Resultados y conclusiones

Este capítulo tiene por objetivo mostrar el grado de cumplimiento de los objetivos del proyecto, realizar una discusión crítica de los resultados obtenidos y profundizar en el análisis de los resultados. Para este fin, se comparan los resultados obtenidos con las métricas de desempeño previstas en la asignatura de preparación. Además, se contrasta con los resultados de investigaciones previas para tomar perspectiva de las conclusiones que aporta al estado del arte y las metas alcanzadas. Al final, se proponen líneas de investigación a futuro u otras mejoras que pueden mejorar el alcance del proyecto.

Evaluación de los KPIs

Análisis de resultados

Entre los resultados principales, se puede mencionar que he experimentado con modelos básicos de aprendizaje automático, basados en arquitecturas de Naive Bayes, Suppor Vector Machine y Random Forest. Los resultados que se han obtenido mediante los procesos de optimización de parámetros, creación de métricas y visualizaciones de curvas ROC y matrices de confusión cumplen con los objetivos de desempeño establecidos. Aunque se tratan de modelos probabilísticos básicos, su utilización puede ser de utilidad como un punto de partida, para mostrar el estado más básico para las tareas de clasificación. Estas arquitecturas se consideran aptas para la tarea de clasificación en casos supervisados, pues el conjunto de datos contiene datos etiquetados. Dado que están basados en la premisa de selección de características, entonces se ha realizado una personalización de la selección y preparación de las características.

A modo de experimentación con Naive Bayes, se ha realizado una aproximación por Bag of Words, por otro lado, mediante Tfidf, con el objetivo de realizar una experimentación del preprocesado que requiere cualquier modelo de NLP. Mediante CountVectorizer se ha creado un pequeño modelo clasificador con el fin de determinar la importancia de ciertas palabras por su alta frecuencia de aparición en el conjunto de datos. Mientras que con Tfidf se vectorizan las palabras para crear una serie de matrices que contienen la frecuencia del documento y la frecuencia inversa de las palabras. que es pues, un proceso completamente diferente dado que resta importancia a las palabras con mayor frecuencia, de modo que las palabras menos comunes adquieren mayor relevancia. Los resultados de esta experimentación han permitido crear una visualización de nube de palabras completa, es decir, sin criterios de diferenciación de etiquetas. Además de un modelo de clasificación con accuracy de 0,59 y 0,65, recall 0,59 y 0,65, precision de 0,6 y 0,69, F1 de 0,58 y 0,66, tiempo de ejecución 90 y 109 segundos para un modelo entrenado en catalán y otro en español respectivamente.

El modelo SVM con accuracy de 0,6 y 0,65, recall 0,6 y 0,65, precision de 0,61 y 0,68, F1 de 0,6 y 0,66, tiempo de ejecución 2863 y 2375 segundos para un modelo entrenado en catalán y otro en español respectivamente.

El modelo Random Forest con accuracy de 0,6 y 0,6, recall 0,6 y 0,6, precision de 0,68 y 0,60, F1 de 0,61 y 0,60, tiempo de ejecución 311 y 314 segundos para un modelo entrenado en catalán y otro en español respectivamente.

A partir de estos datos se puede determinar que SVM trabaja mejor con el conjunto de datos en español y Random forest con el conjunto de datos en catalán.

Pero este resultado está condicionado a los procesos de captura de texto, etiquetado de datos y preprocesado. Por lo tanto, no pueden ser un resultado final. Por este motivo se acompaña de Curvas ROC y de matrices de confusión.

En el modelo Naive Bayes entrenado en catalán y el entrenado en español presentan un área de la curva del 0,79. En detalle, se ve que el modelo en catalán es propenso a confundir la clase negativo con positivo y tiende ligeramente a confundir la clase neutral con positivo. Mientras que el modelo en español suele confundir la clase positivo con negativo y viceversa.

El modelo SVM entrenado en catalán y el entrenado en español presentan un área de la curva del 0,70 y 0,73 respectivamente. En detalle, se ve que el modelo en catalán es propenso a confundir la clase negativo con positivo y tiende ligeramente a confundir la clase neutral con positivo o negativo. Mientras que el modelo en español suele confundir la clase positivo con negativo y viceversa.

El modelo Random Forest entrenado en catalán y el entrenado en español presentan un área de la curva del 0,65 y 0,62 respectivamente. En detalle, se ve que el modelo en catalán es propenso a confundir ligeramente la clase negativo con positivo y tiende ligeramente a confundir la clase neutral con negativo. Mientras que el modelo en español suele confundir la clase positivo con negativo y viceversa.

En el análisis de los resultados obtenidos en los modelos de aprendizaje profundo se han priorizado el aprendizaje por época con 0,93 y con pérdida de 0,14 en la epoch 599 durante el entrenamiento. En la fase de pruebas 0,74 de accuracy y 0,5 de loss. En la fase de validación 0,75 de accuracy y 0,5 de loss para el caso de BETO.

Mientras que en el modelo RoBERTa-ca, el aprendizaje por época con 0,75 y con pérdida de 0,44 en la epoch 599 durante el entrenamiento. En la fase de pruebas 0,73 de accuracy y 0,5 de loss. En la fase de validación 0,78 de accuracy y 0,46 de loss para el caso de RoBERTa-ca.

desarrollar un modelo avanzado de NLP mediante Transformers o LLM para la entrega de producto final, y finalmente, entregar los resultados según la planificación de la universidad. Los márgenes temporales de dedicación serán el periodo de la 1º fase de desarrollo para alcanzar el primero de los objetivos y el periodo de la 2º fase de desarrollo para el segundo objetivo.

El objetivo secundario optativo, es experimentar con un modelo de LLM para sistemas con bajos recursos computacionales. Este objetivo es opcional, por lo tanto, su finalización dependerá de la consecución del objetivo secundario principal. La motivación es iniciar un aprendizaje práctico sobre LLM, por tanto, en caso de no llegar a ser concluido, figurará como una línea de trabajo a futuro.

Correlación entre resultados y KPIs

Hasta el momento de redacción de este capítulo, los resultados de las notas de las pruebas de evaluación superan el umbral del aprobado, por lo tanto, es un medio adecuado para finalizar el proyecto en el año lectivo. Aunque todavía hay varias pruebas por entregar el objetivo de aprobar la asignatura parece realizable.

A lo largo de los meses de desarrollo y escritura del trabajo final de grado, se han dedicado una media de dos horas diarias a la realización de diversas tareas que componen, tanto la redacción de la memoria y los entregables, como para el desarrollo del producto final, que lo compone la interfaz de la aplicación, los modelos y las actividades secundarias de investigación y la formación adicional. Las tareas de desarrollo en ocasiones han conllevado un mayor tiempo de dedicación, por lo cual, la redacción de la memoria en ocasiones ha quedado postergada hasta la finalización del desarrollo de las arquitecturas y modelos, dado que una parte depende en medida de la otra.

En el apartado técnico, se han implementado las medidas de optimización del código que habían sido propuestas con antelación, con la motivación de hacer un código más eficaz, rápido y robusto. Para esto he aplicado prácticas de optimización de código para reducir el tiempo de ejecución de cada modelo y en las funciones que forman parte de cada fichero. De hecho, la preocupación por el rendimiento del despliegue de los modelos en la aplicación ha conllevado realizar varios procesos iterativos, con el fin de optimizar el tiempo de carga de los modelos, para esto he probado con diversas bibliotecas como pickle, joblib y torch, sin embargo, la decisión final se ha podido demorar debido a los problemas asociados a la compatibilidad entre bibliotecas y formatos de ficheros con flujos de desarrollo. Esto no quiere decir, que la implementación sea la más adecuada, dado que pueden existir otras metodologías que permitan obtener un mejor resultado con los modelos de Deep Learning.

Además, se ha realizado una inspección detallada del funcionamiento de las funciones que aparecen en todos los scripts y notebooks del proyecto, mediante procesos de debug y la creación de tests unitarios para aportar una mayor seguridad de Porcentaje de código documentado para cada función por cada script. Excepto, para las clases de objetos que conforman las instancias de los modelos BERT, los cuales han sido probados mediante fases iterativas de pruebas y error.

•

El tiempo dedicado a cada entregable ha ido en incremento desde el inicio de la asignatura, en algunos casos se ha tratado de dedicar una hora a cada actividad, pero en muchas ocasiones esto no ha podido ser posible debido a otras fuerzas ajenas al proyecto. Principalmente, se han tenido en cuenta los tiempos y plazos de entrega dispuestos por la universidad, por ello, se ha cumplido uno de los objetivos principales del proyecto donde se puede destacar el cumplimiento del 100% de los plazos de entrega.

No obstante, la planificación previa ha servido de guía hasta los dos últimos meses de finalizar el proyecto, ya que los factores externos han influido en los plazos temporales asignados a cada actividad. En ocasiones, las tareas no han sido finalizadas según lo previsto, ya sea por falta de experiencia o por falta de conocimiento. A partir de mayo, hay un desfase temporal de una semana debido a la falta de planificación previa, cuya influencia a tenido consecuencias en el tiempo de recuperación frente a fallos que han ido apareciendo en las semanas previas a la entrega del producto final.

Conclusiones de la investigación

La investigación mediante programación del lenguaje natural ha permitido determinar tres aspectos importantes. El primero, la dificultad para encontrar información de calidad almacenada en conjuntos de datos abiertos, pues la disponibilidad de estos puede depender de decisiones ajenas de empresas que descatalogan los conjuntos de datos de forma periódica, debido a la aparición de la nueva legislación de Inteligencia artificial y el reglamento de protección de datos. Los repositorios de datos actúan como mediadores entre el desarrollador y los proveedores de la información, por lo tanto, pueden quedar obsoletos o retirados durante el transcurso de una investigación. Esta realidad ha impactado de lleno en la planificación de este proyecto, dado que la temática ha cambiado por completo. Sin embargo, esto ha permitido investigar más sobre el vocabulario que se ha empleado en comentarios con sesgo de género, para crear un criterio de filtrado que ha permitido eliminar comentarios y opiniones que afectan a la percepción de los individuos según su género. En cambio, en el conjunto de datos de la independencia de Cataluña, se ha constatado el elevado número de comentarios de odio y rechazo a ciertos grupos sociales, debido a sesgos cognitivos de los usuarios de la red, quienes, en mayor o menor medida, dejan su particular forma de percibir el contexto de esta situación mediante los mensajes que escriben.

La segunda, la carga de trabajo que puede generar crear clases de cargas de datos e instanciar modelos para actividades de clasificación de sentimientos. Dado que, los principales artículos están orientados a resolver casos basados en el inglés- Y aunque puede haber numerosos proyectos en Internet, se ha optado por hacer una breve investigación por los sitios web de relevancia, con el fin de llegar a determinar el funcionamiento de la arquitectura, los avances del ámbito de aplicación y otras fuentes de información. La arquitectura de este tipo de Transformers es la más sencilla, ya que sólo está compuesta por un encoder que ha sido optimizado en cada caso para reconocer palabras del español y el catalán debido a un corpus de entrenamiento acorde al idioma, que genera un vocabulario extenso, pero que en algunas ocasiones genera malentendidos con el sistema y esto puede ser la causa de las alucinaciones que padecen.

En tercer lugar, la conclusión más importante es aquella que surge tras la meditación sobre los métodos de captura de datos y el etiquetado de clases por parte de expertos. El caso que atañe es que las redes sociales son una fuente de opiniones, ideas y conversaciones que varían con el tiempo, por lo tanto, los datos recabados en un periodo de tiempo no se pueden aplicar en otras ocasiones, aunque la temática sea la misma. De cara a obtener datos de calidad, considero que las redes sociales no son precisamente las adecuadas, dado que son fuentes de desinformación, mentiras y falacias en las mejores de las ocasiones. En algunas instancias del conjunto de datos, se observan comentarios que no están relacionados con el tema en cuestión, en otros casos, las opiniones están sesgadas por ideología, política o creencias de la tradición, por lo tanto, disminuye la calidad del contenido. Además, el aspecto de etiquetado de datos puede tratarse de una actividad con doble filo, dado que una clasificación errónea de los datos desde el punto de partida puede conllevar a crear un sesgo en el modelo, aunque la distribución de las clases haya sido balanceada. Esto quiere decir que, en problemas de clasificación supervisada, no sólo es importante la objetividad del creador del modelo, sino del equipo que ha realizado la tarea de asignación de grupos a cada uno de los comentarios que forman parte del corpus.

Logros del proyecto

Objetivos alcanzados

En el resultado final del proyecto, se confirma que se ha alcanzado el objetivo principal del proyecto, que ha sido diseñado para experimentar con diversas arquitecturas de aprendizaje automático.

En la primera etapa de experimentación se han utilizado modelos de machine learning, para realizar una primera clasificación de sentimientos mediante arquitecturas de Naive Bayes, SVM y Random Forest. Para todos ellos se ha elegido un método basado en TFiDFvectorizer, aunque como primera tarea he asumido un sencillo análisis de la frecuencia de las palabras mediante un BagOfWords, para determinar el vocabulario asociado a perspectiva de género. La técnica de preprocesado ha sido la lemmatización, este método permite conservar cada palabra sin prescindir de prefijos y sufijos, pero tiene un tiempo de ejecución más largo.

Con esta perspectiva de género se ha implementado una segunda fase de limpieza y preprocesado de un conjunto de datos sobre la independencia de Cataluña. En esta fase, de implementación del procesamiento del lenguaje natural se ha diversificado dos líneas de acción, una enfocada en el idioma español y la otra en la lengua catalana. En ambos casos, se hace uso de las mismas etapas de preprocesamiento, salvo aquella que está enfocada en eliminar las palabras vacías, pues tienen que ser dos objetos específicos para cada caso.

Además, se ha precisado de tuberías de optimización y automatización de los procesos de evaluación de modelos, hasta conseguir determinar los parámetros más adecuados para cada caso, con la finalidad de maximizar la accuracy de cada uno de ellos.

En la fase de modelado mediante Deep learning se ha realizado dos líneas de acción, una para cada idioma, con el fin de entrenar dos modelos BERT preentrenados, llamados BETO y RoBERTa. En ambos casos he utilizado una clase para cargar los datos y otra clase para instanciar el modelo. Esta fase ha sido la que más carga de proyecto a tomado, dado que no se ha usado proyectos de terceros, sino que ha sido implementado siguiendo un proceso de aprendizaje, lectura e investigación del estado del arte en modelos de clasificación con otros fines. Durante los entrenamientos de los modelos he apreciado que el conjunto de datos original no era demasiado grande, por lo tanto, el tiempo de ejecución de estos ha sido reducido. Los resultados obtenidos en las métricas han superado significativamente los valores de la accuracy de los modelos de machine learning. Sin embargo, algo que comparten todos los modelos es su capacidad para alucinar cuando clasifican de forma errónea cada uno de los tweets. Para mayor énfasis, se hizo uso de la evaluación de los modelos mediante técnicas de benchmarking, haciendo uso de la librería llamada Tensorboard y la introducción de la librería de tensorflow llamada Hparams, de modo que, se puede realizar un seguimiento y conservar la trazabilidad de cada combinación de hiperparámetros y las métricas obtenidas en cada ejecución.

Los resultados de los modelos BERT se alcanzaron con varios días de demora, en principio, porque era la primera vez que realizaba una tarea similar, la segunda por la falta de previsión de los días festivos de abril y mayo. Esta incidencia ha supuesto una reducción del tiempo destinado a la creación de un RAG (Retrieval Augmented Generation) , de modo que, he optado por crear un Retrieval que permite la recuperación de tweets a partir del contexto de las frases que lo componen, es decir, a diferencia de los modelos BERT que captan la relación semántica de las palabras y permiten agrupar grupos de palabras por su significado, en el caso del Retrieval, se implementa sentence transformers que captan el contexto de las frases de un tweet, por ello, he eliminado los signos de puntuación, para que cada tweet tenga un significado en sí mismo y permita un mejor procesamiento del lenguaje. Con el fin de saber encontrar tweets semejantes, mediante similitud coseno y distancia.

Entonces, a partir de la experimentación con el Recuperador y bases de datos vectoriales, como ChromaDB, es posible realizar una perspectiva crítica del conjunto de datos empleado para el entrenamiento, y prueba de los modelos. En algunos casos, he detectado que los sistemas de inteligencia artificial no concuerdan con las etiquetas que han creado los expertos. Este hecho, me ha hecho reflexionar sobre la importancia que tiene la fase de captura del dato, más incluso que la fase del modelado, pues en gran medida, el éxito de un proyecto de ciencia de datos viene determinado por el tipo de aprendizaje y la calidad de los datos, aquello que en los entornos empresariales denominan el Gobierno del Dato. Sin embargo, la motivación inicial por realizar un proyecto con impacto social es suficiente para no denostar el ámbito de estudio, pues es un tema social de gran impacto para el funcionamiento de la democracia, que es el motor de cambio de las sociedades democráticas en Europa desde tiempos de los antiguos griegos. La soberanía corresponde al pueblo español, es decir, a todos los ciudadanos españoles, quienes votan para elegir a los representantes del poder ejecutivo. En las pasadas y presentes elecciones de Cataluña se establecen los representantes y las líneas de acción que se ejecutarán en los próximos cuatro años, por lo tanto, la relevancia que las opiniones de los votantes puedan tener en periodos previos a unas elecciones son de sumo interés para periodistas, analistas políticos e incluso agentes del poder ejecutivo que participan directa o indirectamente en las candidaturas.

Entre los resultados del proyecto quedan descartados entonces la creación de tres entornos para el despliegue del modelo en fases de producción. Es decir, toda la metodología MLOPS quedó descartada, por varios motivos, entre ellos la falta de experiencia, la falta de un presupuesto asignado a estas actividades. Además, quedaron descartados los objetivos relacionados con la creación de tuberías para el despliegue de los modelos y la implementación de una integración y despliegue continuos (CI/CD).

Meta de la investigación

La finalidad de la investigación ha sido desarrollar un sistema inteligente capaz de clasificar las emociones a partir del contexto de un texto, con mayor o menor precisión, que los sistemas actuales. De todas las facultades humanas, la capacidad de experimentar sentimientos, comprenderlos y saber diferenciarlos, es aquella en las cuales los ordenadores requieren menor confiabilidad, pues es una actitud humana más propia de la consciencia del ser, que la de la inteligencia.

No obstante, al empezar el proyecto no tenía ninguna simpatía hacia delegar estas tareas a los ordenadores, pues es muy probable que no desempeñen estas tareas con eficacia. Este prejuicio está fundamentado en mi experiencia laboral, donde el Customer Success Factor se mide mediante Mistery Customer, en vez de sistemas de clasificación de botones de caras expresando el sentimiento. Esto quiere decir, que en la actualidad las empresas prefieren delegar estas tareas en humanos, debido a su alta inteligencia emocional, ya que los sistemas de NLP comerciales destinados a estas tareas necesitaban continuas mejoras para adaptar los modelos a los nuevos contextos.

El reto de este proyecto ha sido asumir el rol del profesor, el estudiante y el aprendiz que experimenta de forma autónoma, sin llegar a seguir las pautas de una prueba de evaluación estándar. Este proceso de aprendizaje basado en la experiencia práctica pone de manifiesto el carácter pragmático de la ciencia de datos, donde la prueba y el error junto con el contraste de hipótesis basadas en juicios a priori puede llegar a ser un complemento al método científico inaugurado por René Descartes. Que configura el proceso dialéctico de la ciencia de datos, confirma su rápida evolución y la explosión de conocimiento científico, que se está desarrollando a partir del desconocimiento de las redes neuronales profundas y el comportamiento de sus intrincadas arquitecturas.

Al finalizar esta investigación he comprendido, que el estado del arte ha avanzado mucho desde la aparición de las arquitecturas BERT, por lo que se puede llegar a crear sistemas confiables basados en Deep Learning, siempre que se priorice la comprensión del contexto de las frases, en vez de sistemas de clasificación basados en arquitecturas de solo un encoder. Para el caso, la optimización que ofrecen bases de datos vectoriales o de conocimiento basado en grafos.

Retos y limitaciones del proyecto

La clasificación de sentimientos en redes sociales utilizando técnicas de aprendizaje automático (ML), aprendizaje profundo (DL) y modelos de lenguaje grande (LLM) es un área de investigación desafiante y en rápida evolución. Si bien estos enfoques han demostrado un gran potencial para analizar y comprender las opiniones y emociones expresadas en el contenido generado por los usuarios, también conllevan varios retos y limitaciones.

A lo largo de este proyecto, los desafíos encontrados han sido relativamente menores en comparación con las limitaciones enfrentadas. Uno de los principales obstáculos ha sido la falta de tiempo y recursos computacionales adecuados. El entrenamiento de modelos de DL y LLM a menudo requiere una gran cantidad de datos, así como hardware especializado como GPUs o TPUs para un entrenamiento eficiente [3][8]. Debido a las restricciones de tiempo y la falta de acceso a equipos informáticos de alto rendimiento, el alcance del proyecto tuvo que ser reconsiderado.

Además, los costos asociados con el despliegue y mantenimiento de modelos de ML en entornos de producción pueden ser significativos. Factores como el almacenamiento y procesamiento de datos, la monitorización continua del rendimiento del modelo, y la necesidad de reentrenamiento periódico contribuyen a la complejidad y los gastos generales. Teniendo en cuenta estas consideraciones, se tomó la decisión de priorizar el desarrollo y evaluación de los modelos en un entorno de investigación, posponiendo su implementación en sistemas de producción a gran escala.

A pesar de estos desafíos, el proyecto ha logrado avances prometedores en la clasificación de sentimientos en redes sociales. Mediante el aprovechamiento de técnicas de ML, DL y LLM, se han desarrollado modelos capaces de capturar los matices y la complejidad del lenguaje natural. Los resultados experimentales han demostrado mejoras en la precisión y la capacidad de generalización en comparación con los enfoques tradicionales.

Sin embargo, es importante reconocer que aún existen limitaciones inherentes a estas tecnologías. Los modelos de lenguaje a menudo luchan con el sarcasmo, la ironía y otros aspectos sutiles del lenguaje humano. Además, los sesgos presentes en los datos de entrenamiento pueden propagarse a las predicciones del modelo, lo que lleva a resultados injustos o discriminatorios. Abordar estos problemas requerirá esfuerzos continuos en áreas como la recopilación de datos diversos, técnicas de desaprendizaje y el desarrollo de modelos más interpretables.

Análisis crítico de la planificación

La planificación es un aspecto importante en cualquier proyecto, ya que establece una hoja de ruta para alcanzar los objetivos deseados dentro de un marco de tiempo específico. Sin embargo, a medida que el proyecto avanza, es común encontrar desviaciones y cambios que requieren ajustes en la planificación original. En este subcapítulo, analizaremos críticamente la planificación del proyecto, teniendo en cuenta las actividades no previstas y los cambios realizados.

Tras un periodo de reflexión crítica de la planificación considero que, si bien se estableció una hoja de ruta inicial sólida, el proyecto experimentó varias desviaciones y cambios a lo largo de su desarrollo. La capacidad de adaptarse a estos cambios y tomar acciones de mitigación adecuadas fue crucial para mantener el proyecto en marcha y alcanzar los objetivos deseados. Sin embargo, también es importante reconocer las áreas que requieren mejoras, como la estimación del tiempo necesario para ciertas tareas y la identificación temprana de posibles desviaciones. Al aprender de estas experiencias y ajustar continuamente la planificación, se pueden minimizar los impactos negativos y garantizar el éxito general del proyecto.

Análisis crítico de la metodología

La metodología empleada en este proyecto ha sido un proceso iterativo y adaptativo, que ha evolucionado a medida que se han presentado desafíos y oportunidades. A lo largo del desarrollo, se han realizado ajustes significativos en la planificación y el enfoque, con el objetivo de cumplir con los requisitos del proyecto y superar las limitaciones encontradas.

Durante la primera semana de la fase 1, se logró un avance notable en la limpieza de varios conjuntos de datos. Este paso fue crucial, ya que permitió al modelo etiquetar el sesgo de género y la emoción implícita en textos en español, una desviación del plan original que se abordará más adelante. Además, se iniciaron actividades de creación de pruebas para garantizar la eficiencia del código y se realizaron optimizaciones para mejorar su tiempo de ejecución. Estas tareas sentaron una base sólida para las etapas posteriores del proyecto.

En la segunda semana, se llevó a cabo una clasificación de sentimientos y sesgo de género utilizando los algoritmos Naive Bayes y Regresión Lineal. Además, se desarrolló una primera versión de la interfaz con Streamlit, donde se visualizarán los resultados de la comparativa entre modelos de procesamiento del lenguaje natural (PLN). También se realizaron tareas de preprocesamiento de datos para preparar los conjuntos de datos para nuevos modelos de aprendizaje automático. Estas actividades demostraron un progreso significativo hacia los objetivos del proyecto.

Durante la tercera semana, se inició la modelización de clasificadores de texto utilizando SVM, Árboles de Decisión, Random Forest y otras variaciones de Naive Bayes. El objetivo era preparar las métricas para la posterior evaluación y comparación de resultados mediante curvas ROC y matrices de positivos. En esta etapa, se realizaron varias iteraciones sobre la fase de preprocesamiento de datos para adaptar los conjuntos de datos a cada arquitectura, y se llevó a cabo una optimización de parámetros para cada instancia. Este enfoque iterativo permitió afinar los modelos y mejorar su rendimiento.

En la cuarta semana y los días previos a la entrega, se analizaron las métricas de los modelos y los gráficos de las evaluaciones. Además, se finalizó el desarrollo de la interfaz utilizando la librería Streamlit. Para esto, se realizó un diseño manual de la interfaz, aplicando teorías del color y principios de la Gestalt, con el objetivo de crear una experiencia de usuario adaptada a varios perfiles de usuarios. Este enfoque centrado en el usuario es un aspecto destacado de la metodología empleada.

A lo largo de la prueba de evaluación, se inició la redacción de los dos primeros capítulos de la memoria, realizando revisiones del contenido y asignando citas o menciones a papers e investigaciones de otros expertos. Esta documentación paralela al desarrollo es una buena práctica que contribuye a la solidez y transparencia del proyecto.

Sin embargo, es importante destacar algunos cambios significativos en la metodología. En primer lugar, se modificó en el último momento la estrategia de despliegue del modelo, debido a la falta de consenso para elegir un proveedor en la nube y al deseo de utilizar los recursos de pago asociados a las matrículas de la universidad. Además, se consideró que el proyecto no tenía una finalidad comercial, sino que era una experiencia de aprendizaje autónomo. Si bien este cambio puede haber limitado el alcance del proyecto, demuestra una consideración realista de los recursos disponibles y los objetivos del proyecto.

En segundo lugar, se realizó una reacomodación de la temática del trabajo para ajustarse a los requisitos de impacto social de la UOC. Inicialmente, el proyecto no había considerado los objetivos de desarrollo sostenible, el compromiso ético y global, ni el impacto en la diversidad de género. Este cambio se produjo durante la PEC1, cuando el conjunto de datos original orientado al e-commerce quedó descatalogado del repositorio oficial. Aunque este ajuste supuso un desafío, también brindó la oportunidad de alinear el proyecto con principios éticos y de sostenibilidad importantes.

Además, durante el desarrollo del proyecto, se realizó un período de formación sobre modelos de lenguaje grande (LLM) que no estaba previsto inicialmente. Se llevaron a cabo pruebas con diversos servicios como ChatGPT, Ollama de Facebook AI, Gemini Google y Mistral de Perplexity para elegir el proveedor de LLM y su API [3]. Este proceso de aprendizaje y evaluación, aunque no planificado, demuestra la capacidad de adaptación y la voluntad de incorporar tecnologías emergentes relevantes.

En conclusión, la metodología empleada en este proyecto se caracterizó por su flexibilidad y capacidad de adaptación frente a desafíos y oportunidades. A través de un enfoque iterativo, se lograron avances significativos en la limpieza de datos, el desarrollo de modelos, la evaluación de métricas y la creación de una interfaz centrada en el usuario. Sin embargo, también se produjeron cambios importantes, como la modificación de la estrategia de despliegue y la reacomodación de la temática para alinearse con los requisitos de impacto social. Estos ajustes, aunque desafiantes, permitieron al proyecto evolucionar y mantener su relevancia. En general, la metodología demostró ser efectiva para alcanzar los objetivos del proyecto, al tiempo que se adaptaba a las limitaciones y aprovechaba las oportunidades de aprendizaje y mejora.

Desviaciones y cambios implementados

Durante la fase uno del proyecto, se encontraron varias desviaciones y se tomaron acciones de mitigación correspondientes. Una de las principales desviaciones fue la descatalogación del conjunto de datos preseleccionado. Para mitigar este problema, se realizó una búsqueda exhaustiva de fuentes de datos abiertos con temáticas de repercusión social, evaluando cada conjunto de datos según parámetros como el grado de repercusión en el repositorio y la calidad de los datos. Además, se priorizó la búsqueda de datos escritos en catalán o español para adaptarse mejor al contexto del proyecto.

Otra desviación significativa fue la necesidad de optimizar el código y los modelos. Los ciclos iterativos de desarrollo resultaron ser más largos de lo esperado debido a la baja calidad de las métricas y del código inicial. Para abordar este problema, se realizaron varias iteraciones sobre tres arquitecturas diferentes, utilizando librerías como Grid Search, Pipeline y Time para afinar los modelos mediante tuberías de concatenación de funciones. Aunque este proceso llevó más tiempo del previsto, fue esencial para mejorar el rendimiento y la calidad general del proyecto.

Además de las desviaciones mencionadas, se agregó una tarea adicional no prevista inicialmente: la captura de datos de redes sociales como X, donde los usuarios expresan opiniones y sentimientos sobre las próximas elecciones catalanas. Esta tarea se consideró de gran utilidad para actualizar los tweets de los usuarios y obtener un mayor volumen de datos para la programación de un modelo más grande basado en Transformers y Embeddings. La inclusión de esta tarea en la planificación demuestra la capacidad de adaptación del proyecto a las necesidades emergentes y la importancia de mantenerse al tanto de los eventos actuales relevantes.

Sin embargo, también se detectaron algunas deficiencias en la interfaz de usuario y en ciertas funciones que no cumplían con las expectativas de los tests unitarios. Estas deficiencias han requerido atención adicional y posiblemente un ajuste en la planificación para asegurar que se aborden adecuadamente antes de la entrega final del proyecto.

En la segunda fase de desarrollo, se retrasó el análisis de resultados y comparativas, lo que tuvo un impacto en la planificación general. Este retraso de 12 días resalta la importancia de la gestión del tiempo y la necesidad de tener en cuenta posibles contratiempos al establecer los plazos del proyecto.

Tras un periodo de reflexión crítica de la planificación considero que, si bien se estableció una hoja de ruta inicial sólida, el proyecto experimentó varias desviaciones y cambios a lo largo de su desarrollo. La capacidad de adaptarse a estos cambios y tomar acciones de mitigación adecuadas fue crucial para mantener el proyecto en marcha y alcanzar los objetivos deseados. Sin embargo, también es importante reconocer las áreas que requieren mejoras, como la estimación del tiempo necesario para ciertas tareas y la identificación temprana de posibles desviaciones. Al aprender de estas experiencias y ajustar continuamente la planificación, se pueden minimizar los impactos negativos y garantizar el éxito general del proyecto.