módulos de la aplicación quedarán de la siguiente manera.

Ilustración 4. Módulos de la aplicación

Fuente de elaboración propia.

El desarrollo de la arquitectura del sistema estará basado en el modelo de vistas de arquitectura 4+1, Philippe Kruchten (Modelo de Vistas de Arquitectura 4+1, Sin Fecha), basada en 4 + 1 modelo arquitectónico. Estos son vista lógica, vista de desarrollo, vista de proceso, vista física y escenarios.

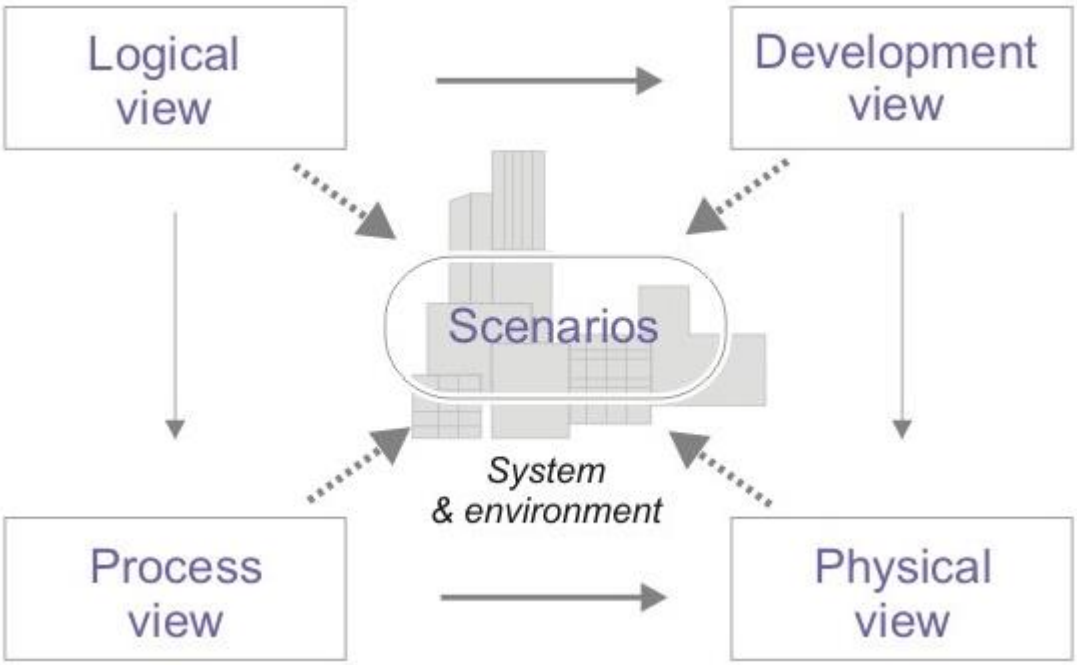


Ilustración 5. Modelo de vista en arquitectura 4+1

Fuente obtenida de (Modelo de Vistas de Arquitectura 4+1, Sin Fecha)

3.4.3 Vista Lógica

La vista lógica (Modelo de Vistas de Arquitectura 4+1, Sin Fecha) describe la funcionalidad y la estructura que mantiene el sistema. La vista lógica se compone de tres diagramas distintos y tiene una gran importancia para la comprensión del sistema por parte de los usuarios finales. Los diagramas en los que se componen son:

1. Diagrama de secuencia
2. Diagrama de clases
3. Diagrama de comunicación

3.4.3.1 Diagrama de secuencia

En el siguiente apartado se realizará un diagrama de secuencias de la aplicación software. El diagrama de secuencia se engloba dentro de la vista lógica. Cada diagrama de secuencia (Diagrama de secuencia, Sin Fecha) mostrará la iteración que existe entre los objetos y el sistema.

1. El diagrama de secuencias para el acceso a la aplicación software será el siguiente

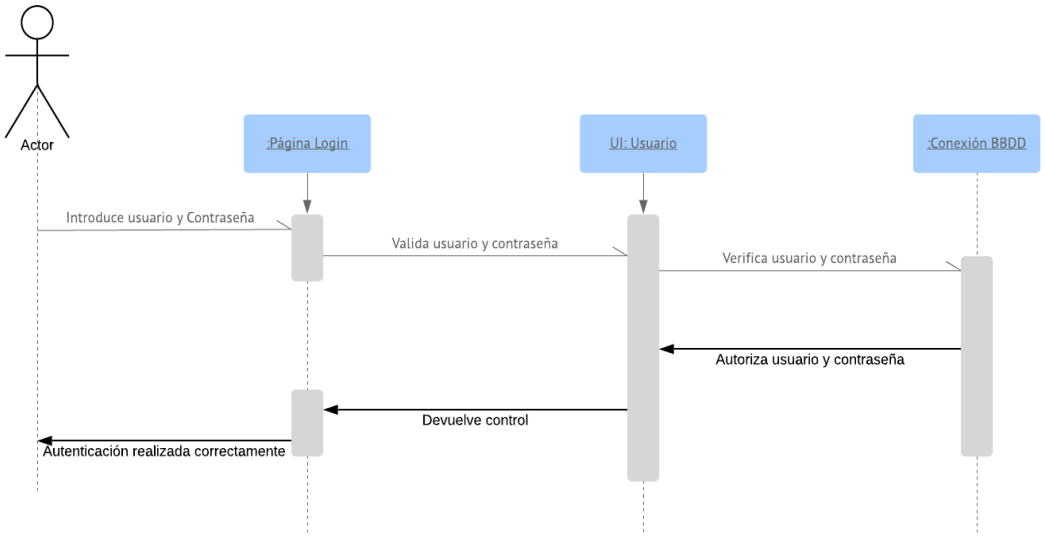


Ilustración 6. Diagrama de secuencia acceso aplicación software

Fuente de elaboración propia.

2. El diagrama de secuencias para la búsqueda de tuits a través de la librería Twitter4J.

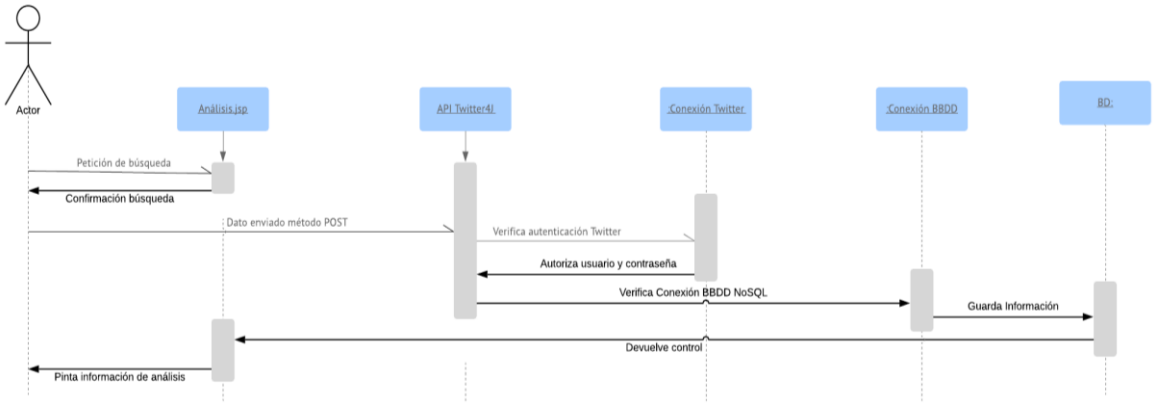


Ilustración 7. Diagrama de secuencia búsqueda de tuits

Fuente de elaboración propia.

3.4.3.2 Diagrama de clase

En el siguiente apartado se detalla el diagrama estructural también conocido como diagrama de clases, donde se detalla la jerarquía de clases de las entidades que se engloban dentro de la base de datos relacional.

"Los diagramas de clase pueden usarse cuando se desarrolla un modelo de sistema orientado a objetos para mostrar las clases en un sistema y las asociaciones entre dichas clases." (Sommerville, 2011, pág. 129)

Las diferentes clases que compone el modelo de la aplicación, representa cada una de las entidades y el conjunto pertenecen a la capa de datos.

Para el desarrollo de la aplicación de análisis de sentimiento, se ha desarrollado una única clase en la que se almacenará la información de usuario.

Ilustración 8. Diagrama de clase

Fuente de elaboración propia

3.4.3.3 Diagrama de comunicación

El diagrama de comunicación es uno de los diagramas que se engloban dentro de la vista lógica de la arquitectura 4+1 y permite las iteraciones entre los distintos objetos del sistema.

1. El diagrama de comunicación para el acceso a la aplicación software será el siguiente

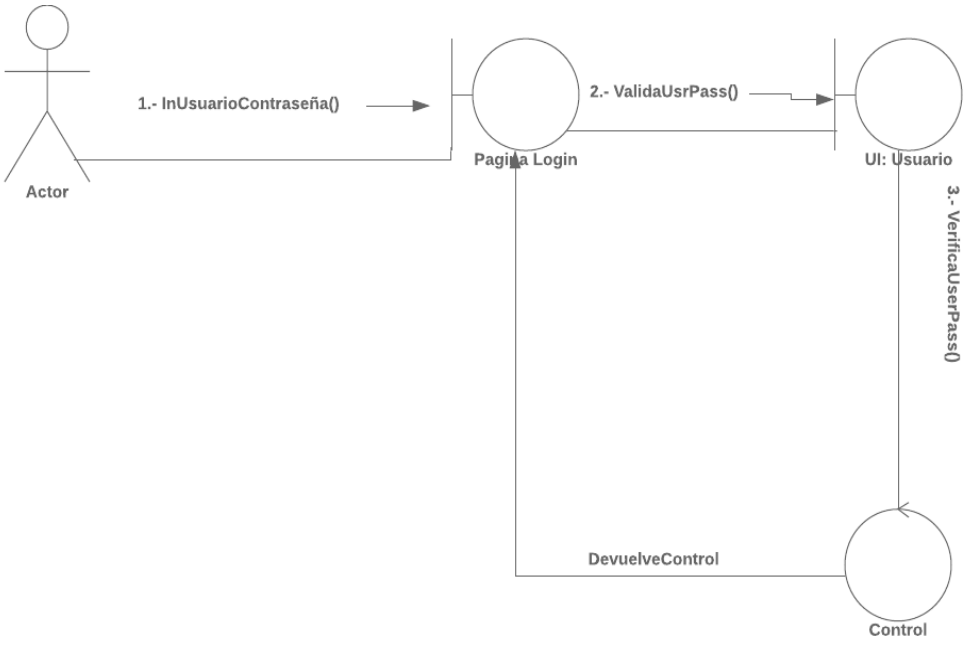


Ilustración 9. Diagrama de comunicación acceso software

Fuente de elaboración propia.

2. El diagrama de comunicación para la búsqueda de tuits a través de la librería Twitter4J.

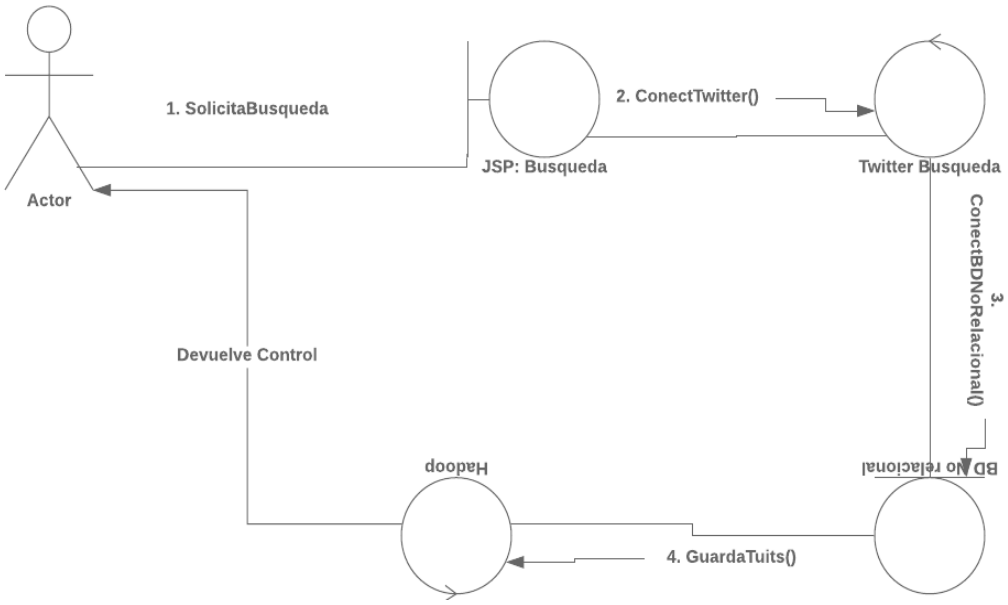


Ilustración 10. Diagrama de comunicación búsqueda de tuits

Fuente de elaboración propia.

3.4.4 Vista de desarrollo

La vista desarrollo (Modelo de Vistas de Arquitectura 4+1, Sin Fecha) describe los componentes que mantiene el sistema. La vista se compone de dos diagramas distintos y tendrá importancia para la comprensión del sistema por parte administradores y desarrolladores de software. Los diagramas en los que se componen son:

1. Diagrama de componentes

3.4.4.1 Diagrama de componentes

El diagrama de componentes es uno de los diagramas que se engloban dentro de la vista de desarrollo. El diagrama de componentes (Diagrama de componentes, Sin Fecha), representa la manera en la cual el sistema software es dividido en componentes. Además, muestra las dependencias que existen entre cada componente del sistema.

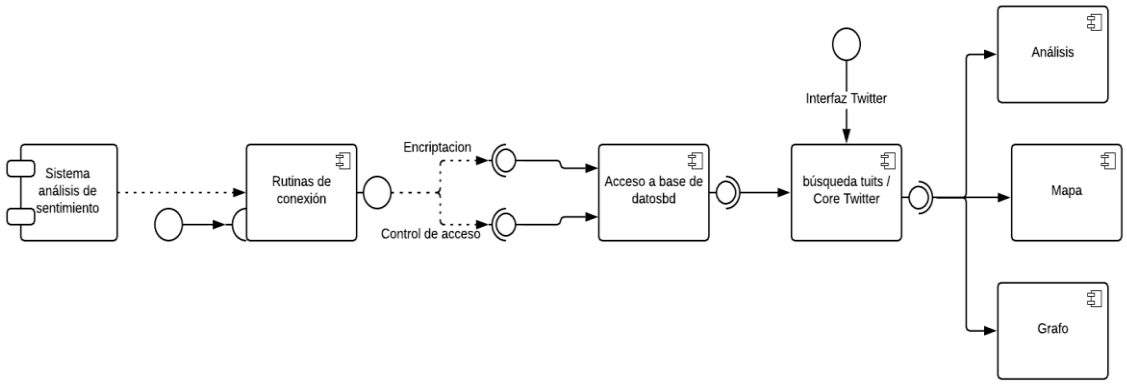


Ilustración 11. Diagrama de componentes - Aplicación software

Fuente de elaboración propia.

3.4.5 Vista de procesos

La vista procesos (Modelo de Vistas de Arquitectura 4+1, Sin Fecha) describe la manera en la cual se comunica los componentes del sistema. La vista de procesos mantiene 1 diagrama y será válido para ingenieros de sistema.

3.4.5.1 Diagrama de actividad

El diagrama de actividades es uno de los diagramas que se encuentra dentro de la vista de procesos. Este diagrama enriquece de forma general la visión de la arquitectura del sistema, ya que muestra la iteración que existe dentro de un escenario.

1. El diagrama de actividad para el acceso a la aplicación software será el siguiente

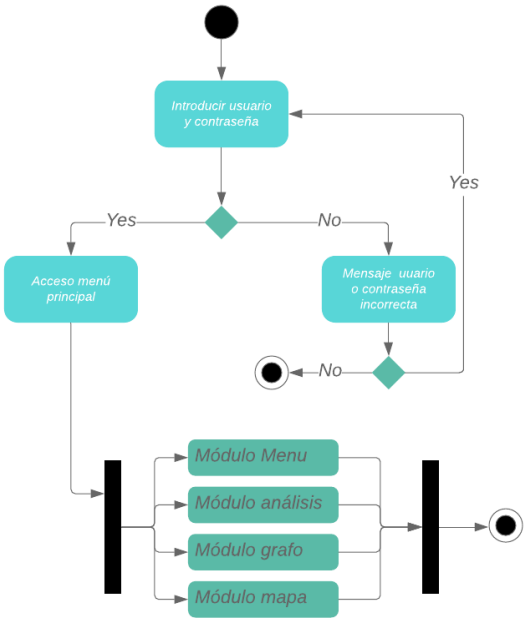


Ilustración 12. Diagrama de actividad acceso software

Fuente de elaboración propia.

3.4.6 Vista de física

La vista física (Modelo de Vistas de Arquitectura 4+1, Sin Fecha) describe la topología que mantiene el sistema. La vista física mantiene 3 diagramas distintos será útil para integradores del sistema y encargados del despliegue.

3.4.6.1 Diagrama de despliegue

En el siguiente diagrama, se representa el proceso de deploy, de los distintos nodos que contiene el sistema, los cuales, tendrán su propia representación y despliegue.

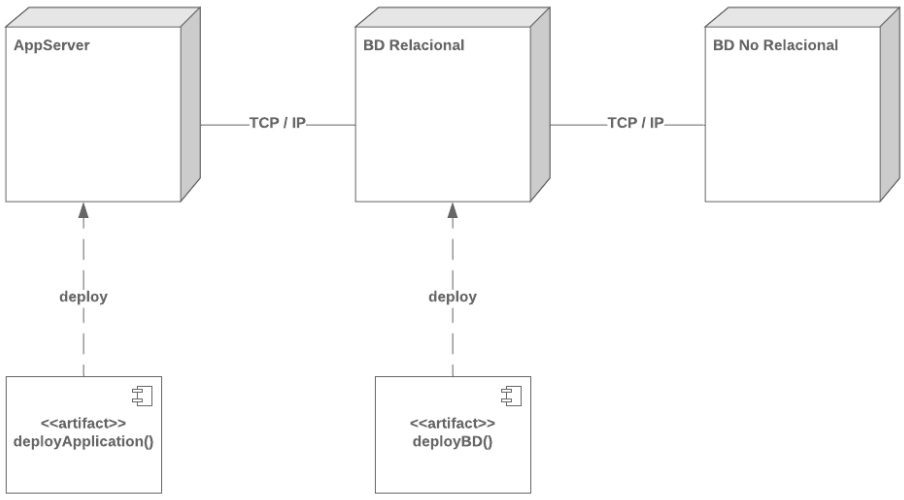


Ilustración 13. Diagrama de despliegue aplicación Análisis de Sentimientos

Fuente de elaboración propia.

3.4.7 Escenarios

Los escenarios (Modelo de Vistas de Arquitectura 4+1, Sin Fecha) describe la iteración entre usuarios y computador.

1. El escenario para el acceso a la aplicación software será el siguiente

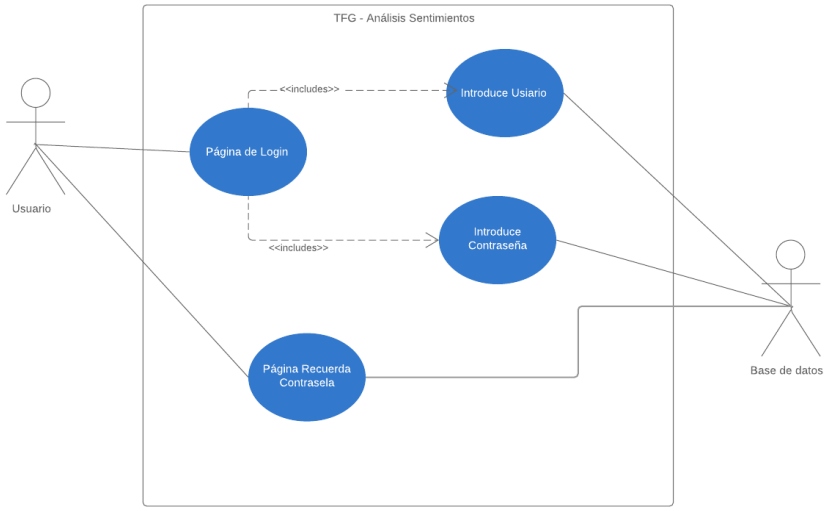


Ilustración 14. Escenario 1 - Caso de uso acceso software

Fuente de elaboración propia

2. El escenario para la búsqueda de tuits será el siguiente

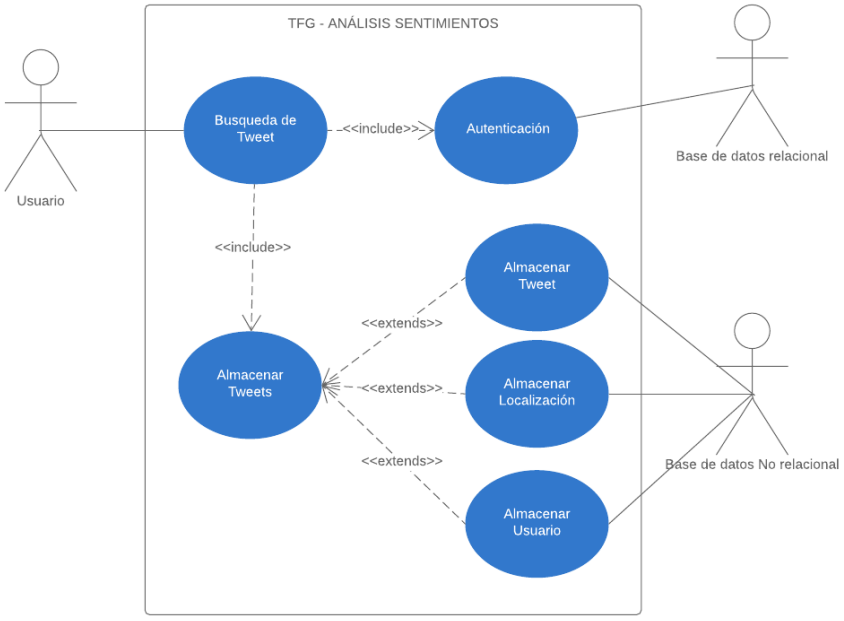


Ilustración 15. Escenario 2 - Caso de uso búsqueda de tuits

Fuente de elaboración propia

3.4.8 Patrones de diseño

En este apartado se explicarán los distintos patrones de diseño empleados para el desarrollo de la aplicación software.

"Cada patrón describe un problema que puede ocurrir una y otra vez en nuestro entorno, y luego describe el núcleo de la solución a ese problema, de manera que la solución pueda ser usada millones de veces, sin hacerlo siempre de la misma manera" (Gamma, Helm, Johnson, & Vlissides, 1994)

3.4.8.1 Patrones de diseño MVC – Modelo, Vista, Controlador

La arquitectura MVC (Gamma, Helm, Johnson, & Vlissides, 1994, pág. 4) – Modelo, Vista, Controlador – es uno de los modelos que suelen utilizarse en el desarrollo de aplicaciones software que desacopla la interfaz de usuario de su funcionalidad. La arquitectura MVC se separa en tres tipos de objetos distintos: El modelo, la vista y el controlador.

- El modelo es el objeto de la aplicación, y contiene las funcionalidades más básicas de la herramienta y los datos.
- La vista es su pantalla de presentación. Muestra y representa la información de los usuarios.
- El controlador define la manera en la que la interfaz de usuario reacciona a los inputs. Gestiona las peticiones de los usuarios.

3.4.8.2 Patrón de diseño Data Transfer Object (DTO)

El patrón de diseño Data Transfer Object, permite transmitir la información de manera eficiente entre el cliente y el servidor por medio de la creación de estructuras de datos que son diferentes del modelo de datos creado.

3.4.8.3 Patrón de diseño Data Access Object (DAO)

En el patrón de diseño Data Access Object, permite separar la lógica de negocio del acceso a la base de datos relacional, habilitando un punto de acceso único a través de la encapsulación de toda la lógica de acceso de datos.

En este patrón, está basado en una clase global y el objetivo principal será proveer a la aplicación software de los métodos necesarios que permitan insertar, borrar, consultar, actualizar los datos ubicados en el repositorio de SQL Server.

3.4.8.4 Patrón de diseño Singleton

En este patrón de diseño, también conocido como patrón de instancia única, se encargará de que una clase únicamente pueda tener acceso único objeto. Su objetivo principal por tanto

será instanciar una clase en la aplicación software, es decir, que una clase solo tenga acceso a una instancia, proporcionando un acceso global.

3.5 Tecnologías utilizadas

3.5.1 Java

Java (Java (lenguaje de programación), Sin Fecha), es un lenguaje de programación de propósito general, concurrente, orientado a objetos, que fue diseñado específicamente para tener tan pocas dependencias de implementación como fuera posible.

3.5.1.1 Librería Twitter4J

Twitter4J es una librería escrita en Java que permite la integración con la API de Twitter. Facilita la autenticación, la obtención y parametrización de datos directamente desde Twitter. La biblioteca, permite hacer búsquedas de una palabra y extraer todos los tuits que la contengan.

3.5.1.2 Librería D3.JS

D3.JS (D3.JS, Sin Fecha), por sus siglas Data Driven Document, es una biblioteca de JavaScript para la visualización datos siguiendo los estándares web utilizando SVG, Canvas and HTML.

D3 combina potentes técnicas de visualización e interacción con un enfoque basado en data-driven para la manipulación de DOM permitiendo diseñar una interfaz visual para los datos iterativa y moderna.

3.5.1.3 Librería Google Maps

La API de Google Maps (Google Maps, 2019), es una biblioteca de JavaScript para la visualización mapas iterativos desarrollada por la empresa Alphabet Inc.

Esta API de Google Maps permite personalizar los mapas con el contenido propio, de manera que se integre de forma eficaz con sus datos a través de las coordenadas de latitud y de longitud.

Existen cuatro tipos de mapas básicos (hoja de ruta, satélite, híbrido y terreno) que puede modificar utilizando capas y estilos, controles y eventos, y diversos servicios y bibliotecas.

3.5.2 Microsoft SQL Server

Microsoft SQL (SQL Server, Sin Fecha), es el gestor de base de datos relacional desarrollado por la empresa Microsoft. Utiliza un lenguaje llamado TSQL por sus siglas Transact-SQL, implementación ANSI del lenguaje SQL, que permite realizar consultas a la base de datos relacional. Para ello, es necesario disponer de la interfaz de SQL Server Management Studio para poder configurar, diseñar y administrar por medio de GUI, todo el entorno de SQL Server.

3.5.3 Librería Stanford CoreNLP

Natural Language Processing Group (Stanford NLP Group, Sin Fecha), se trata de un equipo de profesores y estudiantes pertenecientes a la universidad de Stanford que se encarga de crear y mantener distintos algoritmos que permitan a las computadoras comprender el lenguaje humano.

Para realizarlo, han perfeccionado una biblioteca llamada Stanford CoreNLP y en la que se apoyan para proveer un conjunto de herramientas de análisis gramatical que han sido desarrolladas bajo licencia pública general de GNU v3 permitiendo el procesamiento y reconocimiento del lenguaje natural.

La biblioteca Stanford CoreNLP combina el análisis sintáctico de un texto (permite analizar las estructuras gramaticales de una oración) con el análisis de dependencias.

Está escrito en lenguaje Java y su versión actual permite la utilización en distintos idiomas entre los que se encuentra, inglés, árabe, chino, francés, alemán y español.

Stanford CoreNLP proporciona un conjunto de herramientas para el tratamiento de textos entre las que se incluyen las siguientes:

1. Sistema de resolución de referencia determinista (Stanford Deterministic Coreference Resolution System, Sin Fecha), que implementa un sistema de resolución de referencias

2. Analizador del lenguaje (The Stanford Parser: A statistical parser, Sin Fecha), que resuelve las estructuras gramaticales de las oraciones. Para ello, el analizador del lenguaje se apoya de redes neuronales

3. Etiquetador del lenguaje (Stanford Log-linear Part-Of-Speech Tagger, Sin Feha), que asigna y estructura las oraciones en términos como sustantivo, verbo, adjetivo, etc.

4. Reconocedor de entidades (Stanford Named Entity Recognizer (NER), Sin Fecha), que asigna o etiqueta las formas básicas de la palabra, detectando si una palabra es un nombre de persona, nombre de compañía, etc.

Stanford CoreNLP dispone de algoritmo de clasificación de sentimientos basado en una RNN - red de tensor neuronal recursiva - al que Stanford denomina Sentiment Treebank (Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Sin Fecha) que consigue predecir con mucha precisión los efectos semánticos que se encuentran en un corpus. Además, construye un sistema de representación determinando el sentimiento que desprende el texto realizado en un lenguaje natural a través de cómo las palabras componen el significado de la frase.

3.6 Análisis de la aplicación software con interfaz web

3.6.1 Extracción de datos

El objetivo de este apartado será el de detallar el proceso de extracción de datos mediante la librería Twitter4J.

Para la realización del proceso de extracción de los tweets, es necesario disponer de los parámetros necesarios para establecer la conexión con la API de Twitter. Estos parámetros que se encuentran en el portal de desarrollo de Twitter son:

- **Consumer key** o clave del consumidor, es la clave API que se asocia a la aplicación de Twitter permitiendo la identificación del cliente, también conocido como servicio o sitio web para así acceder a los recursos de este.

- **Consumer secret** o secreto del consumidor es la contraseña del cliente que utiliza para establecer la autenticación con el servicio.

- **Access token** o token de acceso utiliza “consumer key” y “consumer secret” para establecer los privilegios que tiene el cliente, es decir, a qué datos puede acceder y a cuáles no puede acceder.

Los datos extraídos son guardados en un fichero plano en formato CSV para su posterior análisis en la aplicación. Los metadatos que se extraen y se almacenan en el fichero son los siguientes:

- **Usuario**: Usuario que realiza la publicación.
- **Mensaje**: Mensaje extraído de Twitter.
- **Localización**: Localización asociada al tweet.
- **Longitud**: Dato de localización – Longitud.
- **Latitud**: Dato de localización – Latitud.
- **Fecha**: Fecha de realización de la publicación.
- **Análisis de Sentimiento**: Devuelve el sentimiento que expresa el tweet.

En el ejemplo siguiente, se muestra un conjunto o corpus de tweets obtenidos por medio de la aplicación web con interfaz gráfica, realizando una búsqueda de tweets escritos en español y que contengan la palabra “Neymar”.

ID	Location	Date	User	Tweet
1	España	14/06/2019	@ErGarethito	@florenthanos_ Sergi Roberto y De Jong en el medio del campo, y Todibo en defensa, ya te digo que es innegable que es una pésima gestión y que yo nunca lo haría, pero ya sabemos lo que manda el vestuario en el Barça (quieren la vuelta de Neymar) y ya sabemos cómo actúa Bartomeu (pésimo gestor).
2	¿En el tiempo o en el espacio?	14/06/2019	@The_Gabrich	Ojalá Neymar chupando banquillo en París hasta el 2050
3	Málaga	14/06/2019	@CarlosVelsco	RT @juanma_rguez: No hay duda, Neymar es bobo.
4	Buenos Aires, Argentina	14/06/2019	@BarsaARG	Bartomeu informó que se pidieron 35M por falta de liquidez y se van a tener a devolver en un plazo de 6 meses. \nPor esto, el fichaje de Neymar se complica, y mucho.
8	Segovia	14/06/2019	@Aleexxx_77	El Barcelona, sin liquidez, tiene que pedir un préstamo de 35M€ a un banco para pagar la cláusula de Griezmann. No tienen dinero para fichar a Neymar y mientras este haciendo declaraciones de que quiere volver al Barça y el agente de Coutinho enfadado por si le venden. Películon
6	Torrejón de Ardoz	14/06/2019	@PipeDavid_2	RT @quimdomenech: El mejor recuerdo de Neymar en un vestuario: la remontada del Barça a su actual equipo, el PSG. ? ¿otro guiño? https://t...
7	Cartagena	14/06/2019	@Anjelote14	RT @juanma_rguez: No hay duda, Neymar es bobo.
8	Ayamonte, Andalucía, España	14/06/2019	@pepe_casanova	Hacienda 'revienta' el regreso de Neymar al Barcelona https://t.co/sJ1Lld1pvU
9	Barcelona	14/06/2019	@N1Euro	RT @jordi2039: Verano 2019
10	Venezolano	14/06/2019	@Junior10_10	RT @David_Heras: Ha comentado algo interesante Bartomeu respecto al fichaje de Griezmann y es que han tenido que pedir un préstamo de 35M€...
11	Morella	14/06/2019	@bellviso	RT @ViejaEuropa1: Me parecen totalmente respetables todos los motivos que se den sobre por qué Neymar no debe volver a vestir la camiseta d...
12	Carabayllo, Peru	14/06/2019	@OlgerCastroP96	@DavidIbanez5 Neymar besó hasta el pasto y se fue!
13	a girl lives no where	14/06/2019	@CrisMHoyos	@JKNetizen Me encanta la calidad d Neymar, tiene futbol en el, ..tiene el estilo brasileño nadie puede negar q es un TOP y q mejoraba el juego dl FCB pero, el fcb tmb lo mejoraba a el, DICHO ESTO.... SU TRAICION Y FORMA D IRSE M quitan las ganas q vuelva, sobre todo ...
14	unknown	14/06/2019	@brus_guiliz	Neymar por Coutihno y $... Buena oferta ?
15	A Coruña, Galicia	14/06/2019	@Steppen___wolf	@Ramon_AlvarezMM @ViejaEuropa1 El PSG se derrumba sin él, pese a tener en el campo a Mbappé. Y las dudas son sobre Neymar , mientras que sobre Mbappé todo son certezas
16	Asunción-Paraguay	14/06/2019	@enriquepro1	RT @jotajordi13: MESSI, SUAREZ, GRIEZMANN ES BRUTAL.\nPero si llega NEYMAR.... ver al Barça cada partido va a ser algo espectacular.

Ilustración 16. Corpus obtenido mediante Twitter4J a través de la aplicación web. Búsqueda "Neymar"

3.6.2 Análisis de sentimiento

El objetivo de este apartado será evaluar los pasos que realiza la aplicación para la clasificación del sentimiento a través de la librería Stanford CoreNLP.

En la captación del sentimiento, se realiza previamente una limpieza del tweet, para después poder ser catalogado por el clasificador.

- **Normalización del texto**

En la primera fase del proceso de análisis de sentimiento, se abordará teniendo en cuenta la eliminación de las características existentes en el texto que no aportan información para determinar la polaridad del tuit.

Este proceso tratará de normalizar los datos para mejorar la eficiencia y eficacia del clasificador. Para ello, se realizará la siguiente eliminación:

1. Eliminación de saltos de línea
2. Eliminación de URL
3. Eliminación de e-mail
4. Eliminación de hastag
5. Eliminación de menciones

```
1.  public static String LimpiaTweet(String text)
2.  {
3.      text = text.replace("\n", "\\n");
4.      text = text.replace("\t", "\\t");
5.
6.      // Search URL´s
7.      String urlPattern = "((https?|http):((//)|(\\\\))";
8.      Pattern p = Pattern.compile(urlPattern, Pattern.CASE_INSENSITIVE);
9.      Matcher m = p.matcher(text);
10.     int i = 0;
11.     while (m.find()) {
12.         text = text.replaceAll(m.group(i), "").trim();
13.         i++;
14.     }
15.     return text;
16. }
```

- **Clasificación de la polaridad y la intensidad**

En la segunda fase del proceso, la aplicación web evalúa los datos obtenidos, asignando una polaridad positiva o negativa con mayor o menor intensidad a cada tuit dependiendo de la carga emocional que desprenda.

Para la ejecución de este proceso, se ha desarrollado una clase llamada "SentimentAnalyzer" que permite clasificar el sentimiento que desprende cada tweet a través de la librería Stanford CoreNLP.

```
1.  int mainSentiment = 0;
2.  int totalRate = 0;
3.
4.  if (tweet != null && tweet.length() > 0) {
5.     int longest = 0;
6.     Annotation annotation = pipeline.process(tweet);
7.     for (CoreMap sentence : annotation.get(CoreAnnotations.SentencesAnnotation.class)) {
8.         Tree tree = sentence.get(SentimentCoreAnnotations.SentimentAnnotatedTree.class);
9.         int sentiment = RNNCoreAnnotations.getPredictedClass(tree);
10.
11.        String partText = sentence.toString();
12.        if (partText.length() > longest) {
13.            mainSentiment = sentiment;
14.            longest = partText.length();
15.        }
16.
17.        totalRate = totalRate + (sentiment - 2);
18.    }
19. }
```

La clase devolverá los valores enteros comprendidos entre 0 y el 4 que han sido clasificados tal y cómo se muestra en la tabla 5:

Tabla 5. Clasificación del sentimiento

Valor	Etiqueta	Etiqueta corta
0	Very Happy	+HAP
1	Happy	HAP
2	Neutral	NEU
3	Sad	SAD
4	Very Sad	+SAD

3.6.3 Análisis de los datos

El objetivo de este apartado será establecer representaciones visuales a partir de los datos extraídos, así como discutir los resultados de las extracciones de los corpus realizados con la herramienta software desarrollada.

La aplicación cuenta con un módulo de análisis de los datos. Este módulo a su vez se separada en dos bloques.

- **Bloque 1**: En la imagen 17, se muestra el contenido del bloque 1 con un conteo de tweets extraídos por meses. Además, se realiza un desglose del total de tweets categorizados por sentimientos.
  Adicionalmente se añade la funcionalidad que filtra los datos al situar el puntero sobre cada gráfica.

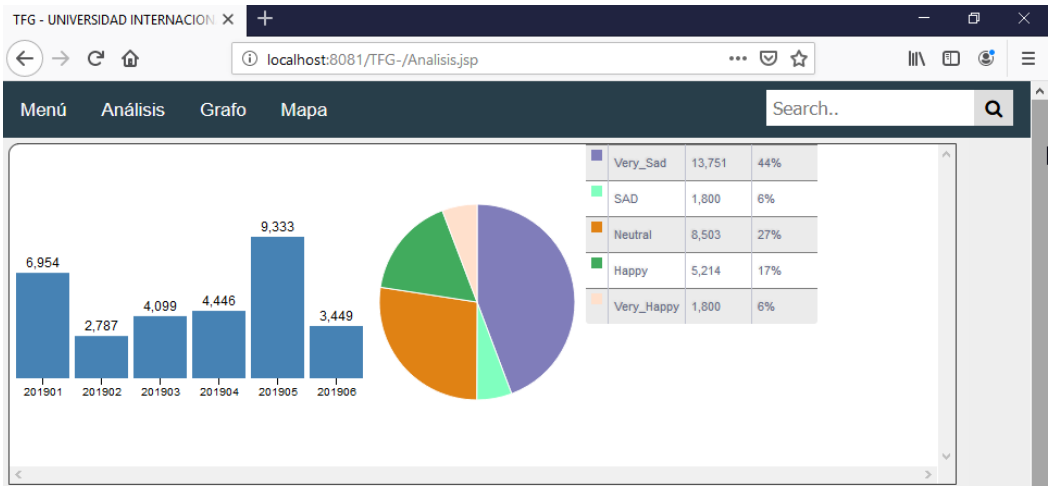
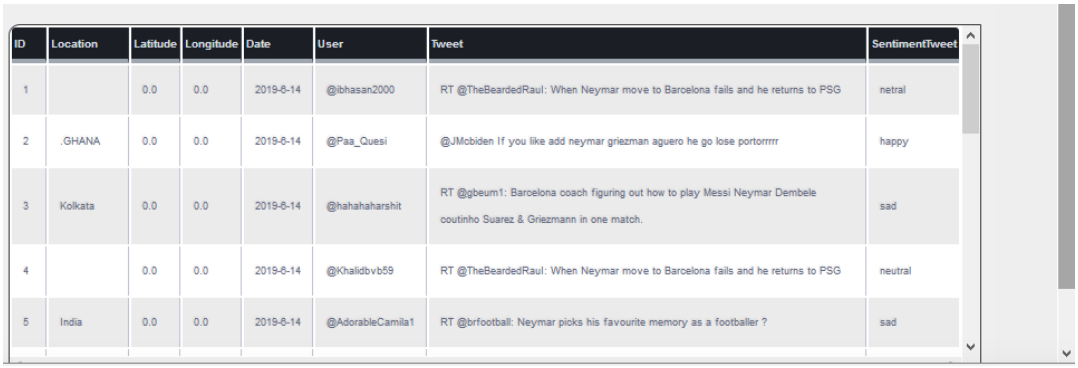


Ilustración 17. Bloque 1 - Gráficos mensuales y desglose por categoría

- **Bloque 2**: En la imagen 18, se muestra el contenido del bloque 2 con una tabla que contiene los tweets que han sido extraídos previamente mediante Twitter4J con la estructura del fichero explicada en el apartado 3.6.2.
  La lectura de los datos se realiza directamente desde el fichero CSV.



ID	Location	Latitude	Longitude	Date	User	Tweet	SentimentTweet
1		0.0	0.0	2019-6-14	@ibhasan2000	RT @TheBeardedRaul: When Neymar move to Barcelona fails and he returns to PSG	netral
2	.GHANA	0.0	0.0	2019-6-14	@Paa_Quesi	@JMcbiden If you like add neymar griezman aguero he go lose portorrrrr	happy
3	Kolkata	0.0	0.0	2019-6-14	@hahahaharshit	RT @gbeum1: Barcelona coach figuring out how to play Messi Neymar Dembele coutinho Suarez & Griezmann in one match.	sad
4		0.0	0.0	2019-6-14	@Khalidbvb59	RT @TheBeardedRaul: When Neymar move to Barcelona fails and he returns to PSG	neutral
5	India	0.0	0.0	2019-6-14	@AdorableCamila1	RT @brfootball: Neymar picks his favourite memory as a footballer ?	sad

Ilustración 18. Bloque 2 - Tabla de tweets extraídos

3.6.4 Análisis en geolocalizado

Para la realización del análisis de la geolocalización de los tweets, la aplicación web con interfaz gráfica utiliza la API de Google Maps.

Al igual que en el caso de Twitter, Google genera una clave o token que será utilizada para tener acceso a los datos de usuario y por tanto para la utilización de Google Maps. La generación de la clave se realiza directamente desde el portal de desarrollo de Google.

En la imagen 18 se muestra la disposición del mapa en la herramienta software con interfaz gráfica.

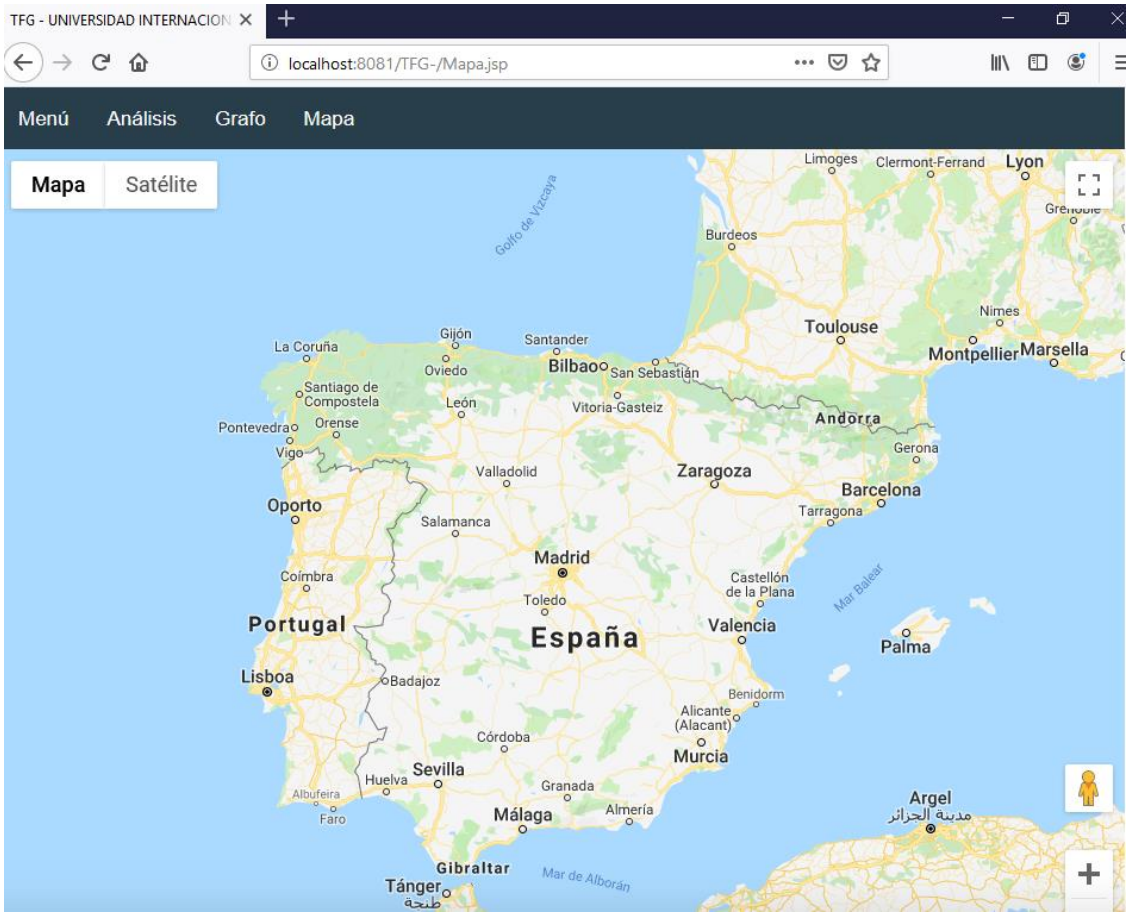


Ilustración 19. Utilización de Google Maps en aplicación web.

3.7 Validación, verificación y pruebas

En este apartado se detalla los procesos de validación, verificación y pruebas realizados sobre la aplicación software con interfaz web.

Este proceso, pretende evaluar y asegurar que el desarrollo cumple con los criterios de especificaciones detallado en el apartado 3.2.

Las pruebas realizadas han consistido en la ejecución de distintas validaciones que han evaluado la consistencia y rendimiento de la herramienta en su totalidad.

La ejecución de las pruebas se ha realizado durante todo el ciclo de vida del desarrollo software centrándose en la realización de pruebas unitarias y pruebas de rendimiento.

Esto ha permitido identificar rápidamente los defectos producidos en el desarrollo, pudiendo atacarlos lo antes posible y manteniendo un control permanente de todos los procesos. Con esta actitud proactiva, se ha evitado problemas derivados de la detección de errores en fases más avanzadas del proyecto.

El ciclo de vida en el desarrollo del software cuenta con tres fases:

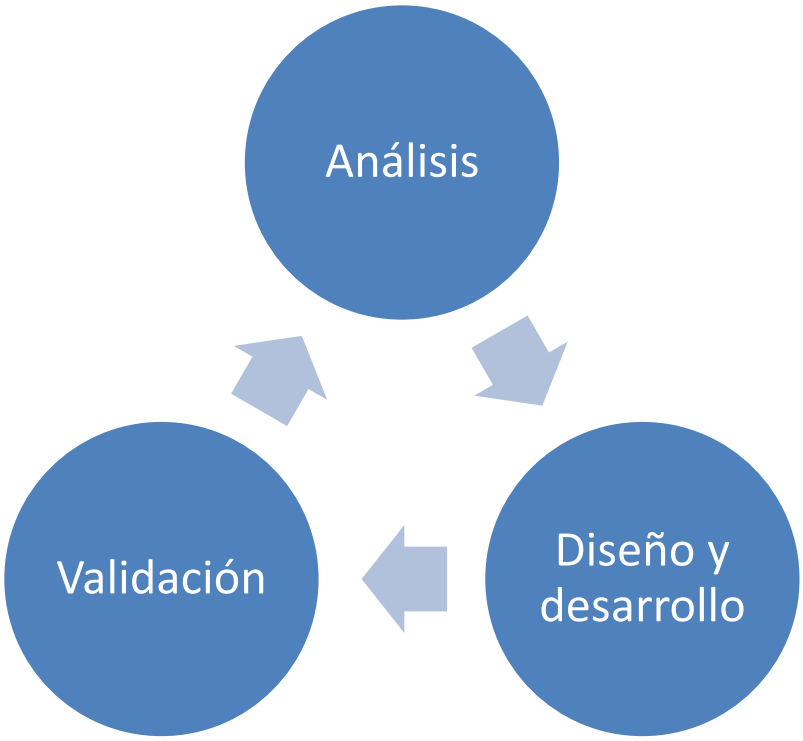


Ilustración 20. Ciclo de vida desarrollo

Fuente obtenida de (Systems Development Life Cycle, Sin Fecha)

3.7.1 Plan de pruebas

El plan de pruebas ha consistido en la realizados de dos tipos de pruebas distintas, las cuales han sido testadas a través de dos herramientas:

- Selenium IDE – Selenium IDE (Selenium, Sin Fecha) es un entorno de pruebas de software que permite testear a través de la mayoría de los navegadores web modernos, aplicaciones web para testear componentes, objetos, etc., de manera que se pueda encontrar errores.
- Apache JMeter – Jmeter (Apache JMeter, Sin Fecha) es un proyecto de apache donde a través de una herramienta software se puede diseñar pruebas de rendimiento, así como medir el comportamiento de la aplicación.

3.7.2 Pruebas unitarias

Para la realización de las pruebas unitarias, se ha utilizado la herramienta Selenium IDE para navegador Chrome. En las pruebas unitarias, se ha testeado el comportamiento de la aplicación junto con la iteración del usuario con la misma.

En la ilustración 21 se observa los pasos que se ha realizado de navegación en la herramienta software y carga de datos en el módulo de análisis y mapas principalmente. La finalización del testeo nos arroja unos resultados satisfactorios sin encontrar errores destacables durante la prueba.

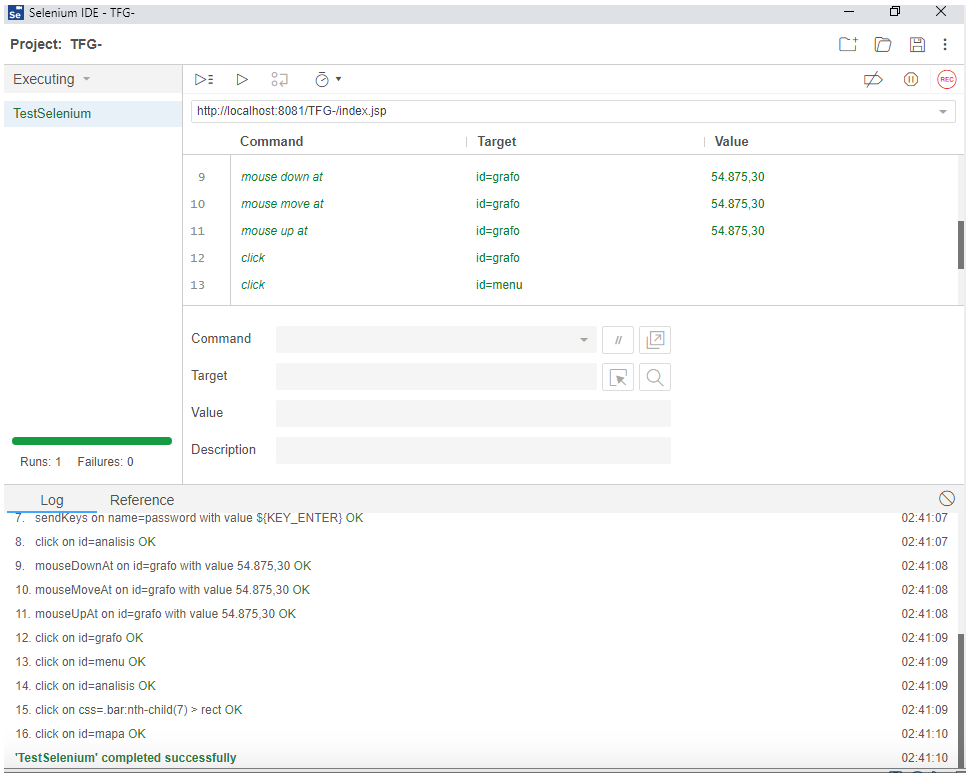


Ilustración 21. Selenium IDE - Chrome. Pruebas unitarias

3.7.3 Pruebas de rendimiento

Dado que el requisito no funcional con identificador 9 dice: “La aplicación requerirá un tiempo de respuesta óptimo y rápido, que no debe superar los 10 segundos en la autenticación”, las pruebas de rendimiento toman un especial interés.

Para la realización de las pruebas de rendimiento, se ha empleado la aplicación de código abierto llamada JMeter.

JMeter, se trata de una herramienta que consiente en la elaboración de pruebas de carga que permitan analizar y medir el desempeño de una aplicación software con interfaz web para así detectar bugs, picos de memoria, etc.

Los resultados obtenidos en las distintas pruebas determinan un correcto funcionamiento del sistema, verificando el cumpliendo del requisito que determina los tiempos de respuesta en el acceso a la herramienta.

En la tabla, se muestra la disposición de dos ejemplos realizados al azar, con carga de 1 usuario y con carga de 100 usuarios:

Tabla 6. JMeter - Tabla comparativa de test con 1 y 100 usuarios

Etiqueta	# Muestras	Media (ms)	Mínimo (ms)	Máximo (ms)	Throughput
Summary Report 1 Usuario	1	2	2	2	32,5/sec
Summary Report 100 Usuarios	100	10	2	276	53,5/sec

- Test con 100 usuarios: Cabe destacar que la herramienta JMeter y por tanto el ejemplo no indica una concurrencia de 100 usuarios si no que simula un pico de accesos de usuarios.

  En la ilustración 19, se demuestra la viabilidad de la herramienta software con interfaz gráfica a través de la recopilación de los datos siguientes:

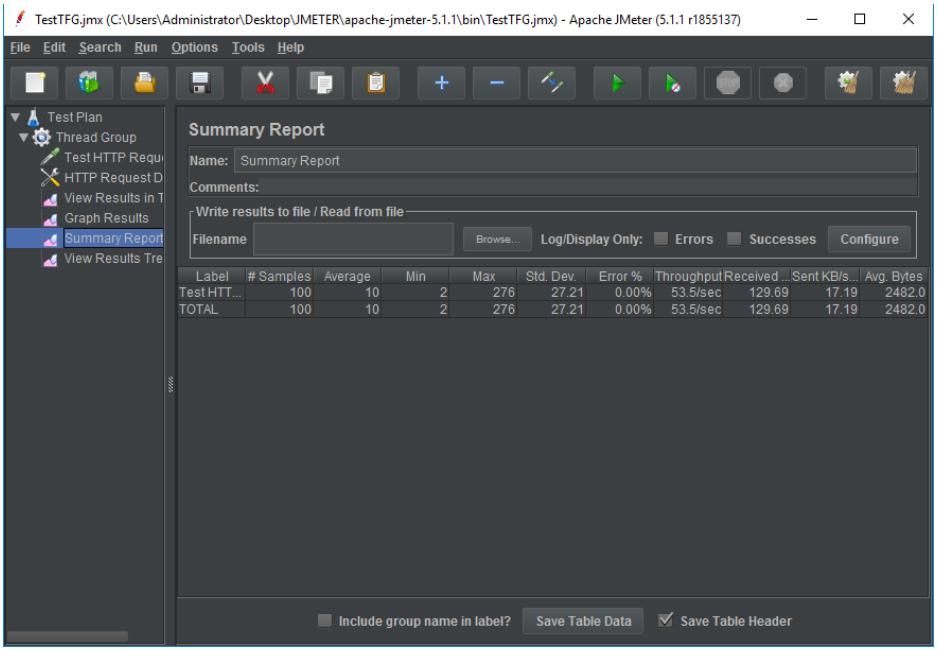


Ilustración 22. JMeter - 100 usuarios

- # Samples – número de peticiones realizadas
- Average – media aritmética de todas las respuestas (Suma el total del tiempo / número de peticiones)
- Min – Mínimo tiempo de respuesta (ms)
- Max – Máximo tiempo de respuesta (ms)
- Std. Desv – Desviación
- Error % - ratio de error. Porcentaje de fallos durante la prueba.
- Throughput – número de peticiones por Segundo que el servidor ha podido manejar durante la prueba.
- KB/Sec – KB por segundo
- Avg. Bytes – promedio de respuesta.

El tiempo mínimo de respuesta ha sido de 2 ms y el máximo de 276 ms. Tal y cómo se muestra en la ilustración 20, las peticiones HTPP han sido verificadas y aceptadas.

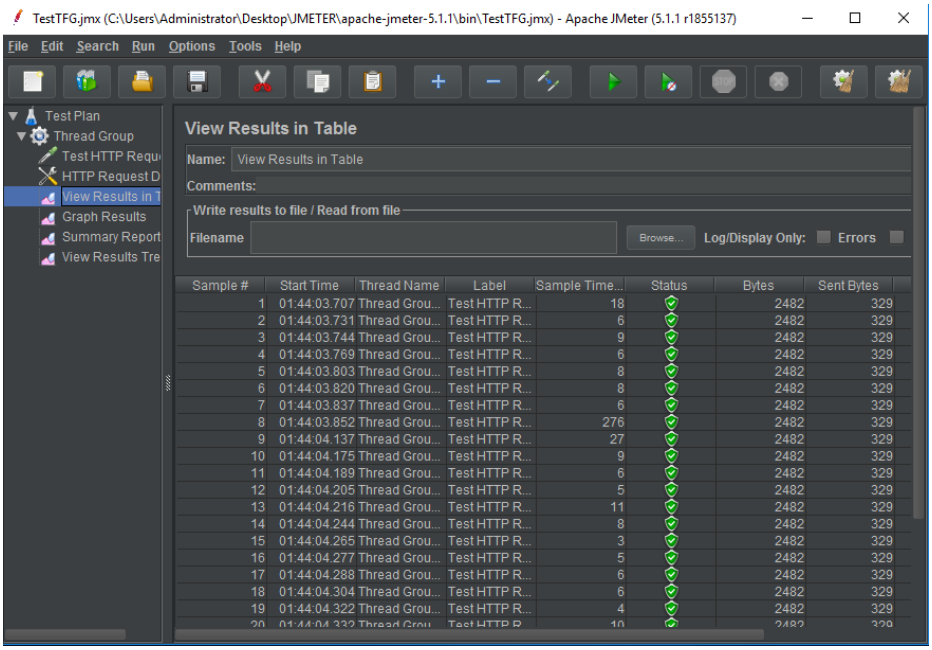


Ilustración 23. JMeter - Resultados HTTP en tabla

4 Conclusiones

Este apartado tiene como finalidad el verificar el cumplimiento de los objetivos definidos en el inicio del proyecto y plasmados en la introducción, dentro del apartado objetivos específicos de este trabajo de investigación.

Se ha realizado un estudio del arte que complementa perfectamente con la visión general del trabajo de fin de grado y cuya base se ratifica sobre la creciente existencia de proyectos de investigación sobre este ámbito de estudio del procesamiento del lenguaje natural y más concretamente sobre trabajos en el área de la minería de sentimientos y opiniones de naturaleza política.

Para la consecución de este trabajo de fin de grado se ha desarrollado una aplicación software con interfaz web basada en Java adaptándola a las recomendaciones de usabilidad y accesibilidad web a través de interfaces amigables y modernas y que cubre perfectamente la estructura de desarrollo. Esta estructura se encuentra definida en tres módulos distintos y que encapsulan la funcionalidad de cada uno de ellos. Cada módulo ha usado técnicas de reutilización del software de grano fino en las que a través de bibliotecas externas ha permitido:

- Extracción de tuits por medio de la biblioteca Twitter4J
- Categorización del sentimiento a través de la biblioteca Stanford CoreNLP
- Usabilidad de interfaz gráfica amigable a través de la biblioteca D3.JS

Este trabajo de fin de grado cubre perfectamente las técnicas y enfoques orientados al análisis de opinión a través de una aplicación software que extrae los tuits y los categoriza en un sentimiento.

El software se ha adaptado a la estructura de desarrollo englobándolo en tres módulos distintos y encapsulando la funcionalidad en cada uno de ellos. Para esto, se ha definido previamente un conjunto de especificaciones técnicas que han dado lugar a un producto que cubre la totalidad de los requisitos funcionales y no funcionales.

Todo el proyecto ha estado englobado bajo el marco de trabajo Scrum, permitiendo aplicar metodología ágil para el desarrollo del software.

Se ha familiarizado con el concepto de control de versiones. El código fuente ha estado controlado con la plataforma que permite el control de versiones en el código GitHub. El control

de versiones realizado en la memoria del trabajo de investigación ha estado controlado con Google drive.

Las pruebas finales realizadas sobre el desarrollo han determinado el correcto funcionamiento del sistema y el cumplimiento de los objetivos inicialmente establecidos.

5 Futuros Trabajos

Este apartado tiene la intención de proponer mejoras en el trabajo de investigación actual que refuercen la calidad en distintas partes de la aplicación software.

El crecimiento experimentado en los últimos años en el paradigma de Big Data y en el análisis de sentimiento ha provocado una irrupción de muchas tecnologías dentro del procesamiento del lenguaje natural y analítica del dato.

Existe una gran variedad de propuestas de mejora que se pueden aplicar al actual trabajo.

- **Análisis de sentimientos**

En el análisis de sentimiento, se han identificado distintas mejoras para el tratamiento del sentimiento de un tuit. En el proceso de análisis, se ha utilizado la librería de código libre Stanford CoreNLP. Sin embargo, esta librería no soporta el análisis de sentimiento en mensajes cuyo idioma sea español.

Por eso se propone la utilización de otras bibliotecas de análisis de opinión que permita la clasificación de textos en español. Las principales API´s que lo permiten son:

1. Google Natural Language API
2. Microsoft Text Analytics API
3. IBM AlchemyAPI

- **Análisis de sentimientos en imágenes**

El desarrollo del software se ha basado en la identificación del sentimiento expresado en un texto. Sin embargo, el crecimiento experimentado en la minería de datos ha permitido que se pueda extraer emociones faciales en imágenes publicadas en las que aparezca la cara de una persona humana.

La dificultad de esto radica en obtener la suficiente información para evaluar el sentimiento que se desprende en la imagen.

Se proponen dos tecnologías que son capaces de soportar en análisis de sentimiento en imágenes. La primera es Microsoft Cognitive Services. La segunda es Face detection de Jquery.

- **Aplicación software**

En el desarrollo de la aplicación software, se han identificado distintas mejoras que serán enumeradas a continuación:

1. Spark Streaming

En primer lugar, sería óptimo disponer de una aplicación que permita la descarga de tuits de manera streaming, mediante un flujo continuo de información, almacenando la fecha y hora de cada tuit. De esta manera, se podrá tener un histórico de información que permita realizar comparativas en tiempo real.

El almacenamiento de datos en el modelo actual se realiza por medio de un fichero, por lo que la aplicación deberá ser adaptada para disponer de una escritura y lectura rápida y óptima.

Una de las herramientas que se propone y que permite realizar estas tareas es Spark Streaming.

2. Redes Sociales

En segundo lugar, se podría dar un valor añadido a la aplicación software, facilitando la incorporación de otras redes sociales en la extracción de la información, con el objetivo de dar una capa analítica más amplia.

El modelo actual cuenta con la recolección de datos desde la red social Twitter. Sin embargo, existen distintas redes sociales con mayor número de usuarios que podrían dar distintas ventajas adicionales al uso de la aplicación.

6 Acrónimos

BLS (Bureau of Labor Statistics). Es traducido como oficina de estadísticas laborales, es la principal agencia investigación de Estados de Unidos encargada de realizar estudios estadísticos de investigación en EE. UU.

D3. (Data Driven Document). Biblioteca de JavaScript para la visualización iterativa de los datos.

GUI. (Graphical User Interface). Interfaz gráfica de usuario.

HTML. (HyperText Markup Language). Lenguaje de programación para la creación de páginas web.

MaxEnt (Máxima Entropía).

NLP. (Natural Language Processing). Es el campo de la computación que estudia las interacciones entre las computadoras y el lenguaje humano.

NER. (Named Entity Recognizer). Es el reconocedor de entidades desarrollado por Stanford.

PB. (Product Backlog).

PBI. (Product Backlog Items).

TFG. Trabajo Fin de Grado. Finalización académica del plan de tipología grado cuya duración es de cuatro años.

TSQL. (Transact SQL). Implementación ANSI del lenguaje SQL desarrollado por la empresa Microsoft.

SVG. (Scalable Vector Graphics). Gráficos vectoriales escalables.

SVN´s (Support Vector Machine). Conjunto de algoritmos de aprendizaje supervisado

SEPLN. Sociedad Española para el Procesamiento del Lenguaje Natural, es la agencia que realiza investigaciones sobre el procesamiento del lenguaje natural.

RNN. (Recursive neuronal network). La red neuronal recursiva es una red neuronal utilizado en la librería Stanford CoreNLP para categorizar los sentimientos asociados a una frase.

7 Bibliografía

Alvarez, R., Garcia, D., Moreno, Y., & Schweitzer, F. (2015). *Sentiment cascades in the 15M movement.* Recuperado el 29 de Mayo de 2019, de https://arxiv.org/abs/1505.03776

Análisis de Sentimiento. (Sin Fecha). *En Wikipedia*. Recuperado el 11 de Mayo de 2019, de https://es.wikipedia.org/wiki/An%C3%A1lisis_de_sentimiento

Apache JMeter. (Sin Fecha). *En Wikipedia*. Recuperado el 11 de Julio de 2019, de https://es.wikipedia.org/wiki/JMeter

Bosch, J. (2000). Design and use of Software Architectures. En J. Bosch, *Design and use of Software Architectures* (pág. 23). Addision-Wesley. Recuperado el 24 de Mayo de 2019

Brandwatch Analytics. (Sin Fecha). *Brandwatch*. Recuperado el 2 de Junio de 2019, de https://www.brandwatch.com/es/blog/analisis-de-sentimiento/

Bureau of Labor Statistics, U.S. Department of Labor. (2019). Occupational Outlook Handbook, Computer and Information Research Scientists. Recuperado el 6 de Mayo de 2019, de https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm

Cambridge Analytica. (Sin Fecha). *En Wikipedia*. Recuperado el 28 de Mayo de 2019, de https://es.wikipedia.org/wiki/Cambridge_Analytica

Chomsky, N. (2019). *En Wikipedia*. Recuperado el 13 de Mayo de 2019, de https://es.wikipedia.org/wiki/Ensamble_(ling%C3%BC%C3%ADstica)

Choy, M., Cheong, M. L., Laik, M. N., & Shung, K. P. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. Recuperado el 18 de Mayo de 2019, de https://arxiv.org/abs/1108.5520

Criado, M. Á. (15 de Mayo de 2011). El 15-M se alimentó de las emociones en Twitter. Recuperado el 28 de Mayo de 2019, de https://elpais.com/elpais/2015/05/15/ciencia/1431702091_568838.html

D3.JS. (Sin Fecha). *En Wikipedia*. Recuperado el 19 de Mayo de 2019, de https://es.wikipedia.org/wiki/D3.js

Dave, K., Lawrence, S., & Pennock, D. M. (2003). *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews.* Proceedings of the 12th

international conference on World Wide Web. Recuperado el 9 de Junio de 2019, de https://www.kushaldave.com/p451-dave.pdf

Deemer, P., Benefield, G., Larman, C., & Vodde, B. (Sin Fecha). *Good Agile*, 1.1. Recuperado el 24 de Mayo de 2019, de http://www.goodagile.com/scrumprimer/scrumprimer_es.pdf

Diagrama de componentes. (Sin Fecha). *En Wikipedia*. Recuperado el 1 de Junio de 2019, de https://es.wikipedia.org/wiki/Diagrama_de_componentes

Diagrama de secuencia. (Sin Fecha). *En Wikipedia*. Recuperado el 2019 de Junio de 1, de https://es.wikipedia.org/wiki/Diagrama_de_secuencia

Erturk, E., & Shi, H. (2016). Natural Language Processing using Hadoop and KOSHIK. Recuperado el 18 de Mayo de 2019, de https://arxiv.org/abs/1608.04434

Gajarla, V., & Gupta, A. (2017). *Emotion Detection and Sentiment Analysis of Images.* Recuperado el 10 de Junio de 2019, de https://www.cc.gatech.edu/~hays/7476/projects/Aditi_Vasavi.pdf

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley. Recuperado el Mayo de 1 de 2019

General Architecture for Text Engineering. (Sin Fecha). *En Wikipedia*. Recuperado el 2019 de Julio de 3, de https://es.wikipedia.org/wiki/General_Architecture_for_Text_Engineering

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying Sarcasm in Twitter: A Closer Look. Recuperado el 18 de Mayo de 2019, de https://pdfs.semanticscholar.org/55e3/6d6b45c91a0daa49234bd47b856470d6825c.pdf

Google Maps. (2019). *Google Maps Documentation*. Recuperado el 19 de Mayo de 2019, de https://developers.google.com/maps/documentation/javascript/tutorial?hl=es

Harfoush, R. (2009). Yes We Did! An inside look at how social media built the Obama brand. Recuperado el 28 de Mayo de 2019

Hootsuite Insights. (Sin Fecha). *Hootsuite*. Recuperado el 2019 de Junio de 2, de https://hootsuite.com/es/productos/insights#

Hootsuite´s. (Sin Fecha). *En Wikipedia*. Recuperado el 28 de Mayo de 2019, de https://es.wikipedia.org/wiki/Hootsuite

Java (lenguaje de programación). (Sin Fecha). *En Wikipedia*. Recuperado el 19 de Mayo de 2019, de https://es.wikipedia.org/wiki/Java_(lenguaje_de_programaci%C3%B3n)

Koppel, M., & Schler, J. (2006). *The Importance of Neutral Examples for Learning Sentiment.* Recuperado el 1 de Junio de 2019, de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.9097&rep=rep1&type=pdf

Kozinets, R. V. (2002). *The Field Behind the Screen: Using Netnography For Marketing Research in Online Communities*. Recuperado el 25 de Mayo de 2019, de https://www.nyu.edu/classes/bkg/methods/netnography.pdf

Lahuerta Otero, E., & Cordero Gutiérrez, R. (Sin Fecha). *Redes sociales: un antes y un después en el comportamiento humano*. Recuperado el 27 de Mayo de 2019, de https://telos.fundaciontelefonica.com/redes-sociales-un-antes-y-un-despues-en-el-comportamiento-humano/

Lozares-Colina, C. (1996). La teoría de redes sociales. *Revista de sociología*, 103-126. Recuperado el 24 de Mayo de 2019, de https://papers.uab.cat/article/view/v48-lozares/pdf-es

Modelo de Vistas de Arquitectura 4+1. (Sin Fecha). *En Wikipedia*. Obtenido de https://es.wikipedia.org/wiki/Modelo_de_Vistas_de_Arquitectura_4%2B1

Naturaleza humana. (Sin Fecha). *En Wikipedia*. Recuperado el 15 de Mayo de 2019, de https://es.wikipedia.org/wiki/Naturaleza_humana

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Recuperado el 13 de Mayo de 2019, de http://www.lrec-conf.org/proceedings/lrec2010/summaries/385.html

Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis.* Foundations and Trends in Information Retrieval. Recuperado el 18 de Junio de 2019

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Recuperado el 8 de Mayo de 2019, de https://arxiv.org/abs/cs/0205070

Pew Research Center. (2019). *En Wikipedia*. Recuperado el 17 de Mayo de 2019, de https://es.wikipedia.org/wiki/Pew_Research_Center

Pew Research Center. (2019). *US survey research 2019.* Recuperado el 16 de Mayo de 2019, de https://www.pewresearch.org/wp-content/uploads/2019/04/FT_19.04.10_SocialMedia2019_topline_methodology.pdf

Procesamiento del lenguaje natural. (Sin Fecha). *En Wikipedia*. Recuperado el 13 de Mayo de 2019, de https://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales

Programación por capas. (Sin Fecha). *En Wikipedia.* Recuperado el 1 de Junio de 2019, de https://es.wikipedia.org/wiki/Programaci%C3%B3n_por_capas

RapidMiner. (Sin Fecha). *En Wikipedia*. Recuperado el 1 de Junio de 2019, de https://es.wikipedia.org/wiki/RapidMiner

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. (Sin Fecha). *Stanford NLP Deep Model*. Recuperado el 18 de Mayo de 2019, de https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf

Reporteros sin Fronteras. (Sin Fecha). Primavera árabe: ¿apogeo de la Web? Recuperado el 29 de Mayo de 2019, de https://www.rsf-es.org/grandes-citas/dia-contra-censura-en-internet/a2011-dia-mundial-contra-la-censura-en-internet/frente-a-la-censura-solidaridad-en-la-red/

Russell, J. A. (1980). *A circumplex model of affect*. Recuperado el 22 de Junio de 2019, de https://www.researchgate.net/publication/235361517_A_Circumplex_Model_of_Affect

Salvetti, F., Lewis, S., & Reichenbach, C. (2006). *Automatic Opinion Polarity Classification of Movie.* Recuperado el 9 de Junio de 2019

Schwaber, K., & Sutherland, J. (2013). *La Guía de Scrum.* Recuperado el 2019 de Mayo de 29, de https://www.scrumguides.org/docs/scrumguide/v1/scrum-guide-es.pdf

Scrum (desarrollo de software). (2019). *En Wikipedia*. Recuperado el 25 de Mayo de 2019, de https://es.wikipedia.org/wiki/Scrum_(desarrollo_de_software)

Selenium. (Sin Fecha). *En Wikipedia*. Recuperado el Julio de 12 de 2019, de https://es.wikipedia.org/wiki/Selenium

Semantria. (Sin Fecha). *En Wikipedia*. Recuperado el 2 de Junio de 2019, de https://en.wikipedia.org/wiki/Semantria

Sociedad Española para el Procesamiento del Lenguaje Natural. (Sin Fecha). *En Wikipedia*. Obtenido de https://es.wikipedia.org/wiki/Sociedad_Espa%C3%B1ola_para_el_Procesamiento_del_Lenguaje_Natural

Sommerville, I. (2011). *Ingeniería de Software.* Madrid: Pearson Education S.A.

SQL Server. (Sin Fecha). *Microsoft SQL Server*. Obtenido de https://es.wikipedia.org/wiki/Microsoft_SQL_Server

Stanford Deterministic Coreference Resolution System. (Sin Fecha). *Stanford NLP Deterministic*. Recuperado el 18 de Mayo de 2019, de https://nlp.stanford.edu/software/dcoref.shtml

Stanford Log-linear Part-Of-Speech Tagger. (Sin Feha). *Stanford NLP Log-linear*. Recuperado el 18 de Mayo de 2019, de https://nlp.stanford.edu/software/tagger.shtml

Stanford Named Entity Recognizer (NER). (Sin Fecha). *Stanford NLP NER*. Recuperado el 18 de Mayo de 2019, de https://nlp.stanford.edu/software/CRF-NER.html

Stanford NLP Group. (Sin Fecha). *Stanford NLP Group*. Recuperado el 18 de Mayo de 2019, de https://nlp.stanford.edu

Stirland, S. L. (2008). Obama´s Secret Weapons: Internet, Databases and Psychology. *Wired*. Recuperado el 18 de Mayo de 2019, de https://www.wired.com/2008/10/obamas-secret-w/

Systems Development Life Cycle. (Sin Fecha). *En Wikipedia*. Recuperado el 22 de Junio de 2019, de https://es.wikipedia.org/wiki/Systems_Development_Life_Cycle

Taboada, M., Brooke, J., Tofiloski, M., & Voll, K. (2011). *Lexicon-Based Methods for Sentiment Analysis.* Recuperado el 10 de Junio de 2019, de https://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00049

The Stanford Parser: A statistical parser. (Sin Fecha). *Stanford NLP Parser*. Recuperado el 18 de Mayo de 2019, de https://nlp.stanford.edu/software/lex-parser.shtml

Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Recuperado el 8 de Mayo de 2019, de https://arxiv.org/abs/cs/0212032

Twitter. (Sin Fecha). *En Wikipedia*. Recuperado el 28 de Mayo de 2019, de https://es.wikipedia.org/wiki/Twitter

Varela, I. (2019). El Tribunal Constitucional nos rescata de una cacicada de los partidos. Recuperado el 23 de Mayo de 2019, de https://blogs.elconfidencial.com/espana/una-cierta-mirada/2019-05-23/tribunal-constitucional-politicos-presos_2018278/

Vilares, D., Thelwall, M., & Alonso, M. A. (2015). The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. *Journal of Information Science*. Recuperado el 23 de 05 de 2019, de http://www.scit.wlv.ac.uk/~cm1993/papers/MegaphonePreprint.pdf

Wang, H., Dogan Can, A. K., & François Bar, S. N. (2012). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. Recuperado el 18 de Mayo de 2019, de https://www.aclweb.org/anthology/P12-3020

We are social. (2019). *Digital in 2019 España*. Recuperado el 28 de Mayo de 2019, de https://wearesocial.com/es/digital-2019-espana

We are Social and Hootsuite´s. (2019). *Global Digital Report 2019.* Recuperado el 10 de Mayo de 2019, de https://wearesocial.com/global-digital-report-2019

Weka. (Sin Fecha). *En Wikipedia*. Recuperado el 2019 de Julio de 3, de https://es.wikipedia.org/wiki/Weka_(aprendizaje_autom%C3%A1tico)