Universidad Nacional Autónoma de México

Facultad de Ingeniería

Algoritmo para la identificación y análisis de patrones de publicidad dirigida utilizada en tendencias actuales

T E S I S

para obtener el título de:

Ingeniero en Computación

presenta:

Donovan Riaño Enriquez

Director de Tesis:
Dr. Guillermo Gilberto Molero-Castillo

Ciudad Universitaria, Cd.Mx., 2022

Dedicatoria

Para mis ...

Seres queridos más cercanos, para mis padres, mis amigos, grandes profesores y personas que me han ayudado en este periodo universitario.

Para mi Madre Claudina Enriquez que desde que tengo memoria, a jalones y estirones me educó y priorizó por mi educación, mi mejor futuro dentro de sus posibilidades. A mi padre Oscar Riaño por también educarme y encaminarme al ejercicio que fue vital en poder despejar mi mente y mejorar mi concentración. A mi hermana Fernanda Fayad que a pesar de tener diferencias, también ha velado por mí, me ha escuchado y hemos salido adelante.

A mis amigos que siempre nos apoyamos en los buenos y malos momentos, en el estudio y en la diversión. A mis amigos de otras carreras, en especial a los de Ingeniería Industrial, porque tuve la oportunidad de tener un trabajo hace un tiempo que me introdujo hacia el emprendimiento, los negocios y a mejorar mis habilidades blandas.

A la M. I. Tanya Ricci por apoyarme y escucharme cuando más problemas tuve, por ser mi profesora, amiga y tutora, así como por hacer que mi estancia en la universidad sea más amena.

Al Dr. Guillermo Molero-Castillo por creer en mi desde que nos conocimos, por observar mi potencial y desarrollarlo, en las publicaciones que realizamos, como en la revista Programming and Computer Software, por escuchar mis ideas y orientarme con esos proyectos personales que surgían. Además de ser mi director de este trabajo de investigación.

A la UNAM en general, por tener buenas instalaciones, el mejor campus de la Ciudad de México, por su oferta cultural, educativa y deportiva, por las competencias en las que pude participar, así como los entrenamientos deportivos de gimnasia y baloncesto. Pero especialmente agradezco a Rodrigo Piñon por colaborar en el programa de Servicio Social, creíste en mí y sentamos las bases del algoritmo para múltiples propósitos, por lo que fuiste fundamental en esto amigo.

Finalmente, quiero agradecer el apoyo recibido para este trabajo de investigación, a través de una beca, al proyecto PAPIIT IA104122.

Resumen

Hoy en día, las empresas eligen mostrar su contenido publicitario en páginas web, utilizando los recursos que ofrece la tecnología, ya sea a través de imágenes llamativas, animaciones e incluso vídeos. El propósito es garantizar que sus ideas transiten hacia las audiencias objetivo, esto es, a las personas que puedan consumir sus productos o servicios. Por lo tanto, en la actualidad y posterior a la pandemia por COVID-19, la tecnología ha revolucionado el marketing digital, teniendo como mayores exponentes de efectividad a las páginas web y las redes sociales. Este documento muestra los resultados logrados como parte del proyecto de investigación aprobado por el Comité de Titulación de la División de Ingeniería Eléctrica de la Facultad de Ingeniería. **Objetivo.** Se describe la implementación de un algoritmo para la identificación y análisis de patrones de publicidad dirigida en tendencias actuales, utilizando expresiones regulares. **Método.** La implementación del algoritmo fue en Python, apoyado por las bibliotecas Tesseract, Pillow y OpenCV, así como por el paquete Tesseract-OCR. La implementación se inició desde un punto antes del proceso OCR (reconocimiento óptico de caracteres). Además, se utilizó Selenium para realizar el deslizamiento automático en el navegador web, tomándose en cuenta la diferencia de altura entre los diferentes tamaños de monitores que hay en el mercado. Cabe señalar que las pruebas se realizaron en tres tipos de navegadores web: a) Google Chrome, b) Mozilla Firefox y c) Safari. **Resultados.** Derivado del trabajo de investigación realizado, se logró la publicación de dos artículos de investigación. El primero en la revista *Programming and Computer Software*, indizado en JCR. El segundo en *LACCEI* (Latin American and Caribbean Consortium of Engineering Institutions), indizado en Scopus. En ambos casos fue posible detectar anuncios publicitarios en la web de manera exitosa. **Conclusión.** Para cada tema de tendencia, hay diferentes formas de llegar al público y a cada uno con diferentes intenciones. Lo importante es observar la reacción que se genera en el público y específicamente en los consumidores potenciales, para diseñar una mejor estrategia de marketing. En este sentido, a través de este tipo de mecanismos, como el algoritmo implementado, se pueden identificar con éxito las tendencias y experiencias de otras marcas con campañas publicitarias exitosas. Este enfoque podría ser útil para que los anunciantes puedan saber cuál fue el anuncio o promoción que causó mayor efecto en los usuarios, en virtud del incremento del marketing digital para el mercado de consumo.

Índice general

Capítulo 1

Introducción

1.1. Contexto de la investigación

A lo largo del tiempo, la tecnología y de manera particular Internet han tenido un gran impacto en el mundo publicitario, los mercados y los modelos de negocio. Antes, la publicidad se hacía de acuerdo a los medios tradicionales. Sin embargo, este escenario dio un giro diferente con la era del Internet. Asimismo, este fenómeno estimuló un impacto en la publicidad, donde las empresas identificaron la utilidad e importancia de utilizar este mecanismo como herramienta para lograr un mayor alcance en un público objetivo, es decir, a través de publicidad dirigida.

Así, en los últimos años, 2020 y 2021, se han visto tendencias que han sacudido al mundo como nunca antes. Una de estas tendencias, que ha venido a cambiar cualquier negocio u organización, además de su impacto en la salud, fue la llegada de la pandemia por coronavirus (COVID-19) [1]. Una enfermedad altamente contagiosa, peligrosa y compleja de manejar debido a su inicio asintomático y la complejidad de su detección temprana [2]. Una de las acciones que tomaron los líderes mundiales fue quedarse en casa para evitar, en lo posible, el contacto y, por ende, posibles contagios. Esto mientras se lograba obtener alguna vacuna o una forma de sobrellevar la situación. Sin embargo, ante estas medidas desesperadas, sumada a la caída de la economía mundial, el marketing digital no se detuvo, tomó mayor impulso, ganando notoriedad generalizada en la sociedad actual.

Esta notoriedad del marketing digital se debe, en parte, a las medidas tomadas por los gobiernos, quienes advirtieron a la población en general mantener una distancia razonable, extremar las medidas de higiene y exponerse lo menos posible al contacto con otras personas [3]. Estas medidas, evidentemente, generaron miedo e incertidumbre en la población, lo que propició, por un lado, un constante bombardeo mediático sobre el tema y, por otro, que las grandes empresas aprovecharan la coyuntura para impulsar sus ventas a domicilio, por ejemplo, artículos de primera necesidad, limpieza y desinfección, cubrebocas, medios de entretenimiento para el hogar, alimento para mascotas, entre otros.

Por otro lado, mientras muchos negocios que dependían del contacto humano tuvieron

una gran caída en su economía, grandes empresas, como supermercados, farmacias e incluso marcas de ropa, aprovecharon la incertidumbre para lanzar promociones y nueva mercadería [4]; consiguiendo así inundar los medios publicitarios para reducir el sedentarismo y garantizar la máxima higiene. La Figura 1.1 contrasta el resultado de la llegada de la enfermedad en combinación con el marketing digital invasivo, que impulsó el número de ventas de diferentes productos relacionados con el tema, representando amplias ganancias totales.

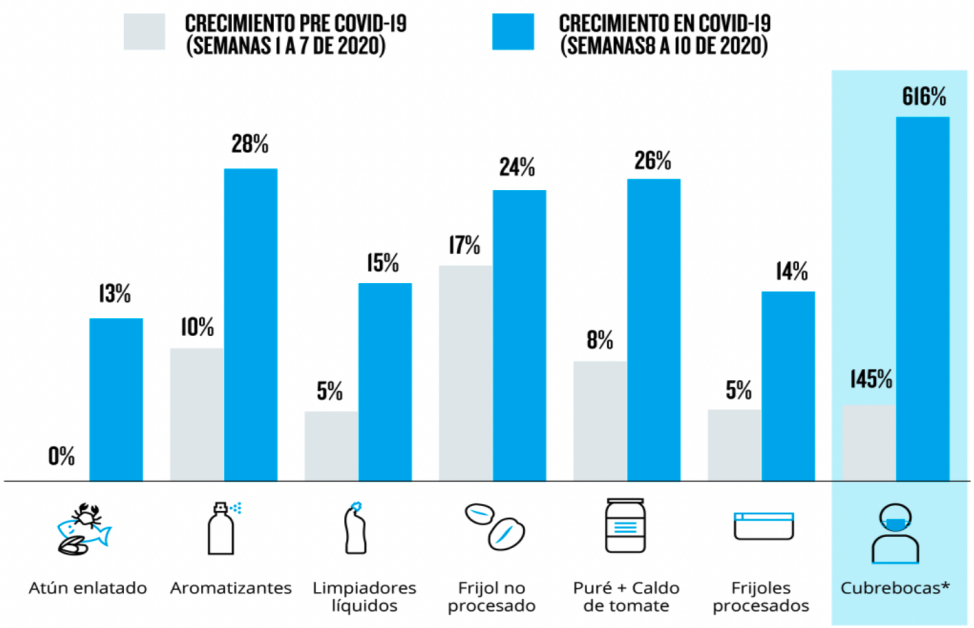


Figura 1.1: Comparación de índices de ventas antes y durante la presencia de COVID-19. Fuente: [5].

Un ejemplo de la implementación de nuevos modelos de negocio, y aprovechando las circunstancias, son las empresas de confecciones, que casi de inmediato incursionaron en la producción de cubrebocas, lo que no solo incrementó sus ganancias, sino que las dejó bien posicionadas, en términos de aceptación, ya que sabían gestionar estratégicamente el marketing. Este ejemplo, dada la situación de COVID-19, es un caso extraordinario, donde algunas empresas buscaron aprovechar las tendencias para ser rentables y mantener perfiles de aceptación a través de los consumidores potenciales [6], dado que en el mundo empresarial tener una buena imagen es vital e incluso más importante que ofrecer productos de calidad.

Otro ejemplo de aprovechamiento de las tendencias actuales, que no necesariamente tiene que ver con el caso de la pandemia de 2020 y 2021, fue *Pride*, también conocido como *LGBT Pride Day*, que es un día del año elegido por el movimiento LGBT (abreviación de las palabras Lesbianas, Gais, Bisexuales y Trans) para afirmar el sentimiento de pertenencia personal, que se genera al mostrar públicamente identidades y orientaciones de género

tradicionalmente marginadas y reprimidas. El propósito fue hacer visibles sus reclamos y su presencia en la sociedad. Ante esto, empresas como Puma, Levis, C&A, Adidas, Swatch, entre otras, aprovecharon los ideales y colores del movimiento afín al orgullo, bajo el lema: "El amor nos une ahora más que nunca", con una propuesta de prendas de vestir, calzado colorido y accesorios.

Como es bien sabido, algunas empresas y estrategas de marketing digital han logrado integrarse y posicionarse exitosamente con el público. Para esto, estas empresas analizaron tendencias mundiales o individuales para vender más y hacer más notoria su imagen. Estas empresas a su vez, a pesar de haber sufrido pérdidas por el confinamiento de la población, tienen en el marketing digital una gran herramienta de apoyo que las ha mantenido a flote, demostrando que una buena estrategia publicitaria marca la diferencia, superando a la competencia.

1.2. Problema de investigación

En los últimos años, la explosión tecnológica ha venido a cambiar la forma de hacer negocios y la humanidad ha aprendido a adaptarse, convivir y aprender de ellos. Estos avances tecnológicos han acortado, cada vez más, los tiempos de nuevos desarrollos y descubrimientos, basados en la comprensión del presente y pensando en el futuro. Precisamente, un campo de conocimiento que actualmente está impulsando nuevos desarrollos es la Inteligencia Artificial [7], donde muchas empresas son cada vez más conscientes de su importancia para incorporar esta tecnología a sus modelos de negocio. Esto permitirá obtener patrones relevantes a partir de los datos generados, así como apoyo en el proceso de la toma de decisiones [8].

En este sentido, el marketing ha sido una de las áreas que ha evolucionado con la tecnología [9] y ha empoderado a las empresas para hacer crecer las ventas y, en consecuencia, mejorar su posicionamiento en el mercado global. Por otro lado, el neuromarketing se basa en la parte de la psicología del consumidor, los colores que reflejan los estados de ánimo, las formas en que los consumidores encuentran más agradable la presentación del producto, las posiciones que toman hombres y mujeres cuando compran solos, con amigos o familiares, entre otras características.

En el caso comercial, el marketing digital está creando nuevas formas para que las empresas interactúen con los consumidores y ofrezcan nuevas formas de comunicación, lo que genera mayores ingresos y una mejor productividad. En este tipo de marketing, las tecnologías de la información se incluyen como parte de las campañas publicitarias a través de diversas aplicaciones de software [10], con las cuales es posible medir las ganancias y el comportamiento de los usuarios de Internet, por ejemplo, la aceptación de anuncios y la tasa de rebote.

Sin duda, la adopción de la tecnología en la vida actual es una realidad, de la cual se

depende cada vez más y que ha cambiado la manera de trabajar, estudiar, ofrecer productos y servicios, y la creación de nuevas líneas de negocio. Ahora, con la tecnología se están creando mayores oportunidades y empleos relacionados. Motivo por el cual, a través de este trabajo de investigación, se buscó la implementación de un algoritmo para la identificación y análisis de patrones de mensajes dirigidos por algunas marcas que han logrado incrementar sus ventas de manera exitosa, dadas las tendencias actuales, con el mensaje de concientizar a sus consumidores.

1.3. Objetivos

1.3.1. Objetivo general

Implementar un algoritmo para la identificación y análisis de patrones de publicidad dirigida en tendencias actuales, utilizando expresiones regulares.

1.3.2. Objetivos específicos

- Diseñar el algoritmo para su funcionamiento en diferentes navegadores web.
- Validar el funcionamiento del algoritmo a través de la detección de publicidad utilizada por diferentes marcas.

1.4. Justificación

Una tecnología importante de la computación es la Web, la cual desde su concepción ha sufrido cambios debido a la revolución y creación de nuevos servicios. En la actualidad, la Web se clasifica en tres etapas [11, 12]: i) la primera se enfoca en la creación de la web, protocolos y navegadores web, como clientes; ii) la segunda la define el aumento de redes sociales, aplicaciones móviles y el cómputo en la nube; y iii) la última, que es la etapa venidera, promete una mayor revolución que las anteriores, descentralizando todo el contenido de la web, creando perfiles únicos, dando un mayor auge a las criptomonedas e integración a los NFTs (Token no fungibles), y el Metaverso (concepto que denota la siguiente generación de Internet).

Hoy en día, la tríada entre NFTs, el ecosistema Web3 y el Metaverso, las cuales son tecnologías que comparten características en común, han permitido una importante sinergia para el despegue de la Web 3.0, que de acuerdo con la literatura actual tienen un amplio potencial que demanda una gran cantidad de conceptos técnicos y de otros dominios, como: sistemas distribuidos, economía, finanzas, organizaciones, emprendimiento, marketing, criptomonedas, criptografía, arte, propiedad intelectual, redes, seguridad, renderización, realidad virtual, realidad aumentada, programación descentralizada, entre otros.

Por otro lado, la adopción de las nuevas formas de hacer negocios y marketing, a través de los anuncios en los sitios web, ha sido la principal fuente de ingresos de empresas como Google y Facebook, ocasionando la irrupción de la privacidad de los usuarios, que en repetidas veces no se han sentido cómodos, no solo con la publicidad invasiva, sino la publicidad dirigida, que se basa en predecir los gustos, necesidades e intereses de sus usuarios.

Asimismo, las estrategias de mercado comenzaron a resultar más efectivas con los métodos de medición de Google, como la tasa de rebote, pago por click, y la publicidad dirigida, que es la que se enfoca en las necesidades, gustos e intereses de los cibernautas. Hoy en día, algunas empresas cambian con frecuencia sus estrategias, enfocándose en la experiencia del usuario, calidad del producto, tiempos más rápidos de entrega, pero sobre todo, en la psicología del consumidor, lo que terminó de moldear la manera de demandar productos y servicios a través de tiendas online y aplicaciones móviles. La pandemia simplemente ahondó más este hito, y por las condiciones de confinamiento y distanciamiento, atrajo y aseguró a los clientes nuevos y existentes.

Anterior al periodo de la pandemia se encontraban *slogans* que aludían al cuidado del bolsillo o descuentos enfocados al ahorro y bienestar. Ahora, se observa que los anuncios cambiaron a: "Quédate en casa", "Cuida de ti y tu familia", desplazando los descuentos y priorizando la atención y la salud como nueva estrategia, así como los movimientos sociales, como el día del orgullo LBGT con los *hashtags* #Pride, reverenciando al orgullo, promoviendo la inclusión, con el objetivo de que las personas se sientan cómodas y libres de comprar sus productos.

1.5. Organización del documento

El documento está organizado de la siguiente manera, el Capítulo 2 presenta los fundamentos de la evolución del marketing digital y sus necesidades, las tendencias actuales, así como las nuevas tecnologías que se incorporan a este escenario. El Capítulo 3 describe el método utilizado, como propuesta de solución, con los pasos definidos para la implementación del algoritmo de identificación y análisis de patrones de publicidad dirigida en tendencias actuales y el uso empleo de expresiones regulares. El Capítulo 4 presenta los resultados obtenidos a partir de las pruebas realizadas en navegadores web, como Chrome, Firefox y Safari; y el Capítulo 5 resume las principales conclusiones y el trabajo futuro.

Se presenta además dos anexos, en los que se incluye información relacionada sobre el trabajo de investigación realizado. En el Apéndice A se presenta el artículo de investigación publicado en la revista *Programming and Computer Software*, indizado en JCR, con un factor de impacto de 0.801 para el 2021. El título del trabajo publicado es *Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm* [13]. El Apéndice B muestra un segundo artículo de investigación, *Algorithm for Identification and Analysis of Targeted Advertising used in Trending Topics*, que fue publicado en *LACCEI* (Latin American and Caribbean Consortium of Engineering Institutions), indizado

en Scopus, esto como parte de las mejoras al trabajo realizado en la publicación anterior [14].

Capítulo 2

Marco teórico y estado del arte

En los últimos años, la explosión tecnológica ha venido a cambiar la forma de hacer negocios y la humanidad ha aprendido a adaptarse, convivir y aprender de estos. Estos avances tecnológicos han acortado, cada vez más, los tiempos de nuevos desarrollos y descubrimientos, basados en la comprensión del presente y pensando en el futuro. Precisamente, un campo de conocimiento que actualmente está impulsando nuevos desarrollos es la Inteligencia Artificial, donde muchas empresas son cada vez más conscientes de su importancia para incorporar esta tecnología a sus modelos de negocio. Esto permitiría obtener patrones relevantes a partir de los datos generados, así como el apoyo en el proceso de la toma de decisiones.

2.1. Marketing digital

En el pasado, el propósito principal del *marketing* fue coordinar los medios de comunicación, hacer tratos para que las personas o las empresas tuvieran una buena reputación sobre los productos que se anunciaban, o sobre las ideas que se planeaban vender, obviamente con el objetivo de generar más ventas. Ahora, con todas las herramientas que ofrece la tecnología, los usuarios utilizan los motores de búsqueda para encontrar lo que quieren, así como para acceder a las críticas y comentarios que hacen la misma comunidad.

Como es lógico pensar, las estrategias de marketing cambian por completo a raíz de los nuevos diseños web, que pueden dar lugar a páginas web dinámicas, en las que las personas pueden interactuar, principalmente con los sitios web, generando información sobre los propios usuarios, como sus gustos e intereses, así como las necesidades de la sociedad en general.

Dentro del marketing digital, el principal objetivo es el usuario [15]; por lo que, hoy en día una estrategia digital debe incluir todos los espacios relevantes para que el usuario interactúe, buscando personas que influyan en sus opiniones para entrar en la misma red de usuarios (internautas), generando mayor fuerza en los datos para nuevas ideas o productos. Esto se refleja en los avances de los motores de búsqueda, que de acuerdo con la experiencia adquirida, pueden volverse cada vez más invasivos, aplicando la psicología del

consumidor.

Después de analizar la estructura, diseños, colores y técnicas de marketing, la idea básica de los anuncios en la web es llamar la atención de personas de cualquier manera posible, usando *banners*, palabras clave o enunciados que roben la atención a primera vista; y, en algunos casos, con el uso de ventanas de cierre manual, conocidas como *pop-ups*, para que una vez que tengan la atención del usuario puedan decir quienes son y qué ofrecen.

2.2. Necesidades de marketing digital

La expansión de la Inteligencia Artificial en los negocios se refleja en el creciente desarrollo de aplicaciones y servicios, que van desde el procesamiento del lenguaje natural hasta el reconocimiento de imágenes. Con el paso del tiempo, este tipo de tecnología ha logrado una mejor capacidad de implementación en las diversas actividades de las organizaciones. Sin embargo, aún no ha sido posible desarrollar autonomía para gestionar todos los procesos que requiere el negocio [16].

Realizar el reconocimiento y análisis de los procesos de negocio es importante para identificar cuellos de botella, desviaciones y otro tipo de problemas. Por lo tanto, es necesario desarrollar nuevos mecanismos para obtener mayores beneficios. Entre las acciones que la Inteligencia Artificial podría abarcar en los negocios destacan [17]: el procesamiento del lenguaje natural, el análisis de satisfacción, el reconocimiento óptico de caracteres, el reconocimiento de imágenes y rostros, la mejora de los procesos de venta a través del aprendizaje automático, entre otras.

Por otro lado, respecto a la publicidad en los navegadores web, cada vez va en aumento, el almacenamiento y la interpretación de la gran cantidad de datos recopilados por los vectores de búsqueda ponen en riesgo la privacidad de los usuarios. Por lo tanto, estos grupos de datos también podrían proporcionar información útil para una mejor gestión de la publicidad dirigida.

2.3. Tendencias actuales

Las tendencias actuales o *trending topics* son estadísticas recopiladas por sitios web o redes sociales, principalmente Twitter, Facebook e Instagram [17], donde se analizan los intereses de los usuarios o las publicaciones que generan mayor controversia. Así, es posible conocer las demandas y necesidades de los usuarios y en qué momentos del día, mes o año es más factible persuadirlos para captar su atención.

En la actualidad las empresas pueden aprovechar los trending topics a su favor, para dar a conocer a la audiencia o promocionar sus productos, pero lo más importante es

crear una buena reputación para generar confianza y acercamiento con los consumidores. En la Figura 2.1 se muestra un diagrama de una red semántica de marketing orientada a trending topics, que se centra en la experiencia de las empresas hacia los consumidores actuales. Esta red semántica relaciona: i) quién ofrece, ii) qué producto o servicio ofrece, y iii) de qué manera llama la atención (oferta o promoción). Estos tipos de publicidad se distinguen al momento de navegar por Internet, los cuales algunos de estos fueron objeto de análisis en este trabajo de investigación.

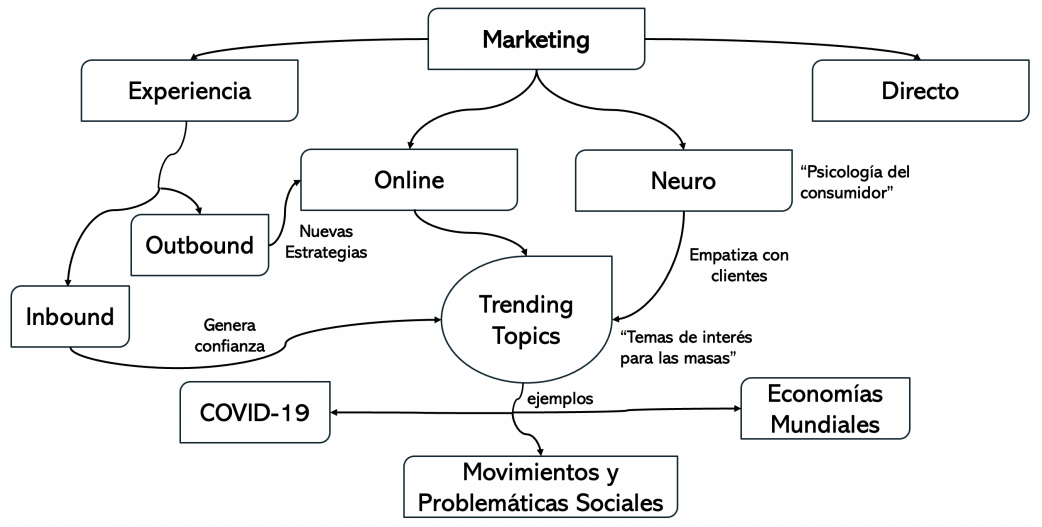


Figura 2.1: Red semántica de marketing orientada a trending topics. Fuente: [18].

En la red semántica se tiene al antagonista de 2020, el *coronavirus*. Para lo cual, empresas como Adidas o Levis han lanzado a la venta mascarillas. Otras empresas como Walmart o Liverpool han potenciado, como la mayoría de las empresas del mundo, el comercio electrónico, pero estas destacan la parte de '*quedarse-en-casa*' o '*cuidar-a-nuestros-clientes*', que a su vez, si las personas se mantienen saludables después de este evento u otros, seguirán regresando a sus tiendas, siendo parte del ciclo de consumo.

En cuanto al nodo del orgullo o *Pride*, Adidas volvió a ser protagonista, quien lanzó una colección de ropa y calzado que simboliza tal movimiento. Además, esta empresa ha sabido interactuar con las necesidades de los usuarios, fabricando productos de origen reciclado, cuidando así el medio ambiente, que también es tema de discusión en los últimos años. Por otro lado, la caída de la economía mundial y de los propios países es otro tema que ahora se ha convertido en tendencia.

El confinamiento por COVID-19 tuvo como consecuencia, en algunos países, la disminución de la producción, siendo parte fundamental del Producto Interno Bruto de cada país [19]. Por lo que, se han tomado medidas y acuerdos entre el gobierno, la banca y las

empresas para generar acciones que retomen la producción y fomenten la inversión.

En este sentido, los sitios web que subsisten de la publicidad ofrecen un breve contenido de noticias, educación o entretenimiento y sus anuncios son proporcionados por Google. En estos, los anuncios cambian a la orden del día. En los casos de las tiendas informativas, corporativas y en línea, estas incluyen sitios como MSN, Amazon, Sanborns, Adidas y otros, cuyo propósito es ofrecer tendencias actuales en productos o servicios. Mientras que en el caso de las redes sociales y derivados como Facebook, Instagram, LinkedIn, Twitter y otros, son sitios que ofrecen interacción con los usuarios y anuncios relacionados con el perfil del usuario.

Por lo tanto, en todos los sitios web mencionados, el usuario se convierte en un posible cliente al ingresar a la publicidad de la empresa o comercio, donde es probable que encuentre más ofertas o promociones, por esta razón, cada vez más productos o servicios aumentan la probabilidad de compra, separado de un anuncio en la web. Esta posibilidad de compra también se debe a que un anuncio en la web puede permanecer en la memoria durante mucho tiempo, lo que posteriormente provocará la búsqueda o compra del producto o servicio.

2.4. Privacidad de los usuarios

En la actualidad, la información y la privacidad se encuentran en una etapa crítica [20]. Empresas como Facebook, Twitter, Google, Amazon, entre otras, han incluido en sus sitios web, aplicaciones o motores de búsqueda para el acaparamiento de cantidades masivas de datos. Estas empresas proporcionan "gratuitamente" sus servicios, a cambio utilizan los datos de los usuarios a conveniencia, por ejemplo, para propósitos publicitarios. Entre los pilares para tener una adecuada seguridad de la información se destacan [9]:

- Autenticación. Para identificar la entidad comunicante y la fuente de datos.
- Control de acceso. Para evitar el uso no autorizado de los recursos.
- Confidencialidad. Para proteger los datos contra la divulgación no autorizada.
- Integridad. Para garantizar la no alteración o destrucción de los datos de forma no autorizada.
- No repudio. Para acreditar el origen de los datos o su entrega.
- Disponibilidad. Para garantizar la continuidad de la accesibilidad y uso por parte de las entidades autorizadas.

Estos servicios se brindan a través de mecanismos de seguridad solos o combinados, tales como cifrado, firmas digitales, mecanismos de control de acceso, mecanismos de integridad de datos, intercambio de autenticación, llenado de tráfico, control de enrutamiento,

entre otras acciones relacionadas con la criptografía y seguridad de la información.

En cuanto a la publicidad en los navegadores web, al ser cada vez mayor, se pone en riesgo la privacidad de los usuarios. Es decir, al buscar en la web cierto tipo de servicios o productos, hay una saturación de información en el sitio web donde se posiciona. El problema no solo influye en el manejo de la información, sino también en la lentitud del flujo de información por temas publicitarios. Así, al tener tanta información, bajo un tratamiento estricto, se proporciona conocimiento útil para una mejor gestión de la publicidad dirigida [9].

De esta forma, se puntúa y acerca un anuncio web más específico según las demandas de cada usuario, teniendo en cuenta, por ejemplo, la actualización de Google. Esta actualización de los servicios de Google, en materia de publicidad, incluye ajustes manuales de los gustos de cada usuario, por supuesto, sugeridos con la información recopilada por *PageRank*, *trackers*[1] y las *cookies* de los sitios web. La Figura 2.2 muestra un extracto de la notificación de dicha actualización, como parte del servicio, que incluye la configuración de perfiles, contraseñas, contactos, información, entre otros.

Figura 2.2: Configuración de publicidad relacionada con una cuenta de usuario en Gmail.

Esta extensión se basa en el principio de RegExp (expresiones regulares) y tiene en cuenta la navegación web de los usuarios, por lo que, la hipótesis es incluir las marcas que más pagan por un anuncio en la web, o incluso las que más venden en un determinado mercado, así como los productos, servicios y sus marcas secundarias [21].

Como ejemplo, la Figura 2.3 muestra el motor de búsqueda de Google para la creación automática de perfiles de usuario en función de la navegación web. Por ejemplo, es relevante un usuario masculino, con un rango de edad entre 18 y 24 años, cuyo interés es

[1]En marketing, son indicadores y rastreadores de la efectividad de las campañas publicitarias dirigidas

comprar en línea, realizar actividades al aire libre, desarrollar aplicaciones móviles, ver películas y otros. Este usuario, mediante la recolección de datos a través de *cookies*, *trackers* e información contenida en el perfil de Gmail, puede ser utilizado para agrupar perfiles con características comunes y, por lo tanto, recibir publicidad de diversas ofertas y promociones.

Figura 2.3: Buscador de Google para la creación automática de perfiles de usuario en función de la navegación web.

Con base en lo anterior, cuando se interactúa constantemente con los servicios de Google, es recomendable revisar con frecuencia las políticas de datos y privacidad sobre los ajustes realizados por *Mountain View*, con el fin de tener un mayor conocimiento sobre la seguridad y uso de la información.

2.5. Bloqueadores de anuncios en navegadores web

En los últimos años, se han incrementado los esfuerzos para implementar bloqueadores de publicidad en la web [22]. Estos se desarrollaron en un inicio para el navegador Firefox, que con el paso de los años fue mejorando hasta lograr un navegador sin anuncios, pero las empresas buscaron la forma de seguir enviando publicidad. En consecuencia, en poco

tiempo, comenzaron a crearse controversias debido a las pérdidas significativas de posibles ingresos económicos [7]. Ante esto, se realizaron también mejoras en los bloqueadores de publicidad, conocidos actualmente como Ad-blockers [23], comenzando una disputa entre los anuncios y la privacidad.

Otro término que también se usa actualmente es *Trackers*. Estas herramientas almacenan información a través de *cookies*, y estas, a su vez, proporcionan la ubicación de las búsquedas realizadas. Así, a medida que crecía la demanda y las posibilidades de bloquear la publicidad, se implementaron nuevos Ad-blockers en la mayoría de los navegadores web, así como en las aplicaciones para dispositivos móviles, la publicidad ha encontrado la manera de realizar anuncios sobre las búsquedas de los usuarios.

Un dato importante a tener en cuenta es que, en los últimos años, Google controla el 85 % del negocio mundial de publicidad en buscadores y alrededor del 50 % de toda la publicidad online [24]. Las personas y la sociedad, en general, ven a Google como un servicio [25], pero detrás hay una tecnología que contiene funciones específicas e incluye extensiones exclusivas y restrictivas.

En este sentido, la publicidad y las compras online son cada vez más demandadas y rápidas, por ejemplo, las compras online ahorran tiempo y distancia. Sin embargo, el uso de sitios web para este tipo de compras sirve a las empresas para aprender de las experiencias y necesidades de los usuarios; traduciéndose como patrones de comportamiento en diferentes extractos sociales, ya sean locales o regionales.

2.6. Selenium

Selenium, conocido también como Selenium Webdriver, es una herramienta para automatizar procesos en diferentes navegadores web [26]. Su propósito es mejorar el soporte para la detección de problemas en cualquier navegador web [27]. Esta herramienta permite probar cualquier navegador web para obtener datos de código HTML, cambiar, abrir y moverse entre pestañas de las ventanas del navegador, retroceder o avanzar según el historial de prueba, cambiar el tamaño de las ventanas, tomar capturas de pantalla, completar campos, hacer clic en partes de un sitio web, entre otros.

Estas tareas son aplicables a los lenguajes de programación Java, Python, C, Ruby, Perl y JavaScript. Los sistemas operativos que admite Selenium son Windows, MacOS y Linux, cada uno con sus respectivos paquetes y entornos de desarrollo integrados (IDE) [28].

Por otro lado, en la actualidad existe interés en utilizar el Reconocimiento Óptico de Caracteres (OCR) en la detección de publicidad web [25], sin embargo, aún se necesitan más esfuerzos para cubrir todas las estrategias de marketing digital. Por lo tanto, es necesario buscar la automatización de este proceso y generar una estrategia correcta para la

identificación y clasificación de lo que se ofrece y quién lo ofrece.

2.7. Trabajos relacionados

Uno de los trabajos enfocados a la detección de publicidad en la web fue [29], donde se presentó un algoritmo que realiza el rastreo de la web. Consiste en obtener información de los sitios web a través del etiquetado, la página web de prueba fue MSN. A partir de las etiquetas se realizó una clasificación mediante un modelo de probabilidad basado en regresiones logarítmicas. La clasificación se realizó a través de palabras clave a lo largo de la página web, es decir, al principio (B-comienzo), en el medio (I-adentro), al final (L-última), única (U), o afuera (O).

En [30] se describe el análisis de publicidad contextual a través de PageSense, que tiene como objetivo asociar anuncios en páginas web. A través de esta plataforma, se detectan regiones en blanco y se selecciona el área no intrusiva para la colocación de anuncios sin romper el estilo original de la página web. Para el análisis se utilizaron combinaciones y probabilidades bayesianas, que reflejan los porcentajes de anuncios para diferentes tipos de productos o servicios y, por lo tanto, define anuncios molestos y aceptables.

En otro trabajo, en [31], se hizo un análisis basado en distancias euclidianas. Estas distancias fueron con respecto a la forma en que los anuncios son de interés para los usuarios, la búsqueda de productos y la adaptación del perfil objetivo, dividiéndolo por secciones, como salud, deportes, empresa, sociedad, educación, arte, ciencia, informática, entre otros.

Por otro lado, [32] describe el análisis realizado a aproximadamente 500 páginas web, en las que se realizaron pruebas destinadas a detectar tipos de anuncios, pero no contenidos. Entre los tipos de anuncios analizados destacan pop-ups, carruseles, videos, gifs, juegos, stickers o texto. También se analizaron los países de donde provienen los anuncios, la frecuencia, el tamaño y el origen de las URL (Uniform Resource Locator).

Con base en lo anterior, en este trabajo de investigación se presenta la automatización de un algoritmo de reconocimiento de publicidad en la web, utilizando expresiones regulares y vectores de datos en formato CSV (Comma Separated Values). Las pruebas se realizaron en tres navegadores: Chrome, Firefox y Safari. Una característica del algoritmo es su ejecución automática y versátil, ya que no requiere acceder al código de la página web en revisión, y que se trata de una aplicación que opera en segundo plano.

Capítulo 3

Método

El desarrollo del algoritmo fue en Python. Este algoritmo ha sido asistido por las bibliotecas Tesseract [33], Pillow y OpenCV [34], así como el paquete Tesseract-OCR. Para la implementación, previo al proceso OCR, se empleó el reconocimiento óptico de caracteres. Además, se utilizó la biblioteca Selenium para realizar el deslizamiento automático en el navegador web.

En el desarrollo del algoritmo, también se tuvo en cuenta la diferencia de altura entre los diferentes tamaños de monitores del mercado (pantallas de los computadores), por consiguiente, el algoritmo realiza una detección dinámica de la altura de las ventanas, ajustándose así a cualquier tamaño de pantalla. El desplazamiento de la información en este trabajo es vertical.

Cabe señalar que las pruebas se realizaron en tres navegadores web: Google Chrome, Mozilla Firefox y Safari.

3.1. Deslizamiento a través de Selenium

A través de Selenium, se abre una nueva ventana para un navegador compatible con esta herramienta. El terminal pide al usuario la dirección URL del sitio web a analizar. Posteriormente, el navegador bajo prueba se expande a pantalla completa para un escaneo rápido y completo del sitio web.

Como parte del algoritmo, y con el objetivo de detener el proceso por unos instantes, como técnica de programación, se utilizaron hilos. La primera pausa se realiza para permitir la carga de la página web, ya que dependiendo de la velocidad de Internet, e incluso del estado del sitio web, se necesita de un tiempo de carga para realizar la captura de la pantalla correcta, y así posteriormente hacer el análisis mediante el reconocimiento óptico de caracteres (OCR, por sus siglas en inglés).

En el siguiente paso, la ventana se maximiza, antes de deslizarse para capturar y alma-

cenar pantallas. Esto provoca una segunda pausa, que es de 0.5 segundos. Posteriormente, inicia el proceso de captura de pantalla hasta terminar con todo el contenido vertical de la página web. Una vez finalizado el proceso de captura, se cierra automáticamente la ventana del navegador web donde se realizó la consulta.

3.2. Reconocimiento óptico de caracteres

El reconocimiento óptico de caracteres, OCR, permite extraer texto de una imagen con escritura alfabética, independientemente del idioma, tamaño o color del texto, con una alta efectividad. La eficacia del OCR oscila entre el 71 y el 98 %. Este sistema es capaz de alcanzar valores medios del 85.1 % para texto manuscrito y del 90.93 % para texto impreso o digital [35].

Se definió una función específica para la búsqueda de todas las imágenes de las capturas de pantalla guardadas con una ruta establecida. Luego se implementó un ciclo donde se analizan todas las imágenes dentro de la ruta establecida, en el orden en que fueron capturadas, esto debido a la nomenclatura que se utilizó para guardarlas.

Cuando se accede a la imagen deseada, los resultados del reconocimiento óptico de caracteres se guardan en una variable, que consiste en texto al que luego se le da formato, poniendo todo en mayúsculas, separando las palabras, eliminando espacios y caracteres no soportados por el lenguaje SQL (Structured Query Language), como: ', \, ., y &; esto con el objeto de comparar con expresiones regulares.

3.3. Expresiones regulares

Una vez obtenida la información del contenido de la página web y sintetizada en cadenas de texto, se accede al servidor local para validar el contenido del sitio web con una base de datos, que alberga alrededor de 600 palabras clave diferentes basadas en expresiones regulares, distribuidas en tres tablas, y cuyo principal objetivo es validar los siguientes temas:

- Palabras más utilizadas en marketing digital.
- Marcas, considerando sus respectivas submarcas en los productos o servicios que ofrece la empresa.
- El tipo o en qué consiste el producto o servicio.

Estas expresiones regulares corresponden a las palabras de marketing digital más utilizadas en español, las cuales, como se mencionó, se utilizaron para la comparación con las cadenas de texto obtenidas. Las Tablas 3.1, 3.2 y 3.3 muestran un fragmento de las expresiones regulares relacionadas con las palabras más utilizadas en marketing digital,

algunas marcas reconocidas y algunos productos que se publican con mayor frecuencia en México; respectivamente.

Cuadro 3.1: Fragmento de las palabras más utilizadas en marketing digital en español.

Palabra clave	**Plural**	**Acento**	**Carácter**
Ahorro	Ahorros	nulo	nulo
Bajo	Bajos	nulo	nulo
Comprar	nulo	nulo	comprar
Cotiza	nulo	nulo	cotizar
Descuento	descuentos	nulo	%
Dinero	dineros	nulo	$
Especial	especiales	nulo	nulo
Gratuito	gratuitos	nulo	gratis
Hasta	nulo	nulo	nulo
Ilimitado	ilimitados	nulo	nulo
Intereses	intereses	interés	nulo
Internet	nulo	nulo	Web
Oferta	ofertas	nulo	nulo
Plan	planes	nulo	nulo
Precio	precios	nulo	nulo
Producto	productos	nulo	nulo
Punto	puntos	nulo	nulo
Rápido	rápidos	nulo	nulo
Rebaja	rebajas	nulo	nulo
Salud	nulo	nulo	nulo

Cuadro 3.2: Fragmento de algunas de las marcas más vendidas y mejor pagadas de México.

Marca	**Submarca**	**Producto**	**Acrónimo**
Adidas	alphabounce	alphabounce	nulo
Adidas	NMD	NMD	nulo
Adidas	originals	originals	nulo
Apple	iMac	iMac	nulo
Apple	iPad	nulo	nulo
Apple	iPhone	nulo	nulo
Bancomer	BBVA	BBVA	BBVA
Banorte	Banorte	Banorte	nulo
HSBC	HSBC	HSBC	nulo
Levi's	501	501	nulo
Levi's	trucker	trucker	nulo
Levi's	Western	Western	nulo
Mazda	Mazda2	Mazada2	nulo
Mazda	Mazda3	Mazada3	nulo
Mazda	Mazda6	Mazada6	nulo
Microsoft	Azure	Azure	nulo
Microsoft	Office	Office	nulo
Nike	Jordan	Jordan	nulo

Cuadro 3.3: Fragmento de algunos productos publicitados en México.

Palabra clave	**Tipo**	**Concepto**
5G	Internet	Internet
Americano	deportes	entretenimiento
Basquétbol	deportes	entretenimiento
Béisbol	deportes	entretenimiento
Jacket	ropa	vestimenta
Chico	talla	vestimenta
Compacto	auto	automóvil
Ella	género	social
Ellos	género	social
Familia	género	social
Grande	talla	vestimenta
Hatchback	autos	automóvil
Jeans	ropa	vestimenta
Laptop	electrónicos	electrónica
Licuadoras	electrónicos	electrónica
Mediano	talla	vestimenta
Sedán	autos	vestimenta
Smartphones	electrónicos	electrónica
Smartwatches	electrónicos	electrónica
Hoodie	ropa	vestimenta

Estas palabras y símbolos que componen las expresiones regulares son los que se utilizan comúnmente en los anuncios en las páginas web [36]. Además, como la publicidad no solo juega con lo visual, sino también con las letras, tamaños y estilos, se amplió, sobre estas palabras, el rango de búsqueda con plurales, acentos y símbolos referentes a algunas palabras clave.

Es importante señalar que estas expresiones regulares de marca y producto están relacionadas con estudios de marca en México y algunos estudios en Latinoamérica, siendo escalables en el mundo. Para México, las estadísticas de 2019 y 2020 se buscaron en las bases de datos del Instituto Nacional de Estadística y Geografía (INEGI), que es un organismo autónomo del gobierno mexicano responsable de las estadísticas geográficas sobre los recursos, la población y la economía.

Otra fuente de datos fue la Comisión Económica para América Latina y el Caribe (CEPAL), que es una agencia de las Naciones Unidas que permite el acceso a la información en algunos países de América Latina. Las estadísticas arrojaron datos sobre la población económica más grande, es decir, personas con edades comprendidas entre los 25 y 29 años. Con base en estos datos, se identificaron las principales marcas de consumo en ese sector específico.

Por otro lado, la publicidad también juega con palabras relacionadas con las estaciones del año, optando por descuentos en productos que están fuera de temporada, o incluso con eventos o situaciones que se destacan en la región o en el mundo, como los trending topics, que son relevantes para patrocinios de algunas marcas, e incluir figuras públicas o deportistas para promocionar el lanzamiento, producto o marca.

3.4. Pseudocódigo del algoritmo

Con base en lo anterior, la idea principal del algoritmo es utilizar el desplazamiento automático, de manera que se pueda capturar la información contenida en las páginas web. Luego, estas imágenes capturadas son procesadas y transformadas a formato de texto, para luego identificar los anuncios existentes con base en coincidencias con las familias de palabras utilizadas en el marketing digital, definidas para este trabajo como expresiones regulares. Finalmente, se identifican las marcas más publicitadas y que son tendencia en las búsquedas a través de navegadores web. Esto es útil para realizar un seguimiento de estos. A continuación se presenta el pseudocódigo del algoritmo desarrollado:

```
Abrir el navegador con selenium
Ingrese la URL de la página deseada
iteration = 1
while true do:
    Altura de la página = Altura de la página web en píxeles según la función de Selenium
    Altura según el tamaño de pantalla de la computadora en píxeles
    Deslizar = Altura = Iteración
    Tomar capturas de pantalla con Selenium
    Guardar imagen con la nomenclatura designada para reconocer las capturas
    Deslizar el sitio de acuerdo con al número de píxeles del monitor (resizable)
end while
if Deslizar >= Alto del monitor then
    break
    iteración += 1
end if
Cerrar el navegador de prueba de Selenium
Buscar las capturas donde fueron guardadas
Se buscan los ficheros CSV y se pasana a forma de lista todas las rutas de los elementos de este archivo
Se realiza la conexión con la BD
for i in range (ruta donde están las capturas) do
    lista de las imágenes iteradas [i]
    lista - pytesseract.image_to_string (img).upper().split ()
    Separe la cadena de texto devuelta por Pytesseract en muchas cadenas más pequeñas con
    elementos atómicos, siendo palabras, porque el proceso de segmentación se realizó
    Cuando se encuentra un espacio. Estas palabras ahora están parseadas.
end for
PalabrasEncontradas = []
for j in range (list) do
    procedemos a realizar consultas para encontrar cada palabra en la "lista" en las 3 tablas de la BD
    Encontrar = Resultado de las queries devueltas por la base de datos.
    for list [j] do
        Iterar
    end for
    Separar la cadena de texto devuelta por Pytesseract
    Se borran los posibles espacios y se ponen en mayúsculas las palabras
end for
if largo (Encontrar)! -0 then
    palabrasEncontradas.append (lista [jl)
    Borrar las capturas ya analizadas
end if
Imprimir (Número de captura analizada)
Imprimir (Palabras encontradas sin repetir, su número de ocurrencias y tabla de precedencia)
Fin del algoritmo
```

En general, el algoritmo tiene tres etapas principales: i) debe obtener la URL de la página web a analizarse y luego tomar las capturas de pantalla mediante el desplazamiento automático; ii) luego, las imágenes de las capturas de pantalla se procesan en el orden en que fueron tomadas, para convertirlas a texto con la ayuda de OCR; y iii) se analiza el texto, obteniendo como resultado una lista de coincidencias con las expresiones regulares, almacenadas en tablas, sobre las empresas que anuncian, los productos y sus estrategias.

Como restricción, este trabajo no realiza Web Scraping, que es un proceso de recolección automática de datos e información de Internet, comúnmente de páginas web que utilizan código HTML. Además, no se incluyeron para la prueba extensiones añadidas al navegador web, ni ninguna cuenta vinculada para la sincronización con los dispositivos. Asimismo, no se consideraron los anuncios con extensión lateral, debido a que el deslizamiento es vertical, de arriba hacia abajo. Otra restricción fue que las capturas deben tener una buena resolución para que el reconocimiento óptico de caracteres sea efectivo en los resultados finales.

3.5. Mejoras al algoritmo

En una segunda etapa del desarrollo, dado que la propuesta es analizar los trending topics a partir del contenido de información que se genera en páginas web, redes sociales y foros de discusión en línea, se hicieron mejoras al funcionamiento del algoritmo para que periódicamente se pueda conocer hacia dónde se dirige la atención del marketing digital. Para esto, se amplió el funcionamiento del algoritmo a cinco etapas:

i. El usuario ingresa la URL de la página web, objeto de estudio. Esta se procesa para obtener el idioma del sitio web a través de la biblioteca *Beautiful Soup* de Python. A su vez, a través de Selenium, se accede al sitio mediante del navegador web, deslizando el cursor automáticamente, con el fin de dividir su contenido. En cada iteración de deslizamiento, se obtiene una captura de pantalla con una determinada nomenclatura.

ii. Al cerrar el navegador web de prueba, se ubican los archivos de captura de pantalla, previamente almacenados en una carpeta definida.

iii. Para cada captura de pantalla, se extrae el texto que esta contiene mediante reconocimiento óptico de caracteres. Para esto se utiliza la biblioteca *Python pytesseract*. Posteriormente, el texto se divide palabra por palabra, para su posterior análisis.

iv. Para el análisis del texto, almacenado en forma de expresiones regulares en archivos de texto plano, con formato CSV, se ejecuta el algoritmo para identificar palabras clave relacionadas con trending topics, movimientos sociales y COVID-19.

v. Esta identificación de palabras se basa en las coincidencias (intersección) del texto almacenado en los archivos CSV y la lista de palabras clave. Así, con base en esta intersección, se despliega el resumen de las palabras identificadas.

Por otro lado, se amplió la base de datos con palabras clave sobre temas de tendencias actuales. Además, se incluyeron palabras que aparecen antes y después de la palabra clave encontrada, lo que permite conocer cómo las empresas están tratando el tema y cómo lo están aprovechando. Por lo que, esto puede ser útil para desarrollar las propias estrategias de marketing digital, siguiendo los éxitos de los demás y evitando campañas de poco éxito o fracaso.

Capítulo 4

Resultados

4.1. Pruebas

Como parte de las pruebas, para el análisis de la publicidad web se consideraron tres tipos de páginas web dinámicas, probadas en tres tipos de navegadores: Chrome de Google, Mozilla Firefox y Safari de Apple. Los sitios web analizados fueron:

- MSN: `www.msn.com/es-mx`
- Sanborns: `www.sanborns.com.mx`
- Ahorra seguros: `https://ahorraseguros.mx`

La Tabla 4.1 resume los resultados obtenidos para cada URL en cada navegador web, el número total de palabras (expresiones regulares) que aparecen como publicidad en cada sitio web evaluado, el número de capturas de pantalla, y el tiempo de ejecución desde la apertura del navegador web hasta la finalización de la comparación.

Cuadro 4.1: Resultados de la evaluación del algoritmo en tres navegadores web.

Navegador web	**Publicidad**	**Capturas de pantalla**	**Tiempo (seg)**
URL 1: MSN – `www.msn.com/es-mx` –			
Chrome	106	9	11.636
Firefox	103	9	12.144
Safari	118	9	14.539
URL 2: Sanborns – `www.sanborns.com.mx` –			
Chrome	56	4	4.449
Firefox	62	4	4.036
Safari	68	4	4.209
URL 3: Ahorra seguros – `https://ahorraseguros.mx` –			
Chrome	133	9	13.547
Firefox	149	9	14.547
Safari	153	9	14.556

4.2. Discusión

Con base en los resultados obtenidos, se pudo identificar que a través del navegador web Safari, el algoritmo detectó una mayor cantidad de anuncios en comparación con los otros dos navegadores Chrome y Firefox. Esta mejor identificación de los anuncios se debe a que el algoritmo hace un mejor ajuste del contenido de la página web, y, por lo tanto, se tarda más en realizar las pruebas. En el caso de Chrome y Firefox, ambos también detectaron una cantidad importante de anuncios, pero al desplazarse por la página web, se perdía una pequeña cantidad de información.

Para determinar la eficiencia del algoritmo se realizó una revisión visual de la información contenida en las páginas web evaluadas. Esta revisión consistió en contar las expresiones regulares en los anuncios web, que debería coincidir con el número total de palabras detectadas por el algoritmo. La Tabla 4.2 resume los resultados obtenidos de la comparación de coincidencias entre las palabras detectadas por el algoritmo y el total de palabras existentes como parte del contenido publicitario en las páginas web.

Cuadro 4.2: Palabras identificadas por el algoritmo respecto al total de palabras con contenido publicitario.

RegExp/ Coincidencia	**Recuento visual**	**Chrome**	**Firefox**	**Safari**
URL 1: MSN – `www.msn.com/es-mx` –				
Microsoft	22	22	14	22
News	9	9	9	9
IOS	10	10	0	10
Android	10	10	0	10
MSN	12	3	10	2
Rebaja	15	13	12	14
Total	124	106	103	118
URL 2: Sanborns – `www.sanborns.com.mx` –				
$	27	18	23	24
Libros	3	2	1	3
Perfumes	2	2	2	2
Tecnología	3	3	3	3
Videojuegos	2	2	2	2
Total	75	56	62	68
URL 3: Ahorra seguros – `https://ahorraseguros.mx` –				
Seguros	57	48	54	56
Seguro	22	22	22	21
Beneficios	7	5	7	7
Precios	5	4	5	4
Servicios	5	2	1	4
Total	172	133	149	153

En el caso de MSN, URL 1, se logró un notable desempeño del algoritmo en la detección de publicidad web a través del navegador Safari, obteniendo un 95.16 % de confianza. Mientras que los navegadores Chrome y Firefox también alcanzaron porcentajes de confianza significativos, cuyos valores fueron 85.48 y 83.06 %, respectivamente. La diferencia en el nivel de confianza entre los navegadores web evaluados se debe a la forma en que concentran la información en la ventana, evitando cortes en la misma.

Para la URL 2, Sanborns, Safari fue donde alcanzó una efectividad de 90.66 %, Chrome 74.66 % y Firefox 82.66 %. Estos resultados se deben al llamativo diseño visual para el usuario, pero complicado de analizar debido a que algunas palabras se sobreponían, así como a la presencia de logos y palabras con diferente tamaño y tipografía concatenadas entre sí. Por lo tanto, fue una tarea difícil para el OCR y los resultados finales se vieron afectados.

En el caso de Ahorra seguros, URL 3, también se lograron resultados significativos con un 88.95 % de confianza en Safari, 86.62 % en Firefox y 77.32 % en Chrome. En esta

prueba, el factor principal para no lograr un mayor nivel de confianza fue la cantidad de logotipos de diferentes marcas con una variedad de fuentes y fondos. Este fue el principal problema al hacer el reconocimiento óptico de caracteres.

Las particularidades de los resultados obtenidos en las pruebas se deben a los cortes y ajustes de pantalla en el deslizamiento automático, es decir, varía la configuración en cada navegador web, cambiando la forma en que se organiza la información en el sitio web, y esto provoca pérdida de contenido.

Se logró mayor éxito a través de Safari, en comparación con Chrome y Firefox, esto debido a que este navegador web logra hacer, antes del deslizamiento automático para la captura de imágenes, una reorganización más rápida y compacta del contenido de la página web, lo que beneficia el desempeño del algoritmo en la detección de publicidad web.

Un problema importante en la detección de anuncios web se debe a la presencia de textura de fondo, concatenación de información, logotipos y otros diseños visuales dentro de los banners, lo que dificulta la extracción del contenido, malinterpretando las expresiones regulares.

4.3. Pruebas complementarias

Como pruebas complementarias, se analizaron algunas tendencias actuales a partir del contenido de información establecido en algunas páginas web. Por ejemplo, se probó el algoritmo en el sitio web oficial de Chedraui México. Se observó que la tienda comercial informaba a la comunidad que siguen las medidas de sanitización y que el cliente y su bienestar son siempre la prioridad. Esto representa una estrategia popular que ha demostrado ser exitosa en términos de ganar popularidad, además de brindar un sentimiento de confianza en los consumidores. La Figura 4.1 muestra parte del sitio web, objeto de estudio, con información relacionada sobre COVID-19.

Figura 4.1: Sitio web de Chedraui México con información relacionada sobre COVID-19.

Luego de la ejecución del algoritmo, se observó que las palabras relacionadas con la tendencia de COVID-19 fueron las famosas #QuédateEnCasa. También se observó que las expresiones que acompañan a la palabra clave, ubicada en la base de datos, fue que se preocupan por el bienestar de sus clientes, lo que ilustra perfectamente las intenciones de la empresa, que es promover la confianza en los consumidores y ayudar con la tarea de distanciamiento social.

Otra prueba se realizó con el sitio web de la marca alemana Adidas, que en su página oficial de Adidas México tiene una pestaña especializada para ver los servicios relacionados con el tema COVID-19. Esto también contrasta el uso de la coyuntura para lanzar productos de higiene y protección personal, como cubrebocas. La Figura 4.2 muestra un fragmento del sitio web de Adidas México con información relacionada sobre COVID-19.

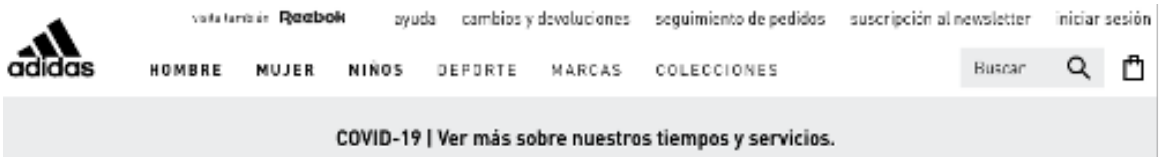


Figura 4.2: Fragmento del sitio web de Adidas México con información relacionada sobre COVID-19.

Al ejecutar el algoritmo, se observó que las palabras clave utilizadas fueron “COVID-19” y “Ver más sobre nuestros tiempos y servicios”, por lo que, la estrategia de mercado está encaminada a la venta de nuevos productos, como cubrebocas. Lo que confirma su estrategia de mercado ante la necesidad del producto.

Dejando de lado el tema de la pandemia por COVID-19, otra tendencia fue Pride, donde al igual que en el caso anterior, diferentes empresas lo vieron como un campo de

oportunidad para orientar el marketing digital para la venta de sus productos, como fue el caso de Calvin Klein (Figura 4.3). Al aplicar el algoritmo se observó que la venta de artículos relacionados con Pride incluye otras palabras antes y después con la intención de captar la atención de las personas y así lograr una venta exitosa. También se observó que en poco tiempo se agotaron algunos productos.

Figura 4.3: Sitio web de Calvin Klein México con información sobre la venta de artículos por la tendencia Pride.

Con base en lo anterior, para cada trending topic existen diferentes formas de llegar al público y cada una con distintas intenciones. Lo importante es observar la reacción que se genera en el público y específicamente en los consumidores potenciales, para así diseñar una mejor estrategia de marketing. En este sentido, a través de este tipo de algoritmos se pueden identificar tendencias y experiencias de otras marcas con campañas publicitarias exitosas.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

Los notables avances en la tecnología contemporánea también traen consecuencias negativas para el usuario final, como la invasión de anuncios en la web. Anuncios que se dirigen en función de las búsquedas, necesidades e intereses de los usuarios.

La publicidad web no consiste únicamente en palabras o frases que captan la atención del usuario con promociones u ofertas. Para los anunciantes es vital que el usuario sepa quién lo está promocionando, independientemente de si el cliente realmente piensa comprar dicho producto, pero lo más importante es llamar su atención y recordar la marca para futuras compras.

El uso de bloqueadores de anuncios refleja que la publicidad dirigida es un mecanismo que no se puede evitar sino moderar. La función de los bloqueadores de publicidad es solo ocultar anuncios, pero los algoritmos de Google logran cada vez más inundar una mayor cantidad de publicidad dirigida.

El uso de expresiones regulares fue útil, además, la implementación de la base de datos facilitó la organización para la detección de publicidad en la web, abarcando más casos de uso de anuncios.

Se encontraron los resultados esperados para las pruebas realizadas, con un porcentaje de efectividad de aceptable a alto, variando de 74.66 % a 95.16 %, y el mayor índice de confiabilidad se dio a través de MSN, debido al diseño simple, tipografía común, tamaños constantes, ausencia de palabras concatenadas, logos y diseños llamativos.

Sin duda, la eficacia del algoritmo en Safari es destacable por su forma de distribuir la información en las pantallas de los usuarios finales.

Usar este enfoque podría ser útil para que los anunciantes utilicen el algoritmo tantas veces como sea necesario, con el fin de saber cuál fue el anuncio, o promoción que causó

mayor efecto en los cibernautas, o en la competencia, en virtud del aumento del marketing digital para el mercado de consumo.

Es importante señalar que recopilar información de los usuarios no es una mala práctica, sino que debe ser con fines que los beneficien. La desventaja de tener publicidad molesta es que distraen la atención de los usuarios y falta una adecuada seguridad de la información.

Para la publicidad dirigida, se podrían cubrir otras opciones que no distraigan, por ejemplo, a través de correos electrónicos, secciones específicas en los navegadores y aplicaciones especializadas, donde los usuarios pueden consultar ofertas publicitarias.

5.2. Trabajo futuro

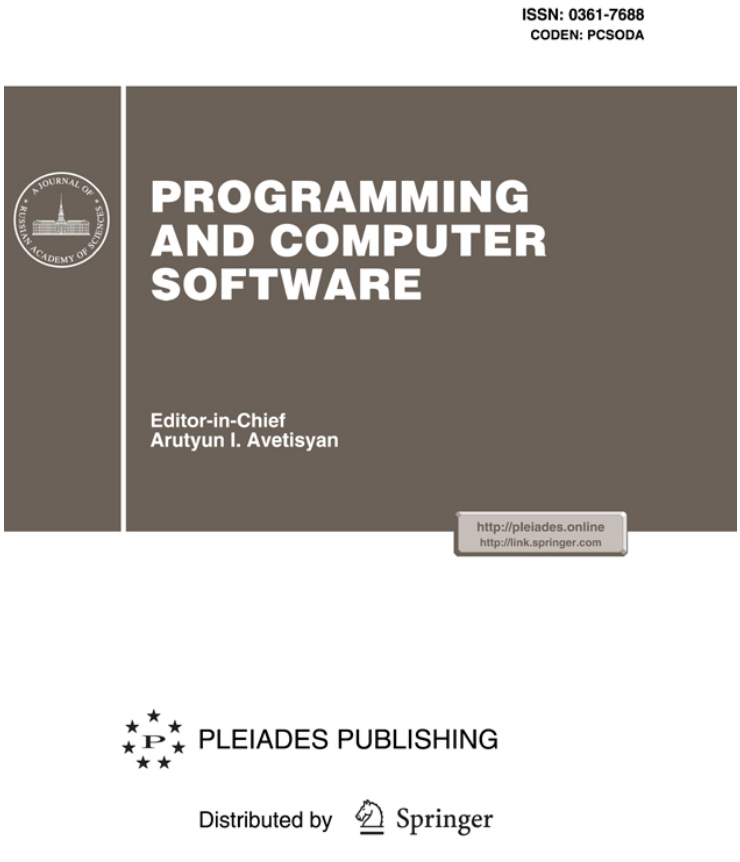
Como trabajo futuro, se pretende incluir más expresiones regulares en la base de datos y hacer una extensión en el algoritmo, es decir, incluir algoritmos de inteligencia artificial capaces de reconocer publicidad con base en patrones de color, tamaño y ubicación de banners, texto en negrita y tipografía, entre otras características en los anunciantes de hoy.

Por otro lado, se tiene contemplado agregar al algoritmo una herramienta para monitorear y comparar movimientos publicitarios de las empresas, con el fin de identificar campañas exitosas basadas en el índice de ventas, la presencia en la web, su enfoque y empatía con usuarios.

Apéndice A

Artículo publicado en *Programming and Computer Software*

Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm. URL: `https://link.springer.com/article/10.1134/S0361768820080162`



ISSN 0361-7688, Programming and Computer Software, 2020, Vol. 46, No. 8, pp. 652–660. © Pleiades Publishing, Ltd., 2020.

Regular Expressions for Web Advertising Detection Based on an Automatic Sliding Algorithm

D. Riaño[a],*, R. Piñon[a],, G. Molero-Castillo[a],***, E. Bárcenas[a],****, and A. Velázquez-Mena[a],*****

[a] *Engineering Faculty, UNAM Circuto Escolar 04360, C.U., Coyoacán, Ciudad de México, CDMX, 04510 Mexico*

**e-mail: donovan20@comunidad.unam.mx*

***e-mail: rodrigo_pinon@comunidad.unam.mx*

****e-mail: gmoleroca@fi-b.unam.mx*

*****e-mail: ebarcenas@unam.mx*

******e-mail: mena@fi-b.unam.mx*

Received April 12, 2020; revised May 16, 2020; accepted July 23, 2020

Abstract—This paper presents the automation of a Web advertising recognition algorithm, using regular expressions. Currently, the use of regular expressions, optical character recognition, Databases, and automation tests have been critical for multiple Software implementations. The tests were carried out in three Web browsers. As a result, the detection of advertisements in Spanish, that distract attention and that above all extract information from users was achieved. The main feature of the algorithm is that automatic and versatile execution does not require access to the code of the page in question and that in the future it can be an application with background operation. Being supported by optical character recognition gives us acceptable efficiency in detecting advertising. Thanks to this identification, it may be possible to generate different applications, both in favor of the user and the brands, always with the aim of improving current online marketing models.

DOI: 10.1134/S0361768820080162

1. INTRODUCTION

In the past, marketing only exists offline and its main objective was to coordinate the media, make deals so that people or even other companies have positive opinions about the products that are advertised, or about the ideas that are planned, to be sold. But now, with all the tools that technology has given us this is over, today users use search engines to find what they want and not only that, but they can also access criticism and comments made by the community.

As is logical to think, marketing strategies change completely as a result of new Web designs, which can lead to dynamic Web pages, in which people can mainly interact with websites, generating information about the tastes of users, their interests, and the needs of society in general.

Within digital marketing, the main objective is the user [1], and therefore the marketing techniques changed their paradigm. Today a digital strategy must include all the relevant spaces for the 'target' to interact, looking for people who influence their opinions to enter the same network of users, and can give more strength to ideas or products. This can also focus on improving search engines and, according to the experience acquired, becoming increasingly invasive in their ways of entering the minds of users.

After analyzing the structure, designs, colors, and marketing techniques, the basic idea of the ads on the Web is to get the attention of people of any possible way, using large banners, striking words, bright colors within the designs or, in some desperate cases, with the use of manual closing banners, so that once they have our attention they can say who they are and what they offer us.

Therefore, a semantic network was designed to understand the current digital marketing trends used in dynamic Web pages. It is important to mention that a semantic network is a form of representation of knowledge through interrelationships in the form of a graph [2]. Figure 1 shows the interrelationships of the semantic network, in which advertisements on Web pages can be divided into three groups: a) sites that subsist on advertising; b) informational, corporate and online stores; and c) social networks or similar.

This semantic network relates: i) who offers, ii) what product or service does it offer, and iii) in what way does it attract attention (offer or promotion). These types of advertising are distinguished when browsing the Inter-

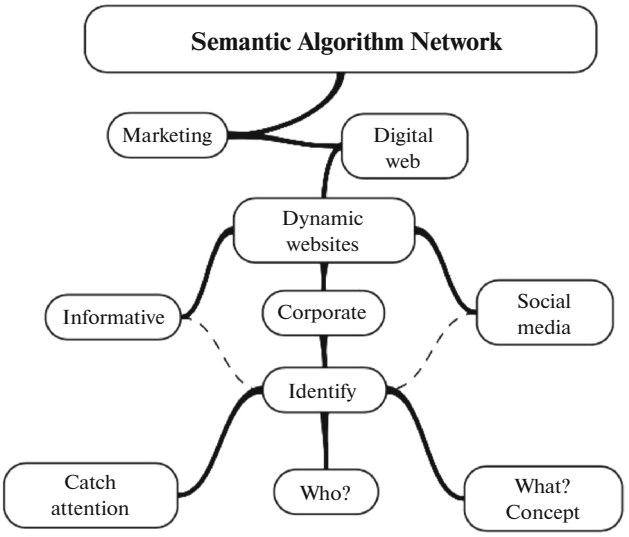


Fig. 1. Semantic network of current Web advertising.

net, which some of these were the object of analysis in this research work.

Regarding advertising in Web browsers, it is increasing every time. The storage, and interpretation of the large amount of data collected by search vectors put users' privacy at risk. Therefore, these data clusters could also provide useful information for better management of targeted advertising.

Websites that subsist from advertising offer a brief content of news, education, or entertainment and their ads are provided by Google. In these, the ads change constantly. In the case of informational, corporate, and online stores, these include sites such as MSN, Amazon, Sanborns, Adidas, and others, whose purpose is to offer current trends in products or services. While in the case of social networks and derivatives such as Facebook, Instagram, LinkedIn, Twitter, and others, they are sites that offer interaction with users and ads related to the user profile.

In all the mentioned websites, the user becomes a possible client when entering the advertising of the company or trade, where it is very likely that he will find more offers or promotions, therefore, more products or services, which increases the possibility of purchase, separated from an advertisement on the Web. This possibility of purchase is also due to the fact that an advertisement on the Web may remain in our memory for a long time, which will later cause the search or purchase of the product or service.

For this reason, these themes were used in the development of the algorithm. This paper presents the automation of a Web advertising recognition algorithm, using regular expressions. The tests were carried out in three browsers: Chrome, Firefox, and Safari. A feature of the algorithm is its automatic and versatile execution, since it does not require access to the code of the Web page in question and that it is an application that operates in the background. As a result, the detection of advertisements in Spanish was achieved, which distracts attention and, above all, can extract information from users when they browse the Internet.

2. BACKGROUND

Given the potential and strategies of current digital marketing, this activity is used more frequently, as an important part of brand loyalty campaigns, since it is a communication channel that achieves the interaction between the potential client and brand. Among the actions focused on online marketing include [3]:

– Customer loyalty.

– Increase the image of the brand and its sales.

– Generate promotions and product tests.

– Encourage the repeat purchase of the product.

– Conduct a direct and personalized communication campaign.

One way to deal with advertising and privacy in Web browsers is through regular expressions (RegExp, Regex, or RE). These expressions are a way of representing character strings that fit a certain pattern [4]. In addition, they are a flexible and efficient mechanism for word processing [5]. Its applications are diverse, for example, validation of form fields, identification of text strings in social networks, search commands, among others [6, 7].

2.1. User Privacy Management

At present, information and privacy management are in a critical stage [8]. Companies such as Facebook, Twitter, Google, Amazon, among others, have included in their websites, applications, or search engines for information hoarding. These companies provide ‘free’ for the service, but use the information of the users, at convenience, for advertising purposes. Among the pillars to have adequate information security stand out [9]:

– Authentication. To identify the communicating entity and the data source.

– Access control. To prevent unauthorized use of resources.

– Confidentiality To protect the data against unauthorized disclosure.

– Integrity. To guarantee the non-alteration or destruction of the data in an unauthorized manner.

– Not repudiation. To give proof of the origin of the data or its delivery.

– Availability. To ensure continuity of accessibility and use by authorized entities.

These services are provided through security mechanisms alone or in combination, such as encryption, digital signature, mechanisms for access control, mech-

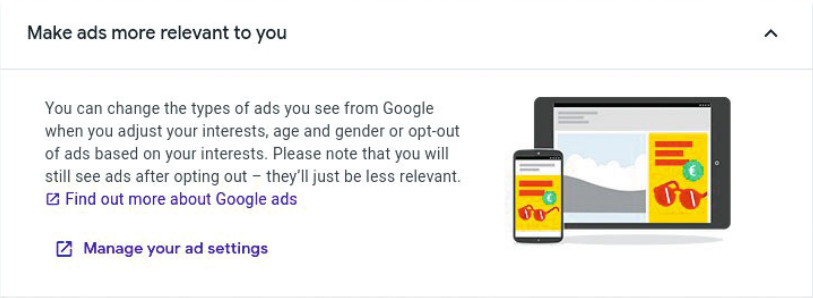


Fig. 2. Advertising settings related to a user Gmail account.

anisms for data integrity, authentication exchange, traffic filling, routing control, among other actions.

With regard to advertising in Web browsers, as it is increasing, it puts the privacy of users at risk. That is, when searching the Web for certain types of services or products, today there is a saturation of information on the website where it is positioned. The problem not only influences information management but also the slow flow of information due to advertising issues. Another important aspect is that user searches generate a large amount of collected data. Therefore, these data, under strict treatment, could also provide useful information for better management of targeted advertising [9].

In this way, a more specific Web advertisement is scored and brought closer according to the interests and needs of each user, taking into account, for example, the Google update. This update of Google services, in terms of advertising, includes manual adjustments to the interests and demands of each user, of course, suggested with the information collected by PageRank, trackers, and cookies. Figure 2 shows an extract of the notification of the said update, as part of the service settings, which includes the configuration of profiles, passwords, contacts, information, among others.

This extension is based on RegExp principle and takes into account the Web browsing of users, by which the hypothesis is to include the brands that pay most for an advertisement on the Web, or even the ones that sell most in a certain market, as well as the products, services, and its secondary brands [10].

As an example, Fig. 3 shows the Google search engine for the automatic creation of user profiles based on Web browsing. For example, a male user is relevant, with an age range between 18 and 24 years, whose interest is to buy online, do outdoor activities, develop mobile applications, watch movies, and others. This user, through the collection of data through cookies, trackers, and information contained in the Gmail profile, can be used to group profiles with common characteristics and, therefore, receive advertising for various offers and promotions.

Based on the above, therefore, when we constantly interact with Google services, it is advisable to review the data and privacy policies frequently about the adjustments made by Mountain View, in order to have greater knowledge about the security and use of our information.

2.2. Ad-Blockers of Web Browsers

In recent years, efforts have been increased to implement Web advertising blockers [11]. These were initially implemented for the Firefox browser, which over the years was improving to achieve a browser without ads. But the companies sought ways to continue sending advertising. Consequently, in a short time, they began to create controversies [1]. Controversies for significant losses of possible income. Therefore, improvements were also made in advertising blockers, currently known as Ad-blockers [12].

Another term that is also currently used is *Trackers* in marketing, which are indicators and trackers of the effectiveness of targeted advertising campaigns. These tools store information through cookies, and these, in turn, provide the location of the searches performed. Thus, as demand grew and the possibilities of blocking advertising, new Ad-blockers were implemented in most Web browsers, as well as in applications for mobile devices, the advertising has found a way to make announcements about user searches.

An important fact to consider is that, in recent years, Google controls 85% of the global search engine advertising business and about 50% of all online advertising [13]. People and society, in general, see Google as a service [14], but behind that, there is a technology that contains specific functions and includes exclusive and restrictive extensions.

In this sense, advertising and online shopping are increasingly demanded and fast, for example, online shopping saves time and distance. However, the use of websites for this type of purchase serves companies to learn from the experiences and needs of users. These are translated as patterns of behavior in different social extracts, either local or regional.

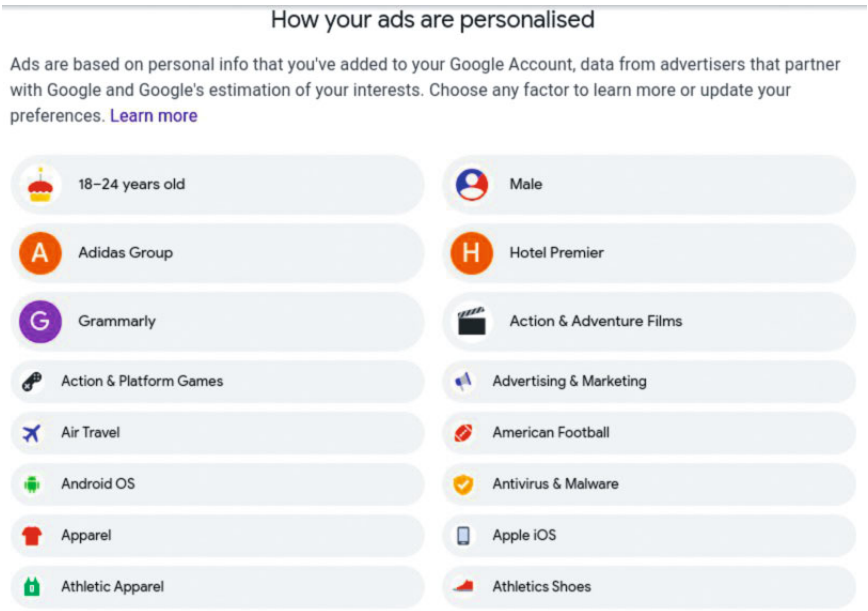


Fig. 3. Google search engine for automatic creation of user profiles based on Web browsing.

2.3. Selenium Automation

Selenium, also known as Selenium Webdriver, is a tool to automate processes in different Web browsers [15]. Its purpose is to improve support for detecting problems in any Web browser [16]. This tool allows testing any Web browser to obtain HTML code data, change, open and move between tabs of the browser windows, return or advance according to the test history, resize the windows, take screenshots, complete fields, clicks on parts of a website, among others.

These tasks are applicable to Java, Python, C#, Ruby, Perl, and JavaScript programming languages. The operating systems that Selenium supports are Windows, Mac OS, and Linux, each with their respective packages and Integrated Development Environments (IDEs) [17].

On the other hand, at present there is interest in using Optical Character Recognition (OCR) in the detection of Web advertising [14], however, more efforts are still needed to cover all digital marketing strategies. Therefore, it is necessary to seek the automation of this process and generate a correct strategy for the identification and classification of what is offered and who offers it.

2.4. Related Works

One of the works focused on the detection of Web advertising was [18], where an algorithm that performs Web crawling is presented. It consists of obtaining information from websites through tagging, the test Web page was MSN. Based on the labels, a classification is made using a probability model based on logarithmic regressions. The classification is made through keywords throughout the Web page, that is, at the beginning (B-beginning), in the middle (I-inside), at the end (l-last), unique (U), or outside (O).

[19] describes the contextual advertising analysis through PageSense, which aims to associate ads on Web pages. Through this platform, blank regions are detected and the non-intrusive area is selected for ad placement without breaking the original style of the Web page. Bayesian combinations and probabilities were used for the analysis, which reflects the percentages of advertisements for different types of products or services and, therefore, defines annoying and acceptable ads.

In another work, in [20], an analysis was made based on Euclidean distances. These distances were with respect to the way the ads are of interest to users, the search for products, and the adaptation of the objective profile, dividing it by sections, such as health, sports, business society, education, art, science, computer, among others.

On the other hand, [21] describes the analysis carried out on approximately 500 Web pages, in which tests aimed at detecting ad types, but not content was carried out. Among the types of ads analyzed, pop-ups, carousels, videos, gifs, games, stickers, or text stand out. The countries where the ads come from, the frequency, the size, and the origin of the URLs (Uniform Resource Locator) were also analyzed.

3. METHOD

The implementation of the algorithm was in Python. This algorithm has been assisted by the Tesseract [22], Pillow, and OpenCV [23] libraries, as well as the Tesseract-OCR package. For the implementa-

tion, we start from a point before the OCR process, that is, optical character recognition. In addition, we used the Selenium library to perform the Automatic sliding in the Web browser.

In the implementation, the height difference between the different sizes of monitors on the market was also taken into account, therefore, the algorithm performs a dynamic detection of the height of the windows, thus adjusting to any screen size. Therefore, the displacement of information in this work is vertical.

It should be noted that the tests were done in three Web browsers: a) Google Chrome, b) Mozilla Firefox, and c) Safari.

3.1. Website Sliding

Through Selenium, a new window opens for a browser compatible with this tool. The terminal asks the user for the URL field. Subsequently, the browser under test expands to full screen for a quick and complete scan of the website.

As part of the algorithm, and with the aim of stopping the process for a few moments, programming threads are used. The first pause is made to allow the page to load, since depending on the speed of the Internet, and even on the state of the page, it takes a loading time to perform the correct screenshot for subsequent analysis using the OCR.

In the next step, the window is maximized, before sliding to capture and store screens. This causes a second pause, which is 0.5 seconds. Subsequently, it starts the screen capture process until it finishes with all the vertical content of the Web page. Once the capture process is complete, the Web browser window where the query was made closes automatically.

3.2. Optical Character Recognition

Optical character recognition, OCR, allows extracting text from an image with alphabetical writing, regardless of the language, size, or color of the text, with high effectiveness. The effectiveness of the OCR ranges from 71 to 98%. This system is capable of reaching average values of 85.1% for handwritten text and 90.93% for printed or digital text [24].

A specific function was defined for the search of all the images of the screenshots saved with an established path. Then, a cycle was implemented where all the images are analyzed within the established route in the order in which they were captured, this due to the nomenclature that was used to save them.

When the desired image is accessed, the OCR results are saved in a variable, which consists of text that is then formatted, placing everything in capital letters, separating the words, eliminating spaces and characters not supported by the language SQL like "', \,. and &", this to be able to compare with regular expressions.

3.3. Regular Expressions

Once the Web content information has been obtained and synthesized into text strings, the local server is accessed to validate the website content with a database, which houses around 600 different words based on regular expressions distributed in three tables, and whose main objective is to validate the following topics:

– Words most used in digital marketing.

– Brands, considering their respective sub-brands in the products or services offered by the company.

– The type or what the product or service consists of.

These regular expressions correspond to the most used digital marketing words in Spanish, which, as mentioned, were used for comparison with previously obtained text strings. Tables 1, 2, and 3 show an fragment of the regular expressions related to the most used words in digital marketing, some recognized brands, and some products that are published more frequently in Mexico; respectively.

These words and symbols that make up regular expressions are those that are commonly used in advertisements on the Web [25]. Furthermore, since advertising not only plays with the visuals but also with the letters, sizes, and styles, the search range with plurals, accents, and symbols referring to some keywords was extended over these words.

It is important to note that these regular expressions of brand and product are related to brand studies in Mexico and some studies in Latin America, being scalable in the world. For Mexico, the statistics for 2019 were searched in the databases of the National Institute of Statistics and Geography (INEGI, by its acronym in Spanish), which is an autonomous body of the Mexican government responsible for geographic statistics regarding resources, population, and the economy.

Another source of data was the Economic Commission for Latin America and the Caribbean (ECLAC), which is an agency of the United Nations that allows access to information in some Latin American countries. Statistics showed data on the largest economic population, that is, people ranging from 25 to 29 years old. Based on these data, the main consumer brands in that specific sector were identified.

On the other hand, advertising also plays with words related to the seasons of the year, opting for discounts on products that are out of season, or even with events or situations that stand out in the region or around the world, which are relevant to sponsorships of some brands, and include public figures or athletes to promote the launch, product or brand.

3.4. Pseudocode

Based on the above, the basic idea of the algorithm is to use automatic scrolling, so that the information

Table 1. Fragment of the most used words in digital marketing in Spanish

Word	Plural	Accent	Character
Ahorro	Ahorros	null	null
Bajo	Bajos	null	null
Comprar	null	null	Compra
Cotiza	null	null	Cotizar
Descuento	Descuentos	null	%
Dinero	Dineros	null	null
Especial	Especiales	null	null
Gratuito	Gratuitos	null	Gratis
Hasta	null	null	null
Ilimitado	Ilimitados	null	null
Interes	Intereses	Interés	null
Internet	null	null	Web
Oferta	Ofertas	null	null
Plan	Planes	null	null
Precio	Precios	null	null
Producto	Productos	null	null
Punto	Puntos	null	null
Rapido	Rapidos	Rápido	Rápidos
Rebaja	Rebajas	null	null
Salud	null	null	null

Table 2. Fragment of some of the best-selling and highest-paid brands in Mexico

Brand	Sub-brand	Product	Acronym
Adidas	Boost	Boost	null
Adidas	NMD	NMD	null
Adidas	Originals	Originals	null
Apple	iMac	iMac	null
Apple	iPad	iPad	null
Apple	iPhone	iPhone	null
Bancomer	BBVA	BBVA	BBVA
Banorte	Banorte	Banorte	null
HSBC	HSBC	HSBC	null
Levis	501	501	null
Levis	Trucker	Trucker	null
Levis	Western	Western	null
Mazda	Mazda2	Mazda2	null
Mazda	Mazda3	Mazda3	null
Mazda	Mazda6	Mazda6	null
Microsoft	Azure	Azure	null
Microsoft	Office	Office	null
Microsoft	Outlook	Outlook	null
Nike	Jordan	Jordan	null

contained in Web pages can be captured. Then these captured images are processed and transformed into text format to later identify the existing advertisements based on filters and matches with the word families used in digital marketing as regular expressions. Finally, the most advertised brands are identified and are trending in searches through Web browsers. This is useful for keeping track of these. Figure 4 presents the pseudocode of the algorithm developed.

In general, as already described in previous sections, the algorithm has three main stages: i) you must obtain the URL of the page, and then take the screenshots through automatic scrolling; ii) the screenshots images are then processed, in the order in which they were taken, to convert them to text with the help of OCR; and iii) the text is analyzed, obtaining, as a result, a list of coincidences with the regular expressions, stored in tables, about the companies that advertise, the products and their strategies.

As a restriction, this work does not make Web Scraping, which is a process of automatic collection of data and information from the Internet, commonly on Web pages that use languages such as HTML, whose data is analyzed for certain needs and purposes [25]. Thus, no personal extensions added to the Web browser were included for the test, nor any linked account for synchronization with the devices. In addition, the ads with lateral extension were not considered, because the sliding is vertical, from top to bottom.

A final restriction is that the captures need to be of a good resolution so the OCR can be effective in the final results.

4. RESULTS

For the analysis of Web advertising, three types of dynamic Web pages were considered, tested in three different browsers: Chrome by Google, Mozilla Firefox, and Safari by Apple Inc. These analyzed websites were:

– MSN: www.msn.com/es-mx

– Sanborns: www.sanborns.com.mx

– AhorraSeguros: https://ahorraseguros.mx

Table 4 summarizes the results obtained for each URL in each Web browser, the total number of words (regular expressions) that appear as advertising on each evaluated website, the number of screenshots, and the execution time since the opening of the Web browser until final comparison.

Based on the results obtained, it was possible to identify that through the Safari browser the algorithm detected a greater number of advertisements compared to the other two Chrome and Firefox browsers. This better identification of advertisements is due to the fact that the algorithm makes a better adjustment of the content of the Web page, and therefore it takes more time to perform the tests. In the case of Chrome and Firefox, both also detected a significant amount of

advertisements, but when scrolling the Web page, a small amount of information was lost.

To determine the efficiency of the algorithm, a visual review of the information contained in the evaluated Web pages was performed. This review consisted of counting the RegGex in the Web ads, which should match the total number of words detected by the algorithm. Table 5 summarizes the results obtained from the comparison of matches between the words detected by the algorithm and the total existing words as part of the advertising content in the Web pages.

In the case of www.msn.com/es-mx, URL 1, a remarkable performance of the algorithm in detecting Web advertising through the Safari browser was achieved, obtaining a 95.16% confidence. While Chrome and Firefox browsers also reached significant confidence percentages, whose values were 85.48 and 83.06%, respectively. The difference in the level of confidence between the evaluated Web browsers is due to the way they concentrate the information in the window, avoiding cuts in it.

For URL 2, www.sanborns.com.mx, where Safari reached effectiveness of 90.66%, Chrome 74.66%, and Firefox 82.66%. These results are due to the striking visual design understandable to human beings but complicated to analyze because some words collided, as well as the presence of logos and words with different size and typography concatenated with each other. Therefore, it was a difficult task for the OCR and the final results were affected.

In the case of https://ahorraseguros.mx, URL 3, significant results were also achieved, with an 88.95% confidence in Safari, 86.62% in Firefox, and 77.32% in Chrome. In this test, the main factor in not achieving a higher level of confidence was the number of logos of different brands with a variety of fonts and backgrounds. This was the main problem when doing optical character recognition.

The particularities of the results obtained in the tests are due to the cuts and screen adjustments in the automatic sliding, that is, the configuration in each Web browser is varied, changing the way in which the information on the website is organized, and this causes loss of content.

Greater success was achieved through Safari, compared to Chrome and Firefox, this because this Web browser manages to do before the automatic sliding for the capture of images, a faster and more compact reorganization of the content of the Web page, which benefits the performance of the algorithm in detecting Web advertising.

A significant problem in detecting Web ads is due to the presence of background texture, concatenation of information, logos, and other visual designs within the banners, which makes it difficult to extract the content, misinterpreting regular expressions.

Table 3. Fragment of some products advertised in Mexico

Key	Type	Concept
5G	Internet	Internet
Americano	Deportes	Entretenimiento
Basquetbol	Deportes	Entretenimiento
Béisbol	Deportes	Entretenimiento
Chamarra	Ropa	Vestimenta
Chico	Talla	Vestimenta
Compacto	Autos	Automóvil
Ellas	Género	Social
Ellos	Género	Social
Familia	Género	Social
Grande	Talla	Vestimenta
Hatchback	Autos	Automóvil
Jeans	Ropa	Vestimenta
Laptops	Electrónicos	Electrónica
Licuadoras	Electrónicos	Electrodomésticos
Mediano	Talla	Vestimenta
Sedan	Autos	Automóvil
Smartphones	Electrónicos	Electrodomésticos
Smartwatch	Electrónicos	Electrónica
Sudadera	Ropa	Vestimenta

Table 4. Results of the algorithm evaluation in three Web browsers

Web Browser	Advertising	Total Screenshots	Time (s)
URL 1: MSN – www.msn.com/es-mx –			
Chrome	106	9	11.636
Firefox	103	9	12.144
Safari	118	9	14.539
URL 2: Sanborns – www.sanborns.com.mx –			
Chrome	56	4	4.449
Firefox	62	4	4.036
Safari	68	4	4.209
URL 3: Ahorra Seguros – https://ahorrcaseguros.mx –			
Chrome	133	9	13.547
Firefox	149	9	14.547
Safari	153	9	14.556

5. CONCLUSIONS

The remarkable advances in contemporary technology also bring negative consequences for the end-user, such as the invasion of advertisements on the

```
Open a browser with selenium
Enter the URL of the desired page
iteration = 1

while true:

    Page Height = Height of the web page in pixels
    according to selenium function
        Height = High computer screen size in pixels
            Slip = Height * iteration
                Capture with selenium tools
                    Save image with the capture number name
                        Slide page according to the number
                        of pixels indicating that Slide
    if Slip> = Page Height:
            break
                iteration +=1
Close selenium browser
Search file where captures were saved
Add in a list all the paths of the elements of this file
Connection is made to the database

for i in range (capture path list):
    image = capture path list [i]
        list = pytesseract.image_to_string (
        img).upper().split ()
            #Separate the text string returned by
            pytesseract into many smaller strings
                #with atomic elements, that are words,
                because the segmentation process was
                carried out
                    #when a space is found. These words
                    are now capitalized.
    WordsFound = []
        for j in range (list):
                #We proceed to make queries to find each
                word in "list" in the 3 tables of the base
                        Find = Result of queries looking
                        for list [j]
        if length (Find)! = 0:
                    WordsFound.append (list [j])
                        Delete Capture already analyzed
    print (Capture number analyzed)
        print (Found Words without repeating, their number
        of occurrences and precedence table)
```

Fig. 4. Pseudocode of the implemented algorithm.

Table 5. Words identified by the algorithm with respect to the total of words with advertising content

RegExp/ Coincidence	Visual Count	Chrome	Firefox	Safari
URL 1: MSN – www.msn.com/es-mx –				
Microsoft	22	22	14	22
News	9	9	9	9
IOS	10	10	0	10
Android	10	10	0	10
MSN	12	3	10	2
Rebaja	15	13	12	14
Total	124	106	103	118
URL 2: Sanborns – www.sanborns.com.mx				
$	27	18	23	24
Libros	3	2	1	3
Perfumes	2	2	2	2
Tecnología	3	3	3	3
Videojuegos	2	2	2	2
Total	75	56	62	68
URL 3: Ahorra Seguros – https://ahorraseguros.mx –				
Seguros	57	48	54	56
Seguro	22	22	22	21
Beneficios	7	5	7	7
Precios	5	4	5	4
Servicios	5	2	1	4
Total	172	133	149	153

Web. Advertisements that are directed based on searches, needs, and interests of users.

Web advertising does not only consist of words or sentences that capture the user's attention with promotions or offers. For advertisers, it is vital that the user knows who is promoting it, regardless of whether the client really plans to buy said product, but the most important thing is to get your attention and remember the brand for future purchases.

The use of Ad-blockers reflects that targeted advertising is a mechanism that cannot be avoided but moderated. The function of advertising blockers is only to hide ads, but Google's algorithms increasingly manage to flood a greater amount of targeted advertising.

The use of regular expressions was useful, in addition, the implementation of the database facilitated the organization for the detection of Web advertising, covering more cases of use of advertisements.

Expected results were found for the tests performed, with a percentage of reliability from acceptable to high, ranging from 74.66% to 95.16%, and the highest reliability rate was given through MSN, due to the simple design, common typography, constant sizes, absence of concatenated words, logos and striking designs.

Undoubtedly, the effectiveness of the algorithm in Safari is remarkable due to its way of distributing information on the screens of the end-users.

Using this approach could be useful for advertisers to use the algorithm as many times as necessary, in order to know which was the announcement, or promotion that caused the greatest effect on cybernauts, or that of their competition, by virtue of increasing digital marketing for the consumer market.

It is important to note that collecting information from users is not a bad practice, but that it should be for purposes that benefit them. The disadvantage of having annoying publicity is that they distract the attention of users and lack of adequate information security.

For targeted advertising other non-distracting options could be covered, for example, through emails, specific sections in browsers, and specialized applications, where users can consult advertising offers.

As future work it is intended to include more regular expressions in the database and make an extension in the algorithm, that is, to include artificial intelli-

gence algorithms capable of recognizing advertising based on color patterns, size, and location of banners, text in bold and typography among other features in today advertisers.

ACKNOWLEDGMENTS

This work was supported by UNAM-PAPIIT IA105320.

REFERENCES

1. Marketing Digital, ¿Qué es el marketing digital?, 2020. http://www.mdmarketingdigital.com/que-es-el-marketingdigital.
2. Redes Semánticas, http://tesis.uson.mx/digital/tesis/docs/9049/Capitulo1.pdf.
3. Marketing Online: Potencial y Estrategias, 2019. http://www.cecarm.com/Guia_Marketing_Online_Potencial_y_Estrategias_-_CECARM.pdf-6120.
4. Pomol, R., González, C., and González, S., Una herramienta didáctica para el aprendizaje interactivo de expresiones regulares, 2013. http://repositorio.uigv.edu.pe/handle/20.500.11818/804.
5. Beltrán, R., El uso de expresiones regulares en la detección de errores escritos: implicaciones para el diseño de un corrector gramatical, 2008. https://dialnet.unirioja.es/servlet/articulo?codigo=4007478.
6. Gallego, A., La jerarquía de Chomsky y la facultad del lenguaje: consecuencias para la variación y la evolución, *Teorema*, 2008, vol. 27, no. 2, pp. 47–60.
7. García, I., Herramienta para la corrección automática de autómatas finitos, 2017. https://riull.ull.es/xmlui/handle/915/5846.
8. Sánchez, J., López, L., and Martíez, J., Solución para garantizar la privacidad en el Internet de las Cosas, *El profesional de la informaciyn*, 2015, vol. 24, pp. 62–70.
9. Ortiz, M., Aguilar, L., and Marín, L., Los desafíos del marketing en la era del big data, *e-Ciencias de la Informaciyn*, 2016, vol. 6, pp. 1–30.
10. Riaño, D., Molero-Castillo, G., Velázquez-Mena, A., and Bárcenas, E., Expresiones regulares para el tratamiento de privacidad de navegadores Web, *Abstr. Appl.*, 2019, vol. **25**, pp. 121–130.
11. Cerezo, P., Ad blocking: el modelo publicitario digital, a revisión, *Cuadernos de periodistas: revista de la Asociaciyn de la Prensa de Madrid*, 2016, pp. 81–89.
12. Londaitz, A., Publicidad en los celulares: publicidad invasiva vs. derecho a la privacidad, *Thesis*, Universidad del Salvador, 2011. https://racimo.usal.edu.ar/4312.
13. Bienvenido a Google, la mejor empresa para trabajar, 2013. http://www.expansion.com/2013/08/23/directivos/1377273795.html.
14. Jarvis, J., Y Google, ¿cómo lo haría?, 2000. https://narrativabreve.com/2013/10/libro-google-jeff-harvis.html.
15. Leotta, M., Clerissi, D., Ricca, F., and Spadaro, C., Comparing the maintainability of selenium webdriver test suites employing different locators: a case study, *Proc. 1st Int. Workshop on Joining AcadeMiA and Industry Contributions to Testing Automation,* Lugano, 2013. https://dl.acm.org/doi/10.1145/2489280.2489284.
16. Gojare, S., Joshi, R., and Gaigaware, D., Analysis and design of selenium WebDriver automation testing framework, *Procedia Comput. Sci.*, 2015, vol. 50, pp. 341–346.
17. Selenium Webdriver, 2017. http://www.tutorialspoint.com/selenium/pdf/selenium_webdriver.pdf.
18. Yih, W., Goodman, J., and Carvalho, V., Finding advertising keywords on web pages, *Proc. 15th Int. Conf. on World Wide Web,* Edinburgh, 2006. https://dl.acm.org/doi/pdf/10.1145/1135777.1135813.
19. Mei, T., Li, L., Tian, X., Tao, D., and Ngo, C., PageSense: toward stylewise contextual advertising via visual analysis of web pages, *IEEE Trans. Circuits Syst. Video Technol.,* 2018. http://dl.acm.org/doi/abs/10.1109/TCSVT.2016.2598702
20. Sánchez, D. and Viejo, A., Privacy-preserving and advertising-friendly web surfing, *Comput. Commun.*, 2018, vol. 130, pp. 113–123.
21. Krammer, V., An effective defense against intrusive web advertising, *Proc. 6th Annu. Conf. on Privacy, Security and Trust,* Fredericton, NB, 2008. https://ieeexplore.ieee.org/document/4641268.
22. Sajjad, K., Automatic license plate recognition using Python and Opencv, College of Engineering, 2010. https://pdfs.semanticscholar.org/bddf/1200eb17f239e4dce2a9cec938eb8cf305f5.pdf.
23. Patel, C., Patel, A., and Patel, D., Optical character recognition by open source OCR tool tesseract: a case study, *Int. J. Comput. Appl.,* 2012, vol. 55, no. 10. https://research.ijcaonline.org/volume55/number10/pxc3882784.pdf.
24. Vallez, M., Keyword research: métodos y herramientas para identificar palabras clave, *BiD: textos universitaris de biblioteconomia i documentació*, 2011, vol. 27, pp. 1–14.
25. Slamet, C., Andrian, R., Maylawati, D., Darmalaksana, W., and Ramdhani, M., Web scraping and naïve Bayes classification for job search engine, *Proc. 2nd Annu. Applied Science and Engineering Conf.,* Bandung, 2018. https://iopscience.iop.org/article/10.1088/1757-899X/288/1/012038/pdf.

Apéndice B

Artículo publicado en *LACCEI*

Algorithm for Identification and Analysis of Targeted Advertising used in Trending Topics.
URL: `https://laccei.org/LACCEI2022-BocaRaton/full_papers/FP57.pdf`

Algorithm for Identification and Analysis of Targeted Advertising used in Trending Topics

Donovan Riaño Enriquez, BSc; Guillermo Molero-Castillo, PhD;
Everardo Bárcenas, PhD; and Rocío Aldeco Pérez PhD
Universidad Nacional Autónoma de México, México
donovanriano@ingenieria.unam.edu, guillermo.molero@ingenieria.unam.edu,
ismael.barcenas@ingenieria.unam.edu, rocio.aldeco@ingenieria.unam.edu

***Abstract*– Today, companies choose to display their advertising content on Web pages, using all the resources that technology offers, whether through eye-catching images, animated images, and even videos. The objective is to ensure that the information not only reaches the public but that their ideas travel to target audiences, to people who can potentially consume your products or services. Thus, at present, technology has revolutionized digital marketing, having as greatest exponents of effectiveness to Web pages and social networks, where the information traffic generated by users is analyzed, for example, through trending topics. A clear example that has been observed in 2020 and 2021 were the trends that have shaken the world, such as COVID-19, changing the way any business or organization works, apart from its impact on health. However, brands, apart from seeking empathy with consumers, designed their advertising campaigns to increase their sales. This paper presents an algorithm for the identification and analysis of advertising patterns directed by some brands that have sought to increase their sales successfully, having as a message the awareness of its consumers through displaying content in banners and advertisements, and thus achieve the attention of users.**

Keywords- Advertising, Digital Marketing, Trending topics, Pandemic, Metaverse, NFTs, Web3.

I. Introduction

Over time, the Internet and Technology have had a great impact on the advertising world, markets, and business models. Before, advertising was only done according to traditional media. However, this scenario took a different turn with the Internet age. Likewise, this phenomenon stimulated an impact on advertising, where companies identified the usefulness and importance of using this mechanism as a tool to achieve greater reach with a target audience, that is, through targeted advertisement.

Thus, in the last two years, 2020 and 2021, trends have been seen that have shaken the world like never before. One of these trends, which has come to change any business or organization, in addition to its impact on health, was the arrival of the coronavirus (COVID-19) pandemic **[1]**.

A highly contagious, dangerous, and complex disease to manage due to its asymptomatic onset and the complexity of early detection **[2]**. One of the actions that world leaders took was to stay at home to avoid, as much as possible, contact and, therefore, possible infections. This while we manage to obtain some vaccine or a way to cope with the situation. Nevertheless, faced with these desperate measures, coupled with the fall of the world economy, digital marketing did not stop, it had greater momentum, gaining widespread notoriety in actual society.

This notoriety of digital marketing is due, in part, to the measures taken by governments, who warned the general population to keep a reasonable distance, maximize hygiene measures and expose themselves as little as possible to contact with other people **[3]**. These measures evidently generated fear and uncertainty in the population, which generated, on the one hand, a constant media bombardment on the subject and, on the other, that large companies took advantage of the situation to boost their home delivery sales, for example, first necessity, cleaning, and disinfection items, face masks, home entertainment media, pet food, among others.

On the other side, while many businesses that depended on human contact had a huge drop in their economy, large companies, such as supermarkets, pharmacies, and even clothing brands took advantage of the uncertainty to launch promotions and new merchandise **[4]**; thus managing to flood the advertising media to reduce sedentary lifestyle and ensure maximum hygiene. Figure 1 contrasts the result of the arrival of the disease in combination with invasive digital marketing, which boosted the number of sales of different products related to the subject, representing large total profits.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.57
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

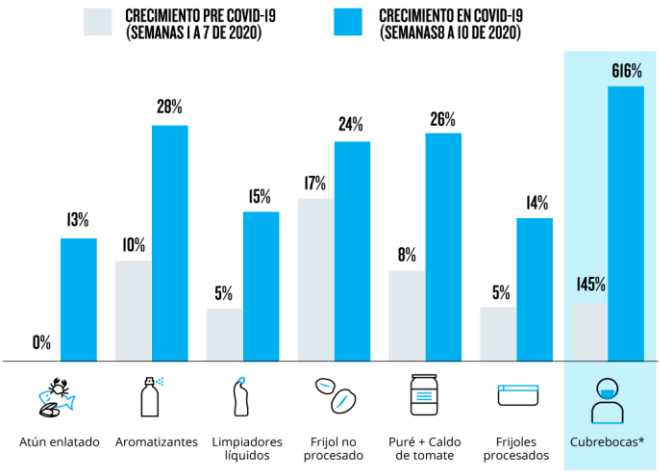


Figure 1. Sales indexes comparison before and during COVID-19. Source: **[5]**

An example of the implementation of new business models, and taking advantage of the circumstances, are the clothing companies, which almost immediately ventured into the production of face masks, which not only increased their profits but also left them well-positioned, in terms of acceptance, since they knew how to manage marketing strategically. This example, given the situation of COVID-19, is an extraordinary case, where some companies sought to take advantage of trends to be profitable and maintain acceptance profiles through potential consumers **[6]**, given that in the business world, having a good image is vital and even more important than offering quality products.

Another example of taking advantage of current trends, which is not necessarily linked to the case of the 2020 pandemic, was Pride, also known as LGBT Pride Day, which is a day of the year chosen by the LGBT movement to affirm the feeling of personal pride, which is generated by publicly displaying sexual and gender identities and orientations traditionally marginalized and repressed. The purpose is to make your claims and your presence visible in society. Given this, companies such as Puma, Levis, C&A, Adidas, Swatch, and others, took advantage of the ideals and colors of the movement related to Pride, under the slogan "Love unites us now more than ever", with a proposal of clothing pieces, colorful footwear, and accessories.

As is well known, some companies and digital marketing strategists have managed to integrate and position themselves successfully with the public. For this, these companies analyzed world or individual trends to sell more and make their image more notorious. These companies in turn, despite having suffered losses due to the confinement of the population, have in digital marketing a great support tool that has kept them afloat, showing that a good advertising strategy makes a difference, surpassing the competition.

This paper presents the identification and analysis of message patterns directed by some brands that have sought to increase their sales successfully, given current trends, with the message of raising awareness towards their consumers. For this, the document is organized as follows, Section 2 presents the background on digital marketing, the need for it in companies, and gives some scopes on trending topics. Section 3 describes the method used as a proposed solution. Section 4 presents the tests and results obtained, based on application examples, and Section 5 summarizes some conclusions and future work.

II. Background

In recent years, the technological explosion has come to change the way of doing business and humanity has learned how to adapt, coexist, and learn from them. These technological advances have shortened, more and more, the times of new developments and discoveries, based on the understanding of the present and thinking about the future. Precisely, a field of knowledge that is currently promoting new developments is Artificial Intelligence **[7]**, where many companies are increasingly aware of its importance to incorporate this technology into their business models. This would allow obtaining relevant patterns from the data generated, as well as assistance in decision-making **[8]**.

A. Digital marketing

Marketing has been one of the areas that has evolved with technology **[9]** and it has empowered companies to make sales grow, and, consequently, improve their positioning in the global market. On the other hand, neuromarketing is based on the part of consumer psychology, the colors that reflect the moods, the shapes, the way in which consumers find the presentation of the product more pleasant, the positions that men take and women when shopping alone, with friends or family, among other characteristics.

In business, digital marketing is creating new ways for companies to interact with consumers and give new forms of communication, which is driving higher revenues and better productivity. In this type of marketing, information technologies are included as part of advertising campaigns through various software applications **[10]**, with which it is possible to measure the profits and behavior of Internet users, for example, the acceptance of advertisements and bounce rate.

B. Digital marketing needs

The expansion of Artificial Intelligence in business is reflected in the growing development of applications and services, ranging from natural language processing to image recognition. With the passage of time, this type of technology has achieved a better implementation capacity in the various activities of organizations. However, it has not yet been

possible to develop autonomy to manage all the processes that business requires **[11]**.

Thus, performing the recognition and analysis of business processes is important to identify bottlenecks, deviations, and other types of problems. Therefore, new mechanisms need to be developed to obtain greater benefits. Among the actions that Artificial Intelligence could cover in business, the following stand out **[12]**: natural language processing, satisfaction analysis, optical character recognition, image and face recognition, the improvement of sales processes through machine learning, among others.

C. Trending topics

Trending topics are statistics collected by websites or social networks, mainly Twitter, Facebook, and Instagram **[13]**, where the interests of users or the publications that generate the most controversy are analyzed. Thus, it is possible to know the demands and needs of users and at what times of the day, month, or year it is more feasible to persuade them to gain their attention.

At present, companies can take advantage of trending topics to their advantage, to make the audience know or promote their products, but the most important thing is to create a good reputation to generate trust and rapprochement with consumers. Figure 2 shows a diagram of a semantic network of marketing-oriented to trending topics, which is focused on the experience of companies towards current consumers.

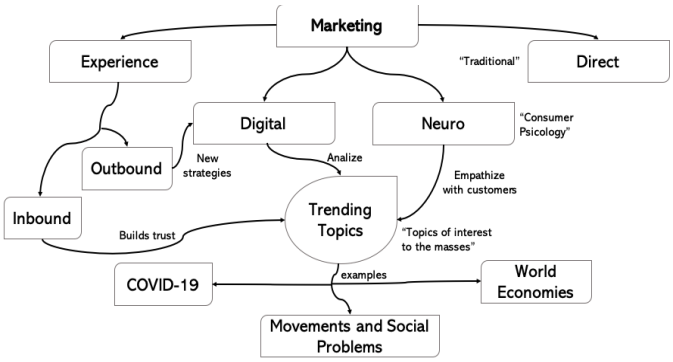


Figure 2. Semantic network of marketing-oriented to trending topics. Source: **[14]**.

In the semantic network, we have the 2020 antagonist, the coronavirus. For which, companies such as Adidas or Levis have launched face masks for sale. Other companies such as Walmart or Liverpool have strengthened, like most companies in the world, electronic commerce, but these highlight the part of "stay-at-home" or "take-care-of-our-customers", which, in turn, if people keep healthy after this event or some others, they will continue to return to their stores, being part of the consumption cycle.

Regarding the node of pride or "Pride", Adidas was once again the protagonist, who launched a collection of clothing and footwear symbolizing such a movement. In addition, this company has known how to interact with the needs of users, manufacturing products of recycled origin, thus taking care of the environment, which is also a topic of discussion in recent years. On the other hand, the fall of the world economy and of the countries themselves is another issue that has now become a trend.

The confinement by COVID-19 had as a consequence the decrease and in some countries, the slowdown of production, being a fundamental part of the Gross Domestic Product of each country **[15]**. Therefore, measures and agreements have been taken between the government, banks, and companies to generate actions that resume production or encourage investment.

D. New technologies

At present, web technology is a discipline that is constantly evolving and growing, which since its inception has undergone changes due to the revolution and creation of new services **[15]**. Today, the Web is classified into three stages **[16] [17]**: i) the first focuses on the invention of the Web, protocols, and Web browsers; ii) the second is defined by the increase in social networks, mobile applications, and cloud computing; and iii) the last one, which is the coming stage, promises a greater revolution than the previous ones, decentralizing all the content of the Web, creating unique profiles, giving a greater rise to cryptocurrencies and integrating NFTs (non-fungible tokens), and the Metaverse (concept denoting the next generation of the Internet).

Nowadays, the triad between NFTs, the Web3 ecosystem, and the Metaverse, which are technologies that share characteristics in common, have allowed an important synergy for the takeoff of Web 3.0, which have a wide potential that demands a large of technical concepts, such as distributed systems, economics, finance, organizations, entrepreneurship, marketing, cryptocurrencies, cryptography, art, intellectual property, networks, security, rendering, virtual reality, augmented reality, decentralized programming, among others. For this reason, mobile applications, cloud computing, artificial intelligence, the Internet of things, and smart devices have demonstrated essential solution capabilities in different areas, successfully adapting, due to their capabilities, to actual society.

In this sense, by combining these technologies, commercial and financial activities, Web browsing, the acquisition and demand for new services, and the creation of new business models can be improved. For example, companies like Gucci **[18]**, Zara **[19]**, Polo Ralph Lauren **[20]** have jumped on the NFT trend, creating their non-fungible

tokens to be used on avatars from different Metaverses (Figure 3). Thus, new sales paths are sought, for example, transferring an advertisement from a platform to a three-dimensional digital universe, within a video game or social network, or even when interacting with its users.

Figure 3. Integration of avatars in the Metaverse. Source: **[21].**

In order to automate transactions, NFTs promise great potential, and it is expected that memberships, promotional codes, tickets, will be encrypted digital keys, all thanks to decentralized application platforms, such as Ethereum, which is the pioneer in digital contracts **[22]**. Another important feature is the possibility of creating unique users on the Web, giving the possibility of accessing Websites through the digital signature of NFTs.

III. Method

The proposal is to analyze, through an algorithm, the trending topics based on the information content that is generated on Web pages, social networks, and online discussion forums. The purpose is to know periodically what people think and opine about social movements, as well as to know where the attention of the masses is directed. From which, in response, brands and corporations establish their advertising and image strategies. The proposed algorithm uses the URL of the Web page, study object, that is, of which we want to know its advertising content. For this, the algorithm bases its operation on five stages:

- First, the URL is entered by the user, which is processed to obtain the language of the website via the Beautiful Soup library of Python. At the same time, through Selenium **[23]**, the site is accessed through the Web browser, sliding the cursor automatically, in order to divide its content. On each swipe iteration, it gets a screenshot with a certain nomenclature.
- Closing the test web browser locates the screenshot files, previously stored in a defined folder.
- For each screenshot, the text contained in it is extracted through optical character recognition (OCR). The Python pytesseract library is used for this. Subsequently, the text is divided word by word, for further analysis.
- For the analysis of the text, stored in the form of regular expressions in plain text files, with CSV format, the algorithm is executed to identify keywords related to Trending Topics, current trends, social movements, and COVID-19.
- This word identification is based on the matches (intersection) of the text stored in the CSV files and the list of keywords. Thus, based on this intersection, the summary of the identified words is displayed.

For the storage of the information, a database was elaborated, with current trends topics, based on the vocabulary used in digital marketing, which could also be scaled to particular topics, managing them by geographic regions. In addition, the words that appear before and after the keyword found were included in the database, thus allowing us to know how companies are treating the issue and how they are taking advantage of it. Undoubtedly, this can be useful to develop our own marketing strategies, following the successes of others and avoiding campaigns of little success or failure.

IV. Tests and Results

As a first case, the treatment of some companies around the issue of the pandemic COVID-19 was analyzed. For example, among the strategies used by supermarkets was to promote the rules of social distancing and sanitation. Thus supporting the indications provided by governments in order to make people understand that they care about the situation and the health of their consumers.

As an example, the algorithm was tested on the official website of Chedraui Mexico. It was observed that the commercial store informs the community that they follow the sanitation measures and that the customer and their well-being are always the priority. A popular strategy that has proven to be very successful in terms of gaining popularity, as well as providing a feeling of confidence in consumers. Figure 4 shows part of the website of the business in question with information related to COVID-19.

Figure 4. Semantic network diagram of marketing-oriented to trending topics.

Figure 5 shows the result, after the execution of the algorithm, where the words related to the COVID-19 trending was the famous #QuédateEnCasa (#stayAtHome). It is also shown that the expressions that accompany the keyword located in the database, was that they care about the

well-being of their customers, which perfectly illustrates the intentions of the company, which is to promote trust in consumers and help with the task of social distancing.

```
Las repeticiones del algoritmo son:
1)COVID -->{'#QuédateEnCasa'}
2)SOCIAL -->set()


Expresiones del tema:


ANTES palabra clave:
Pr
raced
SOC
ntact
#QuédateEnCasa


DESPUES palabra clave:
#QuédateEnCasa
Preocupados
por
el
bienestar
```

Figure 5. Semantic network diagram of marketing-oriented to trending topics.

Another test was carried out with the website of the German brand Adidas, which, on its official Adidas Mexico page, has a specialized tab to see the services related to the COVID-19 issue. This also contrasts the use of the situation to launch personal hygiene and protection products such as face covers. Figure 6 shows a fragment of the Adidas Mexico website with information related to COVID-19.

Figure 6. Fragment of Adidas Mexico website with information related to COVID-19.

When executing the algorithm, it was observed that the words used as context for the key "COVID-19" were "See more about our times and services", which is inferred that the market strategy is aimed at economic exploitation based on the situation that we live in today by disease. Figure 7 shows the results obtained from the official Adidas Mexico website. Undoubtedly, the intention of Adidas is the sale of face covers, which confirms its market strategy in the face of the need for the product. Figure 8 shows advertising related to the sale of face covers.

```
Las repeticiones del algoritmo son:
1)COVID -->{'COVID-19'}
2)SOCIAL -->set()


Expresiones del tema:


ANTES palabra clave:
MARCAS
COLECCIONES
BUSCAR
QA
COVID-19


DESPUES palabra clave:
COVID-19
|
VER
MAS
SOBRE
NUESTROS
TIEMPOS
Y
SERVICIOS.
ANOVE
```

Figure 7. Screen with the results obtained from the Adidas Mexico website.

Figure 8. Adidas Mexico advertising about the sale of face covers.

Leaving aside the issue of the COVID-19 pandemic, another trend was Pride 2020, where, as in the previous case, different companies saw it as a field of opportunity to guide digital marketing for the sale of their products, as was the case of Calvin Klein (Figure 9). When applying the algorithm, it was observed (Figure 10) that the sale of items related to Pride includes other words before and after with the intention of capturing people's attention and thus achieving a successful sale. It was also observed that in a short time some products were sold out.

Figure 9. Calvin Klein Mexico website with information on the sale of Pride 2020 items.

Figure 10. Screen with the results obtained from the Calvin Klein Mexico website.

Based on the above, for every trending topic, there are different ways of reaching the public and each with different intentions. The important thing is to observe the reaction that is generated in the public and specifically in potential consumers, in order to design a better marketing strategy. In this sense, through this type of mechanism, such as the algorithm used, trends, and experiences of other brands with successful advertising campaigns can be identified.

III. Conclusions and Future Work

The usefulness of the algorithm was shown to deal with trending topics and analyze the behavior of the masses in the face of social trends. The purpose is to analyze the market competitors and have a support tool to learn from other advertising movements, both the successful ones and also those that did not transcend, since the idea is to seek feedback based on the mistakes of others or even of their own. Thus, so as not to fail in new marketing strategies.

This work as a tool for the identification of digital advertising can be useful for companies, as a starting point to identify strategies that follow the competition, see the acceptance rate that they manage to achieve, and, therefore, as support to model a counterattack strategy of the changing global market.

On the other side, analyzing social ideals and collective movements can be complex, as they tend to be delicate. However, these are risks that companies take in order to have a good image and reputation, which in turn becomes an increase in their sales or services since it is intended that consumers become the best promoters of the brand by feeling satisfied with the product or service in question.

In this sense, current technology has broken the communication barrier between brands and consumers. Before surveys or censuses were carried out to the general population to find out their opinion, now the interaction is closer to solve the needs of customers, listen to them at all times and improve products or services according to market demand.

Without a doubt, a good market strategy brings with it public acceptance, support, and economic benefits for companies. Therefore, it is important to constantly analyze current trends, the economy, collective ideals, and other related issues, in order to make the most of these situations no matter how complex they are, that is, where some find problems, others they look for solutions.

Part of those solutions is the adoption of new technologies such as Web3, the Metaverse, and NFTs, through which new ways to advertise, acquire and retain customers can be created. In addition, the care of the privacy of the information of the users in the network must be guaranteed.

As far as future work is concerned, it is intended to add to the algorithm a tool for monitoring and comparing advertising movements of companies, in order to identify successful campaigns based on the sales index, the presence on the Web, and their approach and empathy with users.

Acknowledgments

This work was supported by UNAM-PAPIIT IA104122.

References

[1] S. Baker, N. Bloom, S. Davis, K. Kost, M. Sammon, T. Viratyosin. "The unprecedented stock market reaction to COVID-19". Economics Working Paper 20112, 1-16 (2020).

[2] S. Baker, N. Bloom, S. Davis, S. Terry. "Covid-induced economic uncertainty". NBER Working Paper 26983, 1-10 (2020).

[3] R. Cruz, M. Patiño. "Las medidas del Gobierno Federal contra el virus SARS-Cov2 (COVID-19)". Senado de la República. Cuaderno de Investigación 6, 5-36 (2020).

[4] J. Xifra. "Comunicación corporativa, relaciones públicas y gestión del riesgo reputacional en tiempos del Covid-19". El profesional de la información (2), 1--18 (2020).

[5] "Los productos que más consumiremos los mexicanos en cuarentena". Lider empresarial. Last seen January 23, 2021. Source: www.liderempresarial.com/los-productos-que-mas-consumiremos-los-mexicanos-en-cuarentena

[6] T. Aluja. "La minería de datos, entre la estadística y la inteligencia artificial". Qüestiió 25(3), 479-498 (2001) .

[7] "¿Qué es el marketing digital?". Marketing Digital. Last seen January 23, 2021. Source: www.mdmarketingdigital.com/que-es-el-marketing-digital

[8] "Marketing Online: Potencial y Estrategias". Proyecto CECARM, Fundación Integra de Murcia, 1-70 (2014).

[9] M. Ortiz, L. Aguilar, L. Marín. "Los desafíos del marketing en la era del big data". e-Ciencias de la Información 6(1), 1-30 (2016).

[10] P. Quiroga. "¿Qué es la inteligencia artificial y cómo se aplica en los negocios?". Gestión, 1--6 (2018).

[11] L. Rouhiainen. "Inteligencia Artificial, 101 cosas que debes saber hoy sobre nuestro futuro". Alienta Editorial, 5-32, Madrid, Spain (2018).

[12] H. Becker, M. Naaman, L. Gravano. "Beyond trending topics: Real-world event identification on Twitter". Conference on Weblogs and Social Media, 438-441 (2011).
[13] T. Callen, T. "¿Qué es el producto interno bruto?". Finanzas & Desarrollo, 45(4), 48-49 (2008).
[14] D. Riaño, G. Molero-Castillo, R. Piñon, A. Vázquez Mena, E. Bárcenas. "Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm" . Programming and ComputerSoftware, Springer Nature, vol. 46, 652-660, 2020.
[15] "Servicios Web". Ecured. Last seen January 8, 2022. Source: https://www.ecured.cu/Servicios_Web
[16] Díaz H. "¿Qué es un NFT?". Kaspersky. Last seen January 8, 2022. Source: https://latam.kaspersky.com/blog/que-es-un-nft/22918
[17] González de la Cámara J. "El metaverso. La web 3.0". Linked In. Last seen January 8, 2022. Source: www.linkedin.com/pulse/el-metaverso-la-web-30-juan-gonz%C3%A1lez-de-la-c%C3%A1mara
[18] Pinedo, E. "Gucci se sube al tren de los NFT con unas zapatillas virtuales de 12 millones de dólares". Hypertextual. Last seen December 28, 2021. Source: https://hipertextual.com/2021/03/gucci-nft-zapatillas-virtuales-de-12-millones-de-dolares
[19] "Zara lanza su primera colección en formato NFT para sus clientes del metaverso". Crypto news. Last seen January 2, 2022. Source: https://cryptonews.net/es/news/nft/2847657
[20] "Ralph Lauren Announces an Exclusive Partnership with Zepeto". Ralph Lauren Corporation. Last seen January 3, 2022. Source: https://corporate.ralphlauren.com/pr_210825_ZepetoPartnership.html
[21] "Nike tendrá su propio metaverso en Roblox: 'Nikeland' ". Bajo palabra. Last seen January 16, 2022. Source: https://bajopalabra.com.mx/nike-tendra-su-propio-metaverso-en-roblox-nikeland
[22] "Smart contracts: contratos inteligentes para formalizar acuerdos en la era digital". Iberdrola. Last seen December 26, 2021. Source: www.iberdrola.com/innovacion/smart-contracts
[23] S. Gojare, R. Joshi, D. Gaigaware. "Analysis and Design of Selenium WebDriver Automation Testing Framework". Procedia Computer Science, 50, 341-346 (2015).

Bibliografía

[1] S. Baker, N. Bloom, S. Davis, K. Kost, M. Sammon, T. Viratyosin. "The unprecedented stock market reaction to COVID-19". Economics Working Paper, vol. 20112, pp. 1-16, 2020.

[2] S. Baker, N. Bloom, S. Davis, S. Terry. "Covid-induced economic uncertainty". NBER Working Paper 26983, pp. 1-10, 2020.

[3] R. Cruz, M. Patiño. "Las medidas del Gobierno Federal contra el virus SARS-Cov2 (COVID-19)". Senado de la República. Cuaderno de Investigación, vol. 6, pp. 5-36, 2020.

[4] J. Xifra. "Comunicación corporativa, relaciones públicas y gestión del riesgo reputacional en tiempos del Covid-19". El profesional de la información, pp. 1-18, 2020.

[5] "Los productos que más consumiremos los mexicanos en cuarentena". Lider empresarial. Url: `www.liderempresarial.com/los-productos-que-mas-consumiremos-los-mexicanos-en-cuarentena` (Último acceso: mayo de 2022).

[6] T. Aluja. (2001). "La minería de datos, entre la estadística y la inteligencia artificial". Qüestiió vol. 25(3), pp. 479-498, 2001.

[7] "¿Qué es el marketing digital?". Marketing Digital. Url: `www.mdmarketingdigital.com/que-es-el-marketing-digital` (Último acceso: enero de 2022).

[8] "Marketing Online: Potencial y Estrategias". Proyecto CECARM, Fundación Integra de Murcia, pp. 1-70, 2014.

[9] M. Ortiz, L. Aguilar, L. Marín. "Los desafíos del marketing en la era del big data". e-Ciencias de la Información, vol. 6(1), pp. 1-30, 2016.

[10] P. Quiroga. "¿Qué es la inteligencia artificial y cómo se aplica en los negocios?". Gestión, pp. 1-6, 2018.

[11] H Díaz. "¿Qué es un NFT?". Kaspersky. Url: `https://latam.kaspersky.com/blog/que-es-un-nft/22918` (Último acceso: enero de 2022).

[12] J. González de la Cámara. "El metaverso". La web 3.0. Url: www.linkedin.com/pulse/el-metaverso-la-web-30-juan-gonzález-de-la-cámara (Último acceso: enero de 2022).

[13] D. Riaño, G. Molero-Castillo, A. Vázquez Mena, E. Bárcenas."Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algo- rithm". Url: https://link.springer.com/article/10.1134/S0361768820080162

[14] D. Riaño, G. Molero-Castillo, E. Bárcenas, R. Aldeco. "Algorithm for Identification and Analysis of Targeted Advertising used in Trending Topics". LACCEI International Multi-Conference of Engineering, Education and Technology, Boca Ratón, Florida, EUA, 2022. Url: www.laccei.org/LACCEI2022-BocaRaton/full_papers/FP57.pdf

[15] Marketing Digital. "¿Qué es el marketing digital?". Url:http://www.mdmarketingdigital.com/que-es-el-marketingdigital (Último acceso: enero de 2022).

[16] L. Rouhiainen. "Inteligencia Artificial, 101 cosas que debes saber hoy sobre nuestro futuro". Madrid, España: Alienta Editorial, pp. 5-32, 2018.

[17] H. Becker, M. Naaman, L. Gravano. "Beyond trending topics: Real-world event identification on Twitter". Conference on Weblogs and Social Media, pp. 438-441, 2011.

[18] D. Riaño, G. Molero-Castillo, R. Piñon, A. Vázquez Mena, E. Bárcenas. "Regular Expressions for Web Advertising Detection based on an Automatic Sliding Algorithm". Programming and ComputerSoftware, Springer Nature, vol. 46, pp. 652-660, 2020.

[19] T. Callen, T. "¿Qué es el producto interno bruto?". Finanzas Desarrollo, vol. 45(4), pp. 48-49, 2008.

[20] J. Sánchez, L. López, J. Martínez. "Solución para garantizar la privacidad en el Internet de las Cosas". El profesional de la información. vol. 24, pp. 62-70, 2015.

[21] D. Riaño, G. Molero-Castillo, A. Velázquez-Mena, E. Bárcenas. "Expresiones regulares para el tratamiento de privacidad de navegadores Web". Abstraction and Application, vol. 25, pp. 121-130, 2019.

[22] P. Cerezo. "Bloqueo de anuncios: el modelo publicitario digital, a revisión". Cuadernos de periodistas: revista de la Asociación de la Prensa de Madrid, pp. 81-89, 2016.

[23] A. Londaitz. "Publicidad en los celulares: Publicidad invasiva vs. derecho a la privacidad". Tesis. Universidad del Salvador. Url: https://racimo.usal.edu.ar/4312 (Último acceso: Enero, 2022).

[24] "Google, Bienvenido a Google, la mejor empresa para trabajar". Url: www.expansion.com/2013/08/23/directivos/1377273795.html (Último acceso: Enero, 2022).

[25] Jarvis, J., "Y Google, ¿cómo lo haría?". Url: https://narrativabreve.com/2013/10/libro-google-jeff-harvis.html (Último acceso: Enero, 2022).

[26] M. Leotta, D. Clerissi, F. Ricca, C. Spadaro. "Comparing the maintainability of selenium webdriver test suites employing different locators: a case study". Proc. 1st Int. Workshop on Joining AcadeMiA and Industry Contributions to Testing Automation, Lugano, Url: https://dl.acm.org/doi/10.1145/2489280.2489284 (Último acceso: Enero, 2022).

[27] S. Gojare, R. Joshi, D. Gaigaware. "Analysis and design of selenium WebDriver automation testing framework". Procedia Comput. Sci., vol. 50, pp. 341–346, 2015.

[28] "Selenium Webdriver". Url: http://www.tutorialspoint.com/selenium/pdf/selenium_webdriver.pdf (Último acceso: Diciembre, 2021).

[29] W. Yih, J. Goodman, V. Carvalho. "Finding advertising keywords on web pages". Proc. 15th Int. Conf. on World Wide Web, Edinburgh. Url: https://dl.acm.org/doi/pdf/10.1145/1135777.1135813 (Último acceso: Diciembre, 2021).

[30] T. Mei, L. Li, X. Tian, D. Tao, C. Ngo. "PageSense: toward stylewise contextual advertising via visual analysis of web pages". IEEE Trans. Circuits Syst. Video Technol. Url: http://dl.acm.org/doi/abs/10.1109/TCSVT.2016.2598702 (Último acceso: Diciembre, 2021).

[31] D. Sánchez y A. Viejo. "Privacy-preserving and advertising-friendly web surfing". Comput. Commun., vol. 130, pp. 113–123, 2018.

[32] V. Krammer. "An effective defense against intrusive web advertising". Proc. 6th Annu. Conf. on Privacy, Security and Trust, Fredericton. Url: https://ieeexplore.ieee.org/document/4641268 (Último acceso: Diciembre, 2021).

[33] Sajjad, K. "Reconocimiento automático de matrículas usando Python y Opencv". Facultad de Ingeniería. Url: https://pdfs.semanticscholar.org/bddf/1200eb17f239e4dce2a9cec938eb8cf305f5.pdf (Último acceso: Diciembre, 2021).

[34] Patel, C., Patel, A., Patel D. "Reconocimiento óptico de caracteres mediante la herramienta OCR de código abierto Tesseract: un estudio de caso". en International Journal of Computer Applications. Url: https://research.ijcaonline.org/volume55/number10/pxc3882784.pdf (Último acceso: Noviembre, 2021).

[35] Vallez, M. "Keyword Research: métodos y herramientas para identificar palabras clave". BiD: textos universitaris de biblioteconomia i documentació, vol. 27, pp. 1-14, 2011.

[36] Slamet , C., Andrian , R., Maylawati , D., Darmalaksana , W., Ramdhani , M. "Web Scraping and Naïve Bayes Classification for Job Search Engine". 2.ª Conferencia Anual de Ingeniería y Ciencias Aplicadas. Url: `https://iopscience.iop.org/article/10.1088/1757-899X/288/1/012038/pdf` (Último acceso: Noviembre, 2021).