Trabajo Fin de Grado

Optimización de contenidos web: un factor clave del éxito empresarial

Autor

Andrea Fabián Abad

Director

Pilar Olave Rubio

Facultad de economía y empresa
2017

Autor: Andrea Fabián Abad

Director: Pilar Olave Rubio

Título del trabajo: Optimización de contenidos web: un factor clave del éxito empresarial

Titulación: Marketing e investigación de mercados

Modalidad: Análisis de datos

RESUMEN

El uso de técnicas computacionales en el campo del análisis web es obligatorio, con el incipiente desarrollo de negocios web y de ámbitos empresariales en línea, la búsqueda de la obtención de valor radica de un análisis minucioso del usuario y de sus necesidades. El objetivo de este estudio es la investigación de técnicas predictivas en el ámbito del marketing digital y su uso en la optimización de contenidos. Buscando patrones que ayuden al redactor a conseguir mejores resultados de impacto entre sus lectores, entre casi 40.000 artículos de la web ***mashable.com***.

Los días de la semana más óptimos en referencia a acciones, son los fines de semana y los lunes. Gracias a esta información se establece un criterio de publicación sabiendo que pasadas unas 24 horas el artículo tomará un impacto fuerte. Aunque como vemos este no es el único factor que influye en la predicción, por lo que se han propuesto varios métodos de modelización. El modelo que mejor resultado ofrece para predecir la popularidad es el diagrama de árbol, a partir de dos características: Promedio de palabras clave, número de archivos web. ***Estas técnicas consiguen optimizar la redacción de contenidos y facilitan mejores resultados empresariales en lo que al Marketing digital se refiere.***

ABSTRACT

The computational techniques are linked with the web analysis. However with the emerging development of online businesses, the value matches the knowledge of user's needs. The aim of this report is to research the predictive technics within the digital marketing field, and to optimize this content. We have been looking for patterns within around 40.000 mashable.com articles in order to get more shares thanks to the drafting style.

Weekends and Mondays are the optimal days to get a strong impact after 24 hours. Anyway, as we can see it is not the unique feature that contributes this prediction. Therefore we have created some models. The best outcome is provided by Random Forest when predicting the popularity. It is done through two features, Keyword average and number of hits. These techniques are made to enhance the content drafting and to lead better business outcomes at Digital Marketing industry.

ÍNDICE

ÍNDICE DE GRÁFICOS Y TABLAS

INTRODUCCION AL ANÁLISIS DE CONTENIDOS.

El desarrollo de las tecnologías de la información y comunicación (TICS) nos ha llevado a una situación nueva, donde se transmite mucha más información y de una forma más rápida. Es ahí donde entra en funcionamiento la Analítica Big Data, capaz de estudiar tanta información que puede estimar resultados en tan solo segundos. Solo ahora con las herramientas analíticas adecuadas somos capaces de gestionar muchos datos creados por el conjunto de la sociedad, para poder llegar a entenderlos necesitamos un análisis de datos eficiente.

El concepto de **Marketing de contenidos** ha surgido hace apenas unos años, con la aparición de la web 2.0 y de las nuevas necesidades de información de los usuarios que corresponden con una necesidad de actualización cada día más latente. El principal motivo de la investigación es conocer la popularidad de los artículos, noticias y demás publicaciones convertidas en "post", o como se conocen a los elementos web compartidos en los "muros" de redes sociales.

En este nuevo contexto, esta investigación propone un análisis Big Data a través de los datos y metadatos recopilados de una página web, ***mashable.com***, dedicada a la publicación de documentos periodísticos o de artículos de interés. Esta empresa es reconocida mundialmente por publicar sobre tecnología, aunque con el paso de los años ha ido abordando temas muy variados.

El proyecto se enmarca dentro de las técnicas de Minería de datos relacionado con el marketing de contenidos bajo el objetivo de marcar unas pautas y estrategias operacionales en la empresa, para la optimización de la redacción y publicación de estos textos. Las técnicas de minería de datos están basadas en la extracción de información relevante en grandes volúmenes de datos, el objetivo es extraer patrones de comportamiento o tendencias que vuelquen un sentido relacional o explicativo a los datos en un contexto concreto. Los datos bien interpretados generan información, y ésta genera conocimiento al recibir el valor agregado del modelo, según Earl (2000), el modelo se utilizará de forma estratégica, para formar un juicio sobre una cuestión y aportar **valor.**

Aunque en el análisis de datos pueda ser muy variado, este trabajo se centra concretamente en los sistemas de gestión de contenidos y documentos (content and document management). En este entorno del marketing de contenidos y del análisis de resultados, el objetivo de este trabajo es crear un modelo predictivo basado en la popularidad de artículos publicados en ***mashable.com*** a través de la búsqueda de patrones en el diseño de textos que se pueda aplicar a futuras publicaciones, es decir optimizar la redacción de contenidos para conseguir que el elemento sea compartido cuantas más veces mejor.

Este caso de análisis se puede conocer también como minería de textos, el cual hace referencia a un conjunto de métodos y técnicas que permiten sacar la información más significativa de forma automática entre un gran volumen de datos textuales, normalmente no estructurados. Se conoce que aproximadamente el 80% de la información de una empresa está archivado en forma de texto, por lo que este tipo de análisis puede ser realmente importante.

En este contexto hay que hacer referencia a conceptos del marketing digital ya que será necesario usar las técnicas adecuadas para poder llegar a la audiencia máxima posible. Para llegar a comprender la magnitud de este análisis describiremos las estrategias el SEO y SEM conocidas como técnicas que optimizan el resultado en buscadores web, y que reportan una mayor cantidad de tráfico en la web. Se diferencian en que el primero crea tráfico de forma orgánica, es decir, haciendo una correcta utilización de las herramientas del diseño web y adecuándose a los algoritmos que siguen los motores o arañas de búsqueda y por ende sin necesidad de pagar por posicionarse como ocurre en SEM.

Cuando una gran cantidad de datos son recibidos instantáneamente y modelados adecuadamente, permiten establecer predicciones y actuar en consecuencia. El fenómeno del Big Data, que comenzó en las compañías nativas digitales, se ha extendido a otros sectores y correlaciona positivamente con el aumento de la productividad. Algo que ya supieron prever hace cinco años Brynjolfsson y McAfee (2012). Además, ante la creciente aparición de estas compañías digitales que se dedican únicamente a publicar información, este trabajo busca optimizar el diseño y la redacción de textos mediante la popularidad generada por cada artículo.

La audiencia es un indicador de rentabilidad para las agencias de medios, cuanto mayor sea esta mayor será su éxito, este trabajo se basará en la adaptación de textos para que se conviertan en artículos virales al superar un nivel de popularidad, y estableciendo así unos modelos de actuación estratégica que dependan del tema que se está tratando y de la forma en la que se redactan.

Es cierto que actualmente el Big Data es muy mercado muy potencial y beneficioso, durante la realización de mis prácticas de empresa en un medio digital encontré falta de planificación estratégica de contenidos y personalmente opino que para una optimización de la productividad, de la satisfacción del lector y del éxito de las campañas publicitarias es necesaria la aparición de la analítica web o de la minería de textos como estudio de las necesidades y preferencias del usuario o lector, ya que es este el único que te puede llevar al éxito en el ámbito de las comunicaciones digitales.

Esto implica trabajar analíticamente con grandes bases de datos, ya que cada día se generan mayores volúmenes de información y esto continuará creciendo en el futuro, el director ejecutivo de Google, Eric Schmidt (2011) afirmó que "generamos más información en dos días que en toda nuestra historia hasta antes del 2003".

Como es conocido en Marketing los clientes son el centro de toda estrategia y de la misma forma sucede con el Marketing digital, la estrategia de contenidos debe estar definida para un público objetivo, conociendo cuales son las necesidades de información de los usuarios a los que la empresa se dirige. Un artículo publicado en la página web caso de estudio, afirma que la redacción de artículos cada día es más inteligente, conociendo lo que los lectores quieren y cómo lo quieren, es decir adaptándose al público, y no como hacían antes los periodistas que daban su propia opinión, ahora el método ya no es "Push" sino "pull", las ideas de redacción provienen de las reacciones que el público tiene frente al contenido. La información ha de fluir bidireccionalmente para que esto sea posible, hoy en día internet se ha convertido en el lugar donde la gente puede expresarse libremente, aportando opiniones sobre experiencias o productos, la información fluye entre empresas y usuarios. Por lo que, los medios de comunicación digital luchan por aumentar el tráfico, donde los temas más provocativos o sensacionalistas toman relevancia para captar la atención del lector.

MARCO TEÓRICO: LA INFORMACIÓN DE CONTENIDOS COMO ACTIVO ESTRATÉGICO.

La evolución de las tecnologías hace que las empresas creen soluciones o herramientas para recoger datos que afectan a entornos web, empresas como IBM o WebTrends son consideradas pioneras en este ámbito. Herramientas como NetGenesis, hoy integrada en SPSS, han ido evolucionando y proporciona unos niveles de análisis mayores gracias a la generación masiva de datos.

A partir del año 2011 se acuña el término "Big Data", que hace referencia al estudio telemático de grandes volúmenes de datos. Gartner (2011), uno de los principales referentes tecnológicos, lo definió como el "conjunto de activos de información caracterizados por su gran volumen, velocidad y variedad, que exigen formas innovadoras y rentables de procesamiento de la información para mejorar la comprensión y la toma de decisiones". Para que la información sea considera Big Data, deben de cumplir las 5v's establecidas por Doug Laney (2001), volumen, velocidad, variedad, veracidad y valor.

Los datos recopilados de la web son un claro ejemplo del término Big Data, millones de solicitudes son realizadas a los servidores y estos almacenan cada una de ellas en una base de datos, por lo que todas las acciones quedan registradas. Por otro lado, actualmente la forma que tienen las compañías de transmitir su mensaje de forma eficiente es a través de los buscadores y de las redes sociales lo que contribuye a que a mayor público se genere un mayor tráfico web. El posicionamiento de la web también repercute al posicionamiento mental que los consumidores tienen de las marcas, si la web aparece primera en el resultado de búsqueda el consumidor valorará ese posicionamiento aportándole un mayor reconocimiento a la marca.

La analítica digital es un planteamiento estratégico que afecta al entorno empresarial y al de negocio permitiendo el estudio del comportamiento de los visitantes de una página web a lo largo del tiempo, aunque se pueden aportar otras visiones del concepto como la de la Asociación Española de Analítica Web "recopilación, medición, evaluación y explicación racional de los datos obtenidos de Internet, con el propósito de entender y optimizar el uso de la página web de la organización"

La finalidad de este análisis es la toma de decisiones en estrategias de negocio, ya sea a nivel online u offline, pero los datos con los que se trabaja son extraídos en su totalidad de la web.

La analítica digital es un gran entorno de estudio, pero podemos dividirla según el tipo de objetivos que se desean conseguir: analítica de contenido, analítica de audiencia, analítica de comportamiento, analítica de redes sociales, analítica de conversiones, y analítica de publicidad. La analítica web también depende del uso que se vayan a dar a los datos, por lo que se puede diferenciar según la función o necesidad que cubre cada web. Entre ellos podemos encontrar sitios web de contenidos, webs que pretenden crear contactos potenciales (leads) a través de formularios o de redirecciones a una web concreta, pueden ser sitos webs dedicados al e-commerce, en este trabajo me centrare en la **analítica de contenido**.

Se examinarán aproximadamente unos 40.000 artículos publicados en la web mashable.com, de forma que se establezcan unos principios de redacción de contenidos o métodos de producción del texto que conlleven a un aumento de visitantes o lectores, provocando en ocasiones un efecto viral de estos. En este caso la analítica de contenido ayuda a mejorar la usabilidad de los sitios web y de forma que se adapta al usuario y a mejora la experiencia de navegación, es decir, aporta un valor extrínseco a la marca que es percibido por el público.

El Marketing digital puede quedar definido a través de procesos que permiten establecer, promocionar y crear valor para los servicios o productos mediante técnicas en la web. Cada día hay nuevas herramientas y estrategias que abarcan este ámbito, pero en el estudio de la efectividad del contenido hay que destacar las estrategias SEO. Estas afectan directamente sobre la redacción de artículos periodísticos ya que de este depende de su posicionamiento en el buscador, las siglas de SEO provienen de Search Engine Optimization, básicamente se puede definir como la optimización de la página web para conseguir un buen posicionamiento en los buscadores cuando se crea una solicitud de búsqueda avanzada a través de las palabras clave.

Hay ciertos elementos que deben de ser diseñados y analizados para un posicionamiento eficiente, a continuación se recoge una breve explicación de cada uno de ellos y de su funcionamiento.

Palabras claves: conjunto de palabras que se asocian a un tema como forma de consulta. Las palabras clavé son muy importantes ya que se utilizan para categorizar un texto, para indexar la web un buscador y facilitar su búsqueda. Hay que tener en cuenta a las conocidas como palabras vacías, en inglés stop-words, ya que estas son ignoradas por los buscadores, suelen ser pronombres, preposiciones, es decir palabras sin sentido por si solas. Otras palabras están prohibidas por los buscadores para la indexación, palabras vulgares o políticamente incorrectas, estas hay que evitarlas.

La densidad de las palabras clavé es el número de veces que aparecen estas palabras en la página, con respecto al total de palabras. Algunos expertos recomiendan que esta densidad sea del 5% para no saturar el texto, aunque aquí hay diversas opiniones. En caso de exceso, los buscadores lo penalizan como una técnica black hat SEO.

Cuando las palabras clave han sido seleccionadas hay que llevar a cabo una serie de acciones en la propia web, en concreto hay que optimizar el código HTML, dentro de este archivo hay ciertas etiquetas que tienen más relevancia que otras a la hora de ser reconocidas por los motores de búsqueda. La etiqueta tittle (título) es probablemente la más importante ya que si situamos palabras clave en él, este será la fuente principal de búsqueda y aparecerá en la página de resultados, también conocida como SERP (search engine results page), o en la pestaña del buscador. El título es una de las etiquetas que componen la etiqueta head, que también está formada por las etiquetas meta description y meta keyword, que ayudan al buscador a recopilar información del sitio web. Otro de los elementos más relevantes para el uso de palabras clave es la etiqueta anchor text, es un texto en el que se detectan los hipervínculos, por decirlo de otra forma, son palabras o frases incluidas en el contenido que te dirigen a otra página a través de un enlace. Es tan importante porque los robots de búsqueda rastrean los enlaces y reconocen la palabra clave, esta examinación se hace en enlaces entrantes y salientes de la página. Para ganar un puesto mayor en el ranking también es necesario revisar las etiquetas head, que hacen referencia a los encabezamientos del texto, y optimizarlas con palabras

clave. Existen hasta 6 niveles, pero las realmente importantes se conocen como H1, H2, H3.

La optimización en la redacción del cuerpo de texto consiste en incluir una o dos palabras clave en un párrafo pero siempre que tengan sentido, aunque no es tan sencillo, hay muchos otros factores a tener en cuenta. Existen herramientas que se utilizan para la gestión de contenido, los conocidos como CMS (Content Management Systems) sirven para publicar, crear, actualizar, y buscar contenido.

Iñigo Arbildi (2005) realizó una lista de los factores positivos de posicionamiento dentro de la página:

- La densidad de palabras clave en el texto tiene que estar entre un 5% y un 20% del total del texto (de la etiqueta body).
- El tamaño y el tipo de fuente ayudan a remarcar alguna palabra clave.
- El uso de listas o tablas también ayuda y mejor si la palabra clave se sitúa al comienzo.
- El orden de las palabras clave mejor si coinciden con el orden de palabras de consulta.
- El contenido que se presenta mediante gráficos o imágenes es invisible a no ser que utilicemos el atributo alt y que le demos una equivalencia textual además de añadir una palabra clave.
- Como sabemos hay enlaces que pueden hacernos bajar en el ranking, para que estos sean invisibles y el buscador no los tenga en cuenta hay que usar la etiqueta Nofollow después de incluir el enlace en el código HTML.

En resumen, para mejorar esa posición del ranking hay que prestar atención a las páginas de entrada y salida, a los títulos de las páginas, al contenido del sitio, los gráficos y a la estructura del sitio web, más concretamente sobre las palabras clave, los enlaces, el código HTML y el metaetiquetado.

Como hemos visto, no solo las acciones dentro de la propia página son relevantes, también repercuten las acciones externas. Por lo general hay que crear una estrategia con los enlaces, ya que son una de las cuestiones más representativas de las buenas técnicas en SEO, se pretende enlazar el sitio web con otros que ofrezcan información

relevante. Los enlaces son la forma más básica para redirigir tráfico, los inlinks o enlaces entrantes hacen referencia a los que recibe una página, es decir los que apuntan hacia ella. Por otro lado están los enlaces salientes u outlinks que son los que la página envía, o lo que es lo mismo son enlaces que apuntan a otras páginas web. Cuando hay muchos enlaces entrantes se puede decir que la página es popular, porque el algoritmo de PageRank, en el caso de google, es capaz de rastrear todas las rutas de enlace de forma jerárquica, utilizando el método del árbol y asignándole un valor.

Por esta razón es muy importante tener una planificación en la construcción de enlaces, consiguiendo que otras páginas creen enlaces que apunten hacia la web, creando los enlaces cruzados adecuados sin sobrepasarse ya que también está penalizado, no han de ser artificiales. Aunque esto no es lo único que afecta a los resultados de búsqueda, las campañas de publicidad, la frecuencia de actualización del sitio web, o la cantidad de clics recibidos también afectan al posicionamiento.

Por otro lado, se debe de tener en cuenta que hay algunas acciones que empeoran el posicionamiento ya que suelen ser penalizadas por los motores de búsqueda por tratar de engañarlos. A este tipo de técnicas se las conoces como Black Hat.

- Webspam: cuando las páginas de resultados están llenas de páginas que tienen poco valor para el usuario.
- Enlaces transparentes, utilizan el mismo color de fondo para camuflarlos y no molestar al usuario.
- Enlaces ocultos, situados detrás de gráficos u otros elementos
- Enlace engañoso, que lleve a un sitio al que no se quería llegar.
- Keyword stuffing: Es una metodología que consiste en llenar la página web de palabras clave de forma artificial, por ejemplo escondiendo palabras clave en el fondo de la página. En número de keywords en el texto no debería de superar el 20%
- Enlaces discretos, son los que están visibles pero a una escala muy reducida, 1x1 pixel.
- Meta tag Stuffing, consiste en rellenar todas las metaetiquetas con una o varias palabras clave que se repiten.
- Página de entrada que no sea útil para el visitante.

- Enlaces en el signo de puntuación, el buscador lo encontrará y leerá el anchor text en forma de punto final.
- Cloaking, optimizando la página web a través de texto con el mismo color de fondo, consiguiendo que no afecte a la usabilidad del visitante.
- Páginas de redirección que se utilizan únicamente para optimizar SEO y empeoran esta usabilidad.
- Granjas de enlaces o ataques sibilinos creadas para aumentar el posicionamiento de forma artificial mediante enlaces.
- Wiki Spam, consiste en modificar sitios wiki para añadir enlaces a la web.
- Páginas generadas automáticamente o sitios web ladrones que copian el contenido.

El buscador también penaliza los contenidos que hayan sido copiados o duplicados en páginas webs que no sean de los autores del texto, en esto la propiedad intelectual española es contundente, la disposición actualmente vigente es el Real Decreto 1/1996, aprobado el 12 de abril que recoge la Ley de Propiedad Intelectual exigida por la Comunidad Económica Europea, ley que se ha ido adaptando con el paso de los años.

La actualización del sitio web es muy importante, porque con el paso del tiempo la web puede perder relevancia en el ranking de resultados, eso sí, no se debe introducir información que no tenga relevancia en la página. Un objetivo realista puede ser el de actualizar la página una vez por semana.

Todos estos conceptos son importantes porque los métodos de indexación de los buscadores vienen dados por los algoritmos que establecen el orden de búsqueda. Los fundadores de Google se encargaron de crear el algoritmo de búsqueda como la esencia de google, el PageRank de 1999 que recuperaba información de la web, años después y tras muchas modificaciones el algoritmo sigue siendo la base fundamental de las búsquedas en la web. El pasado 8 de marzo del 2017, el algoritmo de google se actualizó, se piensa que la actualización Fred, como se la conoce, trae cambios generados entorno a la detección de contenidos auténticos y útiles para los usuarios,

también se cree que hay nuevos cambios en los anuncios web, por ende se relaciona con la calidad global del sitio web.

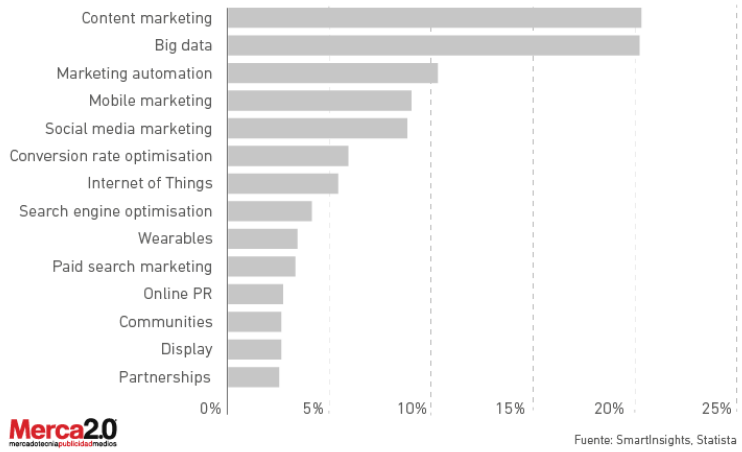
El estudio "Rebooting Ranking Factors. Google.com" de Searchmetrics 2016, muestra como todos estos factores han ido evolucionado respecto a su relevancia dentro del propio ranking. El uso de palabras clave ha ido perdiendo fuerza, casi el 52% de las páginas en el top 20 estudiadas tienen palabras clave en el titulo o en la descripción. Respecto de la experiencia de usuario, el número de enlaces internos es uno de los factores más relevantes en el ranking. Las webs que se encuentran en el top 20 analizadas tienen una media de 1,67 imágenes de más de 200 pixeles por página.

Por otro lado, la correlación entre el uso de redes sociales y el ranking es muy alta, en este caso Facebook es la red social por excelencia para mejorar la indexación en buscadores y para generar mayor tráfico web. Por otro lado el 83% de estas páginas en el top 10 tenían el dominio .com por lo que la función del dominio también es muy importante, además se han ido posicionando mejor los contenidos estructurados y útiles.

Pero no hay que olvidar que el objetivo final de SEO es aumentar los ingresos de la empresa o persona responsable la página web. En este caso, el objetivo será aumentar la popularidad de estos textos, el objetivo del análisis multivariante recae sobre la variable resultado, que ha sido generada mediante el número de veces que ha sido compartido un artículo en las redes sociales.

A este tipo metodologías, cuyo objetivo final es la popularización se le atribuye el nombre de marketing viral. Un efecto viral sobre un artículo llega a conseguir que sea leído por la máxima cantidad de gente en el menor tiempo posible, con lo que consigue un efecto de contagio, ya que todo el mundo queda impactado y desea mostrárselo a sus círculos personales generalmente a través de las redes sociales.

1. Gráfico sector marketing digital



En un artículo publicado en la revista Merca2.0 se recogen las 14 estrategias más efectivas en el sector del Marketing digital, y entre las que ocupan los primeros puestos podemos encontrar las que resultan más efectivas para las empresas. En primer lugar se encuentra el **marketing de contenidos** con un 20.3% de efectividad, seguido de las **herramientas BigData** con un 20.2%, ambas con una distinguida diferencia entre las siguientes estrategias como el 'Mobile Marketing' o el 'Marketing automation'. Este ranking se ha sido generado por SmartInsights.com a través de la herramienta Statista.

Así el marketing de contenidos es el líder en el ámbito digital, aunque sin lugar a dudas para que el contenido sea el mejor y con la mayor audiencia posible, será indispensable que este contenido este analizado y optimizado a los motores de búsqueda. De modo que la información se transmita de la forma más rápida posible, pero sin obviar a las redes sociales. Las técnicas SEO en el marketing de contenidos ayudan a hacer que el contenido sea más interesante, útil y accesible para los usuarios de una página web.

Como hemos comentado el objetivo principal del SEO es la búsqueda de ingresos, pero mediante técnicas SEO lo que se consigue realmente es aumentar el tráfico de la página web. Un buen posicionamiento hace que la web se diferencie entre los competidores y sea más visible. Además, la calidad del contenido ayudará a que las visitas no sean esporádicas, sino que generará **un retorno de lectores cada cierto tiempo**, es decir, el contenido fideliza.

Los datos recopilados de la web se mueven en grandes volúmenes y han de pasar por un software para ser realmente valiosos, cuando la información está estructurada en variables, el análisis facilita el conocimiento de nuestro entorno web y de las posibles decisiones estratégicas que impliquen un valor para la empresa. El hecho de tomar decisiones basadas en hechos reales nos garantiza incidir exactamente sobre la situación

en la que se encuentre a empresa, así el Big data extrae información real con valor predictivo y permite crear estrategias de negocio efectivas.

En este caso de estudio se comprobarán datos de la empresa web ***Mashable.com***, un blog puntero en tecnologías analíticas, la empresa consideró que era importantísimo el estudio de la predicción viral y del análisis de medios digitales. El propio equipo de ***Mashable.com*** creo hace varios años una herramienta conocida como Velocity que explora y recopila información de cómo las personas se relacionan con los contenidos web.

Herramientas como esta que trabaja bajo un algoritmo propio, puede **aportar un valor** que transforme a la empresa exponencialmente, en el caso de mashable.com su popularidad ha aumentado posicionándose como uno de los blogs de habla inglesa más importantes del mundo.

CONTEXTO, DISEÑO Y METODOLOGÍA

3.1. CONTEXTO DE LA EMPRESA

Al ser un caso de estudio adaptado a la realidad hay que detallar los factores fundamentales que explican el funcionamiento de esa página, y ante todo la importancia de la empresa en el sector.

Mashable.com es un sitio web de noticias fundado en 2005 por Pete Cashmore, su actual jefe ejecutivo. La empresa tiene su sede central en Nueva York, además de tener otras oficinas en San Francisco, en Los Ángeles y en Londres. Este medio de comunicación es una fuente líder de información, publica noticias para la nueva generación de jóvenes conectados y es un medio transmisor de cultura digital que inspira a sus lectores.

En su web se define como una compañía global, con medios de comunicación y entretenimiento multiplataforma impulsados por su propia tecnología. Con un récord de 45 millones de visitantes únicos mensuales, y con 28 millones de seguidores en redes sociales se considera como uno de los blogs más influyentes y comprometidos de todo el mundo.

A lo largo de los años ***Mashable.com*** ha ido evolucionando, Pete Cashmore a sus 19 años decidió emprender a través de una compañía digital y para el año 2009 la empresa ya estaba en la lista Forbes de "Top 25 Web Celebs". Han conseguido internacionalizarse, la web está diseñada para los usuarios americanos con la versión web básica (.com), pero además se pueden encontrar las siguientes versiones, mashable Asia, mashable Francia, mashable UK, mashable Australia, mashable India. Así la empresa es capaz de publicar y controlar el contenido más adecuado según la zona geográfica.

La empresa tiene 8,71 millones de seguidores en Twitter por lo que está constantemente en expansión, además en 2016 firmó un contrato de colaboración con el canal de Youtube CineFix, uno de los datos más interesantes sobre esta empresa es la necesidad de innovar y renovarse. Ellos han desarrollado sus propias herramientas, como Velocity *que es capaz de analizar predictivamente la popularidad de un artículo a través de un algoritmo, capacitando a los redactores a posicionar el artículo en la web según la tirada que pueda tener en las redes sociales.* Velocity se creó como una herramienta de depuración de contenidos, pero el desarrollo del modelo predictivo hace que sea una herramienta muy recomendada en cualquier sala de redacción.

Esta herramienta monitoriza la actividad de todas las redes sociales, además ayuda al propio sitio web posicionando artículos según su relevancia social. Han creado una sección nueva llamada watercooler que está controlada por velocity, pues es aquí donde se colocan los artículos de los que más se está conversando en las redes sociales.

Según Haile Owusu, *jefe científico de datos en Mashable.com*, esta herramienta permite reaccionar muy rápidamente ante cambios sociales, cosa que otros equipos hacen a largo plazo. Velocity ha sido galardonada con el premio "Agency of the year" en los premios de marketing de contenidos 2016, por ello esta herramienta se ha convertido en una motivación para el estudio del caso ***mashable.com***.

En estos últimos meses ***mashable.com*** ha ido perdiendo tráfico y ha dejado de ser uno de los blog líderes, como veremos en el análisis de la competencia. Actualmente la empresa tiene un valor estimado de 421.000.000 dólares, lo que lo convierte en un blog muy valioso, es lo suficientemente conocido como para generar $ 57.313 al día ya que

en la página worthofweb.com, dedicada al benchmarking en empresas online, ***mashable.com*** recibe unas 3,821,000 visitas al día y unas 19,104,741 páginas vistas al día estimadas.

Hay 92 enlaces en la página principal, estos se pueden clasificar como enlaces externos: Nofollow (0.2%) o Follow (14%) y los enlaces internos (85.8%). No hay enlaces rotos y la página 404 esta personalizada. Además la versión actual de mashable.com esta optimizada para dispositivos móviles. Mashable.com es una empresa muy popular en las redes sociales y sobre todo los países desde de donde más se accede son Estados Unidos, Reino Unido, India, Canadá y Alemania.

2. Porcentaje de audiencia

Country	Percent of Visitors
United States	47.6%
United Kingdom	6.2%
India	5.6%
Canada	4.1%
Germany	2.5%

Alexa traffic Rank es uno de los rankings más conocidos por las empresas digitales, el ranking perteneciente a la compañía Amazon, muestra el grado de popularidad del sitio web de la empresa, AlexaRank es capaz de monitorear el tráfico web, esta métrica es generada mediante el número de páginas vistas y el alcance de cada una diariamente, por lo que es un ranking que fluctúa continuamente, se asemeja al PageRank de google, puede hacer estimaciones trimestrales de los movimientos que hace en el ranking una web. Esta empresa se sitúa en el puesto 672 a nivel global, ha descendido 93 puestos en 3 meses. Aunque el puesto en el ranking para Estados Unidos es de 276.

La páginas que redirigen el tráfico a la web caso de estudio son google.com con un 15% de visitas únicas, Facebook.com con un 11,3%, reddit.com con 3.2%, yahoo.com con un 3.1% y t.co (twitter) con un 2.9%. Po lo general, los buscadores y las redes sociales son los sitios que más tráfico vuelcan a ***mashable.com***

En el análisis del posicionamiento en buscadores se ha utilizado la web semrush.com como herramienta que analiza la web, en este caso los resultados son todos por búsqueda orgánica porque no hay búsqueda por pago. Además hace un análisis de la publicidad del sitio y de las palabras clavé donde se observa que "Facebook" funciona la que mejor.

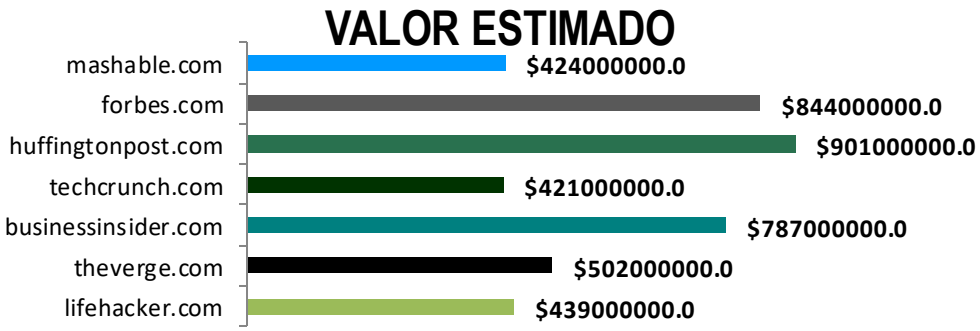
3.2 ANÁLISIS DEL SECTOR EN INTERNET

Hay varias formas de medir la influencia de una empresa en internet, según worthofweb.com y AlexaRank la competencia que tiene una audiencia similar a ***mashable.com*** son medios digitales americanos, como los que se muestran a continuación.

1. lifehacker.com
2. theverge.com
3. businessinsider.com
4. techcrunch.com
5. huffingtonpost.com
6. forbes.com

Para estudiar el sector y la competencia se han generado unos gráficos comparativos que permiten valorar la situación de la web.

3. Gráfico valor estimado



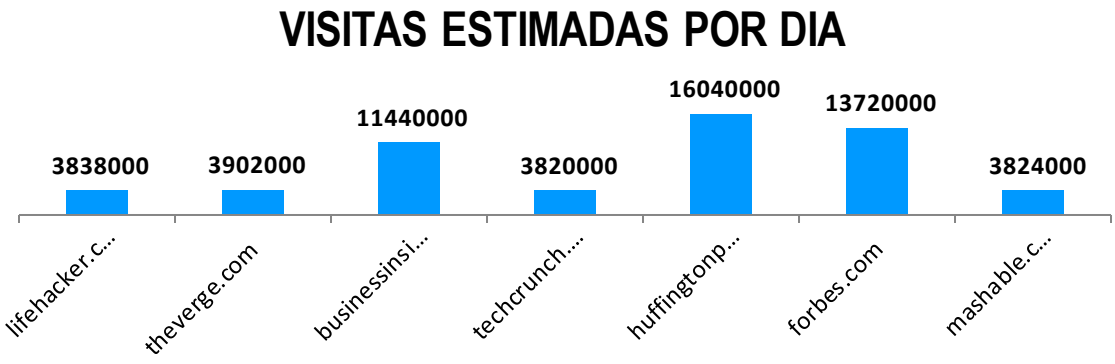
Respecto al **valor de las empresas web** de la competencia hay tres que destacan, huffingtonpost.com, Forbes.com, businessinsider.com, estas son páginas web muy famosas a nivel internacional. Por otro lado quedarían thevergue.com, lifehacker.com, mashable.com, y techcrunch.com siendo esta la de menor valor estimado.

4. Gráfico ingresos al día

En el caso de **los ingresos que reciben al día** cada una de estas empresas, siguen el mismo orden que en el gráfico anterior, con una diferencia, lifehacker.com es el que menos ingresos recibe al día.

Para analizar el tráfico web de la competencia vamos a analizar las siguientes métricas:

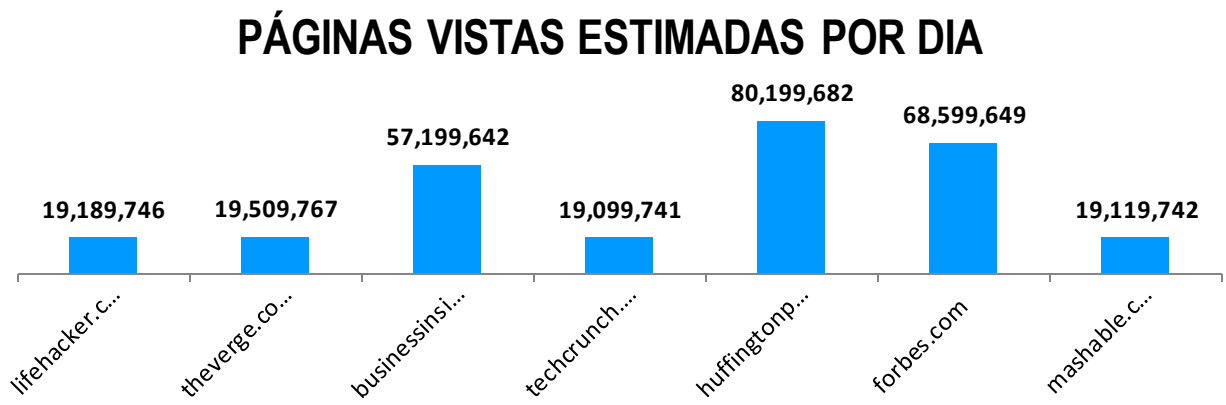
5. Gráfico visitas estimadas



Una visita se produce cuando un usuario entra en la página y envía una solicitud, pero si vuelve a cargarla o navega por ella más tarde también se registrará como otra vista de esa página. Si hablamos de visitas estimadas la web huffintonpost.com destaca con 16.040.000, mientras mashable.com tiene 3.824.000 visitas al día, la penúltima entre estas cinco.

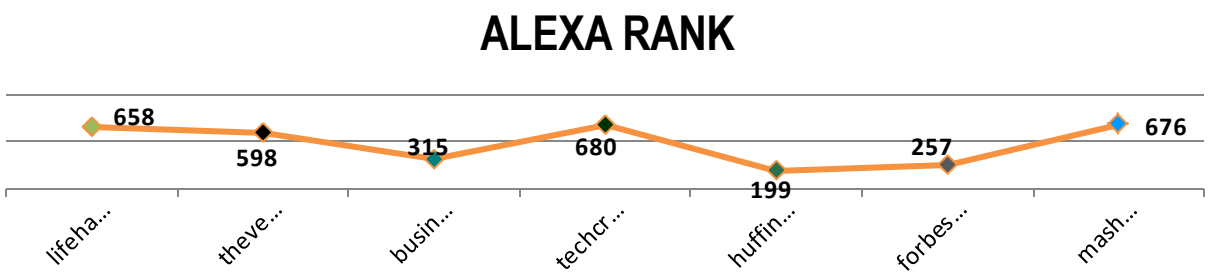
Con **el número de páginas vistas** sucede lo mismo, se contabilizan todas las páginas que se abren en el día. Estas mantienen una relación directa con la cantidad de contenido que se ofrece y la de veces que se abre la página. Aquí mashable.com se posiciona un poco mejor con la cantidad de páginas vistas al día.

6. Gráfico de páginas vistas estimada



El valor de posicionamiento de Alexa Rank se ha consolidado como una herramienta capaz de medir el tráfico de casi todas las páginas y permite hacer un buen benchmarking. El resultado del ranking resulta de dividir los visitantes medios entre el número de páginas vistas.

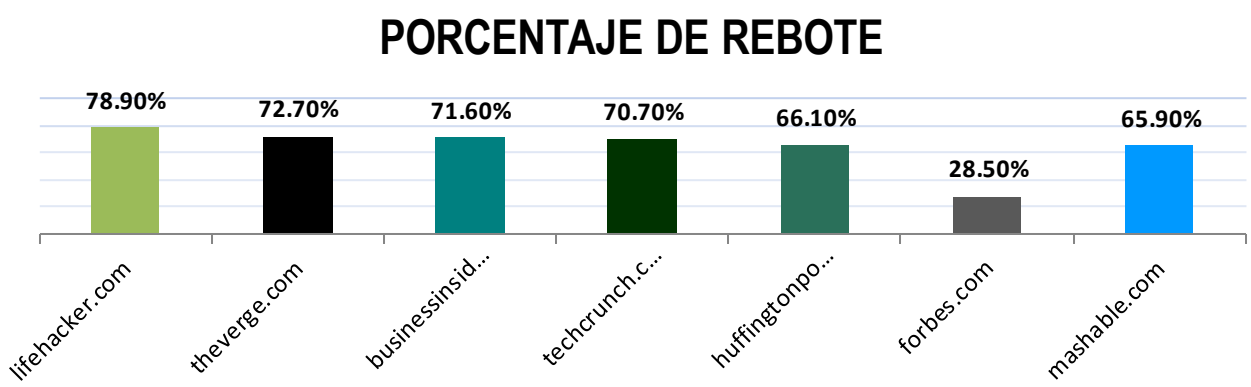
7. Gráfico Alexa Rank



Entre estas empresas la más cercana a #1 es el huffingtonpost.com, esto quiere decir que es la que tiene una mayor combinación entre visitas y páginas vistas en un periodo de tres meses y por ende la mejor posicionada.

El porcentaje de rebote indica la cantidad de veces que se produce una visita única con una sola solicitud en el servidor, la de entrada a la página, en estas sesiones no se puede contabilizar el tiempo ya que no registran ningún hit en el servidor. Tener una tasa de rebote elevada puede ser negativo pero en el caso de ciertas páginas web dónde el contenido se visualiza en una única página es normal que el porcentaje sea elevado, suele suceder en casos de blogs y sitios de noticias.

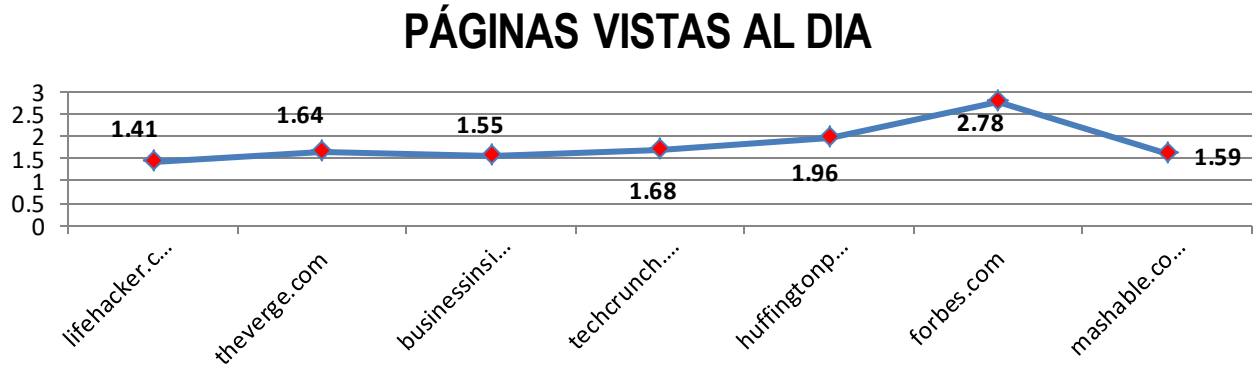
8. Gráfico porcentaje de rebote



En este caso Forbes.com destaca con un porcentaje más reducido, del 28,50%, lo que indica que en su página se producen más interacciones que en las demás. Puede deberse a su diseño, al tipo de contenido o de cómo se ha accedido a él.

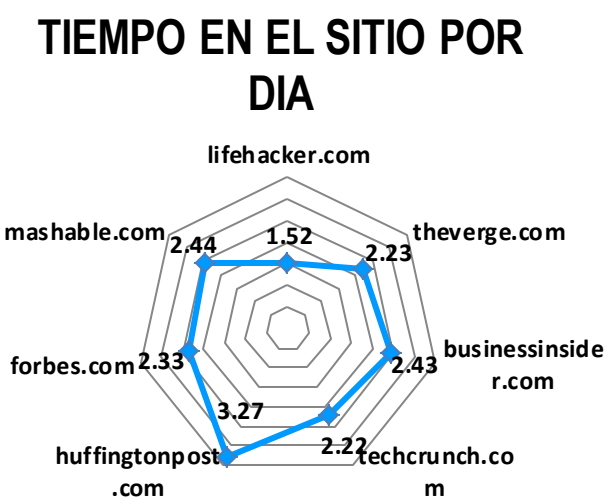
Para Mashable.com el porcentaje es notablemente mayor, con un 65,90% no se posiciona entre las empresas con mayor tasa de rebote. En este caso el acceso a la información suele redireccionar el tráfico web de forma errónea, ya que no consigue que el usuario siga navegando por la web.

9. Gráfico visitas únicas al día



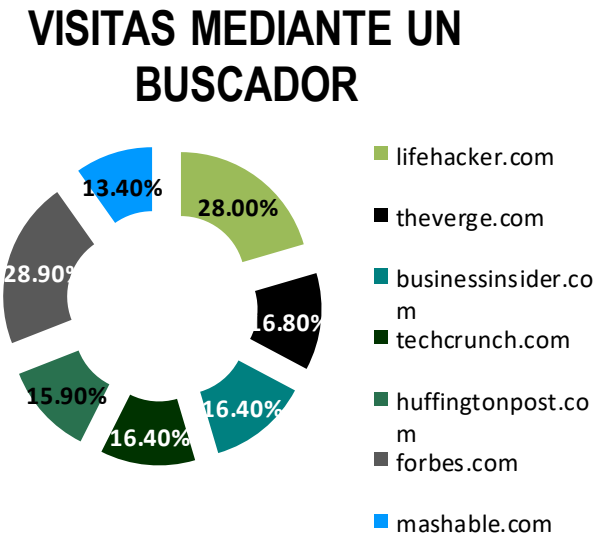
El número de páginas distintas o únicas vistas durante una misma sesión. Como se puede ver Forbes.com tiene un mayor número de páginas vistas únicas, la situación de mashable.com no es tan favorable.

11. Gráfico tiempo en el sitio por día



El tiempo que un usuario pasa en el sitio por día nos indica la actividad de cada web, si se mantienen en el sitio o solo entran y salen. Mientras en mashable.com permanecen 2,44 minutos en lifhacker.com solo lo hacen 1,52 minutos, aunque como en las otras comparaciones el huffingtonpost.com destaca con el mayor tiempo en el sitio con 3,27 que le reporta un mayor retorno (engagement).

10. Gráfico de visitas mediante un buscador



El porcentaje de **visitas que se reciben mediante un buscador** nos puede facilitar una idea general sobre el tráfico web y de cómo accede a la web, en el caso de mashable.com el tráfico de la web se genera más mediante redes sociales que a través de un buscador, por el formato de contenido, sin embargo Forbes.com consigue una tasa alta de visitas a través del buscador, casi igual que de lifehacker.com.

La competencia es realmente fuerte en entornos web, ya que es un sector que se está consolidado por lo que hay muchísimas páginas de todos los tipos. *El posicionamiento de marca en la mente del usuario es muy complicado,* ya que el acceso a tanta información hace muy difícil que te reconozcan y que se conviertan en leads.

3.2 DISEÑO DE ARCHIVOS

Los datos se agrupan en unos archivos conocidos como log files o ficheros de transacciones, que generan información de cómo se usa la web. Estos archivos se crean por el servidor de forma que recogen todas las acciones o peticiones que suceden en un periodo de tiempo en la web.

Al igual que en la analítica tradicional, el análisis web depende de dos tipos de datos, las conocidas como métricas que registran un número, por ejemplo, de acciones en la web que forman una variable cuantitativa, y las dimensiones que hacen referencia a datos cualitativos, por ejemplo la localización o la URL de la página que fue visitada.

Las métricas dan la posibilidad de estudiar la conversión, esto se analiza cuando el usuario ha realizado una acción considerada como importante, por ejemplo que haya compartido la publicación. Los "visitantes o usuarios" son una de las métricas más importantes ya que pueden dar información sobre el tipo de audiencia. También lo son el "porcentaje de rebote" y las "sesiones", tiempo que actúa el usuario ininterrumpidamente.

Aunque como hemos comentado anteriormente, este estudio va más allá de la analítica web. Por lo que las métricas van a ser distintas a las utilizadas comúnmente, al centrarse en un análisis de contenido las características de usuario quedan a un lado para darle importancia a las características del contenido. La primera métrica que se desea resolver es la de distinguir las características que permitan optimizar la redacción convergiendo en unas acciones 'share' óptimas.

Los informes de resultados muestran de forma resumida o visualmente sencilla las conclusiones sobre el funcionamiento de la web, de su audiencia, etc. En la actualidad el nivel tecnológico permite obtener estas métricas en tiempo real, mediante herramientas informáticas se pueden generar datos reales que proporcionen una información estratégica en el momento de tomar una decisión clave.

Por otro lado, la descarga de la base de datos se ha realizado a través de una página web dedicada al aprendizaje automático y a los sistemas de inteligencia, en el UC Irvine Machine Learning Repositorio.

La descarga de la lectura de datos de texto ha sido a través un archivo xls y posteriormente importado a .sav. La base de datos está formada por 39.644 instancias que corresponden con cada una de las páginas web analizadas, y por 61 atributos o variables, está base de datos recoge un total de 94.231 hits, que hacen referencia a la cantidad de archivos, imágenes, sonidos, en general todos los elementos que componen una web y que son solicitados para la correcta lectura de la misma.

Los artículos han sido descargados de la web ***mashable.com***, el contenido ha sido modificado para proteger la propiedad intelectual de los autores, pero los textos completos se encuentran en cada una de las url de la base de datos. Los datos fueron adquiridos por el repositorio el 8 de enero de 2015.

Entre todos los métodos de análisis posibles, en este trabajo se ha optado de forma innovadora con técnicas PLN (procesamiento de lenguaje natural) que es una técnica que ha cogido fuerza en la última década debido a la necesidad de información, sobre cómo interactúa el usuario en entornos web, es un campo de estudio perteneciente a las ciencias de la computación, de la inteligencia artificial y de la neurolingüística. Se basa en el estudio de las interacciones entre el lenguaje humano y el computacional. En la actualidad **estas técnicas se utilizan en Marketing digital** para analizar sentimientos mediante el análisis de textos. Por regla general, este análisis sentimental tiene la finalidad de identificar la objetividad o la polaridad del contenido, por ejemplo, la opinión propia que un escritor tiene en referencia a un tema concreto.

Esta disciplina permite realizar varios tipos de análisis, en este estudio sólo se analizará la subjetividad y la polaridad del contenidos y del título, pero se puede observar que estas no son las únicas. Existen 5 tipos, Clasificación de la subjetividad, Clasificación de la polaridad, Clasificación de la intensidad, Análisis sentimental basado en tópicos/características y Minería de opiniones. Con frecuencia se utilizan de forma conjunta para hacer un análisis de sentimiento más acertado.

La clasificación **de la subjetividad**, es capaz de mostrar si un texto posee la opinión del redactor, es decir, encuentra una carga subjetiva en el contenido del texto. En este estudio también se le asigna una carga positiva, negativa o de no opinión a la subjetividad que expresa el texto, asignando un valor porcentual de la intensidad de la

opinión. Por otro lado, la clasificación de **la polaridad** expresa si un texto está redactado con una carga emocional positiva o negativa, donde la carga positiva en el texto va indicada con un rango entre 0 y 1, mientras que la carga negativa se mueve en un rango de -1 a 0, siendo 0 una carga neutral. Esta métrica se puede ver influida por formas gramaticales como los modales, los negadores y los cuantificadores.

3.2. CARACTERÍSTICAS Y ELECCIÓN DE VARIABLES

Planteado el entorno empresarial y contextual del análisis, se va a realizar un desarrollo metodológico con el objetivo prioritario de optimizar contenidos. Es decir, en cualquier artículo deberemos acometer si es adecuado para la población objeto de interés de los artículos publicados en Mashable.com **entre 2013 y 2014**.

El tamaño de la muestra es de 39644 casos que corresponden a los artículos publicados. La información más relevante para establecer un periodo temporal sobre la base de datos viene marcada por las fechas de publicación de artículos. Respecto al número total de artículos de la muestra, 18199 corresponden al año 2013 y 21445 al año 2014.

El número de variables recogidas en la base de datos es 61, de las cuales 58 son de carácter predictivo, 2 no predictivos y 1 objetivo. La variable objetivo corresponde con el número de veces que se ha compartido el artículo de donde se sacará la previsión de la popularidad.

Las variables que se han recogido en la base de datos explican la tipología del artículo, desde la el número de palabras que posee el título o el contenido, hasta el número de archivos web que posee el artículo, como el número de imágenes, de enlaces, de vídeos u otros elementos. Las palabras se han clasificado según la representación positiva o negativa del contenido. La optimización del tráfico web mediante buscadores ha recaído sobre la variable número de palabras clave en los metadatos que nos ha proporcionado información SEO a nivel redacción de artículo.

Entre las variables que clasifican a los artículos se encuentran los días de la semana que han sido publicados y el tema sobre el que tratan.

El Análisis estadístico descriptivo *nos proporcionará la información suficiente para comprender la estructura de los datos, lo que nos permitirá especificar y plantear el*

modelo. (Véase anexo 1: tabla estadísticos descriptivos) Este modelo servirá para cumplir con el objetivo final facilitando la toma de decisiones, por lo que el AED nos proporcionara la información suficiente para saber diseñar un modelo que prevea la viralización de los artículos.

Además de las variables dadas por la propia página se han creado variables auxiliares para cuantificar si el artículo pertenece o no a alguna de las clasificaciones dadas por la base de datos y/o a alguno de los canales generados por ellos.

ANÁLISIS DE LOS DATOS

Para facilitar el análisis y completar estas clasificaciones, ha sido necesario recodificar en diferentes variables, entre ellas las que hacen referencia al tema del artículo y se pudo observar la falta de identificación de otros canales que también existen en la web, así se generó una variable nueva (canal=otros).

Por otro lado, también se ha decidido recodificar el grado de polaridad y subjetividad, gracias a las técnicas PNL se ha podido clasificar en diferentes grados: polaridad positiva, negativa o neutral mientras que la subjetividad del texto se ha clasificado según el grado de opinión, No opinión, opinión fuerte y opinión débil, ambas de estas variables se han generado para el título y para el contenido.

Además para facilitar el análisis predictivo la variable número de veces que se ha compartido un artículo se ha transformado en una variable binomial, para estimar si un artículo es popular se ha optado por seleccionar la mediana (50% de acciones), convirtiéndose en 0 todos los artículos que reciben <1400 acciones y 1 para los que sean igual o superior a este número.

4.1. GRÁFICOS SECTORIALES DE VARIABLES CUALITATIVAS:

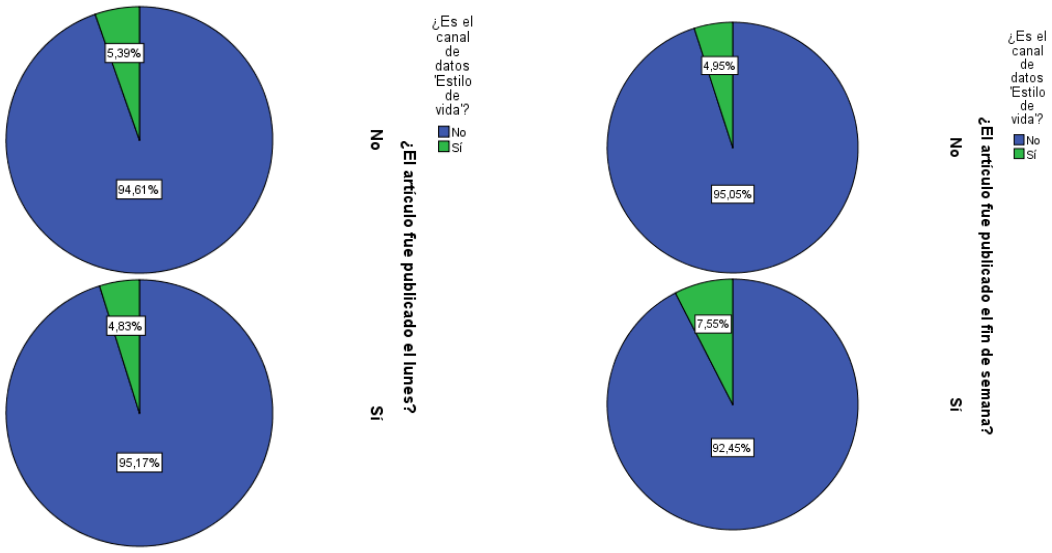
Una de las cuestiones más importantes en un exploratorio de datos son los gráficos sectoriales de las variables categóricas y/o cuantitativas relacionadas, concretamente estamos interesados en relacionar el tipo de canal con el día de la semana que ha sido publicado.

Se están cruzando las variables que indican el tipo de canal o temática de los artículos con el día de la semana que fue publicado. Se han generado cada uno de los cruces en

gráficos sectoriales para explicar su relación. *(Véase anexo 2: gráficos sectoriales de dimensiones)*

CANAL "ESTILO DE VIDA":

12**. Gráficos sectoriales "Estilo de vida"**



De todas las relaciones estudiadas se ha verificado que los publicados el fin de semana aumentan un 3% de artículos publicados sobre estilo de vida con respecto al resto de la semana. Tras analizar los gráficos sectoriales se ha decidido ampliar la información cruzando ambas variables de forma que se consiga entender la relación observando los residuos. En este caso se observa una clara diferencia entre días. En este caso se ve que los residuos corregidos en el artículo publicado el domingo son de 5,8 mientras que para el lunes son -1,8.

CANAL"ENTRETENIMIENTO"

El canal entretenimiento destaca en relación al día de su publicación el lunes, aunque el domingo también recibe una gran significatividad. ***Con una diferencia de más de 4 puntos porcentuales respecto a los demás día.*** *(véase gráfico sectorial en anéxo ENTERTAIMENT: 2. Gráfico sectorial "Entretenimiento")*

CANAL"NEGOCIOS"

Destaca en sus publicaciones a comienzo de la semana, los lunes y los miércoles se publican más artículos sobre negocios. Existe una gran diferencia entre las

publicaciones que se hacen durante la semana y al final de la semana, de hasta un 8%. (véase gráfico sectorial en anéxo BUSINESS: 3. Gráfico sectorial "Negocios")

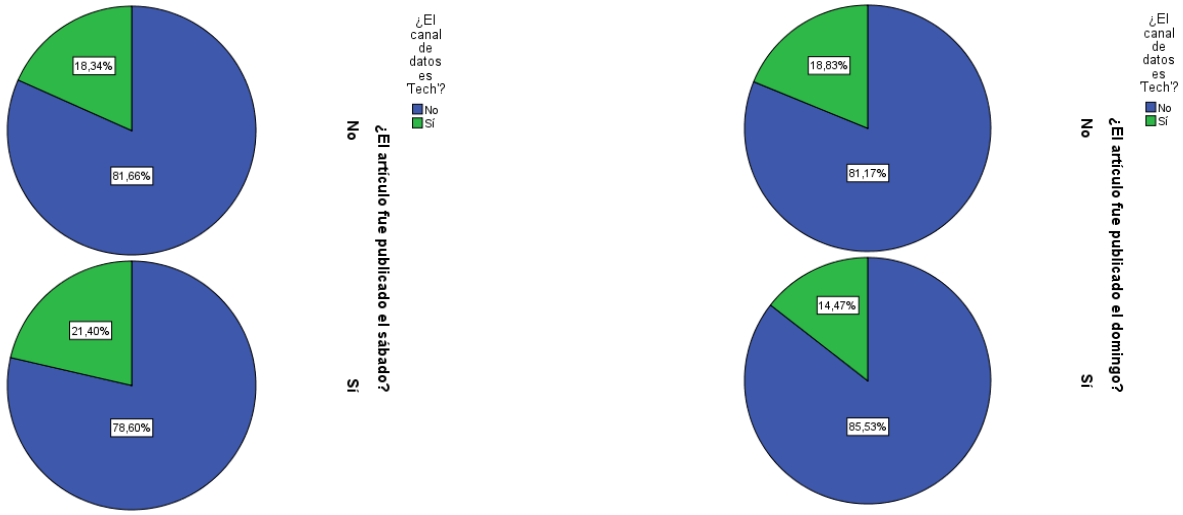
CANAL "MEDIOS SOCIALES"

El martes, el jueves y el sábado se suelen publicar más artículos sobre los medios sociales, lo que sucede en las redes sociales y lo que es tendencia. Tiene sentido que varias veces por semana se actualice este tema y se publique en relación al transcurso de la misma. Aquí las diferencias no son tan notables, pues la relación no difiere en más de 3 puntos porcentuales en ningún caso. (véase gráfico sectorial en anéxo SOCIAL MEDIA: 4. Gráfico sectorial "Social Media")

CANAL"TECNOLOGÍA":

Los martes, miércoles y sobre todo los sábados son los días que más se publica sobre tecnología. Con una gran diferencia de hasta 6% el domingo, los demás día de la semana esta diferencia no es tan notable.

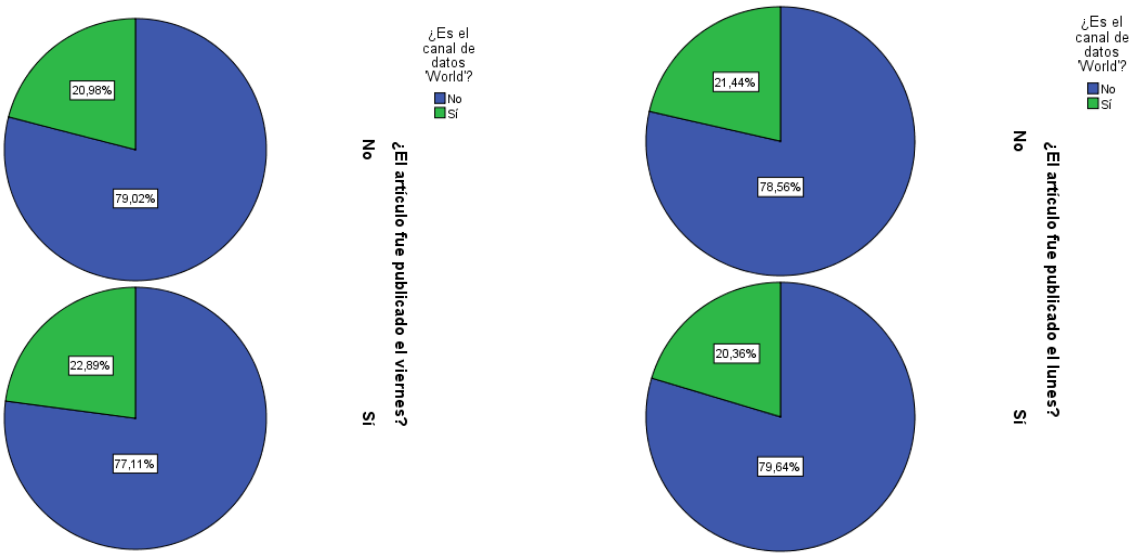
13. **Gráfico sectorial "Tecnología"**



CANAL"MUNDO":

El miércoles, el jueves, el viernes, el sábado y el domingo se publican más noticias sobre el mundo, aunque entre ellos el viernes destaca. Realmente la diferencia no es muy elevada entre la semana puesto que las noticias que suceden en el mundo han de ser publicadas cuanto antes mejor, la diferencia más significativa es apenas el 2% los lunes, los demás días se diferencian en un 1% o menos.

***14*. Gráficos sectoriales "Mundo"**



CANAL "OTROS":

Los artículos que no corresponden a ninguna categoría establecida por la base de datos y que se han clasificado en el tema "otros", que en la página web aparece como "more", suelen ser publicados los fines de semana, mayormente los domingos, existe una diferencia de hasta 7 puntos porcentuales con los primeros días de la semana. (Véase el anexo CANAL OTROS: 7. Gráficos sectoriales "Otros")

Como se observa en el resumen de estas tablas de contingencia hay diferencias muy claras entre publicar un sábado o un lunes sobre tecnología. Esto sucede en todos los casos en los cuales se han realizado dichas tablas para conocer realmente la implicación de la fecha de publicación con el tema de cada caso.

Esta observación pretende optimizar el día de la semana en que se publique cierto tema, para saberlo hay que relacionarlo con la variable objetivo, así podremos saber qué tipo de articulo será más popular por cada día de publicación.

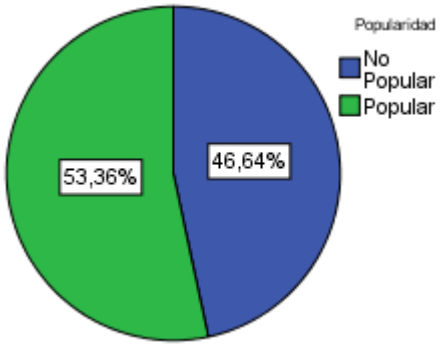
Todas las empresas que se dedican al marketing digital deben de analizar el comportamiento de sus usuarios frente a las diferentes fechas de publicación, pero también lo han de hacer respecto del tema que trata el artículo, ya que un artículo sobre social media o estilo de vida tiene más sentido que sean publicados en fines de semana,

mientras que un artículo sobre negocios no tiene sentido que se publique los fines de semana, ya que durante el tiempo de ocio a poca gente le apetece leer sobre su trabajo.

El análisis de considerar a un artículo popular se establece mediante una hipótesis condicionada, la variable objetivo recodificada en una categórica donde la hipótesis nula son los artículos que no superen las 1400 acciones. Y la hipótesis alternativa de que un artículo sea popular siendo mayor o igual a 1400 acciones.

Este gráfico muestra el comportamiento de los artículos de la base de datos en cuanto a la popularidad de la muestra.

15. Gráfico de popularidad



Pero si deseamos optimizar la publicación teniendo en cuenta el tema y el día que se publique necesitamos comparar el número de veces que más se comparte durante la semana y sobre un tema en concreto. Mediante una tabla dinámica se ha conseguido reunir la información para facilitar y comprender este análisis.

En la tabla multicanal-día, el color verde representa el número medio de acciones más alto para cada tema durante la semana, mientras que el azul muestra las fechas más frecuentes de publicación por temas, como vemos hay días que estos colores coinciden, pero como ya ha sido explicado anteriormente el color verde tiene mayor importancia. *(Véase anexo 6: optimización de fecha de publicación. tabla 2.)*

De aquí podemos concluir que el canal **estilo de vida** recibe el máximo de acciones en **lunes**, mientras que el canal **entretenimiento** lo recibe en **domingo**. El canal **negocios** lo recibe el **sábado**, el canal **social media** está optimizado en **domingo**, el canal **tecnología** lo recibe en **domingo**, el canal **mundo** está optimizado en **sábado** y el canal **otros** se optimiza en **lunes**.

4.2. ANÁLISIS FACTORIAL: REDUCCIÓN DE LA DIMENSIÓN.

En estas bases de datos tan complejas donde la mayoría de variables están definidas de forma similar, como sucede con el número de palabras clave o de imágenes, va a

ser relevante simplificar el análisis definiendo factores que engloben a las variables más correlacionadas entre sí.

Mediante un análisis de los componentes principales que expliquen mayor variabilidad se calcula la matriz de cargas rotadas para extraer las cargas que recibe cada factor, *si la correlación es muy alta en ese factor éste vendrá definido por las variables implicadas en la correlación. (Véase anexo 4: factorial y matriz de componentes rotados)*

La rotación del factor se realiza por el método Varimax, que intenta que cada factor tenga cargas muy altas en valor absoluto o nulas para cada una de las variables, *maximizando (varimax)* la suma de las varianzas de las cargas factoriales al cuadrado.

El determinante de la matriz nos indicará la calidad del ajuste de las cargas factoriales, este debe positivo. En este caso **el determinante es: 0,005** muy cercano a cero, pero adecuado para un modelo factorial. Se han realizado varias pruebas utilizando diferentes conjuntos de variables el determinante en todos los casas era muy pequeño, hay que tener en cuenta que en este modelo las correlaciones parciales son altas, como lo demuestra los coeficientes de la matriz de correlación anti imagen (hay que tener en cuenta que las mayoría de las variables hacen referencia a una idea común).

La varianza explicada por los 7 factores del modelo nos aportará una visión de la cantidad de información que explica el modelo, en este caso es de **58,51%**

La media de adecuación de adecuación de **Kaiser-Meyer-Olkin** es un índice que compara las magnitudes de los coeficientes de correlación observados a las magnitudes de los coeficientes de correlación parcial, resulta de dividir el porcentaje de varianza explicada por el modelo factorial entre el porcentaje de varianza total explicada. Cuanto más cercano a 1 mejor ya que se explicaría mayor varianza. En este caso hay un KMO **(0,521 BAJO)** debido a que casi todas las variables funcionan de forma similar.

16. Tabla KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,521
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	206363,754
	gl	210
	Sig.	,000

Los factores que mejor explican la variabilidad de los datos son:

1. **Contextualización/marco del contenido:** Este factor explica la forma en la que el contenido está redactado, hace referencia a las variables subjetividad del texto, al promedio de polaridad de palabras positivas, a longitud de las palabras, al ratio de palabras positivas y en general al sentimiento de polaridad del texto.
2. **Numero de hits (archivos en web):** Entendiendo como archivos web el número de enlaces o enlaces referenciados, de palabras en el contenido y de imágenes.
3. **Ratio de palabras negativas en el contenido:** este viene definido por una única variable.
4. **Promedio de la peor palabra clave:** es el resultado de las variables promedio de palabra claves y de la peor palabra clave.
5. **Nivel de polaridad absoluto en el título:** También explicado por una única variable, añade importancia al nivel sentimientos en redacción del título de cada artículo.
6. **Número de palabras clave en los metadatos:** Formado por una variable, nos indica la cantidad de palabras que enlazan el tema del artículo con el diseño de la página web y el buscador.
7. **Número de vídeos.** Una única variables compone la carga del factor, este elemento web recibe tanta carga o importancia dentro del modelo que adquiere un factor propio para expresar su relevancia. En la actualidad se consume muchísimo contenido en formato multimedia, solo hay que comprobar la importancia que tiene YouTube como aplicación multicanal.

4.3. CASOS ATÍPICOS (ESTUDIO DE OUTLIER):

Una vez analizada la estructura de los datos, y clasificada en factores se buscarán casos que respondan a comportamientos atípicos dentro de cada factor. Para explicar estos casos debemos establecer un tipo de artículo estándar para ver como difieren estos casos peculiares.

Las características del artículo medio o estándar corresponden con las de mayor frecuencia de aparición en las variables estudiadas, es decir, se generaliza para conocer cuál es la tipología de una publicación media.

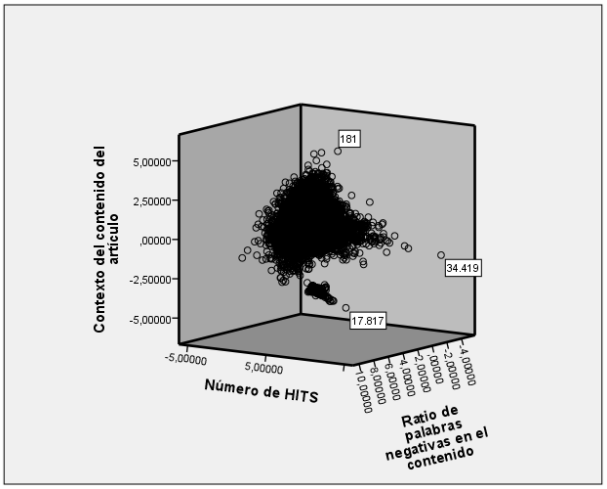
En un artículo estándar hay 10 palabras en el título, y una media de palabras en el contenido de 545,51. Se establece que se incorporan 4 enlaces en el contenido de los cuales 2 suelen ser referenciados a mashable.com, por lo general hay 1 imagen y 0 vídeos. La longitud media de las palabras en el contenido es de 4,5. El número de palabras clave en los metadatos es de 7.

Es más frecuente que los artículos publicados hablen sobre el tema "Mundo" y que hayan sido publicados el miércoles. Además suelen estar redactados con un grado de subjetividad bajo en el contenido, es decir hay una opinión en la redacción pero no es demasiado abusiva, esta opinión generalmente es positiva, ya que el grado de polaridad positiva en el contenido recibe un 88,7%.

El título de estos artículos tiene un grado de subjetividad nulo, ya que en el 45,4% de los casos no hay opinión y con una polaridad neutral en el 50,2% de toda la muestra, aunque destaca la polaridad positiva sobre la negativa. *(Véase los resultados del análisis en el anexo 5: estudio de los outlier, 1. tabla de frecuencias)*

La búsqueda de anomalías se basa en desviaciones en las normas de agrupación, en el caso de los factores se examinaran los casos atípicos por si interfieren de los demás y para distinguir su exclusividad. Este análisis se realiza mediante gráficos de dispersión en función a un modelo regresivo. Hay que tener en cuenta que los atípicos pueden venir dados por errores de procedimiento, por a valores extremos, o por causas no conocidas. Pero es importante tener en cuenta estos supuestos ya que pueden distorsionar el análisis.

17**. Gráfico atípico varianza máx.**



Estudiando los casos atípicos mediante los factores, por la reducción de la dimensión podremos ver los casos que más se alejan de este artículo estándar.

Varianza máxima de los factores: 3 primeros

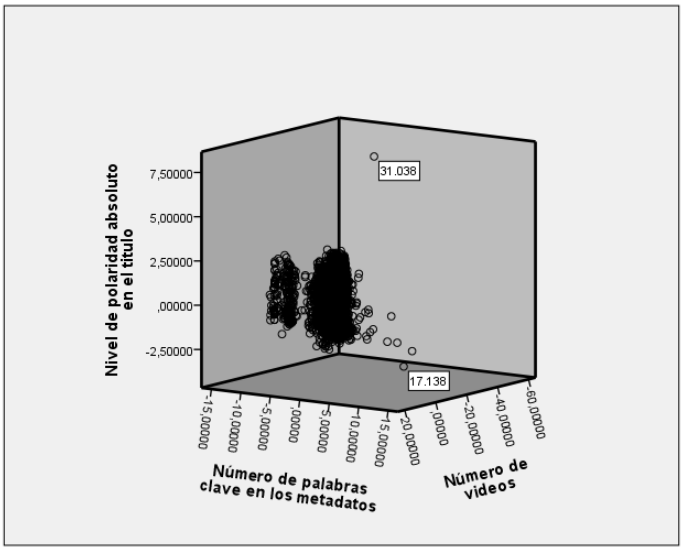
Caso: 34419: "Six People Under Ebola Observation in Madrid Hospital" - Exceso de hits: Posee 119 enlaces y 15 imágenes. Muchos enlaces por toda la repercusión internacional el ébola en Madrid.

Características

La redacción del artículo no es subjetiva ni está muy polarizada, tiene un tono neutral. En general el ratio de palabras negativas no es elevado. Se hace atípico debido a la cantidad de enlaces a los que está vinculado, son noticias que complementan la información de este artículo.

Como se ve en el grafico **hay un grupo significativo de artículos** que se sitúan por debajo del grupo principal, estos artículos comparten características similares, en general reciben una carga en el factor 1: contexto muy bajo, también sucede con el tercer factor, el de ratio de palabras negativas en el contenido, por otro lado si entramos a la página concreta podemos observar que comparten un diseño similar, hay muchos artículos con infografías o imágenes explicativas. Alguno de estos casos son: 894, 918, 1063, 38798, 20182, 1575, 1455, entre otros.

18. **Gráfico atípicos varianza mín.**



Varianza mínima de los factores:

Caso 31038: "Civilians Fleeing Rebel-Held City in Ukraine Are Attacked" - Exceso de carga emocional en el título y defecto de número de videos: La noticia sobre un ataque a ucrania, solo tiene texto y un par de archivos web: Tiene 7 palabras clave en las metadatos y ningún vídeo.

Características:

9 palabras en el título, 1570 palabras en el contenido, con 11 enlaces donde 10 de estos son referencias a artículos de la propia página, posee 51 imágenes en toda la página, 7 palabras clave, y 0 vídeos. Al ser una noticia sobre hecho importante a nivel global se enfatiza mucho en una carga emocional en el título que destaca sobre los demás casos.

En este gráfico también se puede apreciar **una nube de puntos situada a la izquierda** del grupo principal, se observa como cada una de esos puntos varía en función del nivel de polaridad del título, y del número de vídeos, aunque en este sentido varía menos, pero comparten las misma características respecto al factor 6 que depende del número de palabras clave, en general estos artículos tienen pocas palabras en las metadatos.

Estos dos artículos entre varios casos pueden considerarse atípicos, ya que sobresalen en el mapa de dispersión e indican que difieren del resto.

4.4. MODELO DIAGRAMA DE ÁRBOL

El diagrama de árbol es un método basado en el teorema de Bayes (1763), donde los nodos representan los sucesos y las ramas sus probabilidades. A través de las medias de la variable objetivo se generará la situación óptima que permita formular una estrategia de redacción. Por lo que consiste en determinar, a partir de un modelo estimado, el valor que toman las variables endógenas para un valor dado de la variable exógena.

Con los factores y la variables cualitativas como variables independientes y con la variable objetivo como dependiente se genera un árbol de decisiones basadas en la probabilidad condicionada. El árbol se ha generado mediante tres fases de profundidad.

Del análisis de diagrama de árbol obtenemos los siguientes resultados, se han generado 92 nodos, también entendidas como situación de decisión. Como el objetivo es optimizar el número de veces que un artículo ha sido compartido, se utilizará la variable popularidad. *(Véase anexo 7: modelo de decisión predictiva, diagrama de árbol (modelo explicativo)*

19. Nodo ganador

Nodo 41

Categoría	%	n
No Popular	22,9	270
Popular	77,1	910
Total	3,0	1180

El primer nodo en el resultado de ganancia es **el 41**, este recibe el mejor porcentaje de clasificación predictiva como popular, con un 77,1%

Este nodo está compuesto por otras dos tomas de decisiones, que corresponden con las condiciones de los factores **Promedio de la peor palabra clave sea superior a un 0,8622, y de que Número de hits sea superior a 0,47.**

El siguiente nodo es **el 90** con tres condiciones, corresponden al factor **Promedio de la peor palabra clave se encuentre entre (0,157; 0,862), el factor Número de hits ha de estar entre (-0,234, -0,113). También el factor contexto del contenido** situándose un valor superior a -0,458.

El tercer nodo de decisión que resulta ganador es **el 40** y está condicionado con que la variable/factor Promedio de la peor palabra clave sea superior a un 0,862, y que el factor número de hits sea encuentre entre (-0,414; 0,476).

Hay una gran diferencia entre el porcentaje de respuesta del primer nodo de 77,1% al segundo con 71,9%, entre el resto de nodos hay menos diferencia a cada paso.

En resultados generales se han clasificado correctamente el 61,7% de los artículos, mientras que el 77,2% de los artículos populares se han clasificado de forma correcta. Pues la bondad del ajuste es buena como se ve en la tabla de riesgo, con un error típico de 0,002.

20. Tabla de ganancias para los nodos

	Nodo		Ganancia			
Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
41	1180	3,0%	910	4,3%	77,1%	144,5%
90	701	1,8%	504	2,4%	71,9%	134,7%
40	1561	3,9%	1089	5,1%	69,8%	130,7%

4.5. COMPARATIVA DE MODELOS PREDICTIVOS

4.5.1 Modelo discriminante de Fisher

Con todas las variables cuantitativas se hace un análisis que prevé la popularidad mediante la Función discriminante de Fisher. Este modelo simula un análisis de regresión mediante una variable categoría dependiente, y variables cuantitativas independientes para crear grupos condicionados. En este caso la variable dependiente o discriminante minimiza la probabilidad de que la hipótesis nula sea cierta, es decir, que un artículo sea "No popular". (*Para comprobar los resultados del análisis véase anexo 7: modelo de decisión predictiva, discriminante de Fisher, prueba 1. Variables cualitativas)*

El estadístico lambda de Wilks es el cociente entre la suma de cuadrados dentro de los grupos y la suma de cuadrados intra-grupos e inter-grupos. Esto implica una normalización de las variables mediante una distribución de Wilkis basado en un análisis multivariado a través de la hipótesis nula No popular, y la hipótesis alternativa que un artículo sea Popular, si lambda es muy elevada significa que la variable dependiente no discrimina mucho.

En los resultados de clasificación de la primera prueba se ve que el modelo predice un 57,9% de artículos "Populares". Mientras consigue pronosticar correctamente el 61,4% de los casos agrupados originales.

En el resultado de la prueba de discriminación realizada con los factores se obtiene un porcentaje del 51,4% de los casos "Populares", atendiendo así a un 58,5% de los casos clasificados correctamente. *(Véase anexo 7: modelo de decisión predictiva, discriminante de Fisher, prueba 2. Análisis discriminante mediante factores)*

Los resultados de ambas pruebas reflejan una mayor precisión en el modelo predictivo realizado con todas las variables de la base de datos mediante un análisis discriminante paso a paso.

En comparación, cuando se toman las variables los primeros pasos se centran en cuestiones de Palabras clave, haciendo referencia al posicionamiento del buscador como las variables que mejor popularizan la muestra. Mientras que cuando se realizó la prueba con factores se vio que seguía teniendo peso las palabras clave pero rápidamente entraban factores como número de archivos, contexto y polaridad del contenido.

Hay que añadir que en este contraste los resultados de las pruebas con los factores empeoran la precisión de la previsión, esto se debe a la correlación existente entre estos factores.

4.5.2. Modelo Logarítmico Binomial.

La prueba realizada de regresión sobre la variable cualitativa "Popularidad", busca las variables influyentes en el aumento o disminución de la probabilidad de que un artículo sea popular. Las variables que componen el modelo responden a las características de

un artículo web. En este caso el análisis se realiza mediante los factores y las variables cualitativas que nos indican el tema o canal del artículo. La regresión de pasos hacia delante, es decir, que a cada paso se le añaden variables significativas, el primer paso se compone de la variable más importante del modelo.

El modelo se ha generado con 10 pasos, el análisis de **la prueba ómnibus** indica la significatividad de cada paso que realiza el modelo, en este caso cada uno de los pasos es significativo. *(Véase anexo 7: modelo de decisión predictiva, regresión logística binaria, prueba 1.)*

La prueba de Hosmer-Lemeshow nos indica el p-valor del ajuste del modelo, por lo que sí es inferior a 0,05 el ajuste es malo, en este caso se ve como en el paso 3 cambia significativamente.

La tabla de clasificación nos muestra el porcentaje de ajuste de cada paso, prediciendo en el último paso que el 69,6% de los artículos serán compartidos de forma popular. Mientras que el modelo consigue prever el 63,2% de la muestra global.

En el análisis de las variables de la ecuación hay que destacar la influencia de los odds ratio que aparecen en esta tabla en la columna exp (B) pues nos indican la probabilidad de que suceda el caso Popular. La exponencial de Beta es una estimación de lo que se define como ODDS ratio condicional. Este ratio se puede definir como el coeficiente entre la probabilidad de que ocurra la condición de la variable dependiente en la probabilidad de que no ocurra. Los coeficientes nos indican el grado en el que se produce la previsión, por ejemplo en el modelo destaca en este sentido si el canal es Socia Media, ya que este valor indica que la probabilidad es 2,367 veces mayor en el cociente de que sea "Popular" al cociente de que sea "No Popular" siendo artículos del tema "Social Media".

El modelo no consigue aportar mucha fiabilidad en la predicción, por ello se ha decidido realizar una segunda prueba en la que **se le añadan las variables de día de la semana**.

A continuación se realizan las pruebas las de la regresión con más variables cualitativas, para optimizar la predicción. En este caso el modelo se genera en 13 pasos todos

significativos, como muestra **la prueba ómnibus.** *(Véase anexo 7: modelo de decisión predictiva, regresión logística binaria, prueba 2.)*

La prueba de Hosmer-Lemeshow indica un ajuste bastante malo como en el caso anterior, el paso 4 cambia significativamente conforme la chi-cuadrado aumenta. La tabla de contingencia de esta prueba también muestra las diferencias entre lo observado y esperado, a partir de ese paso se convierten en significativas.

Como resumen del modelo se genera la tabla de clasificación, en este caso el último paso predice la popularidad de un artículo con un 71%, aunque esto ya sucedía en pasos anteriores, lo único que gana con cada paso es explicar más el porcentaje global del modelado consiguiendo un 64,5%. En este modelo **la función Exp (B)** se comportan del mismo modo, duplicando la probabilidad de ser popular frente a los que no lo son.

Podemos afirmar que si incluimos más información al modelo se obtienen mejores resultados de aproximación en la predicción, por lo que se piensa que entre el **análisis Discriminante de Fisher y de regresión logística no lanza una predicción muy precisa**, el causante en este sentido podría ser la reducción de la dimensión no siendo óptima, pero en bases de datos tan complejas y estructuradas es difícil de conseguir unas características tan definidas para un concepto tan amplio como es la redacción de artículos digitales.

4.6. CONCLUSIONES DE LA OPTIMIZACIÓN

Haciendo referencia al análisis de las variables cualitativas que aportan una dimensión al caso de estudio se concluye que **el fin de semana se publica más aunque depende del tema,** por otro lado en el análisis de las acciones que reciben diariamente cada tema de artículo cabe destacar que **de forma paralela los artículos reciben mayor reacción durante el fin de semana,** aunque **el lunes también recibe un buen impacto**. Esto es lógico, pues las acciones se producen en un periodo reducido de tiempo tras la publicación. Esta previsión de la publicación y de conocer cuándo se da la respuesta del propio lector es muy recomendable en todos los medios de comunicación.

Tras los resultados factoriales se conoce que la reducción de la dimensión limita un análisis predictivo debido a que la información entre factores esta correlada. Así se puede ver en el análisis discriminante mediante factores, el cual obtiene un resultado de

precisión menor al resultado generado de todas las variables. Es decir que la base de datos está bien estructurada aunque guarden mucha información de una misma idea u objeto de investigación.

En relación al análisis del diagrama de árbol se obtiene que es más probable que los artículos que mantienen la condición del nodo 41 reciban un mayor impacto que otras combinaciones. **Es decir que si redactamos un artículo con un Promedio de la peor palabra clave sea superior a un 0,8622, y de que Numero de hits sea superior a 0,47,** se prevé que tendrá una mayor reacción de acciones que con otras combinaciones, siendo esta **la más óptima.** Pero a través de este árbol se pueden analizar otros supuestos y saber la reacción que esto generaría en los usuarios. El modelo de diagrama de árbol optimiza la redacción de un artículo para que sea popular, obteniendo una predicción del 77,2% de casos considerados populares.

El uso comparativo de distintos modelos de predicción nos ayuda a conocer y seleccionar con mayor precisión los resultados. En Esta tabla se muestran las pruebas realizas y los resultados de clasificación obtenidos.

21. Tabla resumen de los modelos predictivos

MODELOS PREDICTIVOS	CLASIFICACIÓN GLOBAL	CLASIFICACIÓN POPULAR
DISCRIMINANTE DE FISHER		
Variables cuantitativas	61,4%	57,9%
Factores	58,5	51,4%
REGRESIÓN LOGÍSTICA BINOMIAL		
Factores + tema	63,2%	69,6%
Factores + tema + día	64,5%	71%
DIAGRAMA DE ÁRBOL	61,7%	**77,2%**

El análisis de estos contenidos ayuda a los redactores a prever si un artículo con ciertas características será popular o no. Esta **relación entre la tecnología y el marketing** ayuda a las empresas a conocer mejor el tipo de contenido que interesa entre sus lectores y la explotación de los datos puede facilitar a la toma de decisiones estratégicas. En el caso de esta empresa el análisis aporta **un valor añadido al**

departamento de marketing digital facilitando la herramienta perfecta para **crear artículos virales**.

En general, **mashable.com** es un sitio web que puede optimizar sus contenidos a través de la consideración de ciertas características del artículo; de esta forma puede prever si un artículo va a generar un gran impacto en los lectores o no, y así priorizar contenidos y publicarlos estratégicamente. Ayudando así a la viralización de artículos en redes sociales.

La herramienta propia de esta web, **Velocity está basada en un algoritmo generado mediante un análisis de este tipo**, de los que prevén la popularidad a través de las características de los artículos y de su redacción. Por lo que la tecnología nos abre un mundo de posibilidades de optimización siempre teniendo un objetivo de mejora que ayude a la experiencia de usuario.

CONCLUSIONES, LIMITACIONES / RECOMENDACIONES, TRABAJOS FUTUROS

En este trabajo se ha estudiado y analizado la relación entre los temas de interés publicados en artículos de investigación, así como el impacto de las noticas según tema y día de publicación. Así pues, hemos visto la relación entre distintos tipos de canales y su fecha de publicación que consecuentemente nos ha llevado a estudiar los residuos significativos positivos en las tablas de contingencia que indican el día de la semana en el que más se publica dicho canal. Componiéndose una secuencia semanal de publicación para cada tema.

El impacto que tiene la noticia evidentemente estará asociado a las visitas en los días posteriores, como se ha visto en las tablas multicanal. El domingo es el día que más impacto recibe, aunque en general el sábado y lunes también son significativos. Por lo que concluimos que el óptimo de acciones se realiza uno o dos días después de la publicación.

Finalmente es de destacar que en el ámbito del marketing e investigación de mercados las políticas de estudio de marketing de contenidos deben cada día implementarse más en los departamentos de marketing, ya que uno de los ejes centrales de las políticas novedosas en las empresas debe ser estudiar el resultado de las acciones y/o opiniones de sus clientes. En el caso de mashable.com han sabido aprovechar el modelado predictivo para optimizar sus publicaciones y con ellos sus resultados.

Como hemos podido ver, el marketing y la tecnología guardan una relación cada día más importante, esto se muestra en un estudio realizado por la Universidad de Illinois, donde las empresas que se centran en la tecnología logran tener ideas más innovadoras, en cambio las empresas que se centran en las necesidades de mercados suelen obtener ideas más comunes. Por lo que es altamente recomendable que el marketing evolucione al mismo paso que lo hacen las tecnologías, y con ello el estudio de las necesidades de los usuarios, en este caso, de los lectores.

LIMITACIONES / RECOMENDACIONES

Recomendaría el uso una base de datos estructurada en modelos predictivos siempre que se mantenga toda la información posible, pues como hemos visto, la reducción de la dimensión empeora los resultados predictivos. Cuanta más información explicativa

mejor, así el propio desarrollo del trabajo final de grado y la extensión del mismo ha dificultado la explicación visual.

TRABAJOS FUTUROS

Hubiera sido interesante poder trabajar con todas las variables de la base de datos, pero alguno de los análisis no se ha podido realizar mediante SPSS, por lo que la clasificación de la adecuación al tema de cada artículo no ha sido posible de comprobar-

Esta es una de las nuevas cuestiones que afectan al posicionamiento del buscador y por lo tanto un factor muy importante a tener en cuenta, también es una técnica novedosa vinculada al estudio del lenguaje natural y computacional. En la base de datos estructurada aparecían 5 variables con una distribución latent Dirichlet allocation (LDA), que clasifica holísticamente la relación existente entre tema y artículo, mediante el procesamiento del lenguaje natural.

En este sentido el trabajo podría ser ampliado para conseguir una mayor clasificación predictiva de artículos populares, aportando una información complementaria que afecta al posicionamiento web. En este sentido un redactor podría evaluar el grado de aproximación óptimo a cada tema para introducir más o menos información subjetiva y con ello optimizar aún más los resultados.

Por lo que se podría seguir trabajando para optimizar cada una de las características de los artículos individualmente para conseguir maximizar la popularidad, asignándose así unas pautas exactas en la redacción.

BIBLIOGRAFÍA

Agulló. J (2010). Analítica Web y agencias de medios. 2017, de Evoca Sitio web: http://www.evocaimagen.com/cuadernos/cuadernos2.pdf

Albeanu, C. (22 de Julio de 2015). *Journalism.co.uk*. Recuperado el 29 de 02 de 2017, de Inside Velocity: Mashable's predictive social analytics tool: https://www.journalism.co.uk/news/inside-velocity-how-mashable-s-predictive-social-analytics-tool-fits-in-the-newsroom/s2/a565861/

Anónimo. (s.f.). *CreatividadDigital*. Recuperado el 7 de 02 de 2017, de Redacción contenidos web: http://www.creatividaddigital.com.ar/web/e-marketing/redaccion-de-contenidos.htm

AOL. (16 de Diciembre de 2016). *Mashable*. Recuperado el 09 de 04 de 2017, de How data is making journalism smarter than ever: http://mashable.com/2016/12/16/data-smart-journalism/?utm_cid=hp-n-1#ywLjJXc4IsqH

Arbildi, I (2005). "Posicionamiento en buscadores: una metodología práctica de optimización de sitios web". En: El profesional de la información.com, 2005, marzo-abril, v. 14, n. 2, pp. 108-124.

Arribas, B. A. (Febrero de 2014). *iab Spain*. Recuperado el 12 de 04 de 2017, de I Estudio de MEDIOS de Comunicación ONLINE: http://www.iabspain.net/wp-content/uploads/downloads/2014/02/Primer_Estudio_Medios_Comunicacion_Online_IAB_Spain_2014.pdf

Berlanga. V, Rubio. M. J., Vilà. R. (08 de 01 de 2013). *Universitat de Barcelona*. Recuperado el 10 de 05 de 2017, de Cómo aplicar árboles de decisión en SPSS. : http://diposit.ub.edu/dspace/bitstream/2445/43762/1/618361.pdf

Blog Historia de la informática. (17 de Diciembre de 2013). Recuperado el 23 de 03 de 2017, de Reconocimiento de la polaridad semántica: http://histinf.blogs.upv.es/2013/12/17/reconocimiento-de-la-polaridad-semantica/

Dennis, J. (22 de Febrero de 2010). *Illinois News Bureau*. Recuperado el 07 de 04 de 2017, de Business culture steers flow of ideas, study says: https://news.illinois.edu/blog/view/6367/205709

Dobrincu, S. (9 de Mayo de 2017). *crunchbase*. Recuperado el 22 de 05 de 2017, de mashable overview: https://www.crunchbase.com/organization/mashable/contributors

Fernandes. K, Vinagre. P y Cortez. P. Una decisión inteligente proactiva Sistema de apoyo para predecir la popularidad de las noticias en línea. Actas De la 17ª EPIA 2015 - Conferencia Portuguesa sobre Inteligencia Artificial,

Septiembre, Coimbra, Portugal.

Fragoso, R. B. (18 de Junio de 2012). *IBMdeveloperWorks*. Recuperado el 15 de 01 de 2017, de ¿Qué es Big Data?: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

Garbay, J. (27 de Marzo de 2017). *Merca2.0*. Recuperado el 6 de 04 de 2017, de Estas son las 14 estrategias más efectivas en marketing digital: https://www.merca20.com/estas-las-14-estrategias-efectivas-en-marketing-digital/

Gazquez, J. (14 de Noviembre de 2016). *Foromarketing*. Recuperado el 28 de 03 de 2017, de Qué es el Big Data y usos en el Marketing de Contenidos: http://www.foromarketing.com/big-data-marketing-contenidos/

ipmark. (27 de Octubre de 2014). Recuperado el 9 de 02 de 2017, de MEC utilizará la predicción viral de Mashable: http://ipmark.com/mec-utiliza-prediccion-viral-de-mashable/

McAfee, A., Brynjolfsson, E. (2012), "Big Data: the management revolution", Harvard Business Review, vol. 90, n. º 10, p. 61-68.

PuroMarketing. (29 de Julio de 2011). *Marketing de contenidos, clave para la comercialización de las empresas en internet*. Recuperado el 18 de 02 de 2017, de http://www.puromarketing.com/10/10557/marketing-contenidos-clave-para-comercializacion-empresas-internet.html

Raído, G. (23 de Diciembre de 2015). *¿Qué es Big Data? Explicación para Dummies*. Recuperado el 15 de 01 de 2017, de http://www.datacentric.es/blog/bases-datos/que-es-big-data-explicacion-para-dummies/

Schmidt, E. (2011). Conferencia Lake Tahoe en California.

Searchmetrics. (2016). *Rebooting Ranking Factors Google.com*. Recuperado el 06 de 04 de 2017, de http://pages.searchmetrics.com/rs/656-KWJ-035/images/Whitepaper-Searchmetrics-Rebooting-Ranking-Factors-US.PDF?mkt_tok=eyJpIjoiWVRSaE9Ua3pPV00wTXpaayIsInQiOiJkcXZYNEpib1dOY1V1OUhPRkU1MUJcL3JNc25HamZPVGxrSmlZSVhOb040SW1LSHBVRU9URmVWWnRhZGd0QlZ2eE1WMlRhcklK

Semrush. (22 de 03 de 2017). Recuperado el 22 de 03 de 2017, de Visión general de dominio "mashable.com": https://es.semrush.com/info/mashable.com

Wikipedia. (24 de Noviembre de 2016). Recuperado el 21 de 03 de 2017, de Minería de textos: https://es.wikipedia.org/wiki/Miner%C3%ADa_de_textos

Wikipedia. (22 de Mayo de 2017). Recuperado el 14 de 05 de 2017, de Neuro-linguistic programming: https://en.wikipedia.org/wiki/Neuro-linguistic_programming