Análisis de opinión en Twitter sobre la independencia de Cataluña

Samuel Navero Rodriguez

Resumen — El proyecto se basa en la implementación de un sistema de análisis de sentimiento bajo la red social Twitter, sobre las opiniones que publican los usuarios acerca de la Independencia de Cataluña y los presos políticos. El proceso se inicia creando una base de datos, obtenida a partir de una recopilación sobre una cierta cantidad de Tweets, y posteriormente analizado su contenido utilizando varios algoritmos de sentimientos. El objetivo principal es el de averiguar si dichas publicaciones acerca del tema que nos concierne son positivas, negativas o neutras, con la finalidad de obtener una idea sobre la opinión que tienen los usuarios que utilizan esta red social.

Palabras clave — Token,Twitter, Opinion mining, Analisis de sentimiento, Aplicación, Nodejs, Javascript, JSON, ElasticSearch, Kibana, ELK, APIs, Postman, Tweets, Trello.

Abstract — The project is based on the implementation of a system of anàlisis of feeling under the social network Twitter, on the opinions that the users publish it brings over of the Independence of Catalonia and the political prisoners. The process begins creating a database obtained from a summary on Tweets's certain quantity, and later analyzed his content using several algorithms of feelings. The principal aim is of quarrelling if said publicacions brings over of the topic that us concerns they are positive, negative or neutral, with the purpose of obtaining an idea on the opinion that there have the users who use this social network.

Index Terms — Token, Twitter, Opinion mining, Sentiment analysis, Application, Nodejs, Javascript, JSON, ElasticSearch, Kibana, ELK, APIs, Postman, Tweets, Trello.

1 INTRODUCCIÓN

Desde hace ya unos años con la aparición de los dispositivos móviles multitarea, han surgido nuevos tipos de aplicaciones que nos dan la oportunidad de publicar contenidos para que otros usuarios puedan visualizarlo, comentarlo o simplemente expresar su opinión sobre cualquier tema.

De estas aplicaciones, las que más auge han tenido en los últimos años han sido las denominadas redes sociales, que son una herramienta de comunicación entre diversos usuarios bajo un mismo tema. Hoy en día estas redes sociales se han convertido en un lugar en el que se forman y construyen relaciones, se comparten expresiones y conocimientos sobre el mundo que nos rodea y los acontecimientos que van surgiendo a lo largo del tiempo, y que además nos presentan oportunidades para la innovación, la creatividad y el aprendizaje. De entre todas las redes sociales que existen en la actualidad nos centraremos en Twitter [1], una red social que fue lanzada en 2006 y que cuenta con más de 500 millones de usuarios activos en la actualidad, los cuales realizan más de 50 millones de publicaciones al día. Esta red social ha demostrado ser un excelente puente de comunicación entre diferentes usuarios, que permite transmitir información sobre la actualidad, como por ejemplo en revueltas, conflictos políticos o catástrofes naturales. Ya que Twitter nos ofrece testimonios directos e instantáneos desde cualquier parte del mundo, haciendo posible la comunicación bidireccional y convierte a cualquier personalidad en un igual.

La principal función de Twitter es la de conectar personas dentro del mundo virtual, ya sea para construir nuevas conexiones sociales o para mantener las existentes. Permite a sus usuarios publicar y compartir videos, noticias, fotos y sobre todo mensajes de texto con un máximo 280 caracteres para expresar sus opiniones sobre cualquier tema. A estos tipos de publicaciones en Twitter se les denomina comúnmente como Tweets. Una de las principales razones de publicar estos Tweets es la de expresar la opinión que tiene cada persona sobre un tema en concreto, para referirse a ciertos temas en concreto se suele emplear Hashtags, que son palabras que se inician con el símbolo de almohadilla, relacionadas sobre algún tema en concreto al que el usuario quiere hacer referencia, para realizar su publicación. Un ejemplo claro seria #*TFG* para referirse al hashtag "TFG".

El principal objetivo del proyecto tiene relación directa con estos Tweets, más concretamente con los mensajes de texto

- *E-mail de contacto: Samuel.Navero@e-campus.uab.cat*
- *Mencion realizada: Tecnologias de la Información.*
- *Trabajo tutorizado por: Jordi Casas Roma*
- *Curso 2018/19*

que publican diferentes usuarios sobre un tema en concreto bajo ciertos hashtags, para realizar un análisis de sentimientos sobre ellos y poder obtener información sobre la opinión que tienen los usuarios sobre el tema. Para ello nos encargaremos de recopilar diversas publicaciones realizadas por usuarios aleatorios, para después analizarlos utilizando varios algoritmos de análisis de sentimiento, y tras compararlos entre si, elegiremos el que tenga una menor tasa de error para finalmente obtener unos resultados y saber si esos mensajes de texto son de carácter positivo, negativo o neutro.
El tema que he elegido para realizar este proyecto ha sido sobre la independencia de Cataluña y los presos políticos, un problema social que actualmente está en auge en España. Este tema provoca en la actualidad conflictos entre diferentes usuarios de dicha red, los cuales manifiestan opiniones diferentes, por lo que los analizaremos para así poder tener una idea de cual es la opinión global que se tiene en Twitter sobre este tema.
A continuación, en los siguientes apartados veremos los objetivos principales de este proyecto de forma más concreta, así como la metodología escogida que comentaremos en la *sección 4* y la planificación establecida para llevar a cabo el proyecto que describiremos en el *apartado 5*. Continuaremos viendo que herramientas hemos utilizado para la recopilación de mensajes en el *apartado 6*, así como los algoritmos que hemos empleado en los *apartados 6.6* y *6.7* respectivamente, así como la herramienta de la que hemos hecho uso para poder visualizar dichos análisis gráficamente, que comentaremos en el *apartado 6.11*.

2 OBJETIVOS

El principal objetivo del proyecto es el de analizar la opinión que expresan diferentes personas sobre un tema concreto, en una misma red social, más concretamente, la red social Twitter. Los textos que analizaremos serán aquellos relacionados con la independencia de Cataluña y los presos políticos, para así poder tener una idea de cual es la valoración que hacen dichos usuarios.
Los principales objetivos que debemos cumplir para llevar a cabo nuestro proyecto han sido:

1. Diseñar un sistema capaz de capturar datos de Twitter.
2. Diseñar y construir una bbdd para almacenar los diferentes Tweets.
3. Analizar las publicaciones con los diferentes algoritmos de sentimiento.
4. Visualizar los resultados obtenidos.

3 ESTADO DEL ARTE

En la actualidad Twitter es una de las redes sociales más utilizadas en todo el mundo, ya que cuenta con más de 500 millones de usuarios activos y un crecimiento exponencial muy positivo. Todo esto es gracias a la multifunción de esta red social, ya que muchos usuarios acceden no solo para expresar sus opiniones sino también para publicitarse, o incluso, en el mundo empresarial, esta red social es utilizada para conocer las opiniones que tienen sus usuarios sobre sus productos y así poder mejorar en diversos aspectos de su empresa, como calidad, marketing entre otros, gracias a estos feedbacks directos del usuario.
Para poder obtener estos resultados existen muchas aplicaciones, que utilizan algoritmos de análisis de sentimiento, y que nos permiten saber si una opinión es favorable, negativa o incluso neutra.

En esta línea encontramos por ejemplo la empresa Brandwatch, que pone a disposición del cliente y a partir de previo pago, la plataforma Brandwatch Analytics [2] que según su propio eslogan "Detecta, monitoriza y analiza millones de conversaciones para entender mejor lo que piensan tus clientes". Por lo que permite segmentar y analizar las opiniones de los usuarios a empresas como Mapfre, Grupo Bimbo, Ikea o Microsoft, entre otras, para así poder conocer mejor los gustos e inquietudes de sus usuarios y por lo tanto, ofrecerles un mejor producto y servicio.

4 METODOLOGIA

Para llevar acabo este proyecto se ha utilizado la metodología Scrum [3], que se basa en la realización de entregas parciales y regulares del producto final, siempre teniendo en cuenta la disponibilidad de los recursos, así como la dificultad que conlleva su implementación. Por lo tanto, se ha dividido el proyecto en pequeños desarrollos, cada uno con sus recursos asignados y validados. De manera que al finalizar el tiempo establecido para cada parte se hace un análisis de su estado actual, para planificar y realizar la siguiente etapa o si por lo contrario redefinir esta etapa si no se ha podido finalizar y por qué motivos se ha producido el retraso.
Como herramienta para el control de tareas se ha decidido utilizar Trello [4], una herramienta que dispone de varios estados que podemos establecer (Non Started, In Progress, Blocked, Done, etc.) para cada tarea o subtarea generada en los Sprints, y que también permite hacer comentarios, poner fechas limite, enlaces, subir archivos, agregar checklist o crear etiquetas para poder documentar todo el proceso de desarrollo de las tareas, y así lleva un buen control del proyecto.
En cuanto a lenguajes de programación y herramientas para la implementación del proyecto, se han utilizado NodeJS para la conexión con Twitter, descarga de los Tweets y desarrollo de los algoritmos de sentimiento. Para almacenar los JSON que contienen los Tweets se ha utilizado ElasticSearch [5] por su gran fluidez, combinado con Postman [6] para realizar pruebas y con Kibana [7] para finalmente generar los gráficos de los resultados obtenidos.

5 PLANIFICACIÓN

Para empezar este proyecto se ha planteado una

duración aproximada de unas 300 horas de trabajo repartidas entre los meses de octubre a diciembre de 2019 para poder desarrollar completamente el producto.
Al utilizar una metodología ágil, el proyecto se ha dividido en tareas más pequeñas llamadas Sprint con una duración determinada. Al finalizar el tiempo de cada sprint se valora si la tarea se ha podido realizar correctamente y si no ha sido el caso, porque se ha retrasado o que complicaciones se han producido. En este proyecto se ha tenido que modificar la planificación diversas veces para poder adaptarse a los inconvenientes que han ido produciéndose. Estos cambios los veremos al final de este apartado, y también en el apartado "*A1. Tabla planificación*" donde se observa la tabla de la planificación detallada.

Las tareas en las que se ha dividido el proyecto y que conforman cada uno de los sprints son las siguientes:

1. Buscar productos/aplicaciones similares a lo que se propone, caracterizarlos y compararlos.
2. Autoaprendizaje de los lenguajes y herramientas necesarias para llevar a cbo el proyecto.
3. Investigación sobre Twitter, sus usos y que herramientas tiene disponibles para poder reallizar el desarrollo que se desea.
4. Preparar la implementación de la base de datos a partir de los datos recogidos previamente.
5. Analizar y testear con datos reales el correcto funcionamiento de nuestra base de datos.
6. Preparar la implementación de los algoritmos para realizar el estudio de nuestra base de datos.
7. Analizar y testear los algoritmos realizados para obtener el que mejores resultados proporciona.
8. Mejorar la implementación de nuestro programa de análisis con el algoritmo ganador y el total de tweets recopilados.
9. Generar gráficos con los resultados obtenidos del análisis de sentimientos, para poder observarlos de manera más fácil.
10. Escribir la memoria y preparar la presentación oral en función de los resultados y conclusiones obtenidas.

Como hemos comentado en algunos sprints se ha tenido que hacer una replanificación temporal, por falta de recursos, de tiempo o por una planificación inicial incorrecta.
De la planificación inicial se ha retrasado uno de los sprints, en concreto el de Aprendizaje técnico, debido a que aunque originalmente se habían planteado dos semanas de dedicación a esta tarea, a lo largo del proyecto se ha visto que se debían ir asumiendo más conocimientos técnicos de nuevas herramientas, y que por lo tanto era un sprint que ocuparía la mayor parte de la planificación. Pero aun así, al ser una tarea que se podía realizar en paralelo con la resta de tareas no ha penalizado al resto de los sprints.
En la tabla también observaremos que en dos de los sprints se ha conseguido recortar el tiempo de desarrollo, pero esto no significa que las tareas hayan resultado más sencillas. Por ejemplo en el caso del sprint de Implementación de los algoritmos para el estudio de los datos, al ser una tarea con bastante peso, finalmente y debido a ciertas circunstancias, se han podido dedicar más horas de las planteadas al día inicialmente, lo que ha hecho finalizar antes la tarea, aunque las horas invertidas hayan sido mayores de lo planificado.
Debido a esta circunstancia también se ha tenido la posibilidad de dedicar los días de más que han quedado colgados de este sprint, a la siguiente tarea de Test y análisis de los algoritmos para escoger el mejor, lo que ha sido todo un acierto, ya que al ser una tarea muy manual en la que comparamos sobre una cantidad de tweets, los resultados obtenidos con los dos algoritmos respecto a el sentimiento que otorgamos nosotros manualmente para ver la tasa de acierto de estos algoritmos, con la planificación inicial que se había planteado hubiese faltado tiempo de este sprint. Pero al reorganizarlo por la tarea anterior, finalmente se ha podido llevar a cabo a tiempo.

6 DESARROLLO

En este apartado explicaremos más detalladamente que procesos hemos seguido para llegar a nuestro objetivo final, dividido en diferentes partes.

6 .1 ACCESO API DE TWITER

Uno de los primeros pasos, ha sido la creación de una cuenta en Twitter, no como usuario normal, sino como desarrollador para poder llevar a cabo nuestra aplicación. Para esto Twitter requiere al desarrollador autentificarse por medio de OAuth [8], y así obtener la cuenta con la que podremos utilizar la API de Twitter, a través de un acceso token, para tener la cuenta asociada a Twitter, que nos permitirá realizar las diferentes consultas y obtener los datos que deseemos. Esta API nos ofrece en su versión gratuita dos opciones para descargarnos los datos, Search API y Streaming API.
Con Search API tenemos la posibilidad de obtener los Tweets de hasta siete días de margen, con respecto a la fecha en la que se realiza la consulta. Esto a su vez es una de las limitaciones que tiene este tipo de usuarios, ya que si se quiere obtener datos con fecha de más de una semana, Twitter no lo permite a menos que se realice un pago por consulta. Search API nos permite además filtrar por lenguaje y localización. La segunda opción, llamada Streaming API, nos proporciona una conexión permanente con los servidores de Twitter, la cual tiene una restricción en la cantidad de Tweets recibidos que nunca será mayor a 50 Tweets por segundo.
En nuestro proyecto finalmente hemos elegido la primera opción Search API, ya que nos ofrecía la posibilidad de ir recopilando información día a día, el gran inconveniente de este método es que existe una limitación sobre la descarga de información ya que Twitter solo permite descargar 100 Tweets al día para cada Hashtag, esta circunstancia nos ha retrasado bastante, ya que para obtener una cantidad optima de Tweets hemos tenido que ejecutar nuestra aplicación cada día.

6 .2 OBTENCION DE DATOS

Para evitar la redundancia de datos hemos asociado a cada Tweet descargado un identificador, ya que si no lo hiciéramos así, estaríamos repitiendo Tweets y analizando lo mismo todo el tiempo. El periodo de sondeo de Tweets que hemos empleado en nuestra aplicación ha sido la comprendida entre el 5/11/2018 hasta el 31/12/2018. Hemos elegido este periodo ya que en él transcurren varios acontecimientos, como por ejemplo el discurso del rey o el del presidente de la Generalitat, que suelen generar gran controversia entre los adeptos a la red social, y por lo tanto un número de Tweets mayor a analizar. Para saber la opinión de los usuarios de Twitter sobre el tema la independencia de Cataluña y los presos políticos, nos hemos ayudado de diferentes hashtags o etiquetas como son:

- #21D
- #RepublicaCatalana
- #Cataluña
- #Catalunya
- #PresosPoliticos
- #independenciacat
- #Puigdemont

Para realizar nuestro estudio en total nos hemos descargado más de 22.000 Tweets de los cuales, más de 14.000 han sido de ámbito global sobre el territorio de España, y el resto, unos 8.000 aproximadamente, han sido recopilados de manera local, exactamente de las ciudades de Barcelona, Madrid y Sevilla.

Para poder descargar esta gran cantidad de Tweets, hemos elegido hacerlo bajo Nodejs una herramienta que se ha utilizado durante la carrera, y que proporciona un entorno de ejecución multiplataforma, de código abierto, y que funciona bajo el lenguaje de programación JavaScript, que además nos permite acceder a una enorme cantidad de librerías de código abierto, gracias a su gestor de paquetes NPM (Node Package Manager).

6 .3 ALMACENAMIENTO

Para llevar a cabo el almacenamiento tanto de los datos como de los resultados, hemos adquirido el conocimiento de ELK [9], un conjunto de herramientas de gran potencial de código abierto que nos permitirán tratar los datos recopilados de una manera optima y eficaz. Dentro de este conjunto de herramientas contamos con ElasticSearch, un motor de búsqueda y análisis, de código abierto, escalable, y que nos ayuda a manejar una gran cantidad de datos de una manera rápida. Además para poder comprobar este proceso rápidamente y así detectar errores a tiempo, hemos utilizado Postman una herramienta que nos ayuda a realizar peticiones a APIs y así probarlas de una manera rápida y simple. Para poder visualizar toda la información recopilada hemos descubierto que Kibana es la herramienta que mejor se complementa con ElasticSearch. Con esta herramienta podremos exponer los resultados en forma de gráficos e histogramas.

6 .4 ELASTICSEARCH

Seguidamente cuando ya tenemos almacenados y analizados los diferentes Tweets con los diferentes algoritmos de sentimiento, hemos utilizado una herramienta llamada ElasticSearch, un servidor de búsqueda que nos permite indexar y analizar en tiempo real grandes cantidades de datos de manera distribuida, nos permite almacenar documentos ya sean estructurados o no y poder indexar todos los campos de estos documentos casi en tiempo real. Las ventajas que nos proporciona ElasticSearch son:

- Compatibilidad con Java.
- Gran velocidad de respuesta.
- Es distribuido, lo que lo hace fácilmente escalable y adaptable a diferentes situaciones.

Utiliza objetos JSON como respuesta, por lo que es fácil de invocar desde varios lenguajes de programación.

6 .5 ANALISIS DE SENTIMIENTOS

Una vez podemos conectarnos con los servidores de Twitter, y se ha decidido que hashtags queremos utilizar, nos disponemos a elegir el algoritmo de sentimiento. En este momento nos encontramos con dificultades, ya que la mayoría de los algoritmos son en inglés, aun con este inconveniente elegimos dos algoritmos diferentes en castellano, para poder tener varias opciones y poder compararlos y saber cuál es el que menor error nos da. Los algoritmos que estudiaremos serán el Multilang-sentiment y Lorca.js.

6 .6 MULTILANG-SENTIMENT

Nuestra primera opción, consiste en un módulo Nodejs que utiliza las listas de palabras AFINN-165, una lista de palabras calificadas con un valor que comprende entre -5 y +5 según si son positivas o negativas, El análisis de sentimientos se realiza mediante una comprobación cruzada de los tokens de cadena con la lista AFINN, con esto se hace un cálculo matemático sencillo que consiste en la suma del valor de cada palabra dividido por el número total de palabras, dándonos como resultado un valor en concreto.

Emoji	N	Position	p_-	p_0	p_+	$\bar{s}$	Name
😂	14,622	0.80	0.25	0.29	0.47	0.22	face with tears of joy
❤	8,050	0.74	0.04	0.17	0.79	0.75	heavy black heart
💕	7,144	0.75	0.04	0.27	0.69	0.66	black heart suit
😍	6,359	0.76	0.05	0.22	0.73	0.68	smiling face with heart-shaped eyes
😭	5,526	0.80	0.44	0.22	0.34	-0.09	loudly crying face
😘	3,648	0.85	0.05	0.19	0.75	0.70	face throwing a kiss
😊	3,186	0.81	0.06	0.24	0.70	0.64	smiling face with smiling eyes
👌	2,925	0.80	0.09	0.25	0.66	0.56	ok hand sign
💞	2,400	0.76	0.04	0.29	0.67	0.63	two hearts
👏	2,336	0.78	0.10	0.27	0.62	0.52	clapping hands sign

Ilustración 1: Tabla de valores del Ranking Emoji Sentiment.

Este algoritmo también utiliza el Ranking de Emoji

Sentiment, que podemos observar en la *Ilustración 1*, y que consiste en una lista de los emoticonos más utilizados los cuales también tienen otorgados una puntuación y que se contemplarían dentro de dicha ecuación y se contarían como una palabra más.

Al ejecutar este algoritmo nos devuelve un objecto con los siguientes valores:

- **Puntuación**: Valor que tienen las palabras en función de la lista AFINN.
- **Comparativo**: Valor total del texto.
- **Token**: El texto se separa por palabras y cada palabra obtiene una marca.
- **Palabras**: Lista de palabras que se encuentran en la lista AFINN y que por lo tanto tienen puntuación.
- **Positivo**: Lista de palabras catalogadas como positivas.
- **Negativo**: Lista de palabras catalogadas como negativas.

```
var sentiment = require('multilang-sentiment');

var result = sentiment('Cats are totally amazing!', 'en', {
    'words': {
        'cats': 5,
        'amazing': 2
    }
});
console.dir(result);    // Score: 7, Comparative: 1.75
```

Ilustración 2: Ejemplo de ejecución con multilang-sentiment.

6 .7 LORCA.JS

Nuestra segunda opción, Lorca.js, es una librería de PNL en español escrita en javascript. Esta libreria nos permite:

- Separar el texto ya sea por frases, palabras o silabas.
- Detectar oraciones pasivas.
- Frecuencia de usos de las palabras.
- Todas las palabras tienen un valor.
- Se puede obtener el valor sentimental de cada palabra.

Para el calculo global del texto utiliza la misma fórmula que en el algoritmo anterior:

$$\frac{\sum Puntuación\ palabras}{Total\ palabras}.$$

También nos ofrece poder concatenar diferentes métodos de análisis para un mismo texto. Lorca.js realiza una traducción semiautomática de la lista original AFINN, se calcula el valor relativo de cada sentencia devolviendo unos valores, en caso de que el resultado sea más grande que 0 nos daría un sentimiento positivo y en caso contrario seria de carácter negativo.

Una vez analizado los tweets almacenaremos los resultados en formato **JSON**, un formato ligero de intercambio de datos basado en un subconjunto del Lenguaje de Programación JavaScript. En estos archivos, guardaremos los campos más importantes que son:

- _id: Identificador del tweet.
- created: Fecha de creación del tweet.
- full_text: El texto que contiene el tweet.
- location: Desde donde se ha escrito el tweet.
- text: La etiqueta que se ha utilizado.
- file: Nombre del archivo que guardamos.
- multilangsentiment: Puntuación otorgada por el algoritmo multi-lang.
- lorca: Puntuación otorgada por el algoritmo lorca.

6 .8 COMPARATIVA ALGORITMOS

El siguiente paso en nuestro proyecto una vez teníamos almacenados unos 14.000 Tweets fue la de comparar manualmente el resultado obtenido de los diferentes algoritmos, para ello hemos analizado manualmente más de 100 Tweets, este proceso ha sido también bastante costoso debido a que hemos hecho la lectura de los 100 Tweets y dándole nosotros mismos la connotación al texto. Cuando se empezó el proceso de comparación nos dimos cuenta que el algoritmo que más coincidía con nuestra valoración ha sido Lorca.js, si bien es cierto que los dos utilizan la misma fórmula matemática Lorca.js nos otorga valoraciones más aproximadas.
En el apartado "A2. Comparativa" se puede observar una pequeña tabla comparativa del análisis manual entre Lorca y Multilang-sentiment.
Una vez elegido el algoritmo, nos propusimos a realizar una búsqueda del mismo tema pero de una manera local, más concretamente queríamos saber que opiniones tienen los usuarios de las ciudades de Barcelona, Madrid y Sevilla.

6 .9 RECOPILACION TWEETS LOCALES

El siguiente paso fue la recopilación de Tweets de las diferentes ciudades y analizarlos, para llevar a cabo este proceso nos centraremos en el campo de localización. En este apartado descubrimos que el campo location es un campo que puede rellenarse de manera automática, donde mediante GPS Twitter puede saber desde que punto se ha generado el Tweet, pero también es posible rellenarlo de manera manual y le da la oportunidad a los usuarios de escribir la localidad que ellos deseen ya sea real o no, por eso hemos encontrado casos donde la localización no correspondía a una ciudad real, un ejemplo claro seria, campo location = "EnMiCasa".

6 .10 POSTMAN

Para poder realizar las querys necesarias nos hemos ayudado de la tecnología de Postman, esta tecnología esta compuesta por diferentes herramientas y utilidades gratuitas, que nos permite realizar diferentes tareas dentro del mundo de las API REST, como son la creación de peticiones a APIs o elaborar test para validar los comportamientos. Esta herramienta la hemos utilizado para hacer peticiones a la API de ElasticSearch y así poder generar colecciones de peticiones para probarlas de una manera rápida y sencilla, así no tendremos la necesidad de tener que ejecutar toda nuestra aplicación, para saber si la query esta bien realizada, o si bien nos devuelve los datos que nosotros realmente queremos, para finalmente añadirlo a nuestra aplicación, gracias a esta tecnología podemos realizar todas las consultas que queramos. Sin duda ha sido una herramienta que nos ha ayudado en nuestro desarrollo del proyecto. A través de una conexión permanente mediante el puerto 9200, hemos podido realizar pruebas sobre ElastichSearch generando colecciones y testeándolas, sin la necesidad de tener que ejecutar todo el código y la pérdida de tiempo que ello conlleva.

6 .11 VISUALIZACION DE LOS RESULTADOS

Nuestro último paso para alcanzar el objetivo final es poder visualizar los resultados finales de nuestra investigación, nos hemos ayudado de Kibana una herramienta que nos permite visualizar y explorar datos que se encuentran indexados en ElasticSearch, es un complemento perfecto en la que podremos creas nuestros diferentes diagramas, además la ventaja que nos ofrece este conjunto de herramientas es que la visualización y análisis se realiza en tiempo real bajo los datos que dispongamos en ElasticSearch, es decir si actualizáramos nuestros datos automáticamente Kibana actualizaría los diferentes diagramas que se vieran afectados por incorporación de los nuevos datos.

7 RESULTADOS

En este apartado mostraremos los resultados obtenidos de los diferentes análisis realizados sobre la opinión que tienen los usuarios de Twitter acerca de la independencia de Cataluña y los presos políticos, bajo el algoritmo de sentimientos que hemos seleccionado, que en este caso ha sido Lorca.js.

7.1 TWEETS GLOBALES

En este primer gráfico vemos que las publicaciones que generan los diferentes usuarios del territorio español tienden a ser negativos, la diferencia es de poco más del 9% con respecto a los comentarios positivos y del 12% respecto a los neutrales, así que podemos afirmar que la tendencia en España según el análisis de sentimientos de Lorca va a ser de carácter negativo.

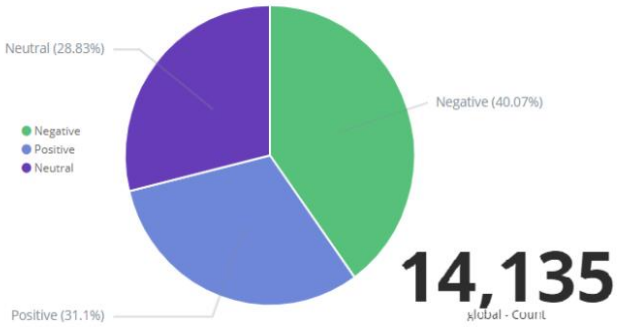

Gráfico 1: Resultados análisis tweets globales.

7.2 RESULTADOS BARCELONA

En este *Gráfico 2* se muestran los resultados de analizar los Tweets de los diferentes usuarios que han realizado sus publicaciones desde la localidad de Barcelona, como se puede observar los comentarios que se generan en esta ciudad son de carácter negativo y es que casi el 50% de ellos lo son, además la diferencia con respecto a los comentarios positivos y negativos supera el 20% en ambos casos.

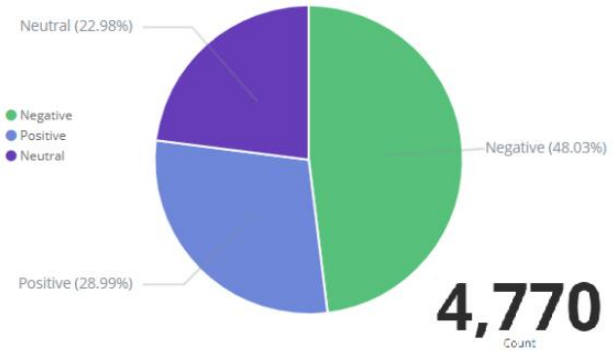

Gráfico 2: Resultados análisis tweets de Barcelona.

7.3 RESULTADOS MADRID

En este *Gráfico 3*, mostramos los resultados obtenidos al analizar las publicaciones que se generan desde la ciudad de Madrid, claramente volvemos a ver aunque con menos porcentaje que en la localidad anterior que las publicaciones vuelven a ser claramente negativas con diferencias que superan el 8% una diferencia menor con respecto a Barcelona pero también hay que tener en cuenta que la cantidad recopilada en Madrid ha sido un poco menor, y ese factor afecta de manera directa a los resultados.

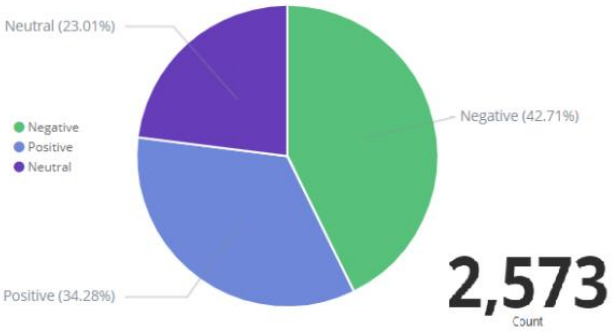

Gráfico 3: Resultados análisis tweets de Madrid.

7.4 Resultados Sevilla

En este *Gráfico 4*, se muestran los resultados de realizar el mismo estudio que en los dos casos anteriores, pero esta vez hemos querido saber la opinión de los usuarios que publican desde Sevilla, lo primero en los que no fijamos es que este caso es bastante diferente respecto a Barcelona y Madrid y es que en este caso se puede observar que la mayoría de publicaciones analizadas son de carácter neutral, más del 60% de los textos concretamente, la diferencia con las publicaciones catalogadas como negativas es mayor al 35% y con respecto a las positivas más del 45%, una gran diferencia con respecto a las otras dos ciudades.

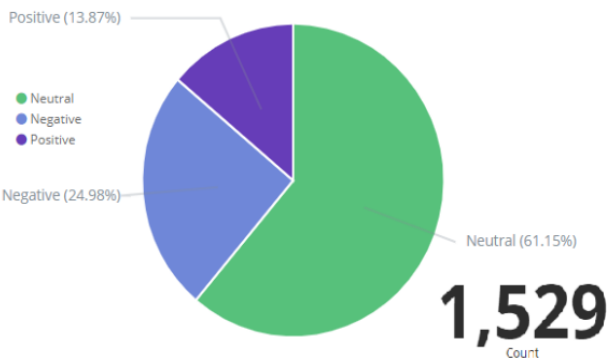

Gráfico 4: Resultados análisis tweets Sevilla.

7.5 Resultados temporales

En este apartado se podrán observar los gráficos obtenidos a partir de los Tweets generados en fechas concretas, y como los sucesos ocurridos en ciertos días han ocasionado un aumento de tweets y de opiniones, respecto a los días en los que no se ha producido ninguno de estos eventos.

En el *Gráfico 5* podemos observar la cantidad de Tweets que se han ido generando en todo el estado español sobre la independencia de Cataluña entre el 5 de noviembre de 2018 y el 5 de enero de 2019.

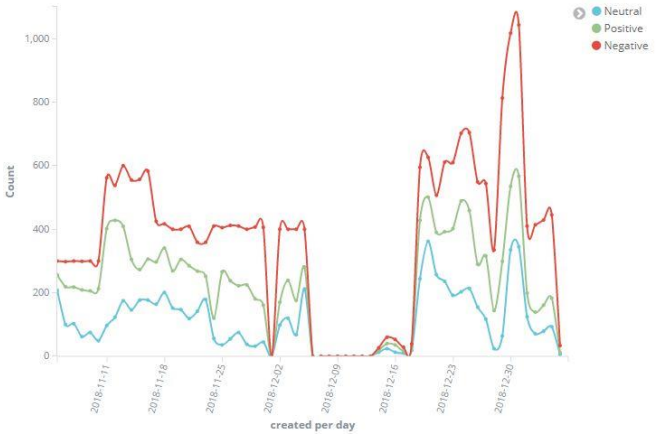

Gráfico 5: Resultados análisis temporal tweets Globales.

Como podemos observar en el gráfico la tendencia marca un gran aumento el día 11 de noviembre de 2018, esto es debido al acontecimiento que ocurrió en Tv3 con la entrevista que le hicieron a Inés Arrimadas, líder de Cs en Cataluña, y que causo bastantes opiniones y discusiones a causa del tema que estamos tratando. La cantidad de Tweets que se generan se mantiene estable hasta llegar a diciembre donde se comentan muy pocos Tweets. Esto es debido a que el tema que había de actualidad eran las elecciones que se produjeron en Andalucía y no tanto el asunto que estamos investigando. Por el contrario, el 2 de diciembre se puede observar como existe un aumento de los comentarios producidos seguramente por los resultados electorales en Andalucía. Después observamos como hay un periodo donde casi no hay comentarios hasta que va llegando uno de los principales acontecimientos que se producen en Cataluña, exactamente el 21D, y es que las intenciones de colapsar la ciudad con coches por parte del grupo CDR generó grandes opiniones en Twitter. Los comentarios van surgiendo en mayor o menor medida hasta el 28 de diciembre, ese día la gente hablaba más sobre el juicio de Ana botella y siete dirigentes más por malversación de pisos. A partir de ese día los Tweets que se van publicando van aumentando, debido al gran acontecimiento que sucede en Cataluña, el discurso del presidente de la Generalitat, este acontecimiento es el que claramente produce más comentarios en Twitter.

En el *Gráfico 6* podemos observar los Tweets que se han ido generando en Barcelona a lo largo de los días desde el 15 de diciembre de 2018 al día 5 de enero de 2019. Lo primero que podemos observar es que el primer aumento se produce el día 20 de diciembre debido a que ese día se produjo un acontecimiento importante, que consistía en la reunión del presidente de la Generalitat con el homólogo español Pedro Sanchez, lo cual creo que generó gran debate en Twitter. El siguiente día que vemos que hay gran cantidad de publicaciones sobre el tema que estamos estudiando es el día 23 de diciembre, donde las publicaciones de diferentes estrellas del espectáculo manifestaron ciertos comentarios sobre la libertad de

expresión.

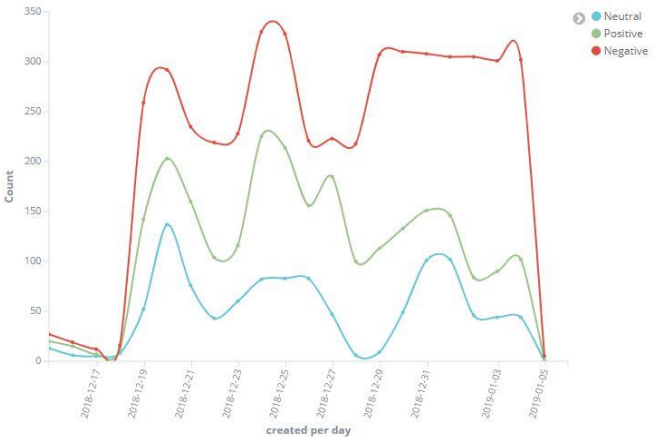

Gráfico 6: Resultados análisis temporal tweets Barcelona.

Finalmente el siguiente aumento se produce en el mismo período que se ha explicado en el gráfico anterior, el discurso navideño del presidente de la Generalitat generó gran cantidad de publicaciones sobre Cataluña.

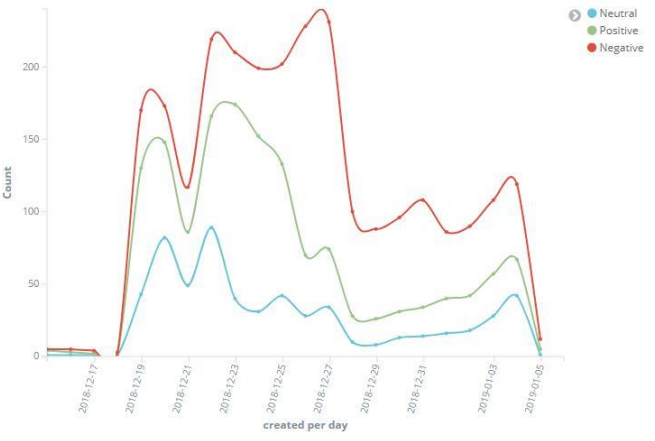

Gráfico 7: Resultados análisis tweets en Madrid.

En el *Gráfico 7* podremos ver la cantidad de Tweets que se van generando en el mismo periodo de tiempo, pero en esta ocasión en la ciudad de Madrid, las subidas y bajadas se van produciendo conforme ocurren los diferentes acontecimientos sociales ya mencionados en los dos gráficos anteriores los resultados obtenidos son muy similares a las de Barcelona en cuanto a publicaciones generadas.

En el *Gráfico 8* podremos observar la cantidad de Tweets que se han ido generando a lo largo del mismo periodo de tiempo, pero en esta ocasión de la ciudad andaluza de Sevilla, se puede observar que existen unas subidas y bajadas de publicaciones muy similares a las de Barcelona y Madrid, salvo que en este caso el volumen general de publicaciones es bastante menor.

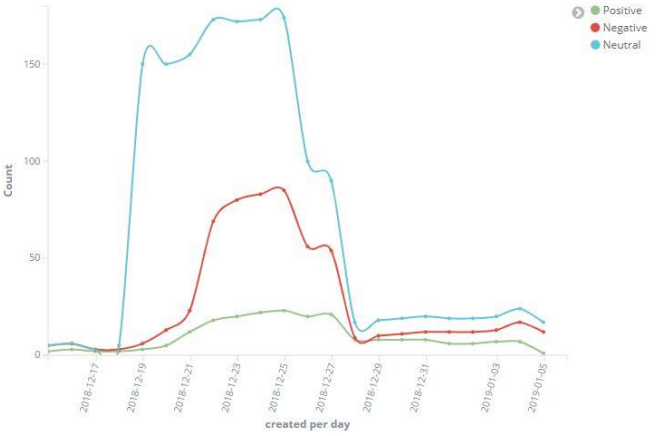

Gráfico 8: Resultados análisis temporal tweets en Sevilla.

En el ***G**ráfico 9* podemos observar la tendencia que tienen las valoraciones que realiza nuestro algoritmo de sentimiento Lorca.js acerca de los Tweets a lo largo del tiempo desde el 5 de noviembre de 2018 hasta el 5 de enero de 2019.

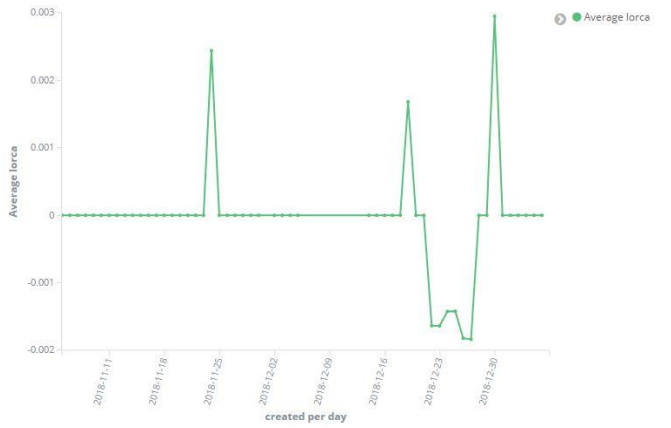

Gráfico 9: Resultados tendencia Lorca.js.

Podemos ver como la nota predominante es el carácter neutro, excepto en fechas señaladas como el 25 de noviembre, el 19 y 30 de diciembre que en estos días predominaron las publicaciones de carácter positivo, seguramente promovidos por los diferentes acontecimientos mencionados anteriormente, cabe destacar que entre el periodo del 21 al 29 de diciembre, el sentimiento que predomino fue de carácter negativo, seguramente debido también a ciertos acontecimientos que se produjeron a raíz de manifestaciones y discursos que ya se comentaron anteriormente.

8. CONCLUSIONES

Una de las primeras conclusiones a las que se ha llegado es la consideración de Twitter no solo como una red social, también como un canal de comunicación ideal para fomentar el debate e incluso para movilizar a sus

usuarios, en varios tweets recopilados, y en el proceso de leer manualmente los tweets, pude observar diferentes textos, donde el mensaje que se quería transmitir eran el de realizar convocatorias de manifestaciones, además de observar que la gran mayoría pretende generar discusión. Esto es debido al comportamiento humano y es que la gran mayoría cuando escriben su opinión y la publican, no todos los demás usuarios coinciden, y estos responden con una opinión diferente generando así el debate. Después de observar los resultados en el apartado anterior podemos llegar a dos conclusiones. La primera conclusión que podemos obtener sobre la opinión que tiene los usuarios de Twitter a cerca de la independencia de Cataluña y los presos políticos es de carácter negativo y neutro, también se puede observar que Madrid es la ciudad que más mensajes positivos genera, en su contra en Barcelona, la opinión mayoritaria ha sido negativa. En general se puede observar que el texto que publican los diferentes usuarios de Twitter no contienen palabras en las que la valoración del texto sea en general de carácter positivo, pero esto no significa que la intencionalidad del mensaje sea negativo, y es que llegados a este punto nuestra segunda conclusión es que una cosa es la valoración del texto por palabras separadas y otra es la intencionalidad y el contexto en las que están escritas. Este es el gran problema del análisis de sentimientos.

LINEAS FUTURAS DE MEJORAS

En la actualidad según el criterio de muchas empresas y personas la información es poder, y es que saber la opinión que tienen los usuarios acerca de un tema en concreto, acontecimiento, o sobre algún producto, nos puede ayudar a generarnos una idea de cuales son los gustos, pensamientos que tienen muchos usuarios. El problema que nos surge y a la vez seria una mejora, es la de poder tener controlado la ironía, el sarcasmo, la ambigüedad, y sobretodo saber bajo que contexto está reflejado ese texto. Si bien no se puede llegar a una solución definitiva, ya que como hemos mencionado anteriormente ni las personas físicas son capaces de ponerse de acuerdo a la hora de clasificar textos en categorías, y es que en mi humilde opinión creo que la escala de categorías positivo, negativo y neutra quedará obsoleta en próximos años. Los nuevos algoritmos de sentimiento podrán clasificar ciertos estados de ánimos como pueden ser la alegría, la tristeza, el enfado o incluso el miedo, según las palabras que vayamos escribiendo. Además en el mundo tecnológico que nos rodea ya existen sistemas capaces de analizar la voz y sistemas de reconocimiento facial que por supuesto ayudaran a construir sistemas totalmente personalizados y que se adaptaran mucho mejor a nuestra manera de pensar. Bajo mi punto de vista los sistemas nos conocerán tan bien que serán capaces de saber si lo que estamos publicando lo hacemos de una manera positiva o negativa. No cabe duda que el futuro que nos espera respecto a estos algoritmos de sentimiento, es que conforme vayan pasando los años, se vayan sofisticando los algoritmos y que se puedan complementar con otros sistemas de análisis llegara a tener una precisión con la que podamos generar estudios más fieles a la realidad que se vive hoy en día.

AGRADECIMIENTOS

En primer lugar me gustaría poder agradecer el tiempo dedicado a mi tutor Jordi Casas Roma a lo largo de todo el proyecto respondiendo a mis dudas, con reuniones asiduas para poder resolver todas las dudas que me han surgido y solventándolas rápidamente para llegar a los objetivos finales. En segundo lugar agradecer a mi familia, en especial a mi pareja Raquel Fernández, por aguantarme en los momentos más difíciles cuando me bloqueaba y sobretodo en los momentos felices cuando se avanzaba. Finalmente no querría olvidarme de mis amigos y todos los que me han dado soporte, se han ido interesando de los avances que llevaba sobre el proyecto y sobretodo aportando ideas que si bien creía que eran descabelladas si me ayudaron para ciertos momentos de bloqueo. A todos, ¡Muchas gracias!

BIBLIOGRAFÍA

[1] I. Twitter, «Es lo que está pasando.,» [En línea]. Available: https://twitter.com/?lang=es. [Último acceso: Septiembre 2018].

[2] Brandwatch, «Brandwatch Analytics: Social Media Listening Platform,» [En línea]. Available: https://www.brandwatch.com/analytics/features/. [Último acceso: Octubre 2018].

[3] ProyectosAgiles.org, «Qué es Scrum,» [En línea]. Available: https://proyectosagiles.org/que-es-scrum/. [Último acceso: 2018 Septiembre].

[4] Atlassian, «Trello Official Page,» [En línea]. Available: https://trello.com/. [Último acceso: Octubre 2018].

[5] Elasticsearch B.V., «Elasticsearch: RESTful, Distributed Search & Analytics,» [En línea]. Available: https://www.elastic.co/products/elasticsearch. [Último acceso: 2018 Octubre].

[6] F. Redondo, «Postman: gestiona y construye tus APIs rápidamente,» [En línea]. Available: https://www.paradigmadigital.com/dev/postman-gestiona-construye-tus-apis-rapidamente/. [Último acceso: 2018 Diciembre].

[7] Elasticsearch B.V., «Getting Started with Kibana,» [En línea]. Available: https://www.elastic.co/webinars/getting-started-kibana. [Último acceso: 2019 Enero].

[8] A. Parecki, «OAuth Community Site,» [En línea]. Available: https://oauth.net/. [Último acceso: Octubre 2018].

[9] Elasticsearch B.V., «What is the ELK Stack?,» [En línea]. Available: https://www.elastic.co/elk-stack. [Último acceso: Octubre 2018].

APENDICE

A1. TABLA PLANIFICACIÓN

A continuación podemos observar la tabla con la planificación inicial planeada, y la que finalmente se llevo a cabo con los cambios especificados en el apartado "*5. Planificación*".

Tarea	Sprint Planificado	Sprint Realizado
Estado del Arte	24/09 – 30/09	24/09 – 30/09
Aprendizaje técnico	1/10 – 15/10	1/10 – 23/12
Implementación de una base de datos	16/10 – 18/11	16/10 – 18/11
Test i análisis de la base de datos	01/11 – 19/11	01/11 – 14/11
Implementación de los algoritmos para el estudio de los datos	20/11 – 22/12	20/11 – 14/12
Test i análiss de los algoritmos para escoger el mejor	22/12 – 05/01	15/12 - 05/01
Implementación final de nuestro programa analitico	07/01 – 13/01	07/01 – 13/01
Memoria i Presentación	14/01 – 27/01	14/01 – 27/01

Tabla 1. Planificación de sprints del proyecto.

A2. COMPARATIVA

En la tabla siguiente podemos observar algunos resultados obtenidos del análisis manual de 100 Tweets, donde podemos encontrar diferentes casuísticas que detallaremos a continuación.

Tweet	Multilang	Lorca	Manual	Sentimiento
Cs presenta mociones-chapuza para que a los #PresosPolíticos se les pueda denegar el indulto.	0 (Neutro)	0 (Neutro)	Negativo	FAVOR
El #9N voté NO y NO.El #1oct voté SI.Cómo yo, miles de personas. La #RepublicaCatalana suma más cuando se desobedece…	-2 (Negativo)	0.0625 (Positivo)	Positivo	FAVOR
Enorme ovación y gritos de #Llibertat en recuerdo a los presos políticos en el #Kursaal no estáis solos.	-1 (Negativo)	- 0.3294117 (Negativo)	Positivo	FAVOR
Jean-Luc Mélenchon y su partido piden la liberación de los presos catalanes y la organización de un referéndum.	0 (Neutro)	- 0.0952380 (Negativo)	Positivo	FAVOR
¿Supongo que Cataluña debería negociar su independencia con Aragón, Valencia, Andorra y Occitania de modo preferente no? Sr. @sanchezcastejon Tal y como exige para Gibraltar.	2 (Positivo)	0.3831932773109244 (Positivo)	Negativo (Irónico)	CONTRA
El Barça permite una pancarta gigante donde	0 (Neutro)	0 (Neutro)	(Neutro)	NEUTRO

se pide abiertamente la independencia de Cataluña				

Tabla 2. Comparativa extraida del analisis manual de los algoritmos utilizados.

En esta muestra de la tabla comparativa, he querido exponer en varios casos determinados los diferentes problemas a los que nos enfrentamos a la hora de utilizar los diferentes algoritmos de sentimientos, así como poder observar cual de los dos algoritmos era más fiable, siempre desde mi punto de vista personal. Esta muestra refleja que aunque existen Tweets que los algoritmos evalúan como negativos, en realidad no lo son, esto es debido a que la gran mayoría de usuarios escriben de una manera cáustica, la gran mayoría utiliza la ironía para expresar sus opiniones, incluso comentarios que son considerados positivos estos en su contexto, no necesariamente son a favor de la independencia de Cataluña, por eso es importante también saber en que contexto están escritas las palabras.