APPLICATIONS OF MACHINE LEARNING IN TWO CASES OF SPANISH POLICY: ELECTIONS 26J AND CATALONIA INDEPENDENCE

by

DAVID PEREA EL KHALIFI

A thesis submitted in conformity with the requirements
for the MSc in Economics, Finance and Computer Science

University of Huelva & International University of Andalusia

November 2017

Aplicación del aprendizaje automático en dos casos de política española: Elecciones 26J e Independencia de Cataluña

David Perea El Khalifi

Máster en Economía, Finanzas y Computación

Ana María Rodríguez Santiago
Universidad de Huelva y Universidad Internacional de Andalucía

2017

Abstract

The political instability that Spanish democracy is suffering in these months is a valuable material for research. This research is based on an analysis of Machine Learning, using unsupervised learning and supervised learning techniques. The latter, together with natural language processing techniques. The analysis is divided into two parts. The study of the general elections of June 26, 2017 to explain the obtaining of political power, and the study of tweets about the Independence of Catalonia to explain the diffusion of political communication in this matter, according to the origin of the users. Obtaining, in the first case, relevant information so that the political parties establish their electoral strategies and the second of the studies outlines the seriousness of the process of Catalan separatism that the media transmits, especially the international ones.

JEL classification: C19; C54; C55; D72; N44

Key words: Catalonia, machine learning, natural language processing, political elections, Spanish policy, tweets.

Resumen

La inestabilidad política que está sufriendo en estos meses la democracia española, constituye un valioso material de trabajo para la investigación. En esta investigación se basa en un análisis de Aprendizaje Automático, utilizando técnicas de aprendizaje no supervisado y de aprendizaje supervisado. Esta última, junto a técnicas de procesamiento de lenguaje natural. El análisis se divide en dos partes. El estudio de las elecciones generales del 26 de junio del 2017 para explicar la obtención del poder político y el estudio de tweets de sobre la Independencia de Cataluña para explicar la difusión de la comunicación política en este asunto, según él origen de los usuarios. Obteniéndose, en el primer caso, información relevante para que los partidos políticos establezcan sus estrategias electorales y en el segundo, se esboza la gravedad del proceso independentista catalán que transmiten los medios de comunicación, en especial los internacionales.

Palabras clave: Aprendizaje automático, Cataluña, elecciones políticas, política Española, procesamiento del lenguaje natural, tweets.

Agradecimientos

Una vez finalizado un trabajo tan arduo y lleno de dificultades como es el desarrollo de un trabajo de fin de máster, es inevitable, que de cierto modo irrumpa un muy humano egocentrismo que lleva a concentrar la mayor parte del mérito del aporte realizado. Sin embargo, el análisis objetivo muestra inmediatamente que la magnitud de ese aporte hubiera sido imposible sin la participación de personas e instituciones que han facilitado las cosas para que este trabajo aporte una contribución valiosa en las investigaciones de este ámbito. Por ello, es para mí un verdadero placer utilizar este espacio para ser justo y consecuente con ellas, expresándole mis agradecimientos.

Debo agradecer de manera especial y sincera a la Profesora Dª Ana María Rodríguez Santiago, del departamento de Economía de la Universidad de Huelva, por aceptarme para realizar este trabajo de fin de master bajo su dirección. Su disposición y disponibilidad a la hora de revisar el mismo, junto a su capacidad para guiar mis ideas ha sido un aporte invaluable en el desarrollo de este trabajo. No cabe duda que su participación ha enriquecido el trabajo realizado.

Quiero expresar también mis más sinceros agradecimientos al director del master en cuestión, D. Emilio Congregado, el haberme facilitado siempre los medios suficientes para llevar a cabo este trabajo de fin de máster. Le quiero dedicar también un agradecimiento especial por la creación de este máster, Master en Economía, Finanzas y Computación (UHU-UNIA). Sin su participación y dedicación activa en el mismo no hubiera tenido la oportunidad de poder plantearme ser un investigador del análisis de datos.

Al mismo tiempo, es de obligado cumplimiento agradecer a los profesores de la universidad de Huelva, José María Millán y Gonzalo A. Aranda-Corral, junto al profesor de la universidad de Málaga, Juan Marcos Castros Boñano, por estar completamente dispuestos a dirigir mi trabajo si este hubiera seguido una línea investigadora de sus competencias.

Entrando en el ámbito personal, debo agradecer a una persona, que ha seguido día a día todo este laborioso proceso, su apoyo y consejos sobre el mismo. Gracias amigo, Miguel Ángel García Díaz.

Una mención especial para quienes han sido mis compañeros, amigos y familia durante estos últimos cuatro meses en una experiencia profesional internacional. Gracias a Futuro Digitale y a toda la gente, tanto italianos como internacionales, que han pasado por mi vida en Terranova da Sibari (Italia). La cual ha sido el escenario de la elaboración de este trabajo. No se puede concebir este trabajo sin su siempre oportuno apoyo, compresión y participación. Grazie mille.

Con permiso de todos, un agradecimiento honorable a mi familia. En especial a mis padres. Gracias a su dedicación a ciegas y al apoyo infalible y confianza en mí, han sido un estímulo constante durante todos mis años de estudios y en el desarrollo de este trabajo. Sin ellos, nada de esto hubiera sido posible. Muchas gracias papis.

Gracias a todos.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

1 Introducción

"¿Pueden las máquinas pensar?". Con esta pregunta, que planteó Alan Turing, uno de los padres de la ciencia de la computación, en su artículo seminal "*Computing Machinery and Intelligence*" (1950), se esbozó el campo de lo que hoy denominamos como Aprendizaje Automático.

Este término, aprendizaje automático, fue acuñado por un pionero estadounidense en el campo de los juegos informático y la inteligencia artificial Arthur Samuel en su artículo seminal "*Some Studies in Machine Learning Using the Game of Checkers*" (1959) donde definió: "Aprendizaje automático: campo de estudio que da a las computadoras la capacidad de aprender sin estar programado explícitamente."

Mitchell (1997) ofreció una definición más detallada del termino aprendizaje automático: "Se dice que un programa de computadora aprende de la experiencia E con respecto a una clase de tareas T y la medida de rendimiento P si funciona en tareas en T, medido por P, mejora con la experiencia E".

Los diferentes algoritmos de aprendizaje automático se clasifican en dos categorías, dependiendo de si hay una "señal" o "retroalimentación" de aprendizaje disponible para un sistema de aprendizaje: Aprendizaje Supervisado, cuando se conoce la clasificación de los datos, y Aprendizaje No Supervisado, si se desconoce la clasificación de los datos (Russel & Norvig, 2003).

Respecto al análisis de nuestro estudio. La explicación de la obtención del poder político y de la difusión de la comunicación política constituye una de las áreas más apasionantes de la investigación en ciencia políticas. En este sentido, recientemente varios acontecimientos políticos relevante han traído como consecuencia la inestabilidad política de la democracia en España, cuyos hechos constituye un valioso material de trabajo para los estudios en estos campos.

Tras el fin del régimen dictatorial del general Francisco Franco (1975) hasta nuestros días (2017), han acontecido numerosos hechos políticos que han pasado a los libros de la historia de España. Desde el periodo de la Transición, con la proclamación del Rey Juan Carlos I (1975), las primeras elecciones generales democráticas (1977), la aprobación de la Constitución (1978), el golpe de estado del 23-F (1981), hasta la democracia de nuestros días, con la entrada de España en la

Comunidad Económica Europea (1986), la entrada del euro en circulación (2002), la crisis económica (2008), la abdicación de Juan Carlos I y proclamación del Rey Felipe VI (2014), entre otros muchos. Los dos acontecimientos políticos a destacar en los dos últimos años en España son, sin duda, las últimas elecciones generales (26 de junio del 2016) y la crisis catalana que rodea al Referéndum de independencia de Cataluña (2017).

Consecuentemente, el objetivo presente de esta investigación es utilizar técnicas de aprendizaje automático para definir la obtención del poder político en las últimas elecciones generales (26J) y para explicar la difusión de la comunicación política en la crisis independentista catalana mediante tweets de medios de comunicación y personajes influyentes de Cataluña, del resto de España e internacionales.

El resto del artículo se organiza de la siguiente manera: En el apartado 2 se define la motivación y justificación que ha llevado al autor a realizar esta investigación. En el apartado 3 se presenta una revisión de la literatura principal en la que se basa la investigación. La metodología seguida en la investigación y los datos utilizados están explicados en los apartados 4 y 5, respectivamente. En el apartado 6 se presentan los resultados obtenidos de la investigación. Por último, las conclusiones obtenidas y la línea definida para futuras investigaciones quedan expuestas en el apartado 7.

2 Motivación y justificación

Se ha decidido estudiar estos dos hechos políticos, las elecciones generales del 26J y la crisis independista catalana, ya que estos dos hechos conjuntos han roto el mayor periodo de estabilidad política de la historia reciente de España, entendiendo por estabilidad política como la de normalidad a nivel institucional; de ahí que se apunte que un sistema político no es estable cuando los distintos elementos que configuran el orden establecido por un régimen de gobierno, se alteran de tal forma que éstos no pueden alcanzar sus objetivos e incumplen así sus compromisos (Sanders, 1981).

Es por esta definición que el período de 40 años de reinado de Juan Carlos I (1975-2014) es considerado uno de los periodos de mayor estabilidad política de la historia reciente de España (Barranco, 2017).

Acerca de nuestro primer análisis, el 26 de junio de 2016, los ciudadanos de España se ven obligados a acudir de nuevo a las urnas, seis meses después de los comicios generales del 2015 (20D), para desbloquear una situación política inédita en el país y elegir al séptimo presidente de la actual democracia. La undécima legislatura ha batido el récord de ser la más corta por la incapacidad de los partidos para alcanzar un acuerdo. No cabe duda que las elecciones generales del 26J han marcado un hito en la historia de la democracia española, al ser las primeras elecciones repetidas de la historia de España (Menéndez, 2016). Los resultados, junto a las elecciones generales del 2015 (20D), han dado lugar a un Parlamento extremadamente dividido en términos de partidos políticos, lo que ha supuesto un gran esfuerzo para establecer un acuerdo de gobierno entre las diferentes fuerzas.

Tras estos dos últimos comicios se ha iniciado un periodo democrático nuevo, caracterizado por una caída severa del sistema bipartidista español (PSOE-PP), donde los partidos se han visto forzados a pactar con un mayor número de fuerzas políticas para poder llevar a cabo sus iniciativas legislativas (Tirado, 2016) .

El estudio realizado de los resultados de las elecciones del 26J analiza los municipios que se consideran ciudades, puesto que el 80% de la población española vive en ciudades (García, 2016). Para ello nos basamos en el término de ciudades que se definió en la Conferencia Europea de Estadística de Praga (1966) como las aglomeraciones de más de 10.000 habitantes (Capel, 1975).

Por otra parte, referente al segundo de los estudios, los movimientos separatistas más persistentes en España se han desarrollado y han permanecido durante largos años de la vida del Estado español moderno, en la región vasca y en Cataluña. Ambos movimientos se han presentado en las zonas de mayor prosperidad (Rodríguez, 2017) . Pero han sido los últimos movimientos del nacionalismo catalán los que han provocado un conflicto institucional a nivel nacional.

La relación entre el Gobierno de España y la Generalitat está pasando por un mal momento (Harguindéguy, Rodríguez-López, & Sánchez, 2017), el más tenso de la democracia. Donde no se puede apuntar la magnitud de sus consecuencias, debido a que aún estamos inmersos en esta crisis institucional. Hará falta un espacio temporal de varios años para poder sacar conclusiones precisas de la dimensión de este conflicto. Lo que no cabe duda es que ha causado una fractura en la sociedad española.

La gravedad del momento es de tal magnitud que, como indicó José Álvarez Junco, catedrático emérito de Historia del Pensamiento y de los Movimientos Sociales y Políticos de la Universidad Complutense de Madrid, lo demuestra el mensaje institucional televisado que pronunció el rey Felipe VI a los españoles para abordar la situación excepcional que se vive en España (Pichel, 2017). En palabras de Felipe VI: "Estamos viviendo momentos muy graves para nuestra vida democrática". Tras los hechos acontecidos desde el referéndum de independencia de Cataluña el 1 de octubre del 2017 (1-O), los españoles son conscientes de la importancia del momento.

La elección de analizar los tweets se debe a que la red social Twitter ha pasado de ser un canal concebido para actualizaciones personales a ser un medio para la comunicación, la conversación y la concertación política (Moya Sánchez & Herrera Damas, 2015). El análisis de la plataforma de marketing para redes sociales Cool Tabs demostró que “el referéndum del 1-O generó casi 12 millones de tweets”. Para el análisis, interesa especialmente conocer la información que se ha transmitido desde el panorama internacional sobre este asunto.

Además, con este artículo se busca generar motivación de investigadores inmersos en el aprendizaje automático u otra rama de las ciencias de la computación, para ayudar a explicar las diversas áreas de investigación de la ciencia política. Pues la ciencia política, al igual que otras ciencias como la psicología, la geografía, la sociología, la economía, la antropología, la comunicación, la jurídica, la historia etc. son áreas de investigación híbridas (Dogan, 1996). Es decir, que todas estas ciencias se relacionan y enlazan con otras para obtener una investigación más precisa. En el estudio se combinan conocimientos de las ciencias de la computación y de las ciencias políticas, para ofrecer una imagen más precisa de la situación.

3 Revisión literaria

En este apartado se describen los principales trabajos de investigación que comparten relación con el estudio presente. En el subapartado 3.1. se detalla las investigaciones que aplican técnicas de aprendizaje no supervisado para análisis políticos, como en el análisis de los resultados electorales del 26J. En el subapartado 3.2 se detalla las investigaciones que analizan tweets, como en el análisis de los tweets sobre la crisis independentista catalana.

3.1 Elecciones 26J

Son varias los trabajos de investigación que analizan aspectos políticos, en especial, elecciones, mediante la técnica de Análisis Factorial (aprendizaje no supervisado).

Las investigaciones en la que más se asemeja al estudio presente son las dos siguientes.

La primera Rectora de una universidad catalana (Universitat Pompeu Fabra), Rosa Virós, catedrática en Ciencias Políticas y de la Administración en la misma universidad, publico en 1979 "*Algunes notes sobre el comportament electoral a Catalunya el 15 de juny de 1977*".

En esta investigación se analiza los resultados en Cataluña de las primeras elecciones generales de la democracia española (15 de junio de 1977), mediante un análisis factorial. Se obtiene como conclusión la existencia de cuatro factores. Se tratan de un factor izquierda-centro, un factor de centro, otro factor de derecha continuista y por último un factor de abstención.

En las siguientes elecciones generales, las segundas elecciones generales de la democracia española (10 de marzo de 1979), se analizan sus resultados mediante un análisis factorial de correspondencias en el artículo seminal "*Geografía electoral española. Una aplicación del análisis factorial de correspondencias de los resultados de las elecciones del 10 de marzo de 1979*" (1980). En el cual el Catedrático del Departamento de Estadística e Investigación Operativa Aplicadas y Calidad de la Universidad Politécnica de Valencia, Rafael Romero Villafranca, dirige la investigación de Luisa Rosa Zúnica Ramajo.

Se obtiene como conclusión la existencia de cuatros factores, que representa una imagen de la sociedad de entonces. Se tratan de dos factores fundamentales (Centralismo-nacionalismo | No nacionalistas {derecha-izquierda}) y otros dos de una importancia menor (Abstención del PSOE | Extremismo-moderación).

Sin embargo, estas líneas de investigación no son solo características del siglo anterior y del ámbito español. En el panorama internacional se publicaron en esta última década interesantes análisis políticos a través del análisis factorial.

"*La calidad de la democracia: Un análisis comparado de América Latina*" (2011) artículo del doctor en Ciencias Políticas y Sociología por la Universidad de Deusto, Mikel Barreda, que compara la calidad de las democracias latinoamericanas.

Se obtiene tres factores tras realizar el análisis factorial. Un factor con la calidad más elevada (Chile, Uruguay, Costa Rica y Panamá), otro factor con los niveles más bajos (Guatemala, Paraguay, Venezuela, Colombia, Honduras y Ecuador) y el último factor con los niveles de democracia más intermedio (países restantes).

"*Forecasting 2016 US Presidential Elections Using Factor Analysis and Regression Model*" (Sinha, Pankaj, et al., 2016) donde mediante análisis factorial se clasifica los parámetros económicos y no económicos responsables de pronosticar el resultado de las elecciones presidenciales de EE.UU. 2016.

El principal factor económico importante en las elecciones presidenciales estadounidenses de 2016 es el crecimiento de la economía, y se encuentra que el factor anti-incumbencia que significa cuánto tiempo el partido en el cargo ha estado controlando la Casa Blanca es un factor no económico importante que probablemente juegue un papel dominante en la elección.

De esta forma, se comprueba que este estudio sigue una línea de investigación relevante desde hace décadas. En la cual muchos investigadores centran sus trabajos.

3.2 Tweets sobre la independencia de Cataluña

Los trabajos de investigación en los cuales se analizan textos de aspectos políticos o se analizan tweets se basan en una línea de investigación muy recientes y novedosa, en vía de desarrollo, que presenta gran atención en el campo computacional. Es por eso que el número de publicaciones sobre minería de texto y procesamiento de lenguaje natural (PLN) han crecido considerablemente en los últimos años, pero son pocas las cuales aportan un análisis desarrollado del mismo.

Sin embargo, aún no hay ningún artículo de este ámbito centrado en el asunto de la independencia de Cataluña del 2017, debido a la brevedad de los acontecimientos y que aún no han concluido.

En el artículo "*Overview of the 1st Classification of Spanish Election Tweets Task at IberEval 2017*" (Gimenez, Baviera, et al., 2017), mediante el algoritmo de K vecinos más cercanos (KNN),

se pretende clasificar los tweets (etiquetados a mano por expertos) sobre las elecciones generales de 2015. Concluyendo que la clasificación de los tweets, cuando los temas son similares, es una tarea difícil que no se realiza con una alta precisión

"*Supervised sentiment analysis of political messages in Spanish: Real-time classification of tweets based on machine learning*" (Arcila-Calderon, Ortega-Mohedano, Jimenez-Amores, & Trullenque, 2017) es un artículo donde 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.

Se concluye que la modelación de las opiniones políticas en Twitter a través del análisis supervisado de sentimientos es una metodología complementaria y necesaria para la contratación y predicción de los resultados electorales. Destacando los nuevos escenarios de diálogo y debate político que se han abierto con la aparición de dos partidos políticos de (Ciudadanos y Podemos) cuya presencia en redes sociales es principal y asociada a un perfil electoral joven hiperconectado y social.

"*Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength Tecnhiques for sentiment analysis in Twitter*" (Baviera, 2017). Este artículo describe dos tipos de estrategias para abordar procesar automáticamente los tweets sin perder fiabilidad. La primera estrategia se basa en los procesos de Aprendizaje Automático Supervisado. Su aplicación requiere integrar algunas herramientas del Procesamiento de Lenguajes Naturales y tomar como punto de partida un corpus clasificado. El segundo enfoque está basado en diccionarios de polaridad. Se concluye que el hecho de tener textos cortos en cada mensaje hace razonable intentar desarrollar métodos que evalúen de forma automática la carga de sentimiento transmitida en cada tweet. El desarrollo de estos procesos aplicados a los mensajes en castellano está siendo objeto de creciente interés por parte de los investigadores en comunicación y del sector de la ingeniería computacional.

4 Metodología

En este apartado se detallan las principales metodologías aplicadas en el análisis. Se emplearán tanto algoritmos de aprendizaje no supervisado como de aprendizaje supervisado. Con este último, conjuntamente, se implantará técnicas de procesamiento de lenguaje natural (PLN).

El subapartado 4.1. se centra en la técnica aplicadas a los resultados electorales del 26J mientras que el análisis de los tweets sobre la crisis independentista catalana se aborda mediante las técnicas que se definen en el subapartado 4.2.

4.1 Aprendizaje No Supervisado

4.1.1 Análisis Factorial

La metodología principal aplicada en este estudio es la desarrollada por el psicólogo y estadístico Spearman bajo el nombre de Análisis Factorial (A.F.), promovida posteriormente por el ingeniero mecánico y psicólogo Thurstone. En este análisis se ofrece la idea principal de las características del método utilizado, basadas principalmente en la obra de Sperman (1904), "'*General intelligence', Objectively Determined And Measured*", y en la de Thurstone (1947), "*Multiple factor analysis*", así como en el artículo de De la Fuente Fernández (2011) "*Análisis Factorial*".

El modelo de A.F. es un modelo de regresión múltiple que relaciona las variables de interés (variables latentes) con variables observadas. En otras palabras, el A.F. es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables. Su propósito último consiste en buscar el número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos (De la Fuente Fernández, 2011).

Consideremos un conjunto de p variables observadas $X = (x_1, x_2, \dots, x_p)$ con sus variables específicas (términos residuales) $x_1, u_2, \dots, u_p$ que se asumen relacionadas con un número dado (k) de variables latentes $f_1, f_2, \dots, f_k$, donde $k < p$, y de cargas factoriales (coeficientes) a_{ij} $\{i = 1, \dots, p; j = 1, \dots, k\}$, mediante una relación del tipo:

$$x_1 = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1k}f_k + u_1 a_{11}f_1$$

$$x_2 = a_{21}f_2 + a_{22}f_2 + \cdots + a_{2k}f_k + u_2$$

$$\vdots$$

$$x_p = a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pk}f_k + u_p$$

O de modo más conciso:

$$x = \Lambda f + u$$

Donde:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}, \qquad f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}$$

Así, con las suposiciones previas, se tiene la comunalidad de la variable x_i (h_i^2), que representa la varianza de la variable x_i y su especificidad (ψ_i) , que representa la parte de la varianza específica de cada variable x_i.

$$Var(x_i) = \sum_{j=1}^{k} a_{ij}^2 + \psi_i = h_i^2 + \psi_i \ (i = 1,2, \dots, p)$$

Además de la matriz de correlación:

$$Cov(x_i, x_I) = Cov\left(\sum_{j=1}^{k} a_{ij}F_j \, , \sum_{j=1}^{k} a_{ij}F_j\right) = \sum_{j=1}^{k} a_{ij}a_{ij} \ \ \forall i \neq I$$

Para efectuar el A.F., es condición necesaria que las variables originales estén intercorrelacionadas porque si no lo estuvieran no se podría explicar nada de las variables. Es por eso que se obtiene la matriz de las correlaciones muestrales, para comprobar si sus características son las adecuadas para realizar un A.F. Existen varios indicadores para analizar la matriz de correlación (KMO test, Bartlett test). En cualquier caso, la hipótesis planteada, siendo R_p la matriz de correlación de las variables observadas, es la siguiente:

$H0: |R_p| = 1 = I$ (variables NO intercorrelacionadas)

$H1: H0\ no\ cierta$ (variables SI intercorrelacionadas)

El criterio para rechazar la hipótesis nula, según el test KMO, es si se obtiene valores de Medida de Adecuación de la Muestra (MSA) bastante altos (>= 0.7 aprox.), mientras que según el Bartlett test, es si se obtienen valores altos de la chi-cuadrado χ^2 o un determinante bajo (<0.05).

Por lo tanto, si alguno de los test rechaza la hipótesis nula esto significa que hay variables con correlaciones altas y se podrá realizar el A.F.

4.1.2 K-means clustering

Esta técnica de agrupamiento, K-means, fue desarrollada principalmente por el estadístico Macqueen en su obra (1967) "*Some methods for classification and analysis of multivariate observations*". La cual junto al artículo de Cambronero y Moreno (2006) "*Algoritmos de aprendizaje: knn & kmeans*" se partirá en este análisis para ofrecer una idea principal de las características de la técnica.

Sigue un procedimiento simple de partición de un conjunto de *n* observaciones en *k* grupos. Donde cada observación pertenece al grupo cuyo valor medio es más cercano.

A partir, de un conjunto de datos $D_n = (x_1, x_2, \dots, x_n)$, donde cada observación es un vector real de d dimensiones, se construye una partición de D_n en k grupos $(k \leq n)$ con el propósito de minimizar dentro de cada grupo la suma de los cuadrados $S = (S_1, S_2, \dots, S_k)$:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2$$

Donde μ_i es la media de puntos en S_i.

4.2 PLN y Aprendizaje Supervisado

4.2.1 Text Mining

La minería textual es un área multidisciplinar basada en la identificación y extracción de nuevo conocimiento a partir de colecciones de documentos o textos, es decir, de información no estructurada, mediante la minería de datos, el aprendizaje automático, la estadística y la lingüística computacional (Brun y Senso, 2004).

El análisis del texto implica, análisis léxico para estudiar la distribución de frecuencia de palabras, análisis de sentimiento, clasificación del texto, y entre otras muchas tareas.

En este apartado se ofrece la idea principal de las características del análisis empleado.

Estudiar la distribución de frecuencia de palabras es útil para poder conocer de qué trata el texto. Sin embargo, hay que tener en cuenta que existen palabras que en un documento ocurren muchas veces, pero que no son importante. Como, según la RAE, las palabras en español: “de”, “la”, “que”, “el”, “en”, etc.

Se puede optar el enfoque de agregar palabras como estas a una lista de palabras y eliminarlas antes del análisis, pero es posible que algunas de estas palabras sean más importantes en algunos textos que otras. Este no es un enfoque muy sofisticado para ajustar la frecuencia de términos para palabras de uso común.

Otro enfoque es observar la frecuencia inversa del texto, que disminuye el peso de las palabras de uso común y aumenta el peso de las palabras que no se usan mucho en una colección de documentos. Este enfoque está destinado a medir la importancia de una palabra para un texto.

Se aplica mediante esta definición:

$$if(term) = \ln\left(\frac{n_{documents}}{n_{documents\ containing\ term}}\right)$$

La ley de Zipf (1940) determinada lengua la frecuencia de aparición de distintas palabras sigue una distribución que puede aproximarse por:

$$P_n \sim \frac{1}{n^a}$$

Donde P_n representa la frecuencia de la n-ésima palabra más frecuente y el exponente a es un número real positivo, en general ligeramente superior a 1. Esto significa que el segundo elemento se repetirá aproximadamente con una frecuencia de 1/2 de la del primero, el tercer elemento con una frecuencia de 1/3 y así sucesivamente (Montemurro,2001).

La ley de Zipf es una ley de potencias, lo que quiere decir que da igual el tamaño del texto que estemos estudiando y que esta proporción en la frecuencia de aparición de las palabras siempre se cumple. Así que da igual que hablemos del conjunto de todo lo publicado en ese idioma, de los libros de un autor determinado, de un libro en concreto, de una revista, o de un artículo. (economíaDigital – Wicho, 2016).

El análisis de sentimientos identificar y extraer información subjetiva del documento. Es una tarea de clasificación masiva de documentos de manera automática, en función de la connotación positiva o negativa del lenguaje ocupado en el documento (Liu, 2007).

Se pueden utilizar las herramientas de minería de texto para abordar el contenido emocional del texto mediante programación. Una de las múltiples opciones es realizar el análisis de sentimiento mediante el paquete *tidytext* del software estadístico R como se muestra en la Figura 1.

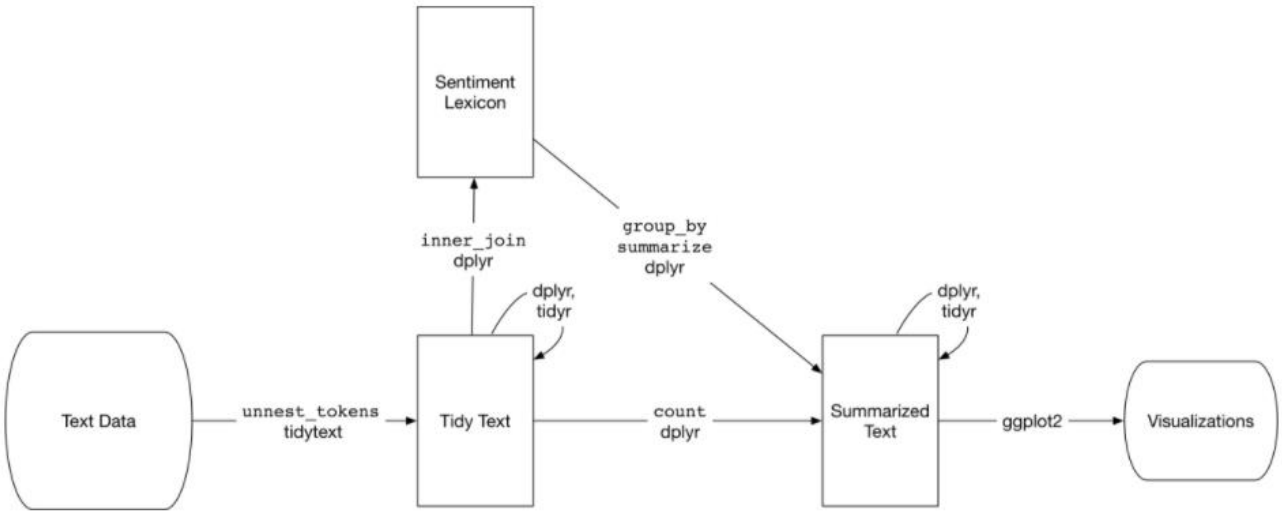

Figure 1. Análisis sentimientos mediante *tidytext* (R). Fuente: (Silge & Robinson, 2017).

4.2.2 K-nearest neighbors

La técnica aplicada; K vecinos más cercanos (KNN), es la desarrollada por los estadísticos Fix y Hodge en su artículo (1951) "*Discriminatory analysis-nonparametric discrimination: consistency properties*", promoviéndose y ampliándose posteriormente en el artículo "*Nearest neighbor pattern classification*" (Cover & Hart, 1967).

En este análisis se ofrece la idea principal de KNN, basadas principalmente en las dos obras anteriores mencionadas y en la de Cambronero y Moreno (2006) "*Algoritmos de aprendizaje: knn & kmeans*".

Sigue un procedimiento en el que una nueva observación se va a clasificar en la clase más frecuente a la que pertenecen sus observaciones más similares.

A partir, de un conjunto de datos $D_n = (X_1, X_2, \dots, X_n)$, donde cada observación pertenece a una clase del conjunto de clases $C_n = (c_1, c_2, \dots, c_m)$, se debe encontrar una función tal que cada x_i es asignada a una clase C_j.

Por lo tanto, seria:
Dado: $D_n = \{(X_1, c_1), (X_2, c_2), \dots, (X_n, c_m)\}$
Siendo $X = (x_1, x_2, \dots, x_n)$ nuevas observaciones a clasificar
Para toda observación ya clasificada (x_i, c_i) calcular $d_i = d(X_i, X) = \sqrt{\sum_{i=1}^{n}(X_i - X)^2}$
Ordenar $d_i (i = 1,2, \dots, N)$ en orden ascendente
Quedarnos con los K casos D_X^K ya clasificados más cercanos a X
Asignar a la X la clase más frecuente en D_X^K.

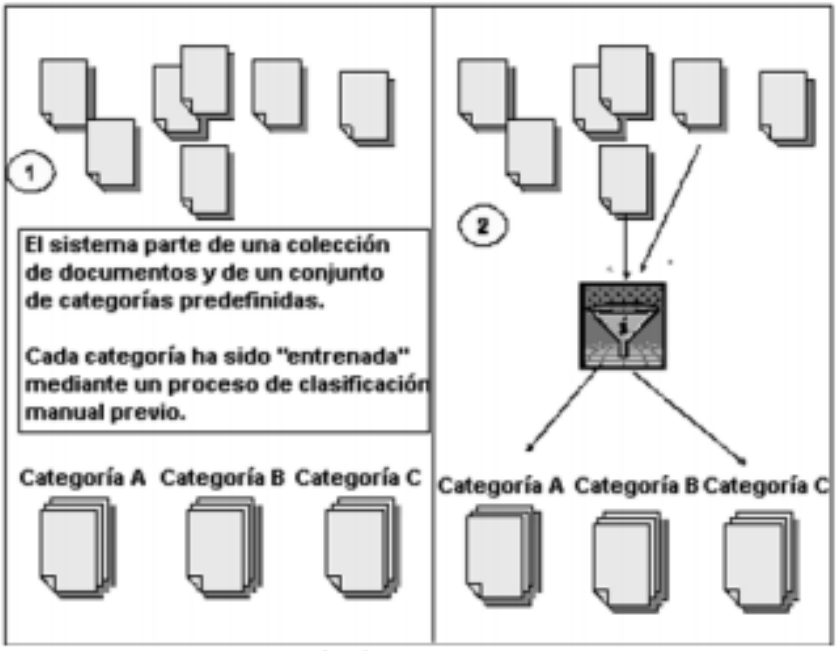

Figure 2. Clasificación texto. Fuente: (Brun y Senso, 2004).

Para emplear este algoritmo, el conjunto de datos se divide al azar en dos conjuntos. El de entrenamiento y el de prueba. El modelo de clasificación KNN se elabora a partir del conjunto de entrenamiento, mientras que el conjunto de prueba determina la precisión del modelo.

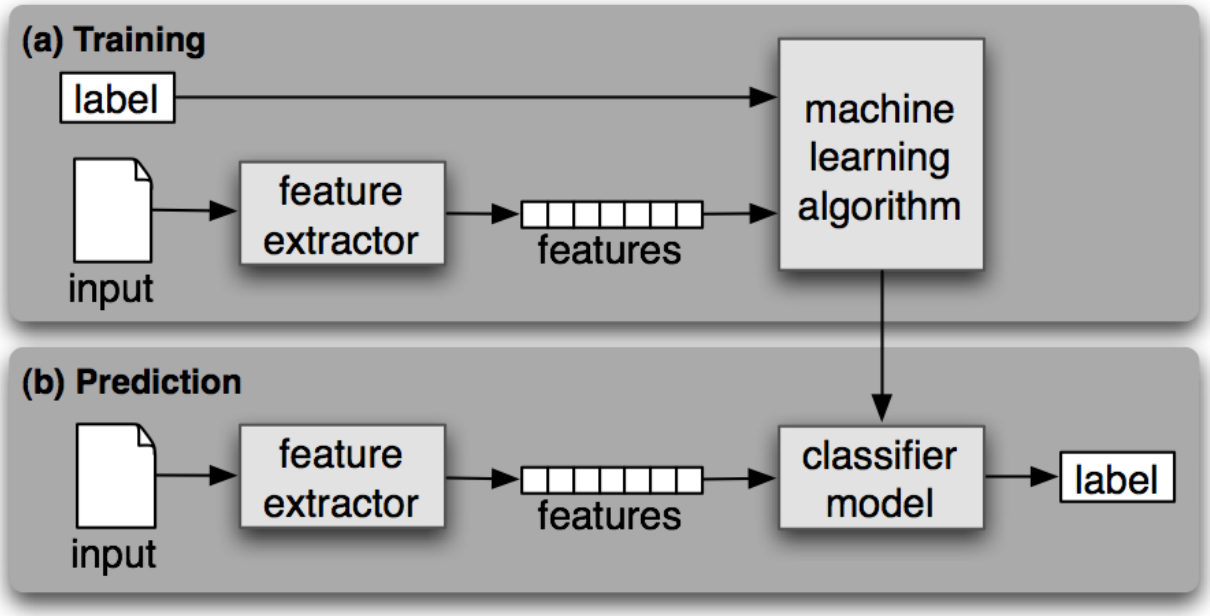

Figure 3. Clasificación Supervisada. Fuente: (Kao & Poteet, 2007).

La información acerca de las predicciones que se obtiene del modelo de clasificación KNN, se recogen en una matriz de confusión (Kohavi and Provost, 1998). En la cual se comparan las clases del conjunto de observaciones predichas versus las clases a la que estas realmente pertenecen. De esta manera, cada fila de la matriz representa el número de predicciones de cada clase, mientras que cada columna representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo las clases.

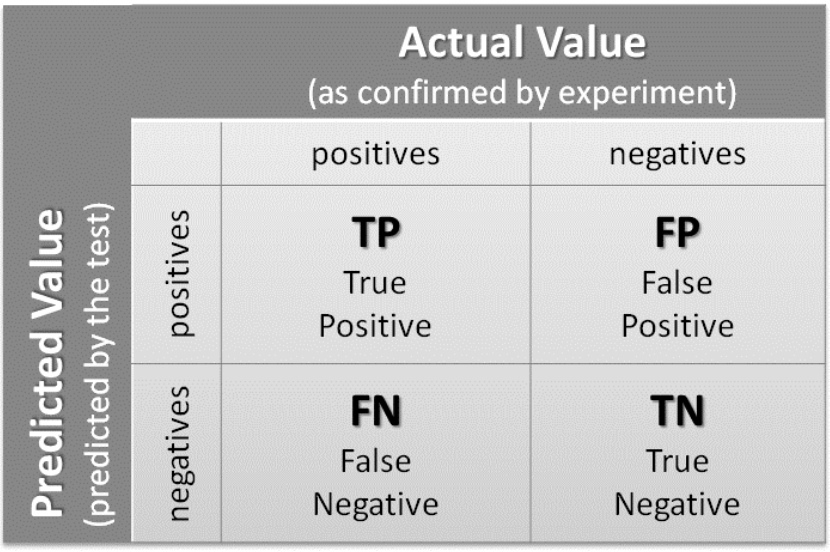

Figure 4. Matriz de confusión. Fuente: (Manthiri, 2017).

Las entradas en la matriz de confusión tienen el siguiente significado en el contexto del estudio:

- TP es el número de predicciones correctas que una instancia es positiva.
- FN es el número de predicciones correctas de que una instancia es negativa.
- FP es el número de predicciones incorrectas de que una instancia es positiva.
- TN es el número de predicciones incorrectas de que una instancia es negativa.

Para aceptar el modelo de clasificación, esta matriz debe generar una precisión considerablemente alta. La precisión (AC) es la proporción del número total de predicciones correctas. Se determina usando la ecuación:

$$AC = \frac{TP + TN}{TP + FN + FP + TN}$$

5 Datos utilizados

Los conjuntos de datos iniciales, de que parte el estudio, se encuentran almacenado en un fichero de texto .csv. En el subapartado 5.1. se detalla cómo está compuesta la base de datos de los resultados electorales del 26J y en el subapartado 5.2. la de los tweets sobre la crisis independista catalana.

5.1 Base de datos: Elecciones 26J

La base de datos de los resultados de las elecciones al Congreso (junio 2016), detallada por municipios, se ha obtenido del área de descargas de la página web de la subsecretaria de dirección general de política interior del Ministerio del Interior (http://www.infoelectoral.mir.es/inicio).

La base de datos está compuesta por el registro de los 8.125 municipios que forman el Estado español y por 64 columnas, de las cuales 51 columnas recogen los votos de cada partido político presentado a las elecciones[1].

Para el análisis se ha realizado una serie de transformaciones de los datos para facilitar el estudio, sin que se pierda información. Se han ordenado los 8.125 registros de los municipios de menor a mayor según su población y se han agrupados en una columna denominada “Partidos minoritarios” los 47 partidos que han obtenido menos representación en el Parlamento. De esta forma se dispone de datos para cuatro partidos políticos (PP, PSOE, Unidos Podemos, Ciudadanos) más la suma de los parditos minoritarios.

Estos cuatro partidos políticos (PP, PSOE, Unidos Podemos, Ciudadanos) son los que han obtenido mayores escaños en el Parlamento. Mientras que los otros cinco partidos políticos (ERC-CAT SÍ, CDC, EAJ-PNV, EH BILDU, CCA-PNC) que han obtenido representación en el mismo, se agrupan junto al resto de partidos en “Partidos Minoritarios”, al tener muy pocos escaños. Toda esta información de los comicios del 26J se puede comprobar en la Figura 5.

[1] Cabe señalar que de los 8.125 municipios solo se analizan los que son considerados como ciudades (>10.000 habitantes). Por lo que el análisis se compondrá de 751 registros.

Figure 5. Resultados Elecciones Generales 2016 (Congreso). Fuente: RTVE.

Es curioso destacar, que esos cinco partidos minoritarios en el Parlamento (ERC-CAT SÍ, CDC, EAJ-PNV, EH BILDU, CCA-PNC) son todos partidos políticos de ideología nacionalistas. Nacionalistas catalanes, nacionalistas vascos y nacionalistas canarios.

Al mismo tiempo, se dispone de la información del número total de abstención, de los votos en blanco y de los nulos de cada municipio. Todos estos datos, junto a los de los partidos políticos, se nivelan según el número total de votos y el total del censo electoral (este último solo se utiliza para la abstención). Las columnas del total del censo electoral y del total de los votos permanecerán con sus datos íntegros en números enteros. Estas dos últimas columnas se utilizan como variables de control, ya que al ser las únicas que no tienen valores porcentuales, se englobarán en un solo factor. Si no estuvieran en un mismo factor el análisis no sería muy preciso.

5.2 Tweets sobre crisis independista catalana

La base de datos de los tweets utilizada para el análisis de la crisis independista catalana se ha obtenido de la plataforma para el modelado predictivo y para competencias de análisis, Kaggle (https://www.kaggle.com). La base de datos ha sido aportada por José Berengueres, doctor en robots de inspiración biológica por el Instituto de Tecnología de Tokio.

Esta base de datos en cuestión está compuesta por 129.093 registros que recoge tweets de carácter general de diferentes medios de comunicación y personalidades influyentes tanto del panorama catalán, como del resto de España e internacional. Se dispone de 44 cuentas de usuarios en total.

Así se podrá analizar los sesgos en la información de las noticias sobre el conflicto catalán, según el origen de los usuarios estudiados[2].

El espacio temporal que compone los tweets es desde el 27 de marzo del 2009 hasta el 27 de octubre del 2017. Sin embargo, el espacio temporal que se analiza es desde el 1 de septiembre de 2017 hasta el 27 de octubre de ese mismo año. La elección de este espacio temporal se debe a que los meses de septiembre y octubre han sido lo más intensos del nacionalismo catalán y por lo tanto los de mayor interés. Así mismo, se pretende que las noticias sobre el atentado terrorista de Cataluña (17-18 de agosto 2017) no afecten al análisis.

Una vez establecido el espacio temporal se seleccionan los tweets que traten el tema de la independencia catalana. Por lo que la base de datos a analizar cuenta finalmente con un total de 19.928 registros. En la Figura 6 se aprecia la dimensión de la base de datos que se ha seleccionado. Se divide en tres subgrupos según la nacionalidad de los usuarios. El dataset "0" representa la base de datos general sin ningún filtro. El dataset "09-10/2017" recoge los tweets de septiembre y octubre del 2017 y el dataset "Catalonia" los tweets que se refieren al proceso independista catalán.

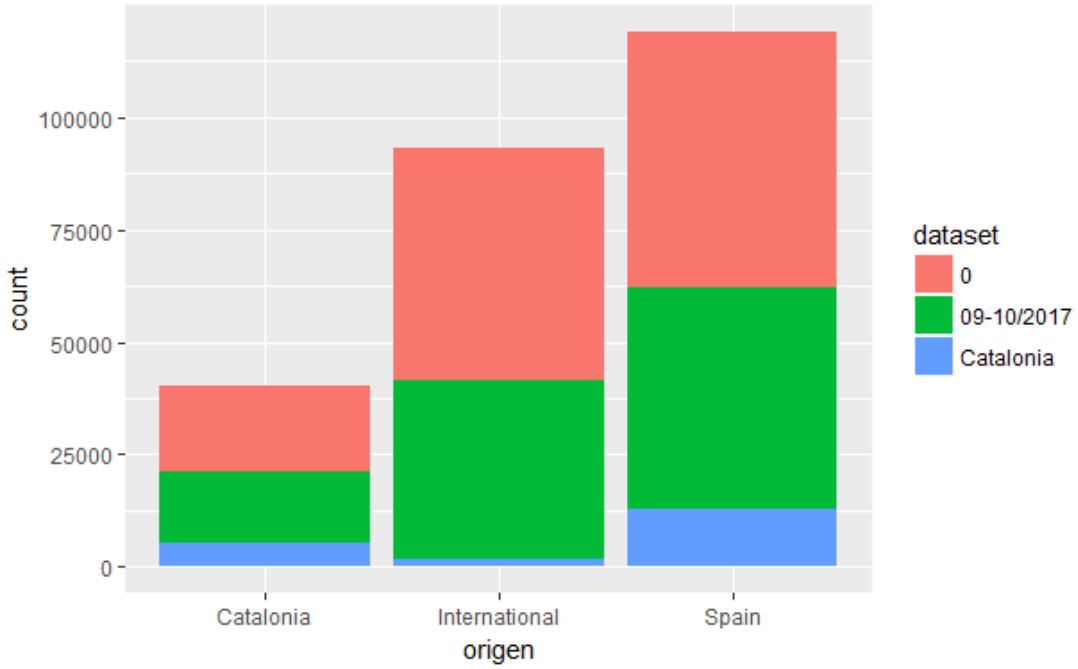

Figure 6. Tamaño base de datos tweets

[2] Los tweets han sido agrupados en tres clases, según el origen de los usuarios. Estas clases pueden ser tanto España, Cataluña o Internacional. De esta forma se conoce las etiquetas de cada registro. Ver Apéndice.

Los tweets sobre la independencia catalana forman un conjunto de tweets bastante considerable respecto al total: 12.826 tweets de 19 usuarios españoles (excepto los catalanes), 5.228 tweets de 7 usuarios catalanes y 1.874 tweets de 18 usuarios internacionales[3].

6 Resultados

A continuación, se presentan los resultados obtenidos a partir de las bases de datos definidas en el apartado 5 "Datos utilizados", tras aplicar las técnicas detalladas en el aparato 4 "Metodología". El análisis completo ha sido realizado con el software estadístico R. Se comienza con los resultados de las elecciones del 26J y se concluye con los tweets sobre el independentismo catalán.

6.1 Elecciones 26J

En primer lugar, se realizaron los test necesarios para comprobar si se cumple uno de los requisitos necesarios para realizar el análisis factorial, que las variables seleccionadas se encuentren altamente intercorrelacionadas. Según el test de Bartlett, se obtiene que las variables están altamente interrelacionadas, ya que se obtiene un p-valor bajo (<0.05), como se muestra en la Tabla 1. Por lo tanto, se cumple la condición necesaria para realizar el A.F.

Table 1. Test de Bartlett

```
$chisq
[1] 21002

$p.value
[1] 0

$df
[1] 45
```

Una vez que se ha determinado que el A.F. es una técnica apropiada para analizar los datos, se debe determinar qué número de factores es el adecuado. Mediante un análisis de paralelos, cuyo resultado se presenta en la Figura 7, se obtiene que son cuatro los factores a explicar.

[3] Los tweets están escritos en tres idiomas: castellano, catalán e inglés. Por lo que se precisa realizar el análisis en los diferentes idiomas.

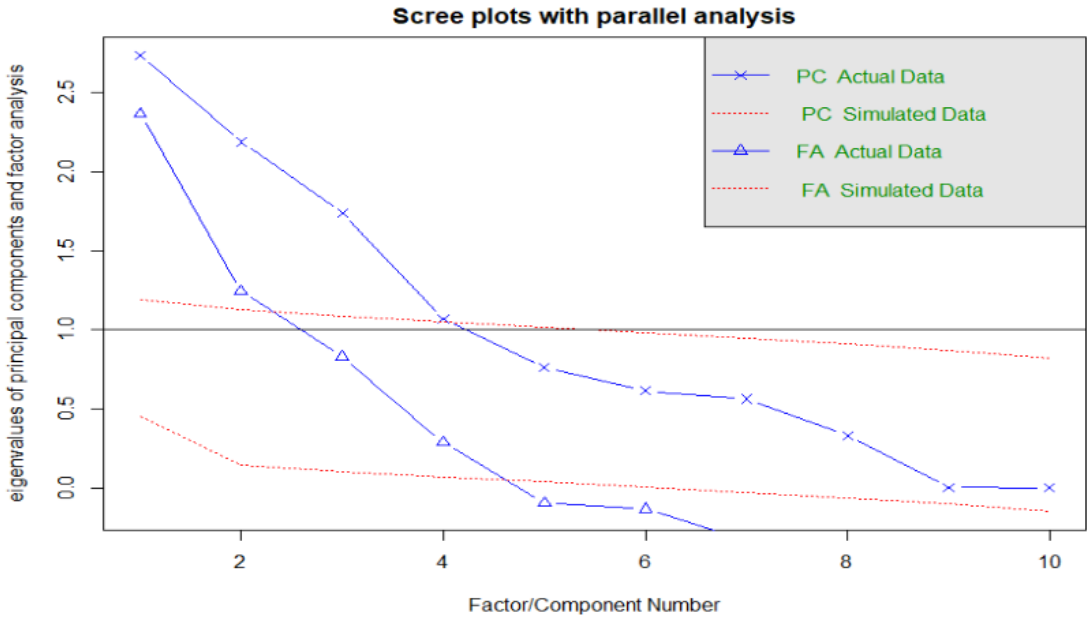

Figure 7. Análisis de paralelos.

A partir de la información que se ha obtenido se genera el A.F., que devuelve la siguiente solución factorial no rotada, por ejes principales, y con cuatro factores (Tabla 2).

Table 2. Solución factorial.

```
Standardized loadings (pattern matrix) based upon correlation matrix
                        PA1   PA2   PA3   PA4   h2      u2 com
Total.censo.electoral  0.07  0.97  0.22  0.07 1.00 -0.0014 1.1
Total.votantes         0.08  0.97  0.21  0.08 1.00  0.0033 1.1
abscención_ptge       -0.28 -0.15  0.48 -0.08 0.33  0.6673 1.9
blanco_ptge           -0.02 -0.18  0.53  0.16 0.34  0.6597 1.4
nulos_ptge             0.40 -0.29  0.38  0.45 0.60  0.3999 3.7
pp_ptge                0.73  0.02 -0.31  0.28 0.71  0.2910 1.7
psoe_ptge              0.50 -0.17  0.33  0.10 0.39  0.6084 2.1
podemos.iu_ptge        0.35 -0.01  0.32 -0.53 0.51  0.4899 2.5
cs_ptge                0.41  0.15 -0.45  0.01 0.39  0.6095 2.2
minorit_ptge          -1.14  0.05 -0.08  0.25 1.37 -0.3654 1.1

                       PA1  PA2  PA3  PA4
SS loadings           2.62 2.08 1.26 0.68
Proportion Var        0.26 0.21 0.13 0.07
Cumulative Var        0.26 0.47 0.60 0.66
Proportion Explained  0.39 0.31 0.19 0.10
Cumulative Proportion 0.39 0.71 0.90 1.00

Mean item complexity =  1.9
Test of the hypothesis that 4 factors are sufficient.

The degrees of freedom for the null model are  45  and the objective function was  28.2
The degrees of freedom for the model are 11  and the objective function was  17.2

The root mean square of the residuals (RMSR) is  0.03
The df corrected root mean square of the residuals is  0.06

Fit based upon off diagonal values = 0.99
```

En la solución se aprecia que se ha obtenido una aceptable varianza (media h2) que logra explicar el modelo, incluso con una alta cantidad de factores. Se obtiene que con 4 factores (de 10 variables originalmente) se logra explicar un 66,4% de la varianza. El modelo a este nivel es aceptable en su capacidad explicativa.

La agrupación de los 4 factores queda representada en la Figura 8, de donde se pueden sacar conclusiones de gran interés. Es por ello que no será necesario realizar el análisis de factores con rotación ortogonal ni oblicua.

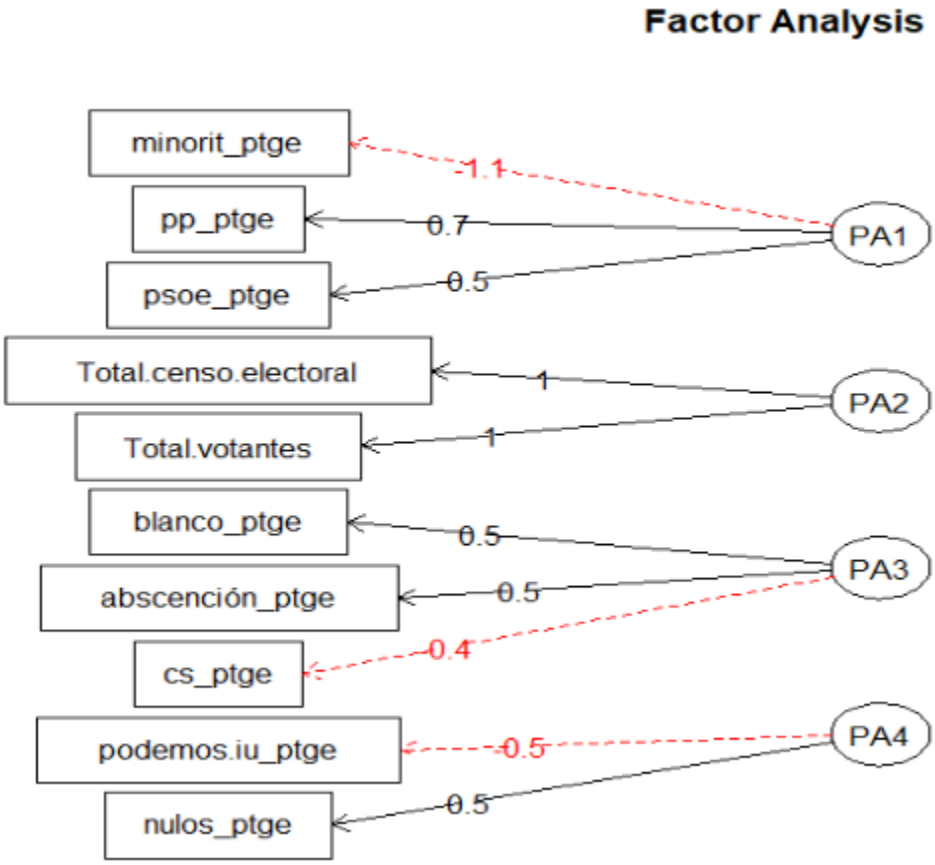

Figure 8. Análisis Factorial.

Del análisis factorial realizado se interpreta que las 10 variables están repartidas en cuatro grupos:

PA1: el primer factor está compuesto por tres variables (minorit_ptge, pp_ptge, psoe_ptge) que recoge el porcentaje de votos a los partidos minoritarios, al Partido Popular y al Partido Socialista Obrero Español.

Este grupo representa a los partidos clásicos y a los partidos regionalistas, principalmente. Por lo tanto, este factor se trata de la **población tradicional**.

Es interesante recalcar que cuando las votaciones a los partidos mayoritarios PP o PSE aumentan, se reducen los votos a los partidos minoritarios, los cuales son los partidos regionalistas, nacionalistas o independentistas. Y viceversa, cuando los otros partidos obtienen mejoras en sus resultados, los grandes partidos, PP y PSOE, empeoran. Esto se explica por el comportamiento de la población tradicional, la cual solo contempla entre sus opciones votar a los partidos PP o PSOE, o dirigir su voto a los partidos regionalistas de su comunidad. Pero apenas se da el caso de que una persona que vote al PSOE, en otras elecciones vote al PP o viceversa, suelen barajar otras opciones, como la que es votar a un partido de su región. Un claro ejemplo es las Islas Canarias o País Vasco, con Coalición Canaria y el Partido Nacionalista Vasco (PNV). En esas comunidades autónomas, la población clásica suele votar a uno de los partidos del bipartidismo o, como segunda opción, a sus respectivos partidos nacionalistas.

PA2: el segundo factor engloba dos variables (Total.censo.electoral, Total.votantes). Las cuales se tratan de número enteros donde recoge el número total de habitantes de una ciudad que pueden votar y el total de votos. Por lo tanto, este factor se trata del censo de población. Pero como se comentó, este factor solo sirve como control, para comprobar si realmente se agrupan en un mismo factor. Así que una vez comprobada dicha agrupación, no será relevante para el análisis.

PA3: el tercer factor agrupa a tres variables (blanco_ptge, abstencion_ptge, cs_ptge) que representan el porcentaje de los votos en blanco, la abstención y el partido de Ciudadanos. Se puede definir este factor como la **población descontenta**.

Esta población es la población tradicional que está cansada del gobierno de los partidos principales. Es por esa razón que sus opciones son no ir a participar en el proceso democrático, por lo cual se abstienen de votar, o ejercen su derecho al voto, pero lo realizan en blanco, ya que ningún de los partidos le convence. Y otra de las opciones es votar al partido político Ciudadanos, el cual se presenta como un partido de ideología similar a los partidos partícipes del bipartidismo, pero con aire fresco de renovación del panorama político.

Además, es necesario señalar que cuando las abstenciones o los votos en blanco disminuyen, aumentan los votos a Ciudadanos y viceversa. Los que se abstienen si deciden votar, votarán a

Ciudadanos, pero no en blanco, ya que para eso prefieren abstenerse, para que así los votos en blanco no beneficien a los grandes partidos. En cambio, los que votan en blanco, si deciden votar a un partido, será mayoritariamente a Ciudadanos, pero no abstenerse, ya que entienden que votar es una obligación y un derecho que tenemos. Por último, los que votan a Ciudadanos, pero prefieren contemplar otra opción, se inclinarán entre abstenerse o votar en blanco.

PA4: el cuarto y último factor agrupa dos variables (podemos.iu_ptge, nulos_ptge) que representan el porcentaje de los votos al partido de Unidos Podemos y los votos nulos.

Un voto nulo es un voto mal realizado, de forma que conlleve su nulidad. Puede considerarse que un voto es nulo cuando el sobre contiene más de una papeleta de dos candidaturas, cuando la papeleta está marcada más de una vez o cuando en el sobre hay algún objeto o frase que no debería estar ahí. El voto nulo es considerado como un voto travieso y casi siempre es voluntario.

Este factor engloba a la **población indignada**. La cual está cansada de la gestión política de los grandes partidos y apoyan por una alternativa diferente en la cual se reconstruya y renueve el gobierno.

También es curioso destacar que la población indignada solo se mueve entre dos opciones; votar en nulo o votar a Unidos Podemos. Por lo tanto, cuando los votos nulos disminuyen, los resultados de Podemos mejoran, y viceversa. Esto se explica, porque este grupo de personas está muy concienciado con las dificultades históricas que ha soportado España para llegar a disponer de elecciones democráticas y está muy comprometido con sus causas. Es por eso que mayoritariamente deciden ir a ejercer su derecho a voto, aunque vote en nulo o al partido político Unidos Podemos.

Una vez identificados e interpretados los factores, se procede a realizar una agrupación de dichos factores mediante el algoritmo denominado K-means clustering, basándose en el A.F. realizado.

Para la elección del número de clústeres existentes se representa la suma de cuadrados intragrupo (Figura 9), donde se observa que a partir de tres clústeres no se produce ninguna mejora significativa en la minimización de la suma de cuadrados intragrupo. Es decir, que con un clúster más (en este caso cuatro) no se obtiene una variabilidad considerable con respecto a una unidad menos de clúster (tres en este caso). A parte de precisión se tiene que tener en cuenta la simplicidad.

De esta manera se selecciona el número de clúster inferior que reduzca considerablemente la suma de cuadrados intragrupo. Por lo tanto, el número de clústeres óptimo para esta base de datos será de tres clústeres.

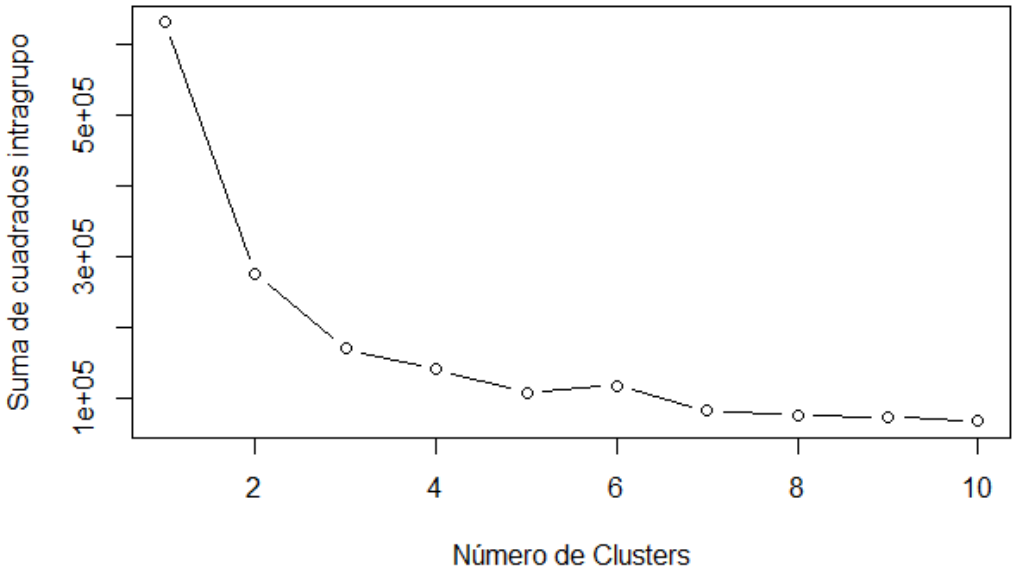

Figure 9. Suma de cuadrados intragrupos.

Al aplicar el algoritmo K-means con tres clústeres a todas las variables se obtienen la Figura 10.

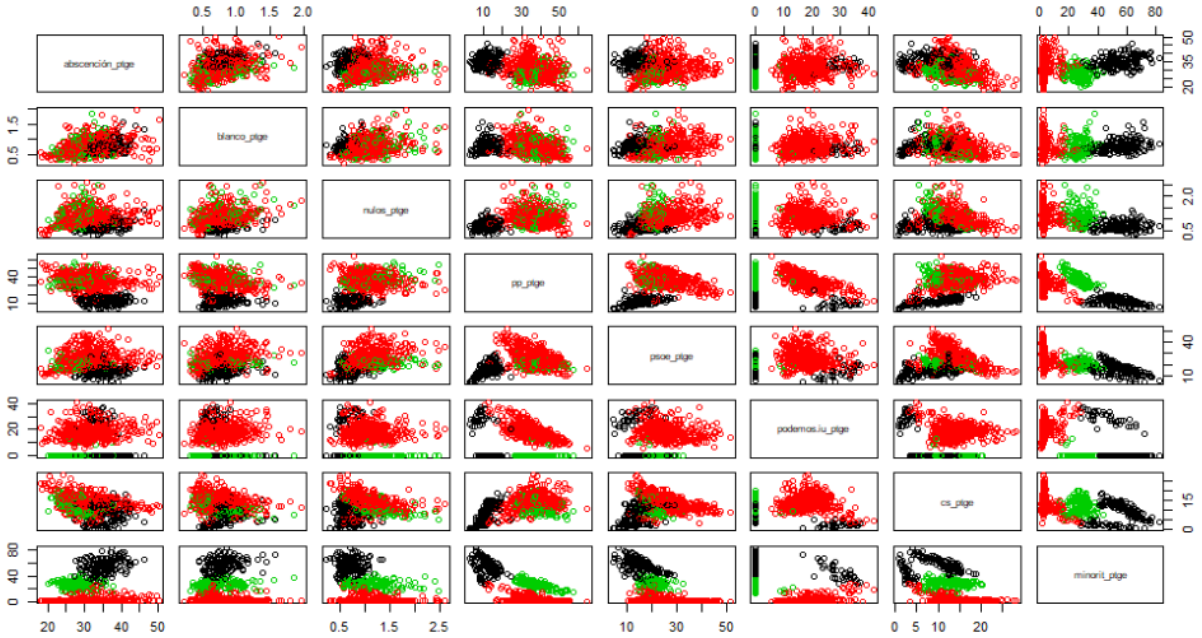

Figure 10. K-mean clustering.

Para poder ver el resultado de manera más clara, la Figura 11 representa la gráfica correspondiente a uno de los pares de variables, PSOE y PP, por ser donde mejor se aprecian los tres grupos de población obtenidos a partir del análisis factorial y componente principales.

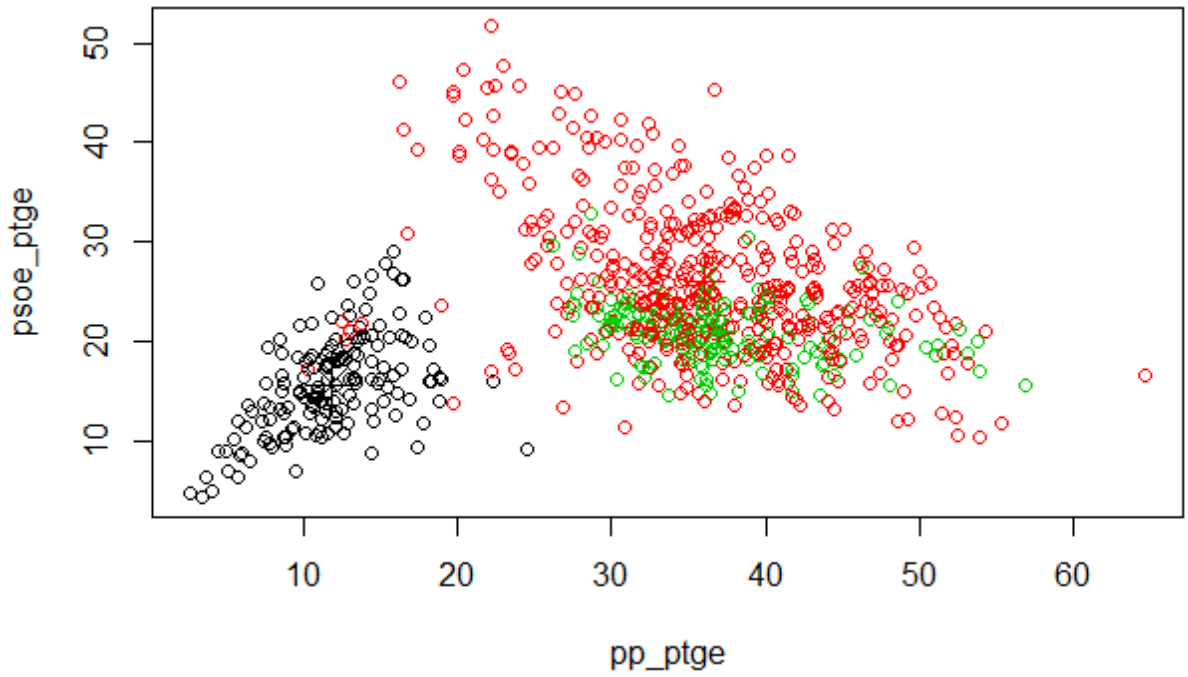

Figure 11. K-mean clustering PSOE-PP.

Se selecciona este par de variables, psoe_ptge y pp_ptge, ya que son los dos partidos políticos que obtienen representación alta en todas las ciudades. Ya que, si se toman como referencia partidos con porcentaje de votos muy bajos en varias ciudades, no se podría explicar de manera clara los diferentes grupos de población visibles, al tener muchos de sus puntos cerca del eje de coordenadas en sus combinaciones de gráficas.

En la representación gráfica de la Figura 11 se observan tres grupos de datos: el color negro representa a la población indignada, que son los que menos votan a los partidos PSOE y PP. El color rojo se refiere a la población tradicional que es la que más vota a ambos partidos[4]. Dentro de esta nube de puntos también se diferencia un pequeño grupo, representado en color verde, que

[4] Cabe señalar que esta nube de puntos es algo más dispersa que el resto debido a que recoge también los votos a los partidos minoritarios (nacionalistas, regionalistas).

hace referencia a la población descontenta. Ya que, realizando una comparación con la población indignada, es la población descontenta la que más vota al PP y al PSOE, pero la que menos en comparación con la población tradicional. Resulta curioso que esa pequeña nube de puntos se acerque más a la ideología de derechas que a la de izquierdas. Esto explicaría por qué la mayoría de los votantes del partido Ciudadanos se recogen en ese grupo, dada su orientación política más cercana a derechas que izquierdas.

De esta manera, en base al análisis llevado a cabo se puede afirmar que la sociedad española se divide en tres grupos: población tradicional (PA1), población descontenta (PA3) y población indignada (PA4), recogiendo 55,76%, 34,21% y 10,03% del total de la población, respectivamente.

6.2 Tweets sobre la independencia catalana

En el análisis de los tweets sobre la independencia catalana, tras limpiar la base datos de registros y caracteres extraños, se ha obtenido cierta información relevante que se detalla a continuación.

Es conveniente precisar que en este análisis se da por supuesto que la publicación de tweets y el interés en el suceso tienen una relación directa. Asimismo, el interés y la tensión del suceso también tienen una relación directa. Por lo tanto, a mayor publicación de tweets, mayor es la tensión del momento.

En el espacio temporal establecido, del 1 de septiembre al 27 de octubre de 2017, se define un crecimiento notable de los tweets, especialmente en días puntuales. En la Figura 12, se aprecia que los días más tensos, donde más tweets se publicaron, fueron los días: 1 de octubre (referéndum de independencia), 10 de octubre (declaración de independencia suspendida), 21 de octubre (Rajoy propone al Senado la aplicación del artículo 155 de la Constitución Española) y 27 de octubre (Proclamación de la República Catalana y paralelamente la aplicación del artículo 155).

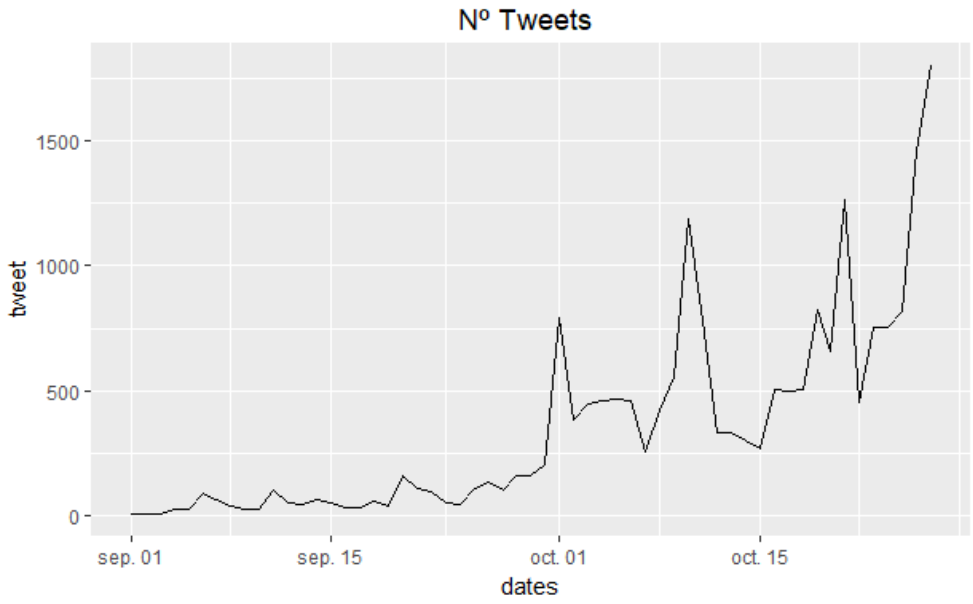

Figure 12. Evolución temporal tweets.

En ese mismo espacio temporal las publicaciones de los tres grupos de usuarios (españoles, catalanes e internacionales) siguen una tendencia similar, como se aprecia en la Figura 13. Llama la atención que el número de tweets de los medios catalanes tiende a ser superior al de los internacionales. Sin embargo, los días previos y el día del referéndum del 1 de octubre, los tweets internacionales superan a los catalanes. Esto demuestra el eco y el impacto que causó en el panorama internacional la jornada electoral de ese día. Siendo el 1 de octubre el día que los usuarios internacionales publicaron más tweets sobre este asunto. Seguido por el 27 de octubre.

En cambio, los tweets catalanes se asemejan mucho en número de tweets publicados a los internacionales, hasta que los días previos a la proclamación de la República Catalana y paralelamente, la aplicación del artículo 155 (27 de octubre), los tweets catalanes se despuntan siguiendo muy de cerca al número de tweets publicados por el resto de usuarios españoles. A pesar de que el número de los usuarios españoles analizado es mucho mayor que el número de usuarios catalanes analizados. 12 usuarios de diferencias. Esto demuestra el gran interés que suscitó por parte de los usuarios catalanes la posible proclamación de la República Catalana y la aplicación del artículo 155.

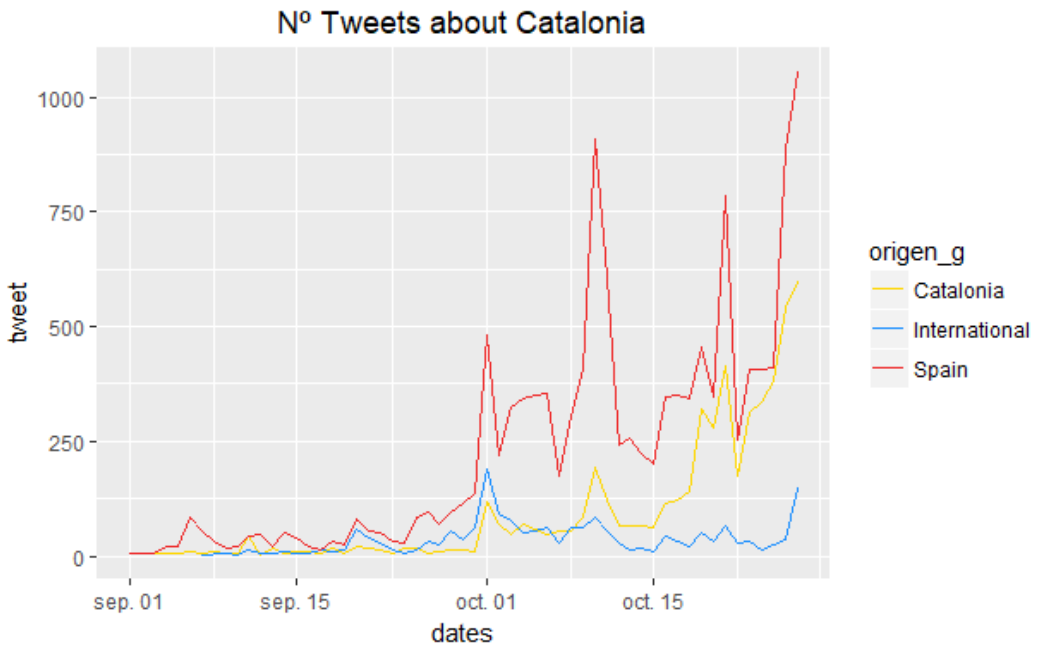

Figure 13. Evolución temporal tweets según origen.

En cuanto al análisis de frecuencia de las palabras empleadas, se obtiene que los tres grupos analizados utilizan un vocabulario similar entre sí. Muchas palabras se repiten, aunque estén en idiomas diferentes, como se puede apreciar en las siguientes tres figuras, donde se representan las 10 palabras más usadas por cada grupo. Se mostrarán las frecuencias de uso, en vez del número, para hacer un análisis estandarizado.

En la Figura 14 se visualiza que los usuarios españoles han utilizado sobre todo los términos siguientes (en un orden descendente): Cataluña, independencia, directo, gobierno, declarar, aplicar, España, video, elecciones y articulo.

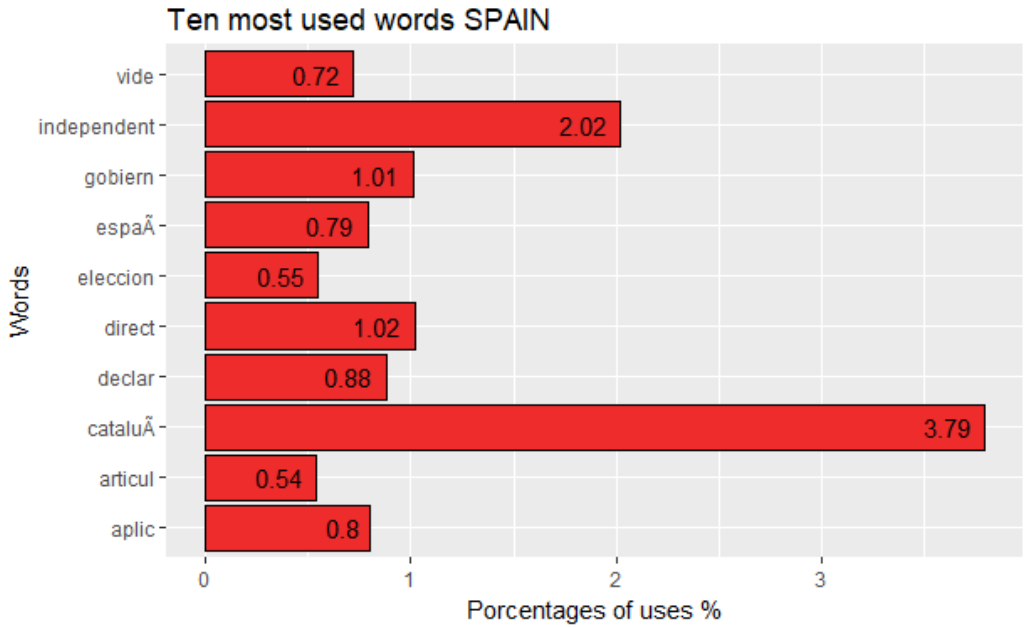

Figure 14. 10 palabras más usadas en España. Fuente: Elaboración propia.

En la Figura 15 se visualiza que los usuarios catalanes han utilizado sobre todo los términos siguientes (en un orden descendente): directe, Rajoy, NCI, govern, espanyol, DUI, aplicar, articulo, independència y convocar.

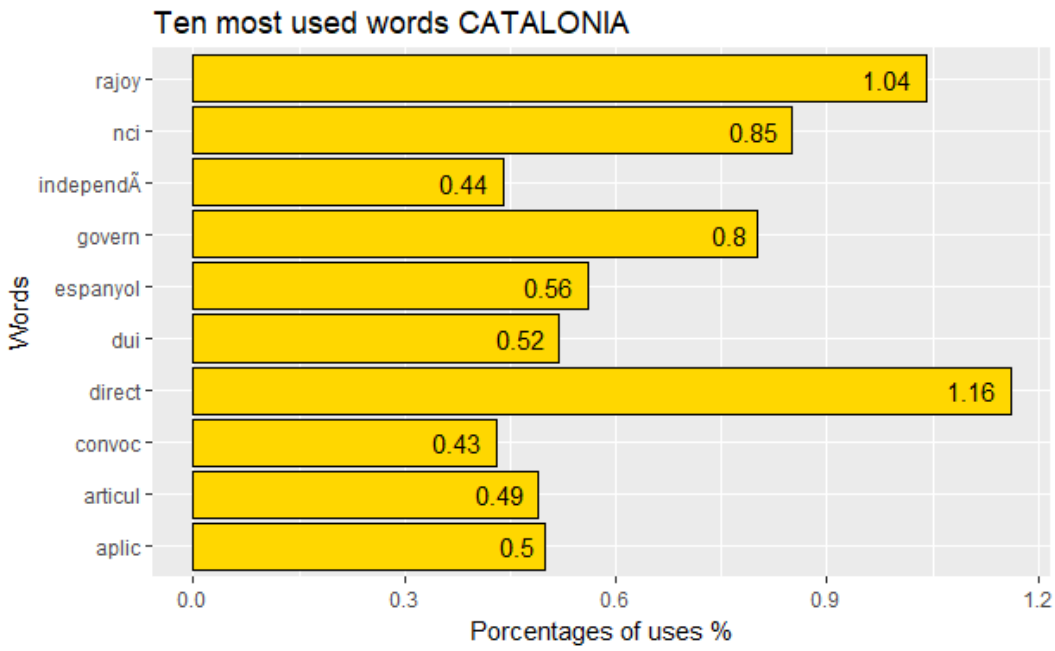

Figure 15. 10 palabras más usadas en Cataluña. Fuente: Elaboración propia.

En la Figura 16 se visualiza que los usuarios internacionales han utilizado sobre todo los términos siguientes (en un orden descendente): vote, referéndum, police, Cataluña, government, declare, lead, Rajoy, España e independent.

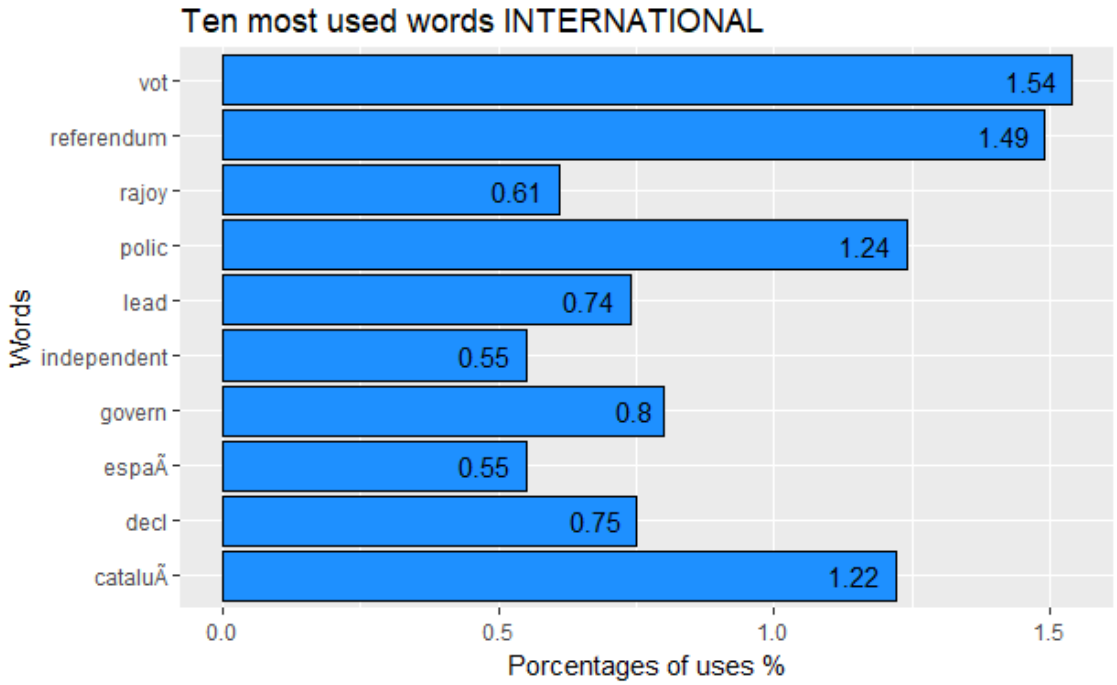

Figure 16. 10 palabras más usadas en Internacional. Fuente: Elaboración propia.

Los resultados muestran el vocabulario que se esperaba que fuera el más utilizado. Aunque cabe destacar el gran uso del término “directo”, tanto por parte de los catalanes y de los españoles. Lo que significa que ambos medios estaban constantemente informando de los sucesos al instante que ocurrían. También es destacable, por parte de los españoles, la aparición en este análisis de la palabra “video”. Lo que significa que gran parte de la información contiene un video para reforzar el mensaje.

En cuanto, a las palabras más utilizada por parte de los medios internacionales, que es el análisis que más interesa, se aprecia que el vocabulario más utilizado (vote, referéndum) se refiere a la jornada electoral que se celebró el 1 de octubre. Aunque es interesante destacar que también se hizo mucho eco de la presencia policial en ese día especialmente.

Continuando con el análisis de la frecuencia de palabras, se comprueba que, en los tweets, tanto catalanes, internacionales como españoles, se cumple la Ley de Zipf. Es decir, que se cumple que

la segunda palabra más usada en cada categoría aparece la mitad de veces que la palabra más usada, la tercera palabra más usada aparece un tercio que la más usada y así consecutivamente.

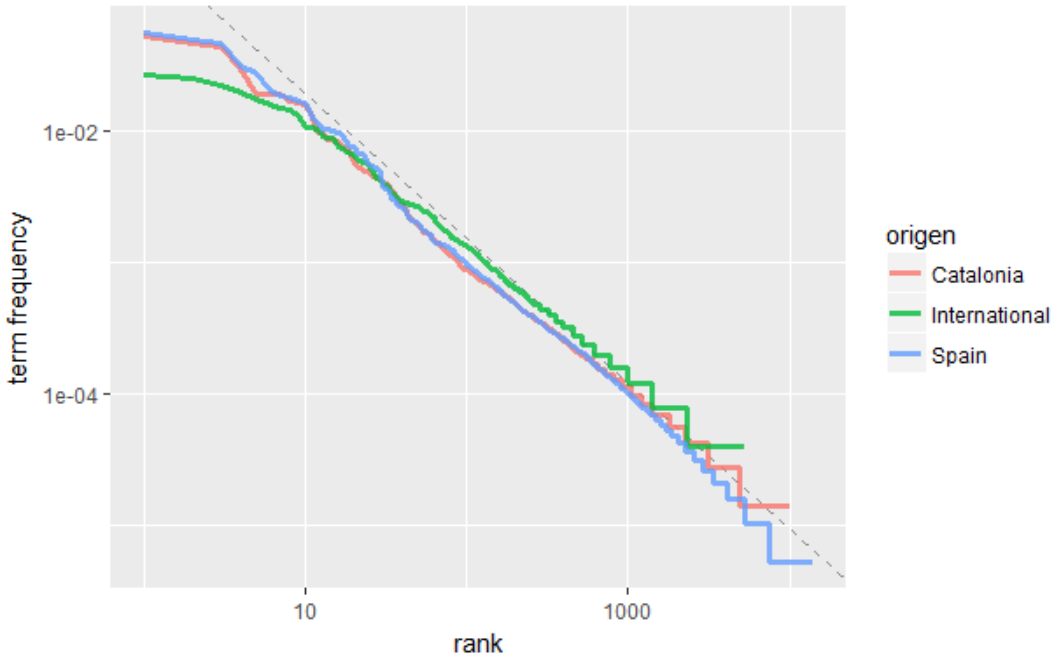

Figure 17 Ley de Zipf (escala logarítmica).

En la Figura 17 se puede ver que los tres grupos son similares entre sí (especialmente Cataluña y España), y que la relación entre el rango (la posición de la palabra en el orden de mayor a menor) y la frecuencia de las palabras tienen pendiente negativa.

Sin embargo, no siguen exactamente a la regresión potencial. Se desvía especialmente en el extremo izquierdo. Aun así, se puede comentar que la recta es bastante aproximada a los datos reales y esta recta se aproxima más a los mismos a medida que aumenta el número de palabras calculadas. Además, al tratarse de una ley aproximada no pretende calcular con exactitud el número de veces que se repetirá una palabra en el texto, pero sí su correlación. Y en este caso su correlación es alta, ya que ambos, la recta y los datos reales, descienden sistemáticamente.

Las desviaciones que vemos aquí en el alto rango no son infrecuentes, en cambio sí sería más inusual si se tuvieran desviaciones en el rango bajo (Zipf, 1949). En estos tweets se usa un porcentaje menor de las palabras más comunes que muchas colecciones de lenguaje.

Sin embargo, este análisis de la frecuencia de palabras no es suficiente para poder sacar conclusiones sobre la información que ha transmitido los medios internacionales sobre la crisis

catalana. Por lo que se realizó un análisis de sentimientos sobre los tweets de los usuarios internacionales.

Antes de nada, se ha decido realizar el análisis de sentimiento mediante el léxico de Bing Liu y sus colaboradores (Hu & Liu, 2004). Descartándose el léxico de Finn Årup Nielsen (AFINN) y el de Saif Mohammad y Peter Turney (NRC). Esta elección se ha tomado debido a que, como se puede apreciar en la Figura 18, que comparando los tres léxicos conjuntamente, el léxico Bing y el AFINN tienen trayectorias relativas similares, mientras que el léxico NRC tiene una trayectoria relativa más dispar. No obstante, se descarta el léxico AFINN, ya que tiene valores absolutos más acentuados, mientas que el léxico Bing sigue unos valores absolutos más condensados. Consiguientemente, con el léxico Bing se obtiene un análisis más homogéneo y no tan de extremos.

Figure 18. Léxicos de sentimientos.

De esta manera, según el léxico Bing, se obtienen las 10 palabras que más contribuyen tanto al sentimiento negativo como positivo, representadas en la Figura 19.

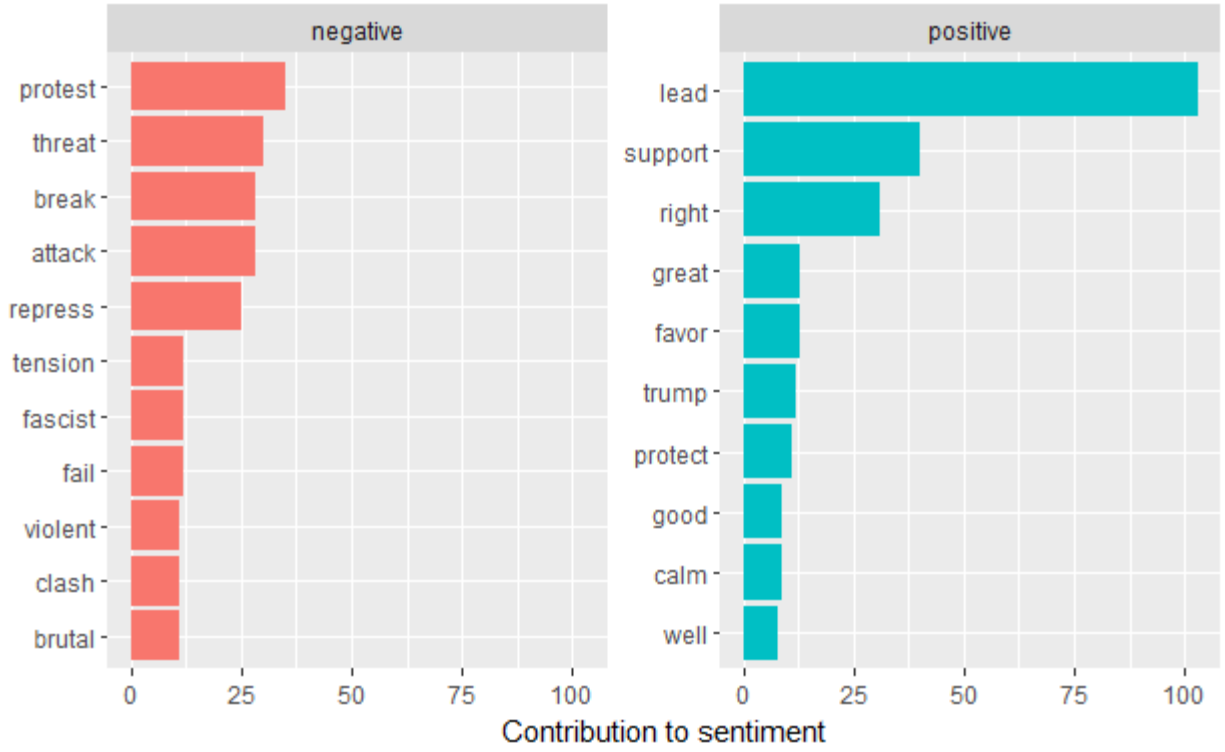

Figure 19. Contribución al sentimiento (individual).

Las palabras que más contribuyen al sentimiento negativo son: protesta, amenaza, romper, ataque, reprimir, tensión, fascista, fallar, violento, conflicto y brutal.

Las palabras que más contribuyen al sentimiento negativo son: protesta, amenaza, romper, ataque, reprimir, tensión, fascista, fallar, violento, conflicto y brutal.

Mientas que las que más contribuyen al sentimiento positivo son: dirigir, apoyo, derecho/justo, grande, aprobación, Trump/triunfo, proteger, bueno, calma y bien[5].

A partir de las palabras negativas se interpreta que se está en una situación extrema donde se impone la fuerza. Mientras que con las palabras positivas se interpreta que toda la situación está controlada en un perfecto orden.

[5] Aclarar que el término “right” se refiere a derecho o a justo, mientras que el término “trump”, se refiera al presidente de los EEUU, Donald Trump, o a la palabra triunfo.

En la Figura 20 presentada a continuación, se puede comparar de manera más sencilla la contribución al sentimiento de las palabras usadas.

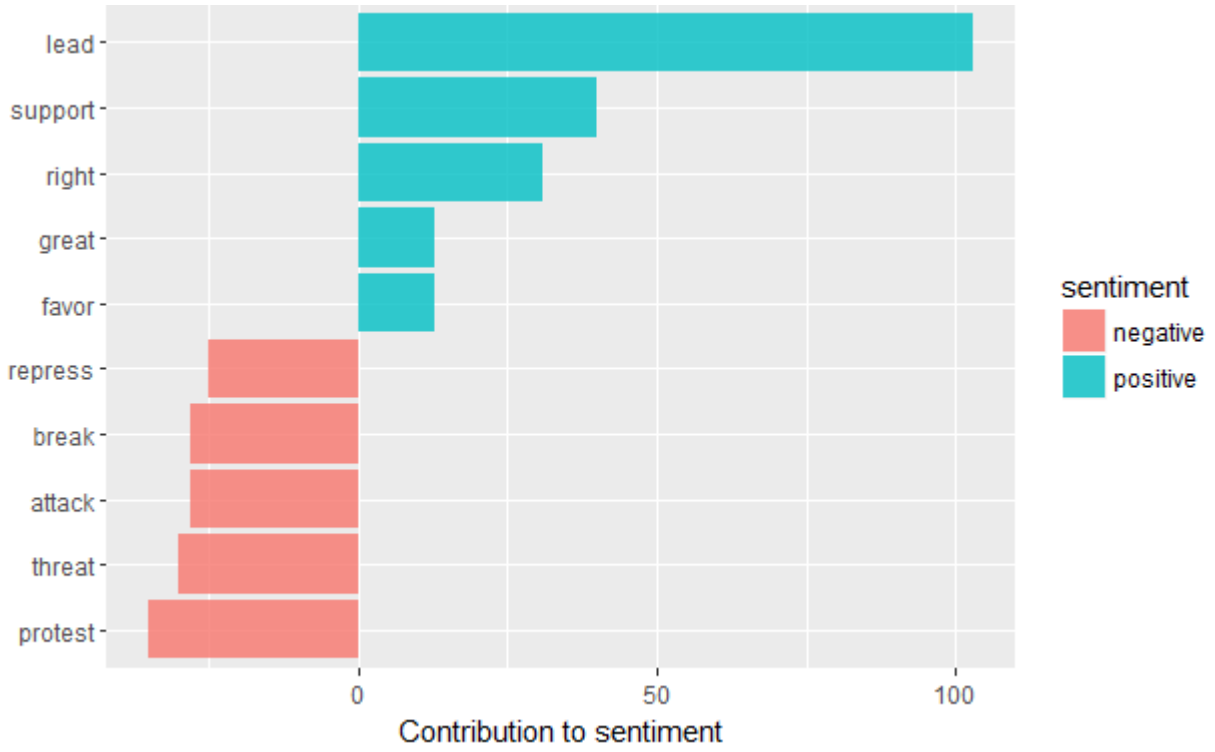

Figure 20. Contribución al sentimiento (conjunto).

El término “lead” tiene un peso considerable. Supone más de la mitad, que el término que le prosigue, “support”. El inconveniente del sentimiento positivo es que solo tres términos, los dos anteriores junto al término “right” son los que más contribuyen al mismo. Ya que los términos que le prosiguen tienen un peso pequeño. En cambio, en el sentimiento negativo, los términos siguen una contribución similar y considerable. Lo que significa que la información que transmiten los tweets de los medios internacionales está cargada de palabras negativas.

Esta afirmación, se aprecia mejor en una nube de palabras (Figura 21). Donde se aprecia que las palabras negativas tienen un peso similar entre ellas, mientras que la positivas tienen pesos más irregulares. Además, aparece un número mayor de palabras negativas que positivas.

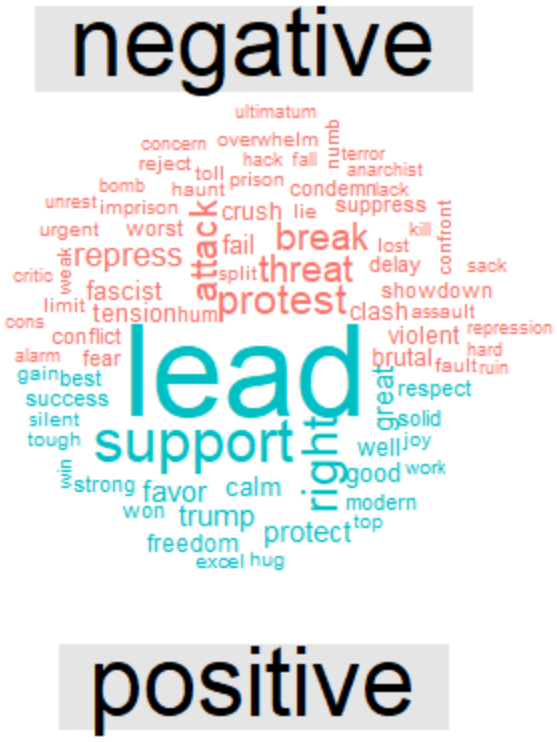

Figure 21. Nube de palabras.

Una vez, realizado el análisis de sentimiento, se decide por llevar a cabo un análisis mediante el aprendizaje supervisado. Se tiene como objetivo comprobar si a partir de la información que existen en los tweets de algunos de sus usuarios, se pueden reconocer las tres clases de usuarios según su origen (catalanes, españoles e internacionales)

Los usuarios que se han seleccionado son todos de habla hispana. Para que la clasificación se realice según el mensaje que transmiten y no por el idioma. Estos usuarios son: eldiario.es, un medio de comunicación digital español de información general con ideología de izquierdas, La Vanguardia, un diario de información general editado en Barcelona con ideología de centro-derecha, y CNN en español es un canal de televisión estadounidense sin una ideología definida dirigido a Latinoamérica, el Caribe y al público hispano en Estados Unidos.

De este modo, para comprobar el reconocimiento de los tweets se ha empleado la técnica de clasificación de los k vecinos más cercanos (KNN). Se pretende predecir la clase a la que pertenecería cada registro con un modelo donde el 80% de los registros de la base de datos corresponde a entrenamiento y el 20% restante de prueba. Obteniendo la siguiente matriz de confusión mostrada en la Tabla 3, con una precisión del 61,55%:

Table 3. Matriz de confusión.

	Actual		
Predictions	Catalonia	International	Spain
Catalonia	165	5	148
International	0	33	17
Spain	15	3	103

[1] 61.55419

Se puede deducir que de los 180 tweets de La Vanguardia (Cataluña) se predicen correctamente 165 tweets, mientras que los 15 restantes como del eldiario.es (España). 41 de los tweets de CNN en español (Internacional) se predicen correctamente 33 tweets, mientras 5 tweets se predicen como de Cataluña y 3 como de España. Por último, 268 de los tweets de eldiario.es se predicen correctamente solo 103 tweets, mientras que 17 se predicen como internacionales y 148 como de Cataluña.

El dato más curioso es que para los tweets de España se predicen bastante mal. Tiene mayores predicciones falsas que verdaderas.

De esta manera, de los 489 tweets solo predice correctamente 301 tweets en su grupo. Es decir, como se ha comentado, se ha acertado un 61,55%. Un porcentaje con el que no se puede afirmar que este modelo realice una predicción suficientemente precisa para confiar en él. Además de que la dimensión de las clases no es muy homogénea y esto puede afectar al análisis.

Por lo tanto, según los tweets de La Vanguardia (Cataluña), eldiario.es (España), CNN en español (Internacional), no se reconoce con gran fiabilidad a que clase pertenecen.

7 Conclusiones y futuras investigaciones

El análisis realizado es realmente útil para reflejar el panorama político actual que vive España. Al mismo tiempo, el presente estudio revela también algunas posibles líneas de investigación en el campo del análisis computacional de gran interés.

El primer análisis, de los resultados de las elecciones generales del 26J, explica la obtención del poder político e identifica cómo se divide la sociedad española según sus votos, a partir de un análisis factorial y posterior elaboración de clústeres. Esta información es de valor para los partidos políticos que, conociendo cómo se divide la sociedad española, podrán diseñar estrategias políticas adecuadas para obtener más votos del resto de la población.

El bipartidismo y los partidos minoritarios, aunque no se encuentran ante un panorama pesimista, deben tener en cuenta que existe una parte considerable de la población que no apoya sus programas, por lo tanto, deberían adoptar medidas distintas para conseguir atraer los votos principalmente de los votantes descontentos, ya que los indignados se les asemejan menos.

En cambio, Ciudadanos, como representante de la población descontenta, y Podemos, de la población indignada, pueden aprovechar esta ramificación de la población para conseguir atraer votos de la población tradicional e indignada y de la población descontenta, respectivamente. También es una buena estrategia dirigir sus esfuerzos de obtención de apoyo a la población que se encuentra en el mismo grupo, pero votan otra opción. Es decir, Ciudadanos deberá conseguir los votos de la población que se abstienen a votar y que votan en blanco. Sin embargo, a Unidos Podemos le será más conveniente conseguir a los votantes que votan en nulo.

Un estudio más exhaustivo, usando datos relacionados con diferentes convocatorias electorales repodría dar resultados de mayor fiabilidad. Se aspira a extender este análisis no sólo a las últimas elecciones, sino a las anteriores convocatorias tanto generales, como autonómicas y municipales. Además de relacionar este estudio con variables económicas.

Por otro lado, en cuanto al análisis de los tweets procedentes de cuentas internacionales sobre la independencia de Cataluña, realizado a partir de técnicas de frecuencias de palabras, análisis de sentimientos y clasificador de los vecinos más cercanos, se puede identificar la gravedad del proceso de independencia catalana. Referirse a esta situación política como una crisis institucional es acertado, debido a la tensión del momento que se refleja en los tweets, a pesar de que se aprecia cierta información que da a entender que esta situación está bajo control.

De todas formas, cabe destacar que el análisis ha sido realizado sin que la crisis haya concluido, por lo que es aún considerado muy prematuro. Hasta el momento de realización se ha podido

estudiar sólo una parte del conflicto. Por ello, se reanudará este análisis una vez que pase un espacio temporal considerable para poder evaluar la situación por completo.

Además, se deberá ampliar el análisis de sentimiento no solo a los tweets internacionales, de lengua inglesa, sino también a todo el ámbito de interés. Por lo tanto, se tienen que analizar los sentimientos de los tweets que estén publicados tanto en castellano como en catalán. El desarrollo de dichos procesos aplicados a los textos en una lengua distinta al inglés, está siendo objeto de creciente interés por parte de los investigadores en comunicación y del sector de la ingeniería computacional.

Referencias

Arcila-Calderon, C., Ortega-Mohedano, F., Jimenez-Amores, J., & Trullenque, S. (2017). *Supervised sentiment analysis of political messages in Spanish: Real-time classification of tweets based on machine learning*. Profesional De La Información, 26(5), 973-982.

Barranco, A. P. (2017). Albert CARRERAS y Xavier TAFUNELL (dirs.), *estadísticas históricas de España. siglos XIX-XX*. 2. ª edición, revisada y aumentada. Madrid, fundación BBVA, 3 vols., 2005, 1.435 pp. (incluye CD-ROM). Revista De Historia Industrial. Economía y Empresa, 15(31), 179-184.

Barreda, M. (2011). *La calidad de la democracia: Un análisis comparado de américa latina*. Política y Gobierno, 18(2), 265-295.

Baviera, T. (2017). *Técnicas para el análisis de sentimiento en twitter: Aprendizaje automático supervisado y SentiStrength.* Revista Dígitos, 1(3), 33-50.

Cambronero, C. G., & Moreno, I. G. (2006). *Algoritmos de aprendizaje: Knn & kmeans*. Inteligencia En Redes De Comunicación, Universidad Carlos III De Madrid,

Capel, H. (1975). *La definición de lo urbano*. Estudios Geográficos, 36(138), 265.

Cool Tabs. (2017). *El referéndum del 1-O genera casi 12 millones de tweets*. Retrieved 19/10, 2017, from https://blog.cool-tabs.com/es/referendum-cataluna-1o-en-twitter/

Cover, T., & Hart, P. (1967). *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13(1), 21-27.

De la Fuente Fernández, Santiago. (2011). *Análisis factorial*. Universidad Autónoma De Madrid,

Dogan, M. (1996). *Political science and the other social sciences*. A New Handbook of Political Science, 97-132.

economíaDigital - Wicho. (2016). *El misterio de la ley de zipf y el lenguaje*. Retrieved 09/11, 2017, from http://www.microsiervos.com/archivo/ciencia/misterio-ley-zipf-y-lenguaje.html

Eíto Brun, R., & Senso, J. A. (2004). *Minería textual*. El Profesional De La Información,

Felipe VI. *Mensaje institucional del rey sobre Cataluña*. (2017, 03/09). [Video/DVD]

Fix, E., & Hodges Jr, J. L. (1951). *Discriminatory analysis-nonparametric discrimination: Consistency properties*. *Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties,*

García, F. (2016). *Compacidad y densidad de las ciudades españolas*. EURE (Santiago), 42(127), 5-27.

Giménez, M., Baviera, T., Llorca, G., Gámir, J., Calvo, D., Rosso, P., et al. (2017). *Overview of the 1st classification of Spanish election tweets task at ibereval 2017*. Paper presented at the Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Murcia, Spain, September, 19.

Harguindéguy, J., Rodríguez-López, E., & Sánchez, A. (2017). *Inter-governmental conflicts between Spain and Catalonia.* Revista Española De Investigaciones Sociológicas, 158, 79-96.

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. Paper presented at the Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168-177.

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining* Springer Science & Business Media.

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data* Springer Science & Business Media.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.* Paper presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1. (14) pp. 281-297.

Manthiri, A. (2017). *Multi class confusion matrix*. Retrieved 01/11, 2017, from http://www.expression-templates.org/tutorials/expression-web-4-ver2-2.pdf

Menéndez, M. (26/06/2016). *Elecciones generales 2016 | las claves del 26J, las primeras elecciones repetidas de la historia de España* - RTVE.es. Retrieved 09/20, 2017, from http://www.rtve.es/noticias/20160626/claves-del-26j-primeras-elecciones-repetidas-historia/1363008.shtml

Mitchell, T. M. (1997). *Machine learning*. 1997. Burr Ridge, IL: McGraw Hill, 45(37), 870-877.

Montemurro, M. A. (2001). *Beyond the Zipf–Mandelbrot law in quantitative linguistics*. Physica A: Statistical Mechanics and its Applications, 300(3), 567-578.

Moya Sánchez, M., & Herrera Damas, S. (2015). *Cómo puede contribuir twitter a una comunicación política más avanzada*. Arbor, 191(a257), 774.

Pichel, M. (2017, 04/09). *Por qué el discurso del rey Felipe VI sobre el referéndum por la independencia de Cataluña muestra la gravedad de la crisis que vive España*. BBC Mundo,

Provost, F., & Kohavi, R. (1998). G*uest editors' introduction: On applied research in machine learning.* Machine Learning, 30(2), 127-132.

RAE, R. A. E. (2015). *Corpus de referencia del español actual (CREA) - listado de frecuencia.* Retrieved 09/11, 2017, from http://corpus.rae.es/lfrecuencias.html

Rodríguez, L. A. (2017). *La crisis económica de la unión europea y los movimientos separatistas brexit.* Internaciones, 4(12), 23-39.

RTVE. (2016). *Resultados elecciones generales 2016 al congreso*. Retrieved 27/10, 2017, from http://resultados-elecciones.rtve.es/generales/congreso/

Russel, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach*, 2003. Prentice Hall Pearson Education Inc., Upper Saddle River, New Jersey, 7458, 116-119.

Samuel, A. L. (1959). *Some studies in machine learning using the game of checkers*. IBM Journal of Research and Development, 3(3), 210-229.

Sanders, D. (1981). *Patterns of political instability Macmillan.*

SENDRA, J. (1981). *Geografía electoral y elecciones en España*. Paper presented at the Anales De Geografía De La Universidad Complutense, (1) pp. 285-293.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* " O'Reilly Media, Inc.".

Sinha, P., Srinivas, S., Paul, A., & Chaudhari, G. (2016). *Forecasting 2016 US presidential elections using factor analysis and regression model.*

Spearman, C. (1904*). "General intelligence," objectively determined and measured.* The American Journal of Psychology, 15(2), 201-292.

Thurstone, L. L. (1947). *Multiple factor analysis.*

Tirado, A. (2016). Declive del bipartidismo y emergencia de los nuevos partidos. un análisis del mapa electoral en las capitales de provincia tras las elecciones generales de 2015. Federación Española Sociología. Universidad Pablo De Olavide, GT 42 Grupo de Trabajo para Estudiantes de Sociología, 13.

Turing, A. M. (1950). *Computing machinery and intelligence*. Mind, 59(236), 433-460.

Villafranca, R. R., & Ramajo, L. R. Z. (1980). *GEOGRAFÍA ELECTORAL ESPAÑOLA Una aplicación del análisis factorial de correspondencias de los resultados de las elecciones del 10 de marzo de 1979*. Reis, (9), 139-167.

Virós, R. (1979). *Algunes notes sobre el comportament electoral a Catalunya el 15 de juny de 1977.* Papers: Revista De Sociología, (12), 83-114.

Zipf, G. (1949). *Human behavior and the principle of least effort* Cambridge, MA: Addiston-Wesley.

Apéndice

Users of Twitter profiles		
SPAIN	**CATALONIA**	**INTERNATIONAL**
Agencia EFE Noticias	Carles Puigdemont	Banksy
Antena 3 Noticias	CNI Catalunya	BBC Breaking News
Cadena Ser	El Nacional.cat	Bloomberg
Cinco Días	El Periódico	Breitbart News
COPE	La Vanguardia	Businessweek
Diario ABC	RAC1	CNN
El Confidencial	VilaWeb	CNN en Español
EL MUNDO		Julian Assange
EL MUNDO TODAY		La Sampa in English
El PAÍS		Libération
eldiario.es		Nigel Farage
elEconomista.es		Russia Today (RT)
Ernesto Ekaizer		Sky News
Expansión		The Australian
Federico Jiménez Losantos		The Economist
Julia Otero		The Guardian
La 1 de TVE		The New York Times
La Razón		Washington Post
La Sexta Noticias		