



**Para citar este artículo:** Sánchez-Holgado, P., Martín Merino, M., & Blanco Herrero, D. (2020). Del *data-driven* al *data-feeling*: análisis de sentimiento en tiempo real de mensajes en español sobre divulgación científica usando técnicas de aprendizaje automático. *Anuario Electrónico de Estudios en Comunicación Social "Disertaciones"*, 13(1), 35-58. <https://doi.org/10.12804/revistas.urosario.edu.co/disertaciones/a.7691>

# DEL DATA-DRIVEN AL DATA-FEELING: ANÁLISIS DE SENTIMIENTO EN TIEMPO REAL DE MENSAJES EN ESPAÑOL SOBRE DIVULGACIÓN CIENTÍFICA USANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

From Data-Driven to Data-Feeling: Sentiment Analysis in Real-Time of Messages in Spanish about Scientific Communication Using Machine Learning Techniques

Do data-driven ao data-feeling: análise de sentimento no tempo real de mensagens em espanhol sobre divulgação científica usando técnicas de aprendizagem automática

**Patricia Sánchez Holgado**, *Universidad de Salamanca (España)*

[patriciasanc@usal.es](mailto:patriciasanc@usal.es)

**Manuel Martín Merino**, *Universidad Pontificia de Salamanca (España)*

[mmartinmac@upsa.es](mailto:mmartinmac@upsa.es)

**David Blanco Herrero**, *Universidad de Salamanca (España)*

[david.blanco.herrero@usal.es](mailto:david.blanco.herrero@usal.es)

*Recibido: 17 de febrero de 2019*

*Aprobado: 10 de octubre de 2019*



## RESUMEN

Los cambios producidos en los últimos años en cuanto a modelos de comunicación social han llevado a todos los sectores a adaptarse a los nuevos medios para alcanzar a su público. La comunicación de la ciencia no es una excepción. La manera en que se distribuyen contenidos sobre ciencia está cambiando debido a la presencia creciente de tecnologías, y la red social Twitter se ha convertido en un importante aliado debido a su gran volumen de usuarios. En el presente trabajo, se utilizan técnicas de aprendizaje automático para desarrollar un clasificador —que funciona en tiempo real— de sentimiento relacionados con mensajes publicados en Twitter. Para ello, se descargaron 200 000 tweets destinados a construir un corpus de entrenamiento limpio y procesado de 10 000 textos etiquetados, la mitad positivos y la mitad negativos, sobre ciencia en español. El corpus permite entrenar el modelo de aprendizaje automático y construir un prototipo OpScience, capaz de determinar el sentimiento de mensajes publicados en Twitter en tiempo real. Los resultados relacionados con la exactitud del clasificador corresponden al 72 %. Estos resultados pueden ayudar a darle mayor valor a temas de la comunicación científica en un espacio de debate social y predecir intereses o tendencias futuras, como se pudo comprobar en una prueba en enero de 2019.

**Palabras clave:** análisis de sentimiento, aprendizaje automático supervisado, Twitter, comunicación científica.

## ABSTRACT

The changes produced in recent years in social communication models have meant that all sectors have had to adapt to new media to reach their audiences. The communication of science is no exception. The distribution of contents about science is adapting to an increasing presence of technologies, and the social network Twitter has become a necessary ally due to its large volume of users. In this paper, machine learning techniques are used to develop a sentiment classifier of messages posted in real-time on Twitter. To this end, 200 000 tweets were downloaded to build a training corpus of 10 000 clean and processed labeled texts, half positive and half negative, about science in Spanish. This corpus allows the training of the machine learning model and builds a prototype, OpScience, able to determine the sentiment of messages posted on Twitter in real-time. The accuracy results obtained by the classifier is around 72 %. This can help to assess issues of scientific communication in a space of social debate and predict future interests or trends, as observed during the test in January 2019.

**Keywords:** Sentiment analysis, Twitter, supervised machine learning, scientific communication.



## RESUMO

As mudanças produzidas nos últimos anos nos modelos de comunicação social têm levado a todos os setores a se adaptar aos novos meios para alcançar a seu público. A comunicação da ciência não é uma exceção. A maneira em que se distribuem conteúdos sobre ciência está adaptando-se a uma presença crescente de tecnologias, e a rede social Twitter se tem convertido em um importante aliado devido a seu grande volume de usuários. Neste trabalho se utilizam técnicas de aprendizagem automática para desenvolver um classificador de sentimento de mensagens publicados em tempo real no Twitter. Para isto, descarregaram-se 200 000 tweets destinados a construir um corpus de treino limpo e processado de 10 000 textos etiquetados, metade positivos e metade negativos, sobre ciência em espanhol. O corpus permite treinar o modelo de aprendizagem automático e construir um protótipo, OpScience, capaz de determinar o sentimento de mensagens publicados no Twitter em tempo real. Os resultados de concordância do classificador situam-se em um 72 %. Isto pode ajudar a valorar temas de comunicação científica em um espaço de debate social e predecir interesses ou tendências futuras, como se conseguiu comprovar em uma prova em janeiro de 2019.

**Palavras-chave:** análise de Sentimento, aprendizagem automática supervisionada, Twitter, comunicação científica.

---

## Introducción

Twitter es uno de los mayores servicios de información en tiempo real de Internet, y es continuamente alimentado por millones de usuarios procedentes de todo el mundo. Cada uno de esos usuarios publica mensajes de texto con sus opiniones, inquietudes, información o simplemente con su devenir diario. Se trata de una gran cantidad de datos que, en su mayor parte, son públicos y despiertan el interés de los investigadores de datos por su enorme potencial de generar conocimiento. Así, es un reto tratar la gestión de los datos masivos en la red y, con ello, buscar maneras de comprender la sociedad a través del análisis de los datos.

La cultura de los datos promete ser un revulsivo para las estrategias empresariales y parece que llegará a tocar a todos los sectores para modificarlos sustancialmente, y así crear un nuevo sistema de toma de decisiones basadas en datos (*data-driven*). Sin embargo, al utilizar métodos y técnicas computacionales aplicadas a la investigación en ciencias sociales podemos ir un paso más allá, no solo para estudiar los datos, sino para analizar cuáles son los sentimientos que aportan (*data-feeling*), de manera que podamos pasar de decisiones basadas en datos a decidir con base en lo que indican los sentimientos de los mensajes y así percibir a través de los datos digitales lo que los usuarios opinan o piensan en el mundo real.



Este trabajo trata el estudio del sentimiento que generan los mensajes de texto publicados en la red social Twitter —los tweets— mediante el desarrollo de un clasificador en tiempo real que hace uso de técnicas de aprendizaje automático, denominado OpScience. Nos centramos en un tema en concreto para este estudio: la comunicación de la ciencia, un área que aún está descubriendo todo lo que la web 2.0 y las redes sociales pueden ofrecer para mejorar su despliegue y llegar a más diversos y numerosos públicos.

En la primera parte del artículo, hablaremos brevemente de los antecedentes que nos han llevado a plantear esta investigación, como el estudio de la opinión pública a gran escala y su aplicación en la comunicación de la ciencia a través de la red social Twitter. A continuación, se repasan los estudios previos relevantes en el campo del análisis de sentimiento en ciencias sociales y el procesamiento del lenguaje natural, en específico en torno al idioma español. Posteriormente, planteamos la metodología del estudio, centrada en el desarrollo de un corpus específico destinado al entrenamiento de un modelo de aprendizaje automático supervisado. Este modelo alimenta el prototipo OpScience, que clasifica los mensajes publicados en Twitter en tiempo real, indicando la polaridad del sentimiento (positivo-negativo). Se exponen los resultados obtenidos de la evaluación del modelo clasificador y el caso de prueba del prototipo con datos nuevos en tiempo real, para finalizar con las conclusiones obtenidas.

## Antecedentes

### La comunicación de la ciencia

En los últimos años, los cambios profundos en los modelos de comunicación social han provocado que todos los sectores tengan que adaptarse a los nuevos medios para seguir estableciendo contacto con su público. La comunicación de la ciencia también está inmersa en ese proceso. La incorporación de las redes sociales en las estrategias de divulgación científica aporta a los propios investigadores tres elementos fundamentales para su desarrollo profesional: *visibilidad*, pues la capacidad de alcance de los mensajes compartidos en las redes resulta exponencial; *impacto*, porque el número de académicos e investigadores que pueden acceder a un trabajo científico tiene influencia en la posibilidad de ser citado y, a mayor impacto, mayor posibilidad de obtener citas; y, por último, *identidad digital*, pues a través de la proyección social del investigador, configurada por la visibilidad y el impacto conseguidos, se construye una sólida identidad digital.

Los públicos interesados en los avances de la ciencia exigen cada vez más información, y los científicos están adoptando roles que hasta no hace mucho eran territorio exclusivo de los profesionales de la comunicación de la ciencia (Brossard & Scheufele, 2013). Las tecnologías se incorporan a los procesos de trabajo como forma de ampliar la difusión y el impacto de las investigaciones (Côté & Darling, 2018; Darling, Shiffman, Côté & Drew, 2013; Pont-Sorribes, Cortiñas-Rovira, & Di Bonito, 2013). Las redes sociales forman parte ya del ecosistema habitual de los medios de comunicación de nuestra sociedad, por lo que su incorporación en las estrategias de comunicación y divulgación de la ciencia, aunque lenta, también va progresivamente en aumento. Los comunicadores de ciencia usan cada vez más la tecnología digital y las redes sociales (Ribas, 2012), para fomentar debates en la esfera pública sobre temas de ciencia y divulgación en general.



En tiempos de posverdad, la ciencia gana peso y los mecanismos que afectan a la opinión pública han sido ampliamente estudiados (Fowks, 2017). Sin embargo, el análisis de la polaridad de opiniones en Twitter es un campo que despegó y del que existe poca investigación previa en idioma español.

### Estudio de la opinión pública a gran escala

El análisis de la opinión pública a partir de datos a gran escala, extraídos de plataformas digitales y medios sociales, ha generado interés como objeto de estudio desde hace un par de años (Bollen, Mao, & Pepe, 2011; O'Connor, Balasubramanian, Routledge, & Smith, 2010; Whitman Cobb, 2015).

Los expertos en minería de opinión han realizado esfuerzos por estudiar la correlación entre los resultados de las encuestas de opinión y los mensajes de Twitter a través de una herramienta específica: el análisis de sentimiento. En uno de estos estudios se observó que un detector de sentimiento basado en Twitter llega a replicar la confianza del consumidor (O'Connor et al., 2010). Sin embargo, la mayoría de estos estudios se basa en la clasificación manual o en el análisis automatizado de contenido con la utilización de diccionarios que etiquetan palabras (por ejemplo, al dar un valor *a priori* negativo o positivo a cada palabra).

El desarrollo de este trabajo persigue la creación de un prototipo de herramienta de análisis denominado OpScience, con un enfoque de aprendizaje automático, supervisado y aplicado a grandes volúmenes de datos generados en tiempo real en la red social Twitter. Estos nuevos enfoques que unan métodos computacionales y *big data* se empiezan a abordar en la investigación de la comunicación en las ciencias sociales y en las consultoras privadas para el seguimiento de la opinión pública, estudios políticos y marketing hace poco. Recientes publicaciones muestran el camino a seguir (Arcila-Calderón, Barbosa-Caro, & Cabezuelo-Lorenzo, 2016), y abren nuevas perspectivas para su aplicación en otros sectores como el de la comunicación científica.

### Microblogging y Twitter

Tres billones de personas alrededor del mundo expresan sus pensamientos y opiniones de modo habitual a través de redes sociales. Twitter es una de esas redes sociales; se trata de un servicio de *microblogging* que combina características de blog, mensajería instantánea y redes sociales. Además, este servicio ha crecido de manera exponencial desde su lanzamiento en 2006. Los usuarios de Twitter generan contenidos basados en textos breves, de un máximo de 280 caracteres —hasta noviembre de 2018 eran 140—, sobre cualquier tema y en tiempo real. Un mensaje o tweet puede alcanzar en pocos minutos una audiencia muy elevada, gracias a que los usuarios comparten, o retuitean, nuevamente los mensajes en una red casi sin fin. El número total de usuarios mensuales activos a principios de 2018 alcanzaba la cifra de trescientos treinta millones. Esta cifra supone un volumen de unos quinientos millones de tweets por día. En España, la cifra de usuarios roza los cinco millones de usuarios.<sup>1</sup> Es una red usada para compartir información y para describir cualquier actividad diaria (Java, Song, Finin, & Tseng, 2007), permite expresar opiniones e intereses en tiempo real (Esparza, O'Mahony, & Smyth, 2012) y su influencia alcanza prácticamente todas las áreas de la vida social, política, económica y educativa, y cubre cualquier tema de debate —deportes, cultura, ocio, ciencia, industria— (Kwak, Lee, Park, & Moon, 2010).

---

1 Más información en: <https://www.omnicoreagency.com/twitter-statistics/>



Es por ello que gran parte de los trabajos que analizan el uso de redes sociales en la comunicación de la ciencia se centran en Twitter: por ejemplo, en cómo los periodistas y los gabinetes de comunicación de los centros de investigación utilizan esta red en específico y otras redes sociales (Kahle, Sharon, & Baram-Tsabari, 2016; Montenegro & Escudero, 2013; Quiñónez Gómez & Sánchez Colmenares, 2016); el uso que científicos y académicos hacen de Twitter (Bonetta, 2009); los estudios de caso de las propias universidades (Campos-Freire & Rúas-Araújo, 2016); o el impacto de Twitter en la publicación científica (Liang et al., 2014; Mandavilli, 2011).

## Twitter como herramienta de comunicación científica en España

Lo que convierte a Twitter en una red relevante es su volumen de usuarios, la generación libre de contenidos y la manera en la que es posible compartir información en tiempo real. La inmediatez es una de las ventajas principales, pero también tiene como desventaja la saturación. Por ello, como herramienta de comunicación científica tiene un enorme potencial y comienza a tomar protagonismo, pero a su vez requiere de un uso eficiente. Un estudio publicado por *Nature* en 2014 ya avanzaba que el 50 % de los usuarios de Twitter que son investigadores lo utilizan para hacer seguimiento de debates sobre su área, y un 40 % lo utiliza como medio para comentar avances o debatir (Baker, 2015).

Como herramienta de comunicación de la ciencia, se ha estudiado el uso de Twitter tanto en revistas científicas (Arcila-Calderón, Calderín-Cruz, & Sánchez-Holgado, 2019; Segarra-Saavedra, Tur-Viñes, & Hidalgo-Marí, 2017), como desde el punto de vista de los agentes y las redes de comunicación implicados en el manejo diario (Pérez-Rodríguez, González-Pedraz, & Alonso Berrocal, 2018). En este último punto, se ha observado que el uso de Twitter no está enfocado en específico en la divulgación de contenido científico, sino en la promoción de productos y eventos. Además, si se analiza el flujo de contenidos, es posible identificar como influenciadores más relevantes, no a los científicos de prestigio o a las instituciones de investigación de España, sino a comunicadores y divulgadores con elevada presencia pública, así como a las entidades que promueven la cultura científica como FECYT, Agencia SINC o el Museo Nacional de Ciencia y Tecnología.

Twitter es la red más usada por periodistas de ciencia para la promoción de sus artículos (Pont-Sorribes et al., 2013). Las universidades y centros de investigación también lo utilizan, y los científicos que buscan ganar impacto (Peters, Dunwoody, Allgaier, Lo, & Brossard, 2014). Los comunicadores de ciencia usan cada vez más la tecnología digital y las redes sociales (Ribas, 2012), y además está demostrado que los primeros datos sobre una primicia científica o técnica se hacen públicos en Twitter (Jarreau, 2015). El uso de Twitter como estrategia de comunicación científica muestra una tendencia al alza con la apertura cada vez mayor de perfiles de usuarios de esta área (Segarra-Saavedra et al., 2017).

Por otro lado, la opinión que se muestra en Twitter tiene una vinculación directa con la actualidad científica nacional e internacional (Pérez-Rodríguez et al., 2018). Esto es un detalle importante para tener en cuenta en el estudio de la opinión en tiempo real en la red social, ya que es un punto de anclaje con el cual comparar el devenir diario y sacar conclusiones. Las redes sociales tienen un importante papel en la comunicación científica y en el estudio de los sentimientos de los mensajes y opiniones que se emiten. Estos datos son relevantes para evaluar el interés del público y detectar pulsos de la conversación sobre temas candentes.



## Estado de la cuestión

En 2010, se empezó a utilizar el *microblogging* como corpus —conjunto cerrado de textos destinado a la investigación científica— válido para el análisis de sentimiento, y se estableció así una nueva línea de trabajo que puede aportar una fuente de datos de gran valor y potencial futuro (Pak & Paroubek, 2010). Posteriormente se realizó una profunda puesta en situación en un campo que se presentaba como una tendencia imprescindible en la investigación con análisis de sentimiento (Henríquez Miranda & Guzmán, 2017; Martínez-Cámara, Martín-Valdivia, Ureña-López, & Montejo-Ráez, 2014).

El análisis de los mensajes de Twitter está siendo utilizado como fuente primaria para múltiples investigaciones, que abarcan desde el papel que juegan los distintos tipos de usuarios en la difusión de la información (Cha, Benevenuto, Haddadi, & Gummadi, 2012), hasta aplicaciones de clasificación y recuperación de la información (García Esparza, O'Mahony & Smyth, 2012; Lee et al., 2011; Yerva, Miklós & Aberer, 2012), análisis sociológicos (Chen, Yen & Hwang, 2012) y análisis semánticos (Alonso Berrocal, Gómez Díaz, Figuerola, Zazo Rodríguez, & Cordon García, 2012; Narr, De Luca, & Albayrak, 2011).

En cuanto a la clasificación del sentimiento de los mensajes en Twitter, existen antecedentes de estudios que lo han aplicado en diferentes temáticas e idiomas con algunos hallazgos de interés para este trabajo, como son: la inclusión de emoticonos como elemento relevante de apoyo al contexto, y como apoyo para aumentar la precisión del modelo (Go, Bhayani, & Huang, 2009); el uso de los *hashtags* como sistema de etiquetas con las que apoyarse para construir el clasificador (Pak & Paroubek, 2010); el uso de un enfoque supervisado del problema en la misma línea de etiquetas con *hashtags* y revisión de características lingüísticas (Kouloumpis, Wilson & Moore, 2011); el uso de un enfoque basado en la semántica, demostrando que la eliminación de las *stop words* —palabras sin aparente carga de significado, como los artículos 'uno', 'una', 'el' o 'la'— no es un paso necesario y que incluso puede tener efectos negativos (Saif, He, & Alani, 2012); y la demostración que en conjuntos de entrenamiento grandes ( $n = 4000$ ) las muestras aleatorias dan mejores resultados (Van Zoonen & Van der Meer, Toni, 2016).

Un avance interesante en las técnicas de aprendizaje automático fue incorporar un mecanismo de votos y la importación de puntuaciones del texto para crear un híbrido entre las técnicas basadas puramente en el léxico y las aproximaciones *machine-learning* (Li & Liu, 2012).

La investigación centrada en la construcción de modelos supervisados de aprendizaje automático para clasificar textos en idioma español todavía es escasa, pero va en aumento. Podemos ver estudios realizados utilizando opiniones sobre restaurantes (Dubiau & Ale, 2013), críticas de cine con aprendizaje no supervisado (Cruz Mata, Troyano Jiménez, de Salamanca Ros, & Ortega Rodríguez, 2008), o con descarga libre sobre todo tipo de texto en Twitter, aplicando también corpus con emoticonos como novedad (Sidorov et al., 2013).

La mayoría de estudios que clasifican textos en español tienen su origen en el proyecto TASS (*sentiment analysis at global level*), un taller de participación organizado por la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) cada año, que tiene como finalidad promover la investigación y el desarrollo de nuevos algoritmos, recursos y técnicas, y que se aplica al análisis de opiniones en español (García Cumbreñas et al., 2016). Para ello, se construyó un gran corpus en español que marcó el inicio de las investigaciones más concretas en esta área, en 2014, conocido como COST (*Corpus Of Spanish Tweets*), que genera anualmente varios ejemplos (Díaz-Galiano,



Martínez-Cámara, García-Cumbreras, García-Vega & Villena-Román, 2018; Hurtado, Pla & Buscaldi, 2015; Martínez-Cámara, Díaz-Galiano, García-Cumbreras, García-Vega, & Villena-Román, 2017; Rosá, Chiruzzo, Etcheverry, & Castro, 2017).

Sin embargo, en el campo de la comunicación de la ciencia en español no existen análisis específicos previos o intentos de clasificación de mensajes con base en el sentimiento, por lo que vamos a intentar tratar este tema desde una perspectiva basada en un corpus creado a tal efecto.

La pregunta de la que partimos es la siguiente: ¿podemos analizar una parte de los datos públicos disponibles en la red social Twitter para conocer actitudes, opiniones y sentimientos en torno a los temas de comunicación de la ciencia que se comparten en esa red, y avanzar hacia la predicción de tendencias o comportamientos futuros?

## Metodología

En el desarrollo del prototipo OpScience se utiliza aprendizaje automático supervisado, es decir, en la generación de un conjunto de datos de entrenamiento (corpus) compuesto por una instancia (mensaje) y una clase (sentimiento positivo o negativo). De este corpus se obtiene un modelo que puede ser capaz de predecir a qué clase pertenecerá una nueva instancia de entrada. En este caso, se trata de un modelo de clasificación binaria (positivo/negativo).

Los métodos que se basan en aprendizaje automático representan el problema de extracción de relaciones como una tarea de clasificación, y utilizan algoritmos de aprendizaje para extraer información relevante de los textos automáticamente. Sin embargo, la ejecución de estos algoritmos implica un alto coste computacional, y sus resultados dependen de la disponibilidad de grandes conjuntos de datos previamente etiquetados (corpus). Los ejemplos que van a entrenar a los algoritmos deben representarse con un formato apropiado para hacerlo de la mejor forma posible. El objetivo del aprendizaje supervisado es crear una función que sea capaz de predecir cuál sería el valor de un elemento de entrada después de haber sido entrenado con el clasificador de sentimiento. En este proceso, se utilizaron Python 2.7 y las librerías Tweepy,<sup>2</sup> NLTK<sup>3</sup> y ScikitLearn.<sup>4</sup>

En el desarrollo de este trabajo se realizaron tres tareas: 1) la construcción de un nuevo corpus de textos de tema científico en lenguaje español, clasificado por sentimiento positivo o negativo, para el entrenamiento del modelo de aprendizaje automático; 2) una vez completado el corpus, se entrenó el modelo al dividir el corpus en *training* y *test* (70 % - 30 %) en dos pasos: el entrenamiento, para que el modelo aprenda al empleando un conjunto de datos manualmente etiquetado, y la validación, proceso en el cual se pueden ajustar los parámetros del modelo y que sirve para evitar problemas, como por ejemplo el sobreentrenamiento, que aparece cuando el modelo aprende los datos de entrenamiento, pero no es capaz de generalizar y, por lo tanto, de predecir correctamente nuevas muestras; 3) por último, se ha implementado el prototipo clasificador de sentimientos en tiempo real y se ha puesto a prueba con datos que no ha visto previamente. La metodología, por tanto, se divide en tres etapas de trabajo.

---

2 Más información en: <http://www.tweepy.org/>

3 Más información en: <https://www.nltk.org/>

4 Más información en: <http://scikit-learn.org/>



### Fase 1. Creación del corpus de textos científicos en español

i. Adquisición de los datos: Como ya se ha avanzado previamente, se utiliza la red social Twitter para construir un corpus que permita entrenar modelos de aprendizaje automático y desarrollar el clasificador de sentimientos. Twitter permite actualmente enviar mensajes cortos de un máximo de 280 caracteres. Para la descarga de datos, nos conectamos con la aplicación que facilita Twitter a los desarrolladores: Application Programming Interface (API) Twitter,<sup>5</sup> utilizando la versión API Streaming (conexión en tiempo real) y la versión API Rest (histórico de datos) en la búsqueda de información.

Los datos que suministran las API están en formato [JSON](#). El conjunto es un listado de tweets que contiene metadatos seleccionados del *tweet object*:<sup>6</sup> el mensaje (*text*), junto con la información sobre el usuario que lo emite (*user*, *screen\_name*), fecha de creación (*created at*), identificador, objetos multimedia, geolocalización, etc.

Se realizó una descarga de datos en varios períodos entre marzo y junio de 2018. Para filtrar mensajes sobre temas de ciencia se utilizaron como palabras clave las siguientes: *ciencia* y *comunicaciencia*, que son habituales en mensajes de divulgación, así como otras palabras relacionadas con eventos especializados en comunicación de la ciencia, como son los hashtags de “Pint of Science” (*PintofScience*, *PoS*, *PoS18* y todas sus variaciones por provincias españolas) y “Ciencia en redes” (*CienciaenRedes*, *CnR18*), de los más populares en España.

En la búsqueda de un mayor volumen de mensajes para lograr el equilibrio entre positivos y negativos, se utilizó la versión API Rest para la descarga de mensajes de cuentas concretas que se había detectado que contenían más diversidad de sentimiento: *maldita ciencia*, *cienciamento*, *ciencia con futuro*, *ciencia por el pueblo*, *stop pseudo ciencias*, *materia ciencia*, *agencia dicyt*, *escépticos*, *agencia sinc* y *principia*.

El *dataset* (conjunto de datos) total obtenido, previos a la limpieza para la formación del corpus, estaba formado por más de 200 000 tweets, como se puede ver e detalle en la tabla 1.

**Tabla 1. Total dataset de mensajes descargados**

	Idioma: Español
Tweets descargados en Streaming	171 459
Tweets descargados en Rest	37 292
Total de tweets descargados	208 751

Fuente: Elaboración propia.

5 Más información en: Twitter Application Manager: <https://apps.twitter.com/>

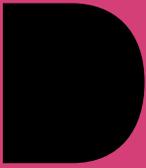
6 Tweet Object, más información en: <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html>



ii. Preprocesamiento de los datos: El dataset estaba compuesto por los mensajes en bruto, por lo que necesitaba limpieza previa antes de etiquetarlo. Para entrenar el modelo, se estableció como volumen de tweets objetivo un corpus de al menos 5000 mensajes etiquetados como positivos y 5000 mensajes etiquetados como negativos (10 000 en total), por lo que al finalizar esta etapa de preprocesado, en el momento en que se alcanzó esa cifra, se detuvo el proceso de etiquetado y se archivó el resto del dataset.

Se realizó un preprocesado de varios pasos, de manera que la entrada de datos de cada etapa fuera la salida de la etapa anterior, y filtramos los atributos que nos interesan para conformar el corpus de texto de temas científicos:

- a. Se almacenaron los tweets en texto csv, y la descarga de los datos se almacenó en archivos .txt, por lo que una vez tenemos todos los ficheros guardados, estos se deben unificar en formato csv para tratarlos y poder hacer posteriormente clasificaciones manuales.
- b. Formatos UTF/ANSI: Revisión del *dataset* total de los formatos para asegurarnos que todo esté en formato de codificación de caracteres UTF8, para que no tengamos posteriormente incidencias.
- c. Idioma español: Revisión de aquellos elementos que no son textos escritos en castellano directamente, por ejemplo, aquellos mensajes que incluyen una palabra clave como “ciencia”, pero están escritos en otros idiomas, especialmente lenguas autonómicas o portugués, eliminamos todos los que no son inequívocamente idioma español.
- d. Textos en minúsculas: Pasamos todo el texto a minúsculas para que esté plenamente unificado;
- e. Retweets: Revisión de los mensajes para suprimir cualquier retweet que exista dentro de ellos. En los *scripts* principales se había filtrado esta opción, pero no fue el caso en las descargas desde REST, por lo que aún podíamos tener mensajes no originales en los datos que se debían eliminar del *dataset*.
- f. Supresión de posibles duplicados con R: Para evitar tener posibles elementos repetidos, hacemos un primer barrido utilizando R y R Studio por rapidez. En las descargas de los datos hay muchos elementos con el mismo contenido que otros, debido principalmente a que los usuarios los comparten desde medios de comunicación y hay seguimiento o respuestas. Por ello se intentó eliminar los que fueran idénticos a través de un *script*.
- g. Otros preprocesamientos: Algunas mejoras que pueden resultar necesarias y efectivas para mejorar el rendimiento del clasificador han sido las siguientes: corregir errores gramaticales, eliminar números irrelevantes o cifras que no aportan valor al texto en lo referido a clasificar su sentimiento, eliminar fechas y determinados signos de puntuación, emoticonos, caracteres especiales, caracteres repetidos varias veces consecutivas o interjecciones sin valor para nuestro objetivo. Este paso de carácter general se puede realizar junto al siguiente, la clasificación manual, que ya es exhaustiva.
- h. Tokenización: El proceso de tokenizar —esto es, separar los textos en tokens o componentes léxicos— se realiza dentro del clasificador para unificar tareas.
- i. Clasificación manual del sentimiento del texto: En el último paso, debimos realizar un proceso manual con el *dataset* de 10 000 textos, puesto que se clasificaron todos los elementos según el sentimiento identificado como positivo o negativo. En esta misma revisión manual, uno por uno, filtramos también aquellos mensajes cuya temática no fuera exclusivamente del sector de la ciencia. Es el paso más largo y laborioso de todos, ya que no hay otra manera de hacerlo sino a través de dos codificadores humanos, con un perfil formativo de estudiantes de doctorado y habituados al manejo diario de Twitter. Los codificadores siguieron instrucciones de un libro de códigos para clasificar en positivo o negativo los tweets.



El primer codificador realizó el etiquetado de los 10 000 textos en su totalidad, mientras que el segundo codificador realizó el etiquetado de un 10 % de los textos. Se realizó una prueba de acuerdo con intercodificadores a través del coeficiente *Alpha de Krippendorff* (Krippendorff, 2011), obteniendo un valor  $\alpha = 0,89$ , calculado con el programa SPSS y la macro escrita por Andrew Hayes (Krippendorff & Hayes, 2007). Se considera como suficientemente confiable un valor  $\alpha \geq .70$  (Krippendorff, 2004). Aquellos tweets que no tuvieron acuerdo entre los dos codificadores finalmente fueron descartados para el entrenamiento del modelo.

Las negaciones en una oración cambian su sentido. Aparentemente es una tarea sencilla buscar los negadores (“no”, “pero”, “es falso que”, etc.). Sin embargo, su distribución en el texto puede ser demasiado variable. No se han suprimido los términos negativos en este proceso, ni se han modificado. Queremos observarlos en su contexto y tal como se redactaron en su origen.

El producto final de esta fase es el listado de documentos (oraciones) que llamamos corpus, compuesto por 5000 positivos y 5000 negativos (figs. 1 y 2). El corpus definitivo utilizado se encuentra accesible en abierto en un repositorio en línea en Github: <https://github.com/patriciasanc/opsience>.

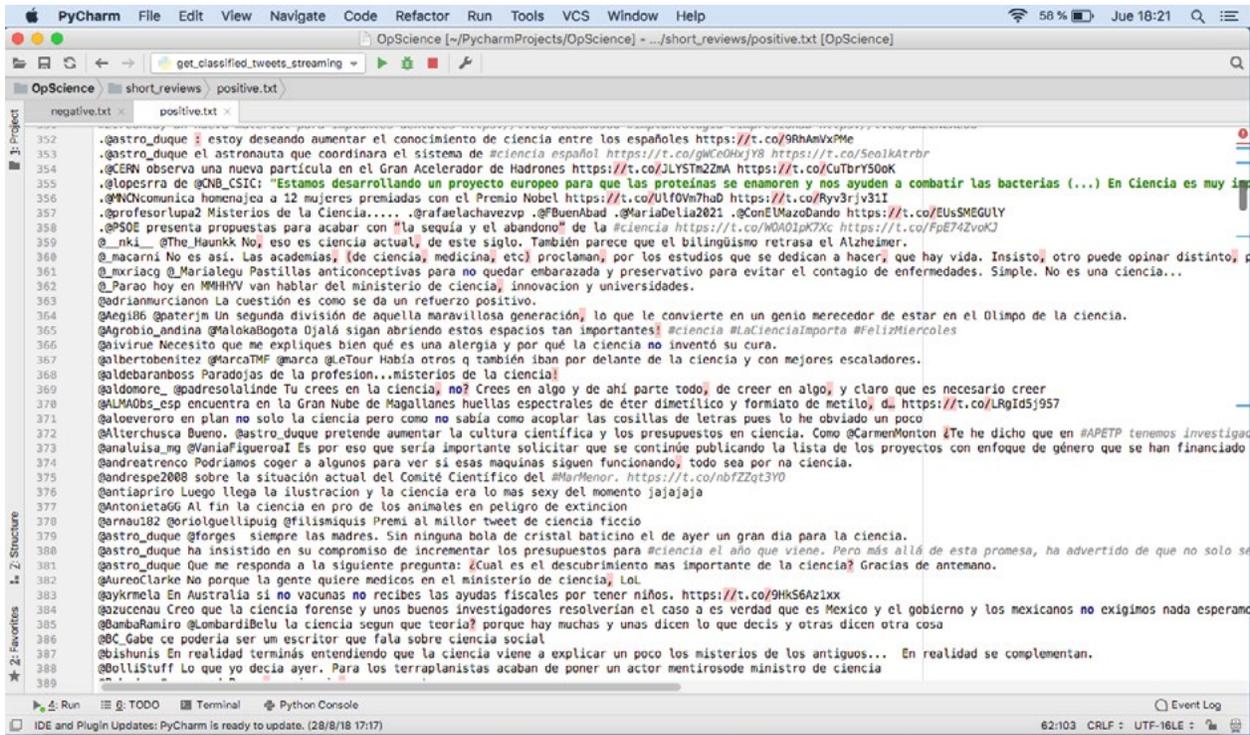


Figura 1. Fragmento del corpus de etiquetados positivos

Fuente: Elaboración propia.

```

1 @MareaGranate Si no, nadie nos oye, y demasiado poco importa ya
2 @PortilloFernan, no preste apoyo público a tal disparate. https://t.co/ft3vuiDGt5
3 @NASA anuncia la desintegración de la sonda #Cassini en atmósfera de Saturno
4 @gaitaarino @cientifico estamos gritando mucho por la ciencia
5 @AlfonsoBastiso pregúntele a @GuillemaAnglada, uno de los 10 científicos del año para @NatureNews, si está fuera por "amplitud de miras"
6 @antonio_rvega @yolandamir1 @laplumlilla @bucdesevilla Pero qué barbaridad está diciéndolo? Tiene usted hijos? https://t.co/wXYUCF3SRE
7 @astro_duque en @horas25: Ningún país de Europa en la crisis ha recortado en investigación. https://t.co/dBIaAg3dvt
8 @astro_duque: "Se ha tratado bastante mal al sistema de conocimiento español, a los investigadores de nuestras universidades" https://t.co/KzN1KynNl
9 @CabradoQuimico Hemos elegido ese día para recordar que la Ciencia española está en la cuerda floja. Mucho por hacer...
10 @Calvor1 Exacto. De ahí el torpe intento de negar la existencia del informe: la Real Academia Nacional de Farmacia. https://t.co/DQLHLN5d6h
11 @CarlosChNav @arturogolada @AytoTafalla Nadie tiene que sembrar alarmismos en Navarra a costa de los recursos municipales.
12 @cescept @apetp_ A veces, se llama dureza a hablar claramente. https://t.co/uV7fevVP7D
13 @cescept Apetece el sarcasmo sobre la posibilidad de detener el envejecimiento de golpe. https://t.co/UW06v3IVc
14 @cescept Es espantoso. El fanatismo produce víctimas.
15 @cescept Esperemos que se tomen medidas muy serias contra la difusión de bulos y de pseudociencias.
16 @cherja1 Para nosotros es inadmisible que se juegue con la salud de las personas difundiendo graves errores https://t.co/1nTFUJGcjj
17 @CincoDiascom nos preocupa que el artículo pueda seguir generando una alarma innecesaria.
18 @CNIO_Cancer Desde 2013, la doctora en Biología Esther Calonge se ha quedado en paro en tres ocasiones. https://t.co/011BKUCxqV
19 @CNIO_Cancer Me temo que en la ciencia española de hoy predominan el conformismo, el pánico a equivocarse y el miedo. https://t.co/UiImKwJfph6
20 @CFMadríd Es indignante que profesionales con la formación científica que reciben los farmacéuticos caigan en esto https://t.co/Mj1xtbXyEH
21 @CSanzAndrea @CEPAC_ Hemos protestado contra este tipo de charlas horribles en ediciones anteriores https://t.co/jR442YyNR8
22 @CSanzAndrea @omph.es Esas conexiones esenciales omnipotentes son el certificado de exención de estudios.
23 @Cuervo23662943 En realidad no ha sido reconocida, sino que se le ha asignado un número con fines administrativos. https://t.co/T9VqDyktE0
24 @cupumad @R_vanGrieken Mientras, el resto de la sociedad sufriendo recortes en Sanidad, Educación, Dependencia, Ciencia https://t.co/evZia8LYr0
25 @Daurnmith Si fuera por eso, no habrían invitado a ese ejemplo de anticiedad. Lo quitan de cartel por puras razones de imagen. https://t.co/ZqcapVNVpZ
26 @drfjmed @cescept @CIBBarcelona Anda mal y desencamina a profesionales y pacientes. https://t.co/kP5AFBDEHx
27 @emitiendo No queremos que te calles... preferimos que te unas a nosotros en intentar cambiar TODAS esas malas políticas https://t.co/kAd51NTpwZ
28 @ffranvazna @luciapediatra Y un rechazo clarísimo de las instituciones, empezando por las municipales que le ceden espacios https://t.co/KV1514mRAD
29 @GEPAC_ Es fundamental trabajar para no dejar paso libre ante la opinión pública a los curanderos y vendedores de humo. https://t.co/0XUMHtWyi5
30 @guillermardos ¿Desde cuándo se cierra la Filosofía a otras disciplinas? https://t.co/4TBuW9PE1
31 @hectorjuavp Las barbaridades que hay que leer. El día en que alguien comprenda que hay riesgos importantes asociados. https://t.co/wMezv0a4KV
32 @Immershine JAJAJAJAJ que no servias para ciencia dijo, que se pudra :3
33 @Javier678592645 Es una vergüenza. En vez de promover medidas que realmente complementan los tratamientos https://t.co/1hLKzYp5Mw
34 @joaquina_Sevilla Lo mismo sucede con los actos crueles y atroces: se califican de "inhumanos" https://t.co/1zKK17p5qT
35 @josefemendi @Planeta_Futuro Y los que se escudan tras el término "libertad" para sembrar el alarmismo y la confusión. https://t.co/anBPke9d64
36 @juanfcfcs @empleogob @sanidadgob @cescept @omph.es @apetp_ @ej_molina_c @julioBasulto_DN Es inaceptable, @empleogob.
37 @jvicenteprieto @GranCanariaCab Es repugnante que se pretenda dar respetabilidad a semejante superstición sin fundamento https://t.co/AL068fjXxe
38 @Mivicentemrieto Centrar el problema en el intruismo profesional es muda corta. ¿No vemos los disparates que causan? https://t.co/cnLrUhhFv35n5

```

Figura 2. Fragmento del corpus de etiquetados negativos

Fuente: Elaboración propia.

## Fase 2. Aprendizaje automático supervisado

En estudios previos, también se ha observado que en los algoritmos de clasificación con Naive Bayes se obtienen los mejores resultados para corpus pequeños, pero su *performance* o desempeño decrece levemente para los tamaños de corpus más grandes en comparación con MaxEnt y SVM, con los que resultados mejoran a medida que crece el tamaño de corpus y se obtiene el máximo desempeño por la experiencia (Dubiau & Ale, 2013).

Ahora, entremos un poco más a analizar el entrenamiento del modelo. El conjunto de datos utilizado para construir el modelo de clasificación se denomina “conjunto de entrenamiento”. Cada caso o texto se corresponde con una clase determinada, puesto que los casos de entrenamiento están etiquetados con su atributo de clase; esto es, el sentimiento positivo o negativo. El objetivo que perseguimos es que nuestro clasificador indique el sentimiento positivo o negativo de los mensajes que se van descargando en tiempo real de Twitter, por lo que previamente vamos a entrenar el modelo con el corpus de 10 000 elementos (5000 mensajes etiquetados como positivos y 5000 mensajes etiquetados como negativos). La medición del modelo nos dará información para valorar los resultados obtenidos.

Se separa un  $q$  % de datos para validar los modelos:  $q = 30$  (entrenamiento  $(100 - q) = 70$  %; Validación test ( $q = 30$  %). Los datos se separan al azar para evitar cualquier orden o estructura subyacente: 7000 textos para entrenamiento y 3000 textos para el testeo.



Para construir los algoritmos de entrenamiento del modelo utilizamos las bibliotecas NLTK<sup>7</sup> (Natural Language Tool Kit) (Bird, Klein, & Loper, 2009) y Scikit-Learn en Python (Pedregosa et al., 2011). A su vez, el modelo de aprendizaje automático presentado utiliza varios algoritmos: Original Naive Bayes, Naive Bayes for multimodal models, Naive Bayes for multivariate Bernoulli models, Logistic Regression, Linear Support Vector Classification (svc) y Linear classifiers with stochastic gradient descent -sgd- training.<sup>8</sup> Se ha utilizado una combinación de los algoritmos clasificadores, que es una técnica habitual, conocida como votación por intervalos de características, puesto que se crea un sistema de votación en el que cada algoritmo tiene un voto y la clasificación que tiene más votos es la elegida. Se construyen intervalos para cada caso en la fase de aprendizaje y el intervalo “vota” por un atributo de clase (sentimiento positivo o negativo) en la fase de clasificación. Cada caso es tratado de forma individual e independiente del resto. Se diseña un sistema de votación para combinar las clasificaciones individuales de cada caso por separado.

Otro elemento novedoso que se introdujo en el modelo es la capacidad de obtener un intervalo de confianza que nos indica, dentro de una escala de entre 0 y 1, la seguridad de la clasificación que se ha realizado como positiva o negativa, y que sirve, así, de control para falsos positivos o decisiones dudosas.

Para evitar tener que entrenar el algoritmo cada vez que se ejecute el *script*, se almacenaron los datos con la herramienta pickle de Python.<sup>9</sup> Después de ejecutar el *script*, los resultados de exactitud del modelo se muestran en la tabla 2:

**Tabla 2. Salida de datos del clasificador: exactitud de los algoritmos del modelo**

<code>('Original Naive Bayes Algo accuracy percent:', 72,64276228419655)</code>
<code>('MNB_classifier accuracy percent:', 72,24435590969456)</code>
<code>('BernoulliNB_classifier accuracy percent:', 72,80876494023904)</code>
<code>('LogisticRegression_classifier accuracy percent:', 71,87915006640107)</code>
<code>('LinearSVC_classifier accuracy percent:', 70,4515272244356)</code>
<code>('SGDClassifier accuracy percent:', 71,14873837981408)</code>

Fuente: Elaboración propia.

El clasificador se entrena con el uso de los ejemplos y obtenemos unos resultados de fiabilidad de los algoritmos utilizados, que tal como vemos en la tabla de salida del sistema (tabla 2) están entre un 70 % y un 72 %. Las palabras que son más características en su polaridad también nos las muestra la salida del modelo y las podemos ver en la Tabla 3:

7 Más información de nltk en: <https://www.nltk.org/>

8 Más información de SciKit Learn en: [https://scikit-learn.org/stable/supervised\\_learning.html](https://scikit-learn.org/stable/supervised_learning.html)

9 Más información en: <https://pythonprogramming.net/new-data-set-training-nltk-tutorial/>



Tabla 3. Salida de datos del clasificador: palabras más informativas

Most Informative Features		
mal = True	neg : pos	= 14.1 : 1.0
gobierno = True	neg : pos	= 13.6 : 1.0
potencial = True	pos : neg	= 13.1 : 1.0
galaxias = True	pos : neg	= 12.4 : 1.0
evidencia = True	neg : pos	= 12.1 : 1.0
nuevas = True	pos : neg	= 11.4 : 1.0
peor = True	neg : pos	= 10.9 : 1.0
contratos = True	neg : pos	= 10.9 : 1.0
propiedades = True	pos : neg	= 9.7 : 1.0
mala = True	neg : pos	= 9.6 : 1.0
aves = True	pos : neg	= 9.1 : 1.0
desvela = True	pos : neg	= 9.1 : 1.0
neuronas = True	pos : neg	= 9.1 : 1.0
mentira = True	neg : pos	= 8.9 : 1.0
mecanismo = True	pos : neg	= 8.7 : 1.0

Fuente: Elaboración propia.

### Fase 3: Evaluación del modelo

Los resultados que se han obtenido con este modelo se han valorado con la utilización de métricas como la matriz de confusión, *accuracy* o exactitud, precisión, *recall* o recuerdo y F-Score<sup>10</sup> (Bird et al., 2009).

#### Accuracy

La métrica principal es la valoración de la exactitud o *accuracy* con la que el modelo ha clasificado las instancias en la fase de entrenamiento. Puesto que es un modelo de predicción, nos interesa saber si hace bien las predicciones. La *accuracy* nos muestra cuál es el porcentaje de datos que el modelo ha clasificado correctamente.

$$\text{Accuracy} = \frac{\text{predicciones correctas}}{\text{total de predicciones}}$$

La distribución de las clases que hemos utilizado está equilibrada (5000 positivos y 5000 negativos). Los datos de exactitud (*accuracy*) ya se han apuntado previamente (tabla 2), siendo el más preciso el algoritmo Naive Bayes for multivariate Bernoulli models (72,80 %), seguido de Original Naive Bayes (72,64 %), Naive Bayes for multimodal models (72,24 %), Logistic Regression (71,88 %), Linear classifiers with stochastic gradient descent (SGD) training

10 Métricas nltk disponibles en: <https://www.nltk.org/api/nltk.metrics.html>



(71,15 %) y Linear Support Vector Classification (70,45 %). En el proyecto TASS, se muestra que el nivel máximo de *accuracy* está en cifras similares, rondando el 72 % (García Cumberas et al., 2016). Teniendo en cuenta el conjunto completo de algoritmos del clasificador, el resultado del modelo por votación conjunta es de 72,31 %.

### Matriz de confusión

Compara los valores que se predicen en el modelo con los reales y se representa con una tabla comparativa (tabla 4).

Tabla 4. Matriz de confusión binaria

		Predicción	
		Pos	Neg
Real	Pos	TP	FN
	Neg	FP	TN

Cada columna de la matriz representará el número de predicciones para cada clase realizada por el modelo, mientras cada fila refleja los valores reales por cada clase. Así, los conteos quedan divididos en cuatro clases —TP, FN, FP y TN— que significan lo siguiente: TP (**True Positives**) son el número verdaderos positivos, es decir, de predicciones correctas para la clase +; FN (**False Negatives**) son el número de falsos negativos, es decir, la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II; FP (**False Positives**) son el número de falsos positivos, es decir, la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I; TN (**True Negatives**) son el número de verdaderos negativos, es decir, de predicciones correctas para la clase -.

En nuestro modelo hemos obtenido una matriz de confusión (tabla 5) para el conglomerado de algoritmos según el sistema de votos, para obtener valores totales: TP = 1158, TN = 1047, FP = 465, FN = 342.

Tabla 5. Resultado de la matriz de confusión

	Neg	Pos
Neg	<1047>	465
Pos	342	<1158>

Fila = referencia, Columna = test

Fuente: Elaboración propia.



En este caso, podemos ver que el modelo devuelve 1158 como verdaderos positivos y 1047 como verdaderos negativos. En el otro lado, obtenemos como falsos negativos 342 y como falsos positivos 465. De este modo la *accuracy* equivale a  $(TP + TN) / (TP + TN + FP + FN)$ .

### Precisión

La precisión mide la exactitud de un clasificador. Una mayor precisión significa menos falsos positivos, mientras que una menor precisión significa más falsos positivos. Esto a menudo está en desacuerdo con la recuperación (*recall*), ya que una manera fácil de mejorar la precisión es disminuir el recuerdo o *recall*. Precisión nos indica —de las instancias que se han recuperado en el modelo— qué porcentaje es relevante. Su fórmula es:  $Precisión = TP / (TP + FP)$  (1158 / 1158 + 465). *Precisión vote classifier* = 71,35 %.

- *Precisión de cada algoritmo*: Original Naive Bayes (70,44 %), Naive Bayes for multimodal models (74,07 %), Naive Bayes for multivariate Bernoulli models (69,30 %), Logistic Regression (69,27 %), Linear Support Vector Classification (67,89 %) y Linear classifiers with stochastic gradient descent -SGD- training (66,55 %).

### Recall o recuerdo

Mide la completitud o sensibilidad de un clasificador. Una *recall* más alta significa menos falsos negativos, mientras que una *recall* más baja significa más falsos negativos. Mejorar el recuerdo a menudo puede disminuir la precisión porque es más difícil ser preciso a medida que aumenta el espacio de muestra. El *recall* calcula qué porcentaje de las instancias relevantes fueron recuperadas y su fórmula es:  $Recall = TP / (TP + FN)$  (1158/1158 + 342). *Recall vote classifier* = 77,20 %.

- *Recall de cada algoritmo*: Original Naive Bayes (79,41 %), Naive Bayes for multimodal models (68,35 %), Naive Bayes for multivariate Bernoulli models (80,61 %), Logistic Regression (76,74 %), Linear Support Vector Classification (74,68 %) y Linear classifiers with stochastic gradient descent (SGD) training (78,74 %).

### F-Score

Es conocida como una media armónica. La precisión y el *recall* se pueden combinar para producir una única métrica conocida como medida F, que es la media armónica ponderada de ambas. Su fórmula es la siguiente:  $F-score = 2 * precisión * recall / (precisión + recall)$ . Según los datos del modelo con el algoritmo de votación conjunto podemos calcular:  $F-score = 2 * 0,713 * 0,772 / (0,713 + 0,772) = 74,13 \%$ .

- *F-Score de cada algoritmo*: Original Naive Bayes (74,66 %), Naive Bayes for multimodal models (71,10 %), Naive Bayes for multivariate Bernoulli models (74,53 %), Logistic Regression (72,81 %), Linear Support Vector Classification (71,12 %) y Linear classifiers with stochastic gradient descent (SGD) training (72,13 %).



Resumiendo, los datos que hemos obtenidos en nuestros cálculos, para cada algoritmo, se muestran en la tabla 6:

**Tabla 6. Resumen de métricas de evaluación del modelo creado**

Algoritmos	Accuracy	Precisión	Recall	F-score
Vote Classifier	72,31 %	71,35 %	77,20 %	74,13 %
Original Naive Bayes	72,64 %	70,44 %	79,41 %	74,66 %
Naive Bayes for multimodal models	72,24 %	74,07 %	68,35 %	71,10 %
Naive Bayes for multivariate Bernoulli models	72,80 %	69,30 %	80,61 %	74,53 %
Logistic Regression	71,88 %	69,27 %	76,74 %	72,81 %
Linear Support Vector Classification (svc)	70,45 %	67,89 %	74,68 %	71,12 %
Linear classifiers with stochastic gradient descent - sgd - training	71,14 %	66,55 %	78,74 %	72,13 %

## Resultados

Para comprobar la validez del modelo y ponerlo a prueba, se procedió a recolectar datos nuevos reales y se ejecutó el modelo, conectándonos de nuevo al flujo en *streaming* de Twitter para descargar datos en el mes de enero de 2019. Desde un *script* programado conectado a Twitter, se descargaron mensajes filtrados por las palabras clave “ciencia”, “comunicaciencia” y “divulgación”. El modelo permitió entonces clasificarlos en función de la polaridad positiva o negativa de su sentimiento. Los algoritmos usados son probabilísticos, por lo que la estimación de confianza que hace es de que cada mensaje sea positivo en un porcentaje de probabilidad determinado (máximo 100 %, valor 1), de manera que se muestra si el nivel de confianza es superior al 80 % (valor 1) o si por el contrario es inferior a 80 % (valor 0), lo que permite elaborar predicciones con mayor seguridad.

Se obtuvo un volumen total de 8131 tweets, extraídos durante 48 horas y clasificados por el modelo, según se muestra en la Tabla 7.

**Tabla 7. Volumen de tweets clasificados por sentimiento y nivel de confianza en la validación del modelo.**

Sentimiento	Confianza	Tweets
neg	0	928
neg	1	2818
pos	0	1027
pos	1	3358
Total		8131

Fuente: Elaboración propia.



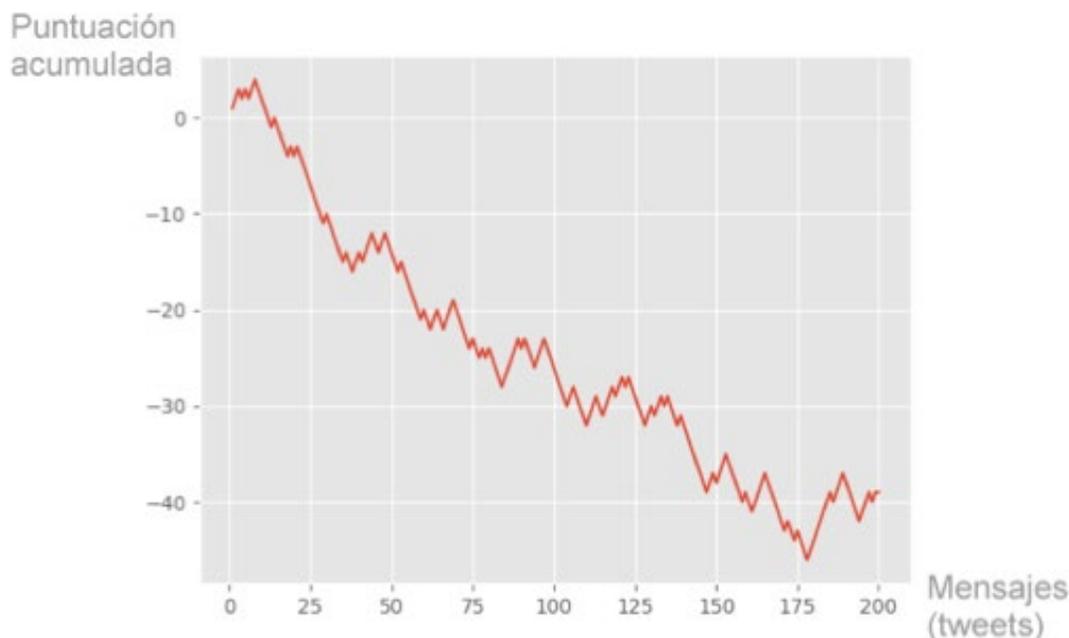
Una muestra de la salida, es decir, de los mensajes clasificados en tiempo real por el modelo al ponerlo en marcha, se puede observar en la tabla 8:

**Tabla 8. Salida de datos clasificados por el modelo en tiempo real**

- Así es como YouTube te ayudará a desengancharte de la plataforma <https://t.co/cyxm7CPH4v> pos 0
- Los científicos descubrieron que todas las personas con ojos azules tienen esta característica única <https://t.co/bdNRKyT5tF> pos 1
- #ciencia Un equipo de científicos de la Universidad de Szeged, en Hungría, descubre un nuevo tipo de neurona <https://t.co/xvddKmjwyq> pos 0
- Saruhashi, la científica que luchó por la igualdad de las mujeres en la ciencia <https://t.co/NO27y0o1EV> pos 1
- ¿Llegó Estados Unidos a la Luna en 1971? Los chinos hallan la respuesta definitiva #Ciencia <https://t.co/PlcOn0fJ36> pos 1
- Una nueva función de una proteína abre una puerta contra el cáncer de mama <https://t.co/n0uRGU4LNH> vía @Elperiodico pos 1
- Así es la oficina inteligente que se adapta a tu estado de ánimo <https://t.co/FLbaNbbnO9> vía @La\_SER Muy buena idea... <https://t.co/Fa97ikXtuP> pos 1
- El 40 % de la gente cree tener recuerdos de bebé, pero son falsos. #ttesi #trauma #terapiapsicologica
- <https://t.co/wLHxwuZsqh> neg 1
- Vamos por mi cuarta Semana Nacional de Ciencia y Tecnología. Ya empiezan los preparativos pos 1
- ¿Sabes cómo llegar a Ciencia al Parque? <https://t.co/CEYTWeMika> pos 0
- ¡Un enigma para la ciencia! Encuentran un tiburón blanco de color rojo (Fotos) <https://t.co/SYaCl7LHf8> <https://t.co/GT1B5pSsX0> pos 1
- La clave del hallazgo de un nuevo método para frenar el envejecimiento #Ciencia <https://t.co/RDw0xOV8Ma> pos 1

Fuente: Elaboración propia.

Para completar el proyecto se incluyó una visualización de los datos en una gráfica en tiempo real (Figura.3), con la utilización de la biblioteca matplotlib para Python. La gráfica muestra, en el eje X, los 200 últimos mensajes descargados y almacenados por el modelo a través de una línea donde podemos ver la tendencia de sentimiento que sigue en el eje Y. En este caso, la tendencia que se observa de sentimiento es variable entre positivo y negativo.



**Figura 3. Visualización en tiempo real del sentimiento detectado. Puntuación acumulada para cada mensaje (1 = positivo, 0 = negativo)**

Fuente: Elaboración propia.

## Conclusiones

Este estudio parte de los mensajes masivos publicados en la red social Twitter sobre temas científicos, para desarrollar un clasificador automático de sentimiento (positivo o negativo) que sea capaz de predecir la polaridad de nuevos mensajes en tiempo real.

El prototipo desarrollado, denominado OpScience, requiere de la construcción previa de un corpus de textos clasificados manualmente con su sentimiento para entrenar un modelo de aprendizaje automático supervisado. La preparación del corpus de textos específicos sobre ciencia implica descargar todos los datos de Twitter y realizar una limpieza, un procesamiento y un etiquetado manual. Según nuestro saber y entender, es la primera vez que se estructura la creación de un corpus específico destinado al análisis de la comunicación de la ciencia en Twitter, lo que aporta una herramienta de trabajo valiosa para futuros experimentos que quieran utilizar estos datos.

En la descarga de datos desde Twitter, se ha podido constatar que la actualidad está muy ligada a las conversaciones que se desarrollan en la red y que, a pesar de filtrar por determinadas palabras clave, los mensajes son muy diversos y cambiantes en función del contexto en el que se emitan. Además, los procesos de filtrado y limpieza del



corpus deben ser adaptados para cada proyecto o temática en particular, puesto que las peculiaridades del idioma, la jerga, la simbología o los complementos de cada mensaje de Twitter son diferentes en función del tópico que se trate. En este hecho radica la principal ventaja de la creación de este corpus, que podrá ser aprovechado en futuras investigaciones sobre comunicación científica.

Para realizar el entrenamiento del modelo de aprendizaje automático se han utilizado seis algoritmos de clasificación (Original Naive Bayes, Naive Bayes for multimodal models, Naive Bayes for multivariate Bernoulli models, Logistic Regression, Linear Support Vector Classification (svc) y Linear classifiers with stochastic gradient descent (sgd) training), que se combinaron para el modelo final, con el objetivo de crear un sistema de votación que aporte en cada caso la mejor solución. El sistema de votación del clasificador ha obtenido unos resultados en su evaluación de: Accuracy = 72,31 %, Precisión = 71,35 %, Recall = 77,2 % y F-score = 74,13 %, lo que se puede comparar con ejercicios recientes como el TASS 2018 que obtiene resultados variables desde 63,1 % hasta 89,3 % según las tareas realizadas (Carlos Díaz-Galiano et al., 2019).

El modelo fue utilizado para comprobar, varios meses después de la obtención del corpus, su validez y utilidad para descargar en tiempo real contenidos publicados en Twitter y evaluar su polaridad positiva o negativa. En 48 horas de enero de 2019 se extrajeron 8131 tweets, 3358 de los cuales fueron considerados positivos y 2818, negativos, con más de un 80 % de confianza.

Líneas futuras de trabajo pueden continuar con la ampliación del corpus a partir de un nivel adicional, de manera que se contemplen sentimientos positivos, negativos y neutros, para valorar si los resultados de precisión mejoran. También se pueden realizar pruebas de clasificación en tiempo real más extensas (en cuanto al tiempo) o para revisar eventos concretos o momentos de especial incidencia de mensajes de comunicación de la ciencia en Twitter. Esta sería la finalidad última de desarrollar y validar este corpus y modelo.

Como limitaciones hay que tener en cuenta que los modelos entrenados con un corpus de textos de temáticas concretas pueden requerir actualización al cabo de un período de tiempo largo. Esto es consecuencia del ritmo habitual de la red social Twitter, cuyos usuarios siguen alimentando la red de mensajes y, así mismo, incorporan nuevas costumbres de habla o nuevas modas culturales.

Este prototipo es un primer paso para profundizar en el análisis de la comunicación de la ciencia en las redes y, especialmente, para descubrir movimientos de opinión e información que puedan resultar útiles a diferentes instituciones o personas, como, por ejemplo, medios de comunicación, investigadores o estudiantes de cara a seleccionar los temas más populares o aceptados en su trabajo.



## Referencias

1. Alonso Berrocal, J. L., Gómez Díaz, R., Figuerola, C. G., Zazo Rodríguez, Á. F., & Cordon García, J. A. (2012). Propuesta de estudio del campo semántico de los libros electrónicos en Twitter. *Scire: Representación y Organización Del Conocimiento*, 18(2), 87-97. Recuperado de <http://eprints.rclis.org/29310/>
2. Arcila-Calderón, C., Barbosa-Caro, E., & Cabezuelo-Lorenzo, F. (2016). Técnicas Big Data: Análisis de textos a gran escala para la investigación científica y periodística. *El Profesional de La Información*, 25(4), 623-631. <https://doi.org/10.3145/epi.2016.jul.12>
3. Arcila-Calderón, C., Calderín-Cruz, M., & Sánchez-Holgado, P. (2019). Adopción de redes sociales por revistas científicas de ciencias sociales. *El Profesional de La Información*, 28(1), 1699-2407. <https://doi.org/10.3145/epi.2019.ene.05>
4. Baker, M. (2015). Social media: A network boost. *Nature*, 518(7538), 263-265. <https://doi.org/10.1038/nj7538-263a>
5. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly. Recuperado de [http://www.nltk.org/book\\_1ed/](http://www.nltk.org/book_1ed/)
6. Bollen, J., Mao, H., & Pepe, A. (2011). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. En: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence (www.aaai.org). Recuperado de <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPaper/2826>
7. Bonetta, L. (2009). Should You Be Tweeting? *Cell*, 139(3), 452-453. <https://doi.org/10.1016/J.CELL.2009.10.017>
8. Brossard, D. & Scheufele, D. A. (2013). Science, New Media, and the Public. *Science*, 339(6115), 40-41. <https://doi.org/10.1126/science.1232329>
9. Campos-Freire, F. & Rúas-Araújo, J. (2016). Uso de las redes sociales digitales profesionales y científicas: el caso de las 3 universidades gallegas. *El Profesional de La Información*, 25(3), 431-440. <https://doi.org/10.3145/epi.2016.may.13>
10. Carlos Díaz-Galiano, M., et al. (2019). TASS 2018: The Strength of Deep Learning in Language Understanding Tasks. *Procesamiento del Lenguaje Natural*, 62, 77-84. <https://doi.org/10.26342/2019-62-9>
11. Cha, M., Benevenuto, F., Haddadi, H., & Gummadi, K. (2012). The World of Connections and Information Flow in Twitter. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 42(4), 991-998. <https://doi.org/10.1109/TSMCA.2012.2183359>
12. Chen, S. C., Yen, D. C., & Hwang, M. I. (2012). Factors influencing the continuance intention to the usage of Web 2.0: An empirical study. *Computers in Human Behavior*, 28(3), 933-941. <https://doi.org/10.1016/J.CHB.2011.12.014>
13. Côté, I. M. & Darling, E. S. (2018). Scientists on Twitter: Preaching to the choir or singing from the rooftops? *FACETS*, 3(1), 682-694. <https://doi.org/10.1139/facets-2018-0002>
14. Cruz Mata, F., Troyano Jiménez, J. A., de Salamanca Ros, F., & Ortega Rodríguez, F. J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento Del Lenguaje Natural*, 41, 73-80. Recuperado de <https://core.ac.uk/download/pdf/16361408.pdf>



15. Darling, E., Shiffman, D., Côté, I., & Drew, J. (2013). The role of Twitter in the life cycle of a scientific publication. *Ideas in Ecology and Evolution*, 6, 32-43. <https://doi.org/10.4033/iee.2013.6.6.f>
16. Díaz-Galiano, M. C., Martínez-Cámara, E., Ángel García-Cumbreras, M., García-Vega, M., & Villena-Román, J. (2018). The democratization of deep learning in TASS 2017. *Procesamiento de Lenguaje Natural*, 60, 37-44. <https://doi.org/10.26342/2018-60-4>
17. Dubiau, L. & Ale, J. M. (2013). Análisis de Sentimientos sobre un Corpus en Español: Experimentación con un Caso de Estudio. En: *14th Argentine Symposium on Artificial Intelligence, ASAI 2013* (pp. 36-47). Recuperado de <http://42jaiio.sadio.org.ar/proceedings/simposios/Trabajos/ASAI/04.pdf>
18. Fowks, J. (2017). *Mecanismos de la posverdad*. Lima: Fondo de Cultura Económica.
19. García Cumbreras, M. Á., Villena Román, J., Martínez-Cámara, E., Díaz Galiano, M. C., Martín-Valdivia, M. T., & Ureña-López, L. A. (2016). Resumen de TASS 2016. En: *TASS 2016: Workshop on Sentiment Analysis at SEPLN Proceedings* (pp. 13-21). Recuperado de [http://ceur-ws.org/Vol-1702/tass2016\\_proceedings\\_v24.pdf](http://ceur-ws.org/Vol-1702/tass2016_proceedings_v24.pdf)
20. García Esparza, S., O'Mahony, M. P., & Smyth, B. (2012). Mining the real-time web: A novel approach to product recommendation. *Knowledge-Based Systems*, 29, 3-11. <https://doi.org/10.1016/J.KNOSYS.2011.07.007>
21. García Esparza, S., O'mahony, M. P., & Smyth, B. (2012). Mining the real-time web: A novel approach to product recommendation. *Knowledge-Based Systems*, 29, 3-11. <https://doi.org/10.1016/j.knosys.2011.07.007>
22. Go, A., Bhayani, R., & Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision*. Recuperado de <http://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf>
23. Henriquez Miranda, C., & Guzman, J. (2017). A review of sentiment analysis in spanish. *Tecciencia*, 12(22), 35-48. <https://doi.org/10.18180/tecciencia.2017.22.5>
24. Hurtado, L.-F., Pla, F., & Buscaldi, D. (2015). ELIRF-UPV en TASS 2015: Análisis de Sentimientos en Twitter. En: *Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015)* (pp. 75-79). Alicante. Recuperado de [http://ceur-ws.org/Vol-1397/elirf\\_upv.pdf](http://ceur-ws.org/Vol-1397/elirf_upv.pdf)
25. Jarreau, P. B. (2015). *All the Science That Is Fit to Blog: An Analysis of Science Blogging Practices*. Louisiana State University. Recuperado de [https://digitalcommons.lsu.edu/gradschool\\_dissertations/1051](https://digitalcommons.lsu.edu/gradschool_dissertations/1051)
26. Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. En: *9th WEBKDD and 1st SNA-KDD Workshop*. San Jose, California: ACM. <https://doi.org/10.1145/1348549.1348556>
27. Kahle, K., Sharon, A. J., & Baram-Tsabari, A. (2016). Footprints of Fascination: Digital Traces of Public Engagement with Particle Physics on CERN's Social Media Platforms. *PLOS ONE*, 11(5), e0156409. <https://doi.org/10.1371/journal.pone.0156409>
28. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11)*, 538-541. Recuperado de <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251?iframe=true&width=90%25&height=90%25>
29. Krippendorff, K. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3), 411-433. <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>



30. Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. *ScholarlyCommons*. University of Pennsylvania. Recuperado de [http://repository.upenn.edu/asc\\_papers/43](http://repository.upenn.edu/asc_papers/43) (no sirve el link)
31. Krippendorff, K. & Hayes, A. F. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1), 77-89. <https://doi.org/10.1080/19312450709336664>
32. Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? En: *Proceedings of the 19th international conference on World wide web - WWW '10* (p. 591). Nueva York: ACM Press. <https://doi.org/10.1145/1772690.1772751>
33. Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011). Twitter Trending Topic Classification. En: *IEEE 2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 251-258). <https://doi.org/10.1109/ICDMW.2011.171>
34. Li, G. & Liu, F. (2012). Application of a clustering method on sentiment analysis. *Journal of Information Science*, 38(2), 127-139. <https://doi.org/10.1177/0165551511432670>
35. Liang, X., et al. (2014). Building Buzz: (Scientists) Communicating Science In New Media Environments. *Journalism and Mass Communication Quarterly*, 91(4), 772-791. <https://doi.org/10.1177/1077699014550092>
36. Mandavilli, A. (2011). Trial by Twitter. *Nature*, 469, 286-287. Recuperado de <https://www.nature.com/news/2011/110119/pdf/469286a.pdf>
37. Martínez-Cámara, E., Díaz-Galiano, M. C., García-Cumbreras, A., García-Vega, M., & Villena-Román, J. (2017). Resumen de TASS 2017. En: *TASS 2017: Workshop on Semantic Analysis at SEPLN Proceeding* (pp. 13-21). Recuperado de <http://www.sepln.org/workshops/tass/>.
38. Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A., & Montejó-Ráez, A. (2014). Sentiment analysis in Twitter. *Natural Language Engineering*, 20(1), 1-28. <https://doi.org/10.1017/S1351324912000332>
39. Montenegro, V. & Escudero, H. (2013). Las redes sociales y la difusión de la tecnología y la innovación. En: *III Congreso Internacional de Comunicación Pública de la Ciencia*. Santa Fe, Argentina. Recuperado de <http://studylib.es/doc/7718559/untitled---copuci-2017>
40. Narr, S., De Luca, E. W., & Albayrak, S. (2011). Extracting semantic annotations from twitter. En: *Proceedings of the fourth workshop on Exploiting semantic annotations in information retrieval - ESAIR '11* (p. 15). Nueva York: ACM Press. <https://doi.org/10.1145/2064713.2064723>
41. O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. En: *Association for the Advancement of Artificial Intelligence. Proceedings of the Fourth International AAI Conference on Weblogs and Social Media* (pp. 122-129). Recuperado de <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842>
42. Pak, A. & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. En: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 1320-1326). Recuperado de [http://www.lrec-conf.org/proceedings/lrec2010/pdf/385\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf)
43. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12, 2825-2830. Recuperado de <http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>



44. Pérez-Rodríguez, A. V., González-Pedraz, C., & Alonso Berrocal, J. L. (2018). Twitter como herramienta de comunicación científica en España. Principales agentes y redes de comunicación. *Communication papers: media literacy and gender studies*, 7(13), 95-111. Recuperado de <https://dialnet.unirioja.es/servlet/articulo?codigo=6442315>
45. Peters, H. P., Dunwoody, S., Allgaier, J., Lo, Y.-Y., & Brossard, D. (2014). Public communication of science 2.0: Is the communication of science via the “new media” online a genuine transformation or old wine in new bottles? *EMBO Reports*, 15(7), 749-753. <https://doi.org/10.15252/embr.201438979>
46. Pont-Sorribes, C., Cortiñas-Rovira, S., & Di Bonito, I. (2013). Challenges and opportunities for science journalists in adopting new technologies: the case of Spain. *SISSA-International School for Advanced Studies, Journal of Science Communication*, 12(3). <https://doi.org/10.22323/2.12030205>
47. Quiñónez Gómez, H. & Sánchez Colmenares, M. (2016). Uso de Twitter en el periodismo científico. Los casos de los diarios El Nacional y El Universal en Venezuela. *Question*, 1(52), 212-231. Recuperado de <http://perio.unlp.edu.ar/ojs/index.php/question/article/view/3490>
48. Ribas, C. (2012). La divulgación y la comunicación de la ciencia, en la encrucijada. *Sociedad Española de Bioquímica y Biología Molecular (SEBBM)*, 173, 10-12. Recuperado de [https://www.academia.edu/4630023/La\\_comunicación\\_de\\_la\\_ciencia\\_en\\_la\\_encrucijada](https://www.academia.edu/4630023/La_comunicación_de_la_ciencia_en_la_encrucijada)
49. Rosá, A., Chiruzzo, L., Etcheverry, M., & Castro, S. (2017). RETUYT en TASS 2017: Análisis de sentimiento de Tweets en Español utilizando SVM y CNN. En: *TASS 2017: Workshop on Semantic Analysis at SEPLN* (pp. 77-83). Recuperado de <http://arxiv.org/abs/1710.06393>
50. Saif, H., He, Y., & Alani, H. (2012). Semantic Sentiment Analysis of Twitter. En: Cudré-Mauroux P. et al. (eds). The Semantic Web – ISWC 2012. ISWC 2012. Lecture Notes in Computer Science. *International Semantic Web Conference (ISWC 2012)*, 7649. 508-524. [https://doi.org/10.1007/978-3-642-35176-1\\_32](https://doi.org/10.1007/978-3-642-35176-1_32)
51. Segarra-Saavedra, J., Tur-Viñes, V., & Hidalgo-Marí, T. (2017). Uso de Twitter como herramienta de difusión en las revistas científicas españolas de Comunicación. En: *7ª Conferencia internacional sobre revistas de ciencias sociales y humanidades, Revista Mediterránea de Comunicación*. Recuperado de [http://thinkepi.net/notas/crecs\\_2017/J\\_16\\_30\\_Segarra.pdf](http://thinkepi.net/notas/crecs_2017/J_16_30_Segarra.pdf)
52. Sidorov, G. et al. (2013). Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. En: Batyrshin I., González Mendoza M. (eds.) *Advances in Artificial Intelligence. MICAI 2012. Lecture Notes in Computer Science*, 7629. Springer, Berlin, Heidelberg. Doi: [10.1007/978-3-642-37807-2\\_1](https://doi.org/10.1007/978-3-642-37807-2_1)
53. Van Zoonen, W. & Van der Meer, Toni, G. L. A. (2016). Social media research: The application of supervised machine learning in organizational communication research. *Computers in Human Behavior*, 63, 132-141. <https://doi.org/10.1016/J.CHB.2016.05.028>
54. Whitman Cobb, W. N. (2015). Trending now: Using big data to examine public opinion of space policy. *Space Policy*, 32, 11-16. <https://doi.org/10.1016/J.SPACPOL.2015.02.008>
55. Yerva, S. R., Miklós, Z., & Aberer, K. (2012). Quality-aware similarity assessment for entity matching in Web data. *Information Systems*, 37(4), 336-351. <https://doi.org/10.1016/J.IS.2011.09.007>