ANÁLISIS SUPERVISADO DE SENTIMIENTOS POLÍTICOS EN ESPAÑOL: CLASIFICACIÓN EN TIEMPO REAL DE TWEETS BASADA EN APRENDIZAJE AUTOMÁTICO

Supervised sentiment analysis of political messages in Spanish: Real-time classification of tweets based on machine learning

Carlos Arcila-Calderón, Félix Ortega-Mohedano, Javier Jiménez-Amores y Sofía Trullenque

Carlos Arcila-Calderón es profesor de la *Universidad de Salamanca* (España), miembro del *Observatorio de Contenidos Audiovisuales* (*OCA*) y editor de la revista *Disertaciones*. Es doctor europeo en Comunicación, Cambio Social y Desarrollo por la *Universidad Complutense de Madrid* y *Master en Data science* y *Master en Periodismo* por la *Universidad Rey Juan Carlos*. Con anterioridad desarrolló su carrera en la *Universidad del Rosario* y en la *Universidad del Norte* de Colombia, y en la *Universidad de Los Andes* y la *Universidad Católica Andrés Bello* de Venezuela.
http://orcid.org/0000-0002-2636-2849

carcila@usal.es

Félix Ortega-Mohedano es profesor de la *Universidad de Salamanca*, miembro del *Observatorio de Contenidos Audiovisuales* (*OCA*), director del *Master* en *Comunicación Audiovisual: Investigación e Innovación* y secretario académico del *Instituto Universitario de Investigación en Ciencias de la Educación* de la *Universidad de Salamanca* (*Usal*). Doctor en Comunicación, Cultura y Educación y licenciado en Economía por la *Usal*. Ha desarrollado su trabajo en investigación de audiencias, metodologías de investigación, estructura del sistema audiovisual y las industrias culturales, televisión, educación y comunicación, publicando artículos y libros a nivel nacional e internacional.
http://orcid.org/0000-0003-2735-4813

fortega@usal.es

Javier Jiménez-Amores es alumno del *Master Universitario en Investigación en Comunicación Audiovisual* (*Muica*) de la *Universidad de Salamanca* (España), graduado en Comunicación Audiovisual por la misma universidad y técnico superior en Imagen y Sonido. En el plano profesional es fotógrafo y diseñador gráfico freelance además de comunicólogo. Anteriormente ha trabajado en el área de comunicación del *Master universitario en Diseño y Comunicación* (*Mudic*) de la *Escuela Superior de Diseño e Ingeniería de Barcelona* (*Elisava*); y como fotógrafo para la compañía hotelera *BeLive Hotels*, en Bayahibe (República Dominicana).
http://orcid.org/0000-0001-7856-5392

javieramores@usal.es

Sofía Trullenque es alumna del *Master Universitario en Investigación en Comunicación Audiovisual* (*Muica*) de la *Universidad de Salamanca*, graduada en Comunicación Audiovisual por la misma universidad y Técnico de Educación Infantil. En el plano profesional es comunicóloga e investiga sobre el consumo de nuevos medios, especialmente tablets y smartphones en la infancia. Anteriormente se ha formado en el área de comunicación en *Replay Serveis Audiovisuals* elaborando cobertura de actos y programas de televisión.
http://orcid.org/0000-0002-5143-2431

strullenque@usal.es

Universidad de Salamanca, Facultad de Ciencias Sociales
Campus Miguel de Unamuno, Edificio FES
Av. Francisco Tomás y Valiente, s/n. 37007 Salamanca, España

Artículo recibido el 10-3-2017
Aceptación definitiva: 8-5-2017

Resumen

Se describe y evalúa la aplicación de la técnica *análisis supervisado de sentimientos* en comunicación política a través de un clasificador en tiempo real de opiniones políticas en tweets en español utilizando técnicas de aprendizaje automático (*machine learning*), tanto en un ordenador local como usando computación distribuida comercial para problemas de datos masivos (*big data*). Describimos las técnicas y métodos emergentes asociados y analizamos las oportunidades que para la comunicación política representan estas innovaciones.

Palabras clave

Análisis supervisado de sentimientos; Opinión política; *Twitter*; Aprendizaje automático; *Big data*; Datos masivos; Tweets políticos.

Abstract

This article describes and evaluates the application of the *supervised sentiment analysis* in political communication through a real-time classifier of political opinions in Spanish tweets using machine learning techniques, both on a local computer and using distributed computing for big data problems. We describe the associated emerging methods and techniques and analyze the opportunities that these innovations represent for political communication.

Keywords

Supervised sentiment analysis; Political opinion; *Twitter*; Machine learning; Big data; Political tweets.

Arcila-Calderón, Carlos; **Ortega-Mohedano, Félix**; **Jiménez-Amores, Javier**; **Trullenque, Sofía** (2017). "Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático". *El profesional de la información*, v. 26, n. 5, pp. 973-982.

https://doi.org/10.3145/epi.2017.sep.18

1. Introducción

En la última década ha surgido un interés creciente por el estudio de las opiniones políticas utilizando datos a gran escala producidos por medios sociales en nuestro entorno socioeconómico (**Cobb**, 2015; **O'Connor** *et al.*, 2010; **Bollen**; **Mao**; **Pepe**, 2011). Sin embargo, la mayoría de estos trabajos se han fundamentado en la clasificación manual y/o en el análisis automatizado de contenido utilizando diccionarios que principalmente etiquetan palabras (por ejemplo, dando un valor a priori negativo o positivo a cada palabra) (**Leetaru**, 2012; **Feldman**, 2013). Otras aproximaciones realizadas desde el aprendizaje automático supervisado o *supervised machine learning* (**Vinodhini**; **Chandrasekaran**, 2012) o directamente derivados de la inteligencia artificial, son escasas en la investigación en las ciencias de la comunicación (**Van-Zoonen**; **Van-der-Meer**, 2016).

Existen contadas aplicaciones comerciales en comunicación como las utilizadas por los equipos de campaña político-comunicativa presidenciales del presidente Obama mediante la solución de *big data* de *Oracle* (**Mariño-Angoso**, 2015; **Sorrells**, 2012).

En áreas de conocimiento de las ciencias sociales, en sentido extenso, y en particular en las instituciones públicas y centros de investigación sociológica y política, así como en empresas y consultoras privadas de comunicación política, opinión pública, estudios políticos y marketing electoral, estas técnicas han empezado a despertar un interés notable. Además, son nuevos los esfuerzos tecnológicos dedicados a enlazar y relacionar el análisis automatizado de sentimientos basado en *machine learning* con tecnologías *streaming* o de transmisión en vivo, capaces de ofrecer a su vez una cantidad importante de datos que pueden ser analizados y evaluados de forma automatizada, ofreciendo informes y conclusiones que asesoren la construcción de discursos y la planificación mediática de partidos políticos, medios de comunicación y comunicaciones corporativas.

Se analiza el potencial de estos métodos y técnicas, explicando la evolución y aplicación de un clasificador en tiempo real de opiniones políticas en español con técnicas de aprendizaje automático, implementadas tanto en un ordenador en servicio local como usando computación distribuida comercial de mayor capacidad e inmediatez de análisis.

Se expone de forma específica cómo se puede implementar y evaluar la técnica de análisis supervisado de sentimientos en el campo de la comunicación política. Esta metodología y técnica representa un instrumento único para el contraste predictivo de los resultados electorales futuros en cualquier país o región, y en particular en nuestro entorno de habla hispana (permitiendo corregir, por ejemplo, el sesgo de las encuestas o anticipar las tendencias más significativas con indicadores adelantados).

La metodología que se presenta en este artículo ha sido implementada como prototipo bajo el nombre de *Autocop*, disponible de forma gratuita en software libre y desarrollado desde el *Observatorio de Contenidos Audiovisuales* (*OCA*) de la *Universidad de Salamanca*, como prueba de concepto para transferencia del conocimiento (Plan *TCUE* 2015-2017 Fase 2).
http://ocausal.imbv.net/proyecto-autocop-es

El prototipo permite la realización de análisis longitudinales para detectar cambios en los indicadores tendenciales asociados a los partidos políticos y sus candidatos, así como comparar estos cambios con los acontecimientos cotidianos. Esto se da como resultado de los avances científico-técnicos en el campo de las ciencias sociales en convergencia con las ciencias de la computación, la inteligencia artificial y los *datos masivos* que recientemente se están implementando en España e Hispanoamérica (**Arcila-Calderón**; **Barbosa**; **Cabezuelo**, 2016).

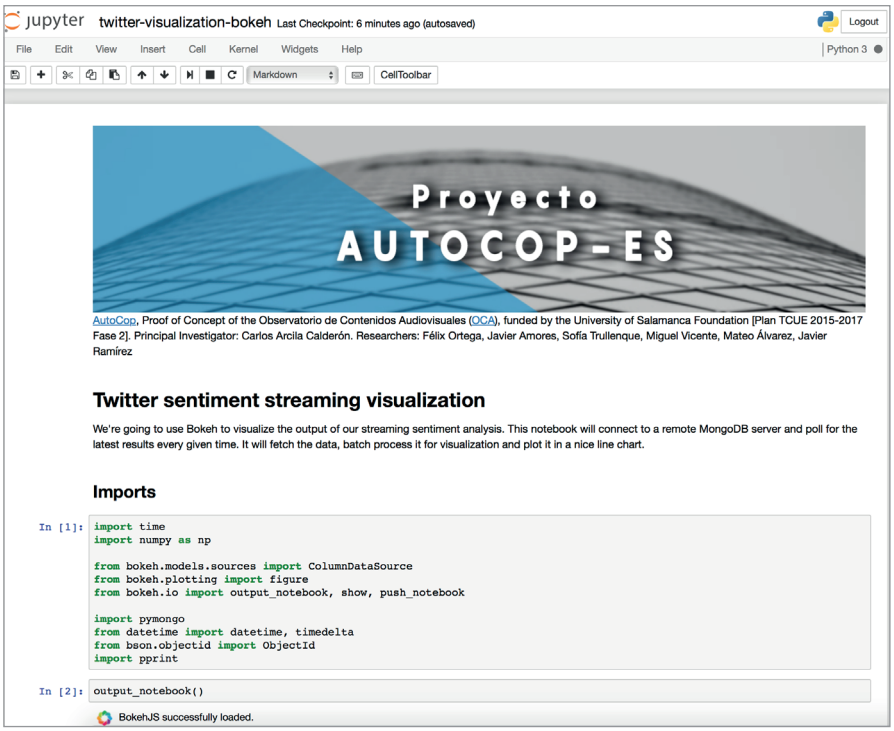

Figura 1. Proyecto *Autocop*
http://ocausal.imbv.net/proyecto-autocop-es

El análisis supervisado de sentimientos se ha llevado a cabo recientemente en otros idiomas diferentes al inglés, como el esloveno (**Bučar**; **Pov**; **Žnidaršič**, 2016), sin embargo salvo por los trabajos de **García-Cumbreras** *et al.* (2016) y **Hurtado**, **Pla** y **Buscaldi** (2015), existe una escasa investigación y por lo tanto aplicación académica y empresarial centrada en la elaboración de modelos supervisados de aprendizaje automático para clasificar textos políticos en lengua española. En este sentido, la utilización de estos modelos predictivos (usando algoritmos de clasificación como el *Naive Bayes*) y la aplicación del clasificador que se explica en este texto permite a cualquier investigador, empresa o consultora independiente en el campo de la comunicación política que implemente esta técnica, conectarse al flujo de datos de *Twitter* en tiempo real (utilizando el *API Streaming*) para predecir y visualizar el sentimiento de cada tweet y de conjuntos agregados de mensajes.

En este artículo se explican los enfoques tradicionales de análisis de sentimientos que vienen aplicándose en comunicación política (análisis manual y automático con diccionarios), para luego explicar en qué consiste el análisis supervisado de sentimientos políticos tanto a pequeña escala (en local) como para problemas de *datos masivos* y mayor procesamiento computacional. Finalmente, se discuten cuáles son los posibles usos y aplicaciones de la técnica explicada, así como sus implicaciones teóricas.

2. Análisis de sentimientos en la comunicación política

El análisis de sentimientos (*sentiment analysis*, *SA*) es una de las principales técnicas de estudio de datos textuales a gran escala (*big data*) empleadas en la investigación en ciencias sociales y en comunicación política. Su objetivo es reconocer y evaluar el valor emocional existente detrás de los textos analizados, a través de su estructura, clasificándolos en *positivos*, *negativos* o *neutros*. En la actualidad, esta metodología se aplica principalmente en la interpretación de los textos difundidos en medios sociales como *Twitter*.

Se analiza el vocabulario del texto mediante el uso de un ordenador que a través de la implementación de un diccionario de *lexicons* procese, reconozca y evalúe las cargas emocionales contenidas en el mensaje. En este sentido, es comprensible que el análisis de sentimientos sea empleado en la monitorización de las redes sociales en tiempo real, sirviéndose de las nuevas herramientas computacionales que simplifican y agilizan los procesos de automatización con el propósito de configurar escenarios teóricos y coetáneos de la opinión pública respecto a determinados temas. Con frecuencia el análisis de sentimientos es confundido con la minería de opinión (*opinion mining*), pero ésta se dedica a la detección de la polaridad, y la identificación de emociones es habitualmente explotada con el mismo objetivo (**Cambria** *et al.*, 2013). Ambas técnicas son diferentes pero complementarias.

La tarea de identificar el sentimiento predominante en un texto escrito es una labor compleja incluso para un ser humano formado. Por este motivo, el análisis de sentimientos automatizado requiere un desarrollo y perfeccionamiento deliberado continuo que ha sido planteado desde dos enfoques:

- aproximaciones semánticas (**Turney**, 2002);
- técnicas de aprendizaje computacional (**Pang**; **Lee**; **Vaithyanathan**, 2002).

Los enfoques semánticos se caracterizan por el uso de diccionarios de sentimientos (*lexicons*) con orientación de polaridad u opinión. Estos sistemas pre-procesan el texto y lo dividen en palabras, comprobando posteriormente la aparición de los términos del *lexicon* para asignar el sentimiento de polaridad del texto mediante la suma de los valores de polaridad ponderada de los términos ("guerra"= -2; "amor"= +3). Estos sistemas incluyen un tratamiento más o

menos avanzado de términos modificadores (muy, poco o demasiado) que incrementan o reducen la polaridad de los términos a los que acompañan, así como la inclusión de términos inversores o negadores (no o tampoco), que invierten la polaridad de los términos a los que afectan. Este método fue el empleado por **Turney** (2002), uno de los pioneros en el uso aplicado del análisis de sentimiento automatizado, en este caso orientado al análisis de reseñas sobre servicios y productos.

> El análisis de sentimientos se emplea en la monitorización de las redes sociales en tiempo real, sirviéndose de medios informáticos que simplifican y agilizan los procesos

Como se explicará en detalle en el siguiente apartado, las aproximaciones de análisis basadas en aprendizaje automático (*machine learning*) consisten en entrenar un clasificador usando un algoritmo de aprendizaje supervisado a partir de una colección de textos anotados en combinación con otro tipo de características semánticas que intentan modelar la estructura sintáctica de las frases, la intensificación, la negación, la subjetividad o la ironía, entre otras variables. Una de las primeras aproximaciones a este enfoque fue el presentado por el trabajo de **Pang**, **Lee** y **Vaithyanathan** (2002), donde se utilizaba el aprendizaje supervisado para el análisis de sentimientos polarizado aplicado a reseñas de películas, clasificándolas en positivas o negativas.

Si bien ambas aproximaciones relativas al análisis de sentimientos perduran en la actualidad, partir del trabajo de estos autores gran parte de la investigación que implementa estas metodologías y técnicas se orienta al análisis e interpretación de microblogs de uso masivo como *Twitter*. En este sentido, **Bermingham** y **Smeaton** (2010) realizaron un estudio sobre cómo los textos breves son más compactos y explícitos en cuanto a la proyección del sentimiento, concluyendo, a partir del análisis de más de 60 millones de tweets, que, en términos de análisis *big data*, es más fácil clasificar los sentimientos de los textos breves difundidos en los microblogs que en otras estructuras léxicas. Otros investigadores como **Bakliwal** *et al.* (2012) han centrado su investigación en elaborar una función para identificar y clasificar los sentimientos presentes en los mensajes de *Twitter*, partiendo de un corpus previo de tweets precodificados.

A medida que evolucionan los métodos de análisis científico asociados a la minería de datos y se perfeccionan las técnicas de análisis de sentimiento automatizado enfocadas a las redes sociales y microblogs, el interés por su aplicación en las ciencias de la comunicación y en particular en las ciencias políticas viene incrementándose en la última década de manera exponencial. Su potencial reside en su capacidad para calibrar y construir indicadores avanzados de opinión pública, en tiempo real. Es por este motivo que muchas de las investigaciones más actuales que implementan estos métodos se centran en la observación de los sentimientos presentes en plataformas como *Twitter* acerca de determinados temas-objetos de actualidad o *hashtags*, que servirían para realizar predicciones de carácter político, casi siempre relacionadas con procesos electorales y/o eventos político-comunicativos. En este sentido, la propia administración de Obama se valió del análisis de sentimientos automatizado para sondear la opinión pública sobre sus políticas y mensajes de campaña antes de las elecciones presidenciales de 2012. En esta línea, **Tumasjan** *et al.* (2010), en el marco de las elecciones federales alemanas de 2009, trataron de comprobar si *Twitter* era usado por sus usuarios como foro para el debate y discusión política, y si el flujo de información que ofrecían sus mensajes podía ser usado como representación válida de los sentimientos políticos de la sociedad germana. En el estudio se analizaron 104.003 tweets con contenido relacionado con los políticos y/o los partidos políticos de las seis partes representadas en el Parlamento alemán, sirviéndose de *LIWC*, un software de análisis de texto. Sus resultados concluyeron que *Twitter* sí es utilizado para la deliberación política y refleja plausiblemente la opinión pública, constituyéndose como un indicador válido de los sentimientos políticos (**Tumasjan** *et al.*, 2010).

Otros autores como **Choy** *et al.* (2011) han utilizado el análisis de sentimientos basado en diccionarios para predecir el porcentaje de votos de los candidatos en las elecciones de Singapur en el año 2011. En su investigación recopilaron 16.616 tweets de la API de *Twitter* durante la campaña de agosto de 2011 y crearon un corpus personalizado corrigiendo el sesgo que provoca el uso de corpus estandarizados en estos análisis. Sus resultados predijeron quiénes serían los dos candidatos más votados, aunque erraron al predecir el vencedor final de las elecciones, por un pequeño margen de votos.

> Las aproximaciones de análisis basadas en aprendizaje automático (*machine learning*) consisten en entrenar un clasificador usando un algoritmo de aprendizaje supervisado a partir de una colección de textos anotados

Bermingham y **Smeaton** (2011) también trataron de utilizar *Twitter* para predecir los resultados electorales de Irlanda de 2011 mediante el uso de un clasificador de sentimientos supervisado. Con su modelo consiguieron una precisión en la clasificación efectiva de éstos del 65%.

Otro modelo y metodología más reciente de identificación de sentimientos políticos a través de *Twitter* es el análisis en tiempo real. Mientras que el análisis de sentimientos tradicional tarda días o semanas en completarse, estos nuevos sistemas analizan los sentimientos políticos presentes en el tráfico de *Twitter*, entregando resultados de forma prácticamente continua e inmediata. Uno de los trabajos pioneros en el uso de este tipo de análisis sería el de **Wang** *et al.* (2012), quienes desarrollaron un modelo de análisis de sentimientos en tiempo real en las elecciones presidenciales estadounidenses de 2012. Su modelo interpreta resultados al instante de cómo determinados eventos pueden afectar

a la opinión pública. Para el estudio se recopilaron más de 36 millones de tweets en *streaming* de los que se extrajo el sentimiento político usando la solución *Amazon Mechanical Turk*. Su sistema consiguió un 59% de precisión en la clasificación efectiva de tweets.

En relación con el análisis de sentimientos basado en técnicas de aprendizaje automático, es relevante señalar que en la actualidad existen únicamente ejemplos primigenios de aplicación a la comunicación política en redes sociales en nuestro contexto cultural y científico. Uno de los escasos modelos probados es el de **Bakliwal** *et al.* (2013), quienes proponen un sistema de análisis de sentimientos con orientación política fundamentado en el aprendizaje automático supervisado. A partir de un experimento que realizaron sobre 2.624 tweets recopilados en los días previos a las elecciones irlandesas de 2011, los autores pusieron a prueba un sistema clasificador de 3 tipos de sentimientos: negativo, positivo y neutro, siendo éste además capaz de identificar y etiquetar correctamente los sarcasmos. El experimento reportó un 61,6% de precisión, un resultado ligeramente superior a los estudios que presentan enfoques no supervisados o basados tan sólo en diccionarios de sentimientos para extraer conclusiones de sentimientos globales.

> Este procedimiento hace uso de algoritmos de clasificación, y su implementación en comunicación política permite el análisis de textos políticos de forma rápida evitando el sesgo producido por codificadores o por diccionarios con categorías *a priori*

En la última década el interés por el análisis de sentimientos computerizado ha seguido una senda constante e *in crescendo*, sobre todo en el campo de la comunicación política, siendo cada vez más numerosas las investigaciones que se valen de esta metodología-técnica. Existe sin embargo todavía una reveladora carencia de investigaciones y aplicaciones metodológicas de análisis en tiempo real y, más aún, de análisis de sentimientos en *streaming* basados además en aprendizaje automático supervisado. Estos sistemas pueden aportar grandes ventajas comparativas a la investigación social, comunicacional y política, permitiendo a los equipos de análisis político-comunicativo anticiparse en la detección e interpretación temprana de la eficacia de sus estrategias comunicativas y políticas prácticamente en tiempo real.

3. Análisis supervisado de tweets políticos

A diferencia del análisis de sentimientos manual con codificadores humanos o del análisis automático asistido por ordenador con el uso de diccionarios, el análisis supervisado de sentimientos utiliza procedimientos del aprendizaje automático supervisado (*supervised machine learning*) para generar modelos basados en datos previamente etiquetados y así poder predecir con un significativo grado de fiabilidad el sentimiento de los mensajes. Este procedimiento hace uso de algoritmos de clasificación y su implementación en comunicación política permite el análisis de textos políticos de forma rápida evitando el sesgo producido por codificadores o por diccionarios con categorías *a priori* incapaces de detectar los temas-temáticas, el contexto o las ironías-sarcasmos.

En el análisis supervisado de sentimientos se puede transformar el modelo de clasificación durante el proceso predictivo si dicho modelo es además alimentado y enriquecido con más textos etiquetados que mejoren su ajuste, lo que resulta de máxima utilidad en el estudio de la comunicación política por la volatilidad y rápida transformación inherente de los actores (gobernantes, partidos políticos, políticos, ciudadanos, etc.) y temas (elecciones, escándalos, etc.). Adicionalmente, esta técnica puede ser implementada en tiempo real gracias a las capacidades y posibilidades de conexión con datos a analizar de medios sociales asociados al debate socio-político como *Twitter*. Aunque inicialmente estas investigaciones emergentes han sido poco precisas e influyentes en sus conclusiones predictivas mediante el análisis de *Twitter* (**Madlberger**; **Almansour**, 2014), esta técnica ha sido utilizada recientemente para la predicción de eventos en diferentes campos (**Kranjc** *et al.*, 2015; **Preethi**; **Uma**; **Kumar**, 2015) como las finanzas (**Smailović** *et al.*, 2013), el estudio de redes sociales (**Sluban** *et al.*, 2015) y los resultados electorales (**Smailović** *et al.*, 2015), con resultados predictivos más que prometedores.

El análisis supervisado de sentimientos políticos puede además ser ejecutado utilizando recursos libremente disponibles como *Python* (versiones 2.7 y 3.4) y la interfaz de programa de aplicación (API) de *Twitter* (*REST* y *Streaming*):

- el API *REST* permite descargar y filtrar el histórico de mensajes de los últimos 7 días, con lo cual se pueden recolectar mensajes políticos para poderlos clasificar manualmente y que alimenten el modelo;
- con el API *Streaming* se puede realizar la conexión al flujo constante de *Twitter* en tiempo real (limitado al 1% de todos los mensajes producidos en ese momento). Todos los mensajes de *Twitter* se obtienen se forma semi-estructurada en formato JSON, lo que permite ejecutar filtros sobre las consultas, por ejemplo, de fechas, idiomas, lugares geográficos o etiquetas incluidas en el texto a analizar del mensaje.

Seguidamente, se pueden utilizar las bibliotecas *Natural Language Tool Kit* (*NLTK*) y *SciKit-Learn* en *Python* para desarrollar un clasificador que entrene un modelo de aprendizaje supervisado con diferentes algoritmos (p. e. *Original Naive Bayes*, *Naive Bayes* para modelos multimodales, *Naive Bayes* para modelos multivariados de Bernoulli, regresión logística, *Linear Support Vector Classification* y clasificadores lineales con entrenamiento de gradiente estocástico dependiente –SGD-). El clasificador se puede entrenar fácilmente usando ejemplos de tweets políticos positivos y negativos en español u otros idiomas (también se podrían incluir "neutros" o "imparciales"), lo cual genera a su vez un modelo que es el que finalmente permite predecir el sentimiento de los nuevos tweets. Si dividimos nuestros datos etiquetados (los tweets codificados manualmente) en dos corpus, uno de entrenamiento (*training*) con el 70% de los mensajes y otro de testeo (*testing*) con el 30% restante,

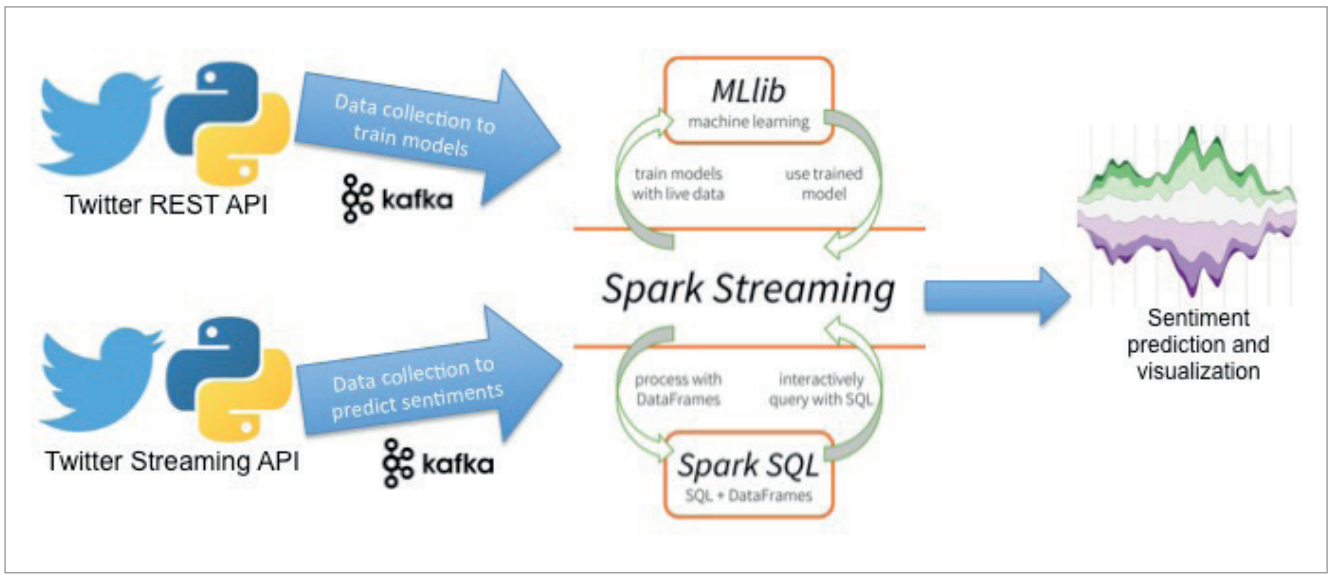

Figura 2. Análisis supervisado de sentimientos políticos en *Twitter* usando computación distribuida en *Spark*

podemos evaluar la capacidad predictiva final del modelo implementado. En otras palabras, se evalúa la fiabilidad con que el modelo acierta en su predicción, la cual suele estar en torno al 70% con algoritmos como el *Naive Bayes*.

Asimismo, podemos incluir en el modelo predictivo un intervalo de confianza en la predicción que indique en escala 0 a 1 con qué seguridad un tweet predicho como "positivo" es efectivamente "positivo", controlando así el error tipo I o la cantidad de falsos positivos con respecto al modelo que hemos desarrollado.

Por ejemplo, el clasificador se puede utilizar conectándolo al flujo de datos de *Twitter* en tiempo real (utilizando el *Streaming* del API disponible) y filtrando tweets escritos en español sobre partidos políticos españoles (#PP, #Podemos, #PSOE, #Ciudadanos) para predecir el sentimiento de cada tweet generado en tiempo real y visualizar automáticamente los resultados (usando por ejemplo la biblioteca *Matplot* para *Python*), incluyendo las predicciones de sentimientos políticos con altos intervalos de confianza (>0,80). Con este clasificador automático, los investigadores y/o las empresas consultoras dedicadas al estudio de la opinión pública, disponen de una metodología-tecnología más que relevante para probar la predicción de resultados electorales futuros en países de habla hispana mediante el análisis de estos flujos de información y *big data.* Estos análisis permiten a investigadores, académicos y consultoras políticas realizar análisis longitudinales durante largos períodos de tiempo con el objeto de detectar y predecir en tiempo real el sentimiento político en *Twitter* en relación con sus clientes políticos o corporativos, y así comparar con los acontecimientos cotidianos. Estos análisis en tiempo real permiten además detectar las crisis comunicativas de sus clientes en la fase primigenia de la "ola comunicativa de crisis" permitiendo reaccionar a estos con mayor proactividad para atemperar la misma y corregir convenientemente antes de que se transformen en "crisis comunicativas catastróficas".

4. Análisis distribuido para grandes cantidades de datos políticos

El análisis supervisado de sentimientos se puede realizar directamente en un ordenador personal, como se ha explicado en el apartado anterior. El cómputo local tiene sin embargo serias limitaciones si los volúmenes de datos y variables a analizar son del orden y dimensión *big data*, lo que requiere de almacenamiento escalable y computación distribuida. En este sentido, la ejecución de análisis de datos en *streaming* en plataformas distribuidas ha sido un reto en el complejo y cambiante panorama de los *datos masivos* (**Turck**; **Hao**, 2016). La incorporación de herramientas como *Apache Kafka* ha permitido que el actual software abierto más extendido para la computación distribuida *Apache Spark* cubra esta brecha con *Spark Streaming* (*Spark Kafka Integration*, 2016), que puede leer código en *Scala* o también en *Python* (con el módulo *PySpark*). Por lo anterior, una versión distribuida del clasificador permite realizar el análisis de sentimientos con *Apache Spark Streaming* en máquinas virtuales utilizando modelos entrenados con *Spark Machine Learning*. Este procedimiento es escalable usando servicios comerciales como los de *Amazon Web Services* (*AWS*), y sus servicios de almacenamiento (*Amazon S3*) y de cómputo distribuido (*Amazon Elastic Computing Cloud*, EC2), creando un conjunto flexible de instancias conectadas en la nube para calcular el análisis (lectura y escritura de datos directamente de o hacia S3), que permiten diseños viables y escalables de problemas de computación distribuida (figura 2).

5. Discusión y aplicaciones

La confrontación y el debate generado en España en los últimos años y en particular con la aparición en escena de dos partidos políticos de dimensión estatal (*Ciudadanos* y *Podemos*) cuya presencia en redes sociales es principal y asociada a un perfil electoral joven hiperconectado y social,

ha abierto nuevos escenarios de diálogo y debate político. La batalla dialéctica de estos nuevos partidos ha tenido y tiene lugar, al menos en una parte significativa, en las redes sociales, en particular en la plaza mayor de debate que es *Twitter*. En este contexto, la investigación en comunicación política y, en especial, de la opinión pública en tiempo real es un factor clave para planificar intervenciones políticas y sociales, y reaccionar oportunamente con proactividad a sus efectos. El análisis supervisado de sentimientos resulta un instrumento esencial para modelar los mensajes positivos y negativos en *Twitter* a lo largo de su cadena de valor comunicativa. Este instrumento de análisis permite predecir el sentimiento en las discusiones en tiempo real sobre temas directamente relacionados con los intereses de los partidos políticos y su agenda comunicativa. En este sentido, la técnica que describimos en este artículo presenta las siguientes innovaciones en el campo de estudio de la comunicación política: un clasificador de sentimientos y opiniones basado en aprendizaje automático:

- orientado a contenidos políticos transmitidos por redes sociales;
- orientado a contenidos en español;
- orientado a predecir el comportamiento electoral en España y otros países de habla hispana;
- que puede ser escalable para afrontar problemas de dimensión *big data.*

Los métodos y servicios computacionales implementados en el clasificador descrito pueden ayudar a los investigadores, a las consultoras y las empresas privadas en el campo de la opinión pública, el marketing y los estudios políticos y gubernamentales a estudiar grandes cantidades de tweets políticos en español ejecutando análisis de sentimientos en tiempo real sin las limitaciones de los enfoques basados en diccionarios. Todos estos métodos e instrumentos requieren de conocimientos y habilidades de programación básica, si bien utilizando scripts ya desarrollados se pueden implementar fácilmente. Esto permite que no se necesiten conocimientos matemáticos para ejecutar los modelos de aprendizaje automático más sencillos. Sin embargo, una comprensión teórica de los algoritmos es condición *sine qua non* para que se pueda realizar una interpretación de calidad de la técnica y sus resultados, lo que sugiere que estos estudios deben ser llevados a cabo por equipos inter-disciplinares o entrenados específicamente en estas metodologías y técnicas analíticas.

Por otro lado, en el caso de los servicios comerciales (p. e. *AWS*), los investigadores o empresas que usen esta técnica deberán considerar los costes financieros adicionales que el alquiler de estos servicios de computación distribuida representa en todo proyecto. El análisis supervisado de sentimientos está diseñado para que cualquier investigador entrenado pueda implementarlo sin dificultad, aunque trabajar con equipos interdisciplinares como hemos indicado (informáticos, estadísticos, expertos en lingüística computacional, expertos en comunicación social y política, etc.) puede mejorar la calidad de los resultados e interpretaciones y reducir los costes de los recursos necesarios. El procedimiento descrito para monitorizar los tweets políticos en *streaming* puede ayudar a mejorar las predicciones electorales, a probar los enfoques teóricos tradicionales y emergentes en la investigación de la opinión pública que requieren de datos longitudinales en series temporales. Esta técnica también puede contribuir al contraste de hipótesis en estudios experimentales que necesitan entradas en tiempo real para crear o adaptar estímulos, y validar sus modelos teóricos.

> Una versión distribuida del clasificador permite realizar el análisis de sentimientos con *Apache Spark Streaming* en máquinas virtuales

Específicamente, en el campo de la comunicación política, los principales interesados en usar el análisis supervisado de sentimientos son:

- Investigadores: los académicos pueden utilizar la técnica para llevar a cabo estudios descriptivos y correlacionales.
- Empresas de estudios de opinión pública: pueden utilizar el clasificador para monitorizar estados de opinión sobre temas políticos. Además, pueden utilizarlo para la construcción de indicadores avanzados de predicción de intención de voto, notoriedad, y popularidad y aceptación de candidatos.
- Consultoras políticas: para complementar la monitorización y estudio que hacen a hechos políticos en el país y planificar las estrategias de comunicación y persuasión política (marketing político), especialmente en ocasiones donde el análisis e interpretación de sentimientos derivados de intervenciones políticas requiere análisis y conclusiones proactivos en tiempo real.
- Partidos políticos: pueden utilizar la técnica para la planificación de sus acciones en función de los sentimientos sobre determinados temas públicos y candidatos electorales.

Las aplicaciones prácticas que nos ofrece el análisis supervisado de las opiniones políticas en tiempo real con técnicas de aprendizaje automático o de evaluación tendencial vía el análisis de *datos masivos* ya vienen siendo utilizadas desde la campaña de 2012 en las elecciones presidenciales en EUA del finalmente presidente reelecto Obama, a través de una solución *big data* de *Oracle* (con un coste medio de 3 a 4 millones de dólares EUA) y la implicación de equipos interdisciplinares de análisis y redacción de informes de campaña. El objetivo de estos equipos es simplificar el acceso a los datos, descubrir, interpretar y predecir de forma rápida la eficacia de los mensajes, la inversión final en los distintos soportes comunicativos, y a su vez custodiar y gobernar de forma segura todos los datos analizados (**Mariño-Angoso**, 2015). El director de la campaña electoral presidencial demócrata de 2012 en EUA, Jim Messina, contó en su equipo de campaña con más de 100 personas directamente especializadas en el análisis de datos (multiplicando por 5 los recursos y personas asignados a estas tareas en la campaña de 2008). La implementación de soluciones de base de datos *HPVertica MPP* (*massive parallel processing*) y modelos predictivos con *R* y *Stata* contribuyeron al éxito final. El visionario e innovador director de campaña afirmó que querían

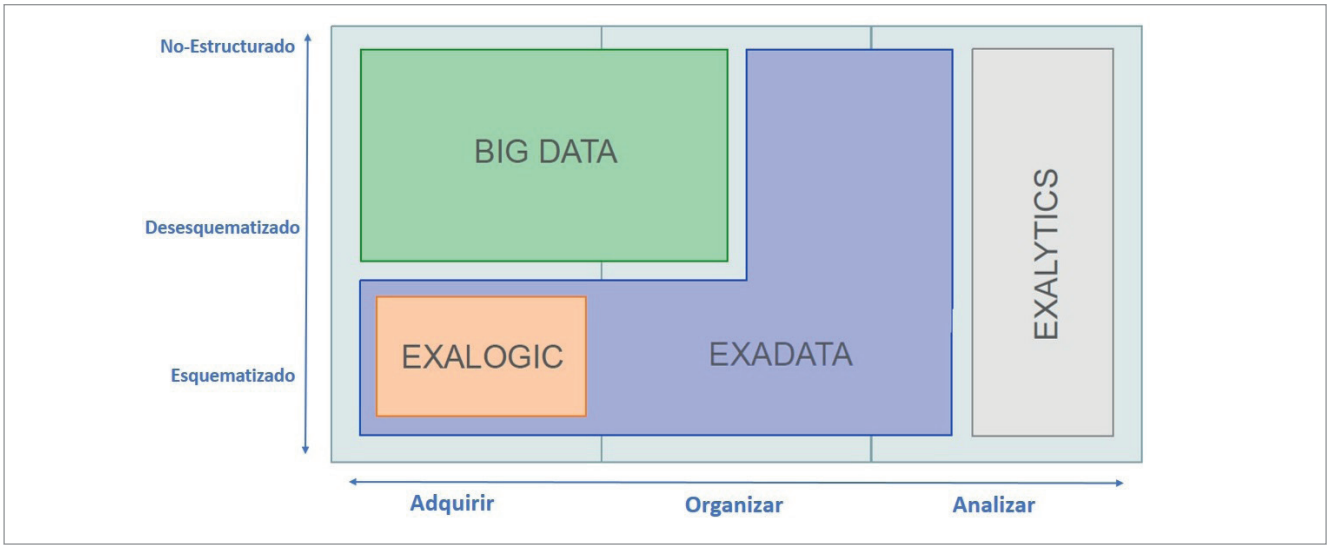

Figura 3. Solución *big data* de *Oracle* de la campaña presidencial de Obama 2012.
Fuente: Elaborado a partir de **Mariño-Angoso** (2015) y **Sorrells** (2012).

> "medir todo lo que pasaba para verificar que todas las decisiones se estaban tomando de forma fundada e inteligente" (**Lampitt**, 2013; **Sorrells**, 2012).

La figura 3 ilustra la solución *big data* de *Oracle* utilizada por el equipo de Obama

En conclusión, los retos y oportunidades que las soluciones *big data* y en particular que el análisis supervisado de sentimientos políticos con las metodologías y técnicas descritas nos lleva a afirmar que vivimos tiempos de *data-analysis-revolution* (DAR) en el análisis, interpretación y gestión de la comunicación política y corporativa. Las debilidades y amenazas, pero más relevantes fortalezas y oportunidades que estas metodologías y técnicas ofrecen a los investigadores en comunicación anticipan una revolución en el análisis científico de los procesos comunicativos asociados a partidos políticos y corporaciones empresariales. La tabla 1 presenta de forma sintética un análisis DAFO focalizado en el análisis supervisado de sentimientos con técnicas de aprendizaje automático e inteligencia artificial aplicado a la comunicación política.

La modelación de las opiniones políticas en *Twitter* a través del análisis supervisado de sentimientos es una metodología complementaria y necesaria para la contrastación y predicción de los resultados electorales en español, en España y países Latinoamericanos. En un contexto donde la calidad de las encuestas está siendo constantemente re-evaluada por los expertos y por la misma opinión pública, por su falta de precisión, el análisis supervisado de sentimientos se nos presenta como ese "escáner en 3D en tiempo real" adelantado que pueda complementar los análisis tradicionales de muestras representativas para predecir resultados e intención de voto. Además, la predicción de sentimientos políticos en tiempo real a lo largo de un período prolongado de tiempo en *Twitter* permitirá realizar análisis longitudinales para detectar

Tabla 1. Matriz DAFO del análisis supervisado de sentimientos en comunicación política

Debilidades	**Amenazas**
Perfeccionamiento de los modelos teóricos. Mejora del reconocimiento de los mensajes por encima del 70%. Interpretación correcta mediante *Lexicons* de aprendizaje continuo.	No presencia de todo el electorado-individuos en el ágora pública a analizar. No existencia de grupos interdisciplinares formados para una correcta implementación de la metodología. Existencia de grupos organizados alineados con los *stakeholders* que puedan perturbar la bondad estadística de la muestra a recoger en los contenidos digitales.
Fortalezas	**Oportunidades**
Proactividad y construcción de indicadores avanzados en tiempo real. Ventajas comparativas en operatividad, diseño muestral y coste. Robustez de la metodología, y calidad de la información y análisis facilitado. Tiempo de análisis reducido y prácticamente en tiempo real. Acceso a una muestra altamente representativa, e incluso de toda la población. Diseño experimental en local o en sistemas *big data* en la nube en función del volumen de datos y temporalidad requerida.	Implementación en el análisis y diseño de la planificación de la comunicación política y corporativa de gobiernos, partidos y empresas. Análisis adelantado de sentimientos políticos, o marca-notoriedad de campañas con el objeto de corregir proactivamente los dislates comunicativos. Mejora de la metodología y calidad final mediante la asociación con procesos tecnológicos de inteligencia artificial.

cambios en los temas como "personajes-candidatos" y en las variables analizadas a través del tiempo para un tema-temática o partido político. De esta manera se podría contrastar la relación de estos cambios con eventos puntuales, lo que a su vez permitiría llevar a cabo políticas de intervención basadas en la investigación y en el análisis de datos con una mayor proactividad y reactividad temporal.

6. Agradecimientos

Los autores agradecen a la *Fundación General* de la *Universidad de Salamanca* y al *Plan TCUE* [2015-2017 Fase 2] la financiación obtenida para el desarrollo de la prueba de concepto: *Clasificador en tiempo real de opiniones políticas en español con técnicas de aprendizaje automático* (*Autocop*). También agradecen a Mateo Álvarez y Javier Ramírez el desarrollo del código distribuido en *Spark*. Finalmente, agradecen a Miguel Vicente y a los miembros del programa de *Opinión Pública y Medios de Comunicación* del *Máster Universitario de Investigación en Comunicación Audiovisual* (*Muica*) de la *Universidad de Salamanca* su participación en el proceso de evaluación de las técnicas explicadas en este artículo.

7. Bibliografía

Arcila-Calderón, Carlos; **Barbosa-Caro, Eduar**; **Cabezuelo-Lorenzo, Francisco** (2016). "Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística". *El profesional de la información*, v. 25, n. 4, pp. 623-631.
https://doi.org/10.3145/epi.2016.jul.12

Bakliwal, Akshat; **Arora, Piyush**; **Madhappan, Senthil**; **Kapre, Nikhil**; **Singh, Mukesh**; **Varma, Vasudeva** (2012). "Mining sentiments from tweets". En: *Proceedings of the WASSA 2012*, pp. 11-18.
http://www.aclweb.org/anthology/W12-3704

Bakliwal, Akshat; **Foster, Jennifer**; **Van-der-Puil, Jennifer**; **O'Brien, Ron**; **Tounsi, Lamia**; **Hughes, Mark** (2013). "Sentiment analysis of political tweets: Towards an accurate classifier". En: *Proceedings of the Workshop on language in social media Lasm2013*, *Association for Computational Linguistics*, pp. 49-58.
http://www.aclweb.org/anthology/W13-1106

Bermingham, Adam; **Smeaton, Alan** (2010). "Classifying sentiment in microblogs: is brevity an advantage?". En: *Proceedings of the 19th ACM Intl conf on information and knowledge management*, pp. 1833-1836.
https://core.ac.uk/download/pdf/11309792.pdf

Bermingham, Adam; **Smeaton, Alan** (2011). "On using *Twitter* to monitor political sentiment and predict election results". En: *Proceedings of the Workshop on sentiment analysis where AI meets psychology* (*Saaip*), *Ijcnllp* 2011, pp. 2-10.
https://www.aclweb.org/anthology/W/W11/W11-3702.pdf

Bollen, Johan; **Mao, Huina**; **Pepe, Alberto** (2011). "Modeling public mood and emotion: *Twitter* sentiment and socio-economic phenomena". En: *Proceedings of the Fifth Intl AAAI conf on weblogs and social media*. *AAAI-Icwsm 2011*. Association for the Advancement of Artificial Intelligence, pp. 450-453.
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2826/3237

Bučar, Jože; **Povh, Janez**; **Žnidaršič, Martin** (2016). "Sentiment classification of the Slovenian news texts". En: *Proceedings of the 9th Intl conf on computer recognition systems* (*Cores 2015*), pp. 777-787.
https://doi.org/10.1007/978-3-319-26227-7_73

Cambria, Erik; **Schuller, Björn**; **Liu, Bing**; **Wang, Haixun**; **Havasi, Catherine** (2013). "Knowledge-based approaches to concept-level sentiment analysis". *IEEE intelligent systems*, v. 28, n. 2, pp. 12-14.
https://www.computer.org/csdl/mags/ex/2013/02/mex2013020012.pdf

Choy, Murphy; **Cheong, Michelle**; **Laik, Ma-Nang**; **Shung, Koo-Ping** (2011). "A sentiment analysis of Singapore presidential election 2011 using *Twitter* data with census correction". Report: *arXiv:1108.5520*.
https://arxiv.org/ftp/arxiv/papers/1108/1108.5520.pdf

Cobb, Wendi-N-Whitman (2015). "Trending now: Using big data to examine public opinion of space policy". *Space policy*, v. 32, pp. 11-16.
https://doi.org/10.1016/j.spacepol.2015.02.008

Feldman, Ronen (2013). "Techniques and applications for sentiment analysis". *Communications of the ACM*, v. 56, n. 4, pp. 82-89.
https://goo.gl/xW9veE
https://doi.org/10.1145/2436256.2436274

García-Cumbreras, Miguel-Ángel; **Villena-Román, Julio**; **Martínez-Cámara, Eugenio**; **García-Morera, Janine** (2016). "The evolution of the Spanish opinion mining systems". *Procesamiento de lenguaje natural*, v. 56, pp. 33-40.
http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5284/3078

Hurtado, Lluís F.; **Pla, Ferran**; **Buscaldi, Davide** (2015). "ELiRF-UPV en TASS 2015: Análisis de sentimientos en *Twitter*". En: *Proceedings of TASS 2015: Workshop on sentiment analysis at Sepln*, pp. 75-79.
http://ceur-ws.org/Vol-1397/elirf_upv.pdf

Kranjc, Janez; **Smailović, Jasmina**; **Podpečan, Vid**; **Grčar, Miha**; Žnidaršič, Martin; **Lavrač, Nada** (2015). "Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform". *Information processing & management*, v. 51, n. 2, pp. 187-203.
https://goo.gl/cnJYJ1
https://doi.org/10.1016/j.ipm.2014.04.001

Lampitt, Andrew (2013). "The real story of how big data analytics helped Obama win". *Infoworld*, 14 February.
http://www.infoworld.com/article/2613587/big-data/the-real-story-of-how-big-data-analytics-helped-obama-win.html

Leetaru, Kalev-Hannes (2012). *Data mining methods for the content analyst: An introduction to the computational analysis of content*. New York: Routledge. ISBN: 978 0 415895149

Madlberger, Lisa; **Almansour, Amai** (2014). "Predictions based on *Twitter* —A critical view on the research process". En: *Data and software engineering* (*Icodse*), *2014 Intl conf.*

IEEE, pp. 1-6.
https://doi.org/10.1109/ICODSE.2014.7062667

Mariño-Angoso, Miguel (2015). "Oracle big data and webdata. What we are doing and future tendencies. Big data environment and Oracle solutions". En: *Conference proceedings, Media metrics and webdata task force, Cost action conference Webdatanet*, University of Salamanca, pp. 23-30.
http://www.webdatanet.eu/data/167

O'Connor, Brendan; **Balasubramanyan, Rammnath**; **Routledge, Bryan-R.**; **Smith, Noah A.** (2010). "From *tweets* to polls: Linking text sentiment to public opinion time series". En: *Proceedings of the 4th Intl AAAI conf on weblogs and social media*, *Icwsm 2011*, pp.122-129.
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842

Pang, Bo; **Lee, Lillian**; **Vaithyanathan, Shivakumar** (2002). "Thumbs up?: sentiment classification using machine learning techniques". En: *Procs of the ACL-02 Conf on empirical methods in natural language processing*, v. 10, Association for Computational Linguistics, pp. 79-86.
http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf

Preethi, Peter G.; **Uma, Vilma**; **Kumar, Ajit** (2015). "Temporal sentiment analysis and causal rules extraction from *tweets* for event prediction". *Procedia computer science*, v. 48, pp. 84-89.
https://goo.gl/v1Pn3T
https://doi.org/10.1016/j.procs.2015.04.154

Sluban, Borut; **Smailović, Jasmina**; **Battiston, Stefano**; **Mozetič, Igor** (2015). "Sentiment leaning of influential communities in social networks". *Computational social networks*, v. 2, n. 1, pp. 1-21.
https://doi.org/10.1186/s40649-015-0016-5

Smailović, Jasmina; **Grčar, Miha**; **Lavrač, Nada**; **Žnidaršič, Martin** (2013). "Predictive sentiment analysis of *tweets*: A stock market application". *Human-computer interaction and knowledge discovery in complex, unstructured, big data. Lecture notes in computer science*, v. 7947, pp. 77-88.
http://first.ijs.si/FirstShowcase/Content/pub/HCI-KDD-2013.pdf
https://doi.org/10.1007/978-3-642-39146-0_8

Smailović, Jasmina; **Kranjc, Janez**; **Grčar, Miha**; **Žnidaršič, Martin**; **Mozetič, Igor** (2015). "Monitoring the *Twitter* sentiment during the Bulgarian elections". En: *IEEE Int conf on data science and advanced analytics* (*DSAA*).
https://goo.gl/s5WtYA
https://doi.org/10.1109/DSAA.2015.7344886

Sorrells, Amy (2012). "How big data and social won the election". *Oracle social spotlight*, 9 November.
https://blogs.oracle.com/socialspotlight/how-big-data-and-social-won-the-election

Spark Kafka Integration (2016). *Spark Streaming + Kafka Integration Guide*.
http://spark.apache.org/docs/latest/streaming-kafka-integration.html

Tumasjan, Andranik; **Sprenger, Timm O.**; **Sandner, Phillipp G.**; **Welpe, Isabell M.** (2010). "Predicting elections with *Twitter*: What 140 characters reveal about political sentiment". En: *Pros of the 4th Intl AAAI conf on weblogs and social media* (*Icwsm*), v. 10, n. 1, pp. 178-185.
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441

Turck, Matt; **Hao, Jim** (2016). *Big data landscape 2016 (Version 3.0)*.
http://mattturck.com/big-data-landscape-2016-v18-final

Turney, Peter D. (2002). "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews". En: *Procs of the 40th Annual meeting on Association for Computational Linguistics*, pp. 417-424.
http://www.aclweb.org/anthology/P02-1053.pdf

Van-Zoonen, Ward; **Van-der-Meer, Toni** (2016). "Social media research: The application of supervised machine learning in organizational communication research". *Computers in human behavior*, v. 63, pp. 132-141.
https://doi.org/10.1016/j.chb.2016.05.028

Vinodhini, Gina; **Chandrasekaran, R. M.** (2012). "Sentiment analysis and opinion mining: A survey". *International journal of advanced research in computer science and software engineering*, v. 2, n. 6, pp. 282-292.
https://goo.gl/c1hWpy

Wang, Hao; **Can, Dogan**; **Kazemzadeh, Abe**; **Bar, François**; **Narayanan, Shrikanth** (2012). "A system for real-time *Twitter* sentiment analysis of 2012 us presidential election cycle". En: *Procs of the ACL 2012 System demonstrations. Association for Computational Linguistics*, pp. 115-120.
http://www.aclweb.org/anthology/P12-3020