Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos

Prediciendo la personalidad mediante el Procesamiento del Lenguaje Natural.

Trabajo Fin de Máster

Tipo de trabajo: Piloto Experimental

Presentado por: del Campo Peinado, Álvaro

Director/a: Blanco Valencia, Xiomara Patricia

Resumen

La personalidad es un concepto el cual podría conocerse mejor si se analizara desde la perspectiva de las nuevas tecnologías como el procesamiento del lenguaje natural.

En el presente trabajo se expondrá el diseño, el entrenamiento y la validación de un modelo capaz de predecir la personalidad de un sujeto a partir de unas distintas interacciones publicadas en la red social Twitter en forma de texto. El modelo se basó la etapa de diseño en el indicador de la personalidad publicado por Myers-Briggs en 1944 y se elaboró a partir de un conjunto de datos de 1794016 instancias en el cual se recogen publicaciones de diferentes usuarios, así como su tipo de personalidad según este indicador.

En la fase de resultados se evaluó cada uno de los modelos de manera individual. Con los resultados se concluyó que los modelos construidos de manera independiente para cada dimensión de la personalidad tienen un mejor rendimiento que los que pretenden diferenciar la personalidad del individuo completamente definida.

Para ello fue necesario el uso de diferentes algoritmos y diferentes técnicas del procesado de texto que iluminaron la manera en que se obtienen los mejores resultados a la hora de abordar este problema.

Palabras Clave: Personalidad, NLP, Myers-Briggs, Machine Learning, Data Science.

Abstract

Personality is a concept which could be better understood if it were analyzed from the perspective of new technologies such as natural language processing.

In the present work, it is explained the phases of designing, training, and validating of a model capable of predicting the personality of a subject from an input based on text written in different interactions published in the social network Twitter. The model based its design on the personality indicator published by Myers-Briggs in 1944 and was made using a data set of 1794016 instances in which publications from different users are collected, as well as their personality type according to this indicator.

In the results phase, it was possible to appreciate how the models built independently for each dimension of the personality had a better performance than those that seek to differentiate a completely defined individual's personality.

To prove this, different algorithms and text processing techniques were used and so illuminated the way in which the best results are obtained when addressing this problem.

Keywords: Personality, NLP, Myers-Briggs, Machine Learning, Data Science.

Índice de contenidos

Índice de tablas

Índice de figuras

1. Introducción

La personalidad y el Indicador Myers-Briggs

La personalidad es un concepto abstracto el cual puede ser estudiado desde diversos enfoques. Actualmente la Real Academia Española (RAE) ofrece definiciones para la personalidad como “diferencia individual que constituye a cada persona y la distingue de otra” o “conjunto de características o cualidades originales que destacan en algunas personas”. Al escuchar estas definiciones podríamos plantearnos preguntas como ¿Cuáles son las características que definen las personalidades? o ¿Qué cualidades albergan las personalidades?

A lo largo de la historia, diferentes autores han tratado de contestar estas preguntas a través de distintitos modelos y teorías relacionadas con la personalidad. Estas teorías variaban en función de las dimensiones personales que se estudiaban a la hora de definir la personalidad de un sujeto. Sobre esto se fundamenta una gran diversidad de modelos de la personalidad y diferentes teorías que intentan acercar de manera cuantitativa, qué aspectos conforman la personalidad y en qué influyen estas dentro del marco de la psicología diferencial.

Algunas de las principales teorías formuladas han sido, entre otras, la teoría de la personalidad basada en la psique y la confrontación de los impulsos y pulsiones con la realidad elaborada por Freud (1923), en la cual, la personalidad va evolucionando en diferentes etapas debido a los conflictos generados en cada una de ellas.

Con un enfoque basado en la experiencia individual nace la teoría de los constructos personales de Kelly (1955). En dicha teoría el autor remarca que la personalidad se constituye como un sistema jerarquizado de constructos personales dicotómicos. Mientras que estas teorías se han considerado obsoletas y han sido rechazadas, otras tienen un apoyo creciente, una de ellas es la teoría del Big Five de Costa y McCrae (1996), la cual, aún siendo presentada por primera vez en 1933 no ha tenido gran apoyo hasta hace poco tiempo.

Esta teoría de la personalidad agrupa las personalidades con base a las cinco grandes dimensiones de la personalidad: apertura a la experiencia, conciencia, extraversión, amabilidad y neuroticismo.

Actualmente, debido a este resurgimiento de las teorías de la personalidad basadas en modelos multidimensionales nos encontramos en un momento de máximo interés en la teoría de la personalidad de Carl Gustav Jung *(1921)*. Carl Gustav Jung propone en su obra

Psychologische Typen (1921) que las personalidades son configuradas por dos grandes partes, una visible que sirve para adaptarse al medio y otra oculta, en la que se encuentran todas las conductas que no son admisibles por el sujeto.

Carl Jung defiende que las personalidades se conforman mediante el énfasis que el sujeto hace de las cuatro funciones psicológicas básicas: sentir, intuir, percibir o pensar. Estas funciones son modificadas por la actitud introvertida o extravertida que tenga el sujeto.

Con esto Carl Gustav Jung distingue, en su obra *Psychologische Typen (1921)* ocho tipos de personalidad psicológicos que se construyen por la combinación de las cuatro funciones psicológicas básicas y dos actitudes personales:

- **Introvertido-Pensativo:** Se incluyen personas con énfasis en comprender su propio ser. Este tipo de personalidades se corresponde con personas con una gran actividad intelectual, los cuales pueden experimentar problemas al relacionarse o interactuar con otras personas. Este tipo de personas se mantienen tenaces y muestran resiliencia ante la consecución de nuevos retos y objetivos.

- **Extravertido-Pensativo:** Las personas que se agrupan dentro de este tipo de personalidad están interesados por comprender la realidad. La mentalidad de este tipo de personalidad está orientada a algún objetivo y las acciones que ponen en marcha tienen una razón bien definida, pudiéndose considerar algunos casos como maquiavélicos o tiranos con otras personas. Este tipo de personas únicamente acepta como verdades, las cosas que se pueden confirmar con suficiente evidencia.

- **Introvertido-Sentimental:** Las personas Introvertido-Sentimentales son inaccesibles, aunque muestran impresión de autonomía y armonía. Este grupo de personalidades muestra una alta solidaridad hacia otras personas, sin embargo, les es difícil establecer relaciones sociales con otras personas, esto las hace muy atentas hacia las necesidades del resto, pero hacen todo lo posible por pasar inadvertidas y permanecer fuera del foco de la atención.

- **Extravertido-Sentimental:** Convencionales y buscadores del éxito. Estas personas tienen muchas capacidades y facilidades a la hora de comprender a otras personas, así como a la hora de empatizar o crear nuevas relaciones sociales. Estas personas tienen tendencia a ser buenos comunicadores, sin embargo, se les dificulta separarse del resto y poner la atención en actividades introspectivas. Del mismo modo, a este tipo de personalidades les cuesta lidiar con la falta de éxito, llegando a sufrir por no

tener la suficiente atención o al no conseguir los logros que se proponen cuando se trata del estatus social.

- **Introvertido-Sensitivo:** Inmersos en sus sensaciones internas. Las personas que se clasifican dentro de este grupo de la personalidad tienen una gran creatividad y habitualmente se identifican como artistas o músicos. Consideran muy valioso todo lo tangible a través de los sentidos y, por ello, dan importancia primordial a la experiencia.

- **Extravertido-Sensitivo:** Con interés en estudiar los fenómenos externos. Son personas curiosas en cuanto a todo lo que les rodea, buscando el funcionamiento de todo lo material. Este tipo de personas pueden tener tendencias hedonistas y anteponer a otros objetivos la búsqueda del placer.

- **Introvertido-Intuitivo:** Este grupo se caracteriza por ser personas comprometidas con sus ideas internas y de carácter soñador. Son Imaginativas y viven dentro de sus propias ideas, haciéndoseles a veces complicado poner los pies en la tierra o ser realistas.

- **Extravertido-Intuitivo:** Capaces de adelantarse a las nuevas tendencias. Las personas con una personalidad extravertida-intuitiva tienen tendencia a ser abiertos a nuevas experiencias o buscar nuevas aventuras. Les gusta adentrarse en nuevos retos y enfrentarse a nuevos estímulos. Al igual que otras personalidades, son obstinados en cuanto a cumplir sus objetivos y, una vez lo realizan, se marcan un nuevo objetivo que dé dirección a su vida olvidando rápidamente el anterior. No suelen prestar demasiada atención a las personas a su alrededor.

Valiéndose de la teoría de la personalidad de Carl Jung (1921), Katharine Cook Briggs y su hija Isabel Briggs Myers confeccionaron el Inventario Tipológico de Myers-Briggs, -MBTI (1944). Este inventario recoge gran parte de las ideas expuestas anteriormente y se formula como una nueva herramienta psicológica que consiste en un cuestionario auto-reportado, el cual, evalúa la percepción del entorno del sujeto y su sistema de toma de decisiones con el fin de proporcionar un modelo de personalidad acorde a su modo de interactuar con la realidad y de percibir lo que les rodea.

La prueba utiliza cuatro escalas dicotómicas que se construyen basándose en las dos actitudes (Extravertida e Introvertida) y las cuatro funciones psicológicas establecidas por Carl Jung (Sentir, Pensar, Intuir y Percibir).

Tabla 1. - Dicotomías de la Personalidad Evaluadas por el Indicador Myers-Briggs.

Dimensión ¿Qué Evaluar?	Dicotomías	
Modo de enfocar la atención y obtener energía.	(E) Extrovertido	(I) Introvertido
Modo de percibir y tomar la información	(S) Sensitivo	(N) Intuitivo
Mecanismos para la toma de decisiones	(T) Pensativo	(F) Emocional
Orientación con respecto al mundo exterior	(J) Juez / Calificador	(P) Perceptivo

Elaboración Propia

Con este criterio encontramos las cuatro categorías que conforman una personalidad según el inventario tipológico de Myers-Briggs (1944), a saber, Extroversión-Introversión, Sensorial-Intuitivo, Pensamiento-Emocional y Calificador-Perceptivo.

Cada una de estas categorías es representada por una letra de la manera en que se muestra en la Tabla 1, por lo que el tipo de personalidad del sujeto será una combinación de cuatro letras cada una correspondiente a una de las cuatro dicotomías Extravertido-Introvertido, Sensitivo-Intuitivo, Pensativo-Emocional y Juez-Perceptivo. Con esto se crea un espectro de 16 tipos posibles de personalidad, los cuales son generados mediante las combinaciones de pares opuestos de esas cuatro dimensiones.

Para el entendimiento de futuras partes de este trabajo es importante clarificar la manera en que se codifican las diferentes personalidades según el indicador Myers-Briggs (1944) con un ejemplo. Como se ha dicho, la personalidad de un individuo se genera mediante la combinación de las cuatro dimensiones en forma de palabra, dando lugar a uno de los 16 posibles tratamientos de la personalidad, por lo que si los resultados del indicador para el individuo son Extravertido (E), Intuitivo (N), Pensativo (T) y Perceptivo (P) su personalidad quedará codificada como la combinación de estas cuatro letras, en este caso (ENTP). Como último ejemplo, en caso de que la personalidad del individuo fuera Introvertida (I), Sensitiva (S), Emocional (F) y Juez (J), la personalidad de ese individuo quedaría codificada como ISFJ.

Figura 1. – Los 16 tipos de Personalidades Diferentes por el Indicador Myers-Briggs.

Human Development Solutions (2016)

Aunque el indicador Myers-Briggs cuenta con una precisión del 75% según indica su propio manual y no fue aceptado en su momento como una prueba de personalidad fiable (Messick, 1985), este fue ganando popularidad a partir de 1957 como herramienta destinada a la rama de la psicología laboral, teniendo buena acogida con el fin de medir el desempeño laboral desde un punto de vista psicológico. Algunos estudios han mostrado la utilidad del cuestionario, especialmente cuando se relaciona con aspectos laborales en los que influyen la diferenciación personal (Adrian Furnham et al., 1993).

Este inventario ha mostrado un buen desempeño a la hora de predecir aptitudes ligadas con el mundo laboral. Algunos ejemplos son el liderazgo, las capacidades comunicativas o las aptitudes analíticas de una persona (McCaulley et al., 1990). Con todo esto, ha mejorado la percepción con respecto al indicador Myers-Briggs, siendo cada vez más aceptado por diferentes colectivos científicos. Debido a todo lo anterior, este inventario se ha convertido en

una de las herramientas más utilizadas para estimar la personalidad de un sujeto de manera rápida y sencilla, parte de ello gracias a la herramienta web *16Personalities*.

Esta herramienta online permite realizar una autoevaluación de la personalidad mediante la adaptación del inventario y su futura digitalización. La adaptación del indicador Myers-Briggs consiste en un cuestionario de 60 ítems de respuesta cerrada múltiple en el cual se presentan diferentes afirmaciones y situaciones sobre las que el usuario debe contestar en una escala del 1 al 7, siendo 7 estoy de acuerdo y 1 no estoy de acuerdo. Del mismo modo, este sitio web integra módulos de análisis y evaluación de las capacidades individuales, realizando un interesante acercamiento del estudio de la personalidad y las diferencias individuales a través de las nuevas tecnologías.

La puesta en marcha de esta herramienta ha dado lugar a que actualmente se cuente con multitud de registros relacionados con diferentes personalidades. Esto ha dado lugar a un interés emergente en realizar modelos de aprendizaje que puedan ser entrenados para predecir personalidades mediante este alto volumen de registros e información adicional de los sujetos involucrados en las pruebas.

1.1 Justificación

Por todo lo planteado anteriormente se entiende la dificultad para cuantificar y definir la personalidad de un sujeto, ya que, el tiempo de aplicación de las herramientas psicológicas y baterías de tests actuales de cara a localizar e identificar la personalidad de un individuo es muy elevado. Algunas de estas dificultades se deben a la complejidad del concepto de personalidad, pero muchas otras son causa de que, en su gran mayoría, las herramientas de las que se disponen por el momento se basan en costosas baterías de tests las cuales, además de tener un alto tiempo de aplicación, es complicado encontrarlas digitalizadas.

1.2 Planteamiento del trabajo

Entendiendo esos conceptos básicos y las dificultades que se plantean, el objetivo del presente trabajo es desarrollar un modelo machine learning capaz de inferir la personalidad de un sujeto a través de pequeños fragmentos de textos escritos por ese mismo usuario. Para ello se trabajará con diferentes modelos de machine learning, entrenados a partir de un dataset de más de un millón de instancias. Con este pretexto, se espera encontrar el mejor modelo posible de machine learning, el cual, en pocos segundos, sea capaz de predecir la personalidad de un individuo.

Este acercamiento daría lugar a una reducción de costes a la hora de definir la personalidad de un sujeto y con ello poder darle una utilidad futura a esta información. Del mismo modo, el hecho de que se pudiera reducir el tiempo en que se realiza una evaluación de la personalidad podría ayudar a entender mejor este concepto y dar lugar a nuevos modelos teóricos de la personalidad en un futuro.

Con respecto a investigaciones pasadas, se espera encontrar un modelo que dé pie a predicciones con mejores métricas que los planteados en otros estudios.

1.3 Estructura de la memoria

En el presente trabajo se describirá el proceso de investigación y profundización en el área de estudio de la personalidad y las técnicas de machine learning a través del procesamiento del lenguaje natural. En los posteriores puntos se describirán las fases de desarrollo, entrenamiento y validación de un modelo de machine learning preparado para aproximar la personalidad de un sujeto en base a las dimensiones establecidas por el indicador Myers-Briggs, a saber, Extravertido-Introvertido, Sensitivo-Intuitivo, Pensativo-Emocional y Juez-Perceptivo.

En primer lugar, estudiaremos las investigaciones recientes de mayor relevancia relacionadas con la elaboración de modelos de machine learning preparados para aproximar las dimensiones de la personalidad de un sujeto a partir de textos escritos por el mismo individuo. De esta manera veremos el emergente interés que rodea a predecir la personalidad a través de técnicas de procesado del lenguaje natural gracias al avance de las nuevas tecnologías.

También estudiaremos algunas investigaciones que, sin referirse directamente a la personalidad, entran en el campo de estudio de la psicología a través de técnicas de predicción y del procesamiento del lenguaje natural. Un ejemplo interesante es el del auge que ha vivido el análisis de sentimientos en diferentes campos de aplicación, especialmente en el corporativo y en el análisis de satisfacción.

A continuación, se expondrá la metodología a través de la cual se ha realizado la elaboración del presente modelo de predicción. La metodología seguirá los pasos comunes para realizar un modelo de machine learning que, aunque serán detallados más adelante, listamos a continuación: Estudio del conjunto de datos, pre-procesamiento de las clases del conjunto de datos, pre-procesamiento del texto de lenguaje natural del conjunto de datos, análisis exploratorio y diseño del modelo.

Por último, detallaremos cada uno de los pasos descritos en la metodología, se mostrarán las diferentes fases llevadas a cabo en el análisis exploratorio de cara a enfocar y dar forma al modelo y, en la fase de resultados se validarán las métricas del modelo elaborado en este trabajo. Estos resultados serán contrastados con investigaciones anteriores en el mismo ámbito de trabajo para, posteriormente, ser apuntadas posibles aplicaciones futuras en las que será útil el uso de los avances logrados en el presente estudio.

2. Contexto y estado del arte

El auge de las técnicas de procesamiento de los lenguajes naturales ha causado una oleada de investigaciones que tratan de entender o evaluar diferentes aspectos psicológicos y cognitivos mediante Deep Learning y modelos de procesamiento del lenguaje (Kong, 2012).

Como hemos hablado anteriormente, diferentes autores no atribuyen un valor estático a la personalidad de los individuos, sino que esta puede fluctuar y cambiar, por lo que, facilitar tareas como obtener un acercamiento de la personalidad, predecir de manera fiable alguna dimensión de esta o de cualquier otro aspecto directamente relacionada con la misma, es algo que ha llamado la atención a diferentes investigadores.

El acercamiento de la tecnología y la estimación de características individuales y psicológicas puede, además, abrir nuevas puertas aún cerradas debido a las herramientas actuales basadas en baterías de tests con un coste de aplicación y de tiempo elevados. Es por esto por lo que cada vez es más frecuente encontrar estas dos ramas de la mano y presenciar el acercamiento actual entre ambas.

2.1 – Aprendiendo la personalidad de un sujeto mediante sus comentarios en Twitter y Yahoo! Answers.

Algunos de estos acercamientos, especialmente relacionados con la predicción de la personalidad de los individuos, se han dado mediante el uso de técnicas de web scraping en redes sociales en las que la interacción entre individuos es alta como, por ejemplo, Twitter o Facebook.

Las redes sociales son un sitio en el que las personas deben presentarse al mundo e interactuar, grabando en ellas parte de su personalidad. Usando esto como premisa, Jennifer Golbeck et al. (2011) trataron de predecir la personalidad de los usuarios de la red social Twitter a través de sus interacciones mediante técnicas de procesamiento del lenguaje natural. En esta investigación se usó información pública accesible a través de los perfiles de diferentes usuarios de la red social para predecir las diferentes dimensiones o tratos del

modelo teórico de la personalidad Big Five (Extraversión, Apertura a la Experiencia, Neuroticismo, Amabilidad y Escrupulosidad).

Cada una de estas dimensiones fue estudiada de manera aislada del resto, obteniendo en el modelo métricas diferentes para cada una de las dimensiones indicador de la personalidad. Goldbeck at al. realizaron un pre-procesamiento del texto basado en la connotación que tienen diferentes palabras y el área al que se aplican, creando así categorías a las cuales podrían referirse las diferentes palabras. Algunas de estas categorías eran por ejemplo (‘You’, ‘Negative’, ‘Friends’ o ‘Work’). Se debe señalar que en el modelo de la personalidad Big Five, se mide la puntuación que un sujeto obtiene en cada una de las dimensiones, por lo que, durante las etapas de diseño, los autores se decantaron por el uso de un modelo basado en los algoritmos de regresión Gaussian Process y ZeroR de la herramienta Weka.

Gracias al modelo, los investigadores pudieron aproximar los niveles de personalidad según las dimensiones del enfoque teórico del Big Five para el cómputo del conjunto de datos. A continuación, con los datos utilizados para el análisis y entrenamiento, fueron capaces de probar el modelo en instancias de manera individual. Con estos acercamientos, Golbeck et al. concluyeron que la personalidad puede ser predicha a través de la información personal que compartirnos en redes sociales, en su caso utilizando comentarios publicados en discusiones generadas en la red social Twitter.

Otros autores también han investigado estas interesantes áreas y descubierto el valor que esconden los datos generados por interacciones entre usuarios en redes sociales, especialmente a la hora de inferir aspectos estrechamente relacionados con la personalidad de un individuo.

Una reciente investigación realizada por Nicolás Olivares et al. (2018) propuso un nuevo modelo de análisis de personalidades basado de nuevo en la teoría de la personalidad Big Five de Costa y McCrae (1996). Para desarrollar este modelo, los investigadores utilizaron 370.000 instancias resultantes de preguntas y respuestas publicadas en el foro comunitario *Yahoo Answers* las cuales fueron extraídas mediante la aplicación de técnicas de web scraping en el portal.

Estas instancias de lenguaje natural fueron tratadas de manera que fuera posible extraer información acerca de las características lingüísticas que tipifican cada uno de los tratos de la personalidad (Extraversión, Apertura a la Experiencia, Neuroticismo, Amabilidad y Escrupulosidad) del modelo Big Five de la personalidad. Para llevar a cabo esta investigación,

los investigadores dividieron y estandarizaron la prueba de personalidad Big Five en 112 adjetivos o descriptores, los cuales se relacionaban con diferentes ítems utilizados para codificar la personalidad mediante baterías de tests.

A continuación, se realizó un baremo de los descriptores para relacionarlos numéricamente con cada una de las dimensiones de la personalidad del modelo Big Five a través de una ecuación. En el estudio se utilizaron quince modelos diferentes sacados a partir de algoritmos de clasificación basados en regresión lineal como Support Vector Machine (SVM), basados en el teorema de Bayes (clasificador bayesiano ingenuo) y de modelos exponenciales (Maximum Entropy) con el fin de alcanzar un modelo que propusiera un mejor acercamiento a la solución del problema planteado inicialmente.

Sus resultados, utilizando la accuracy como métrica, revelaron que es posible entrenar modelos para discriminar entre distintos tipos de personalidad a través de las publicaciones realizadas en comunidades online obteniendo un rendimiento en esta métrica del 0.8 en la dimensión apertura a la experiencia, 0.820 para la dimensión escrupulosidad, 0.833 para la dimensión neuroticismo, 0.846 en la dimensión extraversión y 0.855 en la dimensión amabilidad.

Como en estudios comentados anteriormente, se probó que existen patrones de conductas dentro de la lingüística de los sujetos que son diferenciadoras a la hora de estimar las puntuaciones que se obtendrían mediante una prueba tradicional de personalidad. Esto se logró, entre otros aciertos, gracias al nuevo enfoque basado en la descomposición de los ítems de la prueba en descriptores lingüísticos que generaban clases estructuradas con las que entrenar el modelo.

2.2 – Modelos de machine learning creados a través de la prueba de Myers-Briggs:

Como se ha comentado en el apartado anterior del estudio, el indicador de personalidad de Myers-Briggs y su implementación dentro de la plataforma 16personalities ha supuesto la generación de un gran volumen de datos con clases estructuradas sobre las que trabajar la minería de datos y, a la vez, ha supuesto un interesante punto de partida para aplicar técnicas de inteligencia artificial de una manera más sencilla a las comentadas anteriormente.

Esto ha dado lugar a un mayor interés por inferir conclusiones de datos a través de los resultados proporcionados por el indicador de Myers-Briggs utilizando modelos de machine learning basados en el procesamiento de lenguajes naturales.

Michael C. Komisin (2011) utilizó los resultados de 40 estudiantes en el indicador de personalidad de Myers-Briggs para crear un modelo de aprendizaje automático que distinguiera la personalidad de un sujeto, basándose en la elección de las palabras utilizadas para expresarse.

El modelo Komisin (2011) se construyó mediante clasificadores probabilísticos y no probabilísticos y con ayuda de herramientas de análisis de emociones, características cognitivas y psicológicas como el Linguistic Inquiry and Word Count (LIWC). Este estudio concluyó que las diferentes dimensiones de los tipos de personalidad descritas en el indicador Myers-Briggs no son independientes, sino que interactúan entre sí a la hora de seleccionar las palabras utilizadas. Sin embargo, los resultados del estudio no obtuvieron una precisión superior al 0.70 para ninguna de las dimensiones que evalúa el indicador de Myers-Briggs, a lo cual, el investigador incluyó como una posible mejora del modelo, el uso de un conjunto de datos con un mayor número de instancias.

Alam Sher Khan et al. (2020) publicaron recientemente un nuevo modelo de predicción de la personalidad basado en el modelo Myers-Briggs. Para la realización de la investigación se valieron de un dataset publicado para llevar a cabo una competición en Kaggle. Este dataset cuenta con 8675 instancias, en las cuales, una columna contiene el tipo de la personalidad del sujeto y las otras diferentes publicaciones realizadas por esa misma persona. Con esto, se vieron superadas las limitaciones encontradas en investigaciones anteriores y las debidas al conjunto de datos utilizado, como es ejemplo la realizada por Komisin (2011).

La investigación se propuso con el fin, no solo de lograr inferir la personalidad del individuo con fines de evaluación, sino con un enfoque relacionado con cómo esta afecta en el rendimiento laboral. Del mismo modo que en investigaciones anteriores, algunos de los problemas iniciales que debían superarse aún tenían que ver con el conjunto de datos, en este caso, se debían al desbalanceo de las clases de personalidad.
Para la elaboración de los modelos se utilizaron algoritmos de clasificación supervisada como el clasificador bayesiano ingenuo, support vector machines (SVM), de regresión logística y gradient boosting. Los mejores resultados fueron los resultantes de este último modelo realizado con el algoritmo XGBoost, con el cual obtuvieron para la métrica de precisión un rendimiento del 0.870 en la dimensión Introversión-Extraversión, un 0.923 en la dimensión Sensitivo-Intuitivo, un 0.890 en la dimensión Pensativo-Emocional y un 0.858 en la dimensión Calificador-Perceptivo.

2.3 – Otras investigaciones a través de NLP aplicadas a la psicología.

Los estudios de procesamiento del lenguaje natural en el ámbito de la psicología no solamente tienen como objetivo la predicción de la personalidad. Actualmente nos encontramos en un momento en el que algunas líneas del aprendizaje automático se han focalizado en el análisis del sentimiento en los textos. Estos estudios han ido ganando interés, sobretodo en lo relacionado a reseñas y opiniones de clientes o cuestionarios de satisfacción con un campo de respuesta abierta. Este tipo de interacciones genera datos desestructurados que deben ser procesados mediante técnicas relacionadas con el procesamiento de los lenguajes naturales.

Este tipo de estudios ha ayudado a las empresas a realizar estudios de segmentación de mercado acerca de diferentes productos, ayudando a entender cuáles son las preferencias de los clientes con respecto a una gama de productos o a un negocio concreto. Actualmente, las técnicas de minado de opiniones y su futuro procesamiento a través de técnicas de aprendizaje automático se encuentran entre uno de los principales pilares a la hora de dotar de inteligencia de negocio (B. Liu, 2012).

Este tipo de avances corporativos están dando lugar a investigaciones como la llevada a cabo por Franco Chiavetta et al (2020). En este estudio, se extrajeron 8255 reseñas realizadas a libros adquiridos en el portal de venta online Amazon. Este estudio fue llevado a cabo únicamente con reseñas realizadas en italiano y una de las innovaciones que propuso, fue la de llevar a cabo el modelo a través de uno basado en el acercamiento a la creación de un lexicón italiano inspirado en el léxicon desarrollado para el inglés de SentiWordnet.

Con esto se desarrolló un modelo basado en términos semánticamente similares con una base científica ontológica. Su modelo mostró unas métricas de precisión (Accuracy) del 0.829 mostrando un mejor rendimiento para la clasificación de reseñas positivas con un desempeño de 513 reseñas positivas correctamente clasificadas como positivas frente a 87 reseñas positivas incorrectamente clasificadas como negativas (0.855). Mientras que para las reseñas negativas se clasificaron correctamente 482 reseñas como negativas y 118 reseñas negativas fueron incorrectamente clasificadas como positivas (0.820).

Pero el análisis de sentimientos no sólo ha sido un área de estudio aplicado desde el mundo corporativo y a diferentes mercados, sino que este tipo de estudios también han sido un objeto de estudio en auge en los últimos años con el enfoque científico de estudiar las diferencias individuales. De esta manera y, con un especial interés en estudiar las consecuencias de la crisis sanitaria del COVID-19 iniciada en 2020 Daniel M. Low et al. (2020) se propusieron

estudiar el impacto de la crisis con respecto a los sentimientos expresados por personas en las redes sociales.

Para la realización del estudio, Daniel M. Low utilizó técnicas de minado de texto sobre comentarios realizados en diversas páginas del foro de opinión Reddit, especialmente las dedicadas a la expresión de opiniones relacionadas con emociones derivadas de problemas mentales como pueden ser la depresión, la esquizofrenia o la tendencia al suicidio.

En el estudio se recopilaron comentarios publicados entre 2018 y 2020 y provenientes de 826,961 usuarios diferentes. A través del aprendizaje automático supervisado fueron capaces de categorizar las publicaciones de Reddit para posteriormente analizarlas e interpretarlas según los diferentes grupos de personalidades.

Los resultados encontrados en la investigación mostraron como sentimientos negativos derivados de la situación vivida por la pandemia se incrementaron. De esta manera la expresión de soledad y otros síntomas negativos como estrés económico o desorden alimenticios se incrementaron durante 2020, especialmente debido a la situación vivida. El autor concluyó que los comentarios relacionados con el tema del suicidio y la soledad se duplicaron durante la pandemia. Igualmente se vio un incremento en las publicaciones de los individuos con predisposición a la ansiedad, durante los dos meses anteriores a la pandemia (enero y febrero de 2020).

Todas estas investigaciones prueban que el acercamiento entre las técnicas de inteligencia artificial y el estudio de características psicológicas y diferenciales tienen un largo y emocionante campo que explorar ya que, aunque se ha demostrado que es posible inferir algo tan abstracto como los sentimientos o la personalidad, aún se desconoce cuánto conocimiento relacionado con capacidades y aptitudes psicológicas ocultan los millones de interacciones que son generadas diariamente a través de redes sociales.

Estas investigaciones pueden llegar a ser avances realmente importantes de cara a poder identificar grupos vulnerables a través de las redes sociales, medio de comunicación y expresión que está en auge en estos últimos años. De esta manera, el procesamiento de lenguaje natural permitiría ayudar a los profesionales a identificar los inicios de un posible trastorno mental, como puede ser la depresión o la ansiedad.

En la tabla 2 se establece, a modo de resumen, un esquema de cada una de las investigaciones anteriormente repasadas, así como las características diferenciadoras de cada una de ellas:

Tabla 2 – Comparación de los estudios que han inspirado la investigación.

Autor	Año	Área de Estudio	Origen de los Datos	Algoritmos
Golbeck et al.	2011	Personalidad (Modelo Big Five)	Minado de texto realizado en comentarios de Twitter	Gaussian Process y ZeroR
Nicolás Olivares et al.	2018	Personalidad (Modelo Big Five)	Preguntas y respuestas del portal Yahoo Answers	Suport Vector Machines (SVM), Naive Bayes, Maximum Entropy.
Michael C. Komisin	2011	Personalidad (Indicador Myers-Briggs)	Cuestionarios escritos.	Suport Vector Machines (SVM)
Alam Sher Khan et al.	2020	Personalidad (Indicador Myers-Briggs)	Kaggle (Minado de texto de Twitter)	Regresión Logística, Suport Vector Machines (SVM) y Gradient Boosting.
Franco Chiavetta et al.	2020	Sentimientos expresados en opiniones	Minado de reseñas en Amazon.	Clasificación Binaria.
Daniel M. Low et al.	2020	Sentimientos expresados en interacciones persona-persona.	Minado de comentarios de Reddit.	Stochastic Gradient Descent Linear Classifier y Support Vector Machines (SVM)

Elaboración propia.

3. Objetivos concretos y metodología de trabajo

3.1. Objetivo general

Como hemos introducido en apartados anteriores, el objetivo del presente trabajo consiste en elaborar un modelo de machine learning capaz de predecir la personalidad de un sujeto a partir de textos escritos por ese mismo individuo, evitándose así el tener que llevar a cabo pruebas de evaluación psicológica de alto coste y tiempo de intervención.

Para llevar a cabo el modelo y ampliar el conocimiento adquirido en investigaciones anteriores, se hará uso de un dataset publicado recientemente en Kaggle con 1794016 instancias de sujetos evaluados a través del indicador Myers Briggs y publicaciones realizadas por los mismos. Del mismo modo se hará uso de técnicas de psicología del lenguaje para llevar a cabo la etapa de pre-procesamiento del texto y lograr unas métricas de validación del modelo superiores a las logradas en anteriores investigaciones.

3.2. Objetivos específicos

Para lograr el objetivo propuesto será necesario desglosar este en pequeñas etapas y requisitos que deben verse satisfechos, a destacar los siguientes:

- Explorar y analizar el conjunto de datos de manera efectiva para llevar a cabo las posteriores etapas de un modo ordenado, mostrando de manera explicativa, clara y visual las características del conjunto de datos utilizado para entrenar el modelo y llevar a cabo el trabajo.

- Identificar y aplicar las diferentes etapas de procesamiento de manera coherente y con un claro enfoque hacia la discriminación de las diferentes dimensiones expuestas en el indicador Myers-Briggs.

- Generar y tipificar las variables derivadas del texto de lenguaje natural de manera efectiva para facilitar al algoritmo la tarea de clasificación de manera que se pueda ajustar el conjunto de datos y los parámetros del modelo en la mejor dirección posible.

- Diseñar un modelo de machine learning capaz de identificar y predecir las diferentes dimensiones de la personalidad de un individuo a través de textos escritos por el mismo sujeto.

- Determinar los parámetros y la estructura de los algoritmos utilizados y qué mejores resultados obtiene a la hora de realizar la tarea de clasificación. Describir estos de manera que pueda ser replicado el estudio en futuras investigaciones.

3.3. Metodología del trabajo

El trabajo siguió un método experimental sobre el conjunto de datos definido anteriormente, con el fin de localizar el mejor modelo a la hora de inferir la personalidad de un sujeto según el indicador Myers-Briggs a través de comentarios escritos por él mismo. Para ello se llevaron a cabo las siguientes etapas de manera ordenada y bien definida:

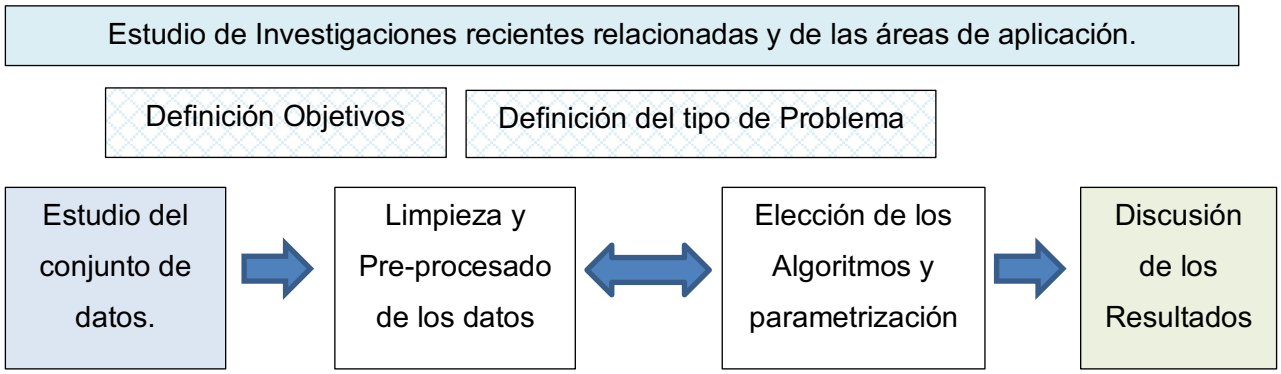

Figura 2 – Diagrama de la metodología.

(Elaboración propia)

- En primer lugar, se situó el presente trabajo, se recogieron diferentes investigaciones realizadas anteriormente, las cuales sirvieron para situar nuestro estudio y marcar objetivos realistas. Fue importante el estudio de las investigaciones anteriores con las que se sitúa de nuestro trabajo, a saber, la personalidad y el procesamiento del lenguaje natural. Una vez se hubo alcanzado el grado de conocimiento suficiente en ambos campos gracias al estudio y a la recogida de información, se marcaron los objetivos que ya hemos comentado anteriormente y que serán el eje principal del trabajo.

- A continuación, se realizó el proceso de familiarización con el conjunto de datos. Aun conociendo de manera superficial en qué consisten estos, los fueron estudiados

profundamente. Esto implicó, estudiar el balanceo del dataset, las clases y la codificación de estas. Del mismo modo, se analizaron los comentarios del dataset, los cuales, han sido minados de la red social Twitter. Este primer acercamiento fue importante, además de para la facilitación de futuras etapas, para contextualizar los datos, entendiendo que suponen la base fundamental sobre la que se va a sustentar todo el modelo y siendo algo imprescindible para avanzar en el diseño del modelo de manera coherente.

- La etapa previa facilitó la tarea de localizar errores redundantes contenidos en el conjunto de datos y dará pie a la limpieza y el pre-procesamiento de los datos. Esta etapa se dividió a su vez tres fases. En primer lugar, se realizó el procesamiento de las clases del dataset. Las clases definidas fueron los 16 tratos de la personalidad establecidos en el indicador Myers-Briggs, y sus divisiones en sub-clases, las cuales fueron definidas como cada una de las dimensiones de la personalidad.

- En segundo lugar, se realizó el procesamiento y la limpieza de los comentarios mediante técnicas de procesamiento del lenguaje natural. Una vez se hubo finalizado estas dos fases, se pasó a la tercera, que consistió en la reducción de los datos eliminando instancias que no aportan información suficientemente útil.

- A continuación, contando con un conjunto de datos limpio, comenzó la etapa de diseño. Se definieron las variables aplicando los algoritmos de vectorización necesarios para ello. Así mismo, se empezaron a tomar decisiones con criterio sobre la cantidad de datos necesarios para optimizar el modelo, la longitud de los comentarios, el tipo de palabras pueden facilitar al modelo la tarea de clasificar correctamente y lo más importante, el tipo de problema al que nos enfrentamos, con el fin de elegir el mejor algoritmo.

- Se comparó el rendimiento obtenido por diferentes algoritmos en la tarea de clasificación. En este punto fue importante revisitar etapas anteriores tantas veces como sea necesario para optimizar el funcionamiento del modelo al mismo tiempo que se parametrizan los algoritmos de manera coherente.

- Por último, se evaluaron y analizaron los resultados obtenidos y se estudió la usabilidad de estos. Con los resultados expuestos de manera clara, se comentan las vías que continúan el camino abierto al mismo tiempo que se exponen cuales han sido las barreras que han impedido obtener los mejores resultados.

4. Desarrollo específico de la contribución

Para llevar a cabo el trabajo, se utilizará el dataset actualizado de tipos de personalidad Myers-Briggs publicado en Kaggle. Este dataset es una actualización del conjunto de datos utilizado en el mismo sitio web para realizar la competición Inclass de predicción de personalidad llevada a cabo por Kaggle en la cual participaron un total de 45 equipos y 178 competidores.

El conjunto de datos inicialmente utilizado para la competición cuenta con 8675 instancias, cada una correspondiente a un usuario del sitio web Personality Cafe website forums. Cada instancia cuenta con dos columnas. La primera de ellas se corresponde a la clase de personalidad del sujeto según las dimensiones del indicador Myers-Briggs y según el resultado obtenido al llevarse a cabo la prueba de personalidad del sitio web 16personalities. La segunda columna se corresponde con publicaciones realizadas por ese usuario en twitter, las cuales han sido extraídas y almacenadas mediante técnicas de minado de texto basado en el web scraping.

Este mismo dataset ha servido para la realización de diferentes investigaciones anteriormente mencionadas, así como para diferentes estudios y actividades de entretenimiento, convirtiéndose de esta manera en un conjunto de datos con gran aceptación y popularidad dentro de la comunidad de Kaggle.

Como hemos dicho, el dataset utilizado en este trabajo es una ampliación, donde se han visto incrementadas las instancias de 8675 a 1794016. El dataset mantiene la misma estructura que hemos expuesto anteriormente, cada una de estas instancias cuenta con una columna de clase correspondiente a la personalidad del sujeto según las dimensiones que reconoce el indicador Myers-Briggs (1944). La otra columna se correspondiente con la publicación o las publicaciones que el sujeto ha realizado en la red social Twitter. Es importante señalizar en este punto que el dataset únicamente contiene instancias que han sido escritas en inglés y que pueden contener vínculos, emoticonos y otros elementos típicos a la hora de interactuar en redes sociales.

Una vez se ha superado la etapa de familiarizarnos con el dataset, se dará comienzo a la exposición de los siguientes pasos. En primer lugar, se debe puntualizar que tanto para la importación del conjunto de datos, la realización de la limpieza de datos, el análisis exploratorio, la ejecución del pre-procesamiento y el desarrollo del modelo, que serán expuestos en los siguientes puntos del estudio, se ha utilizado la versión 3.8 de Python a través de la herramienta Jupyter Notebook.

En la tabla 2 se introducen las librerías de Python utilizadas para llevar a cabo las diferentes fases de la investigación, aunque más adelante se profundizará en las técnicas y tecnologías utilizadas en cada fase del trabajo según sea pertinente.

Tabla 3 – Librerías utilizadas durante el desarrollo.

Librería Utilizada	Objetivo	Fases de utilización
Pandas	Manipulación del dataset; Exploración del dataset; Limpieza del dataset; Pre-Procesamiento del lenguaje natural.	Limpieza del dataset; Pre-Procesamiento; Análisis exploratorio.
Matplotlib / Seaborn	Generación de gráficos.	Análisis exploratorio; Resultados.
Numpy	Realización de cálculos matemáticos y estadísticos.	Análisis exploratorio; Resultados.
Wordcloud	Pre-Procesamiento del lenguaje natural; Generación gráfica de la nube de palabras.	Análisis exploratorio; Pre-Procesamiento.
nltk	Pre-Procesamiento del lenguaje Natural. División de corpus y StopWords. Tokenización de palabras.	Pre-Procesamiento.
xgboost	Creación y parametrización del algoritmo.	Diseño del modelo; Entrenamiento; Validación del modelo; Resultados.
Scikit-Learn	Creación y parametrización del algoritmo y las fases de pre-procesamiento.	Pre-procesamiento; Estandarización; Entrenamiento; Validación del modelo; Resultados.

Elaboración propia.

Limpieza del dataset y pre-procesamiento de los datos.

En primer lugar, para comenzar a sacar conclusiones acerca del dataset, hemos realizado una primera limpieza del dataset sobre la columna correspondiente a la personalidad. Para ello se han codificado todas las instancias de manera que cada trato de la clase contenga únicamente las cuatro letras que definen una personalidad según el modelo Myers-Briggs.

Se ha de recordar que cada una de las letras se corresponde con cada una de las cuatro dimensiones del indicador Myers-Briggs y siempre en el mismo orden, así pues, como ya ejemplificamos en puntos anteriores del presente trabajo, si los resultados del indicador para el individuo son Extravertido (E), Intuitivo (N), Pensativo (T) y Perceptivo (P) su personalidad quedará codificada como la combinación de estas cuatro letras, en este caso (ENTP) mientras que, si la personalidad del sujeto fuera ISFM las dimensiones de la personalidad que han caracterizado su tipo de personalidad han sido del tipo Introvertida (I), Sensitiva (S), Emocional (F) y Juez (J).

Con esta primera etapa, se han estandarizado cada uno de los 16 tipos de personalidad contenidos en el dataset, eliminando errores causados por una mala tipificación y errores debidos a espacios al comiendo y final de los valores. También se han estandarizado las clases para que aparezcan únicamente en mayúscula, eliminando la aparición de clases codificadas de manera inconsistente.

Tras realizar esta primera etapa, se ha pintado en un gráfico de barras la frecuencia en que se da cada uno de los tratos de la personalidad.

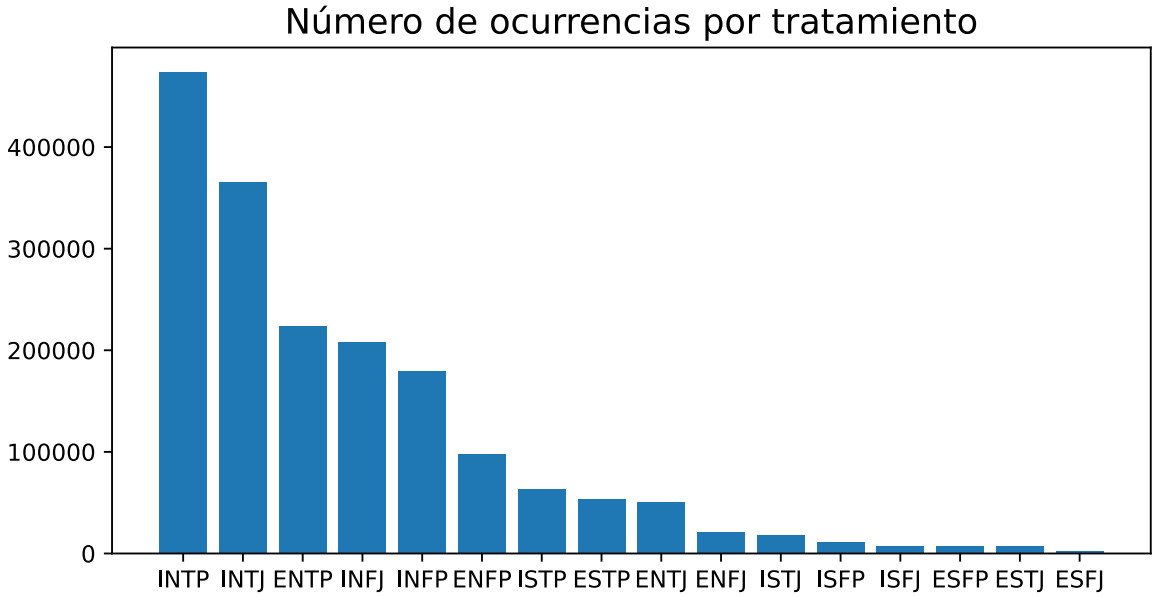

Figura 3 – Distribución de frecuencias para cada tipo de la personalidad.

(Elaboración propia)

En la figura 3 podemos observar la distribución de frecuencias absolutas de cada uno de los tratamientos de la personalidad de entre el espectro de personalidades que se configuran a través de las dimensiones contempladas en el indicador Myers-Briggs (1944).

Como se puede observar en un primer vistazo, predominan los tipos de personalidad cuya primera dimensión de la personalidad es de tipo introvertida (I) sobre los sujetos con esta dimensión de tipo extrovertida (E). Esto mismo sucede con la segunda dimensión del indicador Myers-Briggs, en la que podemos apreciar que, con respecto al modo de percibir y tomar la información, hay una mayoría de instancias en las cuales esta dimensión de la personalidad es de tipo intuitiva (N) frente a una minoría de sujetos que muestran una posición sensitiva (S). Con el fin de explorar la distribución de tratamientos y hacer más relevantes estas diferencias entre cada dimensión de la personalidad se han pintado cuatro gráficos de sectores, cada uno representando una dimensión de la personalidad del indicador Myers-Briggs (1944).

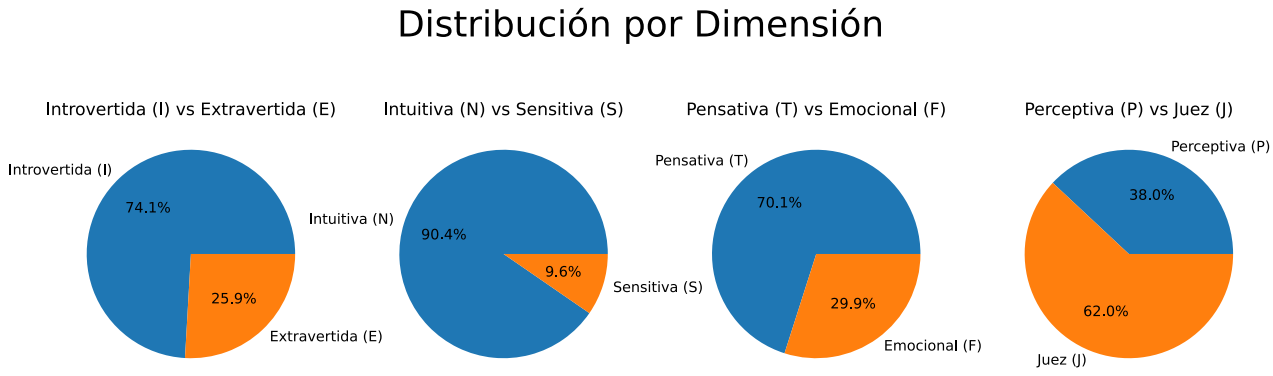

Figura 4 - Distribución de la proporción de cada tratamiento por dimensiones.

Con esta visualización se hacen más evidentes las diferencias en el balanceo en cada una de las cuatro dimensiones. En primer lugar, al poner la atención en la primera dimensión vemos como el número de instancias pertenecientes a la clase Introvertida (I) son aproximadamente el triple (75.8%) que las pertenecientes a la clase Extravertida (E) (24.2%). Esto era algo que ya notábamos con la visualización de barras de la figura 3, pero la diferencia en la distribución se hace más clara. Del mismo modo que en la figura 3 se pudo apreciar la diferencia que había entre las frecuencias absolutas de las clases pertenecientes a la primera dimensión, en la segunda dimensión se hacen aún más notables. Como se puede apreciar en la representación de la figura 4, una gran mayoría de las instancias pertenecen a la clase Intuitiva (N) con un 91.5 por ciento frente al 8.5 por ciento que pertenece a la clase Sensitiva (S). Esta

es la dimensión que cuenta con un mayor desbalanceo y supone un punto importante a tener en cuenta en futuros pasos del estudio, ya que, una de las clases tiene más de 10 veces las instancias de la otra.

Durante el procesamiento del texto de lenguaje natural correspondiente a la segunda columna del dataset se ha utilizado la librería Pandas en combinación de la librería Wordcloud. Con la combinación de estas librerías se han realizado las siguientes etapas en el orden que se muestra a continuación:

1. Conversión de cada ítem a un objeto de tipo cadena de texto.
2. Eliminación de los hipervínculos.
3. Eliminación de los signos de puntuación, interrogación o exclamación.
4. Eliminación de espacios al principio y al final de cada frase.
5. Eliminación de conjuntos de letras no reconocidos como palabras.
6. Conversión de los textos a minúsculas.
7. Eliminación de los tratos de personalidad escritos dentro del texto de los usuarios ya que en algunos comentarios aparecen los resultados que el usuario ha obtenido en la prueba y esto puede generar problemas de sobreajuste en el modelo en futuras etapas de la investigación.
8. Eliminación de palabras vacías (StopWords). Algunos ejemplos de palabras vacías son los artículos, pronombres o las preposiciones. Este paso se ha realizado utilizando la librería nltk y su paquete “corpus” el cual contiene una lista de palabras vacías o stopwords según el idioma. En este caso se ha utilizado la lista de stopwords correspondientes al inglés.
9. Aplicación de técnicas de Lematización. Este paso se ha llevado a cabo mediante la librería nltk, utilizando el paquete stem de la misma. Este paquete contiene los métodos PorterStemmer() y WordNetLemmatizer() que han sido utilizados para llevar a cabo esta etapa de procesamiento.

En este momento la estructura del dataset sigue siendo la misma que la encontrada inicialmente en Kaggle, contando con una columna que muestra el tipo de personalidad de cada sujeto y la columna correspondiente a las palabras que usaremos para entrenar el modelo.

De cara a continuar con la definición de variables y continuar en futuros pasos con la exploración y el análisis del conjunto de datos se van a crear nuevas columnas a partir de la información contenida en el conjunto de datos.

En primer lugar, vamos a crear cuatro nuevas columnas, cada una de ellas representará una dimensión de entre las cuatro que componen un tipo de personalidad. Los valores asignados a estas columnas serán 0 o 1 al tratarse de dimensiones dicotómicas, así pues, por ejemplo, para la primera dimensión del indicador Myers-Briggs, el cual define la personalidad de un sujeto como Extravertida o Introvertida, si el sujeto tiene un carácter Extrovertido, el valor en la columna Sub_Personalidad_1 será 0, en caso de ser introvertido será 1. De manera similar codificamos el resto de las dimensiones en las columnas que nombraremos Sub_Personalidad_2, Sub_Personalidad_3 y Sub_Personalidad_4. Asignando ya valores enteros se ahorrará tiempo y capacidad de procesamiento en futuras etapas como la codificación de etiquetas.

A continuación, vamos a crear una columna que represente el número de palabras del comentario realizado por el usuario una vez llevado a cabo el procesamiento del texto de lenguaje natural. Esta columna nos será útil en futuras fases del trabajo, pero además nos ha permitido visualizar, como se puede ver en la figura a modo de histograma, el gran número de instancias con insuficientes palabras útiles.

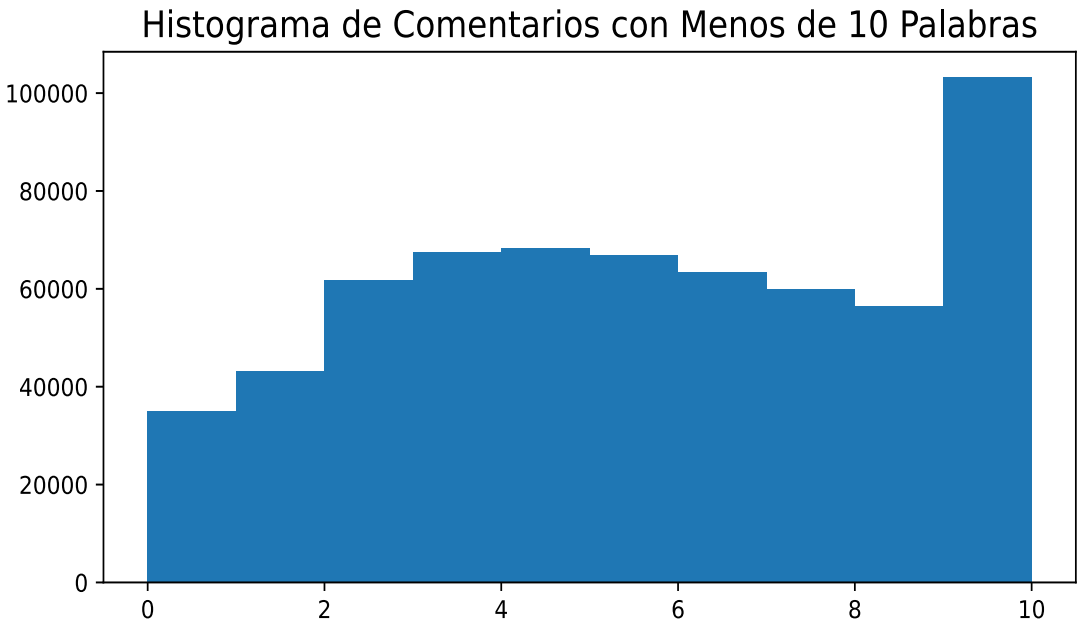

Figura 5 – Distribución de frecuencias para cada tipo de la personalidad.

(Elaboración propia)

En primer lugar, y gracias a la visualización de la figura 5, se han localizado 43266 instancias con tan sólo una palabra útil. Estas instancias no serán tenidas en cuenta en el modelo, ya que, tras explorar ejemplos pertenecientes a las mismas, podemos comprobar que la información contenida en estos comentarios, habitualmente nombres propios e insultos, no es suficiente como para ser tenida en cuenta.

Lo mismo ocurre con las instancias que tan solo contienen 2 o 3 palabras en los comentarios, en estos casos, nos encontramos con 61841 y 67560 instancias respectivamente. Al igual que ocurría con las instancias que tan solo contenían una palabra útil por comentario, estas tampoco serán tenidas en cuenta de aquí en adelante.

Después del análisis de las instancias con menos de 10 comentarios, se ha decidido establecer como criterio el contener más de 7 palabras, una vez realizado el procesamiento del texto, a una instancia para ser incluida en el modelo.

Una vez tomada esa decisión, se han eliminado las instancias en las cuales el número de palabras en la columna de comentarios no superaba las siete una vez aplicadas todas las etapas de pre-procesamiento del texto. Con esto se han reducido el número de instancias de 1794016 a 997486, manteniendo únicamente las que aporten una información más útil al modelo para facilitar al algoritmo la tarea de clasificación.

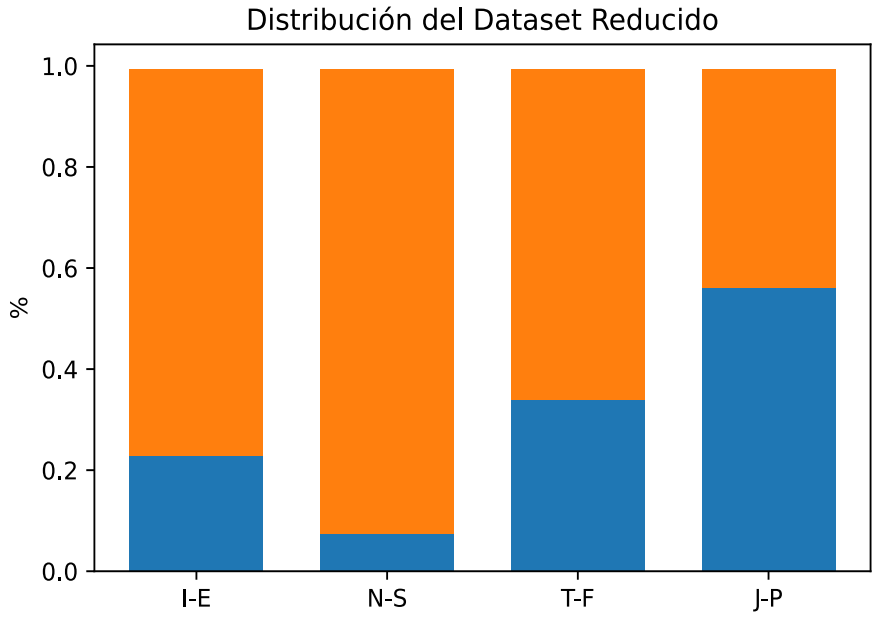

Figura 6 – Distribución de cada dimensión tras la reducción del dataset.

(Elaboración propia)

Sobre este nuevo dataset, se ha vuelto a analizar la manera en que se distribuyen las clases para cada una de las cuatro dimensiones definidas en el indicador Myers-Briggs. Los resultados obtenidos se pueden apreciar en la figura 6.

Al comparar los resultados de las distribuciones de frecuencias obtenidas a partir del dataset con el total de instancias (Figura 4), con el gráfico obtenido con el nuevo dataset reducido, el cual cuenta con 997486 instancias y que ha sido resultante de realizar el primer pre-procesamiento de texto y de filtrar el número de instancias para tener en cuenta, observamos que las personalidades con una dimensión emocional (F) tienen una mayor proporción de instancias con respecto a la clase pensativa (T). Del mismo modo vemos como la cuarta dimensión de la personalidad del indicador Myers-Briggs ha quedado más balanceada que en los datos crudos anteriormente analizados. En este caso ahora ambas están próximas al 60% en el caso de la clase Perceptiva y 40% de la clase Juez.

Es interesante darle importancia a este punto, ya que, aunque en este momento las distribuciones hayan cambiado debido a la eliminación de instancias que no aportaban información suficiente bajo el criterio establecido para el trabajo, en futuras etapas del estudio, como en el diseño del modelo, se tendrá en cuenta la noción de cómo las distribuciones de las clases de personalidad para cada dimensión cambian tomando como variable el número de palabras por comentario. Esto será estudiado más adelante para tomar una decisión sobre la manera en que se incluirá, se balanceará y trabajará con la variable “Numero_de_Palabras” por comentario.

Una vez ha sido limpiado el conjunto de datos tanto en la columna de clases como en los diferentes atributos y se han codificado las variables correspondientes a las diferentes dimensiones de la personalidad, así como el número de palabras por comentario para cada una de las instancias, se da comienzo a la fase de análisis exploratorio de los datos de cara a sacar conclusiones firmes para el diseño del modelo. En primer lugar, se centrará la atención en el análisis de las clases, es decir, la columna tipo de personalidad y las diferentes columnas correspondientes a cada dimensión.

Análisis exploratorio de las clases.

Aunque ya se ha avanzado en esta dirección, se va a continuar sacando conclusiones en esta línea. En primer lugar, se ha dado con un problema que habrá de ser resuelto para el correcto ejercicio de clasificación por parte del modelo. Este problema se refiere al desbalanceo correspondiente a las personalidades con un tipo de personalidad extravertido en la primera dimensión para la cual apenas encontramos un 22.3% de las instancias.

Esto mismo ocurre de una manera aún más pronunciada en la segunda dimensión de la personalidad, la cual se divide en una dicotomía con valores Sensitivo o Intuitivo. En este caso el problema lo encontramos en las personalidades con una dimensión Sensitiva, las cuales únicamente representan el 9% de las instancias totales del dataset.

Con respecto a estas dos dimensiones se deberán aplicar más adelante técnicas de balanceo de los datos de cara a poder facilitar la tarea de clasificación.

Las dimensiones 3 y 4, que se refieren al carácter pensativo o emocional y juez o perceptivo respectivamente, cuentan con un balanceo suficiente como para trabajar con ellas en futuras etapas, sin embargo, más adelante se evaluará si también se mejoran los resultados aplicando técnicas de balanceo de las clases en dichas dimensiones.

Para continuar, vamos a comprobar la correlación de Pearson que existe entre las diferentes dimensiones de la personalidad al ser determinada una personalidad completa. Para ello vamos a analizar la correlación entre todas las columnas a las cuales hemos nombrado anteriormente como Sub_Personalidad_1 (La dimensión que evalúa el carácter Extrovertido (E) o Introvertido (I)), Sub_Personalidad_2 (Carácter Intuitivo (N) frente al Sensitivo (S)), Sub_Personalidad_3 (Carácter Pensativo (T) frente al Emocional (F)) y Sub_Personalidad_4 (Perceptivo (P) frente a Juez (J). Recordemos que cada una de estas columnas tiene un valor de 0 o 1 representando así una de las dos dicotomías de la manera en que se muestra en la tabla 4.

Tabla 4 – Categorización de las diferentes dimensiones del indicador Myers-Briggs.

	0	**1**
Sub_Personalidad_1 (IE)	Extraversión (E)	Introversión (I)
Sub_Personalidad_2 (NS)	Sensación (S)	Intuición (N)
Sub_Personalidad_3 (TF)	Emoción (F)	Pensamiento (T)
Sub_Personalidad_4 (JP)	Percepción (P)	Juicio (J)

Elaboración Propia

Con esta categorización podemos empezar a sacar conclusiones acerca de la manera en que las diferentes dimensiones se relacionan entre sí en los diferentes tipos de personalidad completos compuestos por las cuatro dicotomías.

En primer lugar, vamos a elaborar una matriz de calor a través de los valores de correlación de Pearson resultantes de evaluar las relaciones entre las diferentes dimensiones a través de los valores de las columnas de Sub_Personalidad. Los resultados que se comentarán a continuación pueden ser consultados de manera visual en la figura 7.

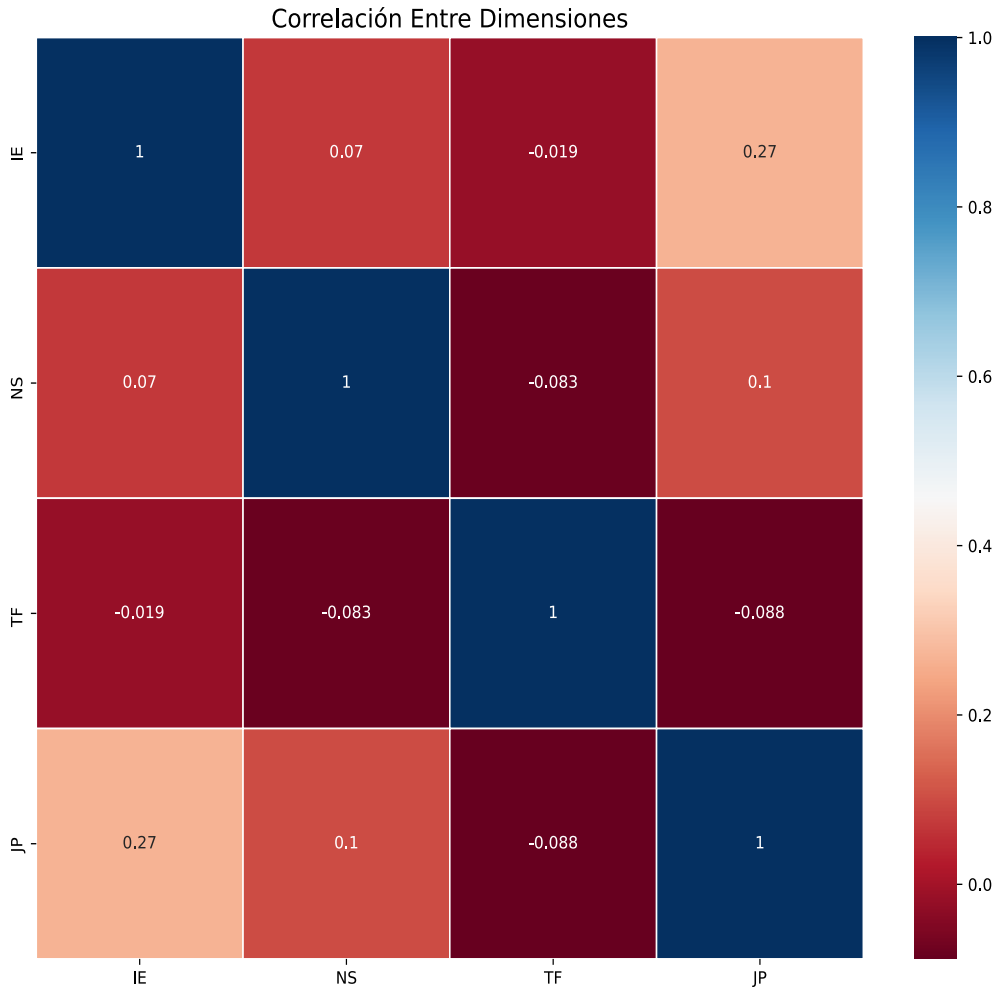

Figura 7 – Matriz de calor elaborado a partir de la correlación de Pearson entre dimensiones de la personalidad.

(Elaboración propia)

Estos resultados nos muestran que la mayoría de las dimensiones no muestran una relación lineal entre sí. Las correlaciones muestran valores positivos y negativos (relación inversa) pero en su mayoría son cercanos a 0. Podemos apreciar que la única correlación con un valor superior a 0.1 es la que existe entre la primera dimensión del indicador (Introversión-Extraversión) y la cuarta dimensión (Juicio-Percepción). Esta correlación tiene un valor de 0.27, lo cual no sería significativo de no ser porque al ser evaluado este valor con el resto de las correlaciones se muestra una diferencia significativa con respecto al resto de relaciones entre dimensiones.

Este resultado es interesante para ser tenido en cuenta a la hora de evaluar el modelo ya que puede iluminar el proceso de interpretación de los resultados de las métricas a la hora de clasificar estas dos dimensiones de la personalidad.

Análisis exploratorio del lenguaje natural.

De cara a poder continuar con el análisis exploratorio de nuestro conjunto de datos, se van a evaluar, junto a las variables relacionadas con los tipos y dimensiones de la personalidad, las generadas a partir de los pasos anteriores de procesamiento del lenguaje natural.

Vamos a recordar que las columnas generadas hasta ahora relacionadas con el lenguaje natural son las elaboradas a partir de los posts escritos por cada uno de los usuarios en Twitter. Se ha pre-procesado el texto de manera que no cuente con palabras vacías ni otros elementos que interfieran con el correcto funcionamiento de los algoritmos, por lo que, en este punto, la columna relacionada con los posts (a la cual se ha nombrado como “Post_Limpio”) sólo contiene palabras útiles. Por tanto, pueden llegar a ser consideradas atributos de importancia para la clasificación. Por otro lado, también se cuenta con la columna correspondiente al número de palabras de la columna “Post_Limpio” a la cual se ha nombrado como “Numero_de_Palabras”.

Una vez ejecutadas las etapas anteriores, se va a realizar una primera nube de palabras generada a partir de todos los comentarios contenidos en las 997486 instancias del dataset sin discriminar tipos de personalidad. Con esta primera nube va a ser posible la identificación de palabras frecuentes y útiles para el modelo utilizadas por los usuarios.

La nube de palabras de la Figura 8 ha sido generada a partir de las palabras contenidas en la columna “Post_Limpio” de todas las instancias.

Figura 8 – Nube de palabras generada a partir del dataset pre-procesado.

(Elaboración propia)

Una vez mostrada la nube de palabras y, para facilitar al lector la tarea de situarse en este punto del trabajo es apropiado recordar que la muestra de comentarios de nuestro dataset se encontraba escrita íntegramente en inglés, y, las palabras en otros idiomas no habrán sido identificadas como parte del corpus de nuestra variable. Continuando con el análisis, en la Figura 8 se pueden observar las palabras que se han utilizado con más frecuencia en los comentarios analizados en el presente trabajo.

En la nube de palabras, la frecuencia de palabras es representada por medio del tamaño con el que cada palabra se muestra. También conviene aclarar que la orientación de cada palabra tiene simplemente un carácter visual. Con estas nociones básicas, llama la atención el uso del plural frente al singular al referirse a “people” frente a “person”, además, esta palabra es la que muestra una mayor frecuencia de uso. La utilización del plural frente al singular puede denotar un mayor número de comentarios con expresiones de juicio general frente a las de descripción de experiencia individuales.

Este resultado puede estar relacionado con diversas dimensiones del indicador Myers-Briggs, especialmente las relacionadas con la introversión y la de juicio frente a percepción, dos dimensiones que pueden basar su modo de expresión en el plural mayestático o plural de modestia.

Continuando con el análisis de esta primera nube de palabras, es interesante observar cómo algunos de los verbos que se utilizan con mayor frecuencia tienen relación con las dimensiones que se definen en el modelo Myers-Briggs. Algunos ejemplos contenidos en la Figura 8 y que representan verbos que pueden ayudar a identificar una dimensión de la personalidad son “feel” (sentir), “think” (pensar), “mind” (importar) o “question” (preguntarse / cuestionar).

Debido a la importancia futura de esta primera etapa de análisis de los comentarios y la escritura de los individuos contenidos en el dataset, para detectar cada uno de los diferentes modelos de personalidad distintos, se decide recrear una nube de palabras en la que se puedan reflejar cuáles son las palabras más utilizadas en el corpus para cada una de las dimensiones de la personalidad del modelo Myers-Briggs. Al generar las nubes de palabras para cada una de las clases (Introvertida, Extravertida, Intuitiva, Sensitiva, Pensativa, Emocional, Juez, Perceptiva) los resultados obtenidos facilitarán la toma de decisiones futuras a la hora de afinar el modelo y poder orientar la investigación.

Dimensión 1.1 Introvertida:

- *Verbos frecuentes en la clase introvertida: "Think" (Pensar) "Feel" (Sentir), "Make" (Hacer), "Want" (Querer) y "Know" (Saber).*
- *Sustantivos frecuentes en la clase introvertida: "Friend" (Amigo), "Time" (Tiempo), "Thing" (Cosa) o "People" (Personas).*

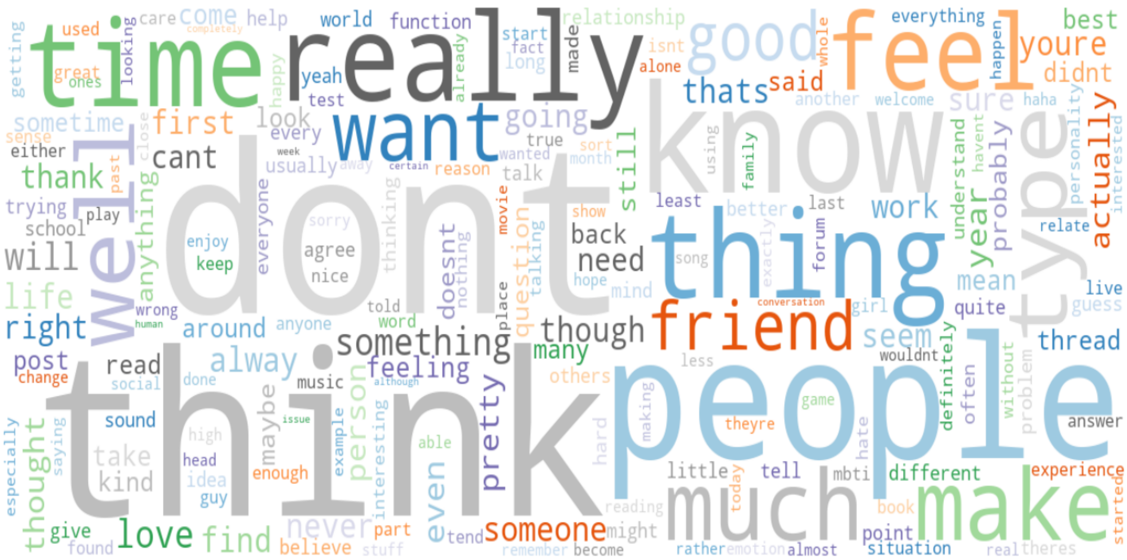

Figura 9 – Nube de palabras para la clase introvertida

(Elaboración propia)

En la nube de palabras que se muestra en la figura 9 se puede apreciar la tendencia a hablar de pensamientos, sentimientos y emociones lo que denota su interés por la introspección. Sin embargo, no suelen mostrar sus propios sentimientos con tanta facilidad como podría darse en los individuos más extrovertidos. Se observan también palabras como "Time" y "Really" lo que nos da a entender también su tendencia a ser estratégicos y pensar en el futuro, sin centrarse en el aquí y ahora.

Dimensión 1.2 Extravertida:

- Verbos Frecuentes en la clase extravertida*: "Think" (Pensar) "Want" (Querer), "Make" (Hacer), "Know" (Saber), "Love" (Querer), "Mean" (Significar) y "Thank" (Agradecer).*
- Sustantivos frecuentes en la clase extravertida*: "Friend" (Amigo), "Time" (Tiempo), "Love" (Amor) o "People" (Personas).*
- *Adjetivos frecuentes: "Good" (Bien) y "Well" (Bueno).*

Figura 10 – Nube de palabras para la clase Extravertida

(Elaboración propia)

En la nube de palabras que se muestra en la figura 10, apreciamos sustantivos y verbos que se utilizan habitualmente cuando estamos en contacto con otras personas, como pueden ser “Friend”, “Love” o “People”. Esto nos lleva a pensar que son personas que tienen una amplia interacción con los individuos de su entorno.

Uno de los verbos más frecuentes en este tipo de personalidades es “Thank” lo que nos da a entender la alta frecuencia de interacciones con los demás.

Sería interesante remarcar también dos adjetivos que aparecen con alta frecuencia en este tipo de personalidades que son “Good” y “Well” lo que nos indica un lenguaje optimista e inferimos también una respuesta aprobatoria frente a los comentarios de otras personas.

Dimensión 2.1 Intuitiva:

- Verbos frecuentes en la clase intuitiva: “Think” (Pensar) “Feel” (Sentir) y “Know” (Saber).
- Sustantivos frecuentes en la clase intuitiva: “People” (Personas), “Friend” (Amigo) o “Love” (Amor).

Figura 11 – Nube de palabras para la clase Intuitiva

(Elaboración propia)

En la nube de palabras que se muestra en la figura 11, podemos apreciar el uso repetido de las palabras “Think”, “Feel” “People” y “Know” lo que nos indica un gran interés por su parte a los sentimientos y pensamientos, no solo por los propios sino también por los ajenos.

Dimensión 2.2 Sensitiva:

- Verbos frecuentes en la clase Sensitiva: “Think” (Pensar), “Find” (Encontrar), “Make” (Hacer), “Know” (Saber) y “Love” (Querer).
- Sustantivos frecuentes en la clase Sensitiva: “Friend” (Amigo), “Year” (Año), “Love” (Amor), “People” (Personas) y Work (“Trabajo”).

Figura 12 – Nube de palabras para la clase Sensitiva

(Elaboración propia)

En la nube de palabras que se muestra en la figura 12, lo que primeramente nos llama la atención es la baja frecuencia de la palabra “Feel”. Las personas que son más propensas a ser sensitivas, cabría esperar que hablasen de sus sentimientos y los expresasen, pero a través de este estudio se observa que la tendencia a expresar emociones a través de las Redes Sociales es bajo. Sin embargo, cabe intuir que la información que tienen más en cuenta es la que perciben a través de sus sentidos y no lo que podrían intuir.

Dimensión 3.1 Pensativa

- Verbos frecuentes en la clase Pensativa: “Think” (Pensar), “Know” (Saber), “Want” (Querer), “Make” (Hacer) y “Need” (Necesitar).
- Sustantivos frecuentes en la clase Pensativa: “Thing” (Cosa), “Time” (Tiempo) o “People” (Personas).
- Adverbio frecuente en la clase Pensativa: “Really” (Realmente).

Figura 13 – Nube de palabras para la clase Pensativa

(Elaboración propia)

En la nube de palabras que se muestra en la figura 13, se observa una tendencia a hablar de pensamientos y no tanto de sentimientos (baja frecuencia de las palabras “Love” y “Feel”), esto se puede ver en la alta aparición de palabras como “Think”, “Know” y “Thing” que son de las más comunes en estos individuos.

Se aprecia también una alta frecuencia en la palabra “Really” lo que da a entender que se expresan con cierta seguridad y no de manera impulsiva. Por ello, a la hora de tomar decisiones tienden a ser más racionales.

Dimensión 3.2 Emocional.

- Verbos frecuentes en la clase emocional: “Think” (Pensar) “Feel” (Sentir), “Know” (Saber) y “Love” (Amar).
- Sustantivos frecuentes en la clase emocional: “Friend” (Amigo), “Time” (Tiempo), “Love” (Amor) y “People” (Personas).

Figura 14 – Nube de palabras para la clase Emocional

(Elaboración propia)

En la nube de palabras que se muestra en la figura 14, se ve una tendencia muy marcada en ciertas palabras como son “Think”, “People”, “Know”, Time” y “Friend” quedando en segundo plano las demás palabras, lo que indica que este tipo de personas tienden a expresar con mayor intensidad sus pensamientos y sentimientos. Dada la frecuencia en palabras como “People” y “Friend” se puede ver la importancia que tienen las demás personas de su entorno.

Dimensión 4.1 Juez.

- Verbos frecuentes en la clase Juez: “Think” (Pensar) “Feel” (Sentir), “Make” (Hacer), “Know” (Saber) y “Love” (Querer).

- Sustantivos frecuentes en la clase Juez: “Time” (Tiempo), “Love” (Amor) o “People” (Personas), “Friend” (Trabajo), “Thing” (Cosa).

Figura 15 – Nube de palabras para la clase Juez

(Elaboración propia)

En la nube de palabras que se muestra en la figura 15, se observa una alta frecuencia en palabras que tienen que ver con la opinión o sensación que puede tener el individuo sobre el exterior, por ejemplo, alta aparición de las palabras pensar, saber y sentir. Este tipo de personas tienen la tendencia a juzgar o sacar conclusiones. También es interesante la alta frecuencia de la palabra “Dont” que da a entender que es común que expresen sus juicios, pudiendo ser estos positivos o negativos.

Dimensión 4.2 Perceptiva

- Verbos frecuentes en la clase intuitiva: “Think” (Pensar) “Want” (Querer) “Know” (Saber) y “Feel” (Sentir).
- Sustantivos frecuentes en la clase intuitiva: “People” (Personas”, “Friend” (Amigo) y “Time” (Tiempo).

Por último, la nube de palabras que se muestra en la figura 16, corresponde a las personas incluidas en la clase Perceptiva. En este tipo de personalidades es común que den mayor importancia a la manera en la que sus sentidos perciben los fenómenos externos, dándonos esto a entender el uso repetido de las palabras “Think” y “Feel”.

Figura 16 – Nube de palabras para la clase Perceptiva

(Elaboración propia)

Una vez analizadas las nubes de palabras para cada clase de la personalidad podemos concluir que las palabras más frecuentes para todas las clases son de uso común, algunas de estas palabras son “Think” (Pensar), “People” (Personas), “Time” (Tiempo) y “Feel” (Sentir). Se puede observar en diferentes clases de la personalidad que los verbos “Think” (Pensar) o “Find” (Encontrar) son utilizados con frecuencia, esto es algo que tiene coherencia si se tiene en cuenta que los comentarios analizados provienen de la red social Twitter la cual es utilizada con frecuencia para expresar opiniones y diferentes pensamientos.

Como se puntualizó al llevar a cabo el análisis de la nube de palabras general con todas las instancias del dataset, el sustantivo “People” (Personas) tiene la mayor frecuencia de uso de entre los sustantivos, esto vuelve a ser algo que cobra sentido al contextualizar el origen de los datos. También llama la atención el uso de sustantivos abstractos como “Time” (Tiempo) o “Love” (Amor). Este tipo de palabras han sido de alta frecuencia y de uso común en todas las clases de personalidad analizadas especialmente en las dimensiones que tienen un carácter calificador o Intuitivo.

Del mismo modo, al observar palabras que han sido utilizadas con menos frecuencia que las anteriores, pero que pueden aportar un valor diferencial a cada una de las dimensiones, nos encontramos con la frecuencia del uso de la palabra “School” (Colegio) en la clase introvertido, la palabra “Family” (Familia) en la clase emocional y la palabra “Best” (mejor) en la clase intuitiva. Es oportuno recalcar también el uso de vocabulario coloquial en la clase Extravertida con palabras como por ejemplo “Cool” (Guay).

Estos resultados son interesantes, pero, debido a la poca diferenciación entre las palabras más usadas por cada uno de los diferentes caracteres de cada una de las dimensiones de la

personalidad del modelo Myers-Briggs (1944) no aportan suficiente luz como para poder tomar decisiones concluyentes. Esto puede deberse a que, una personalidad está formada por el conjunto e interacción de las distintas dimensiones de la personalidad.

En este punto, se han analizado las diferentes palabras que más han sido utilizadas en los comentarios para cada una de las dimensiones y anteriormente, ya se ha mostrado la frecuencia con que se repiten los tipos de la personalidad en el conjunto de datos. Para continuar con la exploración del catálogo de datos, vamos a analizar el número de palabras que contiene de media cada comentario para una de las clases de personalidad que estamos considerando.

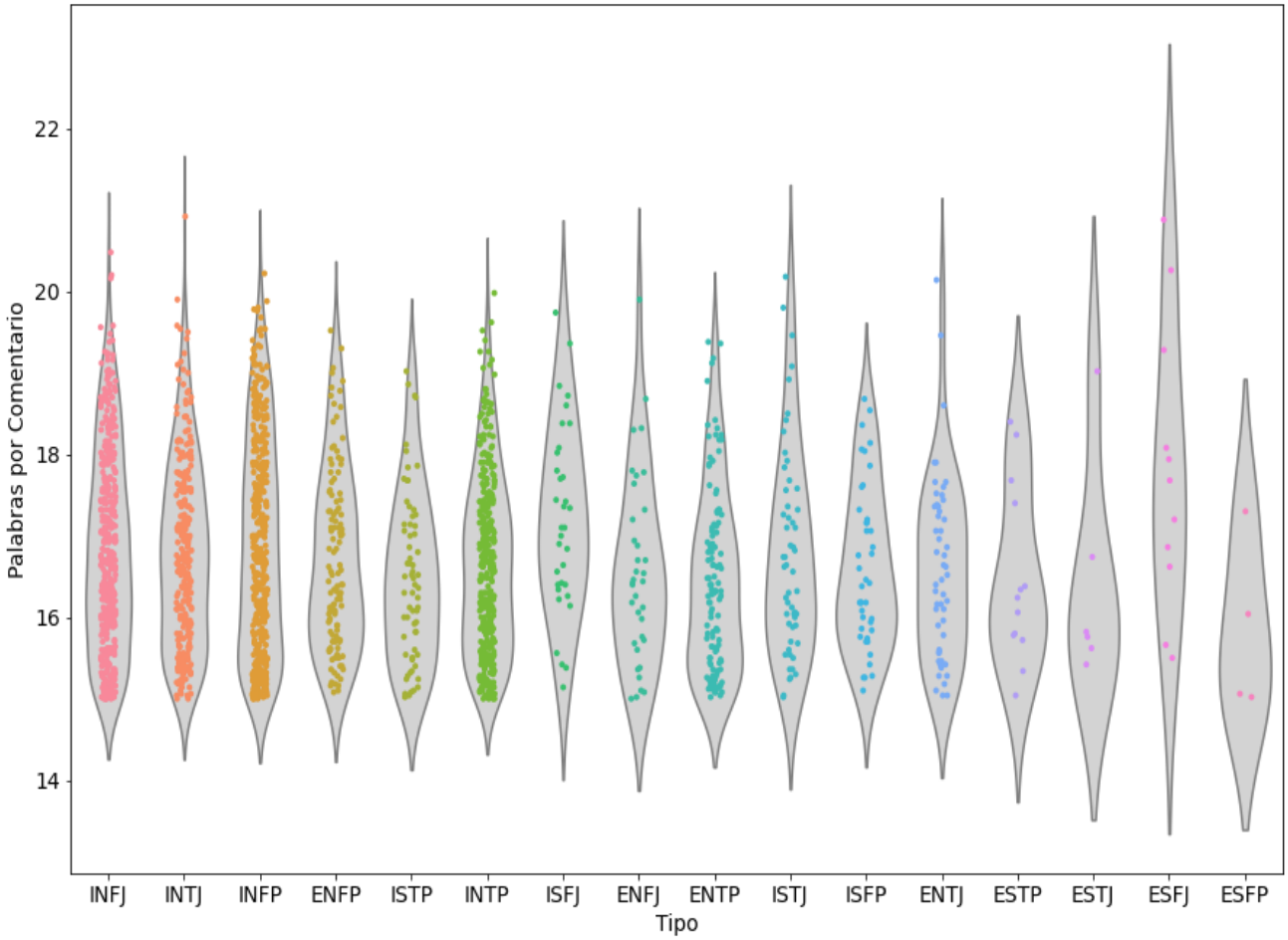

Figura 17 – Diagrama de Violines mostrando la cantidad de palabras por comentario según los diferentes tratos de personalidad.

(Elaboración propia)

En primer lugar, es preciso recordar que, en el dataset, únicamente se ha tenido en cuenta las instancias que contaban con un mínimo de 7 palabras una vez procesado el texto de los

comentarios. Teniendo esto en cuenta, en la figura 17 se puede ver como las clases que con más frecuencia se repiten son INFJ, INTJ, INFP e INTP.

Estos resultados van en la línea de lo que ya comentado en el primer análisis y que se puede revisar en la Figura 3. Del mismo modo, y, al igual, en la misma línea de los resultados mostrados anteriormente en la Figura 3, se puede comprobar que las clases ESTP, ESTJ y ESTP cuentan con un número de instancias muy inferior al resto de clases. Por ello, en este punto, y como se verá más adelante, se va a empezar a introducir la cuestión de si es correcto considerar las clases del modelo como el conjunto de las cuatro dimensiones, obteniendo con ello un conjunto de datos notablemente desbalanceado en cuanto al número de instancias pertenecientes a cada una de las clases.

La otra opción consiste en dividir el problema en cuatro problemas de clasificación binaria y encontrar una solución de clasificación binaria con cuatro dimensiones, ya que esto generaría instancias más balanceadas. Como se ha dicho, más adelante se profundizará en este aspecto tanto en el apartado del modelo enfocado al modelo de machine learning, como en los futuros apartados de resultados y discusión de estos.

Volviendo a lo mostrado en la Figura 15, se puede apreciar cómo, en general no hay grandes diferencias entre las distintas clases, si bien es cierto que al profundizar en el gráfico sí podemos encontrar detalles que llaman la atención. Al observar la figura correspondiente a la personalidad ESFJ, se puede apreciar como la distribución de las diferentes instancias tiene un rango más elevado al resto de personalidades, sin embargo, esta personalidad cuenta con pocas instancias en comparación con otras, por lo que los resultados no llegan a ser concluyentes. Del mismo modo, se observa que la personalidad ESFP concentra el número de comentarios de sus ejemplos en cantidades menores que en el resto de las personalidades evaluadas. Esto, es tenido en cuenta junto al hecho de que, se cuenta con pocas instancias de este tipo de personalidad en la muestra aleatoria con que se ha realizado el gráfico.

De cara a poder evaluar de manera más concluyente los resultados con respecto al número de palabras por comentario y tipo de personalidad, se evaluará el número de comentarios por pares en cada una de las dimensiones. Esto podrá iluminar el análisis con respecto al número de palabras utilizadas para cada una de las clases de las diferentes dimensiones que evalúa el indicador Myers-Briggs (1944).

Esta aportación será interesante para poder tener en cuenta de una u otra manera la variable número de palabras por comentario, la cual podrá complementar a la evaluada anteriormente y que arroja información sobre el contenido de cada uno de los comentarios tras ser pre-procesado el texto.

En primer lugar, como podemos observar en la Figura 18 la cual muestra las distribuciones de las instancias en la primera dimensión del indicador Myers-Briggs, la cual evalúa si un individuo tiene una predisposición más introvertida o extravertida, se puede observar, que, en contra de lo que se podría suponer, los comentarios de personalidades introvertidas tienen una media de palabras ligeramente mayor (16.78) a los comentarios realizados por individuos con un tipo de personalidad extravertida en esta dimensión (16.60). Llama también la atención como los comentarios de individuos con una personalidad de tipo extrovertido en esta dimensión se distribuyen de manera menos central que en la clase introvertida.

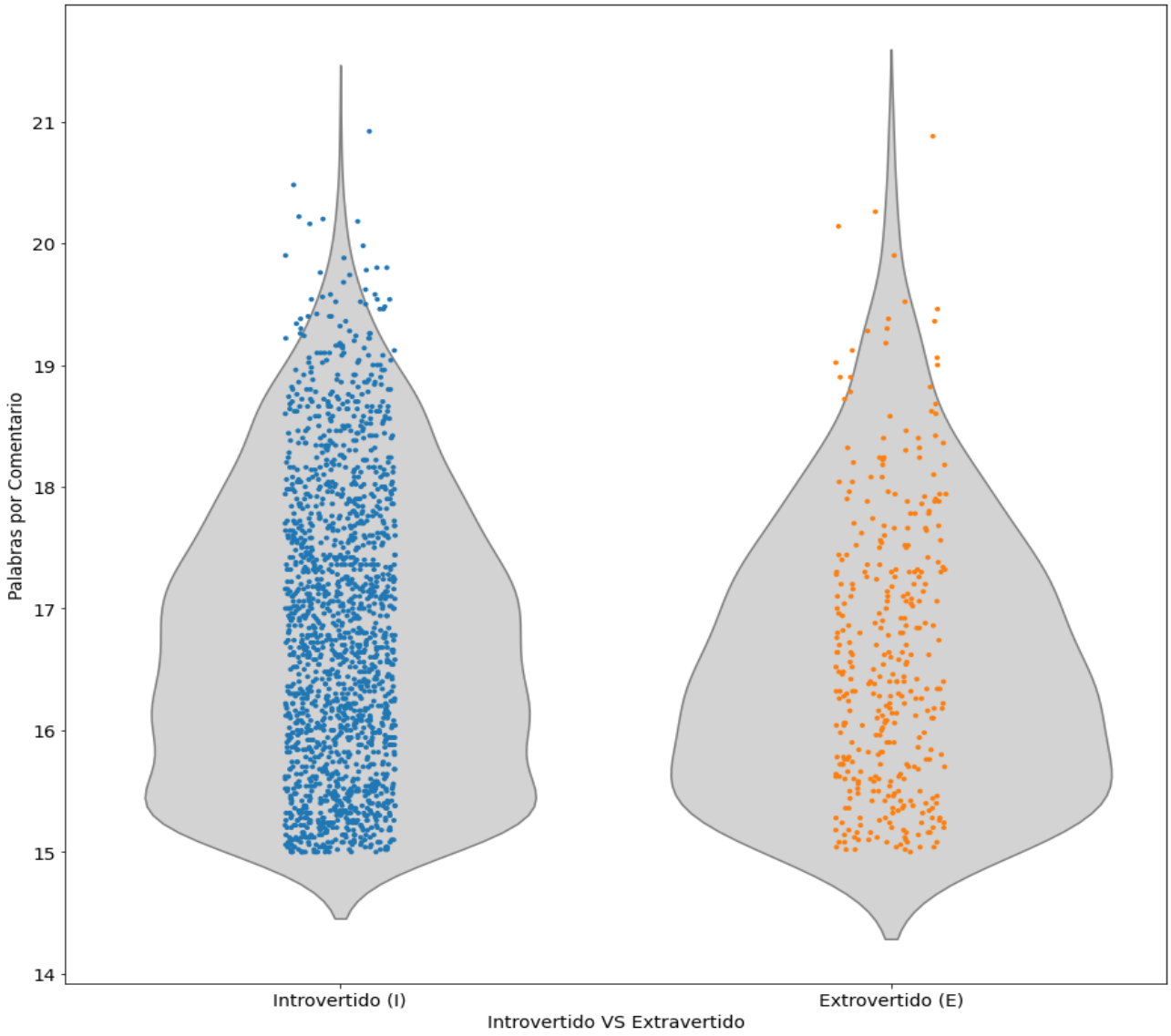

Figura 18 – Diagrama de Violines mostrando la cantidad de palabras por comentario en la dimensión Introvertido-Extravertido.

(Elaboración propia)

En la Figura 19 podemos observar las distribuciones del número de palabras por comentario según la dimensión que valora si el individuo muestra una disposición intuitiva o sensitiva. En esta dimensión no se observan diferencias notables ni en cuanto a la distribución ni al número de palabras para uno u otro modo de la dicotomía. Los resultados de las medias de palabras tanto para los individuos intuitivos con una media de 16.75 palabras por comentario, como para las sensitivas con 16.73 son similares.

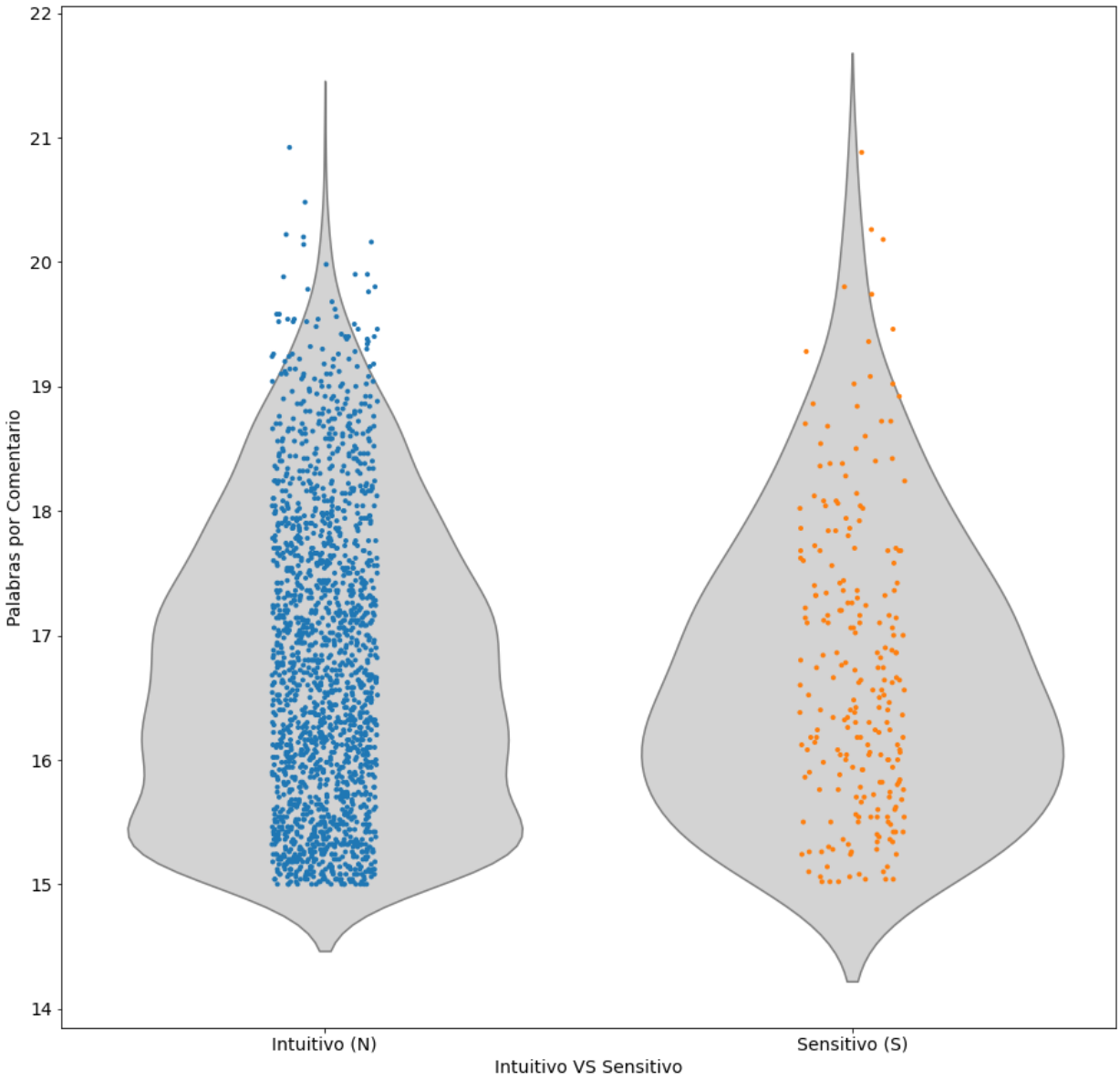

Figura 19 – Diagrama de Violines mostrando la cantidad de palabras por comentario en la dimensión Intuitivo-Sensitivo

(Elaboración propia)

Con respecto a la dimensión que evalúa si el individuo tiene un carácter emocional o pensativo, encontramos ligeras diferencias en cuanto al número de palabras por comentario. En primer lugar, la media de palabras por comentario de las instancias clasificadas como emocional es de 16.84, un valor algo mayor que para las instancias clasificadas con carácter pensativo 16.65. En la figura 20 podemos apreciar como el número de instancias con un número de palabras por comentario mayor a 18 abunda en la clase emocional mientras que en los individuos con carácter pensativo se aglomeran los comentarios en torno a las 15 y 17 palabras.

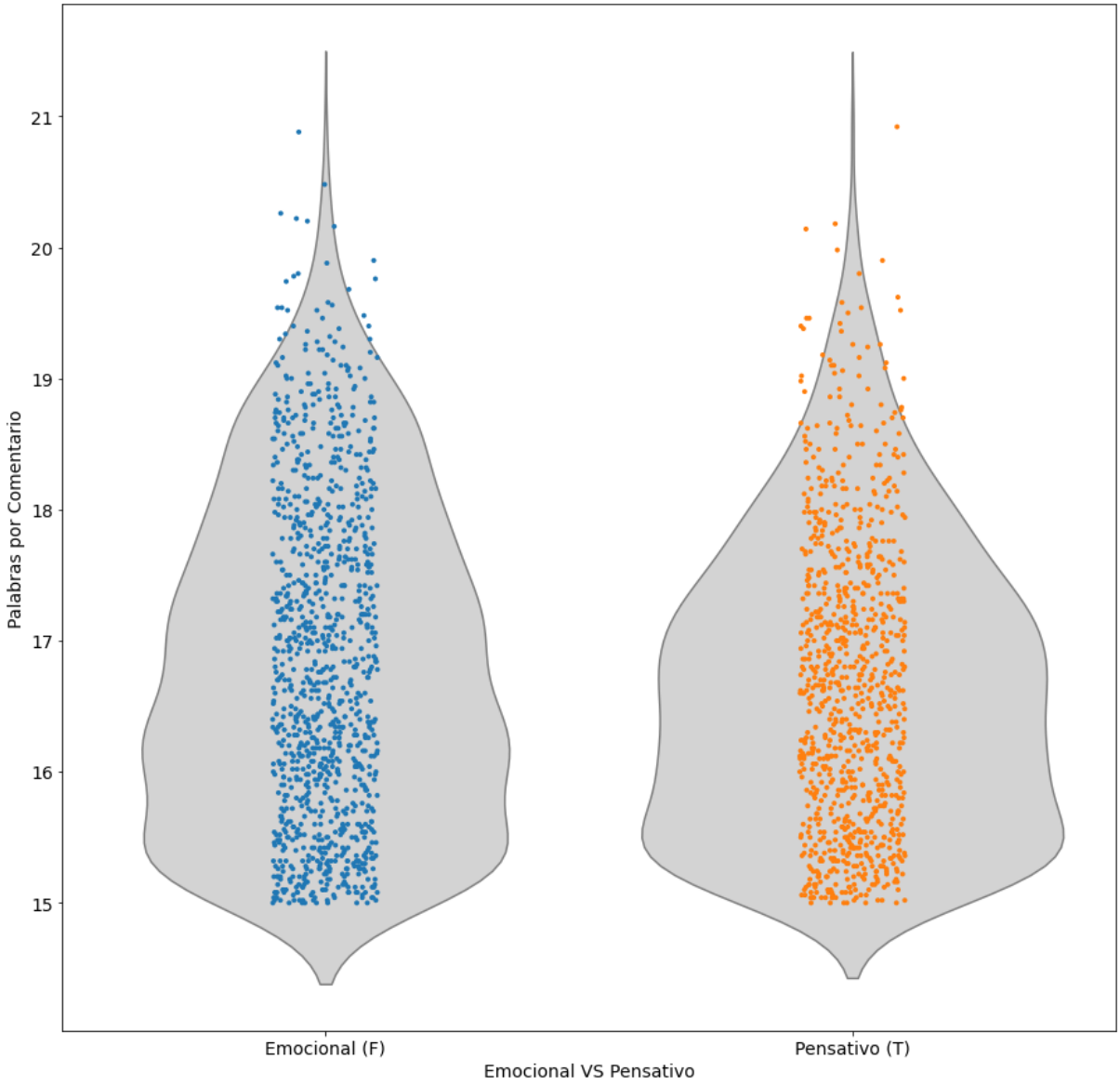

Figura 20 – Diagrama de Violines mostrando la cantidad de palabras por comentario en la dimensión Emocional-Pensativo.

(Elaboración propia)

Por último, al evaluar la última dimensión que clasifica el carácter del individuo como juez o perceptivo, en la figura 21 encontramos que la distribución para la personalidad de tipo perceptivo aglomera la mayoría de sus instancias en la parte baja de la forma en torno a las 15 o 16 palabras por comentario. Por el contrario, las instancias de tipo juez, tienen la mayoría de sus instancias distribuidas entre los 16 y los 17 comentarios. Con respecto a los máximos, es interesante observar que, aunque la cola de las personalidades de tipo juez se alarga más, la densidad de instancias con más de 18 palabras por comentario es mayor en las personalidades de carácter perceptivo.

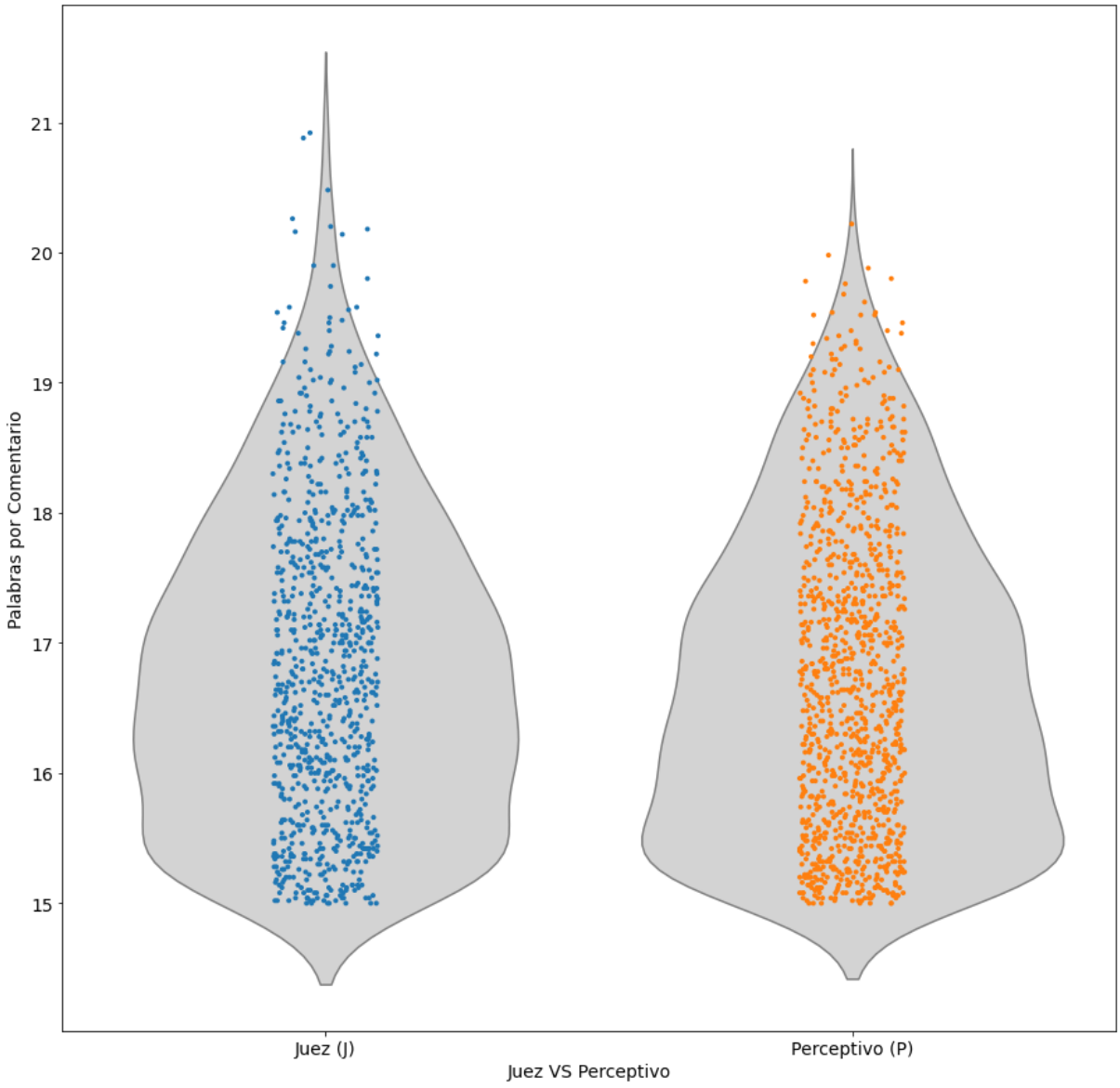

Figura 21 – Diagrama de Violines mostrando la cantidad de palabras por comentario en la dimensión Juez-Perceptivo.

(Elaboración propia)

Una vez se ha tomado nota de las fases previas de análisis exploratorio se puede llevar a cabo la etapa del diseño del modelo de Machine Learning. Para ello vamos a partir del dataset resultante tras la etapa de pre-procesamiento y análisis exploratorio del dataset. En este punto cabe hacer una descripción de la estructura de datos que usaremos a lo largo del modelo:

Tabla 5 – Columnas del Dataset preparado para el modelo de aprendizaje.

Nombre de la Columna	Información que contiene	Tipo de dato
Personalidad	El tipo de personalidad completo del individuo formado por la combinación de una de las palabras, por ejemplo “INTJ”.	Cadena de Texto
Sub_Personalidad_1	El tipo de personalidad relativo a la primera dimensión del indicador Myers-Briggs. (Extravertido o Introvertido)	Numero entero (0 o 1)
Sub_Personalidad_2	El tipo de personalidad relativo a la segunda dimensión del indicador Myers-Briggs (Intuitivo o Sensitivo)	Numero entero (0 o 1)
Sub_Personalidad_3	El tipo de personalidad relativo a la tercera dimensión del indicador Myers-Briggs (Emocional o Pensativo)	Numero entero (0 o 1)
Sub_Personalidad_4	El tipo de personalidad relativo a la cuarta dimensión del indicador Myers-Briggs (Juez o Perceptivo).	Numero entero (0 o 1)
Post_Limpio	Lista de palabras del comentario realizado una vez aplicado el procesamiento del texto comentado anteriormente	Numero entero (0 o 1)
Número_de_Palabras	Número de palabras en el comentario una vez aplicado el procesamiento del texto mencionado anteriormente.	Entero

Elaboración Propia

Como dijimos anteriormente, en este punto del trabajo, nuestro dataset cuenta con 997486 y de aquí en adelante trabajaremos con todas ellas. Como se mencionó anteriormente, la resolución del problema estaba comprometida a decidir si se afrontaba con un problema de clasificación multiclase en el cual usaríamos con las clases los tratos de personalidad completos con las 4 dimensiones contenidos en la columna personalidad o si se afrontaba como cuatro problemas de clasificación binaria para cada una de las dimensiones que contempla el indicador. Finalmente, y tras el estudio de ambas vías se decidió profundizar en cada una de ellas para tomar conclusiones basadas en los resultados de las diferentes métricas obtenidas.

Para poder utilizar la columna “Post_Limpio” como variable debemos transformar cada conjunto de palabras en un vector que represente la ocurrencia o no ocurrencia de una palabra en un post, y el número de veces que ocurre la misma. Por ello y de cara a poder hacer el texto reconocible por los futuros algoritmos de machine learning que utilizaremos, se ha aplicado el countVectorizer de la librería de Python Scikit-Learn. Esta transformación aplica las dos etapas que se comentaban anteriormente. En primer lugar, pasa las palabras a un vector el cual representa la frecuencia con que ocurre cada una de las palabras. Esto generará multitud de atributos en el conjunto de datos basado en las palabras que aparecen en los comentarios preprocesador de los usuarios. Los métodos transformación del countVectorizer se han realizado sobre la columna Post_Limpio para convertir esta columna en una nueva matriz de frecuencia de palabras. Con esto ya contamos con un vector el cual ha almacenado cada palabra y su frecuencia de término para cada comentario. A continuación, se ha aplicado el algoritmo TF-IDVectorizer de la librería Scikit-Learn, el cual nos proporcionará, a partir del vector generado con countVectorizer, el peso de una palabra con respecto a las veces que aparece por comentario y con respecto al resto de comentarios. Estos pesos serán tenidos en cuenta por el algoritmo a la hora de identificar la importancia de cada atributo.

Como fue descrito en apartados anteriores del trabajo, se han codificado las clases a través de la etapa de pre-procesamiento MultiLabelBinarizer de la librería de Python Scikit-Learn. En esta fase, estamos considerando las clases como las contenidas en las columnas “Sub_Personalidad_1”, “Sub_Personalidad_2”, “Sub_Personalidad_3” y “Sub_Personalidad_4”. Cada una de esas columnas nos generará un problema de clasificación diferente en el que las clases serán las dicotomías correspondientes a esa dimensión de la personalidad. Con MultilabelBinarizer hemos codificado cada una de esas columnas dos nuevas columnas que indican, la dicotomía, mediante el nombre de la columna y la pertenencia, mediante el valor de la instancia en esa columna con un 1 (Pertenencia) o 0 (No Pertenencia) tal y como se mostró en la tabla 4.

También es importante solucionar los problemas correspondientes al balanceo del dataset. Al contar con un nuevo de instancias suficiente. Para cada uno de los dos modelos se llevará a cabo una técnica de balanceo diferente, esto es debido al número de instancias con menor representación en cada uno de los casos. En primer lugar, para el modelo en el cual vamos a tener en cuenta las dimensiones para llevar a cabo cuatro modelos de clasificación binaria vamos a realizar el balanceo mediante la técnica de under-sampling, la cual consiste en igualar el número de clases existentes seleccionando una muestra aleatoria de instancias en la clase dominante de igual tamaño al tamaño del número de instancias de la con un menor número de ocurrencias. Como se pudo ver en el análisis exploratorio, la división de instancias según su dimensión nos va a permitir realizar una técnica de balanceo basado en under-sampling. La dimensión más afectada es la correspondiente a la que categoriza el carácter en uno intuitivo o sensitivo, sin embargo, seguimos contando con un dataset de 95758 instancias, lo cual se puede considerar como un número suficiente de instancias para llevar que ingieran los diferentes algoritmos.

Con respecto a la resolución del problema enfocada a la clasificación de los 16 tipos de personalidad completos, no se considera una buena opción utilizar una única técnica de balanceo basado en el under-sampling. Esta decisión se toma debido a lo restrictiva que resulta la clase ESFJ, la cual, se encuentra representada únicamente por 1425 instancias en este punto del trabajo. Debido a esto, se decide utilizar una técnica combinada de over-sampling y under-sampling. Para ello, vamos a tomar como referencia la clase

En este punto nuestros datos están preparados para ser ingeridos por el algoritmo. Como se ha mencionado, este problema se podría definir como cuatro clasificaciones binarias. Para solucionar la clasificación, por tanto, hemos escogido el algoritmo de regresión logística de la librería Scikit-Learn. ENTJ de la cual, contamos con 12536 instancias. Esta clase se considera que cuenta con un número suficiente de ocurrencias como para que el algoritmo consiga establecer cuales son los atributos diferenciadores que la conforman. Para llevar a cabo el balanceo habiendo tomado esa referencia, en primer lugar, vamos a tomar muestras aleatorias del resto de clases que cuentan con un número de instancias superior (INTP, INTJ, ENTP, INFJ, INFP, ENFP, ISTP, ESTP, y ENTP). Una vez hemos igualado las instancias de esas clases al número de instancias de la clase ENTJ mediante la creación de muestras aleatorias de 12536 instancias se va a continuar realizando la técnica de over-sampling para las muestras que cuentan con una representación de instancias inferior (ISTJ, ISFP, ISFJ, ESFO, ESTJ y ESFJ). Con esto ya se cuenta con los conjuntos de datos balanceados tanto para resolver el problema de clasificación multiclase y de clasificación binaria por dimensiones.

Con ello hemos preparado al completo los conjuntos de datos que van a ser utilizados para llevar a cabo las tareas de entrenamiento de los modelos y futura validación de estos. Para ello, se dividirá el dataset en dos conjuntos de entrenamiento y validación. Para ello se han creado dos muestras aleatorias a partir del dataset que se tuvo en primer momento, la primera de ellas, la cual usaremos para el entrenamiento contará con un 80 por ciento de las instancias elegidas aleatoriamente. El segundo conjunto, el cual se reservará para la validación, cuenta con el 20 por ciento restante.

Con esto se tiene todo lo necesario para poder probar el modelo a través de los diferentes algoritmos escogidos. Para exponer los resultados de una manera más ordenada, se va a mostrar los resultados por separado, para los modelos basados en una clasificación multiclase y para los basados en cuatro clases binarias.

Algoritmos utilizados y resultados obtenidos en los modelos de clasificación multiclase.

Tabla 6 – Resultados como modelo multiclase sobre los 16 tipos de personalidad.

Algoritmo	Precisión (Accuracy)
CatBoost Classifier	0.688
SVC (Support Vector Classifier)	0.652
XGBoost Classifier	0.636
Logistic Regression	0.599
Random Forest	0.480
Multinomial Naive Bayes	0.427

Elaboración Propia

En la tabla 6 se pueden observar las métricas obtenidas en los diferentes modelos a la hora de resolver el problema de clasificación multiclase. Los algoritmos utilizados han sido algunos de los más utilizados para llevar a cabo modelos de clasificación. La métrica utilizada ha sido la precission “Accuracy” la cual es una métrica adecuada teniendo en cuenta que el dataset ha sido correctamente balanceado. Se pude observar que la precisión oscila entre el 0.427 obtenido por el modelo basado en el algoritmo multinomial Naive Bayes y el máximo de 0.688 de rendimiento obtenido por el modelo basado en el algoritmo support vector classifier de la librería Scikit Learn.

Entre medias encontramos los resultados obtenidos por los modelos basados en algoritmos basados en Support Vector Classifier (0.652), en el algoritmo de Scikit learn XGBoost, basado en Gradient Boosting (0.636), el generado por el modelo de Regresión Logística (0.599) y el

modelo de ensemble learning generado a través del algoritmo Random Forest de la librería de Python de Scikit Learn (0.48).

Estos resultados han obtenido menos rendimiento y permiten un menor rango de mejora que los obtenidos por los modelos basados en las cuatro dimensiones. En estos modelos, se recuerda que vamos a abordar el problema como un problema de clasificación diferente para cada una de las dimensiones. Para ello se probará, del mismo modo que en los resultados del modelo multiclase, el modelo a partir de diferentes algoritmos los cuales pueden consultarse de manera más visual en la tabla 7.

Algoritmos utilizados y resultados obtenidos en los modelos de clasificación basado en clasificaciones binarias.

Tabla 7 – Resultados (Accuracy) de los modelos basados en dimensiones.

Algoritmo	Parámetros	IE	NS	TF	JP
GMXBoost Classifier	n_estimators = 200 max_depth = 2 nthread = 8 learning_rate = 0.2	0.770	**0.931**	0.689	0.583
CatBoost Classifier	loss_function=MultiClass task_type='GPU' verbose=False	0.771	0.924	**0.694**	0.579
Linear Regression Classifier	max_iter=3000 C=0.5 n_jobs=-1	**0.771**	0.925	0.692	**0.588**
Random Forest	max_depth=2	**0.771**	0.925	0.659	0.563

Elaboración Propia

Para comenzar, se puntualizará que se ha utilizado, al igual que en los algoritmos de los modelos multiclase, la precisión (Accuracy) como métrica de evaluación. Al observar estos resultados y de un simple vistazo, ya se puede observar que el rendimiento de los modelos basados en la clasificación por dimensiones ha tenido un mayor desempeño que los basados en la clasificación multiclase. Es interesante el apreciar como ningún algoritmo ha obtenido métricas superiores al resto en todas las dimensiones, sino que cada uno de ellos ha obtenido métricas superiores al resto en una dimensión específica. Comenzando por de arriba abajo analizaremos cada uno de los algoritmos.

En primer lugar, el algoritmo GMXBoost fue escogido debido al buen rendimiento que ha tenido en modelos expuestos durante la investigación del estado del arte. Para poder optimizar los resultados se ha recorrido el dataset con este algoritmo utilizando la técnica de grid_search que dé con unos mejores resultados utilizando un método de Kfold con 10 splits. Esta hiperparametrización ha dado con los parámetros que mejor rendimiento ofrecen al algoritmo los cuales pueden observarse en la columna parámetros de la tabla 7. Con ello el modelo ha clasificado con una precisión del 93,1 por ciento las instancias en la dimensión que caracteriza la personalidad de intuitiva o sensitiva. Este resultado ha sido satisfactorio y es el de mayor rendimiento obtenido en todo el modelo, incluso superando a los obtenidos en los modelos evaluados en el estado del arte.

Aunque esta dimensión también ha tenido un buen desempeño en otros algoritmos, no se ha mejorado esta métrica. Sin embargo, el algoritmo CatBoost Classifier de la librería Scikit-Learn de Python ha obtenido unos resultados superiores en la dimensión Pensamiento-Emoción, siendo estos los de mayor rendimiento frente al resto de algoritmos. Aunque esta métrica no ha sido tan buena como la lograda en la dimensión Intuición-Sensación, ha tenido un buen desempeño, siendo capaz de clasificar correctamente el 69,4 por ciento de las instancias.

Los modelos basados en los algoritmos RandomForest y Linear Regression han tenido ambos una precisión similar del 0.771 en la dimensión Introversión-Extraversión. Estos resultados son los mejores obtenidos por todos los modelos de clasificación basados en las dimensiones. Del mismo modo, el modelo basado en el algoritmo de regresión lineal ha tenido la mejor métrica de precisión al clasificar correctamente el 58,8 por ciento de las instancias de la dimensión Juez-Perceptivo. Esta métrica no se considera como buena teniendo en cuenta que el modelo está clasificando instancias con únicamente dos clases y deberá ser un punto para abordar en futuras líneas de trabajo.

Con estos resultados se puede concluir como las diferentes dimensiones tienen atributos de importancias diferentes y, como se vio durante la fase de análisis apenas han mostrado relación entre sí una vez se ha llevado a cabo la tarea de clasificación a partir de los algoritmos.

Por último, y de cara a iluminar el modo en que los diferentes algoritmos se han dado importancia a los atributos, o, como en este caso, a las diferentes palabras del modelo, se ha analizado el resultado de las importancias que han recibido las palabras que más han influido a la hora de clasificar las instancias. Para ello se han extraído el peso en importancia para cada palabra en cada uno de los modelos de clasificación basados en dimensiones y se ha realizado la media para cada una de las palabras. Los resultados, que serán expuestos a continuación pueden observarse pintados en un gráfico de barras horizontales en la figura 22.

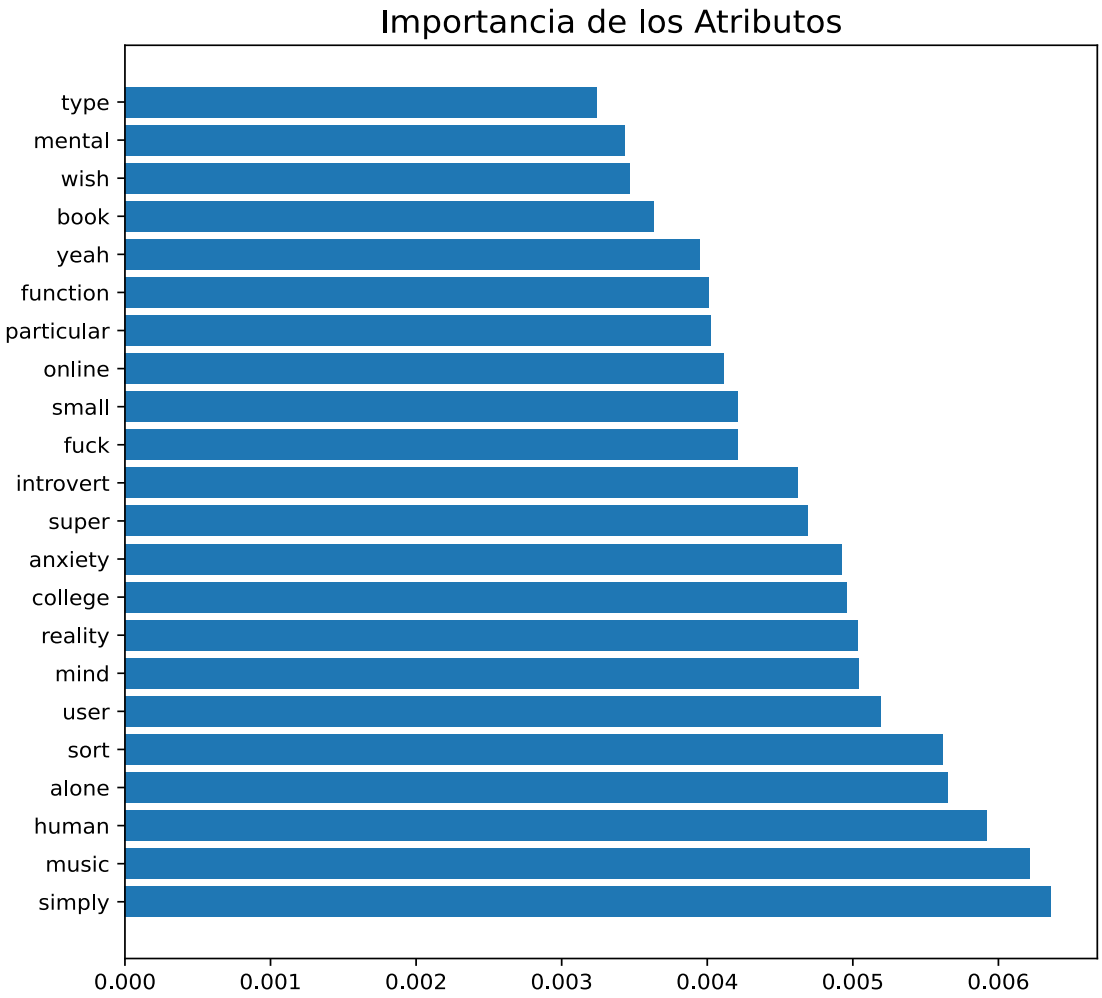

Figura 22 – Gráfico de barras horizontales representando la importancia de los atributos baremada por los diferentes modelos.

(Elaboración propia)

Al analizar la importancia de las palabras con mayor peso para los algoritmos nos encontramos que las palabras que mayor frecuencia tenían al analizar las nubes de palabras no figuran como palabras de importancia para la tarea de clasificación de los algoritmos a excepción de la palabra “type” (Tipo). Esto es interesante ya que rompe con algunas conclusiones iniciales acerca de como los verbos se relacionaban con las diferentes dimensiones de la personalidad. Sin embargo, los únicos verbos que han sido de importancia a la hora de clasificar las diferentes dimensiones de la personalidad han sido “mind” (Importar) “wish” (Desear). Aunque estos verbos fueron expuestos en las nubes de palabras, no formaban parte de los usados con mayor frecuencia y por ello, parecen haber tenido un carácter diferenciador para los diferentes algoritmos. En línea con lo anterior, si se puede

apreciar como algunas de las palabras de mayor importancia tienen un carácter coloquial como es el caso de “yeah”, “super” o “fuck” como ya se apuntaba que podrían ser un rasgo diferenciador en la manera de expresarse de los usuarios de diferentes personalidades.

5. Conclusiones y trabajo futuro

5.1. Conclusiones

En el trabajo se ha tratado de diseñar un modelo que permita predecir la personalidad de un sujeto a partir de textos escritos por esa persona. La motivación del trabajo era la de poder, en un primer lugar, abaratar el proceso de evaluación de la personalidad, y, en segundo lugar, el generar un modelo que fuera capaz de identificar las diferentes dimensiones de la personalidad de un individuo y, aunque no haya sido posible dar con un modelo de predicción capaz de sustituir la validez de los métodos tradicionales, se han realizado interesantes avances.

Como fue expuesto, numerosos autores se han interesado en esta línea de trabajo y han propuesto sus diferentes modelos, los cuales contaban con métricas que mostraban buen rendimiento en la tarea de clasificación de diferentes tipos de personalidad a partir de texto escrito por usuarios en redes sociales y portales web. De entre estos modelos fueron mencionados dos que tenían una estrecha relación con el indicador Myers-Briggs (1944). Estos son los modelos elaborados por Kosimin (2011) y Alam Sher Khan (2020).

Estos modelos obtuvieron un buen rendimiento en la tarea de clasificación basada en las dimensiones del indicador Myers-Briggs, siendo de entre los modelos de especial interés el diseñado por Alam Sher Khan (2020), el cual, utilizando gradient boosting como algoritmo, alcanzó cifras entre el 0.85 y 0.9. Aunque el modelo de este trabajo no ha conseguido alcanzar un rendimiento general superior a ese modelo, sí que se ha conseguido superar el rendimiento del modelo de Kosimin (2011), el cual contaba con una precisión del 0.70. Con respecto al modelo de Alam Sher Khan (2020) se puede identificar una mejora significativa en la dimensión Intuitivo-Sensitivo, para la cual su modelo tuvo un rendimiento del 0.89 mientras que en el presente se ha obtenido una métrica de precisión del 0.931 a partir del modelo que ha sido diseñado con el algoritmo XGBoost y con la parametrización que se definió durante la fase de resultados.

Esto es un avance que se debe tener en cuenta ya que parece que este modelo se comporta especialmente bien en esta dimensión. Estos resultados iluminan la posibilidad de mejorar

modelos en los cuales esta dimensión no había sido clasificada con tanta precisión. La mejora del modelo de este trabajo con respecto a esa dimensión de la personalidad puede apreciarse no solo en los resultados del modelo de gradient boosting sino también con el que se evaluó de regresión lineal.

Se puede concluir que los modelos con un mayor rendimiento han sido los que han clasificado independientemente cada una de las dimensiones. Esto genera pensar que igualmente, las etapas anteriores de procesamiento y exploración del lenguaje deben ser de gran importancia a la hora de trabajar con cada una de las dimensiones. Es posible que el haber tratado los datos desde un comienzo de una manera específica para cada una de las dimensiones hubiera mejorado el rendimiento de las métricas para las diferentes categorías de la personalidad que se evalúan en el indicador Myers-Briggs.

En este caso se realizó un modelo total para poder concluir la mejor solución a la hora de afrontar estos problemas, y, gracias a ello se ha dado con un punto de gran importancia a la hora de poder revisitar puntos anteriores del trabajo y mejorar los resultados obtenidos. De la misma manera se puede concluir que el éxito a la hora de llevar a cabo la predicción de la personalidad reside en buscar una solución que sea flexible y no cerrarse a una sola línea. Explicando esto, ya se ha profundizado en como el concepto de la personalidad supone un reto a la hora de ser abordado, y esto se ha podido apreciar también al observar las palabras que los modelos han tomado como más importante. Al mismo tiempo que el contexto cambia, el modo de expresarse en redes sociales también lo hace, por lo que ningún modelo basado en procesamiento del lenguaje natural que pretenda predecir la personalidad estará actualizado si no se ha reformulado de nuevo recientemente.

Por otra parte, se ha dado con la importancia de parametrizar correctamente los algoritmos al llevar a cabo el proceso de entrenamiento y validación del modelo. Esto deberá ser tenido en cuenta en futuros trabajos si se quieren alcanzar las métricas más altas posible en el rendimiento de los algoritmos y es algo que también ha sido de gran importancia y un punto de exposición en trabajos anteriores.

5.2. Líneas de trabajo futuro

Con respecto a las líneas del trabajo que el presente modelo ilumina, se puede concluir en primer lugar que la predicción de la personalidad debe de ser planteada analizando con cautela las diferentes dimensiones que el enfoque teórico escogido haya definido. En el caso del indicador Myers-Briggs deben se puede trabajar en mejorar los resultados de manera

independiente en cada una de las dimensiones. El proceso de un futuro modelo debe tener una interacción muy elevada entre etapas, teniendo que evaluar las palabras más importantes que ha escogido un modelo y volviendo atrás para volver a analizar y procesar las instancias en relación con los resultados obtenidos. Este proceso interactivo debería ser el punto clave para abordar estos modelos de predicción ya que tanto el modo de expresarse de los usuarios en redes sociales como el propio concepto de la personalidad y sus dimensiones son variable en el tiempo y pueden dar pie a cambios relevantes a la hora de estandarizar y baremar las palabras.

La formulación de un léxicon especializado en esta tarea y que permita la predicción de la personalidad en otros idiomas como el castellano también resultarían tareas de gran interés en el futuro. En esta línea no se han encontrado estudios que lleven a cabo tareas similares con resultados relevantes.

Todos estos resultados pueden dar pie a diversas aplicaciones de gran interés tanto a nivel personal como a nivel empresarial y corporativo. Al tener un modelo de predicción que ayude a clasificar las personalidades se podrían facilitar o mejorar tareas de reclutamiento en las empresas, ya que, como ha sido expuesto durante el trabajo, los diferentes caracteres de la personalidad tienen una alta relación con el desempeño de diferentes tareas. Un interesante avance sería el de poder incluir modelos de predicción de la personalidad en herramientas como LinkedIn en las cuales, además de contar con publicaciones realizadas por los usuarios, se puede extraer información relativa al puesto de trabajo, a la disposición de un sujeto a mantenerse en un empleo o a cambiar con más frecuencia. Algunas de las dimensiones evaluadas como la extraversión se relacionan estrechamente con estos conceptos, así como con puestos de trabajo como puede ser el de un comercial. La intuición y la percepción también son dimensiones que se han relacionado con la toma de decisiones en trabajos expuestos en el estudio y esto también sería información interesante y útil si se pudiera predecir en este tipo de entornos.

Por último y en relación con el campo de la psicología, sería interesante el poder evaluar la personalidad de manera dinámica en el tiempo. Como ya se ha expuesto la personalidad es un concepto dinámico el cual actualmente se evalúa de manera puntual y bajo costes altos de tiempo. Con un modelo capaz de predecir la personalidad a partir del texto escrito se podían abrir vías a una nueva forma de evaluación de la personalidad y con esto definir nuevos enfoques teóricos que sean capaces de definir de una manera más científica este difícil concepto.

6. Bibliografía

Adrian Furnham, Paul Stringfield, Personality and work performance: Myers-Briggs type indicator correlates of managerial performance in two cultures, Personality and Individual Differences, Volume 14, Issue 1, 1993, Pages 145-153.

Alam Sher Khan, Hussain Ahmad, Muhammad Zubair Asghar, Furqan Khan, Saddozai, Areeba Arif, and Hassan Ali Khalid. 2020. Personality Classification from Online Text using Machine Learning Approach. International Journal of Advanced Computer Science and Applications 11, 3 (2020).

B Liu. Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, vol.5,no.1, pp.1-167,May 2012.

C. Sumner, A. Byers, R. Boochever, and G. J. Park. 2012. Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In 2012 11th International Conference on Machine Learning and Applications, Vol. 2.

College, M. (1993). The Utility of the Myers-Briggs Type Indicator. Review of Educational Research, 63(4), 467-488

Espina, A., & Figueroa, A. (2017). Why was this asked? automatically recognizing multiple motivations behind community question-answering questions. Expert Systems with Applications, Vol. 80, No. 1, pp. 126-135.

Gjurkovic, M.; Snajder, J. Reddit: A gold mine for personality prediction. In Proceedings of the Second Workshop on Computational Modelling of People's Opinions, Personality and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; pp. 87–97.

Golbeck, J.; Robles, C.; Edmondson, M.; Turner, K. Predicting personality from Twitter. In Proceedings of the IEEE Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011

James W. Pennebaker and Laura A. King. 1999. Lin- guistic styles: Language use as an individual differ- ence. Journal of personality and social psychology 77(6):1296.

James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological aspects of natu- ral language use: Our words, our selves. Annual review of psychology 54(1):547–577.

Jung, C. G. (1928). Psicological types. En Contributions to Analytical Psychology. London: Routledge & Kepan Paul, 295-312.

Komisin, M.; Guinn, C. Identifying personality types using document classification methods. In Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, Marco Island, FL, USA, 23–25 May 2012; pp. 232–237.

Kong, E., & Farrell, M. (2012). Facilitating knowledge and learning capabilities through neuro-linguistic programming. International Journal of Learning, 18(3).

Kong, E. (2012). The potential of neuro-linguistic programming in human capital development. Electronic Journal of Knowledge Management, 10 (2), pp.131-141.

Myers, I. B. (1962). The Myers-Briggs type indicator manual. Princeton, NJ: The educational Testing Service.

Myers, I. B., McCaulley, M. H., Quenk, N. L. y Hammer, A. L. (1998). MBTI Manual (A guide to the development and use of the Myers Briggs type indicator).

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. IEEE Intelligent Systems, Vol. 32, No. 2, pp. 74-79.

Punnoose, R.; Ajit, P. Prediction of employee turnover in organisations using machine learning algorithms, A case for Extreme Gradient Boosting. Int. J. Adv. Res. Artif. Intell. 2016, 5, 22–26.

W. Jin and H.H. Ho, “A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining”, Proc. 26th Ann. Int’l Conf. Machine Learning, pp. 465-472, 2009.

Wilde, D. (2011). Psychological teamology, emotional engineering and the Myers-Briggs Tipe Indicator. In S. Fukuda (Ed.), Emotional Engineering. New York: Springer London, 365-374.