

# Herramienta para el análisis de la opinión en tweets periodísticos

Lage García, Lola

Curso 2013-2014

Director: Horacio Saggion

GRADO EN INGENIERÍA INFORMÁTICA



Universitat  
Pompeu Fabra  
Barcelona

Escola  
Superior Politècnica

## Trabajo de Fin de Grado



Quiero dedicar este proyecto en primer lugar a mis padres, con los que no puedo compartir la alegría de finalizar mis estudios. Sé que estarían muy orgullosos de mí.

A mis hermanos por el apoyo y los ánimos que he recibido siempre de ellos tanto en mis estudios como en el resto de mi vida personal.

A mis sobrinos, con los que no he pasado mucho tiempo en los últimos años debido a la distancia que nos ha separado, pero que siempre tengo en mi mente.

A mi gran amigo Alfonso Bayard. Aunque ya no está entre nosotros, sé que le hubiese alegrado mucho verme acabar la carrera.

Y especialmente a mi novio Luis por su apoyo y ayuda durante el último año de carrera, y sobre todo por su paciencia y comprensión durante las últimas semanas del desarrollo de este trabajo. Sin duda sin él todo hubiese sido más complicado.



## Agradecimientos

Quiero agradecer en primer lugar a mi tutor del proyecto, Horacio Saggion, por haberme permitido realizar este trabajo con él, sobre todo teniendo en cuenta que la distancia que nos ha separado durante la realización del mismo ha sido de casi 600 kms. El haberme dado la opción de comunicarnos mediante e-mail y videoconferencia me ha permitido volver a mi ciudad natal antes de lo previsto. Además, debo agradecerle también su diligencia a la hora de contestar a mis mensajes y de resolver mis dudas.

Por otro lado, quería agradecer también a diferentes grupos con los que me he puesto en contacto vía e-mail durante la ejecución del proyecto y que me han facilitado los recursos utilizados. Concretamente, a Elhuyar Fundazioa por haberme proporcionado la lista de interjecciones y el léxico de polaridad; a las autoras del estudio “*Análisis lingüístico de expresiones negativas en tweets en español*” por enviarme la lista de emoticonos etiquetados con polaridad; y a los organizadores del Taller de Análisis de Sentimientos de la Sociedad Española para el Procesamiento del Lenguaje Natural, por haberme proporcionado los corpus de entrenamiento y de test, así como por hacer público los estudios que los participantes realizaron para dicho taller. Y en general, a todas aquellas personas que se dedican a la investigación y comparten su conocimiento con el resto del mundo.

Por último, agradecer a todos los profesores que he tenido por todo lo que me han enseñado y por su disponibilidad cuando ha sido necesario, así como a todos mis compañeros a los que siempre recordaré con cariño, especialmente a Daniel Naro e Iván Latorre por su compañerismo y su generosidad.

Muchas gracias a todos.



## **Resumen**

Durante la última década la cantidad de información subjetiva existente en Internet ha crecido exponencialmente. Debido a la aparición de las redes sociales, así como la proliferación de blogs y sitios web donde los usuarios pueden generar contenido, Internet se ha convertido en una fuente de información aún más valiosa de lo que era. Este nuevo modelo conocido como web 2.0 ha despertado un gran interés en diferentes sectores.

La comunidad científica encuentra una tarea interesante, desde el punto de vista del procesamiento del lenguaje natural, el poder extraer automáticamente las opiniones y sentimientos de los mensajes volcados por los usuarios en la red.

Siguiendo esta línea de investigación, en este trabajo se ha creado una herramienta capaz de realizar el análisis de sentimientos en mensajes escritos en Twitter. Concretamente, se ha centrado en tweets periodísticos sobre diferentes temas de actualidad política escritos por cuatro de los más importantes periódicos nacionales.

## **Abstract**

Over the last decade the amount of subjective information on the Internet has grown exponentially. Due to the emergence of social networks, as well as the proliferation of blogs and web sites where users can generate content, Internet has become a source of even more valuable information than it was. This new model, known as web 2.0, has aroused great interest in different sectors.

From the point of view of natural language processing, the scientific community find it an interesting task to obtain automatically the opinions and feelings of the messages posted by users on the network.

Following this line of research, in this work it has been created a tool able to perform sentiment analysis on tweets. In particular, we have focused on journalistic tweets on topics of current policy written by four of the major national newspapers.





## Prólogo

En cada curso de la carrera se tratan diferentes asignaturas, habiendo siempre alguna que en un curso te atrae más que el resto. Según se acerca la hora de realizar el trabajo de fin de grado te vienen a la cabeza todas estas materias y es difícil decidirse por un tema concreto. En mi caso la idea surgió al cursar durante el último curso la asignatura optativa de Aplicaciones Inteligentes para la web.

En esta asignatura aprendí entre otras cosas, el funcionamiento de los sistemas recomendadores que he utilizado tantas veces en Internet, o cómo realizar resúmenes automáticos de textos. El contenido de la materia era muy práctico y además de la teoría implementamos pequeños sistemas que me parecieron muy interesantes.

Por este motivo decidí realizar mi trabajo final de grado sobre una idea que ya se había trabajado en otros proyectos anteriores, el análisis de sentimientos de mensajes de Twitter. Mi profesor me recomendó unir este interés con algún tema que me atrajera. Debido a que me gusta leer la prensa, y principalmente las noticias relacionadas con la política, decidimos realizar una herramienta para el análisis de sentimientos en tweets y aplicarla sobre aquéllos escritos por diferentes periódicos nacionales. La idea me gustó y allí nació la semilla de este proyecto.



# Índice

	Pág.
Resumen .....	vii
Prólogo .....	ix
Listado de figuras .....	xiii
Listado de tablas .....	xv
1. INTRODUCCIÓN .....	1
1.1 Contexto.....	1
1.2 Objetivo .....	2
1.3 Estructura del documento .....	3
2. ESTADO DEL ARTE .....	5
2.1 Generación de léxicos de polaridad .....	6
2.2 Análisis en redes sociales .....	6
2.3 Normalización de textos .....	7
3. RECURSOS UTILIZADOS .....	9
3.1 Corpus .....	9
3.2 Listado de abreviaturas .....	11
3.3 Léxico de polaridad .....	11
3.4 Listado de interjecciones .....	13
3.5 Listado de emoticonos .....	13
4. HERRAMIENTAS UTILIZADAS .....	15
4.1 Twitter4j .....	15
4.2 Freeling .....	15
4.3 WEKA .....	17
5. ARQUITECTURA DEL SISTEMA .....	19
6. RECOLECCIÓN DE TWEETS .....	21
7. ANÁLISIS – CLASIFICACIÓN .....	23
7.1 Algoritmo .....	23
7.2 Carga de recursos .....	24
7.3 Procesamiento del corpus .....	24
7.4 Generación de instancias .....	25
7.5 Transformación de instancias .....	26
7.6 Generación y evaluación del clasificador .....	28
7.7 Procesamiento de tweets .....	30
7.8 Clasificación de tweets .....	30

	Pág.
8. TEMÁTICA DE LOS DATOS ANALIZADOS .....	31
9. VISUALIZACIÓN DE RESULTADOS .....	35
9.1 Generación de datos .....	35
9.2 Estructura de la web .....	35
9.3 Representación gráfica de la información .....	38
10. ANÁLISIS DE LOS RESULTADOS OBTENIDOS .....	41
11. DISCUSIÓN .....	49
11.1 Problemas encontrados .....	49
11.2 Trabajo futuro .....	49
11.3 Conclusiones .....	50
Bibliografía .....	53
Anexo 1. Temas de actualidad tratados .....	55
Anexo 2. Gráficas de los resultados obtenidos .....	57

## Lista de figuras

	Pág.
Fig. 1. Tweet extraído del corpus .....	10
Fig. 2. Muestra del fichero de abreviaturas .....	11
Fig. 3. Muestra del fichero léxico de polaridad .....	12
Fig. 4. Muestra del fichero de interjecciones con polaridad .....	13
Fig. 5. Muestra del fichero de emoticonos .....	14
Fig. 6. Demo on-line de Freeling .....	17
Fig. 7. Interfaz gráfica de WEKA .....	18
Fig. 8. Arquitectura del sistema. Parte 1 .....	19
Fig. 9. Arquitectura del sistema. Parte 2 .....	20
Fig. 10. Fichero de configuración del módulo Crawler .....	21
Fig. 11. Fichero XML con los tweets recolectados por el módulo Crawler .....	22
Fig. 12. Ejemplo de fichero ARFF .....	26
Fig. 13. Ejemplo de fichero ARFF filtrado .....	27
Fig. 14. Resumen de la evaluación del clasificador .....	29
Fig. 15. Página principal de la interfaz web .....	36
Fig. 16. Sección “Proyecto” de la interfaz web .....	37
Fig. 17. Sección “Contacto” de la interfaz web .....	37
Fig. 18. Representación gráfica de la información en la interfaz web .....	38
Fig. 19. Código HTML necesario para embeber un tweet en la web .....	39
Fig. 20. Resultado obtenido sobre el tópico Conflicto de Ucrania .....	41
Fig. 21. Ejemplo de tweets etiquetados correctamente como <i>Positivos</i> en el tópico Conflicto de Ucrania .....	42
Fig. 22. Ejemplo de tweets etiquetados correctamente como <i>Negativos</i> en el tópico Conflicto de Ucrania .....	43
Fig. 23. Ejemplo de tweets que no siguen el patrón de aparición en el léxico en el tópico Conflicto de Ucrania .....	43
Fig. 24. Ejemplo de tweets clasificados en la categoría <i>Ninguno</i> en el tópico Conflicto de Ucrania .....	44
Fig. 25. Resultado obtenido sobre el tópico Abdicación del Rey Juan Carlos I .....	44
Fig. 26. Ejemplo de tweets clasificados correctamente como <i>Positivos</i> en el tópico Abdicación del Rey Juan Carlos I .....	45
Fig. 27. Ejemplo de tweets clasificados correctamente como <i>Negativos</i> en el tópico Abdicación del Rey Juan Carlos I .....	45
Fig. 28. Ejemplo de tweet clasificado como <i>Neutro</i> en el tópico Abdicación del Rey Juan Carlos I .....	46
Fig. 29. Ejemplo de tweets clasificados en la categoría <i>Ninguno</i> en el tópico Abdicación del Rey Juan Carlos I .....	46
Fig. 30. Ejemplo de tweets clasificados incorrectamente en el tópico Abdicación del Rey Juan Carlos I .....	47
Fig. 31. Ejemplo de tweets que incluyen negación clasificados incorrectamente en el tópico Abdicación del Rey Juan Carlos I .....	47



## Lista de tablas

	Pág.
Tabla 1. Resumen del contenido del corpus .....	9
Tabla 2. Distribución de sentimiento en el corpus .....	10
Tabla 3. Definición de la tabla que contiene los tweets en la BBDD .....	30
Tabla 4. Cantidad de mensajes recolectados de Twitter .....	31
Tabla 5. Ranking de hashtags citados por los periódicos ABC y El País .....	31
Tabla 6. Ranking de hashtags citados por los periódicos El Mundo y La Razón ...	32
Tabla 7. Definición de la tabla Tópicos en la BBDD .....	34
Tabla 8. Contenido parcial de la tabla Tópicos de la BBDD .....	34
Tabla 9. Número de tweets identificados por tópico .....	34





# 1. INTRODUCCIÓN

## 1.1 Contexto

Durante los últimos años, debido al desarrollo de la llamada Web 2.0, la cantidad de información subjetiva disponible en Internet ha crecido exponencialmente. Cada vez son más los sitios en los que los usuarios pueden crear y compartir contenido. Las redes sociales concretamente, son una fuente muy valiosa de información ya que aglutina gustos, preferencias y opiniones de millones de usuarios alrededor del mundo. Del mismo modo, miles de sitios permiten hoy en día que los usuarios aporten sus opiniones y valoraciones sobre productos o servicios.

Conseguir obtener esta información subjetiva que se encuentra repartida por la web es una tarea interesante desde el punto de vista del procesamiento del lenguaje natural, pero además es una tarea de gran interés y con un gran valor para las empresas (Saggion and Funk, 2009) y el poder político. Hasta hace unos años, las empresas no tenían forma de saber lo que los compradores opinaban sobre sus productos o servicios si no era a través de encuestas u observando las ventas realizadas. Hoy en día, gracias a internet, esta información está al alcance de cualquiera ya que los usuarios continuamente expresan sus opiniones públicamente e incluso se dejan influir por las mismas para tomar sus decisiones de compra. Igualmente, los partidos políticos pueden utilizar esta información para conocer la opinión de los ciudadanos sobre diversos temas, e incluso para predecir la intención de voto frente a unas elecciones.

El mundo de la publicidad también ha dado un giro completo. Las empresas hoy prefieren invertir dinero en comprar espacios de publicidad en sitios donde saben de antemano que los usuarios pueden estar interesados en los productos o servicios que ofrecen. De hecho, la prensa ha perdido gran parte de sus ingresos por publicidad por este motivo (Larrañaga, 2010). Es obvio que para una compañía es más eficaz invertir en publicidad en sitios donde los usuarios expresan buenas opiniones sobre ella o que muestran interés sobre productos o servicios similares a los que ellos ofrecen, que en sitios donde se da el hecho contrario.

La minería de opiniones o análisis de sentimientos (AS) se enmarca dentro del procesamiento del lenguaje natural (PLN), y se refiere a la aplicación de este último para extraer la información subjetiva que se encuentra en un texto, de modo que éste pueda ser clasificado como positivo o negativo. Existen numerosos estudios, algoritmos y técnicas que se centran en este tipo de análisis.

De un tiempo a esta parte ha ido tomando especial interés el hecho de poder adaptar estos algoritmos para poder aplicarlos sobre textos relativamente cortos y con ciertas características especiales. Esto es debido a la importancia de las opiniones escritas por

los usuarios en redes sociales como Twitter, en la que los mensajes tienen una limitación de longitud de 140 caracteres.

Con la introducción de los mensajes SMS en los móviles y la limitación de caracteres que existía, los usuarios se fueron acostumbrando a utilizar abreviaturas a la hora de escribir. Esta costumbre sigue existiendo actualmente en el uso de aplicaciones de mensajería instantánea, o en mensajes vertidos en redes sociales, sobre todo cuando existe limitación en el número de caracteres como ocurre en Twitter. Por otro lado, el uso extendido de emoticonos, así como la costumbre de repetir letras innecesariamente para expresar cierto énfasis, o la existencia común de errores ortográficos, ha convertido el lenguaje utilizado en un lenguaje “especial” que debe ser pre-procesado antes de poder ser analizado con los algoritmos y técnicas convencionales. Además, Twitter particularmente, utiliza ciertos caracteres especiales que tienen un significado específico en dicha red social. Por ejemplo, el símbolo @ se antepone al nombre de un usuario para mencionarle o el símbolo # precediendo a una palabra (hashtag) indica que es un tema que se está tratando en la red.

La Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)<sup>1</sup> celebra desde hace más de veinte años un congreso anual. Éste tiene como finalidad potenciar el desarrollo de las diferentes áreas de estudio relacionadas con el procesamiento del lenguaje natural (PLN), así como ayudar a la difusión de las investigaciones que lleva a cabo la comunidad científica y exponer las posibles aplicaciones reales en este campo.

Tanto es el interés suscitado por el tema del análisis de sentimientos y de reputación online en redes sociales, que en 2013 se ha realizado la segunda edición del evento satélite a dicho Congreso “Taller de Análisis de Sentimientos en la SEPLN” (TASS)<sup>2</sup>, cuyo objetivo es fomentar la investigación en el campo del análisis de sentimiento en los medios sociales en el idioma español. Principalmente, desea fomentar el desarrollo de nuevos algoritmos y la utilización de los ya existentes, para implementar sistemas cada vez más complejos que puedan analizar sentimientos en opiniones obtenidas de Twitter.

## 1.2 Objetivo

El objetivo del presente trabajo es desarrollar una herramienta que permita analizar los sentimientos de mensajes escritos en la red social Twitter. Concretamente, se centra en aquéllos que tratan temas de actualidad política publicados por cuatro de los más importantes periódicos nacionales: El País, El Mundo, La Razón y ABC.

Estos mensajes son clasificados en una de las siguientes cuatro categorías: positivo, neutro, negativo o ninguno de los anteriores.

---

<sup>1</sup> <http://www.sepln.org/>

<sup>2</sup> <http://www.daedalus.es/TASS>

El resultado se puede visualizar a través de una interfaz web a la que puede accederse tanto desde el PC, como desde tablets y teléfonos móviles. En esta interfaz, el usuario puede elegir entre diferentes temas y se muestra una gráfica comparativa de los cuatro periódicos donde se indica el número de tweets clasificados en cada categoría. Asimismo, permite consultar el contenido de los tweets clasificados y la polaridad predicha por el sistema.

## **1.3 Estructura del documento**

Después de este primer capítulo introductorio, en el siguiente se realizará un recorrido por diferentes estudios de investigación que se han realizado en los últimos tiempos en el campo del análisis de sentimientos. Entre éstos se incluyen estudios de generación de léxicos de polaridad, normalización de textos, así como el análisis centrado en redes sociales.

En el capítulo 3 se presentarán los recursos que se han utilizado para el desarrollo de la aplicación, y en el 4 se introducen las herramientas empleadas para su programación.

A continuación en el capítulo 5 se explica la arquitectura del sistema completo.

Durante el capítulo 6 se detalla cómo se ha realizado la recolección de tweets de la red social Twitter para su posterior análisis, y cómo se han almacenado.

Seguidamente se encuentra el capítulo más extenso del trabajo. En él se detalla el algoritmo utilizado para la creación del clasificador que predecirá la polaridad de los tweets.

En el capítulo 8 se indica el número de tweets recolectado por el sistema. Esta información se divide por periódicos, indicando para cada uno de ellos los temas de actualidad más frecuentemente tratados.

Los resultados obtenidos por la herramienta se pueden visualizar en una interfaz que es explicada en el capítulo 9.

A lo largo del capítulo 10 se realiza un análisis de las predicciones hechas por el clasificador en dos de los diez temas de actualidad tratados.

Por último, se exponen unas conclusiones sobre el trabajo realizado, así como problemas que se han encontrado a lo largo del desarrollo del proyecto y posibles trabajos futuros.



## 2. ESTADO DEL ARTE

Durante la última década se han realizado muchos trabajos en el campo del análisis de sentimientos. Si nos centramos en la tarea de la obtención de la polaridad de un texto, éste puede referirse a diferentes niveles: texto completo, oración, o incluso a las diferentes entidades nombradas en el texto.

Algunos trabajos de investigación importantes que estudian el análisis de polaridad a nivel de documento utilizan para ello críticas escritas por los usuarios en la web. Turney, (2002) las analiza centrándose en cuatro dominios diferentes (automóviles, bancos, películas y destinos vacacionales). Presenta un algoritmo de aprendizaje no supervisado para clasificar dichas opiniones como recomendadas (thumbs up) o no recomendadas (thumbs down). Estas clasificaciones las predice calculando la media de la orientación semántica (SO) de las frases del texto que contienen adjetivos o adverbios. Esta SO consiste en estimar, haciendo uso de buscadores de páginas webs, la Información Mutua Puntual (PMI) entre los términos a analizar y dos palabras que representan inequívocamente una orientación semántica positiva y negativa (en este caso escoge las palabras “excellent” y “poor”). La PMI mide estadísticamente la posible aparición de una palabra a partir de la aparición de otra.

Otro estudio destacado sobre el problema de la clasificación de opiniones en positivas o negativas, lo realizan Pang et al., 2002, utilizando como datos las críticas de películas encontradas en la web. El hecho de que el usuario además de escribir una opinión, pueda evaluar con un número de estrellas la película en cuestión, hace que no sea necesario etiquetar manualmente cada una de las opiniones como positivas o negativas. En su estudio utilizan tres algoritmos ya utilizados anteriormente para tareas como la clasificación de textos por tema: Naive Bayes, maximum entropy (MaxEn) y support vector machines (SVM). La conclusión obtenida es que, a pesar de que la precisión del resultado del uso de métodos de aprendizaje automático supera los *baselines* producidos manualmente por un humano, éstos no tienen un rendimiento tan bueno como el que se obtiene al tratar el problema de categorización por tema, convirtiendo por tanto el problema de análisis de sentimiento en una tarea más compleja.

Estos primeros estudios únicamente consideran el aprendizaje a partir de ejemplos con una polaridad positiva o negativa, ignorando los ejemplos que muestran un sentimiento neutro. Existen estudios como el de Koppel et al., 2006, en el que se muestra la importancia que tiene el uso de ejemplos neutrales en el proceso de aprendizaje, demostrando una mejor distinción entre polaridad positiva y negativa si se hace uso de éstos.

## **2.1 Generación de léxicos de polaridad**

Encontramos estudios orientados a generar léxicos de palabras en las que éstas se encuentran anotadas con su correspondiente polaridad. Rao et al., 2009, realizan un estudio en el que tratan el problema de detectar la polaridad de las palabras como un problema semi-supervisado de propagación de etiquetas en un grafo. Para ello utilizan como recurso la base de datos léxica en inglés Wordnet, y el diccionario de sinónimos de OpenOffice.

Debido al gran número de personas hispanohablantes en todo el mundo, existen diversos estudios (Pérez-Rosas, V, et al., 2012; Brooke, J, et al., 2009) en este ámbito para el desarrollo de léxicos de polaridad en español, ya que para el inglés ya existe un recurso lingüístico llamado SentiWordNet desarrollado para la comunidad científica (Esuli & Sebastiani, 2006) de gran utilidad aunque con cierta complejidad de uso. Esto es debido a la necesidad de realizar una desambiguación de la palabra antes de poder ser utilizada. En (Saggion and Funk, 2010) proponen una solución a este problema.

## **2.2 Análisis en redes sociales**

En los últimos años, debido a la importancia de las redes sociales en el campo de la minería de opiniones, muchos investigadores se han dedicado al estudio de las mismas. Existen interesantes estudios realizados sobre la información recogida de los comentarios escritos en Twitter. Bollen et al., 2010, han llegado a estudiar la posibilidad de predecir los resultados del mercado de valores a partir de los sentimientos expresados en esta red.

Un artículo interesante que examina el funcionamiento de los clasificadores en la minería de opinión de tweets en español, es el de Grigori Sidorov et al., 2013. Exploran diferentes configuraciones para ver cómo cada una afecta a la precisión de los algoritmos de aprendizaje automático. Experimentan con los algoritmos de Naive Bayes, Decision Tree y SVM, dado que éstos ya han presentado buenos resultados para el idioma inglés. En sus configuraciones tienen en cuenta diferentes tamaños n-gram, la longitud del corpus, el número de clases de sentimientos, corpus balanceado vs. corpus no balanceado y diferentes dominios para entrenar y testear (teléfonos móviles y política). En sus conclusiones determinan que la mejor configuración corresponde al uso de unigramas como features, un número tan pequeño como se pueda de clases (positivo y negativo), un tamaño de al menos 3000 tweets en el conjunto de entrenamiento (un tamaño superior no incrementa la precisión significativamente), un corpus no balanceado muestra una ligera mejoría en los resultados y el clasificador con más precisión es el SVM. Además, concluyen que el hecho de entrenar el sistema con tweets de un dominio diferente al que posteriormente se utilizará, empeora significativamente la precisión de los resultados, llegando a bajar del 85,8% al 28.0% en la prueba realizada con SVM.

En España, gracias a la SPLN y al TASS que se nombraba en el capítulo uno de este documento, diversos grupos de investigación españoles han presentado sus algoritmos sobre el análisis de sentimientos en Twitter en idioma castellano, dando así un impulso a esta línea de investigación.

Saralegi y San Vicente, 2013 consiguieron en este taller los mejores resultados en la tarea de análisis de sentimiento a nivel global de tweet. El método de aprendizaje supervisado que presentan usa un clasificador SVM que construyen con la herramienta WEKA. Esta solución incluye un procesamiento basado en conocimiento lingüístico para preparar los features que utilizará el clasificador. Dicho procesamiento incluye lematización y etiquetado POS realizado con Freeling, etiquetado de polaridad para el que construyen su propio léxico de polaridad, tratamiento de emoticonos y de negación. Además, para aumentar su precisión realizan un pre-procesamiento del texto de los tweets en el que realizan correcciones ortográficas.

## **2.3 Normalización de textos**

El pre-procesamiento de tweets se ha convertido en una fase importante en la tarea de análisis de sentimientos en Twitter. Se puede considerar que los usuarios han desarrollado una nueva forma de expresión que incluye el estilo de abreviaturas de los SMS, así como variantes léxicas, letras repetidas, uso de emoticonos, empleo de mayúsculas para añadir énfasis, etc. Existen numerosos trabajos sobre la normalización de este tipo de textos cortos para el idioma inglés (Han et al., 2011), pero sin embargo no existen tantos para el español. Por este motivo, se está fomentando esta área de investigación desde la SEPLN y se ha realizado en 2013 Tweet-norm<sup>3</sup>, un taller centrado en la tarea de normalización de tweets y desde el que ponen a disposición del público un corpus de tweets con problemas de normalización. Los diferentes algoritmos y recursos utilizados por los participantes pueden ser consultados en las Actas<sup>4</sup> del “XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural”.

---

<sup>3</sup> <http://komunitatea.elhuyar.org/tweet-norm/>

<sup>4</sup> <http://www.congresocedi.es/images/site/actas/ActasSEPLN.pdf>





### 3. RECURSOS UTILIZADOS

#### 3.1 Corpus

Uno de los recursos más importantes utilizados en el proyecto es el corpus con el que se entrena y testea el clasificador desarrollado. Este corpus, que se hizo público una vez finalizó el taller de análisis de sentimientos TASS, es el utilizado por los participantes en dicho workshop.

Está compuesto por 68.017 mensajes de Twitter en español escritos por periodistas, políticos y famosos de diferentes países de habla hispana. Está dividido en dos conjuntos, el de entrenamiento y el de test. A continuación se muestra una tabla<sup>5</sup> resumen.

Atributo	Valor
Tweets	68.017
Tweets (test)	60.798 (89%)
Tweets (entrenamiento)	7.219 (11%)
Temas	10
Idiomas en tweets	1
Usuarios	154
Fecha comienzo (entrenamiento)	2011-12-02 T00:47:55
Fecha fin (entrenamiento)	2012-04-10 T23:40:36
Fecha comienzo (test)	2011-12-02 T00:03:32
Fecha fin (test)	2012-04-10 T23:47:55

*Tabla 1. Resumen del contenido del corpus.*

Para cada mensaje, la información contenida en el corpus es:

- su identificador en la red social Twitter
- nombre de usuario que lo ha escrito
- fecha y hora
- lenguaje: siempre será español
- tema del que trata: política, entretenimiento, economía, música, fútbol, cine, tecnología, deportes, literatura u otros.
- polaridad global del tweet: para ello se utiliza una escala de 5 niveles: muy positivo (P+), positivo (P), neutro (NEU), negativo (N) y muy negativo (N+). Además, también puede etiquetarse con la etiqueta NONE en caso de no

<sup>5</sup> Fuente: Villena-Román, Julio, and Janine García-Morera. "TASS 2013-Workshop on Sentiment Analysis at SEPLN 2013: An overview."

incluirse en ninguno de los anteriores, es decir, de no indicar ningún sentimiento.

- polaridad de las entidades nombradas en el tweet (si las hubiera). Pueden o no coincidir con la polaridad global.

El formato en el que se distribuye el corpus es en XML y la estructura de cada mensaje se muestra a continuación.

```
<tweet>
  <tweetid>142378325086715906</tweetid>
  <user>jesusmarana</user>
  - <content>
    <![CDATA['Portada 'Público', viernes. Fabra al banquillo por 'orden' del Supremo;
      Wikileaks 'retrata' a 160 empresas espías. http://t.co/YtpRU0fd]]>
  </content>
  <date>2011-12-02T00:03:32</date>
  <lang>es</lang>
  - <sentiments>
    - <polarity>
      <value>N+</value>
      <type>AGREEMENT</type>
    </polarity>
    - <polarity>
      <entity>Wikileaks</entity>
      <value>N+</value>
      <type>AGREEMENT</type>
    </polarity>
    - <polarity>
      <entity>Fabra</entity>
      <value>N+</value>
      <type>AGREEMENT</type>
    </polarity>
  </sentiments>
  - <topics>
    <topic>política</topic>
  </topics>
</tweet>
```

*Figura 1. Tweet extraído del corpus*

La distribución de sentimiento dentro del corpus se muestra en la siguiente tabla.

Sentimiento	Frecuencia (training)	Frecuencia (test)
P+	22,44%	34,12%
P	4,12%	2,45%
NEU	8,45%	2,15%
N	16,91%	18,56%
N+	12,51%	7,5%
None	23,58%	35,22%

*Tabla 2. Distribución de sentimiento en el corpus.*

Debido a que en nuestro trabajo trataremos únicamente cuatro categorías: Positivo, Negativo, Neutro y Ninguno, hemos sustituido las etiquetas N+ por N y P+ por P antes de empezar a trabajar con el corpus.

## 3.2 Listado de abreviaturas

Como ya se ha comentado con anterioridad, los usuarios de Twitter han adoptado un lenguaje similar al que se utiliza frecuentemente en los mensajes SMS. Esto es debido a su limitación en cuanto a número de caracteres que se puede escribir en un mensaje (140).

Por este motivo, se ha decidido realizar un pre-procesamiento del texto contenido en los tweets para detectar las abreviaturas más comúnmente utilizadas y sustituirlas por su palabra completa. De este modo, posteriormente la palabra puede ser reconocida y tratada correctamente.

Esta sustitución se realizará basándose en un fichero que contiene 63 líneas, una por cada pareja de “abreviatura – palabra completa”. Las abreviaturas que contiene el fichero han sido recogidas de distintas páginas web<sup>6</sup>. A continuación se puede observar una pequeña muestra de dicho fichero.

<b>salu2</b>	saludos
<b>tmb</b>	también
<b>tp</b>	tampoco
<b>ppio</b>	principio
<b>dl</b>	del
<b>hsta</b>	hasta
<b>knto</b>	cuanto
<b>stoi</b>	estoy

*Figura 2. Muestra del fichero de abreviaturas*

## 3.3 Léxico de polaridad

Para conocer la polaridad de las palabras que contiene un tweet se utiliza un léxico de polaridad en castellano que nos ha sido proporcionado por la Fundación Elhuyar<sup>7</sup>. Éste fue creado a partir de diferentes fuentes e incluye alrededor de 5.200 palabras tanto positivas como negativas, etiquetadas con su polaridad. Dicho recurso fue conocido a partir de su participación en el taller TASS del año 2013 comentado con anterioridad.

<sup>6</sup> <http://www.servicios.movistar.com.pe/oye-sms-diccionario-sms.html>  
[http://es.wikipedia.org/wiki/Diccionario\\_SMS](http://es.wikipedia.org/wiki/Diccionario_SMS)

<sup>7</sup> [http://komunitatea.elhuyar.org/ig/files/2013/10/ElhPolar\\_esV1.lex](http://komunitatea.elhuyar.org/ig/files/2013/10/ElhPolar_esV1.lex)

A continuación se puede observar una muestra del contenido del fichero.

<b>acongojar</b>	N
<b>aconsejable</b>	P
<b>acordar</b>	P
<b>acortarse</b>	N
<b>acosar</b>	N
<b>activo</b>	P
<b>acuchillar</b>	N
<b>acuerdo</b>	P

*Figura 3. Muestra del fichero léxico de polaridad*

Para tomar la decisión sobre qué fichero se utilizaría en el trabajo, se estudiaron varias alternativas. Las más importantes se explican a continuación.

Se descartó la posibilidad de usar el léxico de polaridad de marzo de 2012 de la Universidad del Norte de Texas, elaborado según lo descrito en su artículo (Pérez-Rosas, V, et al., 2012) sobre este tipo de recursos en español. Este léxico está recogido en dos ficheros de texto plano, donde uno de ellos ha sido anotado automáticamente (2.496 palabras) y otro manualmente (1.347 palabras), siendo este último considerado más robusto por sus autores. Estas anotaciones que indican la polaridad de las palabras se refieren al idioma inglés. Sólo las primeras 100 palabras de ambos ficheros han sido etiquetadas además por españoles nativos con la polaridad correspondiente al idioma español. Se observa que en muchos casos las etiquetas son opuestas en los dos lenguajes tratados. Por este motivo, se ha creído que etiquetar el resto de palabras de una forma automática a partir de la polaridad indicada para el idioma inglés, pudiera no ser acertado. Este motivo hizo que fuera descartado para este trabajo.

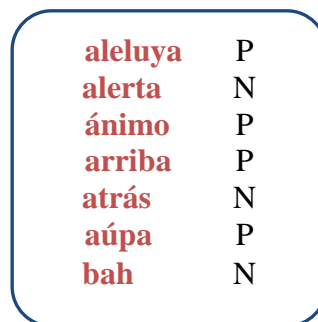
Otro recurso de polaridad en castellano que se ha estudiado y que ha sido descartado, es el del Centro de Investigación en Computación (CIC) del Instituto Politécnico Nacional de México que es explicado en profundidad en el artículo sobre minería de opiniones en tweets (Sidorov, Grigori, et al., 2013). Este léxico al que llaman SEL (Spanish Emotion Lexicon) está compuesto por 2.036 palabras, las cuales se encuentran etiquetadas con una de las seis emociones básicas del ser humano: alegría, enfado, tristeza, sorpresa o disgusto. Además, en otra columna se indica un número al que llaman PFA (Probability Factor of Affective use), y que indica en un rango de 0 a 1 la probabilidad de que esa palabra se refiera a la categoría de emoción indicada. El método para calcular esta probabilidad es explicado en otro de sus artículos (Sidorov, Grigori, et al., 2012). El hecho de que el etiquetado se haya realizado a mano convierte este léxico en un recurso robusto, pero habría que determinar exactamente cómo de útil resulta la probabilidad que indican en cada palabra y cómo tratar aquellas palabras etiquetadas como sorpresa,

ya que pueden referirse a un sentimiento positivo o negativo. Este hecho fue clave para descartarlo en este trabajo.

### 3.4 Listado de interjecciones

Debido a que las interjecciones son palabras que expresan sentimientos muy vivos, de dolor, alegría, tristeza, asombro, alarma, etc. son importantes a la hora de realizar el análisis de sentimientos de un texto. Por este motivo, tal y como se explica en el artículo del grupo Elhuyar del TASS 2013 citado anteriormente, pueden ser utilizadas para ayudar a decidir si un texto es positivo o negativo. El listado de interjecciones utilizado por los autores del artículo en su investigación y que se ha conseguido obtener para su uso en este trabajo, contiene 76 diferentes expresiones etiquetadas como positivas o negativas. La fuente principal de este listado ha sido una página web<sup>8</sup>.

En la siguiente figura se puede ver una muestra de este fichero.



<b>aleluya</b>	P
<b>alerta</b>	N
<b>ánimo</b>	P
<b>arriba</b>	P
<b>atrás</b>	N
<b>aúpa</b>	P
<b>bah</b>	N

*Figura 4. Muestra del fichero de interjecciones con polaridad*

### 3.5 Listado de emoticonos

Los emoticonos son caracteres que representan una emoción. Son muy extendidos en la web y son de gran valor a la hora de realizar el análisis de sentimientos, ya que al igual que con las interjecciones, pueden ayudar a averiguar si un texto es positivo, negativo o neutro.

Por este motivo, se ha utilizado un listado de emoticonos que nos han facilitado (Villar Rodríguez, E. et al., 2013) y que cuenta con 123 líneas, en la que cada una tiene una expresión regular que representa un emoticono y alguna de sus variaciones, así como una etiqueta para indicar si esta expresión muestra una emoción positiva, negativa o neutra. Gracias a este recurso se podrán identificar los emoticonos en los tweets analizados.

En la siguiente figura se puede observar una muestra del contenido del fichero.

---

<sup>8</sup> [http://www.solosequenosenada.com/gramatica/spanish/listado/lista\\_10\\_interjeccion.php](http://www.solosequenosenada.com/gramatica/spanish/listado/lista_10_interjeccion.php)

<b>:-&amp;</b>	N
<b>(: ;)-\*(\))+</b>	P
<b>(: ;)-\*</b>	P
<b>(: ;)-D</b>	P
<b>=P</b>	P
<b>:- (S)+</b>	N
<b>:-\ </b>	N

*Figura 5. Muestra del fichero de emoticonos*

## 4. HERRAMIENTAS UTILIZADAS

### 4.1 Twitter4j

La aplicación desarrollada necesita acceder a la información de Twitter para recolectar los tweets que posteriormente se analizan y clasifican. Para ello se ha utilizado una librería gratuita y open source llamada Twitter4j<sup>9</sup> escrita en Java. Aunque es una librería no oficial de Twitter, se presenta en la documentación para desarrolladores<sup>10</sup> existente en su web. Ésta permite realizar de una forma sencilla la integración de nuestra aplicación con la API de Twitter. La versión utilizada ha sido la 4.0.1.

Twitter utiliza el protocolo de autorización OAuth (Open Authorization). Este protocolo abierto permite que una aplicación externa acceda a Twitter en nombre de un usuario sin que ésta conozca las credenciales de su cuenta. De este modo hace de llave de entrada a Twitter de una forma segura para el usuario, ya que su contraseña no necesita ser conocida por un tercero para actuar en su nombre. Twitter4j permite compatibilidad con OAuth.

Se ha creado una cuenta de Twitter que será la que se utilice para acceder y recolectar los tweets. Por otro lado, es necesario registrar la aplicación en Twitter<sup>11</sup> para que cuando nos conectemos sepa qué aplicación es la que está intentando acceder a sus datos. Para ello, al realizar el registro nos asigna un par de identificadores llamados *consumer key* y *consumer secret* que identifica inequívocamente la aplicación. Además, obtenemos el *access token* y el *access token secret*, que permiten realizar peticiones a la API de Twitter en nombre de la cuenta creada. Estos identificadores son necesarios para realizar la conexión con la API de Twitter y recolectar los tweets.

Una vez realizada la conexión a la API se pueden realizar búsquedas de tweets por diferentes criterios. Nuestro sistema realiza las búsquedas utilizando el nombre de usuario de las cuentas oficiales de los periódicos. Para ello, hay que tener además en cuenta que Twitter impone unas restricciones<sup>12</sup> en cuanto al número de peticiones que pueden hacerse a su API en un determinado periodo y en cuanto a la antigüedad<sup>13</sup> de los tweets obtenidos.

### 4.2 Freeling

FreeLing es una librería de código abierto para el procesamiento multilingüe automático desarrollada por la Universidad Politécnica de Catalunya (UPC) y el Centro de

---

<sup>9</sup> <http://twitter4j.org/en/index.html>

<sup>10</sup> <https://dev.twitter.com/docs/twitter-libraries>

<sup>11</sup> <https://apps.twitter.com/>

<sup>12</sup> <https://support.twitter.com/articles/160385>

<sup>13</sup> [https://dev.twitter.com/docs/api/1.1/get/statuses/user\\_timeline](https://dev.twitter.com/docs/api/1.1/get/statuses/user_timeline)

Tecnologías y Aplicaciones del Lenguaje y del Habla (TALP). Se distribuye gratuitamente bajo licencia GPL. y proporciona funciones de análisis y anotación lingüística de textos. Al ser de código abierto, permite a los programadores utilizar los recursos por defecto (diccionarios, lexicones, gramáticas, etc.), ampliarlos para dominios específicos, o incluso desarrollar otros nuevos para diferentes idiomas. Actualmente el español es uno de los lenguajes soportados.

El uso de Freeling en aplicaciones de procesamiento del lenguaje natural, ahorra un gran trabajo al programador. Por esta razón se ha extendido su uso y puede verse cómo en diferentes artículos relacionados con el análisis de sentimientos, los autores lo han escogido para realizar sus investigaciones. Es un código robusto y tiene un buen rendimiento en términos de velocidad en tratamiento de datos en el mundo real.

Algunos de los servicios ofrecidos por Freeling son:

- División en oraciones.
- Tokenización de texto.
- Análisis morfológico.
- Detección de entidades nombradas.
- Reconocimiento de fechas, números, moneda y magnitudes físicas como velocidad, peso o temperatura, entre otras.
- Etiquetado gramatical, más conocido como part of speech (POS) tagging en inglés.

Freeling utiliza en el etiquetado gramatical para el castellano un conjunto de etiquetas que está basado en la propuesta del grupo EAGLES<sup>14</sup> (Expert Advisory Group on Language Engineering Standards) para la anotación morfosintáctica de corpus y lexicones de cualquier lengua europea.

En este trabajo Freeling ha sido utilizado para dividir el texto en palabras o tokens, obtener su lema y etiquetar cada una de ellas con su correspondiente categoría gramatical.

En la siguiente imagen puede verse un ejemplo realizado en la demo<sup>15</sup> que Freeling pone a disposición de los usuarios en su web. En él puede observarse en color rojo las etiquetas EAGLES que indican la categoría gramatical de las palabras. Por ejemplo en la palabra “el”, el código DA0MS0 significa:

D - Categoría: Determinante

A - Tipo: Artículo

---

<sup>14</sup> Expert Advisory Group on Language Engineering Standards (EAGLES):  
<http://www.ilc.cnr.it/EAGLES96/home.html>

<sup>15</sup> Demo de Freeling: <http://nlp.lsi.upc.edu/freeling/demo/demo.php>



- 0 - Persona: El 0 es el valor por defecto, con excepción de los valores determinantes que tienen el valor 1, 2, ó 3.
- M - Género: Masculino.
- S - Número: Singular.
- 0 - Poseedor: Sólo se determina para los determinantes posesivos, que podrán tomar los valores 1, 2 ó 3.

El significado de cada código puede encontrarse en la web de EAGLES citada anteriormente.

**FreeLing 3.1**  
AN OPEN-SOURCE SUITE OF LANGUAGE ANALYZERS

**Write your sentences**  
El gato come pescado.

**Analysis options**

- ☒ Multiword detection
- ☒ Number recognition
- ☒ Date/Time recognition
- ☒ Quantities, ratios, and percentages
- ☒ Named Entity detection
- ☐ Named Entity classification
- ☐ Phonetic encoding
- ☒ No sense annotation
- ☐ WN sense annotation: Frequency sorted (MFS disambiguation)
- ☐ WN sense annotation: PageRank sorted (UKB disambiguation)

**Select language**: Spanish  
**Select output**: Morphological Analysis

**Submit**

**Analysis Results**  
**Sentence #1**

El	gato	come	pescado	.
el	gato	comer	pescado	.
DA0MS0	NCMS000	VMIP3S0	NCMS000	Fp
1	1	0.994868	0.608233	1
		comer	pescar	
		VMM02S0	VMP00SM	
		0.00513197	0.391767	

Figura 6. Demo on-line de FreeLing

Destacar que en su web dispone de un foro que es de gran utilidad, ya que los desarrolladores de esta aplicación responden de una forma rápida y eficaz a las preguntas de los usuarios. También la documentación es bastante completa y se distribuyen ejemplos de cómo integrar algunas de sus funcionalidades en las aplicaciones a través de su librería.

## 4.3 WEKA

WEKA (Waikato Environment for Knowledge Analysis) es un software gratuito de aprendizaje automático y minería de datos distribuido bajo licencia GNU-GPL. Es una plataforma desarrollada en Java por la Universidad de Waikato, en Nueva Zelanda.

Weka contiene una buena colección de algoritmos para tareas de minería de datos y modelado predictivo. Éstos pueden aplicarse directamente a un conjunto de datos a través de su interfaz gráfica o bien pueden ser llamados desde un programa externo a

través de las API proporcionadas. En este trabajo se ha tomado la segunda opción, concretamente se ha utilizado la API de Java, ya que es el lenguaje que se ha utilizado para programar todo el sistema.

Weka incluye herramientas para el preprocesamiento de los datos (filtros), clasificación (árboles, tablas), clustering, reglas de asociación, y adicionalmente, diversas formas de visualización de los datos, tanto en el inicio del proceso de carga de datos, como después de haber aplicado un algoritmo.

En este proyecto WEKA ha sido utilizado para generar el clasificador que determina la polaridad de un tweet, es decir, lo clasificará como positivo, negativo, neutro o ninguna de las categorías anteriores. Para poder entrenar y testear este modelo del clasificador se utiliza el corpus explicado en el punto 3.1 de este documento.

En la siguiente imagen se puede ver su interfaz gráfica. Cabe destacar la completa documentación existente en su web.

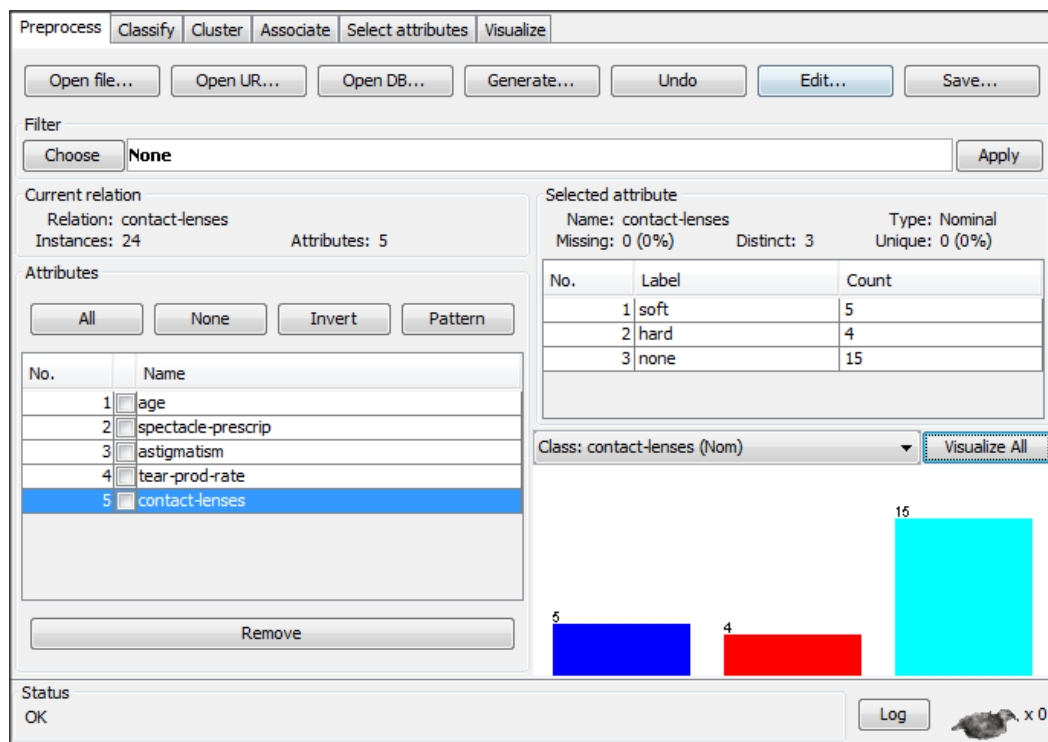


Figura 7. Interfaz gráfica de WEKA

## 5. ARQUITECTURA DEL SISTEMA

El sistema se compone principalmente de dos módulos escritos en Java que se ejecutan en batch, es decir, sin que exista interacción con el usuario durante su ejecución.

El primer módulo se encarga únicamente de recolectar tweets de la red social Twitter y guardarlos en ficheros XML para su posterior tratamiento.

El segundo módulo, núcleo del sistema, se encarga de analizar el contenido de los tweets recolectados por el módulo anterior y de crear el clasificador que predecirá su polaridad. Para ello utiliza el corpus de entrenamiento y de test explicado anteriormente en este documento. Además, este mismo módulo se encarga de realizar las predicciones y almacenar toda la información en una base de datos.

En la siguiente figura se puede ver un esquema general de la arquitectura que incluye los dos módulos. En los siguientes apartados se explica más en profundidad las funcionalidades de cada uno de los mismos.

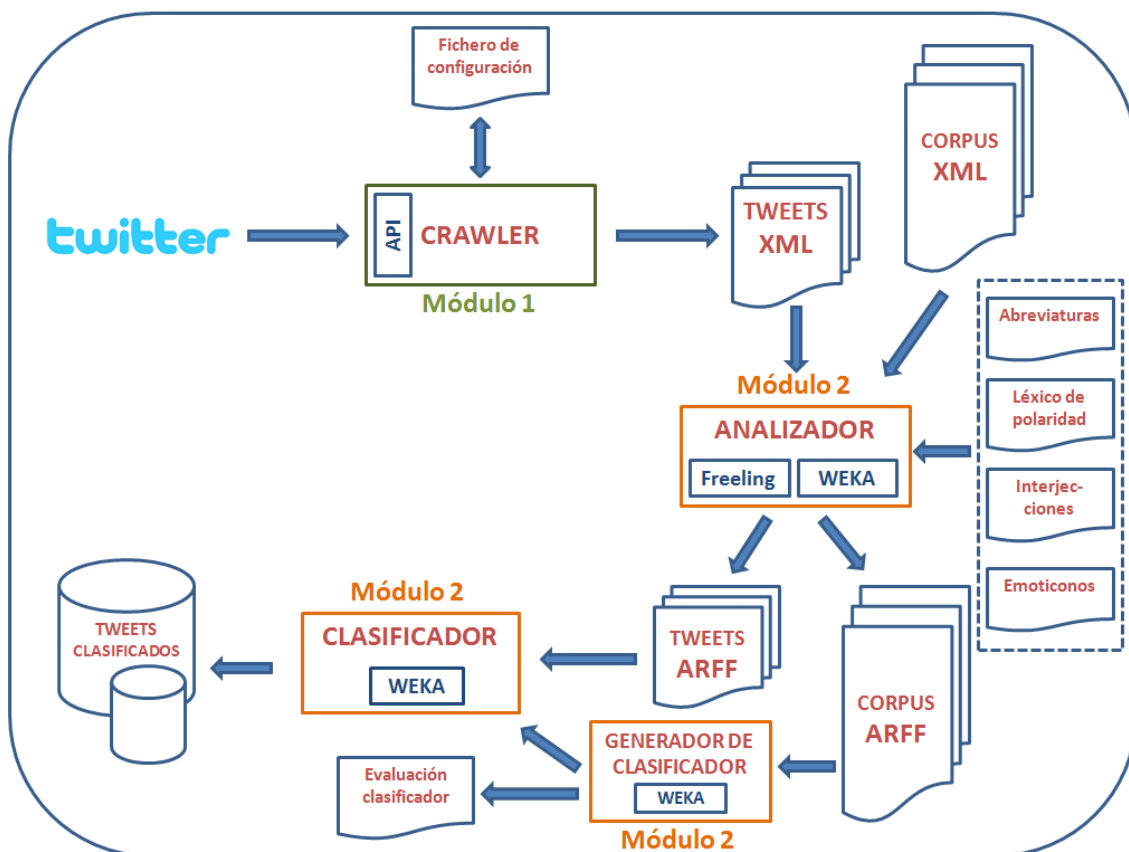
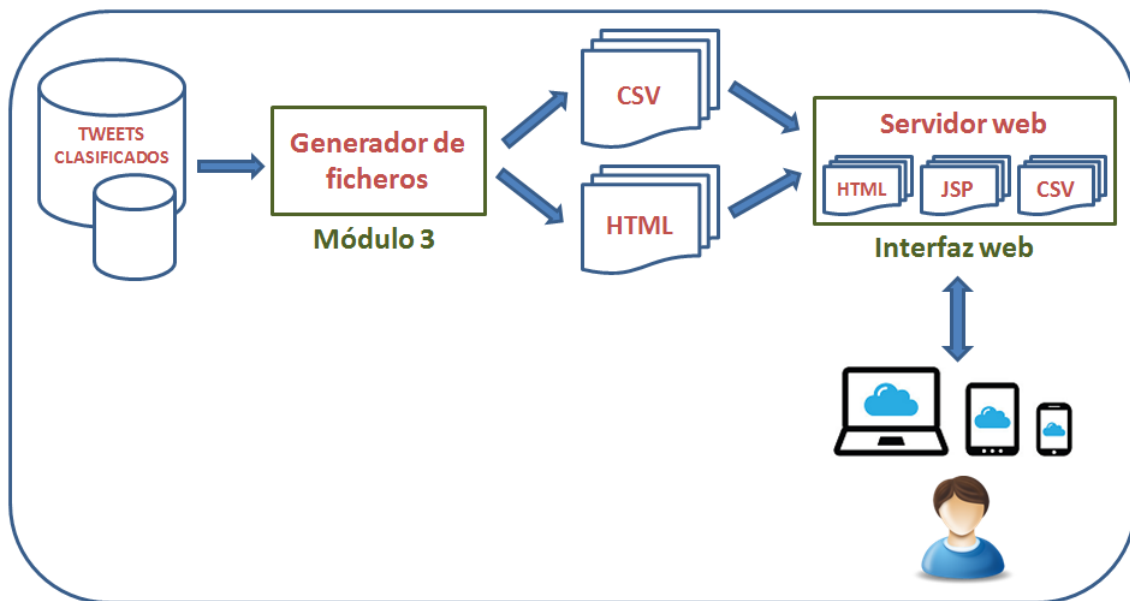


Figura 8. Arquitectura del sistema. Parte 1.

Para la visualización de los resultados se ha implementado una interfaz web que será descrita en el punto 9 de este documento. En esta interfaz el usuario puede elegir entre diferentes temas políticos y ver el resultado del análisis realizado de forma gráfica. Para ello, se ha implementado un tercer módulo en Java que a partir de consultas SQL a la base de datos, genera unos ficheros que serán el origen de datos de la interfaz. Este módulo también se ejecuta en batch.

En la siguiente figura se muestra la arquitectura de esta parte del sistema.



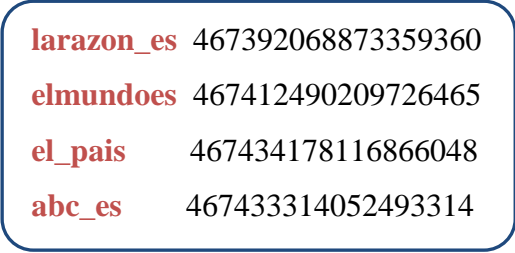
*Figura 9. Arquitectura del sistema. Parte 2.*

## 6. RECOLECCIÓN DE TWEETS

El primer módulo se encarga de recolectar los mensajes publicados en Twitter (conocidos como *status* en la red social) por las cuentas oficiales de los medios de comunicación escogidos. Para ello se utiliza la librería `twitter4j`<sup>16</sup>, una librería escrita en Java que permite fácilmente integrar una aplicación Java con Twitter.

Este módulo utiliza un fichero de configuración llamado `users.txt` que contiene 4 líneas, una por cada cuenta oficial de Twitter de los periódicos de los que se quiere obtener la información. Además, en cada una de las líneas, existe un número de 18 dígitos que indica el código del último tweet que fue recolectado por nuestro sistema para ese usuario. De este modo, en la siguiente ejecución del mismo, el sistema comienza a recuperar los tweets posteriores a dicho código.

El fichero de configuración es leído cada vez que se ejecuta la aplicación, y modificado a la finalización de la misma. A continuación se muestra un ejemplo del contenido de dicho fichero.



```
larazon_es 467392068873359360
elmundoes 467412490209726465
el_pais    467434178116866048
abc_es     467433314052493314
```

*Figura 10. Fichero de configuración del módulo Crawler.*

Los mensajes o *status* son guardados por el sistema en ficheros XML. Se genera un fichero por cada uno de los usuarios de Twitter indicados en el fichero de configuración.

Aunque se obtiene mucha información relacionada con el tweet, la única información que se guarda de cada uno es la siguiente:

- Id del tweet.
- Texto del mensaje
- Fecha de publicación

A continuación se muestra un ejemplo de uno de estos ficheros XML.

---

<sup>16</sup> <http://twitter4j.org/en/index.html>

```

- <tweets>
  - <tweet>
    <tweetid>443014731784536064</tweetid>
    - <content>
      <![CDATA[Los 9 bulos principales sobre el 11-M http://t.co/ey9EOg3vPK Todos fueron
      desmontados en los 100.000 folios del sumario #recuerdo11M]]>
    </content>
    <date>Mon Mar 10 14:25:17 CET 2014</date>
  </tweet>
  - <tweet>
    <tweetid>443009726830620672</tweetid>
    - <content>
      <![CDATA["Venezuela ya no es un país democrático y la gran movilización es para que
      haya todavía elecciones", dice Vargas Llosa http://t.co/NZ6LjD6bXz]]>
    </content>
    <date>Mon Mar 10 14:05:24 CET 2014</date>
  </tweet>
  - <tweet>
    <tweetid>443000964405280768</tweetid>
    - <content>
      <![CDATA[El ministro de Interior, sobre #Ceuta: "Hubiera sido mejor no lanzar las pelotas
      de goma" http://t.co/eREadr8mXx @elpais_politica]]>
    </content>
    <date>Mon Mar 10 13:30:35 CET 2014</date>
  </tweet>
</tweets>

```

*Figura 11. Fichero XML con los tweets recolectados por el módulo Crawler.*

## 7. ANÁLISIS – CLASIFICACIÓN

Este segundo módulo Analizador-Clasificador es el núcleo del sistema y es el más complejo. Se encarga de procesar el corpus de entrenamiento y de test para crear el clasificador que predecirá la polaridad de los tweets recolectados por el módulo explicado en el capítulo anterior.

Utiliza diferentes librerías para realizar sus funciones, entre las que se encuentran:

- **WEKA:** Para crear el clasificador y evaluarlo. Además, transforma los tweets recolectados en un formato reconocido por el clasificador para realizar la predicción de polaridad.
- **Freeling:** Obtiene los lemas de las palabras que forman el tweet y realiza el etiquetado POS de las mismas.
- **JDom:** Se usa para tratar los ficheros XML.
- **MySQL-Connector:** Necesaria para conectar y trabajar sobre la base de datos MySQL.

### 7.1 Algoritmo

Para realizar la clasificación de los tweets se ha decidido tomar como referencia el algoritmo utilizado por los ganadores del taller TASS 2013 (Urizar, XS., et al., 2013). Debido a la limitación de tiempo del proyecto, no se han implementado todas las funcionalidades del mismo.

El algoritmo realiza un tratamiento previo sobre el contenido de los tweets en el que las abreviaturas más comúnmente usadas en Internet son sustituidas por la palabra completa a la que representan. No se incluyen correcciones ortográficas en este tratamiento inicial, tal y como sí hicieron los ganadores de TASS 2013.

Se hace un procesamiento de los tweets utilizando la librería Freeling que incluye lematización y etiquetado POS.

Del contenido de los tweets, únicamente se tienen en cuenta para nuestro clasificador los emoticonos y las palabras que se encuentren en alguna de las siguientes categorías gramaticales: verbos, adverbios, adjetivos, sustantivos e interjecciones, ya que éstas suelen llevar el mayor peso subjetivo en un texto.

El léxico de polaridad sirve para detectar la polaridad de las palabras, siendo únicamente aquellas que aparecen en el corpus de entrenamiento y que se encuentran en el léxico, las que se utilizan como atributos o *features* en el clasificador.

Por último destacar que se usa un clasificador SVM que es construido con la herramienta WEKA utilizando el algoritmo SMO que implementa. Esto se debe a que en el artículo Pang et al., 2002 descrito en el capítulo 2, concluyen que es el que mejor resultado les ofrece. Además, al inicio del proyecto se realizó una prueba con el algoritmo Naive Bayes, que es uno de los tratados en este artículo, y el resultado era ligeramente peor.

En los siguientes apartados se explica más en detalle el algoritmo completo.

## **7.2 Carga de recursos**

El sistema realiza una carga de información a memoria desde los diferentes ficheros (fichero de abreviaturas, léxico de polaridad, interjecciones y emoticonos).

Estos ficheros se encuentran en una carpeta llamada *resources*.

## **7.3 Procesamiento del corpus**

El siguiente paso que se realiza es el procesamiento de los tweets del corpus, tanto de entrenamiento como de test. Por lo tanto, se procede a cargar los corpus en memoria y para cada tweet se realizan las siguientes tareas:

- Pre-procesamiento o normalización en el que las abreviaturas son sustituidas por la palabra completa. Esto se realiza gracias a la ayuda de la información cargada en memoria desde el fichero de abreviaturas.
- El mensaje se divide en palabras o tokens (tokenización) y se guarda para cada una de ellas su lema, y su categoría gramatical (etiquetado POS). Esta tarea se realiza utilizando la librería de Freeling.
- Utilizando las expresiones regulares cargadas desde el fichero de emoticonos, se busca alguna coincidencia dentro del texto del mensaje. Por cada coincidencia, se obtiene la polaridad del emoticono encontrado y se va acumulando en contadores el número de emoticonos positivos y negativos en el tweet.
- Se crea un campo de contenido reducido en el que únicamente se almacenan los lemas de las palabras del tweet que se han etiquetado con una de las siguientes categorías gramaticales: adjetivo, sustantivo, interjección, verbo o adverbio.



Además, para cada una de estas categorías se contabiliza el número de palabras que se han encontrado dentro del tweet de esa categoría concreta.

- Con la ayuda del léxico de polaridad y de interjecciones cargados en memoria, se contabiliza cuántas de las palabras que se incluyen en el campo nuevo creado de contenido reducido, son positivas o negativas. También se contabiliza el número de interjecciones positivas y negativas encontradas. Además, el campo de contenido reducido se reduce aún más, dejando únicamente aquellas palabras que se encuentren registradas en el léxico de polaridad, ya que éstas serán las que ayuden a determinar la polaridad final del tweet.

Por lo tanto, al final de esta parte hemos obtenido para cada tweet la siguiente información:

- |                                      |                            |
|--------------------------------------|----------------------------|
| - Número de interjecciones positivas | - Número de sustantivos    |
| - Número de interjecciones negativas | - Número de adjetivos      |
| - Número de emoticonos positivos     | - Número de adverbios      |
| - Número de emoticonos negativos     | - Número de verbos         |
| - Número de palabras positivas       | - Número de interjecciones |
| - Número de palabras negativas       | - Contenido reducido       |

Esta información será la que posteriormente se utilice para crear el clasificador, convirtiéndose en lo que se llaman atributos o *features*.

## 7.4 Generación de instancias

Una vez obtenida toda la información sobre cada uno de los tweets, se procede a generar un fichero con extensión ARFF (Attribute Relation Format File), formato reconocido por la herramienta WEKA. En este fichero se indican todos los atributos o *features* existentes para cada una de los tweets (instancias), y se añade además la polaridad (clase) como último atributo.

En la figura siguiente se puede ver un ejemplo de un fichero ARFF. Cada tweet es representado por una lista separada por comas. El primer elemento de la lista es el contenido reducido del tweet (tipo de dato string), y el resto de elementos de tipo numérico representan los valores que tienen cada uno de los atributos generados por nuestro analizador. El último valor de la lista es la polaridad del tweet y únicamente puede contener los valores P, N, NEU y NONE.

```

@relation Rel

@attribute content string
@attribute numEmoPos numeric
@attribute numEmoNeg numeric
@attribute numWordPos numeric
@attribute numWordNeg numeric
@attribute numInterPos numeric
@attribute numInterNeg numeric
@attribute numUppCase numeric
@attribute numVerb numeric
@attribute numNoun numeric
@attribute numAdj numeric
@attribute numAdv numeric
@attribute numInterj numeric
@attribute classPolarity {P,N,NEU,NONE}

@data
gracia,0,0,1,0,0,0,0,0,3,0,0,0,P
'corrupto no',0,0,0,2,0,0,0,5,6,1,1,0,N
encantar,1,1,1,0,0,0,0,2,2,0,0,0,P
'bueno abrazo grande ser grandeza',0,0,4,1,0,0,0,3,7,3,2,0,P
'ser gracia bueno amigo',0,0,3,1,0,0,0,1,6,1,0,0,P

```

Figura 12. Ejemplo de fichero ARFF

## 7.5 Transformación de instancias

El siguiente paso es realizar una transformación de las instancias generadas. Concretamente, se convertirá el campo de tipo *String* llamado *content* que guarda el contenido reducido del tweet, en una serie de atributos. Además, la representación de los tweets o instancias pasa de ser una lista de valores, a convertirse en un vector. Este proceso se realiza utilizando el filtro *StringToWordVector* ofrecido por WEKA.

Posteriormente se aplica otro segundo filtro llamado *Reorder* para reordenar los atributos, ya que WEKA necesita que el último atributo sea aquél que queremos predecir, es decir, la clase polaridad.

El algoritmo vuelve a generar otro fichero ARFF con el resultado de aplicar los filtros. En la siguiente figura se muestra el efecto de aplicar los filtros mencionados sobre el fichero de la figura 12.

```
@relation'Rel-weka.filters.unsupervised.attribute.StringToWordVector
```

```
@attribute abrazo numeric  
@attribute amigo numeric  
@attribute bueno numeric  
@attribute encantar numeric  
@attribute gracia numeric  
@attribute grande numeric  
@attribute grandeza numeric  
@attribute ser numeric  
@attribute corrupto numeric  
@attribute no numeric  
@attribute numEmoPos numeric  
@attribute numEmoNeg numeric  
@attribute numWordPos numeric  
@attribute numWordNeg numeric  
@attribute numInterPos numeric  
@attribute numInterNeg numeric  
@attribute numUppCase numeric  
@attribute numVerb numeric  
@attribute numNoun numeric  
@attribute numAdj numeric  
@attribute numAdv numeric  
@attribute numInterj numeric  
@attribute classPolarity {P,N,NEU,NONE}
```

```
@data  
{4 1,12 1,18 3}  
{8 1,9 1,13 2,17 5,18 6,19 1,20 1,22 N}  
{3 1,10 1,11 1,12 1,17 2,18 2}  
{0 1,2 1,5 1,6 1,7 1,12 4,13 1,17 3,18 7,19 3,20 2}  
  {1 1.2 1.4 1.7 1.12 3.13 1.17 1.18 6.19 1}
```

*Figura 13. Ejemplo de fichero ARFF filtrado*

Como se puede observar en la figura anterior, cada tweet que anteriormente era representado por una lista, se ha convertido en un vector. En este vector se indican pares de números, en el que cada par tiene el siguiente significado:

- primer valor: número de atributo al que hacen referencia, teniendo en cuenta que el primer atributo se representa con el valor 0. En este caso el atributo 0 se corresponde con la palabra abrazo.
- segundo valor: en caso de que el primer valor se corresponda con una palabra extraída del contenido del tweet, este valor indica el número de veces que aparece en el texto. Si el primer valor se refiere a alguno de los atributos que hemos creado con nuestro módulo analizador, como por ejemplo *numVerb* (que indica el número de verbos), este segundo valor indicará el número de verbos encontrados en el texto del tweet.

## 7.6 Generación y evaluación del clasificador

Con las instancias generadas correspondientes al corpus de entrenamiento se procede a la creación del clasificador. Concretamente, se utiliza el algoritmo SMO que WEKA utiliza para implementar una SVM.

Una vez creado, se realiza una evaluación del mismo y se obtiene un informe detallado en el cual se distinguen tres apartados:

- **Resumen**: porcentaje global de aciertos y errores cometidos en la evaluación.
- **Precisión detallada por clase**: para cada uno de los cuatro posibles valores que puede ser predicho por el clasificador, se muestra el porcentaje de instancias correctamente predichas (TP: True positives) y el porcentaje de instancias con otros valores que son incorrectamente predichas a ese valor (FP: False positives).

Además, se muestran ciertas medidas que tienen relación con las anteriormente expuestas.

$$Precision = \frac{\text{Instancias correctamente clasificadas con el valor "X"}}{\text{Total de instancias clasificadas con el valor "X"}}$$

$$Recall = \frac{\text{Instancias correctamente clasificadas con el valor "X"}}{\text{Total instancias que deberían haberse clasificado con el valor "X"}}$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

- **Matriz de confusión**: es una matriz que muestra de una forma detallada para cada clase el número de instancias predichas. Tiene dimensiones NxN, donde N es el número de los posibles valores que puede tomar la clase. En este caso los valores son *Positivo*, *Negativo*, *Neutro* o *Ninguno*, y por tanto se obtiene una matriz de 4x4.

A continuación se muestra el resultado de la evaluación obtenida por el clasificador creado.

```

=== Summary ===

Correctly Classified Instances      40757      67.0367 %
Incorrectly Classified Instances    20041      32.9633 %

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                0.812    0.205    0.695      0.812    0.749      P
                0.733    0.170    0.603      0.733    0.662      N
                0.090    0.018    0.100      0.090    0.094      NEU
                0.513    0.087    0.762      0.513    0.613      NONE
Weighted Avg.    0.670    0.150    0.682      0.670    0.664

=== Confusion Matrix ===

      a      b      c      d  <-- classified as
18046  1943   298  1946 |      a = P
 2314 11614   474  1442 |      b = N
   518   633   117    37 |      c = NEU
 5076  5075   285 10980 |      d = NONE

```

Figura 14. Resumen de la evaluación del clasificador

Como se puede observar en la figura, alrededor de un 67% de los tweets son clasificados de forma correcta. Si nos fijamos en las cuatro categorías existentes, vemos que los mejores resultados se obtienen en la categoría Positivo (F-Measure: 0.749) y los peores resultados en la categoría Neutro (F-Measure: 0.094), donde únicamente se clasifican correctamente 117 instancias de las 1.305 existentes.

Analizando la matriz de confusión, averiguamos que a pesar de la cantidad de instancias predichas de forma errónea, para la clase Positivo (a) se han clasificado correctamente la mayoría de las instancias, concretamente 18.046. Otras 1.943 instancias que deberían haberse clasificado como positivas, han sido clasificadas como negativas (b), y casi la misma cantidad, 1.946, han sido categorizadas en la categoría *Ninguno* (d).

Para la clase *Negativo* (b) ocurre lo mismo que con la categoría anterior, es decir, la mayoría de las instancias predichas se han realizado de forma correcta (11.614).

Sin embargo, se observa que la categoría *Neutro* (c) tiene el mayor porcentaje de errores, habiéndose clasificado 518 de estas instancias como positivas, 633 como negativas, 37 como neutras y únicamente 117 como correctas.

En la última y cuarta categoría en la que se clasifican las instancias en las que no se ha encontrado ninguna de las tres polaridades anteriores, podemos observar que 10.980 instancias se han clasificado correctamente, y casi la misma cantidad han sido clasificadas incorrectamente en las otras tres categorías.

Si tenemos en cuenta la distribución de sentimientos a lo largo del corpus que se muestra en la tabla 2 de este documento, podemos realizar las siguientes observaciones.

El hecho de que el porcentaje de tweets pertenecientes a la categoría Neutro en el corpus de entrenamiento sea tan solo de un 8,45% puede ser el motivo de que el clasificador tenga tan mal resultado en esta categoría. Por otro lado, el resultado en esta categoría puede no afectar demasiado al conjunto total debido a que únicamente el 2,15% de los tweets en el corpus de test son de esta polaridad.

Si comparamos nuestro resultado con los obtenidos por los participantes en el taller TASS 2013, podemos comprobar que el resultado es relativamente bueno, ya que en dicho taller los resultados de precisión en esta misma tarea estuvieron comprendidos entre un 23% y un 68,6%, siendo el mejor valor el obtenido por los participantes que presentaron el algoritmo que hemos intentado emular en este trabajo. Además, cabe destacar que la categoría *Neutro* fue también en la que obtuvieron los peores resultados.

## 7.7 Procesamiento de tweets

Una vez creado el clasificador y evaluado, se procede a procesar los tweets que han sido recolectados de Twitter por el módulo recolector explicado en el capítulo 6.

A estos tweets se les aplica exactamente el mismo proceso que a los procedentes del corpus, es decir, lo explicado en los apartados 7.3, 7.4 y 7.5. Con la única diferencia de que en este caso no conocemos la polaridad de los mismos, por lo que el valor de la clase polaridad será desconocido. Al finalizar el proceso, los tweets se guardan en una tabla de la base de datos.

## 7.8 Clasificación de tweets

Una vez analizados y procesados los tweets, se realiza su clasificación. Para ello se utiliza el modelo de clasificador generado y explicado en el apartado 7.6.

A cada registro de la base de datos que representa un tweet, se le añade la polaridad predicha, de modo que al final de este proceso la base de datos tiene una única tabla de datos llamada Tweets con los campos que se indican en el siguiente cuadro.

Campo	Descripción
<b>id</b>	Código identificador del tweet
<b>texto</b>	Contenido del mensaje del tweet
<b>polaridad</b>	Polaridad predicha (P, N, NEU, NONE)
<b>hashtag</b>	Hashtag nombrado en el texto del mismo. Si hubiese varios, sólo guarda el primero que encuentre.

Tabla 3. Definición de la tabla que contiene los tweets en la BBDD

## 8. TEMÁTICA DE LOS DATOS ANALIZADOS

De cada uno de los cuatro periódicos nacionales sobre los que se ha trabajado (El País, ABC, La Razón y El Mundo), se ha extraído mediante el módulo recolector de datos la siguiente cantidad de mensajes. Se muestra, según nuestro clasificador, cómo han sido categorizados.

	Positivo	Negativo	Neutro	Ninguno	TOTAL
<b>El País</b>	1.213	1.655	79	897	3.844
<b>ABC</b>	1.532	1.952	73	1.455	5.012
<b>La Razón</b>	1.168	1.083	58	915	3.224
<b>El Mundo</b>	1.120	1.290	45	1.199	3.654

*Tabla 4. Cantidad de mensajes recolectados de Twitter.*

Por cada periódico, se detalla a continuación el ranking de los 15 hashtags que más ha nombrado y el número de veces que lo ha hecho. También se muestra el número de tweets que no contienen ningún hashtag, que como se puede ver, representa aproximadamente un 80% de los tweets totales.

ABC	
Sin hashtag	3.306
# venezuela	78
#endirecto	74
#ep2014	64
#elreyabdica	58
#champions	41
#hasidonoticia	35
#madrid	31
# ucrania	27
#ampliamos	25
#curioso	24
#ligabbva	23
#ciencia	20
#directo	18
#endosminutos	18
#rusia	15

El País	
Sin hashtag	3.260
#portada	61
#elreyabdica	47
#recuerdo11m	41
#finalchampions	21
#elecciones europeas	19
#rusia	18
#ucrania	14
#marchasdignidad22m	13
#gabrielgarciamarquez	12
#venezuela	12
#cmin	10
#gamonal	8
#22m	8
#depadolfo suarez	7
#melilla	6

*Tabla 5. Ranking de hashtags citados por los periódicos ABC y El País.*

El Mundo	
Sin hashtag	2.572
#tuitopina	161
#envivo	131
#endirecto	99
#ampliamos	69
#encuentrodigital	44
#laportada	43
#bonusparatuiteros	42
#elreyabdica	25
#marchasdignidad22m	14
#elmundoenlibertad	14
#editorial	13
#eleccionesem	10
#22m	9
#11m10aniversario	8
#adolfo Suarez	8

La Razón	
Sin hashtag	2.541
#elreyabdica	32
#ucrania	26
#madrid	24
#den2014	23
#directo	20
#cmin	18
#ere	15
#deporte	13
#entrevista	13
#cosasbuenasdelosviernes	12
#ampliamos	12
#depadolfo Suarez	8
#mikioskero	8
#venezuela	8
#11m10aniversario	7

*Tabla 6. Ranking de hashtags citados por los periódicos El Mundo y La Razón.*

Esta información, junto con un estudio más exhaustivo de los mensajes que no contenían ningún hashtag, ha sido utilizada para decidir los temas de actualidad<sup>17</sup> a tratar en este trabajo:

- Conflicto de Ucrania.
- Fallecimiento de Adolfo Suárez.
- Independentismo catalán.
- Aniversario del atentado el 11-M en Madrid.
- Marchas por la dignidad del 22-M.
- Reforma de la Ley del aborto.
- Asalto a las vallas de Ceuta y Melilla.
- Caso Noós o caso Urdangarín.
- Caso Bárcenas.
- Abdicación del Rey Juan Carlos I

En las tablas anteriores, se puede observar que además de existir un gran número de mensajes sin hashtag en su contenido, se encuentran muchos tweets que contienen hashtags que no ayudan a identificar el tópico que se trata en el mensaje. Por ejemplo: #envivo, #endirecto, #entrevista, #editorial, #ampliamos, etc.

<sup>17</sup> En el anexo 1 puede consultarse una descripción de cada uno de los temas de actualidad expuestos.



Por este motivo se ha realizado un tratamiento de la información existente en la base de datos. Este tratamiento ha consistido en realizar búsquedas por palabras clave en el contenido del tweet que pudiesen ayudar a catalogarlo dentro de alguno de los temas de actualidad política ya comentados. De este modo, se han conseguido etiquetar alrededor de 1.000 mensajes que no contenían hashtag.

Por ejemplo, el siguiente tweet no contiene ningún hashtag pero está relacionado con el caso Noós:

*Tweet: ' la fiscalía se opone a que Urdangarín sea imputado por un presunto delito de blanqueo de capitales'.*

Ante un caso como el ejemplo anterior, se decide asignar un valor manual al campo hashtag de la tabla Tweets de la base de datos. Para ello se ejecuta una sentencia SQL como la siguiente:

*UPDATE tweets SET hashtag="#noos" WHERE hashtag IS NULL AND texto LIKE "%urdangarin%";*

De este modo, se consigue disminuir el número de tweets que no contienen hashtag y que no podemos relacionar con uno de los tópicos a analizar.

Algunas de las palabras claves que se han utilizado en las búsquedas para relacionar tweets con los tópicos tratados, han sido por ejemplo:

Tópico: Asaltos a las vallas de Ceuta y Melilla por parte de inmigrantes. Palabras clave:

- valla + Ceuta
- valla + Melilla
- asalto + valla
- concertinas + Melilla

Aparte de este problema que ha sido tratado, se han detectado casos en los que sobre un mismo tema existen diferentes hashtags que lo hacen referencia. Por ejemplo:

Tópico: Aniversario del atentado del 11-M en Madrid. Hashtags relacionados:

- #11m
- #11-M
- #11m10aniversario
- #11memoria
- #recuerdo11m

Por este motivo se ha decidido agrupar en un mismo tema varios hashtags que lo hacen referencia. Para ello se ha creado una tabla adicional en la base de datos llamada Tópicos con los siguientes campos:

Campo	Descripción
<b>topic</b>	Nombre de un tópico de actualidad
<b>hashtag</b>	Hashtag relacionado con el tópico

*Tabla 7. Definición de la tabla Tópicos en la BBDD*

A continuación se muestra parcialmente el contenido de esta tabla.

Topic	Hashtag
<b>11-M</b>	#11m
<b>11-M</b>	#11-M
<b>11-M</b>	#11m10aniversario
<b>11-M</b>	#11memoria
<b>11-M</b>	#recuerdo11m

*Tabla 8. Contenido parcial de la tabla Tópicos de la BBDD*

Después de este tratamiento, el número de tweets identificados de cada uno de los temas de actualidad a tratar es el siguiente:

Tópico	Número de tweets
Conflicto Crimea	513
Fallecimiento de Adolfo Suárez	235
Abdicación del Rey Juan Carlos I	234
Independentismo catalán	108
Aniversario atentado 11-M Madrid	105
Marchas por la dignidad 22-M	69
Reforma de la Ley del Aborto	64
Asalto a las vallas de Ceuta y Melilla	48
Caso Noós - Urdangarín	45
Caso Bárcenas	40

*Tabla 9. Número de tweets identificados por tópico*

## 9. VISUALIZACIÓN DE RESULTADOS

### 9.1 Generación de datos

Para mostrar los resultados a través de la interfaz web, es necesaria la utilización de un módulo intermedio que genere unos ficheros de texto plano con la información necesaria para componer las gráficas comparativas de los diferentes periódicos.

Este módulo generador de datos está escrito en Java y se ejecuta en batch, al igual que los anteriores. Utiliza la tabla Tópicos de la base de datos que describimos con anterioridad. Gracias a ella y a la información almacenada de los tweets, mediante consultas SQL genera unos ficheros CSV (Comma Separated Values) en los que se indica para cada uno de los tópicos existentes y para cada periódico, el número de tweets encontrados de cada polaridad.

Estos ficheros son ubicados en el servidor web junto con las páginas web que forman la interfaz. Dependiendo del tema de actualidad elegido por el usuario se cargarán unos u otros datos.

### 9.2 Estructura de la web

La web está formada por diferentes secciones a las que se accede por un menú situado en la parte superior de la pantalla. Cada una de estas secciones será explicada en los siguientes apartados.

#### a) Página principal

La página principal, a la que puede accederse desde otra sección pulsando sobre la opción del menú *Analizar*, tiene dos partes:

##### Parte 1:

Ofrece al usuario la posibilidad de elegir, utilizando una lista desplegable, entre diferentes temas de actualidad política sobre los que se ha hecho un análisis.

Una vez elegido el tema y hecho click sobre el botón *Analizar*, se muestra una nueva página en la que el usuario puede ver gráficamente una comparativa del resultado obtenido en los diferentes medios analizados.

## Parte 2:

Muestra una descripción sobre cada uno de los medios de comunicación sobre los que se realiza el análisis y permite acceder a su edición digital mediante un botón situado en la parte inferior.

Análisis de la opinión en tweets periodísticos

Analizar

Proyecto

Contacto

Análisis de la opinión en tweets periodísticos

Elija el tema que le interese analizar

Atentado terrorista del 11-M

Analizar »



**EL MUNDO**  
@elmundoes

Nacido después de la Dictadura (1989), éste destacó como diario de investigación de calidad al destapar casos de corrupción en el Gobierno de Felipe González, sobresaliendo la figura de Pedro J. Ramírez, uno de sus fundadores y director hasta febrero de este año.  
Con sede en Madrid, su línea editorial se postula más en la derecha española, aunque ellos se definen como un diario liberal.  
Según los últimos datos certificados por la Oficina de Justificación de la Difusión (OJD) el promedio de tirada es de 248.463 ejemplares.

Ir a su web »

**EL PAIS**  
@el\_pais

Diario español de línea independiente y plural con sede en Madrid. Fundado por José Ortega Spottorno, su primer número apareció el 4 de mayo de 1976, en plena Transición. Su actitud ante el golpe de Estado del 23-F le consolidó como un diario de izquierdas y valiente. Según los últimos datos certificados por la OJD cuenta con un promedio de tirada de 359.809 ejemplares. Cuenta con una edición digital y pertenece al Grupo Prisa.

Ir a su web »

**ABC.es**  
@abc\_es

Fundado por Torcuato Luca de Tena y Álvarez Ossorio, ABC es un diario español con más de un siglo de antigüedad (1903) y con una línea editorial liberal, conservadora y monárquica.  
Se caracterizó en sus comienzos por su tamaño reducido, formato que sigue manteniendo.  
Según los últimos datos certificados por la OJD cuenta con un promedio de tirada de 198.347 ejemplares.

Ir a su web »

**La Razón**  
@larazon\_es

Diario español con sede en Madrid, fundado en 1998 por Luis María Anson. Su nacimiento provocó una fuerte lucha con el periódico ABC, ya que estaban enfocados hacia el mismo espectro de lectores.  
Tiene una marcada línea conservadora y actualmente suele ser polémico por sus portadas y las opiniones de su director Francisco Marhuenda.  
Según los últimos datos certificados por la OJD cuenta con un promedio de tirada de 119.060 ejemplares.

Ir a su web »

Figura 15. Página principal de la interfaz web

### b) Proyecto

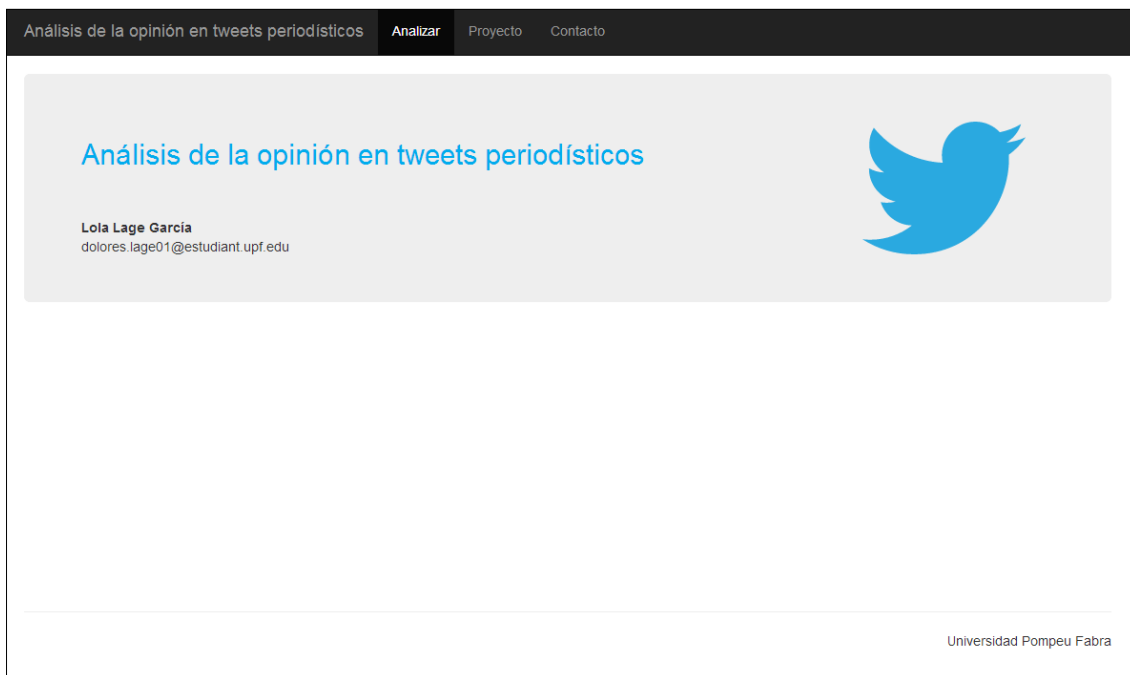
En esta sección se muestra al usuario información sobre este proyecto. Concretamente, la información disponible es el resumen tanto en castellano como en inglés que se encuentra al inicio de este documento.



*Figura 16. Sección “Proyecto” de la interfaz web*

### c) Contacto

En esta página se muestra la forma de contactar con la autora del proyecto.



*Figura 17. Sección “Contacto” de la interfaz web*

### 9.3 Representación gráfica de la información

Para poder crear una web dinámica en la que la información que muestre sea diferente dependiendo del tema de actualidad que el usuario elija, la interfaz se ha desarrollado en JSP con la ayuda de la librería javascript d3.js<sup>18</sup> para la generación de gráficos. Además, para conseguir una interfaz web que se adapte a dispositivos móviles, se ha utilizado el framework Bootstrap<sup>19</sup>.

Los resultados de los análisis sobre un tema elegido por el usuario se muestran con un gráfico de barras comparativo donde aparecen los cuatro medios de comunicación analizados.

Para cada uno de los medios se muestran cuatro barras que representan el número de tweets que ha publicado sobre el tema elegido, representando cada una de las barras una categoría diferente de clasificación: Positivo, Negativo, Neutro y Ninguno.

Para poder ver qué tweets han sido clasificados dentro de cada categoría, el usuario puede pulsar sobre el botón que se encuentra debajo de cada uno de los periódicos representados. Los tweets se mostrarán en la parte derecha de la pantalla con el mismo aspecto que tienen en Twitter. Como información adicional debajo de cada tweet se indica la polaridad que nuestro sistema ha asignado al mismo, de este modo el usuario puede observar si esta clasificación es o no correcta.

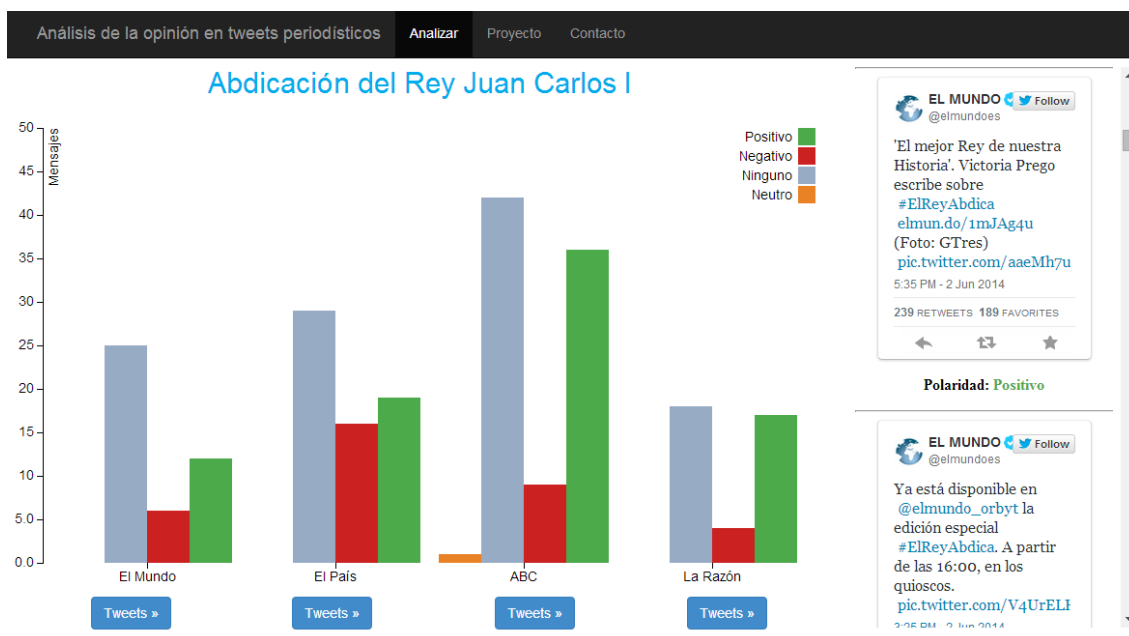


Figura 18. Representación gráfica de la información en la interfaz web

<sup>18</sup> <http://d3js.org/>

<sup>19</sup> <http://getbootstrap.com/>

El hecho de poder mostrar los tweets en la página web con el mismo formato que en la red social, se debe a que Twitter ofrece la posibilidad de embeberlos<sup>20</sup> en una página web de forma sencilla. A continuación se muestra el código HTML que permite incrustar un tweet en la página. La información indispensable para ello es conocer el nombre del usuario que ha escrito el mensaje y el código identificador del tweet, que como se puede ver en la siguiente figura, es lo que se usa para formar la url de enlace.

```
<script async src='http://platform.twitter.com/widgets.js' charset='UTF-8'></script>  
<blockquote class='twitter-tweet' width='150' align='center'  
data-cards='hidden' data-conversation='none'>  
<a href='https://twitter.com/abc_es/statuses/443322993230360577'></a>  
</blockquote>
```

*Figura 19. Código HTML necesario para embeber un tweet en la web.*

---

<sup>20</sup> <https://dev.twitter.com/docs/embedded-tweets>





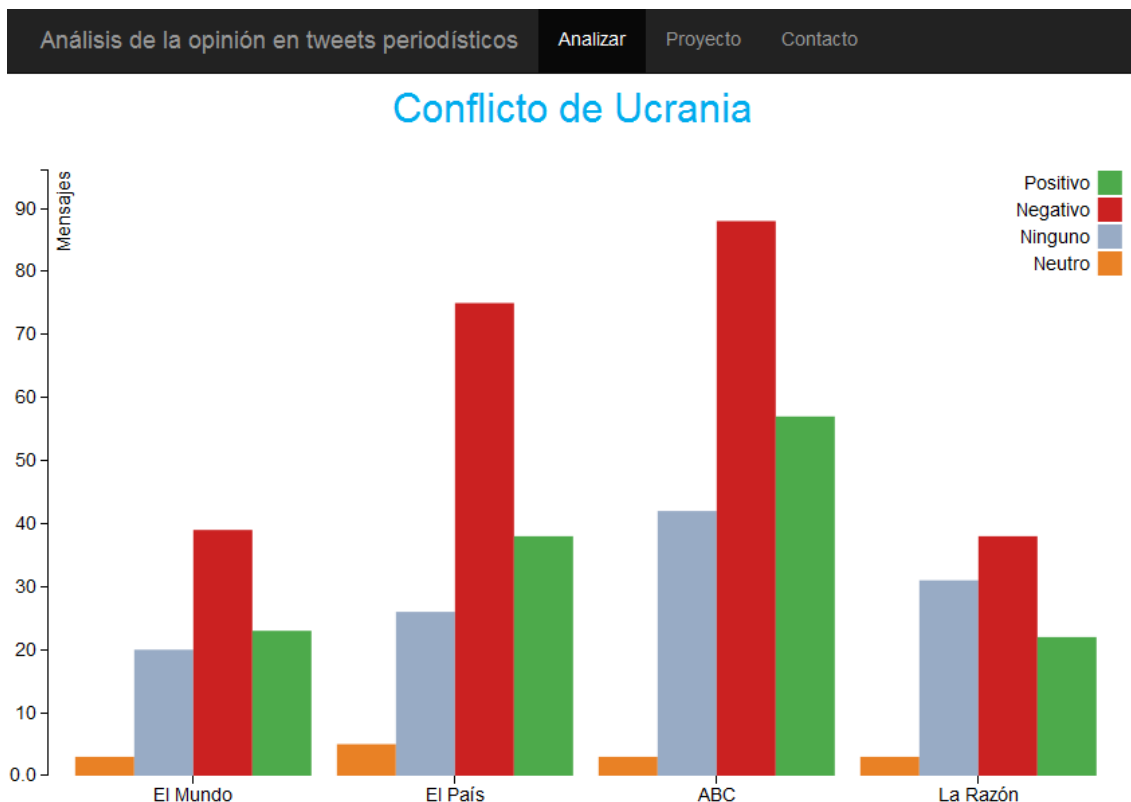
## 10. ANÁLISIS DE LOS RESULTADOS OBTENIDOS

En este apartado se examina el resultado obtenido por nuestro sistema clasificador. En la evaluación del mismo, que se explica en el apartado 7.6, se obtiene un porcentaje de acierto del 67%, por lo tanto sabemos que en sus predicciones existirá un número elevado de errores.

A continuación se muestran los resultados obtenidos en dos de los temas sobre los que más tweets han publicado los periódicos. Se exponen casos en los que el sistema acierta en su predicción, y algunos en los que falla, intentando buscar algún patrón que nos ayude a entender el funcionamiento del clasificador.

Tópico: Conflicto de Ucrania:

En la siguiente imagen extraída de la interfaz web, se puede apreciar que en líneas generales este tópico ha generado tweets de un sentimiento negativo. En este sentido el sistema es creíble, ya que estamos hablando de un conflicto político que ha causado muchos disturbios en los que ha habido muertes.



*Figura 20. Resultado obtenido sobre el tópico Conflicto de Ucrania*

En este tópico analizado, se puede observar que el medio de comunicación que más se ha hecho eco de las noticias referentes al mismo ha sido ABC, y el que menos el periódico El Mundo.

De las cuatro categorías existentes, se puede distinguir claramente que la que cuenta con menos cantidad de tweets clasificados es *Neutro*.

Analizando los mensajes individualmente, se puede apreciar que los mensajes que contienen palabras cuyo lema coincide con el lema de aquéllas que se encuentran en el léxico de polaridad, suelen ser clasificados con la polaridad de las mismas, aunque no siempre se cumple esta regla.

Algunos ejemplos de tweets analizados y clasificados de forma correcta y que parecen seguir esta regla se pueden ver a continuación. Todos ellos contienen palabras que en el léxico de polaridad se consideran positivas. En el primer tweet encontramos las palabras “propone” y “ayuda”, en el segundo “aprobará” y “rápidamente” y en el tercero “apoyan” y “adhesión”.



Figura 21. Ejemplo de tweets etiquetados correctamente como positivos en el tópico Conflicto de Ucrania

Si observamos alguno de los mensajes etiquetados como negativos, vemos que ocurre algo similar. El primer tweet contiene la palabra “suspende”, el segundo “grave” y “sufrir”; y el tercero “violencia” y “estalla”. Todas ellas etiquetadas con polaridad negativa en el léxico.



*Figura 22. Ejemplo de tweets etiquetados correctamente como negativos en el tópico Conflicto de Ucrania*

Existen casos en los que es más difícil averiguar qué atributos han tenido más peso para el clasificador, ya que el patrón que hemos explicado de las palabras y su polaridad no se sigue. A continuación se presentan algunos ejemplos de clasificaciones en las que los tweets contenían palabras incluidas en el léxico y cuya polaridad no ha determinado la clasificación del tweet.

En el primer tweet la única palabra en el léxico es “invadido” y está etiquetada con polaridad negativa. En el segundo encontramos la palabra “disolver” que está etiquetada como negativa. Por último, en el tercer tweet se incluye la palabra “pactan” que se considera positiva.



*Figura 23. Ejemplo de tweets que no siguen el patrón de aparición en el léxico en el tópico Conflicto de Ucrania*

Por otro lado, encontramos muchos tweets que se han clasificado correctamente en la categoría *Ninguno*, ya que no expresan sentimientos de ningún tipo. En el ámbito que estamos analizando este tipo de mensajes son muy habituales.



Figura 24. Ejemplo de tweets clasificados en la categoría *Ninguno* en el tópico *Conflicto de Ucrania*

Tópico: Abdicación del Rey Juan Carlos I:

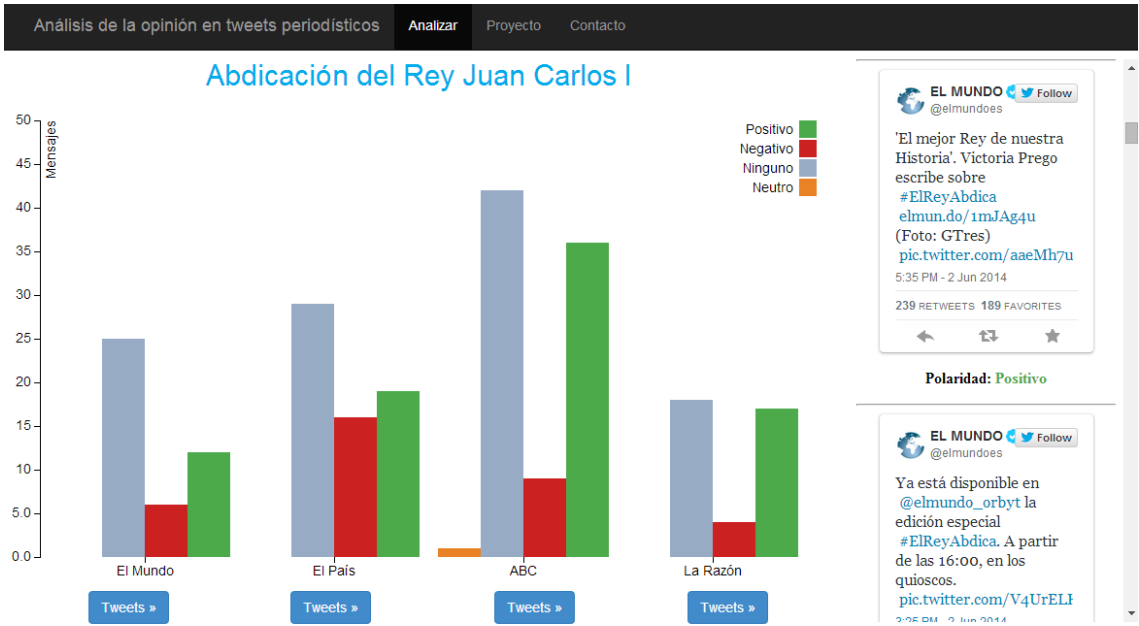


Figura 25. Resultado obtenido sobre el tópico *Abdicación del Rey Juan Carlos I*

Como se puede observar en la gráfica, sobre este tópico en general los mensajes escritos por los periódicos han sido clasificados como positivos, aunque existen también muchos de ellos en los que no se ha encontrado polaridad.

Observando los tweets, realmente este resultado representa correctamente el sentimiento general, ya que la mayoría de los tweets ensalzan la figura del Rey y su importante papel en la época histórica de la Transición Española.

A continuación se muestran algunos de los tweets que se han clasificado correctamente como positivos dentro de esta temática.



*Figura 26. Ejemplo de tweets clasificados correctamente como Positivos en el tópico Abdicación del Rey Juan Carlos I*

En cuanto a tweets clasificados como negativos, encontramos algunos como los siguientes, en los que al igual que en la temática anterior, su clasificación parece depender de la existencia de palabras con polaridad negativa en el léxico ("cruciales", "fallida", "retirada", "cansado", "rompió"), aunque no sabemos en qué grado afectan el resto de atributos que tiene en cuenta el clasificador.



*Figura 27. Ejemplo de tweets clasificados correctamente como Negativos en el tópico Abdicación del Rey Juan Carlos I*

El único tweet en este tópico clasificado como neutro, contiene dos palabras clasificadas con polaridad negativa en el léxico (“contra” y “urgencia”) y una positiva (“aprueba”). Además, como el resto de tweets publicados, no contiene ningún emoticono, por lo que es difícil saber exactamente qué atributos del clasificador han tenido más peso para declinarse por esta predicción.



**Polaridad: Neutro**

*Figura 28. Ejemplo de tweet clasificado como Neutro en el tópico Abdicación del Rey Juan Carlos I*

Haciendo referencia a los mensajes que son etiquetados en la categoría *Ninguno*, se puede decir que aunque en ocasiones el sistema falla, muchos de los mensajes claramente muestran esa falta de polaridad que hace que no puedan ser clasificados en ninguna otra categoría. Por ejemplo:



*Figura 29. Ejemplo de tweets clasificados en la categoría Ninguno en el tópico Abdicación del Rey Juan Carlos I*

Tal y como se observaba en la matriz de confusión mostrada en el apartado 7.6, muchos de los tweets que debían de ser clasificados en el test dentro de las categorías

*Neutro* o *Ninguno*, habían sido clasificados como positivos o negativos. Ejemplos de estas clasificaciones erróneas pueden ser los siguientes.



*Figura 30. Ejemplo de tweets clasificados incorrectamente en el tópico Abdicación del Rey Juan Carlos I*

Se han detectado ejemplos en los que el sistema ha fallado en su predicción y que podrían solventarse agregando un detector de negación en el algoritmo. De modo que en estos casos, podría invertirse la polaridad. A continuación se muestran algunos de éstos.



*Figura 31. Ejemplo de tweets que incluyen negación clasificados incorrectamente en el tópico Abdicación del Rey Juan Carlos I*

Una vez repasados estos dos tópicos se puede observar que aunque el léxico de polaridad tiene una fuerte influencia en la predicción que realiza el clasificador, el resto de atributos también tienen un peso en la misma que debe ser tenido en cuenta.

A pesar de la tasa de error en las predicciones, se ha demostrado que con la herramienta desarrollada y por medio de la interfaz web, es posible captar el sentimiento general que se desprende de los tweets publicados por los periódicos.



## **11. DISCUSIÓN**

### **11.1 Problemas encontrados**

A la hora de realizar el análisis de los mensajes del corpus, el primer problema que he encontrado ha sido el hecho de que los usuarios en Twitter utilicen abreviaturas, repitan letras o cometan faltas de ortografía. Estas palabras a veces no son correctamente clasificadas en su categoría gramatical, y además no pueden encontrarse en el léxico de polaridad. Para obtener unos mejores resultados, habría que hacer un proceso de normalización más exhaustivo que corrigiese el texto antes de empezar su tratamiento. De este modo todas las palabras serían reconocidas por el sistema y etiquetadas dentro de la categoría gramatical correcta.

Otro obstáculo que he encontrado y que es parte del problema existente en la tarea de análisis de sentimiento, es el hecho de la dificultad de clasificar correctamente mensajes que contengan ironía. Ésta es difícil de detectar incluso para el ser humano, por lo que realizarlo de una forma automática es una tarea compleja que no se ha tenido en cuenta en este trabajo.

En cuanto al tratamiento de los mensajes extraídos de Twitter de los periódicos nacionales, el problema inicial disminuye, ya que no suelen realizar faltas de ortografía ni suelen repetirse letras para mostrar énfasis. Sin embargo, surge un problema diferente. Estos usuarios no escriben opiniones personales, sino titulares de noticias en las que el lenguaje suele ser más sobrio y neutro, no añadiendo interjecciones o emoticonos en ellos. Esto hace que algunos de los atributos creados en el clasificador, y que al analizar el corpus de entrenamiento tenían su importancia, dejen de tener peso, no ayudando por tanto a realizar una predicción correcta.

### **11.2 Trabajo futuro**

Como ya se ha comentado en el apartado 2.3 existen estudios de investigación que se centran en la normalización de textos cortos, y concretamente algunos estudian la normalización de tweets, por lo que el problema planteado anteriormente podría solventarse añadiendo algún proceso de normalización antes de tratar los tweets.

En este trabajo, por falta de tiempo, únicamente se han tratado las abreviaturas más comunes en textos cortos, pero esta tarea puede ampliarse con la utilización de diccionarios y ayudar así a obtener mejores resultados.

Otra mejora que puede añadirse al sistema desarrollado, es la detección automática de la ironía, ya que ésta aparece en muchos de los tweets tratados. Existen estudios al respecto, incluso centrados en la red social Twitter, Barbieri and Saggion (2014), que podrían ayudar a mejorar el sistema.

Un trabajo futuro que puede realizarse es el de añadir nuevas categorías a predecir, como por ejemplo: Muy Positivo y Muy Negativo. En este caso, habría que tener en cuenta las estrategias de intensificación y atenuación existentes en gramática. Quizá añadir como atributos el número de retwiteos de un mensaje o el número de hashtags que contiene puede ayudar al clasificador.

Por último, mencionar la posibilidad de realizar un tratamiento de la negación en el contenido del tweet. Existen diversos estudios (Rodríguez E.V., et al., 2014) sobre este tema y es utilizado por diversos participantes en el TASS que demuestran una mejora en los resultados finales de los clasificadores presentados.

### **11.3 Conclusiones**

El objetivo del proyecto era desarrollar una herramienta capaz de detectar el sentimiento en los tweets publicados por diferentes periódicos nacionales. En particular, el propósito era realizar un análisis sobre diferentes temas de actualidad política y mostrar los resultados en una interfaz atractiva para el usuario.

Este objetivo ha sido cumplido. Para ello se han recolectado alrededor de 15.000 mensajes de la red social Twitter durante el periodo en el que se ha desarrollado el trabajo. De entre todos ellos, finalmente se decidió analizar los 10 temas de actualidad política que más repercusión han tenido en la red social.

Aunque la herramienta comete errores en sus predicciones, se ha demostrado que observando la gráfica comparativa de los cuatro periódicos que muestra la interfaz, se consigue detectar el sentimiento general que los medios de comunicación han plasmado en los tweets publicados. Esto me ha causado una gran satisfacción.

A nivel personal, el desarrollo de este proyecto me ha valido para poner en práctica muchos de los conocimientos adquiridos durante mis estudios.

Por otro lado, he tenido que crear un sistema completo que incluye las aplicaciones realizadas en Java, la interfaz web, la conexión con la base de datos MySQL y la gestión del servidor web Apache Tomcat. Y aunque al principio tuve diversos problemas, he sido capaz de ir resolviéndolos por mí misma, de lo cual estoy muy contenta.

Este trabajo me ha servido para adentrarme en el campo del procesamiento del lenguaje natural, el cual considero muy interesante. Desde el punto de vista práctico, también me ha ayudado a adquirir nuevos conocimientos de algunas de las herramientas que ayudan a realizarlo, WEKA y Freeling, así como de programación en entorno web. He aprendido JSP y Javascript, además de haber descubierto el framework Bootstrap que te

ayuda a realizar diseños web atractivos de una forma sencilla. También he conocido más a fondo la librería D3.js para la creación de gráficos.

Por otro lado, he vivido de primera mano la generosidad por parte de diferentes grupos de investigación que no conocía previamente y con los que me he puesto en contacto durante la realización del proyecto. En ningún caso han dudado en compartir conmigo algunos de sus recursos de una forma totalmente desinteresada.

En definitiva, ha sido una experiencia de final de carrera muy positiva.



## Bibliografía

SAGGION, Horacio; FUNK, Adam. Extracting opinions and facts for business intelligence. *RNTI Journal, E (17)*, 2009, vol. 119, p. 146.

RUBIO, Julio Larrañaga. Industria de los periódicos: nuevos modelos económicos y nuevos soportes. *Estudios sobre el mensaje periodístico*, 2010, vol. 16, p. 59-78.

TURNEY, Peter D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002. p. 417-424.

PANG, Bo; LEE, Lillian; VAITHYANATHAN, Shivakumar. Thumbs up?: sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002. p. 79-86.

KOPPEL, Moshe; SCHLER, Jonathan. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 2006, vol. 22, no 2, p. 100-109.

RAO, Delip; RAVICHANDRAN, Deepak. Semi-supervised polarity lexicon induction. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009. p. 675-682.

PEREZ-ROSAS, Veronica; BANEJA, Carmen; MIHALCEA, Rada. Learning Sentiment Lexicons in Spanish. En *LREC*. 2012. p. 3077-3081.

BROOKE, Julian; TOFILOSKI, Milan; TABOADA, Maite. Cross-Linguistic Sentiment Analysis: From English to Spanish. En *RANLP*. 2009. p. 50-54.

ESULI, Andrea; SEBASTIANI, Fabrizio. Sentiwordnet: A publicly available lexical resource for opinion mining. En *Proceedings of LREC*. 2006. p. 417-422.

SAGGION, Horacio; FUNK, A. Interpreting SentiWordNet for opinion classification. En *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*. 2010.

BOLLEN, Johan; MAO, Huina; ZENG, Xiaojun. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011, vol. 2, no 1, p. 1-8.

SIDOROV, Grigori, et al. Empirical study of machine learning based approach for opinion mining in tweets. En *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2013. p. 1-14.

VILLENA-ROMÁN, Julio; GARCÍA-MORERA, Janine. TASS 2013-Workshop on Sentiment Analysis at SEPLN 2013: An overview. In Proceedings of "XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS2013). Madrid. pp. 112-125. ISBN: 978-84-695-8349-4

URIZAR, Xabier Saralegi; RONCAL, Inaki San Vicente. Elhuyar at TASS 2013. In Proceedings of "XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS2013). Madrid. pp. 143-150. ISBN: 978-84-695-8349-4

HAN, Bo; BALDWIN, Timothy. Lexical normalisation of short text messages: Makn sens a# twitter. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011. p. 368-378.

DÍAZ-RANGEL, I.; SIDOROV, G.; SUÁREZ-GUERRA, S. Weighted Spanish Emotion Lexicon. (submitted) (2012)

RODRIGUEZ, Esther Villar; SERRANO, Ana García; RODRÍGUEZ, Marta González. Análisis lingüístico de expresiones negativas en tweets en español. In Proceedings of "XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS2013). Madrid. pp. 112-125. ISBN: 978-84-695-8349-4

BARBIERI, Francesco; SAGGION, Horacio. Modelling Irony in Twitter. *EACL 2014*, 2014, p. 56.

RODRÍGUEZ, Esther Villar; BASTIDA, Ana. I. Torre; GARCÍA-SERRANO, Ana; RODRÍGUEZ, Marta González. TECNALIA-UNED@ TASS: Using a linguistic approach for sen-timent analysis. In Proceedings of "XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS2013). Madrid. pp. 200-205. ISBN: 978-84-695-8349-4

## **Anexo 1. Temas de actualidad tratados.**

- Conflicto de Ucrania: a raíz de la negativa por parte de sus autoridades a la firma de un Acuerdo de Asociación con la Unión Europea (UE) creció el descontento de sus ciudadanos, que mediante protestas consiguieron deponer al presidente Víktor Yanukóvich. El conflicto continúa actualmente generando mucha tensión.
- Fallecimiento de Adolfo Suárez: el 23 de marzo muere en Madrid Adolfo Suárez, presidente del Gobierno en el período 1976-1981. Considerado una pieza clave en la transición española.
- Independentismo catalán: corriente político-social que pretende conseguir la independencia de Cataluña del Estado español, ante la negativa del Gobierno central a que se lleve a cabo un referéndum.
- Aniversario atentado 11-M en Madrid: atentado producido en los trenes de cercanías de Madrid el 11 de marzo de 2004. Fue llevado a cabo por una célula de terroristas yihadistas y causó 192 muertos.
- Marchas por la dignidad del 22-M: miles de personas acuden a Madrid desde diferentes puntos de España el 22 de marzo de este año debido al malestar generado por los recortes realizados por el Gobierno, el aumento del paro y en general, la gestión de la crisis económica.
- Reforma de la Ley del aborto: reforma presentada por el Gobierno del Partido Popular que ha generado una gran controversia en la sociedad española por ser una ley más restrictiva y conservadora que la actual.
- Asalto a las vallas de Ceuta y Melilla: en este tema se tratan los asaltos que han sufrido durante estos últimos meses las vallas fronterizas de Ceuta y Melilla. Los inmigrantes africanos han visto en ellas la puerta de entrada a Europa, a pesar de la presencia de la Guardia Civil.
- Caso Noós - Urdangarín: caso de presunta corrupción política en la que se ha visto involucrado Iñaki Urdangarín, marido de la infanta Cristina. Su nombre proviene del Instituto Noós que éste dirigía junto con su socio Diego Torres. Ha tenido un gran impacto mediático y ha causado un gran daño a la imagen de la Casa Real.
- Caso Bárcenas: caso de corrupción que afecta al partido político que actualmente gobierna España, el Partido Popular. Su nombre proviene del ex - tesorero del partido Luis Bárcenas, que a día de hoy se encuentra en prisión por diferentes delitos, entre los que se encuentra el blanqueo de capitales, falsedad documental y cohecho.

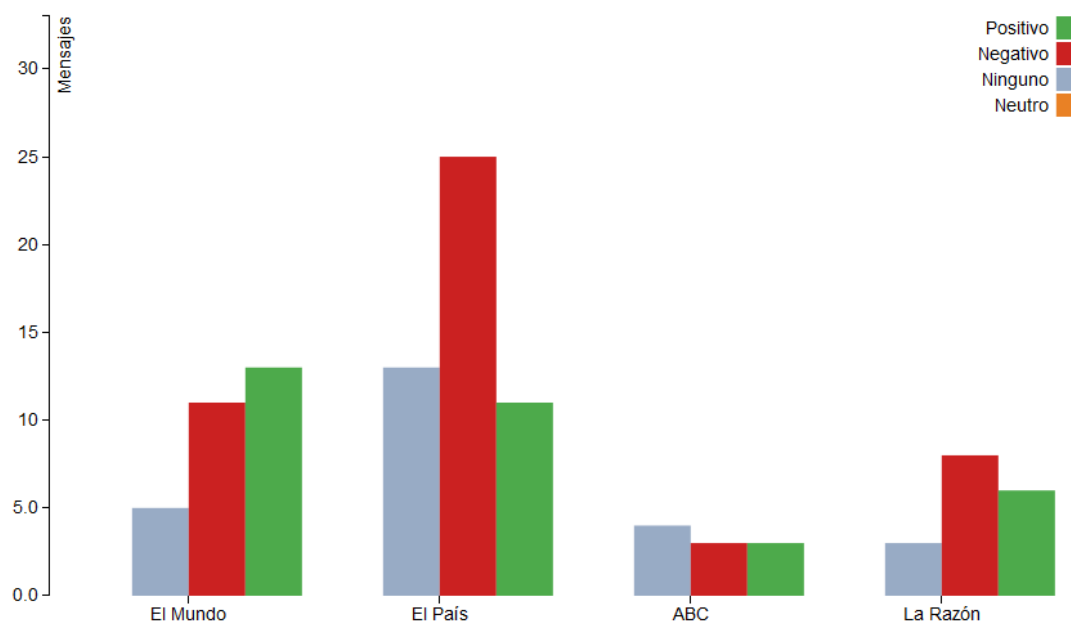
- Abdicación del Rey Juan Carlos I: Juan Carlos I, Rey de España desde 1975 comunica su decisión de abdicar la corona de España, dejando la misma en manos de su hijo Felipe, quien reinará como Felipe VI. Ante tal decisión parte de la sociedad española se manifiesta para pedir un referéndum entre monarquía y república.



## Anexo 2. Gráficas de los resultados obtenidos.

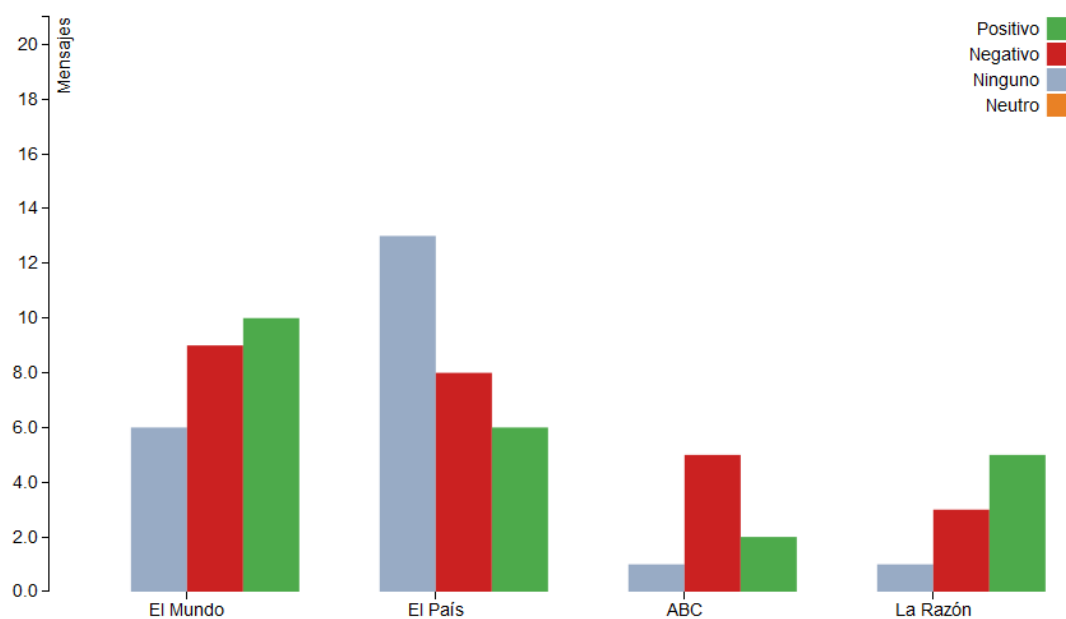
### Aniversario del atentado terrorista del 11-M en Madrid.

#### Atentado terrorista del 11-M



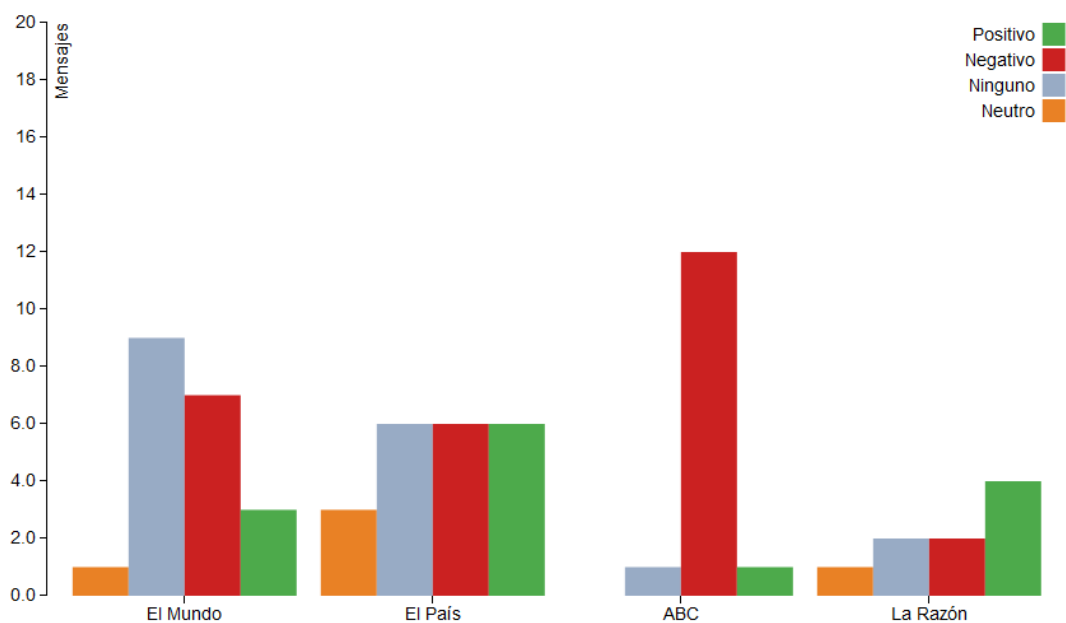
### Marchas por la dignidad del 22-M.

#### Marchas de la dignidad del 22-M



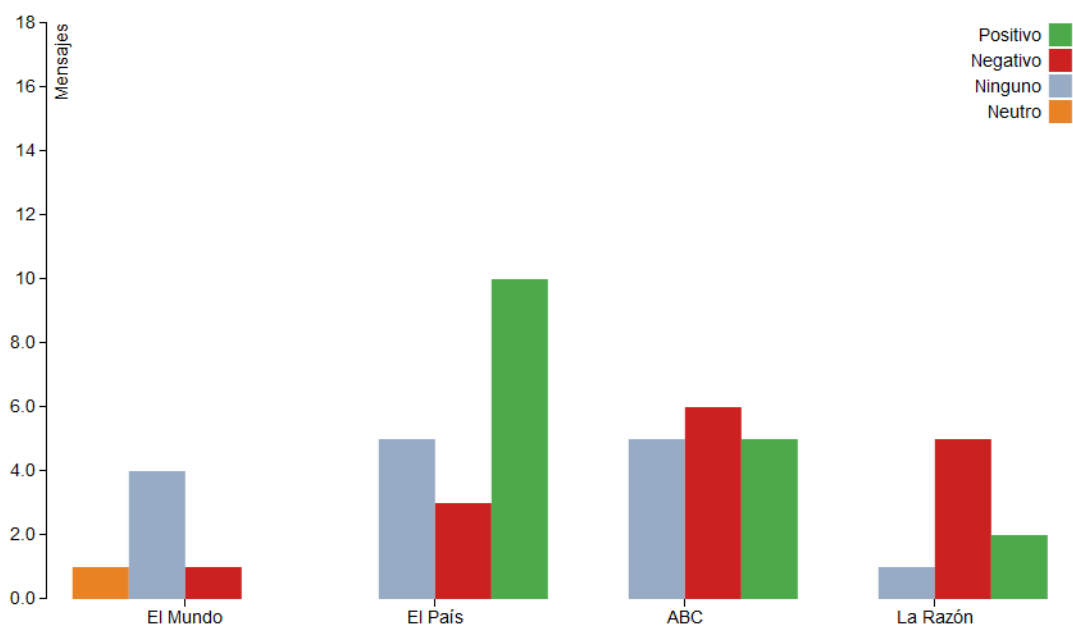
## Reforma de la Ley del aborto.

### Reforma de la Ley del Aborto

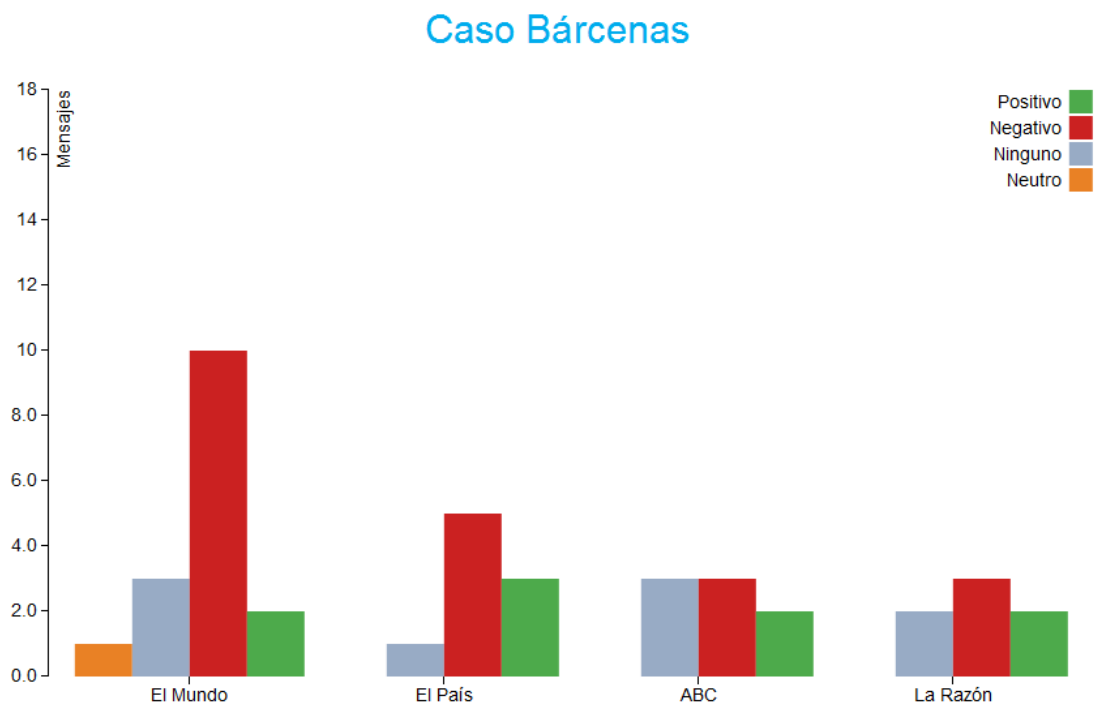


## Asaltos a la valla de Ceuta y Melilla.

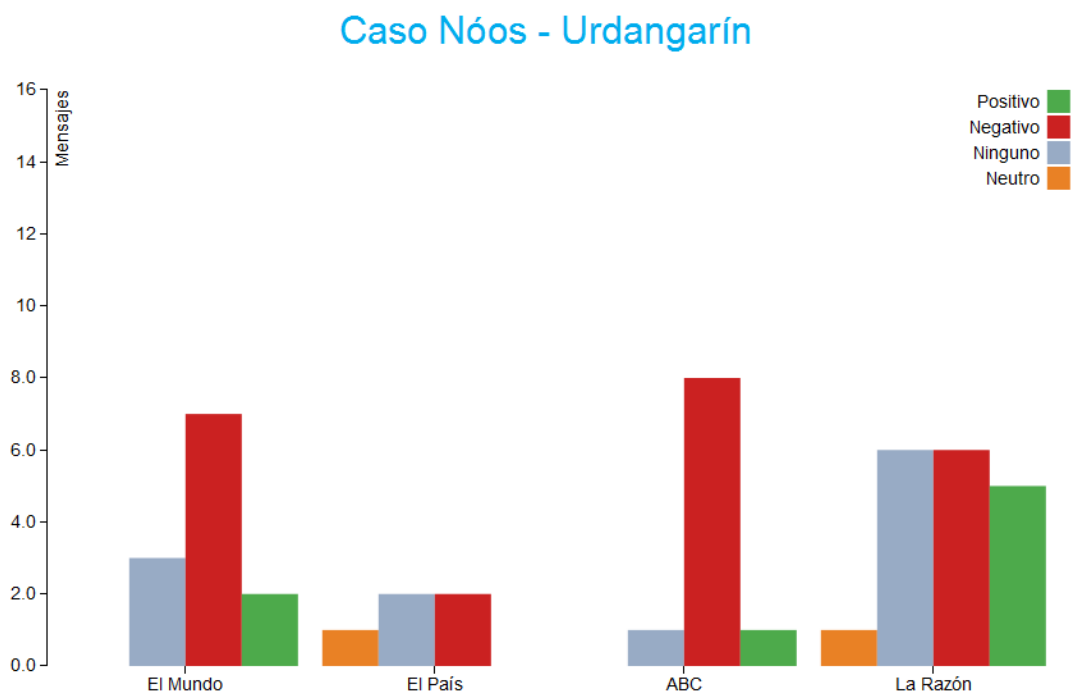
### Asaltos a la valla de Ceuta y Melilla



Caso Bárcenas.

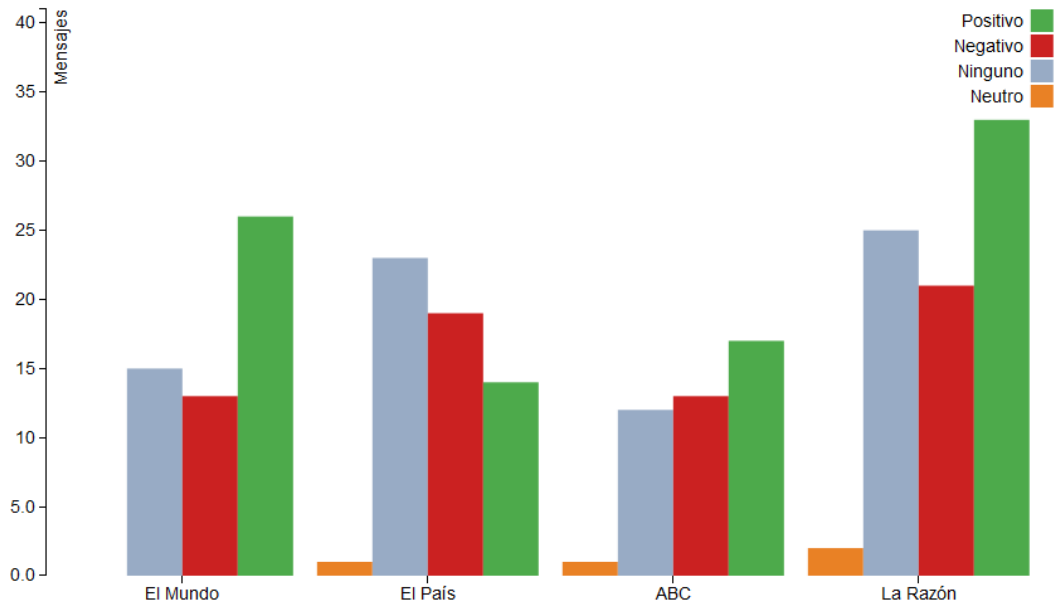


Caso Noós – Urdangarín.



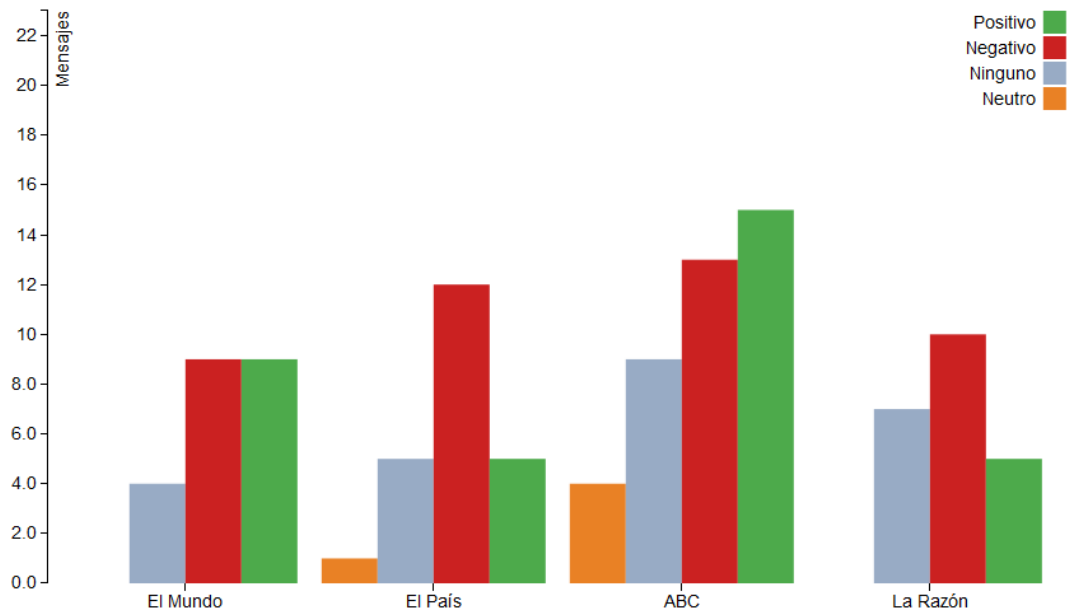
## Fallecimiento de Adolfo Suárez .

### Fallecimiento de Adolfo Suárez

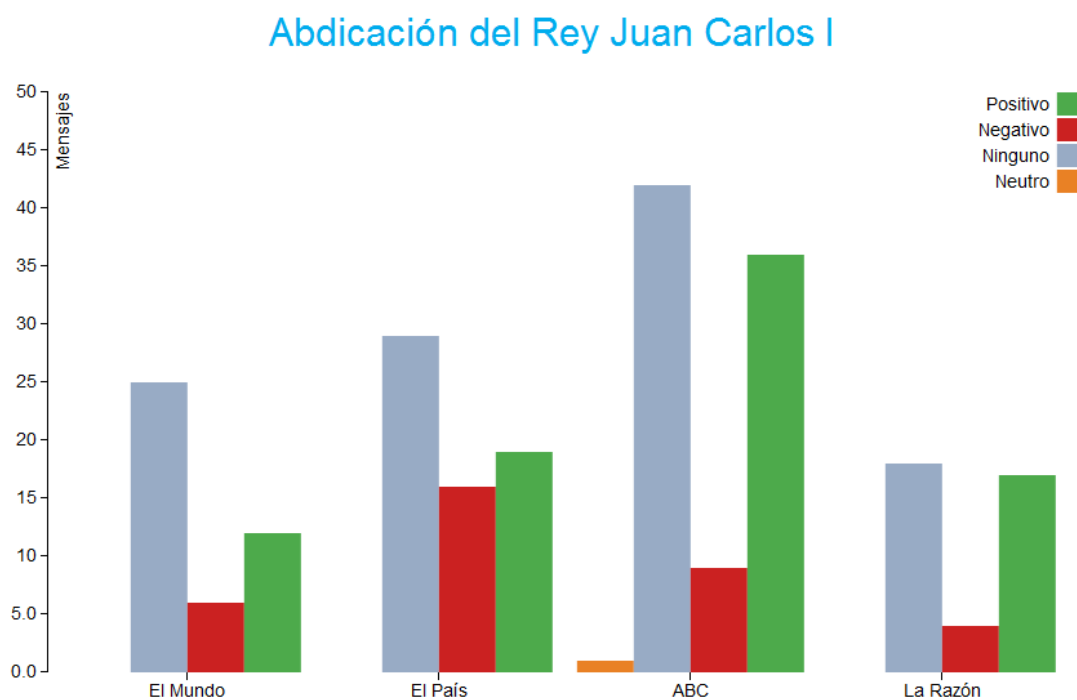


## Independentismo catalán.

### Independentismo catalán



## Abdicación del Rey Juan Carlos I.



## Conflicto de Ucrania.

