UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Propuesta de un clasificador para detectar innovaciones de producto del sector del calzado en Twitter.

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: García Mansilla, Marta

Tutor/a: Doménech i de Soria, Josep

CURSO ACADÉMICO: 2021/2022

Resumen

Las redes sociales como Twitter son una gran fuente de información, ya que la gente comparte opiniones y las empresas se publicitan y comparten su contenido. Dentro de este contenido, se pueden encontrar las innovaciones de producto, que ocupan un importante valor en la sociedad, ya que pueden suponer mejoras tanto sociales como económicas. Un sector en el que se insertan constantes innovaciones de producto es el sector del calzado, debido a que se ve afectado por los constantes cambios de tendencia en busca de seguir la moda. Por tanto, el fin de este trabajo es crear un clasificador capaz de detectar las innovaciones de producto que las empresas del calzado publican en Twitter. Para ello, en primer lugar, mediante minería de datos se extrae información sobre el sector del calzado en Twitter. Puesto que esta información consiste en textos, el siguiente paso es el procesado del lenguaje natural, con el fin de que esta información sea una entrada apta para los algoritmos de aprendizaje automático que se aplican a continuación. Por último, se realiza una comparativa de los resultados que ofrecen los algoritmos, con la finalidad de decidir cuál es el que mejores predicciones ofrece, para posteriormente, desplegar este modelo con el objetivo de que pueda detectar innovaciones de producto.

Palabras clave: innovación, aprendizaje automático, clasificadores de textos, Twitter, Python, sector del calzado.

Resum

Les xarxes socials com Twitter són una gran font d'informació, ja que la gent comparteix opinions i les empreses es publiciten i comparteixen el seu contingut. Dins d'aquest contingut, es poden trobar les innovacions de producte, que ocupen un important valor en la societat, ja que poden suposar millores tant socials com econòmiques. Un sector en el qual s'insereixen constants innovacions de producte és el sector del calçat, pel fet que es veu afectat pels constants canvis de tendència a la recerca de seguir la moda. Per tant, la fi d'aquest treball és crear un classificador capaç de detectar les innovacions de producte que les empreses del calçat publiquen en Twitter. Per a això, en primer lloc, mitjançant mineria de dades s'extrau informació sobre el sector del calçat en Twitter. Com que aquesta informació consisteix en textos, el següent pas és el processament del llenguatge natural, amb la finalitat que aquesta informació siga una entrada apta per als algorismes d'aprenentatge automàtic que s'apliquen a continuació. Finalment, es realitza una comparativa dels resultats que ofereixen els algorismes, amb la finalitat de decidir quin és el que millors prediccions ofereix, per a posteriorment, desplegar aquest model amb l'objectiu que puga detectar innovacions de producte.

Paraules clau: innovació, aprenentatge automàtic, classificadors de textos, Twitter, Python, sector del calçat.

Abstract

Social networks such as Twitter are a great source of information, as people share opinions and companies advertise and share their content. In this content, we can find product innovations, which have an important value in society, as they can lead to both social and economic improvements. A sector in which constant product innovations are inserted is the footwear sector, due to the fact that it is affected by the constant changes in trends in search of follow the fashion. Therefore, the purpose of this work is to create a classifier capable of detecting the product innovations that footwear companies publish on Twitter. To do so, firstly, by means of data mining, information about the footwear sector on Twitter is extracted. Since this information consists of texts, the next step is natural language processing, in order to make this information a suitable input for the machine learning algorithms that are then applied. Finally, a comparison of the results offered by the algorithms is carried out in order to decide which one offers the best predictions, and then this model is deployed to detect product innovations.

Keywords: innovation, machine learning, text classifiers, Twitter, Python, footwear industry.

Tabla de contenidos

Índice de figuras, tablas y ecuaciones

Índice de Figuras

Índice de Tablas

Índice de Ecuaciones

1. Introducción

La innovación es el proceso por el que se pretende implementar elementos nuevos o modificados con la finalidad de mejorarlos. En la actualidad, los nuevos avances tecnológicos provocan el auge de la innovación, lo que permite el desarrollo de los países, ya que mejora el bienestar, la salud de la población y la productividad. Es de gran importancia, por tanto, detectar esta innovación para que se puedan tomar decisiones adecuadas que den paso a la implementación de otras innovaciones, y puedan afrontarse así, retos sociales o económicos.

1.1 Motivación

En estos momentos, las redes sociales son una importante fuente de información a partir de la cual se pueden saber desde los gustos de cada persona hasta sus orientaciones políticas. El interés por obtener información de cualquier dato del entorno provoca la necesidad de analizar estas redes sociales, lo que plantea un reto, ya que se trata de datos no estructurados y desordenados, basados en el lenguaje natural. Puesto que muchas empresas también son usuarios de estas redes sociales, se puede obtener información valiosa sobre ellas, como puede ser la innovación.

El interés por detectar y analizar la innovación viene dado por la necesidad de estar actualizado, para poder utilizar adecuadamente los recursos disponibles y maximizar así los resultados, ya sea con fines sociales como el bienestar, o con fines económicos como la mayor productividad.

Uno de los sectores interesantes para el análisis de innovaciones de producto puede ser el sector del calzado, debido a que se ve afectado por los constantes cambios de tendencias y pretende seguir la moda del momento.

Para la extracción y análisis de estos datos de innovación, existe la opción de varias redes sociales, pero Twitter ofrece un fácil acceso a su interfaz para la obtención automática de datos, y facilita el análisis de información, debido a la forma en la que los usuarios se expresan en esta red social, ya que su contenido se basa en textos breves.

1.2 Objetivos y metodología

El objetivo final que pretende alcanzar este trabajo es:

- Crear y evaluar un clasificador para detectar innovación de producto, específico para el sector del calzado y la red social Twitter.

Para alcanzar este objetivo principal, surgen objetivos específicos, como son:

- Comprender el concepto de innovación, los tipos de innovación en general y la innovación de producto en particular.
- Conocer el contexto del sector del calzado en España.
- Entender el funcionamiento de la red social Twitter y los mecanismos para acceder a los datos.
- Aplicar técnicas de procesado de lenguaje natural para transformar las variables y obtener las características de los tweets de las empresas.
- Implementar y evaluar distintos algoritmos de clasificación.

La metodología que se lleva a cabo para poder alcanzar estos objetivos con éxito es CRISP-DM, del inglés Cross-Industry Standard Process for Data Mining, en castellano Proceso Estándar para la Minería de Datos en todos los sectores. Esta metodología describe 6 fases diferentes, con las que llevar a cabo un proyecto, y las tareas necesarias para realizar cada fase. (Shearer, 2000)

- **Fase I. Comprensión del negocio**
  En esta primera fase se realiza una comprensión de las necesidades del cliente, por lo que se lleva a cabo un estudio del contexto.

- **Fase II. Comprensión de los datos**
  La segunda fase consiste en la obtención del conjunto de datos de Twitter, en etiquetar los tweets con presencia de innovación y en otras tareas que permitan su conocimiento, como un análisis preliminar.

- **Fase III. Preparación de datos**
  En la tercera fase se realizan transformaciones y limpiezas de datos para obtener el conjunto de datos final, como el procesado del lenguaje natural, necesario para que los datos sean aptos para el modelado.

- **Fase IV. Modelado**

  La cuarta fase consiste en la aplicación de técnicas para modelar los datos y poder obtener resultados, para ello, se aplican técnicas de aprendizaje automático supervisado.

- **Fase V. Evaluación**

  En la quinta fase se analizan los resultados del modelado y se estudia en detalle la predicción correcta de cada algoritmo aplicado.

- **Fase VI. Despliegue**

  La sexta y última fase consiste en el despliegue y monitorización del modelo, es decir, su puesta en producción y continuo mantenimiento para que funcione correctamente.

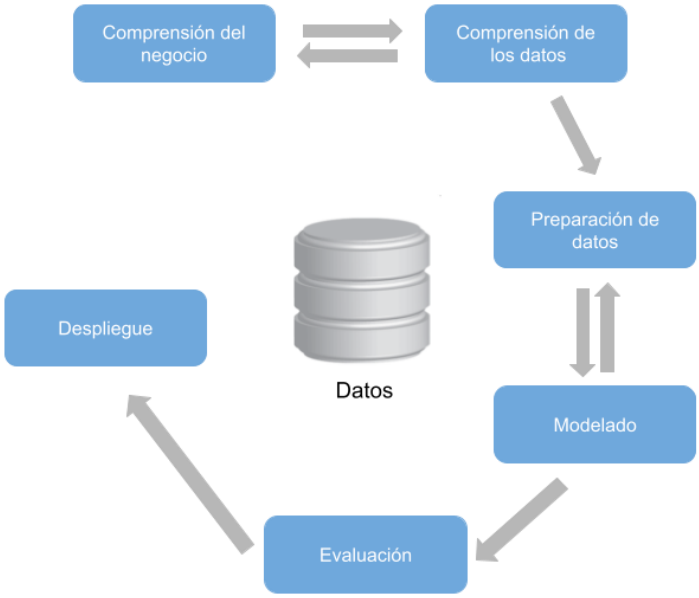


Figura 1. Funcionamiento y fases de la metodología CRISP-DM

Fuente: Elaboración propia

1.3 Estructura

La estructura que sigue el proyecto está directamente relacionada con la metodología que se sigue al realizarlo. En primer lugar, se realiza una introducción en la que se expone la motivación que provoca la realización de este trabajo y los objetivos que pretende alcanzar.

A continuación, en el estado del arte y el análisis del problema, se describe la situación del contexto actual, respecto a los temas que se tratan en el trabajo, y se realiza un análisis de las oportunidades de negocio y otros factores que influyen en el proyecto.

En las tres siguientes secciones se explica el proceso realizado a los datos, comenzando por su obtención y procesado, seguido por la aplicación de algoritmos y los resultados obtenidos tras evaluarlos, y por último el despliegue del modelo entrenado.

Finalmente, en las dos últimas secciones, se exponen las conclusiones obtenidas y la solución propuesta tras la realización del proyecto, junto a los trabajos futuros que incluyen posibles mejoras o la ampliación de objetivos.

1.4 Impacto Esperado

La detección de innovación en el sector del calzado puede suponer mejoras considerables en dicho sector, ya que permite que otras empresas introduzcan innovaciones a partir del conocimiento de las ya existentes. Estas innovaciones pueden suponer desde la mayor rentabilidad de las empresas, hasta la implementación de productos que mejoren la sociedad, por ejemplo, con productos que respeten el medioambiente.

Por tanto, este proyecto se plantea responder la siguiente pregunta: ¿Es posible detectar la innovación de producto de las empresas del sector del calzado a partir del análisis automático de sus tweets?

Por ende, se pretende que en caso de que la respuesta sea afirmativa se obtenga como producto resultante un clasificador capaz de detectar esta innovación en las empresas del sector del calzado, mediante el análisis de sus tweets.

El producto resultante tiene dos usuarios principales, las empresas de este sector que emplearían el clasificador para monitorizar la competencia, con la finalidad de maximizar sus beneficios. Y las instituciones públicas encargadas de regular las políticas del sector del calzado, con la finalidad de analizar los resultados de las políticas existentes y mejorarlas en caso de que fuera necesario.

2. Estado del arte

En este apartado se exponen varios conceptos fundamentales para entender la importancia y necesidad del trabajo realizado. También se analiza la forma actual de recolección de la innovación, se contempla el uso de Twitter para el análisis de información y se da una perspectiva de la situación actual de la economía del sector del calzado.

2.1 Innovación

Uno de los principales referentes y fuente de información sobre la innovación es el Manual Oslo (OECD/Eurostat, 2018), al que se refiere en repetidas ocasiones este trabajo.

Según el Manual Oslo, una innovación es un producto o proceso nuevo o mejorado, o la combinación de ambos, que difiere significativamente de los productos o procesos anteriores, habiendo sido introducido en el mercado o puesto en marcha por la empresa.

La innovación permite mejorar desde el nivel de vida de los individuos hasta el de sectores económicos enteros. Medir adecuadamente la innovación y sus consecuencias, ya sean positivas o negativas, permite a los agentes encargados de tomar decisiones, seguir con su estrategia o cambiarla, en función de qué maximiza más sus beneficios o el bienestar de los individuos. Los efectos más amplios de la innovación contribuyen o dificultan los objetivos de la sociedad, como la mejora del empleo, la salud y las condiciones medioambientales, entre otros retos sociales.

Puesto que la innovación es una actividad dinámica y omnipresente que se da en todos los sectores de una economía, el interés generalizado de los usuarios por comprender lo que impulsa a las empresas, las comunidades y los individuos a innovar y los factores que influyen, impulsan la construcción de un sistema para medir e informar sobre la innovación y la posterior producción de datos, estadísticas, indicadores y análisis en profundidad.

En el Manual Oslo se indica que la innovación puede y debe medirse. Además, se indica que la mayoría de los datos recogidos sobre innovación provienen de encuestas realizadas a todo tipo de empresas y clientes. Esta forma de recolección de datos puede llegar a ser un tanto subjetiva, ya que pueden existir diferencias entre lo que comprenden los posibles encuestados, y lo que realmente se quiere preguntar, ocasionado por las diferencias en el lenguaje o el vocabulario.

Por tanto, la innovación es un constructo subjetivo con el potencial de que la medición dé resultados divergentes, dependiendo de la perspectiva, las creencias y el contexto del encuestado. (Galindo-Rueda & Cruysen, 2016)

Con el fin de evitar conflictos entre la definición formal de una innovación y la de cada encuestado, y poder obtener así resultados más objetivos, el Manual Oslo registra las pautas para la recolección e interpretación de datos relacionados con la innovación. Este manual facilita la comparación entre innovación en el sector empresarial, ofreciendo un estándar a nivel mundial y además permite medirla de forma estadística.

Según las teorías de la innovación, la medición de esta puede orientarse por cuatro dimensiones: conocimiento, novedad, aplicación y creación del valor.

Las innovaciones ocurren a partir de la aplicación práctica de la información y de su comprensión, que son los conocimientos. La principal actividad que permite que se adquieran nuevos conocimientos es la investigación y el desarrollo experimental (I+D). Una vez obtenidos los conocimientos, pueden utilizarse para desarrollar nuevas ideas, modelos, métodos o prototipos que pueden constituir la base de las innovaciones.

La novedad de una innovación está relacionada con sus usos potenciales, determinados por las características de un producto o proceso en comparación con las alternativas. Algunas características pueden medirse objetivamente, como la eficiencia energética o la resistencia de los materiales, pero muchas otras son subjetivas como la satisfacción del usuario o la afinidad emocional, lo cual dificulta considerablemente su medición.

En cuanto a la aplicación, para que una nueva idea se considere innovación es necesario que se implemente y además que sea accesible para los usuarios potenciales. Esta nueva idea debe poseer al menos alguna característica que la organización no haya puesto previamente a disposición de sus usuarios.

La innovación es considerada una actividad económica que requiere recursos, pero con esta inversión se pretende crear un valor, aunque no puede garantizarse, ya que los resultados de la innovación son inciertos.

Por tanto, la innovación puede tener importantes repercusiones en la estructura y la dinámica de los mercados, ya que puede expulsar a los competidores de un mercado o bloquear la entrada de otros nuevos, entre otras causas, como resultado de importantes ventajas de costes, características novedosas del producto o efectos de red.

Concretamente este trabajo se centra en la innovación de **producto**, que consiste en un bien o servicio nuevo o mejorado que posee características significativamente diferentes a las anteriores

de la empresa. Además, las innovaciones de productos pueden utilizar nuevos conocimientos, o basarse en nuevos usos de tecnologías ya existentes. Por tanto, este tipo de innovación aporta un gran valor a la empresa, ya que los objetivos económicos de las innovaciones de una empresa pueden incluir la generación de beneficios, el aumento de las ventas o el conocimiento de la marca, gracias a la innovación de los productos. (Crépon, 1998)

2.2 Twitter

Twitter es una red social de microblogueo, es decir, permite a sus usuarios enviar y publicar mensajes breves llamados tweets, con un máximo de 280 caracteres.

Puesto que se trata de una red social ofrece interactividad entre usuarios. A partir de esta interacción se puede obtener información como quién escribe cada mensaje, cuándo y dónde lo hace, o cuál es su contenido. Esta información se puede observar directamente desde la red social, pero Twitter también ofrece el acceso y recolección de estos datos de forma estructurada.

Además, en esta plataforma los usuarios expresan sus sentimientos y opiniones sobre cualquier tipo de tema, lo que ofrece fuentes de datos potencialmente útiles para la previsión de variables sociales y económicas. Se han realizado estudios sobre el análisis de sentimientos, donde se intenta clasificar la emoción y la opinión humana, y se lleva a cabo de manera satisfactoria. (Bravo-Marquez, Mendoza, & Poblete, 2014)

Por ejemplo, los contenidos de los tweets han ayudado a monitorizar la opinión pública respecto al COVID-19, la vacunación y las políticas relacionadas. (Ntompras, Drosatos, & Kaldoudi, 2022).

También se han utilizado datos de esta red social para describir las preferencias políticas y pronosticar los resultados de las elecciones (Blazquez & Domenech, 2018). Además, el análisis de los tweets ha ayudado a predecir los movimientos del mercado de valores (Fuentes Dávila Otani, 2021)

Otra de las aplicaciones de esta red social ha sido la monitorización de la opinión pública sobre nuevas políticas, que da paso incluso, a poder predecir temas de importancia mundial. (Cody, Reagan, Dodds, & Danforth, 2016)

2.3 Sector Calzado

La industria del calzado consiste en las actividades que hacen posible el diseño, fabricación, distribución, comercialización y venta del calzado. En este sector se encuentran diversos segmentos de producto en los que pueden especializarse las empresas, como son los zapatos de vestir, zapatillas, zapatos para niño, sandalias o zapatos de protección, entre otros. Además, también se emplean gran diversidad de materiales en la fabricación del calzado, como pueden ser la tela, el plástico, el caucho o el cuero.

Concretamente en España, el primer foco se origina en la Comunidad Valenciana en el siglo XIX. Posteriormente, también comienza a tomar gran importancia este sector en las zonas de Almansa en Albacete y Arnedo en La Rioja, entre otras. Todas estas empresas del sector del calzado se representan mediante el código NACE 1520. Los códigos NACE (Nomenclatura de Actividades Económicas) constituyen el estándar de clasificación a nivel europeo para las actividades económico-productivas.

En la actualidad, el sector del calzado en España se ve afectado por la globalización de la economía, ya que provoca un incremento de la competencia, debido a la entrada masiva de productos de otros países, sobre todo de los países asiáticos. Por tanto, para mantener su producción, las empresas españolas intentan tomar algunas medidas, como trasladar parte o totalidad de su producción a otros países con menor coste salarial. O implementar estrategias competitivas relacionadas con el precio, como incidir en la tecnología, diseño o distribución, invirtiendo en la innovación, para obtener productos con mayor valor añadido, que permitan cubrir costes salariales más elevados. (Ybarra Pérez & Santa María Beneyto, 2005)

La sociedad se encuentra en un alto nivel de digitalización, por tanto, la presencia de las empresas en redes sociales puede ayudarles a alcanzar nuevas audiencias o incluso a aumentar el número de clientes potenciales. Es por esto, por lo que las empresas ven redes sociales como Twitter, una plataforma perfecta para la transmisión de su publicidad y, por tanto, también de sus innovaciones.

2.4 Propuesta

Como ya se ha comentado anteriormente, lo habitual es que la información sobre innovación se extraiga a partir de encuestas, pero requiere muchos recursos humanos y temporales, puede provocar resultados subjetivos debido a la diferente comprensión de cada encuestado. Y, además, existe cierta reticencia por parte de las empresas a contestar las encuestas ante la posibilidad de proporcionarle información valiosa a la competencia.

También se ha visto que las redes sociales se han usado en otros trabajos para un análisis de sentimientos e incluso para monitorizar la opinión pública. Por tanto, este trabajo considera las redes sociales como una buena herramienta, a la par que eficiente, para obtener información sobre la innovación. Concretamente, se centra en la red social Twitter, ya que, en esta plataforma las empresas tienen gran presencia y los usuarios expresan sus gustos y opiniones.

Además, el trabajo centra su análisis en el sector del calzado, ya que es un sector que, ante su situación económica, se ve prácticamente forzado a innovar y digitalizarse para poder afrontar las consecuencias de la globalización y las continuas importaciones.

3. Análisis del problema

En este apartado se explica la solución elegida dentro de las posibles, se muestra el plan de trabajo, y el presupuesto necesario para llevarlo a cabo, y finalmente se analizan el marco legal y ético.

3.1 Análisis de posibles soluciones y solución propuesta

El problema que plantea resolver este proyecto es la obtención de un clasificador de innovación en el sector del calzado. Este clasificador supone una oportunidad de negocio tanto en el ámbito privado con las empresas, que quieren obtener información de la competencia y monitorizar sus innovaciones, como en el ámbito público, con las instituciones públicas, cuyo fin es analizar el resultado de las políticas implantadas en el sector.

Como ya se ha comentado anteriormente, la obtención de innovación mediante encuestas puede llegar a resultar un proceso lento y tedioso, ya que pueden pasar meses desde que se realiza la encuesta hasta que se consiguen los datos a analizar, como se puede observar en la última encuesta sobre innovación de las empresas del Instituto Nacional de Estadística, donde los últimos datos son de 2020, pero su publicación no se ha podido realizar hasta diciembre de 2021. Además, muchas empresas prefieren no participar en las encuestas para evitar poner a disposición de la competencia información valiosa. Por tanto, surgen otras opciones a través de las cuales se puede obtener información de sus innovaciones. Puesto que las empresas se quieren dar a conocer para ampliar sus clientes y aumentar sus ventas, utilizan redes sociales y sitios web para publicitarse. Esto supone que redes sociales como Instagram, Facebook, LinkedIn, Twitter o incluso la propia página web de cada empresa sean posibles soluciones para la obtención de innovación en el sector del calzado. Concretamente en este trabajo, se propone como solución extraer la información de Twitter, ya que es una red social donde las empresas tienen gran presencia y los usuarios expresan sus gustos y opiniones mediante breves mensajes.

3.2 Plan de trabajo y presupuesto

Tras clarificar la solución propuesta al problema que se plantea, es necesario realizar un plan de trabajo para que sea posible llevarla a cabo.

Inicialmente es necesario realizar un estudio del contexto actual e informarse sobre varios conceptos para que sea posible la realización del proyecto. En esta primera fase se realiza un estudio de la documentación de la API de Twitter y de Tweepy, una librería que permite el acceso

a Twitter desde Python. También se obtiene información sobre el procesado de textos, y, por último, se realiza una documentación sobre Aprendizaje Automático y sobre la librería de Python Scikit-learn, que permite la aplicación de estos algoritmos.

En la siguiente fase se comienza con la extracción de datos, para ello, se accede a la API de Twitter desde Tweepy y posteriormente se etiqueta el conjunto de datos en función de si hay presencia de innovación o no. Una vez obtenido el conjunto de datos se procede a su limpieza, se obtienen las características a partir de los textos y se balancea el conjunto de datos.

Tras procesar el conjunto de datos, en la última fase, se realizan análisis tanto descriptivos como predictivos, posteriormente se evalúan los modelos y se interpretan los resultados. A continuación, se despliega el modelo y se propone un mantenimiento para garantizar su correcto funcionamiento.

Las diferentes fases junto a la redacción de la memoria se llevan a cabo siguiendo los tiempos que se reflejan en el diagrama de Gannt.

Tareas	Inicio	Final	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17
Estudio contexto actual	28.02.2022	03.04.2022																	
Documentación sobre la API Twitter	28.02.2022	13.03.2022																	
Documentación sobre Tweepy	07.03.2022	13.03.2022																	
Documentación sobre procesado de textos	14.03.2022	20.03.2022																	
Documentación sobre AA y clasificadores	21.03.2022	03.04.2022																	
Documentación sobre Scikit-learn	28.03.2022	03.04.2022																	
Obtención del conjunto de datos	04.04.2022	08.05.2022																	
Extracción de datos	04.04.2022	16.04.2022																	
Etiquetado	11.04.2022	24.04.2022																	
Limpieza de datos y obtención de características	25.04.2022	01.05.2022																	
Balanceado del conjunto de datos	02.05.2022	08.05.2022																	
Análisis predictivo	09.05.2022	22.05.2022																	
Evaluación de modelos	23.05.2022	29.05.2022																	
Interpretación de resultados	30.05.2022	05.06.2022																	
Despliegue del modelo	06.06.2022	12.06.2022																	
Propuesta de mantenimiento	13.06.2022	19.06.2022																	
Redacción de la memoria	04.04.2022	27.06.2022																	

Figura 2. Diagrama de Gannt

Fuente: Elaboración propia

La realización del proyecto tiene una duración aproximada de 17 semanas, realizando una media de 20 horas semanales, lo cual equivale a 350 horas. Puesto que este trabajo corresponde a 12 ECTS, y cada ECTS equivale a 25 horas, el cómputo total de horas se ajusta a los créditos asignados a este proyecto.

Una vez se han planificado las tareas y se sabe el tiempo que requiere cada una de ellas, se puede ajustar un presupuesto en base al precio medio de un científico de datos. Puesto que el salario medio de un científico de datos es de 35.000€ anuales, cada hora valdría alrededor de 19€.

El principal coste del trabajo son los recursos humanos, ya que el software necesario para realizar el trabajo es Python, un lenguaje de programación que posee una licencia de código abierto, por tanto, su coste es gratuito. Y en cuanto al hardware, para realizar el trabajo se emplea un ordenador portátil, por lo que tomando 1000€ como precio promedio de un ordenador y vida útil de 5 años, se podría presupuestar el coste de hardware durante 17 semanas en 65€. Finalmente, el presupuesto total de este trabajo es de 6715€, como se puede observar en la Tabla 1.

Tarea	Tiempo	Coste
Estudio contexto actual	**100h**	**1700€**
Documentación API Twitter	30h	510€
Documentación Tweepy	10h	170€
Documentación procesado textos	20h	340€
Documentación AA y clasificadores	35h	595€
Documentación Scikit-learn	5h	85€
Obtención del conjunto de datos	**100h**	**1700€**
Extracción de datos	25h	425€
Etiquetado	35h	595€
Limpieza de datos	20h	340€
Obtención de características	10h	170€
Balanceado de datos	10h	170€
Análisis de datos	**120h**	**2040€**
Análisis predictivo	40h	680€
Evaluación de modelos	20h	340€
Interpretación de resultados	20h	340€
Despliegue del modelo	20h	340€
Propuesta de mantenimiento	20h	340€
Redacción de la memoria	30h	510€
Hardware	**350h**	**65€**
TOTAL:	**350h**	**6715€**

Tabla 1. Tareas por realizar, tiempo y coste necesarios para llevarlas a cabo

Fuente: Elaboración propia

3.3 Análisis del marco legal y ético

El marco legal es la normativa que debe ser cumplida para garantizar la legalidad de un trabajo, junto al marco ético que indica las pautas a seguir ante posibles dilemas morales. En el marco legal y ético se deben tener en cuenta la protección de datos, la propiedad intelectual, y la ética, entre otros aspectos legales.

Puesto que los usuarios antes de empezar a utilizar Twitter aceptan su política de privacidad, están aceptando que Twitter tenga acceso a sus tweets y demás información publicada. Además, Twitter permite el acceso de estos datos a terceros, como son los desarrolladores que acceden mediante la API. Por tanto, la protección de datos se lleva a cabo según la normativa.
(Política de Privacidad de Twitter, 2022)

La propiedad intelectual permite obtener reconocimiento y explotación económica de una creación del intelecto humano. Para ello, se emplean las licencias y las patentes, este proyecto se basa en una licencia de software comercial, ya que el objetivo es su comercialización en empresas o en el sector público.

En cuanto a la ética, puesto que los datos que se obtienen en el proyecto son datos de empresas, y no personales, no tendría por qué afectar al comportamiento humano con acciones como discriminación entre sexos o sesgo entre razas. Aunque sí puede surgir un dilema moral, ya que, una vez creado el clasificador, este podría ser empleado para provocar actos de competencia desleal, como engaño, imitación o denigración.

4. Obtención y procesado de datos

En este apartado se explica cómo se han obtenido los datos, que características poseen y el procesado necesario, para que posteriormente puedan ser la entrada de un modelo, con el fin de extraer información sobre ellos.

4.1 Obtención de datos

La base de datos del Sistema de Análisis de Balances Ibéricos (SABI) es una herramienta Web que ofrece el acceso a la información general y las cuentas anuales de más de 2,7 millones de empresas españolas.

Se pueden realizar búsquedas en el repositorio filtrando por diversos criterios. En este caso, se filtra por actividad y se marca la fabricación de calzado, representada por el código NACE 1520, de esta forma se obtienen las empresas españolas que facturan en dicho sector. A partir de la información general de las empresas**,** se puede obtener su nombre de usuario en Twitter. Como representación de este sector se escoge una muestra de las 20 empresas con mayor facturación y con presencia en Twitter, para posteriormente obtener 50 tweets de cada una de ellas.

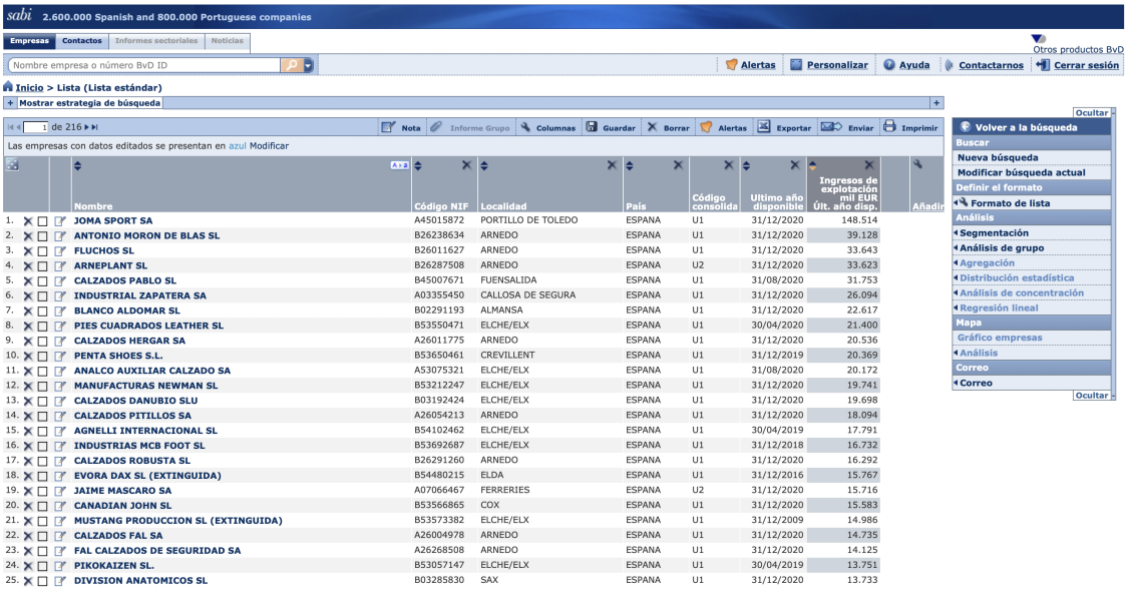


Figura 3. Captura de las empresas del sector del Calzado en SABI

Fuente: Sistema de Análisis de Balances Ibéricos

Las 20 empresas con mayor facturación en este sector y con presencia en Twitter, junto a su CIF (Código de Identificación Fiscal), localidad y nombres de usuario son las siguientes:

Nombre de la empresa	CIF	Localidad	Usuario en Twitter
Joma Sport SA	A45015872	Portillo de Toledo	JomaSport
Industrial Zapatera SA	A03355450	Callosa del Segura, Alicante	Panter_Calzado,
Calzados Hergar SA	A26011775	Arnedo, La Rioja	Callaghan_Shoes,
Pikokaizen SL	B53057147	Elche, Alicante	pikolinos
Calzados Pitillos SA	A26054213	Arnedo, La Rioja	CalzadoPitillos
Disgramarc SL	B03469327	Elda, Alicante	MAGRITshoes
Calzados Robusta SL	B26291260	Arnedo, La Rioja	CalzadosRobusta
División Anatómicos SL	B03285830	Sax, Alicante	DianCalzado
Fal Calzados de Seguridad SA	A26268508	Arnedo, La Rioja	FalSeguridad
Pe Eme & Company Online 1959 SL	B02907186	Elche, Alicante	pmcalzado
Creaciones Alpe SL	B45057874	Santa Cruz del Retamar, Toledo	AlpeWomanShoes
Cutillas Confort SL	B03830825	Elche, Alicante	DrCutillas
Pikostore SL	B53906590	Elche,Alicante	MartinelliShoes
Artesania de Calzado SA	A12093886	La Vall d'Uixo, Castellón	snipe_shoes
Garvalin Calzados SL	B03406931	Elche, Alicante	GARVALIN
Horcajo Investment SL	B26537530	Arnedo, La Rioja	NaturalWorldEco
Calzados Futurmoda SL	B53970398	Elche, Alicante	DAngelaShoes
Pala Vega Internacional SL	B02486405	Almansa, Albacete	LuisGonzaloShoe
Modexpress SL	B53846986	Murcia	calzadomiralles
Calzados Segarra Canos SL	B12230959	La Vall d'Uixo, Castellón	CalzadosSegarra

Tabla 2. Nombre, CIF, localidad y usuario en Twitter de las empresas de las que se extraen tweets

Fuente: Elaboración propia

Una vez obtenidas las cuentas, se procede a la extracción de los tweets, para ello se utiliza la API (Application Programming Interface) oficial de Twitter. Una API es una interfaz que hace posible la comunicación entre componentes de software y permite recuperar así información con código de forma automática. Para la obtención de la información se realizan consultas, conocidas como *endpoints*, que permiten la comunicación con la interfaz. En estas consultas, se indica mediante parámetros la información que se desea extraer, como el usuario del que se quiere obtener información, el período del que se quieren obtener los tweets o en qué campos del tweet se está interesado.

Puesto que esta interfaz maneja grandes cantidades de datos, es necesario una autenticación para garantizar la seguridad de estos. La API de Twitter tiene diferentes métodos de autenticación, pero concretamente en este trabajo se usa el método *OAuth 2.0 Bearer Token*, ya que permite el acceso a la información pública disponible en Twitter, solamente con el Bearer Token, es decir, con la clave asociada a la aplicación, sin necesidad de especificar los tokens de usuario, al contrario que en los otros métodos.

Para obtener las credenciales que permiten el acceso a la API, hay que crear una aplicación, para ello se inicia sesión en Twitter y se accede a la sección de desarrolladores. Tras crear la aplicación, en el apartado de claves y tokens se genera el Bearer Token, necesario para poder realizar las consultas mediante el método de autentificación OAuth 2.0. (Twitter API., 2022)

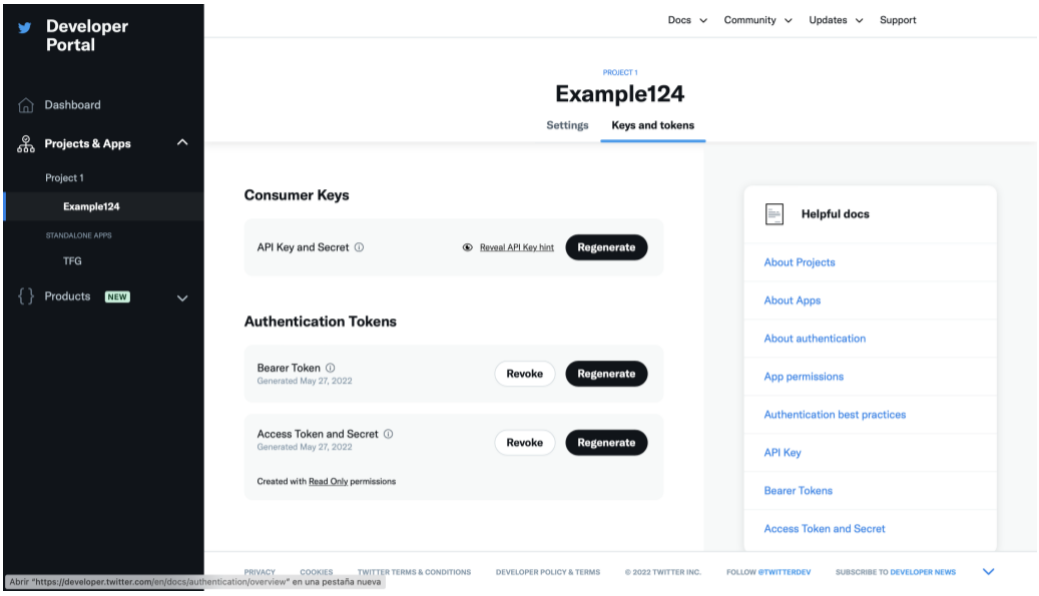


Figura 4. Captura de la sección de claves y tokens dentro de una aplicación

Fuente: Twitter

Para acceder a la API se utiliza una librería de Python llamada Tweepy, que actúa como conector y permite mediante métodos de Python la extracción de tweets. Concretamente, se utiliza el método *Client*, al que se le indica como parámetro el Bearer Token para la autentificación, y posteriormente se utiliza la función *get_users_tweets*, asociada a este método, que corresponde con el *endpoint* "GET /2/users/:id/tweets". Esta función permite obtener los últimos tweets del usuario indicado, pudiendo insertar también como parámetro, el número de tweets a extraer. (Roesslein, 2009)

De esta forma, se extraen los 50 últimos tweets de cada una de las cuentas anteriormente mencionadas, y finalmente se obtiene un conjunto de datos compuesto por 1000 casos.

Una vez obtenidos los 1000 tweets se deben etiquetar manualmente en función de si poseen o no innovación, puesto que el objetivo es que, mediante aprendizaje automático, se aprenda de las características de los textos, para posteriormente poder predecir si existe o no innovación, y para esto, los modelos primero deben aprender qué características de los textos se asocian a la innovación.

El etiquetado de innovación es manual, un tweet se etiqueta como innovación si en él se habla directa o indirectamente de innovación de producto, considerando innovación de producto como un artículo nuevo o mejorado, que difiere significativamente de los artículos anteriores.

id	text	innov	cuenta
1493183145214152711	¿Y si por #SanValentín le envías esta imagen a tu crush? https://t.co/KxLX1XiBwm	0	JomaSport
1493133271659651072	RT @bbmarathon: Seguimos con buenas noticias 🤩 para la @aspanova Bilbao Bizkaia 10K 👏. Las personas inscritas podrán beneficiarse 😎 del 40…	0	JomaSport
1491761617998127110	😍 #JomaTeam #SoloporDeporte https://t.co/8Aqd26qj1C	0	JomaSport
1491687933665026049	@MXiaoPodium @rfetm_tenismesa @galiadvorak Fue un placer tener a dos jugadoras olímpicas en nuestras instalaciones. Joma es la casa del deporte. Fuerte abrazo!	0	JomaSport
1491671037238554625	PODIO!!!! #JomaTeam #ElCorazónDeEspaña https://t.co/0lEDBGM4sc	0	JomaSport
1491371644417613829	🙌 Hola @RFETenis 🆕 Partner en el mundo de la raqueta. 🎾 Conoce todos los detalles https://t.co/7MeFDxftdj 👈 #SoloPorDeporte https://t.co/c7wbCLb17R	0	JomaSport
1491080037671575556	@amesti3 ENHORABUENA! Eres el ganador de dos entradas, te escribiremos por privado.	0	JomaSport
1491035074191831041	RT @JomaSport: Vive un #DerbiJoma en Ipurua @SDEibar 🆚 @Fuenla Síguenos, dale RT al post y gana dos entradas. Anunciaremos el ganador el…	0	JomaSport
1490717509963636743	Vive un #DerbiJoma en Ipurua @SDEibar 🆚 @Fuenla Síguenos, dale RT al post y gana dos entradas. Anunciaremos el ganador el martes a las 17:00 h. #JomaTeam #SoloporDeporte https://t.co/kfm3ODGj56	0	JomaSport
1489596401986285584	👏👏👏 #JomaTeam https://t.co/pMz6YkkS5N	0	JomaSport

Figura 5. Captura de 10 tweets sin innovación

Fuente: Base de datos obtenida de Twitter

id	text	innov	cuenta
1491370727383388164	RT @InterMovistar: ⁉️ ¿Podrías definir las nuevas TOP FLEX en una palabra? 💬CONSÍGUELAS dejando tu palabra en comentarios y siguiendo a…	1	JomaSport
1503355606685696002	Obtén el equilibrio perfecto entre lo deportivo y lo femenino con las nuevas zapatillas Tenerife. Consigue las zapatillas TENERIFE en nuestra tienda online 👜 https://t.co/SycEA3OIZm #pedromiralles #primaveraverano22 #springsummer22 #newcollection #sneakers #zapatillas https://t.co/8FWIRM86ge	1	pmcalzado
1481605000543539202	RT @SeFutbol: 👕 Os presentamos la nueva equipación de la @sefutbol Sala bajo el lema: "Nuestra tradición es el fútbol sala". 👑 📻 Nuestro…	1	JomaSport
1481350301080768516	RT @bbmarathon: Estábamos ansiosos 👑 de poder presentaros la camiseta 👕 de la @bbmarathon🏃‍♂️, con la que pretendemos transmitir la pasión,…	1	JomaSport
1500909536605085697	Dos versiones revolucionarias con las que disfrutarás de comodidad superior y ergonomía a cada paso⚡️✒️ #ForzaSporty VS #VitaEco ¿Con cuál te quedas? ✌️ https://t.co/Fpyn1TwnJr	1	panter_calzado
1499399598988619783	La revolución sostenible en calzado empieza por #VitaEco 🌱 ⚠️Su suela fabricada en poliuretano, una vez acabada la vida útil, se recicla para otros usos como pavimentos⚡️ ¿Has llevado alguna vez un calzado comprometido con el medio ambiente? https://t.co/a7uiHonwdu https://t.co/Rj7BLtkYLD	1	panter_calzado
1495488779989172235	Innovamos para llevar tu comodidad a otro nivel🔥 La combinación de espuma reciclada y tejido PET convierten esta plantilla en la opción ideal para aportar comfort extra a tu calzado 👌 Y lo mejor, están fabricadas de manera sostenible♻️ Descubre más en https://t.co/PRkS2wVnAz https://t.co/u7ZCkiuYgi	1	panter_calzado
1496138238855102468	Somos la primera marca en obtener los sellos Comfort y Functional de Inescop en un mismo producto👏 ✅Comfort: Garantiza la comodidad del calzado. ✅ Functional: Asegura un calzado saludable y confortable. ¿Sabes qué calzado lo ha conseguido? ¡Pronto lo descubrirás!👀 https://t.co/GquzYSMwkb	1	panter_calzado
1498750358943842306	Auténticamente rompedoras🔥 No solo las querrás llevar para realizar estas actividades, también formarán parte de tu día a día 🙌 Descubre todo sobre este revolucionario calzado en https://t.co/eq1qmjfhrn📍 https://t.co/CZjhP1U2BG	1	panter_calzado
1463448346765541377	ROBUSTA NEXT ya está en los medios de comunicación > JobWear Magazine #98 Esta nueva línea representa un gran avance para nuestra marca y actualmente se encuentra desarrollando las últimas fases de producción, para comenzar su comercialización en el primer trimestre del 2022 🔷 https://t.co/HJK9kT95jA	1	CalzadosRobusta

Figura 6. Captura de 10 tweets con innovación

Fuente: Base de Datos obtenida de Twitter

Por tanto, la base de datos resultante consiste en 4 variables, id del tweet, texto del tweet, usuario que escribió el tweet e innovación en el tweet (0 = no hay innovación, 1 = hay innovación). Y 1000 casos.

4.2 Procesado de datos

La principal variable por analizar y la que nos aportará información sobre qué características están relacionadas con la innovación, es el texto del tweet. Puesto que se trata de una variable tipo texto, debemos convertirla a información de tipo numérica, ya que sino no será una entrada apta para los modelos.

La técnica empleada para transformar los datos de tipo texto a tipo numérico es *Count Vectorization*, que consiste en contar las veces que aparece cada palabra en cada tweet, para crear una matriz con tantas filas como tweets y tantas columnas como palabras diferentes haya en el conjunto de todos los textos. De esta forma, la intersección entre filas y columnas representará cuantas veces aparece la palabra de la columna en el tweet de la fila. (Pedregosa, 2013)

Además, para que el número de columnas de la matriz resultante sea menor, anteriormente se procesan los tweets eliminando el texto que no aporta información para el análisis, como son las palabras vacías, es decir, las palabras que no tienen significado, como artículos o pronombres. También se eliminan signos de puntuación, acentos, números, emoticonos, enlaces web, menciones, hashtags y palabras con 1 y 2 caracteres.

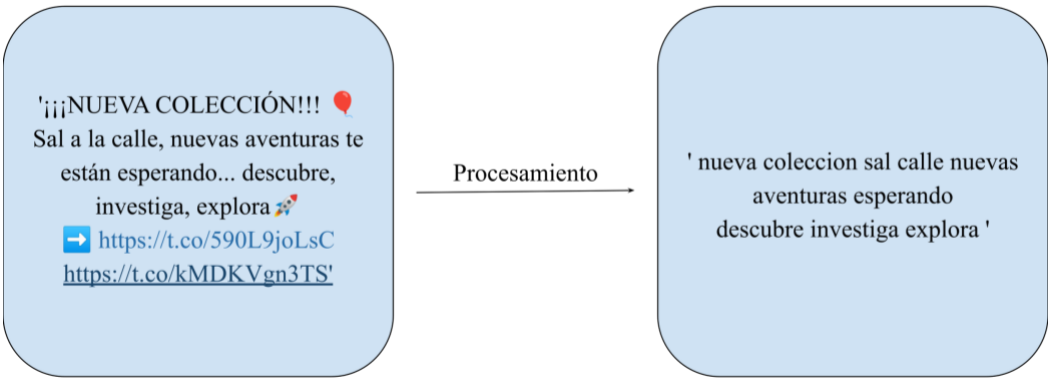


Figura 7. Ejemplo de un tweet antes y después de ser procesado

Fuente: Elaboración propia

Con todo este procesamiento se consigue que al crear la matriz de características el diccionario de palabras diferentes sea mucho menor.

Por ejemplo, si se aplicara el método *CountVectorizer* a los siguientes tweets:

- 'Evita accidentes en tu trabajo con tu nuevo calzado'.
- 'Nueva colección disponible en tu zapatería'.
- 'Te presentamos el nuevo modelo. Cómodo y elegante.'

Tras procesarlos, se obtendría la siguiente matriz:

	accidentes	calzado	coleccion	comodo	disponible	elegante	evita	modelo	nueva	nuevo	presentamos	trabajo	zapateria
0	1	1	0	0	0	0	1	0	0	1	0	1	0
1	0	0	1	0	1	0	0	0	1	0	0	0	1
2	0	0	0	1	0	1	0	1	0	1	1	0	0

Figura 8. Ejemplo de la matriz resultante tras aplicar CountVectorizer

Fuente: Elaboración propia

Una vez creada la matriz de características se eliminan las palabras que solamente aparecen en una frase, ya que de esta forma se reduce el vocabulario, y por tanto, también la cantidad de columnas de la matriz.

La otra variable por analizar, y que posteriormente será la variable salida de los modelos de predicción es la que indica si existe o no innovación en cada tweet.

Como se puede observar en el siguiente gráfico, los tweets con innovación no superan el 10% de los tweets obtenidos, por tanto, nuestro conjunto de datos está desbalanceado. Esto puede provocar que los modelos no sean capaces de aprender las características asociadas a los tweets con innovación, ya que no hay suficientes, y tiendan a predecir nuevos tweets como no innovadores, ya que tienen mayor probabilidad de acierto en su fase de entrenamiento.

Figura 9. Gráfico con el porcentaje de tweets con innovación

Fuente: Elaboración propia

Por tanto, para solucionar esto y evitar que los modelos aporten resultados sesgados, realizamos un procesamiento a la base de datos que nos permita balancearla.

El método usado para realizar este procesamiento es SMOTE, una técnica estadística que permite el equilibrado, ya que toma muestras de las características de cada caso y de las de sus vecinos más próximos, y de esta forma genera nuevos casos de la clase minoritaria.
(Lemaître & Nogueira, 2017)

Tras aplicar SMOTE, se obtiene un conjunto de datos con el 50% de tweets con innovación, es decir, un conjunto totalmente balanceado.

Figura 10. Gráfico con el porcentaje de tweets con innovación tras aplicar SMOTE

Fuente: Elaboración propia

5. Conocimiento extraído y evaluación de modelos

En este apartado se comenta el conocimiento extraído de los datos, es decir, la información que aportan. Primero se realiza un análisis descriptivo de los datos y posteriormente se aplican modelos para realizar un análisis predictivo, por último, se evalúan estos modelos.

5.1 Análisis descriptivo

Al realizar un análisis descriptivo se pretende obtener información del conjunto de datos, como qué relaciones se dan entre los datos o que características poseen.

El conjunto de datos cuenta con 4 variables: código de identificación de cada tweet, texto del tweet, existencia o no de innovación y cuenta que lo publicó.

Puesto que el código de identificación de cada tweet es único por cada fila, no aporta gran información, por tanto, no se tendrá en cuenta en este análisis. Esto deja a 3 variables por analizar, texto del tweet, presencia de innovación y cuenta que lo publicó.

Como ya se comentó anteriormente, el conjunto de datos sobre el que se trabaja contiene un bajo porcentaje de innovación, 9% de tweets con innovación frente a 91% sin innovación. Por tanto, es interesante estudiar qué cuentas dedican más tweets a la innovación.

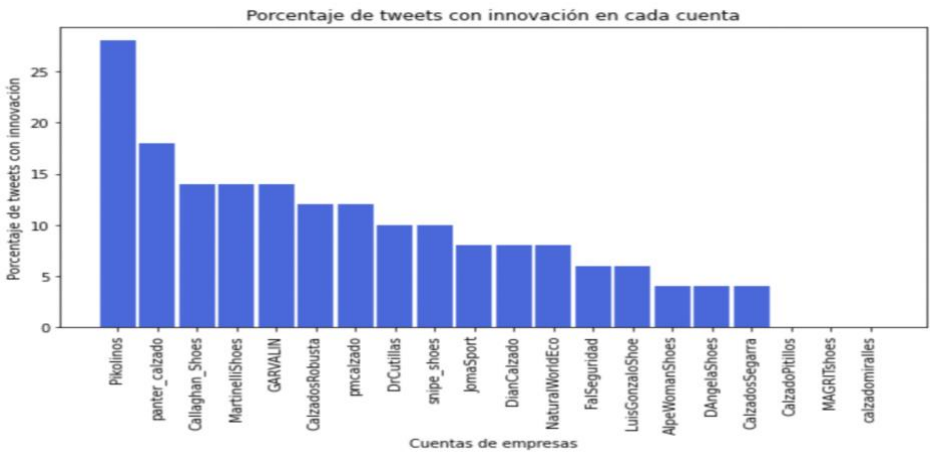


Figura 11. Gráfico de barras sobre porcentaje de tweets con innovación por cuenta

Fuente: Elaboración propia

En este gráfico se puede observar que la cuenta con mayores publicaciones sobre innovación es 'Pikolinos', pero aun así no supera el 30% de innovación en sus tweets. Otras cuentas como 'CalzadoPitillos' o 'calzadomiralles' incluso no tienen ninguna publicación relacionada con la innovación.

A partir de este gráfico se podría afirmar que en general en el conjunto de datos, la mayoría de contenido que comparten las empresas en sus cuentas de Twitter, no se trata de innovación.

Además, existe gran diferencia entre los porcentajes de innovación de algunas cuentas, a partir de lo cual se puede inferir, que las empresas tienen diferentes enfoques, ya que hay algunas de ellas que dedican parte de sus tweets a publicar innovación y otras prefieren publicar otro contenido. Esto provoca por tanto que los datos esten sesgados o no balanceados.

Por otra parte, se analizan las características de los textos que poseen innovación.

Figura 12. Gráfico nube de palabras sobre las palabras más repetidas en los tweets con innovación

Fuente: Elaboración propia

La información que se puede inferir de la nube de palabras es que las palabras que más aparecen en los tweets que contienen innovación son 'colección' y 'nueva', seguidas de 'modelo', 'novedades', 'calzado' y 'zapatilla', entre otras. Estas podrían ser las palabras que influyeran a la hora de predecir si un tweet posee innovación, pero para saberlo con certeza realizamos un análisis predictivo.

5.2 Análisis predictivo

Cuando se realiza un análisis predictivo se pretende que, a partir de los patrones de comportamiento aprendidos en un conjunto de datos, se pueda predecir nueva información.

Para realizar análisis predictivos se utiliza el aprendizaje automático, existen dos tipos de aprendizaje automático, dependiendo de si existen datos etiquetados o no. El aprendizaje automático no supervisado se basa únicamente en las entradas, es decir, en datos no etiquetados.

Por otro lado, el aprendizaje automático supervisado se basa en un corpus etiquetado. En este algoritmo encontramos una primera fase, llamada aprendizaje o entrenamiento, en la que el algoritmo aprende las características asociadas a cada una de las diferentes etiquetas del corpus. Otra de las fases, la predicción, consiste en aplicar el algoritmo ya entrenado a un nuevo volumen de datos sin etiquetar, pero de la misma naturaleza, para poder predecir posteriormente los valores de las etiquetas en función de las características aprendidas.

Dentro de los algoritmos supervisados encontramos dos familias, de regresión o de clasificación. Los algoritmos de regresión proporcionan una salida numérica, mientras que la salida de los algoritmos de clasificación es una categoría. (Baviera, 2017)

Este trabajo se centra en los algoritmos supervisados de clasificación, considerando como variable a clasificar la presencia de innovación en el tweet. Concretamente se aplican los clasificadores citados a continuación.

Para la aplicación de estos clasificadores se utiliza Scikit-learn que es una librería de software libre para el lenguaje de programación Python. Esta librería se utiliza para aprendizaje automático y, por tanto, permite aplicar algoritmos de aprendizaje supervisado.
(Pedregosa, Scikit-learn: Machine learning in Python., 2011)

5.2.1 Máquinas Soporte Vectorial

El algoritmo máquinas de soporte vectorial (SVM) consiste en representar cada caso como un punto en el espacio, para posteriormente crear un hiperplano que divida los puntos en función de la clase a la que pertenecen. El algoritmo SVM tiene como objetivo encontrar el hiperplano que optimice los resultados de predicción, y este hiperplano es el que posee un mayor margen desde los puntos más cercanos de cada clase, también conocidos como vectores de soporte.

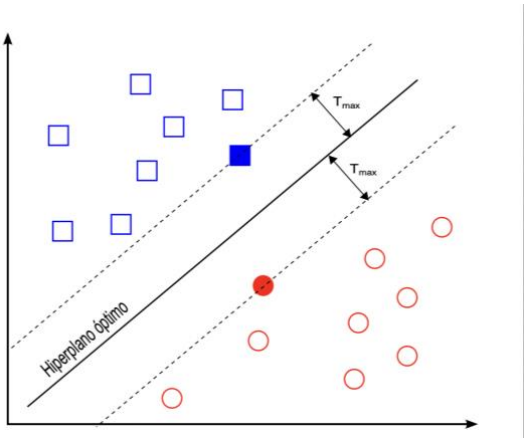


Figura 13. Captura de un hiperplano óptimo

Fuente: Suárez,2014

Este algoritmo recibe ciertos parámetros que permiten que los resultados puedan ser más ajustados al conjunto de datos que se pretende predecir. El parámetro C, es un parámetro de regularización que permite que en el caso de que el algoritmo no sea capaz de dividir los datos de forma precisa en dos clases, se penalicen los datos mal clasificados. Por otra parte, también se toma como parámetro la función kernel, que es la función que clasifica los puntos en una clase u otra, y por tanto le da forma al hiperplano. Las funciones kernel pueden ser lineales, polinómicas, sigmoides, o de base radial o guassiana. (Suárez, 2014)

Tras aplicar el algoritmo, se obtienen los coeficientes de las palabras que más influyen en predecir la presencia de innovación en un tweet. Esto podría indicar que los nuevos tweets que contengan estas palabras serán clasificados como innovación. Las palabras son las siguientes.

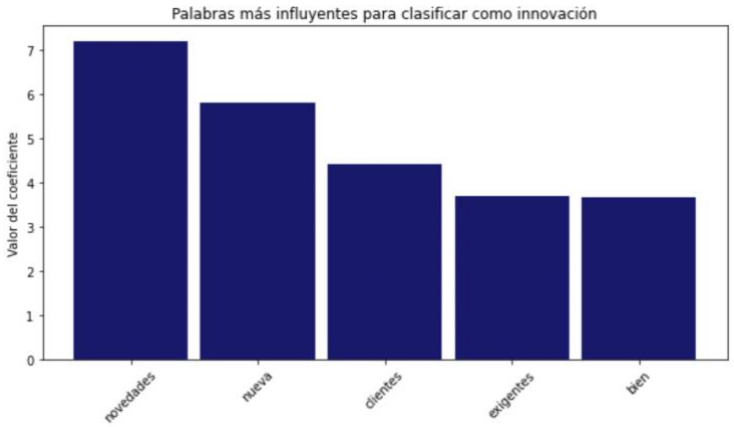


Figura 14. Gráfico de coeficientes más influyentes en soporte vectorial con los datos sin balancear

Fuente: Elaboración propia

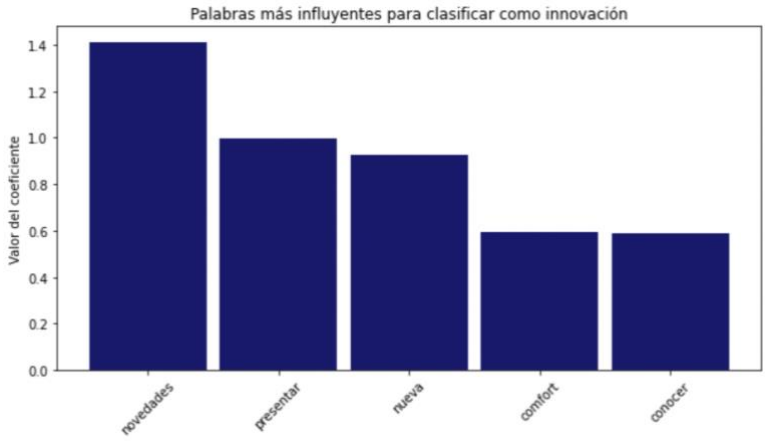


Figura 15. Gráfico de coeficientes más influyentes en soporte vectorial con los datos balanceados

Fuente: Elaboración propia

5.2.2 Regresión logística

La regresión logística consiste en la clasificación de un conjunto de datos basándose en probabilidades. En este modelo, la variable dependiente es la variable a predecir, mientras que las variables independientes son con las que se entrena el modelo. Este algoritmo permite estudiar si la variable a predecir depende o no del resto de variables del conjunto de datos.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

β_k son los coeficientes estimados a partir de los datos

X_k son las variables independientes, las características de los tweets

Y es la variable dependiente, la presencia de innovación

Ecuación 1. Función lineal del modelo de regresión logística

Al entrenar el modelo, se estiman los coeficientes que multiplican a cada una de las variables independientes. En esta estimación, las características que influyen a que un tweet que no contenga innovación (innov = 0) tendrán un valor negativo en sus coeficientes. Mientras que las características que influyen en predecir un tweet con innovación tendrán coeficientes con valores positivos. (Berlanga Silvestre & Vilà-Baños, 2014)

Tras aplicar el modelo, se puede observar que las palabras cuyos coeficientes influyen positivamente a la predicción, es decir, influyen en predecir presencia de innovación, son las siguientes.

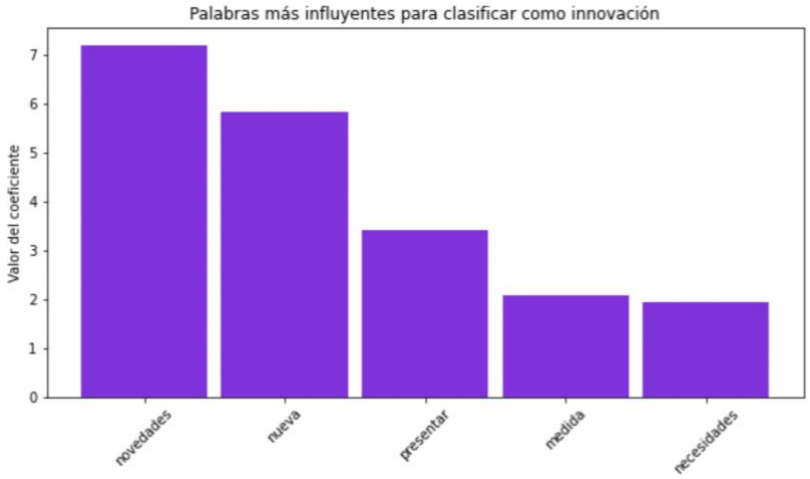


Figura 16. Gráfico de coeficientes más influyentes en regresión logística con los datos sin balancear

Fuente: Elaboración propia

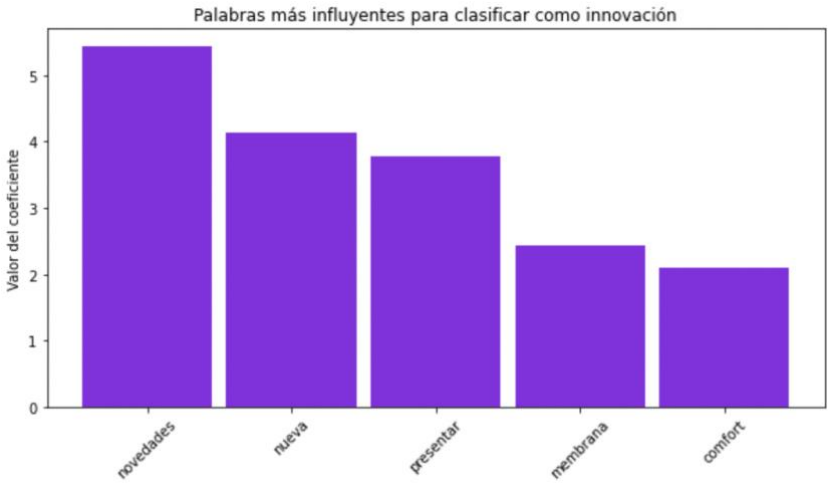


Figura 17. Gráfico de coeficientes más influyentes en regresión logística con los datos balanceados

Fuente: Elaboración propia

5.2.3 Naive Bayes

Estos algoritmos basan su funcionamiento en el teorema de Bayes, que consiste en establecer la probabilidad de un evento en función de lo que ha ocurrido anteriormente. Por tanto, la probabilidad de que ocurra el suceso A teniendo en cuenta lo que ocurre en el conjunto de entrenamiento B, P(A|B), es igual a la probabilidad de que se cumpla el evento A en el conjunto de entrenamiento B, P(B|A), multiplicado por la probabilidad a priori del evento A, y dividido por la probabilidad del conjunto de entrenamiento B.

$$P\ (A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

Ecuación 2. Teorema de Bayes

Por otro lado, el algoritmo supone que las variables del conjunto de datos son totalmente independientes entre sí, es por esto por lo que se llama "Naive", ingenuo en inglés, ya que este hecho rara vez sucede.

Dentro de los algoritmos Naive Bayes encontramos diferentes clasificadores que difieren entre ellos por la distribución que asumen que siguen las variables del conjunto de datos. Entre ellos se encuentran el clasificador Bernoulli, adecuado para variables binarias, el clasificador Gaussiano, para variables con distribuciones continuas, y por último, el clasificador multinomial adecuado para variables discretas y comúnmente usado para clasificación de textos, que es el que se aplica en este proyecto. (Zhang, 2004)

5.3 Evaluación de modelos

Existen diferentes técnicas para la evaluación de modelos como son la validación cruzada o la partición del conjunto de datos en entrenamiento y prueba. Estas técnicas consisten en utilizar cierta cantidad de datos para entrenar el modelo y otra proporción de estos para evaluarlo, es decir, su finalidad es medir la exactitud de predicción del modelo ante nuevos datos a predecir.

En este trabajo se utiliza el método de evaluación que consiste en dividir el conjunto de datos en una partición de entrenamiento y otra partición test, concretamente un 75% de los datos se dedican a la primera partición y un 25% a la segunda. Este método se diferencia de la validación cruzada

en que el proceso de entrenamiento del modelo se aplica una única vez a la partición de entrenamiento, y posteriormente se evalúa prediciendo también una única vez sobre la partición de prueba. La validación cruzada, sin embargo, realiza varias iteraciones en las que cada vez toma una proporción de datos como entrenamiento y otra como test, y en cada iteración va cambiando las particiones.

Tras haber entrenado los modelos con el 75% de los datos, se procede a predecir el 25% de datos restantes, para posteriormente comparar las predicciones realizadas con los valores reales de las etiquetas y poder observar que porcentaje de acierto tiene cada modelo sobre el conjunto de datos.

Puesto que el objetivo de este proyecto es que se detecte la innovación en los tweets, es preferible que el modelo prediga que los tweets contienen innovación, aunque no sea cierto y posteriormente sea necesario corregir manualmente esta clasificación, a que haya tweets con innovación que no sean detectados. Ya que, si un tweet con innovación es clasificado como lo contrario, se omite información valiosa para la empresa.

Por tanto, una medida de evaluación adecuada para valorar qué modelo identifica mejor la innovación es el “Recall”, ya que esta métrica indica el porcentaje de innovación que el modelo ha sido capaz de identificar. La matriz de confusión también es un buen indicador para ver la buena clasificación de la innovación, ya que permite analizar visualmente como han sido clasificados los tweets.

Por otro lado, la curva ROC, del inglés Receiver Operating Characteristic, en castellano Característica Operativa del Receptor, consiste en un gráfico en el que se representan el ratio de falsos positivos en el eje X, frente al ratio de verdaderos positivos en el eje Y. A partir de este gráfico se puede obtener la métrica AUC, del inglés Area Under the Curve, que mide el área bajo la curva ROC e indica la probabilidad de que el modelo pueda distinguir entre la clase positiva y la clase negativa.

Además, también se utiliza como medida para evaluar los modelos el “Accuracy”, que aunque no se centra particularmente en la clasificación de tweets con innovación, indica el porcentaje de predicciones totales que el algoritmo es capaz de realizar correctamente.

A continuación, se muestran los valores de estas métricas para cada uno de los algoritmos aplicados, con los datos balanceados y sin balancear, para posteriormente poder decidir cuál es el que ofrece mejores resultados.

	Accuracy	Recall	AUC
Naive Bayes	91.6%	27%	62.54%
Soporte Vectores	92.8%	36%	67.3%
Regresión Logística	92.8%	41%	69.36%
Naive Bayes (SMOTE)	**86.4%**	**64%**	**76.12%**
Soporte Vectores (SMOTE)	75.6%	40%	68.2%
Regresión Logística (SMOTE)	78.8%	59%	69.89%

Tabla 3. Valores de métricas de evaluación para los diferentes modelos

Fuente: Elaboración propia

Tras observar la Tabla 1, se puede afirmar que todos los modelos mejoran las predicciones positivas después de balancear el conjunto de datos, aunque lo realizan a costa de empeorar las predicciones de tweets sin innovación, lo que provoca que disminuya el valor de la métrica Accuracy. Puesto, que el objetivo es detectar innovación, se asume el aumento de falsos positivos, con la finalidad de aumentar también el número de verdaderos positivos.

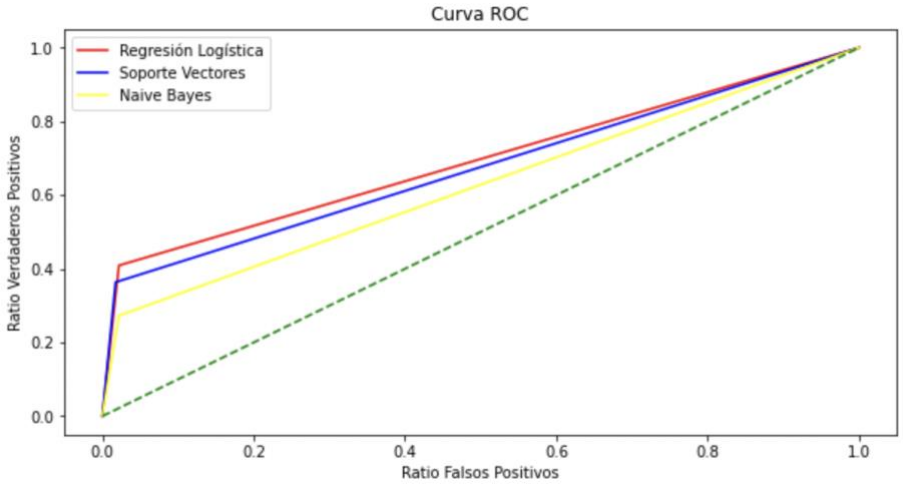


Figura 18. Gráfico con las curvas ROC de los modelos con el conjunto de datos sin balancear

Fuente: Elaboración propia

Además, se puede observar en la curva ROC, que antes de balancear el conjunto de datos, el modelo que mejores predicciones realiza es el de Regresión Logística.

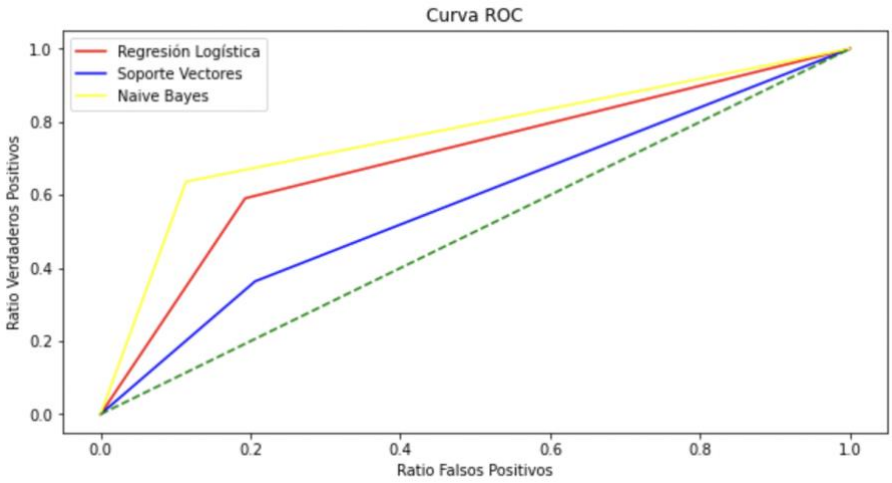


Figura 19. Gráfico con las curvas ROC de los modelos con el conjunto de datos balanceado

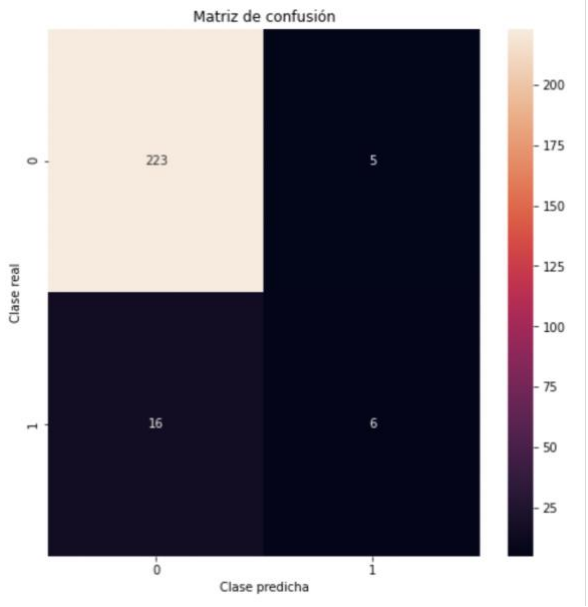
Fuente: Elaboración propia

Pero, sin embargo, una vez balanceados los datos, se observa que el algoritmo Naive Bayes, con un valor AUC del 76.12%, es el que realiza mejores predicciones. Como también se puede observar en el valor de su Recall, ya que acierta un 64% de casos positivos.

Otra forma de observar el resultado de las predicciones más visual es la matriz de confusión. A continuación, se muestran las diferentes matrices de confusión para todos los modelos.

Naive Bayes

Sin Balancear



Balanceado

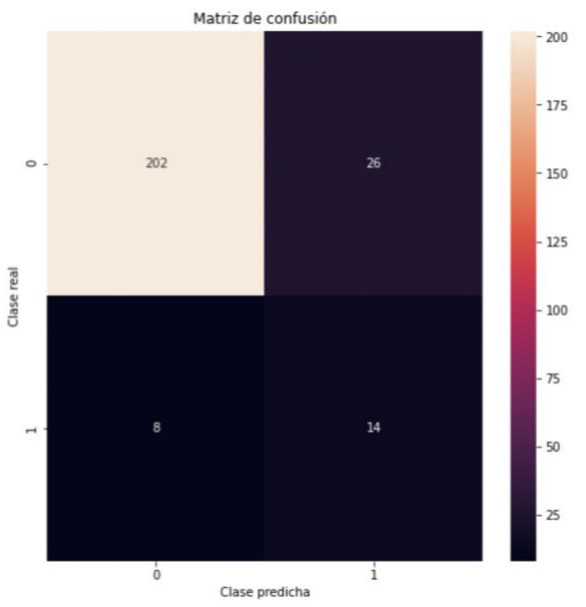
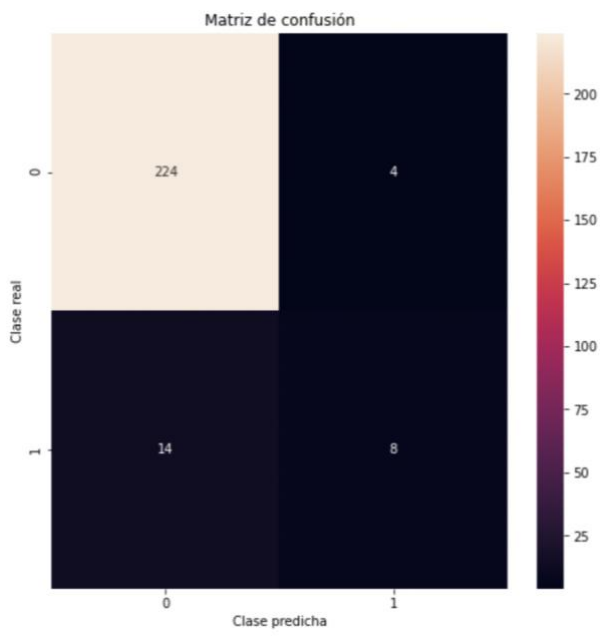


Figura 20. Matrices de confusión resultantes tras aplicar Naive Bayes

Fuente: Elaboración propia

Máquinas Soporte vectorial

Sin Balancear



Balanceado

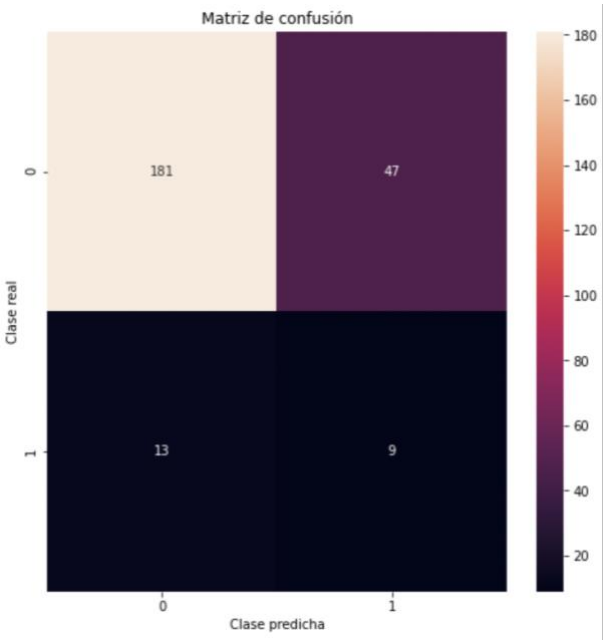
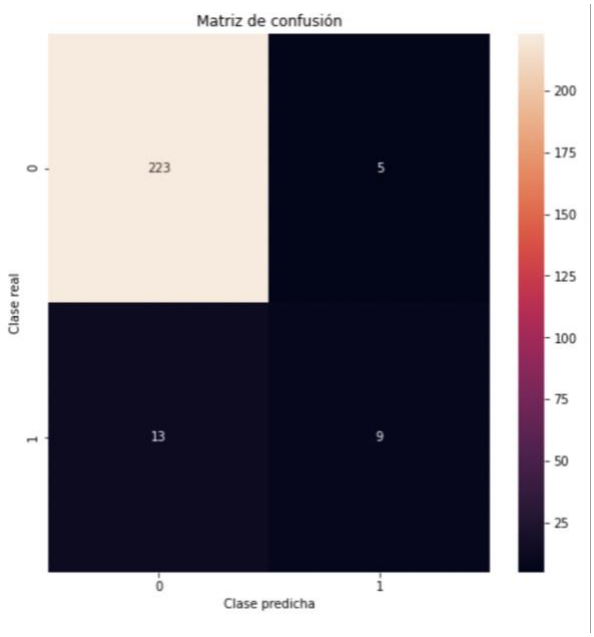


Figura 21. Matrices de confusión resultantes tras aplicar Máquina de Soporte Vectorial

Fuente: Elaboración propia

Regresión Logarítmica

Sin Balancear



Balanceado

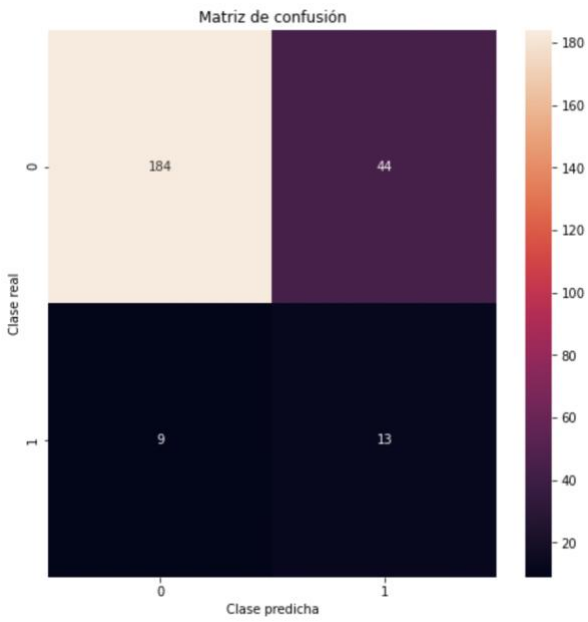


Figura 22. Matrices de confusión resultantes tras aplicar Regresión Logística

Fuente: Elaboración propia

Tras observar las matrices de confusión, podemos concluir también que los tres modelos mejoran las predicciones de verdaderos positivos después de aplicar SMOTE para balancear el conjunto de datos. Además, se puede concluir nuevamente, que el modelo que mejor resultados ofrece para detectar tweets con innovación, es el algoritmo Naive Bayes, habiendo balanceado previamente el conjunto de datos.

También es interesante analizar el coste temporal de cada uno de los algoritmos, para ver cuál es más eficiente.

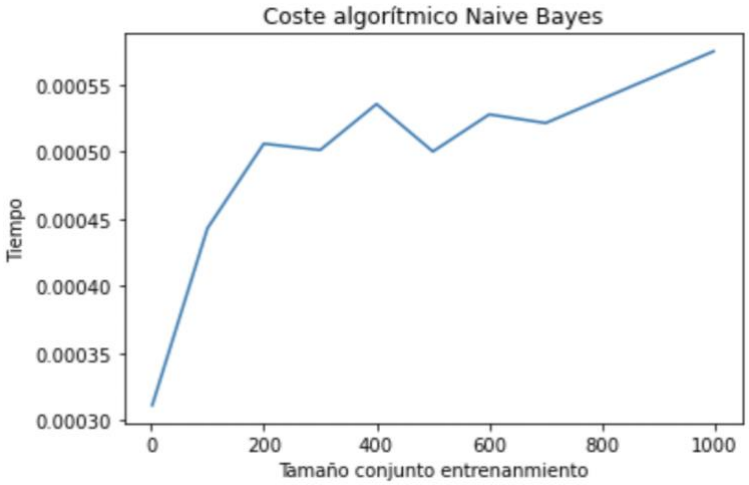


Figura 23. Gráfico del coste temporal del algoritmo Naive Bayes

Fuente: Elaboración propia

El coste del algoritmo Naive Bayes es logarítmico, $O(\log(x))$, ya que mientras aumenta la talla del conjunto de datos de entrenamiento, el tiempo necesario aumenta siguiendo aproximadamente una función logarítmica.

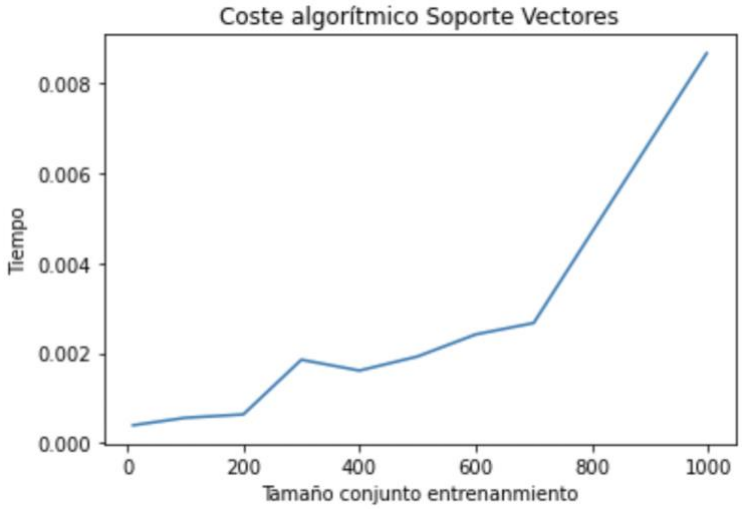


Figura 24. Gráfico del coste temporal del algoritmo Soporte Vectores

Fuente: Elaboración propia

En cuanto al algoritmo Soporte Vectores, su coste es cuadrático, $O(x^2)$, puesto que, al aumentar la cantidad de datos del conjunto de entrenamiento, el tiempo invertido aumenta prácticamente de forma cuadrática.

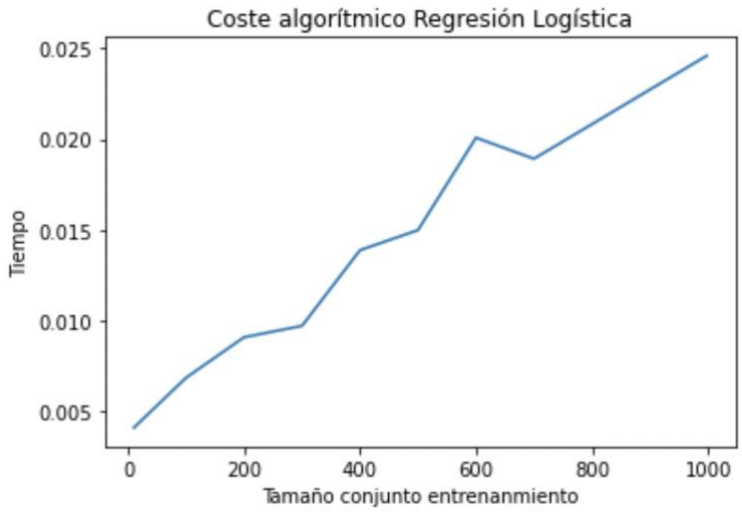


Figura 25. Gráfico del coste temporal del algoritmo Regresión Logística

Fuente: Elaboración propia

El coste del algoritmo Regresión Logística es lineal, $O(x)$, lo cual indica que a medida que crecen los datos del conjunto de entrenamiento, el tiempo en entrenar el algoritmo crece linealmente.

Por tanto, el logaritmo más eficiente a nivel temporal es el algoritmo Naive Bayes, ya que, con conjuntos de datos de mayor volumen, el aumento del coste temporal es menor que con el resto de algoritmos.

Por último, cabe mencionar que los algoritmos se exponen a algunos riesgos como son que los usuarios dejen de confiar en el sistema, debido a que el clasificador detecte un elevado número de falsos positivos, y por tanto el cliente prefiera revisar el Twitter de la empresa a monitorizar manualmente. O en el caso opuesto, que el clasificador obtenga una cantidad elevada de falsos negativos y, por tanto, ocasione malos resultados al cliente, ya que puede perder información de gran valor. Para reducir el impacto de ambos riesgos, la principal solución es mejorar el modelo y asegurarse de su correcto funcionamiento, lo cual puede llevarse a cabo realizando mantenimientos periódicos, como se explicará en el siguiente apartado.

6. Validación y despliegue

En este apartado se exponen diferentes enfoques de aplicación y despliegue para el modelo creado. Además, se indican formas convenientes de cómo realizar un mantenimiento del modelo para asegurar su correcto funcionamiento.

Una posible y valiosa aplicación del modelo sería el uso por parte de las empresas del sector del calzado, con el fin de monitorizar e investigar los comportamientos en innovación de empresas competidoras, para poder capacitarse mejor y conseguir satisfacer las necesidades de los clientes mejor que la competencia. Otra posible aplicación de este clasificador podría ser el uso por parte de instituciones públicas, con el fin de monitorizar la innovación y poder analizar si las políticas implementadas en el sector producen resultados adecuados o necesitan ser mejoradas.

Para comprobar la validez y correcto despliegue del modelo se realiza una prueba en la que se clasifican 15 nuevos tweets de tres empresas diferentes, Pikolinos, NaturalWorldEco y MartinelliShoes. Tras la clasificación se analizan los resultados con el fin de extraer información útil y comprobar si la clasificación se ha realizado correctamente, aunque si se tratara de su verdadero uso, se confiaría en el clasificador y solamente se corregirían los falsos positivos, lo que quiere decir, que los tweets con innovación mal clasificados, es decir, los falsos negativos, no serían detectados, y por tanto sería información útil que no llegarían a conocer bien la institución pública o bien la empresa.

La obtención de estos nuevos tweets se realiza utilizando la API de Twitter como anteriormente, y el procesamiento es el mismo que se realiza al conjunto de datos inicial, es decir, se eliminan las palabras vacías, acentos y números, entre otros caracteres.

El modelo que recibe estos datos como entrada para realizar la clasificación, es el algoritmo Naive Bayes con el conjunto de datos balanceado, puesto que es el modelo que mejores resultados ofrece en su evaluación.

Texto del tweet	Predicción	Clasificación correcta
RT @YorokobuMag: De lejos parece asqueroso, pero de cerca es 😍😍😍 https://t.co/AXPOx6PBOT	No Innovación	Sí
La nueva colección está repleta de modelos versátiles y atemporales, ¿La has visto ya? https://t.co/pptY8IsRnl #ecofriendly #vegan #shoesaddict	**Innovación**	Sí
😍😍😍 https://t.co/AVeWZCSV1K	No Innovación	Sí
La belleza de la fabricación artesanal https://t.co/pptY8IsRnl #artesania https://t.co/hdH7XB6nqp	No innovación	Sí
La primera comunión es un acontecimiento único que los peques recordarán toda su vida, ¿Sabes que tienes que tener en cuenta a la hora de elegirlo? ¡Te lo contamos! https://t.co/BAvUkaDBvr #primeracomunión https://t.co/cp3G8oEUV3	No innovación	Sí

Tabla 4. Resultado de predecir innovación en tweets de la cuenta NaturalWorldEco

Fuente: Elaboración propia

Texto del tweet	Predicción	Clasificación correcta
Avance de la nueva colección de Martinelli. Salones metalizados perfectos para eventos en rosa, oro, plata y negro. #martinelli #newcollection #heels #fashion https://t.co/jlhXIjWg3m	**Innovación**	Sí
¿Cómo limpiar tu zapato de piel Antic? Vuestras dudas quedarán resueltas en 4 sencillos pasos. ¿Qué os parece este contenido? ¿Útil? #howtoclean #zapatoselegantes https://t.co/45PZwuULeJ	No Innovación	Sí
¿Cuántos eventos tienes este año? ¿Más de 3? No lo olvides: nuestro sorteo "The Biggest Shoe Closet" sigue activo recuerda participar en el post del sorteo 👉https://t.co/AqnJXsfp33 #martinelli #sorteoespaña #gratis#premio https://t.co/1nwQ1zoLP1	No Innovación	Sí
@OCarraminana Hola Oscar, sí en CASAS y El Corte Inglés también puedes consultar otros puntos de venta en este enlace: https://t.co/oDhjoEGeX7	No innovación	Sí
La sandalia ideal para los eventos viene de la mano de Martinelli. Tacón metalizado con un estampado bicolor y unas tiras que abrazan el tobillo.	No innovación	**No**

Tabla 5. Resultado de predecir innovación en tweets de la cuenta MartinelliShoes

Fuente: Elaboración propia

Texto del tweet	Predicción	Clasificación correcta
Te traemos el modelo más buscado y en los colores de la temporada 💝 https://t.co/X0ZoTtfzMe https://t.co/4uGx8shGPy	**Innovación**	Sí
Mallorca, una vida intensa bajo el cielo mediterráneo ☁️🌊🌞. Los más atrevidos son capaces de recorrer la isla entera a base de darle al zapato 👟. ¿Eres capaz de convertirte en un aventurero del caminar? https://t.co/2FEZYBhVDa https://t.co/bGJpD5O3I7	**Innovación**	**No**
Uno de los modelos más buscados de la temporada. La combinación perfecta entre comodidad y tendencia. ¿Aún no te has hecho con él? https://t.co/ejgdQ2LAAK https://t.co/j6Jm8bL6Eq	No Innovación	Sí
Ponte tus sandalias y sal a comerte el mundo con la máxima comodidad. ¿Cuáles son tus favoritas? https://t.co/j8oXyyz0fD https://t.co/hCdSGzGO6r	No innovación	Sí
¿Has sentido un flechazo con este modelo? Nosotros sí 😍 Menciona a esa persona que crees que se enamorará al instante de estos Pikolinos… https://t.co/9jrSb4MZY2 https://t.co/Lv1mZEcVtI	No innovación	Sí

Tabla 6. Resultado de predecir innovación en tweets de la cuenta Pikolinos

Fuente: Elaboración propia

Tras la clasificación se obtienen 4 tweets con innovación. Pero hay uno de ellos que ha sido clasificado erróneamente, ya que en realidad no contiene innovación:

"Mallorca, una vida intensa bajo el cielo mediterráneo ☁️🌊🌞. Los más atrevidos son capaces de recorrer la isla entera a base de darle al zapato 👟. ¿Eres capaz de convertirte en un aventurero del caminar? https://t.co/2FEZYBhVDa https://t.co/bGJpD5O3I7 ".

Aunque se trata de un error de predicción, no supone tanto coste para las empresas o las instituciones públicas como lo hace un tweet con innovación que se clasifica erróneamente, ya que en esta situación se lee el tweet y si no posee innovación se descarta, pero si no se detecta el tweet con innovación, esa información ni siquiera llega a oídos del cliente. Esto podría provocar en las empresas una reacción tardía a lo que hace la empresa competidora y ocasionaría posibles pérdidas de oportunidad. Y en cuanto a las instituciones públicas, podría provocar la no identificación del mal funcionamiento de una política y, por tanto, las malas consecuencias al no mejorar esta política pública a tiempo.

Esto ocurre con el siguiente tweet que no ha sido detectado por el clasificador:

“La sandalia ideal para los eventos viene de la mano de Martinelli. Tacón metalizado con un estampado bicolor y unas tiras que abrazan el tobillo.”

A partir de los tweets clasificados correctamente como innovación se puede obtener información de qué productos y qué novedades han introducido las empresas en el mercado. Se pueden analizar los resultados desde dos perspectivas diferentes, como empresa en el sector del calzado que quiere monitorizar la innovación de otra empresa competidora, o como institución pública que quiere analizar los resultados de las políticas implementadas en el sector.

En el caso del siguiente tweet publicado por la empresa Pikolinos, la información que se obtiene es que han incorporado al mercado un modelo de sandalias de los colores que se consideran de moda en la temporada, y estos colores son blanco, verde y marrón.

“Te traemos el modelo más buscado y en los colores de la temporada 💔 https://t.co/X0ZoTtfzMe https://t.co/4uGx8shGPy”

Figura 26. Captura de pantalla de unas sandalias de Pikolinos

Fuente: Página web Pikolinos

A partir del tweet publicado por la empresa NaturalWorldEco, se pueden observar las características de las zapatillas de su nueva temporada, además se puede obtener información útil para las instituciones públicas, ya que es una empresa que trata de cuidar del medio ambiente y que fabrica su calzado con materiales libres de sufrimiento animal.

"La nueva colección está repleta de modelos versátiles y atemporales, ¿La has visto ya? https://t.co/pptY8IsRnl #ecofriendly #vegan #shoesaddict".

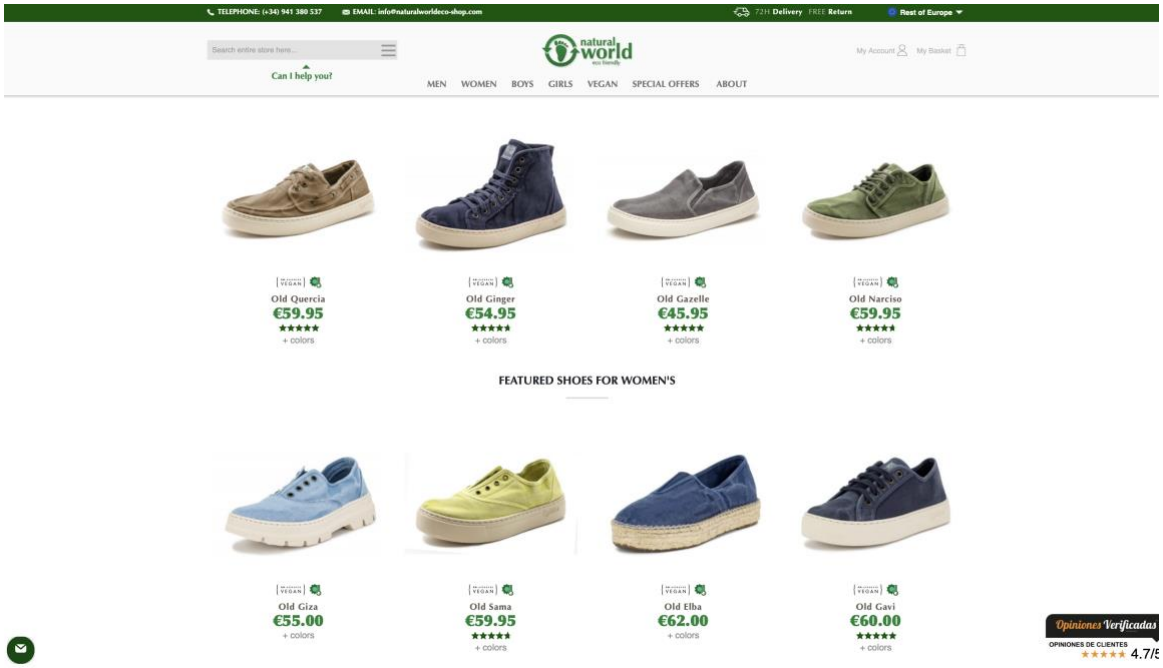


Figura 27. Captura de pantalla de la nueva colección de zapatillas de NaturalWorldEco

Fuente: Página web NaturalWorldEco

Por último, en cuanto al tweet publicado por MartinelliShoes, se obtiene la información de que esta empresa pronto introducirá en el mercado una nueva colección y que entre los zapatos de esta colección habrá unos de tacón metalizados, disponibles en varios colores, rosa, oro, plata y negro.

"Avance de la nueva colección de Martinelli. Salones metalizados perfectos para eventos en rosa, oro, plata y negro. #martinelli #newcollection #heels #fashion https://t.co/jlhXIjWg3m "

Figura 28. Captura de pantalla de un zapato de MartinelliShoes

Fuente: Página web MartinelliShoes

Además del despliegue del modelo, para asegurar el adecuado funcionamiento de este, es necesario un mantenimiento. Este mantenimiento consiste en la actualización del conjunto de datos sobre el que el clasificador aprende. Aunque la parte de extracción de nuevos tweets para mantener actualizada la base de datos de entrenamiento se puede automatizar, hay una parte que por el momento sólo podría realizarse manualmente, esta parte consiste en etiquetar el conjunto de datos en función de si existe, o no, innovación en cada uno de los tweets.

La necesidad de actualización periódica del conjunto de datos de entrenamiento viene dada por la constante, y cada vez más rápida, evolución del lenguaje, ya que la lengua evoluciona a la par que la tecnología, debido a la creación de palabras que definen nuevos conceptos y la constante búsqueda de optimización del lenguaje para poder expresar ideas complejas, sobre todo en el campo de la innovación. Por tanto, si no se actualizara periódicamente el conjunto de datos, a lo largo del tiempo, esto provocaría un decremento significativo del rendimiento, ya que habría nuevas palabras asociadas a la innovación, que no serían detectadas por no haberlas tenido en cuenta en el momento de entrenamiento del modelo.

Un ejemplo de cómo evoluciona el lenguaje, pueden ser algunas expresiones que podrían utilizarse para definir innovaciones hace unos años, como: '*Chachi piruli Juan pelotilla*', o '*Guay del paraguay*', y que hoy en día se encuentran prácticamente en desuso.

7. Conclusiones

En este capítulo se revisa el trabajo realizado y cómo se han alcanzado los objetivos, se especifica el legado proporcionado, se describe la relación con los estudios cursados y se mencionan posibles trabajos futuros.

7.1 Trabajo realizado

Este trabajo ha llevado a cabo la construcción y evaluación de un clasificador para detectar la innovación de producto del sector del calzado en Twitter, con un porcentaje de aciertos superior a 76%.

En cuanto a los objetivos específicos planteados, también se han llevado a cabo exitosamente. Se ha conseguido comprender el concepto de innovación, los tipos de innovación en general y la innovación de producto en particular, además de realizarse un estudio que ha permitido el conocimiento del contexto del sector del calzado en España, para posteriormente poder realizar la obtención automática de datos sobre el sector del calzado en Twitter, accediendo a su interfaz mediante librerías de Python.

Otro de los objetivos conseguidos es la aplicación de técnicas de procesado de lenguaje natural, que han permitido realizar tanto la limpieza de textos de palabras sin significado, para la obtención de mejores resultados, como la extracción de características numéricas, mediante *CountVectorizer*, para que estos datos sean una entrada válida para los modelos.

Por último, también se ha llevado a cabo correctamente la implementación y comparación entre los diferentes algoritmos de clasificación, realizando el método de partición en entrenamiento y test, para detectar el algoritmo con mayor porcentaje de positivos correctamente clasificados.

Este trabajo ha supuesto la aplicación de los conocimientos adquiridos en diversas áreas, ya que se ha realizado un proyecto real desde cero, por tanto, a pesar de los conocimientos generales impartidos en las asignaturas, se ha necesitado ampliarlos realizado un estudio o documentación para la aplicación de las técnicas en este caso concreto.

Además, surgen algunos inconvenientes como el etiquetado subjetivo de innovación, que supone que el resultado pueda estar sesgado por haber entrenado los algoritmos sobre un conjunto de datos clasificado únicamente por una persona, que puede tener su concepto particular de

innovación. En este caso se debería buscar la consistencia del etiquetado, y deberían de realizar este proceso varias personas.

7.2 Legado

Este proyecto supone una mejora tanto para las empresas como las instituciones públicas del sector del calzado, ya que anteriormente, para detectar las innovaciones de producto tendrían que realizarlo manualmente, accediendo a las cuentas de Twitter de cada una de las empresas del sector y analizando uno por uno los tweets para encontrar los que contuvieran innovación. Este nuevo producto les permite ahorrar tiempo y tomar decisiones o actuar en el momento oportuno, ya que en algunas situaciones la detección manual de innovaciones podría suponer no tomar una decisión a tiempo o actuar más tarde de lo debido.

Además, el código creado para este trabajo supone un legado, ya que mediante su utilización se puede reproducir el análisis realizado y pueden obtenerse resultados similares, incluso también puede servir de base para realizar otros estudios relacionados. El código empleado está disponible en un repositorio abierto, al que puede tener acceder cualquier usuario.

Personalmente, este trabajo supone un valioso aprendizaje sobre como afrontar nuevos problemas y tomar decisiones. Además, proporciona una enseñanza sobre organización del tiempo y de los recursos, y ejercita trabajar de forma autónoma.

7.3 Relación del trabajo desarrollado con los estudios cursados

Los conocimientos empleados en este trabajo surgen de asignaturas como “Lenguaje Natural y recuperación de la información”, donde se enseña una idea general de como extraer información de redes sociales y posteriormente analizar el lenguaje natural. También se emplean conocimientos de Aprendizaje Automático, adquiridos en “Modelos descriptivos y predictivos”, junto a técnicas de evaluación o despliegue, impartidas en “Evaluación, despliegue y monitorización de modelos”. Además, para el estudio de clientes potenciales o de la situación actual de los diferentes temas que se tratan en el trabajo, se emplean conocimientos adquiridos en “Comportamiento económico y social”.

Aunque se podría decir que las asignaturas anteriormente mencionadas son las más relacionadas con las tecnologías empleadas, en este proyecto se aplican en menor o mayor medida conocimientos adquiridos en todas y cada una de las asignaturas impartidas, como la metodología a utilizar en un proyecto o la gestión de tareas, conceptos adquiridos en las asignaturas de

"Proyecto", o como los conceptos básicos de programación, adquiridos en las diversas asignaturas de "Programación".

Para la realización de este proyecto, además del conocimiento obtenido en las asignaturas impartidas, se necesitan competencias adquiridas durante los estudios cursados, como pueden ser la "Planificación y gestión del tiempo", el "Aprendizaje permanente", la "Comunicación efectiva" y la "Innovación, creatividad y emprendimiento".

7.4 Trabajos Futuros

Tras finalizar el proyecto, cabe mencionar posibles ampliaciones o mejoras, que podrían haberse realizado en el caso de disponer de mayores recursos, como temporales o humanos.

Una mejora interesante a aplicar sería que el conjunto de datos con el que se entrena el modelo pudiera ser etiquetado por varias personas, para posteriormente comparar los resultados del etiquetado y poder asegurar así su consistencia. Ya que, aunque exista una definición objetiva de innovación, la decisión de si existe innovación o no, es subjetiva y puede variar según la perspectiva que tenga la persona encargada de etiquetar. Además, también sería interesante aumentar el tamaño del conjunto de datos, lo cual conlleva a etiquetar un mayor número de tweets, y requiere de más recursos. En cuanto al procesado del lenguaje natural sería interesante la aplicación de técnicas como "stemming" o lematización con la finalidad de reducir el número de palabras diferentes hallando su lema o raíz.

En cuanto a ampliaciones, se podría implementar una aplicación mediante la cual el cliente tuviera acceso al clasificador, esta aplicación sería de fácil manejo e intuitiva, para que los clientes, ya sean una empresa o una institución pública, puedan realizar la clasificación de nuevos tweets, sin necesidad de tener conocimientos sobre programación. Además, también podría ser interesante que esta aplicación permitiera obtener nuevos tweets de empresas del sector, utilizando la API de Twitter como se ha hecho en el proyecto. El usuario indicaría la cuenta de la empresa a monitorizar y obtendría nuevos tweets que posteriormente podría clasificar.

Otra de las ampliaciones podría ser la clasificación de innovación en empresas de otros sectores. Puesto que los clasificadores solo son capaces de predecir en base a las características que aprenden, y entre sectores hay características diferentes debido a que cada uno tiene palabras específicas. Para ampliar la clasificación a otros sectores, como podría ser el sector textil, se tendría que desarrollar otro clasificador análogo al ya creado, pero obteniendo tweets de empresas de este otro sector.

Además, se podría ampliar el trabajo realizado detectando innovación en otras redes sociales como Instagram o Facebook, donde también tienen gran presencia las empresas. Para lo cual sería necesario entrenar un nuevo clasificador, debido a que, al tratarse de diferentes redes sociales, el tipo de contenido que se publica es distinto y, por tanto, los conjuntos de datos para el entrenamiento tienen diferente naturaleza.

Bibliografía

Baviera, T. (2017). Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. *Universitat de Valencia*.

Berlanga Silvestre, V., & Vilà-Baños, R. (2014). Cómo obtener un modelo de Regresión Logística Binaria con SPSS. *REIRE, Revista d'Innovació i Recerca en Educació*, 105-118.

Blazquez, D., & Domenech, J. (Volume 130 de 2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, págs. 99-113.

Bravo-Marquez, F., Mendoza, M., & Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, 86–99.

Cody, E. M., Reagan, A. J., Dodds, P. S., & Danforth, C. M. (2016). *Public opinion polling with Twitter*. Burlington: The University of Vermont.

Crépon, B. E. (1998). *"Research, innovation and productivity: An econometric analysis at the firm level".*

Fuentes Dávila Otani, R. C. (2021). *Desarrollo de una aplicación de análisis del mercado, basado en el procesamiento del lenguaje natural de la red social Twitter, con Machine Learning para predicción de éxito del lanzamiento de un.* Lima: UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS.

Galindo-Rueda, F., & Cruysen, A. V. (2016). *Comprobación de los conceptos, definiciones y preguntas de las encuestas sobre innovación: Conclusiones de las entrevistas cognitivas con directivos de empresas".* París: OECD.

INE. (s.f.). Obtenido de Encuesta sobre innovación en las empresas: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176755&menu=ultiDatos&idp=1254735576669

Lemaître, G., & Nogueira, F. (2017). *Imbalanced-learn: A Python Toolbox*. Obtenido de SMOTE.

Ntompras, C., Drosatos, G., & Kaldoudi, E. (2022). A high-resolution temporal and geospatial content analysis of Twitter posts related to the COVID-19 pandemic. *Journal of Computational Social Science*, págs. 687-729.

OECD/Eurostat. (2018). *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation.* Paris/Eurostat, Luxembourg: OECD Publishing.

Pedregosa, F. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, págs. 2825-2830.

Pedregosa, F. (2013). *Sklearn*. Obtenido de feature_extraction. text. CountVectorizer.

Política de Privacidad de Twitter. (2022). Obtenido de Twitter: https://twitter.com/content/twitter-com/legal/es/privacy

Roesslein, J. (2009). *Tweepy Documentation.* Obtenido de Tweepy Documentation v3, 5.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing,*, págs. 13-22.

Suárez, E. J. (2014). *Tutorial sobre Máquinas de Vectores Soporte (SVM).* Madrid: UNED.

Twitter API. (2022). Obtenido de Twitter: https://developer.twitter.com/en/products/twitter-api

Ybarra Pérez, J.-A., & Santa María Beneyto, M. J. (2005). *"El sector del calzado en España: retos ante un contexto de globalización".* Boletín Económico de ICE. N. 2838.

Zhang, H. (2004). *The Optimality of Naive Bayes.* Fredericton.

Anexo 1: ODS

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza.**				**X**
ODS 2. **Hambre cero.**				**X**
ODS 3. **Salud y bienestar.**	**X**			
ODS 4. **Educación de calidad.**				**X**
ODS 5. **Igualdad de género.**				**X**
ODS 6. **Agua limpia y saneamiento.**			**X**	
ODS 7. **Energía asequible y no contaminante.**		**X**		
ODS 8. **Trabajo decente y crecimiento económico.**	**X**			
ODS 9. **Industria, innovación e infraestructuras.**	**X**			
ODS 10. **Reducción de las desigualdades**.				**X**
ODS 11. **Ciudades y comunidades sostenibles**.				**X**
ODS 12. **Producción y consumo responsables.**		**X**		
ODS 13. **Acción por el clima.**		**X**		
ODS 14. **Vida submarina.**			**X**	
ODS 15. **Vida de ecosistemas terrestres.**			**X**	
ODS 16. **Paz, justicia e instituciones sólidas.**				**X**
ODS 17. **Alianzas para lograr objetivos.**				**X**

Tabla 7. Objetivos de desarrollo sostenible

Fuente: Elaboración propia

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

Puesto que el resultado de este trabajo es un clasificador de innovación, existen varios objetivos de desarrollo sostenible con los que se relaciona. Uno de los enfoques de este clasificador es el uso por parte de instituciones públicas para analizar el funcionamiento de las políticas implantadas y las consecuencias resultantes. Por tanto, España como miembro de Naciones Unidas, aplica políticas públicas para garantizar el cumplimiento de los 17 objetivos de desarrollo sostenible, aunque concretamente en el sector del calzado hay objetivos más presentes que otros.

Los objetivos de desarrollo sostenible más relacionados con este trabajo son:

- Industria innovación e infraestructuras
- Trabajo decente y crecimiento económico
- Salud y bienestar

El desarrollo sostenible de la industria, tecnología e infraestructuras tiene como resultado un aumento de la calidad de vida en todos los ámbitos de la sociedad. Por tanto, la inversión por parte de las empresas del sector del calzado en infraestructuras resulta en una economía resistente y logra una estabilidad social, mientras que con la inversión en tecnología e industria se fomentan la preservación del medioambiente y la eficiencia energética.

En cuanto al trabajo decente y crecimiento económico, el sector del calzado debe conseguir una transformación de la sociedad hacia un empleo equitativo, inclusivo y sostenible, con buenas condiciones laborales.

Los productos químicos peligrosos y la contaminación del aire, el agua y el suelo provocan un gran número de enfermedades, por tanto, es un objetivo por parte de las empresas del sector del calzado reducir estos productos para asegurar la salud y el bienestar de la población.

También se encuentran otros objetivos de desarrollo sostenible relacionados, aunque en menor medida, estos son:

- Producción y consumo responsables
- Energía asequible y no contaminante
- Acción por el clima

La producción y el consumo sostenible consisten en fomentar el uso eficiente de los recursos y la energía con el fin de proteger el medioambiente y mejorar el acceso a los servicios básicos. Por tanto, la producción responsable por parte de las empresas del sector del calzado provoca una mejor calidad de vida global.

En la actualidad se necesita cada vez más energía, pero la disponibilidad energética no es infinita. Por tanto, se pretende que se lleve a cabo la implementación de energías renovables y eficiencia energética en todos los sectores, como el del calzado, con la finalidad de que el acceso a la energía sea universal.

Todos los lugares del mundo se ven afectados por el cambio climático, ya que produce consecuencias negativas tanto en la economía como en la sociedad, provocando un impacto negativo en la vida de las personas. Por tanto, es de vital importancia que toda la sociedad contribuya a frenar el cambio climático, las empresas como las del sector del calzado pueden hacerlo reduciendo su huella de carbono, con acciones como las anteriormente mencionadas.

Además, indirectamente, este trabajo también está relacionado con ODS como “Agua limpia y saneamiento”, “Vida submarina” y “Vida de ecosistemas terrestres”. Puesto que, si en las empresas de este sector se implantara el uso de materiales reciclables y reutilizables, esto influiría positivamente en todos los ODS anteriormente mencionados.

Anexo 2: Encuesta

Como información complementaria al TFG se ha realizado una encuesta con la finalidad de conocer la opinión de las personas sobre el sector del calzado y la innovación. Para ello se realizan las siguientes preguntas:

1. ¿Cuál es tu edad?
2. ¿Con qué genero te identificas?
3. ¿Consideras importante que tu calzado sea innovador?
4. ¿Cuál es para ti la definición de innovación?
5. ¿Qué características buscas en unos zapatos/zapatillas?
6. ¿Con qué frecuencia compras un nuevo par de zapatos/zapatillas?
7. ¿Compras calzado de segunda mano?
8. ¿Consumes calzado especial u ortopédico?
9. ¿Utilizas un calzado especial en tu profesión?
10. ¿Cuánto crees gastas anualmente en zapatillas?
11. ¿Cuál es el precio máximo que estás dispuesto a pagar por unas zapatillas?

Se han encuestado a 61 personas en un rango de 14 a 73 años.

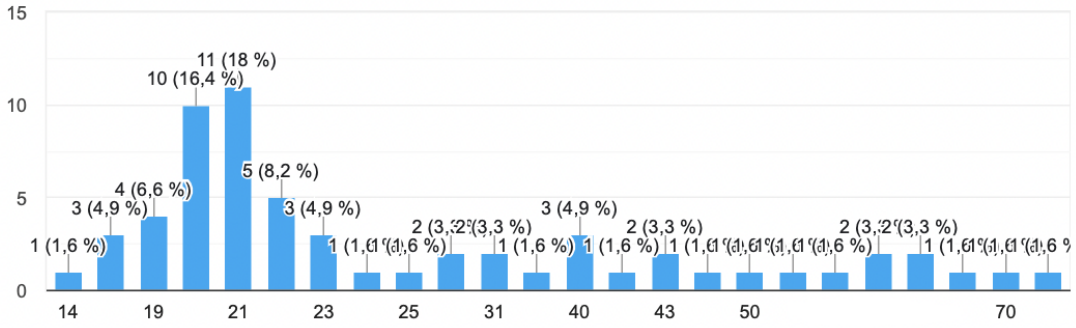


Figura 29. Histograma de edades de las personas que han contestado la encuesta

Fuente: Generado por Google

De los cuales un 39.3% son hombres y un 60.7% mujeres.

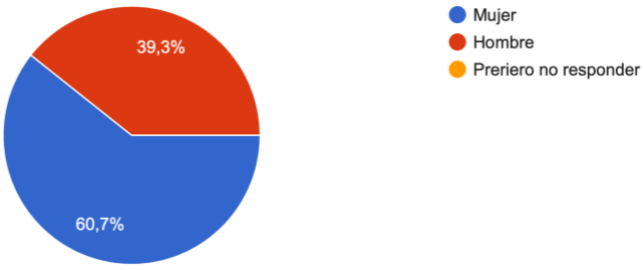


Figura 30. Gráfico con el porcentaje de género de las personas que han contestado la encuesta

Fuente: Generado por Google

Dentro de los encuestados un 59% considera importante que su calzado sea innovador, y el 41% no.

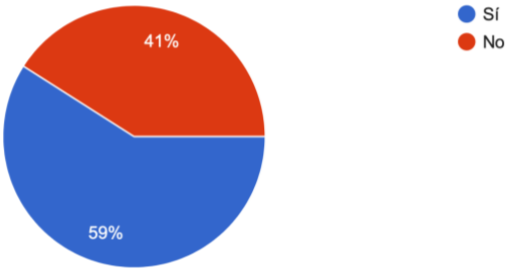


Figura 31. Gráfico con el porcentaje de importancia de innovación para los encuestados

Fuente: Generado por Google

En esta encuesta se realizan diferentes preguntas de las que se puede obtener información para realizar otros estudios, como el conocimiento del porcentaje de personas que consume calzado ortopédico o especial para el trabajo, lo cual proporciona una perspectiva de los segmentos de producto más rentables en los que se podrían especializar las empresas del sector. También se obtiene información sobre lo que los clientes estarían dispuestos a pagar por un calzado o qué características buscan en él. Pero, lo que realmente cabe destacar de esta encuesta son las

diferentes respuestas de los encuestados en cuanto a su concepto de innovación. En esta pregunta encontramos respuestas de todo tipo, desde algo nuevo, diferente, original hasta algo que mejora los materiales, es cómodo, o no se ha visto antes. A partir de esta respuesta podemos inferir que el concepto de innovación de cada persona es muy diferente, por lo que utilizar encuestas para extraer información sobre la innovación puede provocar resultados subjetivos.