Aplicación de técnicas de minería de textos al *frame analysis*: identificando el encuadre textual de la inmigración en la prensa

Javier Álvarez-Gálvez
Universidad Loyola Andalucía
jalvarez@uloyola.es

Juan F. Plaza
Universidad Loyola Andalucía
jfplaza@uloyola.es

José Antonio Muñiz
Universidad Loyola Andalucía
jamuniz@uloyola.es

Javier Lozano Delmar
Universidad Loyola Andalucía
jlozano@uloyola.es

Recibido: 14 de octubre de 2013
Aceptado: 20 de mayo de 2014

Resumen
El creciente aumento de la información *online* hace patente la necesidad de técnicas para el estudio de grandes volúmenes de datos textuales que actualmente se encuentran disponibles. El presente trabajo examina algunas de las posibilidades de las técnicas automáticas de análisis textual (minería de textos) al campo de las ciencias sociales y, específicamente, al de la comunicación. En concreto, se presenta un caso de estudio que muestra de una manera aplicada cómo estos métodos y técnicas de análisis pueden ser usados para la clasificación automática de contenidos y la descripción-caracterización del *framing* textual de artículos de prensa.
Palabras clave: minería de textos, análisis de contenido, medios de comunicación, framing, inmigración

Application of text mining techniques for frame analysis: identifying the textual framing of immigration in the press

Abstract
The growing availability of online information highlights the need of techniques for the study of large volumes of textual data. This paper examines some possibilities of automatic techniques for textual analysis (text mining) to the field of social sciences and, specifically, to communications studies. In this manuscript, we present a case study to show in an applied manner how these methods and techniques for text analysis can be used for automatic content-characterization and description of textual framing in newspaper articles.
Keywords: text mining, content analysis, mass media, framing, immigration

Referencia normalizada
ÁLVAREZ-GÁLVEZ, Javier; PLAZA, Juan F.; MUÑIZ, José Antonio; y LOZANO DELMAR, Javier (2014): "Aplicación de técnicas de minería de textos al *frame analysis*: identificando el encuadre textual de la inmigración en la prensa". *Estudios sobre el Mensaje Periodístico.* Vol. 20, Núm. 2 (julio-diciembre), págs.: 919-932. Madrid, Servicio de Publicaciones de la Universidad Complutense.

1. Introducción

El creciente aumento de la información *online* hace patente la necesidad de técnicas para el estudio de grandes volúmenes de datos textuales que actualmente se encuentran disponibles. Desde el campo de las ciencias de las computación y la lingüística se están desarrollando técnicas de análisis puramente automáticas (es decir, no supervisadas) que tienen su origen en el campo de la minería de datos y que están encontrando su aplicación en materiales textuales (Álvarez-Gálvez, 2012). Aunque la denominada "minería de textos" nace en los años ochenta, los recientes avances informáticos han posibilitado el vertiginoso crecimiento de esta área de estudio. Las nuevas técnicas posibilitan la explotación analítica de grandes volúmenes de información textual de un modo relativamente rápido y sin necesidad de emplear codificadores humanos.

Hasta el momento la mayor parte de las técnicas puramente automáticas de análisis textual se encuentran enfocadas a las ciencias de la computación y la lingüística. Aunque ya existen algunos ejemplos de la aplicación de estas técnicas en el ámbito de las ciencias sociales (Laver, Benoit y Garry, 2003; Slapin y Proksch, 2008; Hopkins y King, 2010), todavía representan un campo muy reciente para los profesionales del ámbito de la comunicación, la sociología y la ciencia política.

El presente trabajo examina algunas de las posibilidades de las técnicas automáticas de análisis textual al campo de las ciencias sociales y, específicamente, al de la comunicación. En concreto, se presenta un caso de estudio que muestra de una manera aplicada cómo estos métodos y técnicas de análisis pueden ser usados para la clasificación automática de contenidos y la descripción-caracterización del *framing* textual de artículos de prensa.

2. Marco teórico

2.1. «Frame analysis», medios de comunicación y encuadres noticiosos

El denominado «frame analysis» (o análisis del marco) es un método multidisciplinar de investigación que se ha empleado para analizar el modo en el que las personas definimos y comprendemos las situaciones dentro de un determinado contexto, así como las acciones que se despliegan dentro de dicho marco. La noción inicial del concepto de frame analysis es atribuida al trabajo de Gregory Bateson (1955) que introduce el término frame para describir el contexto en el cual las interacciones tienen lugar y que define los límites de la comunicación. Posteriormente, dicho concepto será ampliado por el sociólogo Erving Goffmann (1974): "Frame analysis: An essay on the organization of experience". Ahora bien, este concepto ha sido ampliamente desarrollado en otras disciplinas de las ciencias sociales como, por ejemplo, las ciencias de la comunicación, los estudios de comportamiento político y la teoría de los movimientos sociales.

Goffman parte de la idea de que la comprensión de la realidad implica un proceso de construcción, tipificación y categorización social de las experiencias externas. Comprendemos y organizamos nuestro mundo mediante el continuo empleo de «marcos de referencia primarios» (primary frameworks) que nos permiten definir y comprender las distintas situaciones sociales (Goffman, 1974). Se podría decir que el marco de referencia es el material, creado socialmente, a partir del cual conferimos sentido a nuestro mundo. Es decir, un esquema o marco interpretativo que permite a los individuos "situar, percibir, identificar y etiquetar un número aparentemente infinito de sucesos concretos definidos en sus términos" (1974: 23). Los «marcos de referencia primarios» constituyen un elemento fundamental en la cultura de los individuos, ya que de estos, considerados en su conjunto, emerge la comprensión grupal. Son los marcos de referencia de un grupo los que establecen su sistema de creencias específico, su visión del mundo. De ahí que, generalmente, los individuos muestren una considerable resistencia a modificar sus marcos de referencia. Hacen que la realidad cotidiana resulte comprensible, pero al mismo tiempo es gobernada por ellos. Así, la sociedad define el esquema interpretativo que hace posible la comprensión del curso de la acción social, a la vez que establece un sistema de control social a partir de dicho esquema.

Los medios de comunicación de masas, en cuanto que representan una forma específica de conocimiento de la realidad, asumen un punto de vista concreto a la hora de transmitir la información. En efecto, al igual que los actores sociales en cualquier tipo de interacción cotidiana, adoptan un determinado marco o enfoque para explicar los fenómenos sociales. Por este motivo, los medios de comunicación ofrecen su propia visión (enmarcada) de la realidad. En efecto, es de esta misma idea de marco de referencia de Goffman (1974) de donde parte el concepto de «encuadre noticioso». Un concepto fundamental en la teoría del Framing y, su predecesora, la teoría de la Agenda Setting que se encuentra concretamente referido al encuadre de los medios informativos.

Tankard et al. (1991: 3) definen el encuadre noticioso como: "la idea organizativa central del contenido de las noticias, que proporciona un contexto y presenta el asunto a través del uso de la selección, énfasis, exclusión y elaboración". Siguiendo a Entman (1993: 52) podríamos decir que "encuadrar es seleccionar algunos aspectos de la realidad percibida y hacerlos más destacados en el texto comunicativo, de tal manera que consigan promover una definición del problema particular, una interpretación causal, una evaluación moral y/o una recomendación de tratamiento para el asunto descrito". Así, el encuadre de la realidad que realizan los medios de comunicación actuaría aislando cierto material y centrando la atención sobre el objeto que se pretende representar, acentuando algunos de sus rasgos a la vez que se excluyen o eliminan otros. Igartua y Muñiz (2004) asumen que las noticias son una representación de la realidad concebida por el mismo periodista. Así, es importante tener en cuenta la figura del periodista como productor de conocimiento (Rodrigo Alsina, 2005), incluso, más específicamente, como productor de "esquemas interpretativos" que nos permitirán interpretar la realidad en relación a los elementos representados en el marco o encuadre.

2.2. La necesidad del análisis de contenido para la captación del efecto framing

Como indican D'Adamo, García Beaudoux y Freidenberg (2007: 135), además de indicarnos "acerca de qué pensar" y de aumentar la saliencia o activación de unos temas sobre otros, "los medios de comunicación también nos brindan explicaciones sobre las causas y consecuencias relacionadas con las cuestiones destacadas en sus agendas". La teoría del Framing parte de la premisa de que el modo de encuadrar la información que realizan los medios de comunicación afecta a cómo los mensajes son recibidos por el público. Como se ha puesto de manifiesto en recientes trabajos, el encuadre o enfoque de los medios de comunicación puede generar efectos cognitivos y afectivos sobre las actitudes (Igartua et al., 2005; Reese, 2007; Scheufele y Tewksbury, 2007; Kepplinger, Geiss, y Siebert, 2012). Por esta razón, los investigadores del ámbito de las ciencias sociales han dado generalmente una buena acogida a la metodología de «frame analysis», ya que dicha perspectiva facilita la explicación de los efectos de los medios de comunicación sobre las audiencias.

Dentro del campo específico de los estudios sobre los medios de comunicación de masas, el «frame analysis» análisis se dirige al estudio de los efectos producidos por los diferentes elementos del encuadre noticioso. Es decir, se trata de investigar cómo son encuadrados los diferentes elementos (visuales y/o textuales), así como los efectos que dichos elementos pueden producir sobre las opiniones y percepciones de los receptores de la información (es decir, los consumidores de los diferentes medios de comunicación). El denominado "efecto framing" puede tener un impacto en nuestra cognición sobre los diferentes temas que son tratados en los medios e incluso puede cambiar nuestra predisposición hacia estos. Estudios del ámbito de la política y de la comunicación sugieren que los medios tienen la capacidad de decirle a la gente en qué temas pensar (de Vreese and Boomgaarden, 2003; McCombs, 2004). Por ejemplo, el encuadre noticioso de los medios informativos puede afectar a las opciones políticas de los electores (Iyengar y Kinder, 1987), o incluso sobre el desarrollo de prejuicios y actitudes negativas hacia los inmigrantes (Haider-Markel, Delehanty y Beverlin, 2007; Igartua and Cheng, 2009; Igartua, Moral y Fernández, 2011).

Ahora bien, el estudio del efecto framing no resulta sencillo, ya que son difíciles de identificar tanto los elementos como los procesos subyacentes que dan lugar al desarrollo y/o reproducción de opiniones y/o actitudes. Como paso previo para todo análisis del encuadre noticioso, es necesario describir aquellos elementos que definen y caracterizan el encuadre de las noticias (Cappella and Jameson, 1997), es decir, aquellos elementos que son susceptibles de producir efectos en la percepción de los individuos, independientemente del medio que usen (prensa escrita o digital, radio o televisión). Por este motivo, es habitual que desde este enfoque del ámbito de la comunicación se inicien los estudios empleando técnicas de análisis de contenido. Este tipo de técnicas permiten al investigador realizar inferencias acerca de una población de textos previamente definida, para dar respuesta a "qué temas ocurren", "qué relaciones semánticas existen entre los temas ocurrentes", y "qué posiciones de red son ocupadas por tales temas o relaciones temáticas" en textos procedentes de diversos tipos de fuente, mensaje, canal o audiencia (Roberts, 2000). De ahí que el análisis de contenido sea una técnica comúnmente empleada al inicio de los estudios que se centran en el análisis del efecto framing.

Si bien el análisis de contenido también presenta sus limitaciones. Esta técnica suele ser costosa en términos de tiempo y recursos económicos, puesto que requiere el entrenamiento previo de codificadores humanos que serán los encargados de leer y clasificar/categorizar toda la información. Sin embargo, aunque el uso de codificadores humanos suele aportar mayores niveles de validez a la clasificación/categorización realizada, paralelamente reduce la confiabilidad de la misma (Álvarez-Gálvez, 2012). Esto es, aunque los codificadores humanos tienen una gran capacidad para la interpretación del contenido latente de la información, pueden introducir sesgos o incluso perder información manifiesta que no es visible a primera vista, sobre todo cuando se trabaja con grandes cantidades de información.

Ahora bien, en la actualidad contamos con técnicas automáticas que pueden ser empleadas para la clasificación y categorización de contenidos textuales. Desde la perspectiva de las modernas técnicas de *minería de textos* (text mining), la clasificación y análisis de información textual puede ser tomada como un problema de predicción de resultados a partir de un corpus textual (es decir, como conjunto de elementos alfanuméricos). De ahí que estas técnicas puedan emplear los mismos métodos que empleamos para el análisis de datos numéricos a los textos (Weiss et al., 2005). Si bien podemos apreciar diferencias evidentes entre los datos numéricos y los textuales, esencialmente muestran claras similitudes: (1) son datos que se obtienen de una muestra (con independencia del tipo de distribución que presenten) y (2) que tienen cierta carga simbólico-semántica que puede ser extraída a partir del empleo de técnicas estadísticas (Álvarez-Gálvez, 2012). Los datos textuales no presentan las cualidades aritméticas de los datos numéricos, pero sí son susceptibles de ser manejados a nivel binario –esto es, en términos de presencia o ausencia– a partir de matrices de datos.

Aunque el concepto de minería de datos (data mining) está ampliamente extendido en la actualidad, la *minería de textos* (text mining) aún encuentra una modesta acogida por parte de la comunidad científica de las ciencias sociales. A pesar de todo, en nuestros días ya existen técnicas que permiten tratar la información textual a modo de datos numéricos (Laver, Benoit y Garry, 2003; Hopkins y King, 2010; Slapin y Proksch, 2008). Estos estudios evidencian que las técnicas de minería de textos permiten la organización y clasificación puramente automática de documentos, así como la recuperación y extracción de información específica, su evaluación y predicción. En estas técnicas la intervención humana ha quedado reducida al mínimo, reduciéndose así los problemas generados por la codificación humana de la información textual.

Partiendo de las conocidas ventajas de estas técnicas automáticas de análisis textual, en el presente trabajo se analizan las posibilidades de la minería de textos para la identificación y descripción del *framing* textual de artículos de prensa. En concreto, este estudio se dirige a comprobar la utilidad de las actuales técnicas de minería de textos para la clasificación/categorización automática de aquellas *palabras clave* que caracterizan el framing textual de artículos de prensa, así como sus relaciones de proximidad. A modo de estudio de caso, se emplea un conjunto de editoriales de prensa del diario El País que hacían referencia al tema de la inmigración.

3. Método

3.1. Datos y muestra

Se analizarán 365 editoriales de prensa que durante el periodo 1999-2008 trataban el fenómeno migratorio, procedentes de uno de los principales diarios de tirada nacional: El País. Se optaría por seleccionar artículos editoriales debido a que se pretendía analizar –a modo de estudio de caso– el framing textual que se ofrecía en este periódico concreto. De ahí que se seleccionaran artículos de opinión, que permitían identificar el modo de encuadrar desde una determinada línea editorial.

La razón de optar por este periódico se debió a tres criterios estratégicos fundamentales: (1) es un diario con gran seguimiento a nivel nacional, (2) ofrece la posibilidad de tener acceso a las ediciones impresas a través de Internet para el periodo de estudio prefijado, y finalmente, (3) permite emplear los buscadores a modo de filtro para la extracción de texto a partir de palabras clave. Por otra parte, no se querían establecer diferencias en función de la ideología de los diferentes diarios españoles, ya que al ser un estudio exploratorio se trataba de captar, del modo más específico posible, los principales encuadres de la inmigración que se ofrecen en el marco de este diario. Finalmente, los artículos serían clasificados como "editoriales sobre inmigración" si: (a) pertenecían a la sección editorial del periódico; (b) hacían referencia a los términos inmigración o inmigrante*.

3.2. Metodología estadística

Para la clasificación automática y la posterior identificación de las dimensiones latentes que, mediante la agrupación de diferentes categorías, componen los diferentes frames (o encuadres) se emplearán dos técnicas de análisis. En primer lugar, como técnica de clasificación el análisis cluster jerárquico y, en segundo, el análisis de correspondencias múltiples para la reducción dimensional de la información del conjunto textual analizado.

El análisis cluster jerárquico parte del cálculo inicial de la matriz de distancias de los individuos en la muestra (en este caso, los diferentes términos o categorías textuales). Dicha matriz describe las distancias entre los diferentes elementos de la muestra y parte de aquellos más similares –esto es, los más próximos en términos de distancia– para, progresivamente, ir avanzando en la clasificación de otros elementos 'próximos' que acabarán componiendo el conglomerado final. Estos clusters a su vez irán componiendo otros cluster más amplios y a la vez más heterogéneos, hasta llegar al último paso en el que todos los clusters quedan agrupados en un último cluster global que integra a todos los elementos de la muestra.

En el análisis cluster jerárquico se empleó el método de Ward (1963). Este procedimiento de análisis puede describirse como un método escalonado en el que, en cada nuevo paso, se unen los dos grupos para los cuales el incremento del valor total de la suma de cuadrados de diferencias, de cada individuo al centroide del cluster, sea menor. El descrito modelo es el siguiente:

$$E_k = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k - m_j^k)^2 = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k)^2 - n_k\sum_{j=1}^{n}(m_j^k)^2$$

donde denominamos x_{ij}^{k} al valor de la *j*-ésima variable sobre el *i*-ésimo individuo del *k*-ésimo cluster, suponiendo que dicho cluster posee n_k individuos, m_k al centroide del cluster *k*, con componentes m_j^k. Siendo E_k la suma de cuadrados de los errores del cluster *k*, es decir, la distancia euclídea al cuadrado entre cada individuo del cluster *k* a su centroide.

En segundo lugar, a modo de validación del análisis clasificatorio previo, el análisis de correspondencias múltiples se empleó por su adecuación para la identificación de estructuras latentes a partir de datos categóricos. Así, como para la representación de las dimensiones subyacentes de amplios conjuntos de datos en un espacio de distancias euclídeas. Esta tipología de análisis puede entenderse como una extensión de los análisis de componente principales para variables categóricas. Los análisis se llevaron a cabo mediante el paquete de análisis estadístico R.

4. Resultados

4.1. Análisis descriptivo: identificación de categorías relevantes

A partir del conjunto de documentos analizados se obtuvo una matriz de términos compuestas por un total de 5.221.240 palabras, de las cuales 14.825 eran términos únicos (no repetidos) y 356 documentos. En la Figura 1 se presentan los cincuenta los términos más frecuentes en el conjunto de datos: 'gobierno', 'españa', 'ley', 'política', 'europea', 'países', 'marruecos', 'extranjería', 'situación', 'derechos', 'zapatero', 'problema', 'extranjeros', 'social', 'europa', 'población', 'aznar', 'personas', 'ilegales', 'irregulares', 'millones', 'seguridad', 'fronteras', etc. En definitiva, un conjunto de términos que hacía referencia a diferentes temas que han sido relativamente recurrentes en los medios de comunicación durante los últimos años. Entre otros: la ley de extranjería, la política de inmigración, el drama de las pateras o el problema de seguridad nacional.

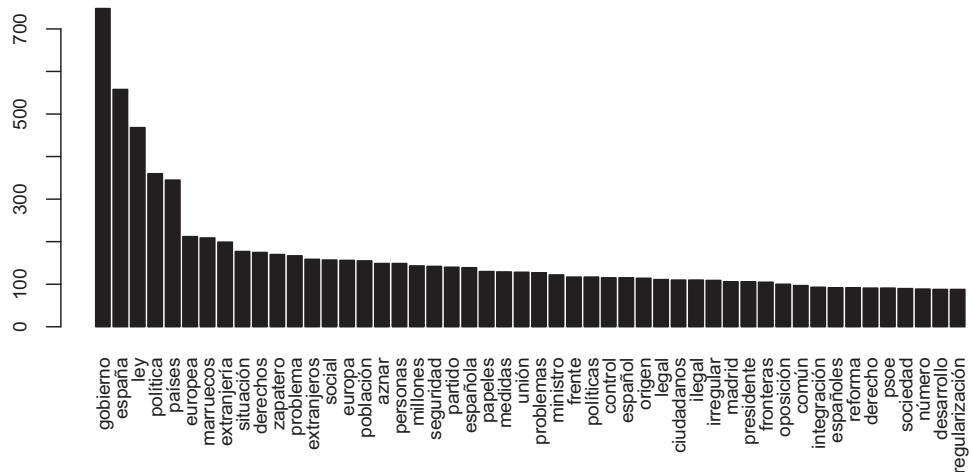

Figura 1. Top 50 términos más frecuentes

Cuando se extrajeron los términos más frecuentes para cada uno de los años de la muestra analizada, en general, no se observaron grandes diferencias durante este periodo. En efecto, si tratamos de clasificar manualmente cuáles son los términos que aparecen en cada año se pueden observar algunas diferencias específicas (ver Tabla A1, Anexo I). Por ejemplo, en el año 1999 se hacía referencia a la atención sanitaria a menores de edad, mientras que las referencias a Europa y a la política común de inmigración no aparecen hasta el año 2002, una vez que ya se comenzaban a percibir los efectos de la Ley Orgánica de Inmigración de 2000. Ahora bien, ¿hasta que punto estamos seguros de que estos son los términos relevantes y no otros? Evidentemente, justo por detrás de los treinta términos con mayor relevancia –es decir, aquellos con mayor frecuencia estadística– que, a modo ilustrativo, han sido seleccionados, se encuentra un conjunto más amplio de categorías que pueden resultar relevantes para el análisis. Si bien, este conjunto no-observado de categorías presenta una menor relevancia en términos de frecuencia de aparición, a fines analíticos específicos podrían resultar relevantes. Sin embargo, teniendo en cuenta que nuestro objetivo analítico se dirige a la identificación de los encuadres generales que han enmarcado el fenómeno de la inmigración en un corpus textual específico, el presente análisis se centrará exclusivamente en la identificación de estos marcos conceptuales de mayor frecuencia y, por consiguiente, más representativos del conjunto textual analizado.

4.2. Proximidad entre categorías

Mediante un análisis cluster jerárquico se identificaron aquellos conceptos que, en términos de proximidad, se encontraban más cercanos. Es decir, mediante esta técnica de análisis multivariable se localizaban las categorías que se encontraban agrupadas. El primer cluster identificaba conceptos relacionados con la Ley de Extranjería (p.e. derechos, legal, papeles, regularización, irregular, etc.). El segundo cluster identificó una mayor gama de categorías que hacían referencia al debate sobre el impacto de la inmigración sobre la economía del país y la sociedad (‘aumento’, ‘población’, ‘desarrollo’, ‘economía’, ‘trabajadores’, ‘integración’, ‘sociedad’). El tercer cluster, identificó claramente el encuadre que hacía referencia a las relaciones España-Marruecos (‘marruecos’, ‘rabat’, ‘relaciones’, ‘melilla’). El cuarto identificó el encuadre sobre el debate político entre gobierno y oposición sobre la política de inmigración. Este encuadre quedó formado por categorías como: ‘oposición’, ‘partido’, ‘presidente’, ‘rajoy’, ‘reforma’ ‘ministro’, etc. En quinto lugar se localizó el encuadre que centraba el debate sobre la necesidad de una política de inmigración común en el marco de la UE (‘europa’, ‘ilegal’, ‘común’). Finalmente, el último encuadre que se encontró fue el referido a la denominada ‘crisis de los cayucos’ de Canarias (‘crisis’, ‘canarias’, ‘control’, ‘fronteras’, ‘medidas’).

4.3. Caracterización de frames

Una vez localizados los términos/categorías que describían los seis encuadres predominantes en el conjunto de textos, se llevó a cabo un análisis de correspondencias. Mediante este segundo análisis se pretendía validar la clasificación inicial realizada mediante el análisis cluster jerárquico. El análisis de correspondencias obtuvo unos re-

sultados similares a los del cluster jerárquico. En efecto, cuando representábamos las diferentes categorías en un espacio bidimensional, la composición de los grupos resultaba relativamente similar. No obstante, mediante esta segunda técnica se posibilitaba un mejor y más complejo ajuste de las categorías en el análisis. Por ejemplo, ahora el encuadre que hacía referencia a la Ley de Extranjería incluía un mayor número de categorías, que se encontraban dimensionalmente próximas a las del control de la inmigración desde la política. El encuadre referido a la política común europea de inmigración quedaba definido justo por encima del cluster de la política nacional. El cluster que agrupaba categorías de debate socio-económico sobre la inmigración quedó separado del encuadre que apuntaba al aumento de la población en el mundo. Y, nuevamente, se localizaba el encuadre sobre la crisis entre las relaciones de España y Marruecos.

4.4. Validación manual de la clasificación

Finalmente, se llevó a cabo una clasificación manual de una muestra aleatoria extraída a partir de la población de textos analizados (N=130). Así, dos codificadores humanos realizarían una clasificación temática de los textos. En concreto, los codificadores recibieron la instrucción de que identificaran el encuadre al que se estaba haciendo referencia en los textos aleatoriamente seleccionados. Se les pedía que leyeran el texto y que extrajeran el tema general al que se hacía referencia.

Como se podía esperar, la clasificación de los codificadores manuales resultó ser más precisa que la realizada por la técnica automática. En conjunto se localizaron nueve encuadres o frames diferentes: (F1) Control de la inmigración ilegal; (F2) Economía y trabajo; (F3) Políticas de integración de inmigrantes; (F4) Relaciones España-Marruecos; (F5) Ley de Extranjería; (F6) Política Común de Inmigración en la UE; (F7) Relaciones internacionales en materia de inmigración; (F8) Tragedia de la inmigración; (F9) Otros temas, de dudosa clasificación. El análisis de fiabilidad de la clasificación manual resultó ampliamente satisfactorio (α=0,92).

	Frame 1	**Frame 2**	**Frame 3**	**Frame 4**	**Frame 5**	**Frame 6**	**Frame 7**	**Frame 8**	**Frame 9**	**Total**
Frame 1	**20**	0	0	0	0	0	0	0	0	20
Frame 2	0	**11**	0	0	0	0	0	0	0	11
Frame 3	0	0	**13**	0	0	0	0	0	0	13
Frame 4	0	0	0	**14**	0	0	0	0	0	14
Frame 5	0	0	0	0	**19**	0	0	0	0	19
Frame 6	0	0	0	0	0	**15**	0	0	0	15
Frame 7	0	0	0	0	0	0	**6**	0	0	6
Frame 8	0	0	0	0	0	0	0	**11**	0	11
Frame 9	0	0	8	0	0	5	4	0	**4**	21
Total	20	11	21	14	19	20	10	11	4	**130**

Tabla 1. Clasificación manual de textos en función de la temática.*

*Frames localizados: (F1) Control de la inmigración ilegal; (F2) Economía y trabajo; (F3) Políticas de integración de inmigrantes; (F4) Relaciones España-Marruecos; (F5) Ley de Extranje-

ría; (F6) Política Común de Inmigración en la UE; (F7) Relaciones internacionales en materia de inmigración; (F8) Tragedia de la inmigración; (F9) Otros temas, de difícil clasificación.

Al comparar la clasificación automática frente a la manual se identificaron ciertas regularidades. El F1 referido al control de la inmigración coincide con el cluster. El F4 que clasifica los textos sobre la crisis en las relaciones España-Marruecos es claramente identificado por la técnica automática, así como el F6 sobre la política común de inmigración europea. El F7 parece estar relacionado con el encuadre referido al aumento de la población a nivel mundial. Por otra parte, mientras que los codificadores humanos identifican tres encuadres diferenciados en F2, F3 y F8, la técnica automática los agrupa en el cluster resultante del análisis de correspondencias. Finalmente, en el F9 se identificaron otros textos que contenían el término inmigración pero que, residualmente, hacían referencia a otros temas (por ejemplo, 'medidas contra el dopaje y el racismo en el deporte' o el 'aumento de casos de SIDA en España').

5. Discusión

Como se demuestra en este estudio, las técnicas analíticas de minería de datos permiten, mediante la categorización y clasificación automática de términos por documento, la identificación de los diferentes encuadres, no obstante, aún están lejos de poder obtener los niveles de validez de la codificación humana (Klüver, 2009; Álvarez-Gálvez, 2012). Por este motivo, estas nuevas herramientas siguen dependiendo de la supervisión de codificadores humanos entrenados que confirmen la validez de los métodos automáticos. Evidentemente, tampoco debemos pensar que el trabajo de los codificadores humanos sea infalible. La cognición humana tiene unos límites. La labor de clasificar/categorizar la información textual puede resultar relativamente sencilla cuando se trabaja con codificadores entrenados y documentos conocidos, pero dicho trabajo se complica proporcionalmente a medida que aumenta el tamaño del corpus textual (esto es, el conjunto de textos a analizar) y la variedad temática. Sin embargo, las técnicas de minería de textos pueden resolver estos problemas y, frente a las técnicas clásicas de análisis de contenido, aportan nuevas ventajas. Entre otras: la posibilidad de manejar corpus textuales de gran tamaño de forma rápida, la realización automática de análisis complejos y la reducción de errores humanos (Álvarez-Gálvez, 2012). Así pues dichas técnicas representan una nueva alternativa para los investigadores del ámbito de las ciencias sociales, así como una ventaja para procesos de toma de decisión en los que se precisa de una rápida respuesta (por ejemplo, en la toma de decisiones políticas, el análisis de las corrientes de opinión en la web y en las redes sociales).

Comparando ambas clasificaciones (manual vs. automática), se puede apreciar que los codificadores humanos encuentran diferencias entre los encuadres F2, F3 y F8, mientras que la técnica automática los agrupa en un solo cluster. Este desajuste entre ambas clasificaciones corrobora que los codificadores humanos identifican diferencias semánticas en estos tres frames que la técnica automática es, inicialmente, incapaz de determinar. Ahora bien, la cuestión que se nos plantea es la siguiente: ¿realmente es este un error de la técnica automática? En principio, lo lógico sería pensar que sí. Sin

embargo, debemos de tener en cuenta que parte de los errores de clasificación en el análisis de fiabilidad se dieron en el F3 (además del F6 y F7), lo cual indica que también la clasificación manual tendría cierto margen de mejora. De hecho, es necesario considerar que mientras que la técnica automática basa su clasificación en el contenido manifiesto de la información, la clasificación humana se basa tanto en contenido manifiesto como latente. Lo que, por una parte, hace la clasificación más válida a nivel general, pero al mismo tiempo, introduce sesgos que dificultan la replicabilidad de los resultados y, por consiguiente, su comparabilidad (Klüver, 2009).

Si bien, en principio, el análisis automático que se ha llevado a cabo puede resultar generalista, debemos considerar que dicho análisis se puede refinar todo lo que se desee. En este estudio de caso hemos trabajado con textos, extraídos de un diario específico, que hacían referencia a la inmigración como tema general, de modo que pudiéramos identificar los encuadres básicos de este fenómeno en los últimos años. Lógicamente, también se podrían seleccionar textos clave sobre encuadres específicos para así, de este modo, estudiar en profundidad la relación entre las diferentes categorías de los encuadres. Seleccionando encuadres predefinidos, se podrían localizar los términos que se encuentran asociados a otros términos que pudieran resultar de interés a nivel analítico y, así estudiar aquellos que dimensionalmente se encuentran más próximos y que suelen aparecer localizados en artículos con determinadas características. Como por ejemplo el término 'patera', que suele aparecer en artículos en los que aparecen las palabras 'subsaharianos', 'marruecos', 'ceuta', 'melilla', 'mar' o 'cayucos', o el término 'delincuencia' suele localizarse en artículos en los que se hace referencia a 'bandas', 'violencia', 'peleas', 'asaltos', etc.

6. Conclusiones

A partir de los resultados podemos afirmar que estas técnicas presentan un elevado grado de fiabilidad en la clasificación y categorización del contenido manifiesto, así como para la identificación del encuadre textual de las noticias y la descripción de sus elementos. Unos resultados que favorecen la replicabilidad de los estudios, al mismo tiempo que se eliminan los sesgos de la clasificación manual. Contando con los límites propios de la clasificación del contenido latente y de las ambigüedades propias del lenguaje, en minería de textos (y en minería de datos, en general), la complejidad del análisis puede incrementarse y mejorarse todo lo que se desee, siempre partiendo de estos límites. Por consiguiente, se hace patente la necesidad de extender el uso de estas herramientas automáticas al ámbito de las ciencias sociales y, especialmente, aquellos estudios en comunicación que quieran encontrarse adaptados a la velocidad de los flujos informacionales de nuestro tiempo.

7. Referencias bibliográficas

ÁLVAREZ-GÁLVEZ, Javier (2012): "Análisis cuantitativo de textos: del análisis de contenido al tratamiento de textos como datos.", en ARROYO MENÉNDEZ, Millán y SÁDABA, Igor (Eds.): *Metodología de la investigación social. Técnicas innovadoras y sus aplicaciones.* Madrid, Síntesis.

BATESON, Gregory (1955): "A theory of play and fantasy. Steps to an ecology of mind" (pp. 177-193). New York, Ballantine.

CAPPELLA, Joseph N., & JAMIESON, Kathleen H. (1997): *Spiral of cynicism. The press and the public good*. New York, Oxford University Press.

D'ADAMO, Orlando; GARCÍA BEAUDOUX, Virginia; y FREIDENBERG, Flavia (2007): *Medios de Comunicación y Opinión Pública*. Madrid, McGraw Hill, 2007.

DE VREESE, Claes & BOOMGAARDEN, Hajo (2003): "Valenced news frames and public support for the UE". *Communications*, 28 (4), pp. 36-381.

ENTMAN, Robert (1993): "Framing: Toward a clarification of a fractured paradigm". *Journal of Communication*, vol. 43, nº 3, pp. 51-58.

GOFFMAN, Erving (1974): *Frame analysis: An essay on the organization of experience*. Cambridge, MA, Harvard University Press.

KLÜVER, Heike (2009): "Measuring Interest Group Influence Using Quantitative Text Analysis". *European Union Politics*, vol. 10 nº. 4, pp. 535-549.

HAIDER-MARKEL, Donald P., DELEHANTY, William, & BEVERLIN, Mathew (2007): "Media Framing and Racial Attitudes in the Aftermath of Katrina". *Policy Studies Journal*, 35 (4), pp. 587–605.

HOPKINS, Daniel & KING, Gary (2010): "A Method of Automated Nonparametric Content Analysis for Social Science". *American Journal of Political Science*, 54, 1, pp. 229-247.

IGARTUA, Juan José y MUÑIZ, Carlos (2004): "Encuadres noticiosos e inmigración. Un análisis de contenido de la prensa y televisión españolas". *Zer: Revista de estudios de comunicación,* 2004, nº, 16, pp. 87-104.

IGARTUA, Juan José y CHENG, Lifen (2009): "Moderating effect of group cue while processing news on immigration. Is framing effect a heuristic process?" *Journal of Communication,* 59 (4), pp. 726-749.

IGARTUA, Juan José; MORAL, Félix; y FERNÁNDEZ, Itziar (2011): "Cognitive, attitudinal and emotional effects of the news frame and group cues on processing news about immigration". *Journal of Media Psychology,* 23 (4), pp. 174-185.

IYENGAR, Shanto & KINDER, Donald (1987): *News that matters: Agenda-Setting and priming in a television age*. Chicago, University of Chicago Press.

KEPPLINGER, Hans Mathias; GEISS, Stefan; & SIEBERT, Sandra (2012): "Framing Scandals: Cognitive and Emotional Media Effects". *Journal of Communication*, 62 (4), pp. 659-681.

LAVER, Michael; BENOIT, Ken; & GARRY, John (2003): "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review*, 97, pp. 311-331.

MCCOMBS, Maxwell (2004): *Setting The Agenda: The Mass Media And Public Opinion*. England, Polity Press. UK, Cambridge.

REESE, Stephen D. (2007): "The Framing Project: A Bridging Model for Media Research Revisited". *Journal of Communication*, 57 (1), pp. 148–154.

ROBERTS, Carl W. (2000): "A conceptual framework for quantitative text analysis". *Quality & Quantity*, 34 (3), pp. 259–274.

RODRIGO ALSINA Miquel (2005): *La construcción de la noticia*. Barcelona, Paidós.

SCHEUFELE, Dietram & TEWKSBURY, David (2007): "Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models". *Journal of Communication*, 57 (1), pp. 9–20.

SLAPIN, Jonathan & PROKSCH, Sven-Oliver (2008): "A Scaling Model for Estimating Time-Series Party Positions from Texts". *American Journal of Political Science*, 2008 52 (3), pp. 705-722.

TANKARD, James; HENDRICKSON, Laura; SILBERMAN, Jackie; BLISS, Kriss; & GHANEM, Salma (1991): "Media frames: Approaches to conceptualization and measurement". Paper presented at the annual convention of the *Association for Education in Journalism and Mass Communication*, Boston, MA.

WARD, Joe H. (1963): "Hierarchical grouping to optimize an objective function". *Journal of the American Statistical Association,* 58, pp. 236-244.

WEISS, Sholom; INDURKHYA, Nitin; ZHANG, Tong; & DAMERAU, Fred (2005): *Text Mining: Predictive Methods for Analyzing Unstructured Information*, Springer.

8. Anexo: Tabla A1. Conjunto de las treinta palabras más frecuentes por año

Rank.	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
1	ley	gobierno	ley	gobierno	gobierno	gobierno	españa	país	país	gobierno
2	gobierno	ley	gobierno	país	aznar	españa	país	españa	españa	país
3	españa	país	españa	españa	país	país	gobierno	países	gobierno	política
4	país	extranjería	extranjería	política	españa	política	países	gobierno	países	rajoy
5	congreso	españa	país	ley	política	países	marruecos	europea	política	zapatero
6	extranjería	derechos	política	países	ley	zapatero	europea	política	ley	españa
7	derechos	política	marruecos	europea	países	extranjeros	política	problema	marruecos	extranjeros
8	grupo	partido	situación	aznar	seguridad	millones	millones	zapatero	personas	países
9	senado	países	papeles	delincuencia	derechos	papeles	población	millones	trabajadores	berlusconi
10	situación	control	regularización	sevilla	extranjería	población	zapatero	unión	europea	partido
11	social	ejido	derechos	extranjería	irregular	europa	europa	europa	medidas	unión
12	hijos	marruecos	irregular	seguridad	oposición	española	problema	medidas	presidente	ley
13	países	población	países	europa	problemas	social	regularización	población	española	crisis
14	población	psoe	personas	marruecos	legal	integración	madrid	extranjeros	español	europea
15	sanitaria	reforma	legal	origen	marruecos	ley	ilegal	presidente	europa	marruecos
16	popular	situación	ilegal	española	reforma	marruecos	situación	canarias	extranjeros	derecho
17	centros	social	pacto	frente	psoe	ciudadanos	unión	español	control	congreso
18	enmiendas	fuerzas	reglamento	problema	delincuencia	legal	melilla	personas	francia	directiva
19	legal	oposición	sociedad	situación	lucha	número	problemas	ley	relaciones	legislatura
20	número	sociedad	vigor	común	problema	situación	social	ciudadanos	economía	sarkozy
21	problema	políticas	aznar	control	social	aznar	extranjeros	común	fronteras	situación
22	española	expulsión	constitucional	irregular	control	derechos	marroquí	europeo	mayoría	debate
23	extranjeros	judicial	frente	medidas	frente	europea	ministro	fronteras	debería	económica
24	socios	proyecto	ministro	social	guerra	laboral	rabat	papeles	desarrollo	origen
25	atención	ministro	rey	canarias	políticas	debería	ceuta	crecimiento	europeo	derechos
26	legislatura	regularización	social	ciudadanos	español	partido	ley	elecciones	frente	fronteras
27	personas	jóvenes	cuestión	aumento	humanos	personas	políticas	ilegales	población	medidas
28	proyecto	personas	entrada	legal	madrid	alemania	comunidades	número	sarkozy	políticas
29	sociedad	problemas	española	personas	partido	españoles	fronteras	trabajadores	seguridad	problemas
30	trabajadores	vida	mafias	extranjeros	personas	origen	desarrollo	áfrica	situación	servicios
	1922/16	3368/34	3366/33	4494/50	3209/36	3533/37	3923/42	4281/50	3148/29	3078/29