UNIVERSIDAD DE HUELVA
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

Depto. Ingeniería de la Información y el Conocimiento

Ciencias de la Computación

TRABAJO FIN DE GRADO

Técnicas de *Deep Learning* para la Detección de Valores Humanos en Argumentos

Autor: Nordin El Balima Cordero

Tutor(a): Jacinto Mata Vázquez

Co-tutor(a): Victoria Pachón Álvarez

julio, 2023

Técnicas de Deep Learning para la Detección de Valores Humanos en Argumentos

A mi abuela y a madre
Gracias por estar en cada momento

TRABAJO FIN DE GRADO - ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

Técnicas de Deep Learning para la Detección de Valores Humanos en Argumentos

Nordin El Balima Cordero
Huelva, julio 2023

Resumen

Los valores humanos se refieren a los principios, creencias y convicciones que las personas consideran importantes y fundamentales en sus vidas. Representan las cualidades y características que guían el comportamiento y las decisiones de los individuos, y reflejan su sentido de lo que es correcto, valioso y significativo en la vida. Los valores humanos son fundamentales para la identidad y la cultura de una sociedad, ya que influyen en la forma en que las personas interactúan entre sí y con el mundo que les rodea. La comprensión de los valores humanos es importante porque nos permite entender mejor a las personas, sus creencias, actitudes y motivaciones.

El objetivo principal de este trabajo es el estudio e implementación de técnicas para la detección automáticas de valores humanos en un argumento, dado por una premisa, una conclusión y una instancia que muestra la postura (a favor o en contra) entre estas. Este dataset es una gran opción para detectar valores humanos pues proporciona una variedad de argumentos que abordan diferentes temas y perspectivas. Para abordar el desafío de detectar los valores humanos en los argumentos, hemos utilizado enfoques basados en modelos preentrenados como BERT y RoBERTa, que son arquitecturas avanzadas de transformers. Estos modelos han sido entrenados en grandes cantidades de datos textuales y han demostrado un rendimiento sobresaliente en una amplia gama de tareas de procesamiento del lenguaje natural (PLN). Hemos aprovechado el conocimiento adquirido por estos modelos en el procesamiento y comprensión del lenguaje para nuestra tarea específica de clasificación de valores humanos.

El trabajo comienza contextualizando mediante el marco teórico los conceptos principales de Machine Learning y Deep Learning así como las estrategias utilizadas como el submuestreo (undersampling) para abordar el desequilibrio en la distribución de clases en el conjunto de datos, lo que nos ha permitido mejorar la capacidad de los modelos para detectar valores humanos, el procesado de los argumentos para estandarizar el texto, técnicas de aumento de datos con back-translation, etc.

Para evaluar los resultados se participó en “The 17th International Workshop on Semantic Evaluation”, más concretamente en la tarea “ValueEval: Identification of Human Values behind Arguments”. Una vez aceptada, nuestra propuesta acabó con la publicación de un paper científico en sus actas.

Concluida la competición se han propuesto un conjunto de nuevas propuestas para mejorar los resultados obtenidos, detallando una comparativa con los resultados previos. En resumen, hemos utilizado modelos preentrenados de vanguardia junto con técnicas de procesamiento de datos para abordar el problema de identificar los valores humanos en los argumentos, lo que ha demostrado ser eficaz en la tarea a resolver.

TRABAJO FIN DE GRADO - ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

Deep Learning Techniques for Detecting Human Values in Arguments

Nordin El Balima Cordero
Huelva, julio 2023

Abstract

Human values represent the beliefs and convictions that people consider fundamental in their lives, guiding their behavior and decisions. They are vital for the identity and culture of a society, and their understanding allows a deeper understanding of people, their beliefs, attitudes, and motivations.

The objective of this work is to study and implement techniques for the automatic detection of human values in an argument. The argument consists of a premise, a conclusion, and an instance showing the stance between these. We have chosen a dataset that offers a variety of arguments addressing different topics and perspectives.

We confronted this challenge using approaches based on pre-trained models such as BERT and RoBERTa, which are advanced transformer architectures. These models have been trained on large volumes of textual data and have demonstrated exceptional performance in a variety of natural language processing (NLP) tasks.

We initiated this work by contextualizing the main concepts of Machine Learning and Deep Learning through a theoretical framework. We also employed strategies such as undersampling to address the imbalance in the class distribution in the dataset. This method allowed us to improve the models' ability to detect human values. In addition, we processed the arguments to standardize the text and used data augmentation techniques with back-translation.

To evaluate our results, we participated in "The 17th International Workshop on Semantic Evaluation", in the task "ValueEval: Identification of Human Values behind Arguments". Our proposal was accepted and published in the proceedings of this workshop.

Once the competition concluded, we have proposed a set of new strategies to improve the results obtained and we have carried out a comparison with the previous results.

In summary, we have used advanced pre-trained models along with data processing techniques to address the problem of identifying human values in arguments. This combination has proven to be effective for the task at hand.

Agradecimientos

Quiero tomar un momento para expresar mi más sincero agradecimiento a todos aquellos que han jugado que me han apoyado en mi viaje académico y personal. Primero, a mis estimados tutores Jacinto y Vicky, que han servido de guía indispensable para poder realizar este trabajo.

A mis queridos compañeros de la universidad, compañeros del día a día que han compartido conmigo tanto los momentos de diversión y alegría como aquellos en los que debíamos apoyarnos para estudiar.

Por último y no menos importante, a mi familia por su apoyo constante.

Gracias

Nordin El Balima Cordero
Huelva, 2023

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

1.1. MOTIVACIÓN

Los valores humanos son fundamentales para entender nuestra sociedad, ya que son los principios, creencias y convicciones que guían el comportamiento y las decisiones de los individuos. Con el llegada de la era digital, cada vez más de nuestras interacciones y discursos se llevan a cabo en línea, lo que genera grandes cantidades de texto en lenguaje natural que reflejan nuestras creencias y valores. Sin embargo, la comprensión y detección de estos valores humanos en los textos es un desafío que aún se está explorando y representa un campo de investigación importante en el procesamiento del lenguaje natural (NLP). La capacidad para identificar y entender los valores humanos en los textos puede tener una amplia gama de aplicaciones, desde el análisis de las tendencias sociales hasta la mejora de las interacciones humanas con las tecnologías de la inteligencia artificial.

1.2. OBJETIVO

El objetivo principal de este trabajo es el estudio e implementación de técnicas para la detección automática de valores humanos en un argumento, dado por una premisa, una conclusión y una instancia que muestra la postura (a favor o en contra) entre estas. Para abordar este objetivo, se han implementado técnicas avanzadas de Machine Learning y Deep Learning, utilizando modelos preentrenados de vanguardia como BERT y RoBERTa.

Para evaluar los resultados de las técnicas desarrolladas se ha participado en el "The 17th International Workshop on Semantic Evaluation"[1], más concretamente en "Task 4. ValueEval: Identification of Human Values behind Arguments"[2]. Cuyo trabajo culminó con la presentación de un artículo científico [2] que ha sido aceptado y publicado. [3]

1.3. ANTECEDENTES Y ESTADO DEL ARTE

El procesamiento del lenguaje natural es un campo de la inteligencia artificial que ha experimentado avances significativos en la última década, con el desarrollo de nuevas técnicas y modelos que pueden entender y generar texto en lenguaje natural con un alto nivel de sofisticación. En particular, los modelos basados en transformers como BERT y RoBERTa han demostrado ser particularmente eficaces en una amplia gama de tareas de *NLP*, desde la comprensión de texto hasta la generación del mismo.

El análisis de sentimientos en texto es un campo bastante amplio, encontramos varios trabajos que abordan el desafío desde diferentes enfoques. En particular, aquí observamos como *BERT*[11] permite analizar las noticias sobre la bolsa para poder así automatizar órdenes, órdenes que serían difíciles de tomar por un humano pues los precios fluctúan con rapidez y surjen miles de noticias por minuto. Por otro lado aquí[3] nos muestran como podemos usar *BERT* para extraer sentencias de un artículo y clasificar los valores en ellas.

Estudios como estos no muestran que los modelos preentrenados pueden ser efectivos en la detección de valores humanos en textos argumentativos. Este trabajo se basa en estas investigaciones previas y busca mejorar y expandir sus resultados mediante la implementación de nuevas técnicas y estrategias.

1.4. ESTRUCTURA DEL DOCUMENTO

El resto de este trabajo se organiza de la siguiente manera:

[1] https://semeval.github.io/SemEval2023/
[2] https://touche.webis.de/semeval23/touche23-web/index.html#submission
[3] https://aclanthology.org/volumes/2023.semeval-1/

- En el Capítulo 2, Marco Teórico, se exponen los conceptos fundamentales del Machine Learning y Deep Learning que se utilizaron en este trabajo, incluyendo las arquitecturas de los modelos BERT y RoBERTa y las técnicas de procesamiento de texto que se emplearon.
- En el Capítulo 3, "The 17th International Workshop on Semantic Evaluation", se describe la participación en la competición y las técnicas y estrategias implementadas para la detección de valores humanos.
- En el Capítulo 4, Propuestas de Mejora, se plantean y se detallan nuevas estrategias y técnicas para mejorar los resultados obtenidos en la competición.
- En el Capítulo 5, Conclusiones y Trabajo Futuro, se presentan las conclusiones de este trabajo y las direcciones posibles para investigaciones futuras.
- Finalmente, en el Anexo, se presenta el artículo científico que ha sido aceptado y publicado en el "The 17th International Workshop on Semantic Evaluation".

Con esto, se busca proporcionar una contribución significativa a la tarea de detectar valores humanos en textos, utilizando para ello las técnicas más avanzadas en el campo del procesamiento del lenguaje natural.

CAPÍTULO 2
Marco Teórico

En este capítulo se exponen y explican los fundamentos teóricos necesarios para entender y abordar la detección automática de valores humanos en textos. En particular, se describen los conceptos fundamentales del aprendizaje automático, las redes bayesianas, el aprendizaje profundo y los transformers, todos ellos pilares del tratamiento automatizado de lenguaje natural y elementos centrales en la propuesta que se plantea en este trabajo. Además, también se abordará el preprocesamiento de texto, un paso fundamental para garantizar que los datos estén en un formato adecuado para su tratamiento.

2.1. CLASIFICACIÓN AUTOMÁTICA DE TEXTOS

La clasificación automática de textos es una tarea esencial en el procesamiento del lenguaje natural *NLP*. Consiste en asignar categorías predefinidas a un texto dado. Los algoritmos de clasificación se entrenan en un conjunto de datos etiquetado, donde cada ejemplo de texto se asocia con una o varias categorías. Una vez entrenado, el algoritmo puede utilizarse para clasificar nuevos ejemplos de texto en esas categorías.

2.1.1. Machine Learning

El Machine Learning o aprendizaje automático es un subcampo de la inteligencia artificial que desarrolla algoritmos que permiten a los sistemas aprender de los datos y mejorar su rendimiento en la toma de decisiones.

Una de las tareas más comunes en el Machine Learning es la clasificación, que consiste en aprender a asignar categorías a nuevos ejemplos basándose en un conjunto de ejemplos previos, denominado conjunto de entrenamiento. Este aprendizaje se realiza mediante la optimización de una función objetivo o función de pérdida, que mide la discrepancia entre las predicciones del modelo y las etiquetas reales de los ejemplos. Podemos diferenciar los siguientes tipos de aprendizaje:

- **Aprendizaje Supervisado:** En el aprendizaje supervisado, cada ejemplo de entrenamiento consta de un par de entrada-salida, donde la salida es una etiqueta asignada a la entrada. El objetivo del algoritmo es aprender una función que, dada una entrada, produzca la salida correcta. En la clasificación de textos, esta función se utilizaría para asignar una o varias categorías a un texto dado. Ejemplos de algoritmos de aprendizaje supervisado son las máquinas de vectores de soporte (SVM) y las redes neuronales[12].
- **Aprendizaje No Supervisado:** En el aprendizaje no supervisado, los datos de entrenamiento consisten solo en entradas sin etiquetas asociadas. El objetivo del algoritmo es descubrir estructuras ocultas en los datos. En la clasificación de textos, esto podría implicar la agrupación de textos similares juntos en grupos o çlusters", sin ninguna etiqueta predefinida para estos grupos. Ejemplos de algoritmos de aprendizaje no supervisado son K-means y el modelado de temas con LDA (Latent Dirichlet Allocation).
- **Aprendizaje Semi-supervisado:** El aprendizaje semi-supervisado se encuentra entre el aprendizaje supervisado y el no supervisado. En el aprendizaje semi-supervisado, algunos datos de entrenamiento tienen etiquetas y otros no. El objetivo del algoritmo es aprovechar tanto los datos etiquetados como los no etiquetados para aprender una función que produzca las etiquetas correctas para las entradas. En la clasificación de textos, esto podría implicar el uso de un pequeño conjunto de datos etiquetado para aprender una función de clasificación inicial, y luego perfeccionar esta función con un gran conjunto de datos no etiquetado.
- **Aprendizaje por Refuerzo:** En el aprendizaje por refuerzo, un agente aprende a tomar decisiones mediante la interacción con un entorno. El agente recibe una recompensa o una penalización basada en las decisiones que toma, y su objetivo es aprender a tomar decisiones que maximicen la recompensa acumulada a lo largo del tiempo. Aunque el aprendizaje por refuerzo se ha aplicado menos en la clasificación de textos que los otros tipos

de aprendizaje, hay algunas aplicaciones interesantes, como el aprendizaje de políticas de clasificación que se adaptan en el tiempo.

Cada uno de estos enfoques tiene sus fortalezas y debilidades, y la elección entre ellos depende de la naturaleza del problema y de los datos disponibles. En este trabajo, nos centraremos principalmente en el aprendizaje supervisado, ya que tenemos un conjunto de datos etiquetado disponible para entrenamiento y validación.

2.1.2. Redes Bayesianas

Las redes bayesianas son un modelo gráfico probabilístico que representa un conjunto de variables aleatorias y sus dependencias condicionales mediante un grafo dirigido acíclico (DAG). Cada nodo del grafo representa una variable aleatoria, mientras que las aristas representan las dependencias entre las variables.

Una red bayesiana permite representar de manera compacta una distribución conjunta sobre un conjunto de variables aleatorias. Cada variable en una red bayesiana tiene una distribución condicional asociada que depende de sus padres en la red. En particular, la probabilidad conjunta de un conjunto de variables en una red bayesiana se calcula como el producto de las probabilidades condicionales de cada variable dada sus padres en el grafo, según la fórmula:

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|Pa(x_i))$$

donde $Pa(x_i)$ denota el conjunto de padres del nodo x_i en el grafo.

La esencia de las redes bayesianas reside en la aplicación del Teorema de Bayes, una afirmación matemática fundamental en la teoría de la probabilidad y la estadística que describe cómo actualizar nuestras creencias sobre las hipótesis a la luz de nuevas pruebas. Se expresa formalmente como:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

donde:

- $P(H|E)$ es la probabilidad posterior, es decir, la probabilidad de la hipótesis H dada la evidencia E.
- $P(E|H)$ es la probabilidad de la evidencia dada la hipótesis, también conocida como la verosimilitud.
- $P(H)$ es la probabilidad de la hipótesis antes de la evidencia, también conocida como la probabilidad previa o a priori.
- $P(E)$ es la probabilidad de la evidencia.

En el contexto de las redes bayesianas, el Teorema de Bayes proporciona una forma de actualizar las probabilidades de las variables a la luz de nuevas pruebas. Si una nueva pieza de evidencia se hace disponible, las probabilidades de todas las variables en la red se actualizan automáticamente para reflejar esta nueva información.

Una de las grandes ventajas de las redes bayesianas es que permiten realizar inferencias a partir de datos incompletos o inciertos. Dado un conjunto de variables observadas, se puede calcular la distribución posterior de las variables no observadas utilizando las técnicas de inferencia de la red bayesiana.

Además, las redes bayesianas proporcionan un marco natural para aprender de los datos. Dado un conjunto de datos de entrenamiento, se puede aprender tanto la estructura de la red (es decir, qué variables dependen de cuáles) como los parámetros de la red (es decir, las probabilidades condicionales asociadas a cada variable). Este aprendizaje puede ser supervisado, si se conocen las categorías a las que pertenece cada ejemplo de entrenamiento, o no supervisado, si no se conocen estas categorías.

Por último, las redes bayesianas son altamente interpretables. A diferencia de otros modelos de aprendizaje automático, como las redes neuronales, las redes bayesianas proporcionan una representación visual y matemática clara de las dependencias entre las variables, lo que facilita su interpretación y análisis.

2.1.3. Deep Learning

Las técnicas de aprendizaje profundo son una extensión de las redes neuronales tradicionales, y su nombre se refiere a la gran cantidad de capas ocultas que estas pueden tener. Al aumentar la profundidad de una red neuronal, se incrementa su capacidad para aprender patrones complejos en los datos como podemos ver en la Figura 2.1.

Las redes neuronales se componen de neuronas artificiales o nodos, organizadas en capas. Cada neurona en una capa está conectada a todas las neuronas en la capa siguiente, y estas conexiones tienen asociados unos pesos, que son los que se ajustan durante el proceso de entrenamiento.

Cada neurona realiza una combinación lineal de sus entradas utilizando los pesos, y después aplica una función de activación no lineal. La salida de una neurona se convierte en la entrada para las neuronas de la capa siguiente. La primera capa de la red recibe como entrada los datos brutos, y la última capa produce la salida de la red.

En las redes neuronales profundas, existen una o más capas ocultas entre la capa de entrada y la capa de salida. La idea es que las primeras capas ocultas aprenden a extraer características de bajo nivel de los datos, como bordes en el caso de las imágenes, y las capas ocultas posteriores combinan estas características de bajo nivel para formar características de más alto nivel.

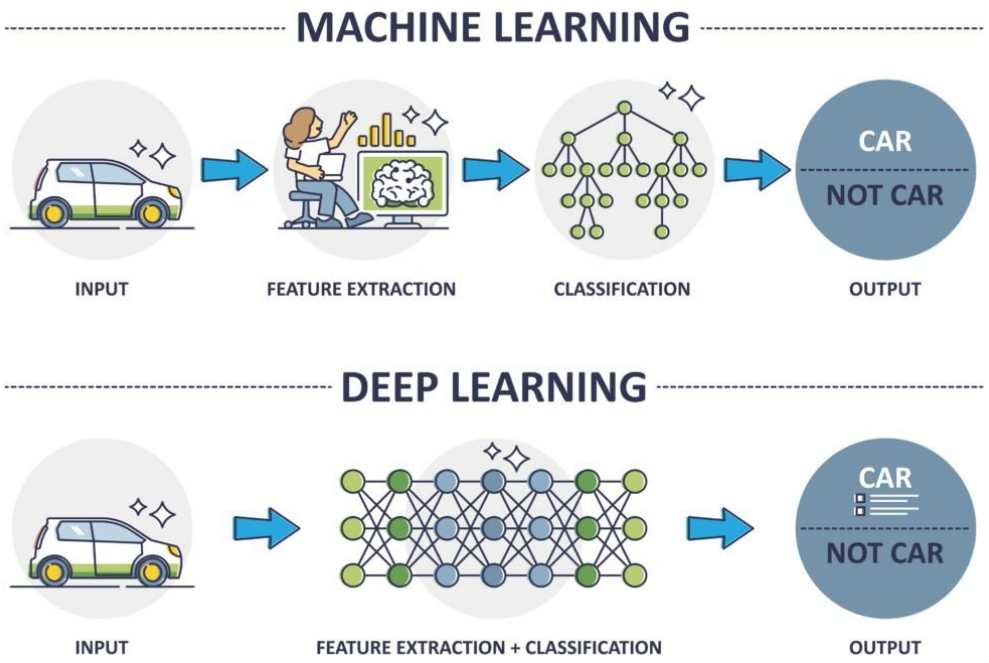

Figura 2.1: Diferencia entre Machine Learning y Deep Learning[1]

2.2. TRANSFORMERS

Los Transformers son una arquitectura de red neuronal propuesta por Vaswani et al. en su artículo 'Attention is All You Need' en 2017 [14]. Este modelo cambió drásticamente el panorama del procesamiento del lenguaje natural (PLN) y el aprendizaje automático en general, ya que ofrece una solución eficiente y eficaz para el modelado de dependencias a largo plazo en los datos secuenciales, como los textos.

[1] https://www.ait.de/en/deep-learning/

2.2.1. Funcionamiento de los Transformers

Los Transformers son una clase de modelos que utilizan mecanismos de atención para pesar la relevancia de las diferentes partes de la entrada para cada elemento de la salida. A diferencia de las redes neuronales recurrentes (RNNs)[13] y las redes de memoria a largo plazo (LSTMs)[7], que procesan las secuencias de manera iterativa, los Transformers son capaces de procesar todas las partes de la secuencia al mismo tiempo, en paralelo, lo que los hace más eficientes y adecuados para las modernas GPUs (véase Figura 2.2).

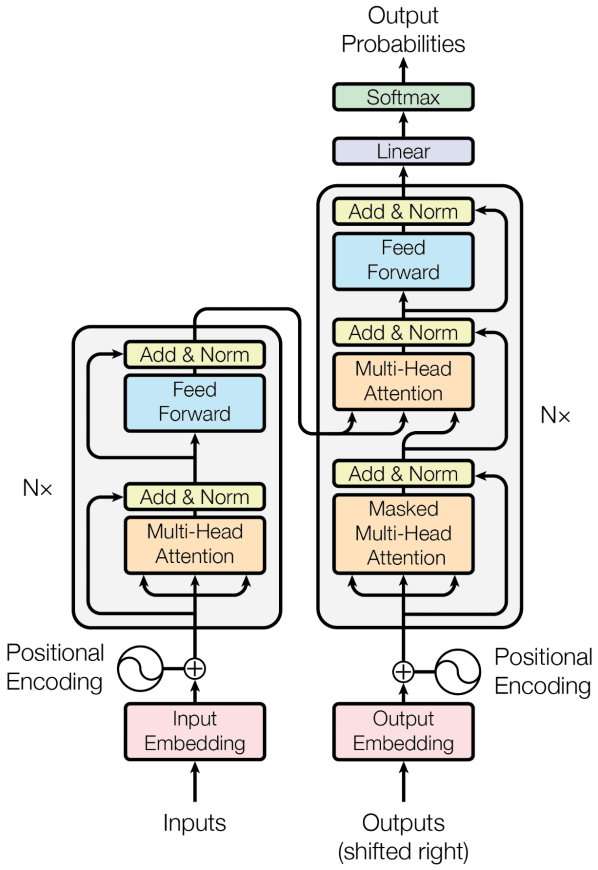

Figura 2.2: El Modelo Transformer[2]

Una de las características principales de los Transformers es la atención de múltiples cabezas, que permite al modelo concentrarse en diferentes partes de la entrada para cada cabeza de atención, y luego combinar la información de todas las cabezas para producir la salida. Este mecanismo de atención de múltiples cabezas es esencial para entender las complejas dependencias y relaciones que pueden existir en los datos de texto.

La arquitectura del Transformer se compone de dos partes principales: el codificador y el decodificador. El codificador lee y procesa la entrada, mientras que el decodificador genera la salida.

- **Codificador:** La entrada se pasa a través de una serie de capas de codificación, cada una de las cuales utiliza la atención de múltiples cabezas y una red feed-forward para procesar la información. Cada posición de la entrada es atendida por todas las demás posiciones para capturar la dependencia de la secuencia completa.
- **Decodificador:** Similar a la capa de codificación pero con una capa adicional de atención de múltiples cabezas para atender las salidas del codificador. Esto permite que cada paso del decodificador tenga acceso a toda la secuencia de entrada.

2.2.2. Ventajas e inconvenientes de los Transformers

Ventajas:

[2] https://www.aprendemachinelearning.com/como-funcionan-los-transformers-espanol-nlp-gpt-bert/

- Capacidad para modelar dependencias a largo plazo en los datos: A diferencia de las RNNs y las LSTMs, que pueden tener problemas para manejar dependencias a largo plazo debido al desvanecimiento del gradiente, los Transformers son capaces de manejar estas dependencias de manera eficiente.
- Eficiencia computacional: Los Transformers son altamente paralelizables, lo que los hace más eficientes que las RNNs y las LSTMs para el entrenamiento en GPUs modernas.
- Flexibilidad: Los Transformers no requieren que los datos sean secuenciales, lo que los hace adecuados para una amplia gama de tareas de aprendizaje automático.

Inconvenientes:

- Necesidad de grandes cantidades de datos: Como los Transformers son modelos muy potentes, necesitan grandes cantidades de datos para entrenar y pueden sobreajustar si los datos son limitados.
- Intensidad computacional: A pesar de su eficiencia en términos de paralelización, los Transformers son modelos intensivos en términos de memoria y cálculo, lo que puede ser un problema en dispositivos con recursos limitados.
- Dificultad para interpretar: Aunque los Transformers pueden producir grandes resultados, su funcionamiento interno puede ser difícil de interpretar y entender.

2.2.3. Aplicaciones de los Transformers

Los Transformers han demostrado ser efectivos en una amplia gama de tareas de PLN, incluyendo la traducción automática, el resumen de textos, la generación de texto, la comprensión del lenguaje natural, la respuesta a preguntas y muchas otras. Además, con la aparición de modelos Transformer como BERT, RoBERTa, GPT-3 y otros, los Transformers se han convertido en el estado del arte en muchas de estas tareas, superando a los enfoques tradicionales y estableciendo nuevos estándares de rendimiento.

2.2.4. BERT (Bidirectional Encoder Representations from Transformers)

BERT (Bidirectional Encoder Representations from Transformers) es una arquitectura de modelo basada en transformadores que ha demostrado un gran rendimiento en una amplia gama de tareas de procesamiento del lenguaje natural. A diferencia de los modelos de lenguaje tradicionales, que solo se basan en el contexto a la izquierda o a la derecha de una palabra, BERT utiliza una estrategia de entrenamiento de enmascaramiento de palabras para aprender representaciones de palabras bidireccionales.

2.2.4.1. Funcionamiento de BERT

BERT se basa en un *Transformer* (el mecanismo de atención que aprende relaciones contextuales entre palabras en un texto). Un *Transformer* básico consta de un codificador para leer la entrada de texto y un decodificador para producir una predicción para la tarea. Dado que el objetivo de BERT es generar un modelo de representación del lenguaje, solo necesita la parte del codificador. La entrada para el codificador de BERT es una secuencia de tokens, que primero se convierten en vectores y luego se procesan en la red neuronal. Pero antes de que pueda comenzar el procesamiento, BERT necesita que la entrada se masajee y decore con algunos metadatos adicionales:

- ***Token embeddings***: Se añade un token [CLS] a los tokens de palabras de entrada al principio de la primera frase y se inserta un token [SEP] al final de cada frase.
- ***Segment embeddings***: Se añade un marcador que indica *Frase A* o *Frase B* a cada token. Esto permite que el codificador distinga entre frases.
- ***Positional embeddings***: Se añade una incrustación posicional a cada token para indicar su posición en la frase.

Esencialmente, el *Transformer* apila una capa que mapea secuencias a secuencias, por lo que la salida también es una secuencia de vectores con una correspondencia 1:1 entre tokens de entrada y salida en el mismo índice. Y como aprendimos anteriormente, BERT no intenta predecir la siguiente palabra en la frase. BERT implementa dos estrategias principales durante su proceso de entrenamiento, que son esenciales para comprender cómo aprende representaciones lingüísticas efectivas:

[3] https://medium.com/@samia.khalid/bert-explained-a-complete-guide-with-theory-and-tutorial-3ac9ebc8fa7c

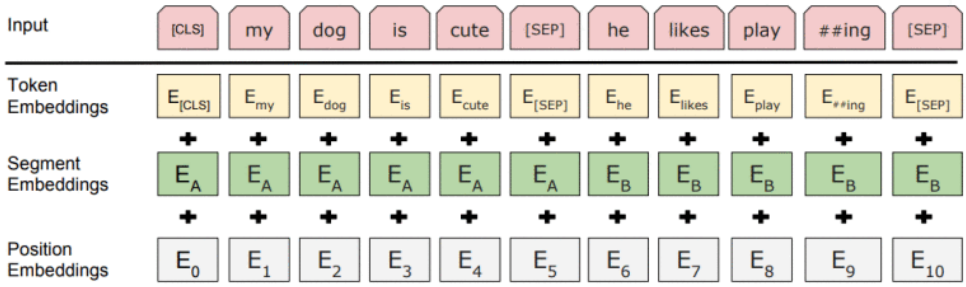

Figura 2.3: Esquema representativo del proceso de BERT[3]

1. **Masked LM (MLM)**

   BERT utiliza un enfoque conocido como *Masked Language Model* (MLM). La idea detrás de MLM es relativamente directa: se enmascara aleatoriamente el 15 % de las palabras en la entrada, sustituyéndolas con un token [MASK]. Luego, se pasa toda la secuencia a través del codificador basado en atención de BERT, y se intenta predecir solamente las palabras enmascaradas, basándose en el contexto proporcionado por las demás palabras no enmascaradas. Sin embargo, el enmascaramiento directo plantea un problema: el modelo solo intentaría predecir cuando el token [MASK] está presente. Para resolver esto, de los tokens seleccionados para enmascaramiento:

   - El 80 % se reemplaza realmente con el token [MASK].
   - El 10 % de las veces, los tokens se reemplazan con un token aleatorio.
   - El 10 % restante, los tokens permanecen sin cambios.

   Durante el entrenamiento, la función de pérdida de BERT considera solo la predicción de los tokens enmascarados e ignora los no enmascarados. Esto resulta en un modelo que converge más lentamente que los modelos de izquierda a derecha o de derecha a izquierda.

2. **Next Sentence Prediction (NSP)**

   BERT también incorpora la predicción de la siguiente oración para comprender las relaciones entre dos frases, esencial para tareas como respuesta a preguntas. Durante el entrenamiento, al modelo se le proporcionan pares de oraciones y aprende a predecir si la segunda oración sigue a la primera en el texto original. La entrada a BERT separa las oraciones con un token especial [SEP]. Durante el entrenamiento, se le suministra al modelo dos oraciones de entrada de manera que:

   - El 50 % de las veces, la segunda oración sigue a la primera.
   - El 50 % restante, es una oración aleatoria del corpus completo.

   BERT debe entonces predecir si la segunda oración es aleatoria o no, asumiendo que una oración aleatoria estará desconectada de la primera. Para hacer esta predicción, toda la secuencia de entrada pasa por el modelo basado en Transformer. La salida del token [CLS] se transforma en un vector con forma 2x1 utilizando una capa de clasificación simple, y se asigna la etiqueta IsNext mediante softmax.

   El modelo se entrena con ambas estrategias, MLM y NSP, simultáneamente, minimizando la función de pérdida combinada de ambos — la combinación potencia la eficacia del entrenamiento.

2.2.4.2. Arquitectura de BERT

La arquitectura de BERT se compone de múltiples capas de transformers apilados. Cada capa de transformador tiene una estructura similar, con una capa de atención multi-cabeza y una red neuronal de feed-forward. La entrada a BERT consiste en una secuencia de tokens, donde se agregan tokens especiales al principio y al final de la secuencia para denotar el inicio y el final del texto.

BERT se entrena en una tarea de lenguaje a gran escala utilizando enormes cantidades de datos no etiquetados. Una vez entrenado, se puede utilizar en tareas específicas de NLP a través de un proceso llamado ajuste fino (fine-tuning), donde el modelo preentrenado se adapta a una tarea específica mediante el entrenamiento con un conjunto de datos etiquetados más pequeño.

2.2.4.3. Ejemplos de uso

BERT ha demostrado un rendimiento sobresaliente en una variedad de tareas de NLP[6], incluyendo el reconocimiento de entidades nombradas, la clasificación de texto, el etiquetado de secuencias y la respuesta a preguntas. Su capacidad para capturar tanto información local como global en los textos ha llevado a mejoras significativas en la comprensión del lenguaje natural.

Algunos ejemplos de uso de BERT incluyen:

En tareas de clasificación de texto, BERT puede analizar el contexto completo de una oración para determinar la categoría a la que pertenece. En el etiquetado de secuencias, BERT puede asignar etiquetas a diferentes partes de una oración, como identificar los sustantivos y los verbos. En la respuesta a preguntas, BERT puede entender el contexto de una pregunta y proporcionar respuestas relevantes basadas en la información del texto.

Los resultados de BERT han sido muy prometedores, superando en muchos casos a los modelos anteriores y estableciendo nuevos puntos de referencia en varias tareas de NLP.

2.2.5. RoBERTa

RoBERTa (Robustly Optimized BERT approach) es una variante mejorada del modelo BERT que se basa en la arquitectura de transformadores y ha demostrado un rendimiento sobresaliente en una amplia gama de tareas de procesamiento del lenguaje natural (NLP). RoBERTa es una versión optimizada y robusta de BERT, que incorpora modificaciones y técnicas adicionales para mejorar su capacidad de modelado del lenguaje.

2.2.5.1. Modificaciones clave en RoBERTa

Aunque RoBERTa se basa en la arquitectura de BERT, Facebook AI introdujo una serie de mejoras en el proceso de entrenamiento y en la metodología de preentrenamiento para optimizar su rendimiento. Las modificaciones principales incluyen:

1. **Eliminación de la predicción de la siguiente oración (NSP)**:

   BERT incluía una tarea llamada "Next Sentence Prediction"(NSP) para preentrenar sus modelos. RoBERTa descartó esta tarea, entrenando en su lugar con secuencias de texto más largas, permitiendo que las oraciones contiguas de un documento entrenaran juntas.

2. **Uso de más datos de entrenamiento**:

   Aunque BERT ya utilizó una combinación de BooksCorpus y English Wikipedia para su entrenamiento, RoBERTa expandió esta base de datos. Incorporó más textos, como OpenWebText, que es un conjunto de datos que replica el utilizado en GPT-2.

3. **Optimización en la formación de lotes**:

   RoBERTa entrenó con lotes más grandes y con secuencias más largas en comparación con BERT. Esto, combinado con el hecho de que se eliminó la tarea NSP, permitió un entrenamiento más eficiente y una mejor representación contextual del texto.

4. **Dinámica de máscaras**:

   En lugar de enmascarar palabras al azar en cada época (como hacía BERT), RoBERTa enmascara dinámicamente palabras de manera diferente en cada época durante el entrenamiento, lo que resulta en un modelo más robusto.

5. **Entrenamiento más extenso**:

   Mientras que BERT tenía un entrenamiento específico dependiendo de su tamaño (por ejemplo, BERT-base fue entrenado durante 1M pasos), RoBERTa se entrenó durante más tiempo, explorando hasta 500k pasos para garantizar una convergencia óptima.

En resumen, aunque RoBERTa mantiene la arquitectura fundamental de BERT, las modificaciones en el proceso de preentrenamiento y en los datos utilizados han permitido a RoBERTa superar a BERT en una serie de tareas de procesamiento de lenguaje natural.

2.2.5.2. Funcionamiento de RoBERTa

RoBERTa utiliza una estrategia similar a BERT, donde aprende representaciones contextuales de palabras basadas en el enmascaramiento de palabras y la predicción de frases siguientes. Sin embargo, a diferencia de BERT, RoBERTa se entrena en un conjunto de datos más grande y utiliza una configuración de entrenamiento más larga. Esto permite a RoBERTa capturar aún más información contextual y mejorar su capacidad para modelar el lenguaje.

Una de las características clave de RoBERTa es su capacidad para realizar preentrenamiento sin discriminación (unsupervised pretraining). En lugar de enfocarse en tareas de lenguaje específicas durante el preentrenamiento, RoBERTa se entrena en una gran cantidad de texto sin etiquetar, lo que le permite aprender representaciones más generales y robustas del lenguaje. En la siguiente Figura 2.4 podemos ver cmo RoBERTa nos da una predicción de la palabra "comiendo" dada una frase cualquiera.

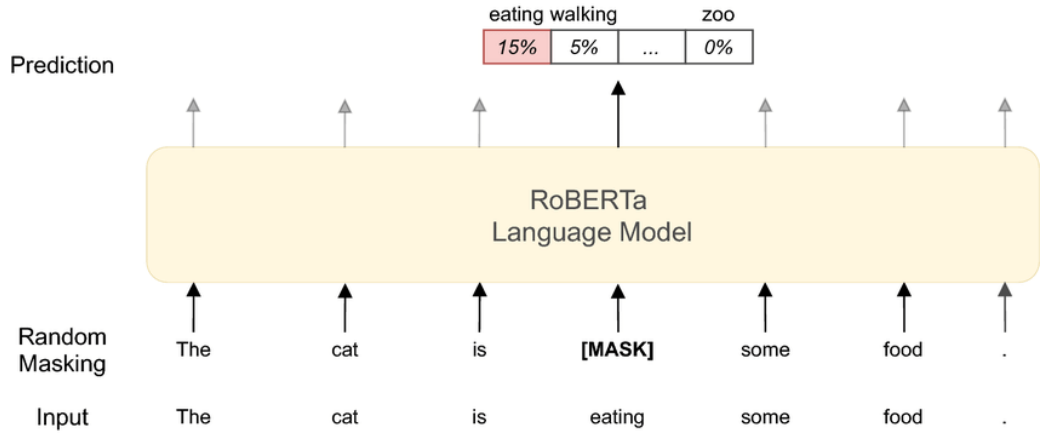

Figura 2.4: Esquema ejemplo de uso de RoBERTa[4]

2.2.5.3. Arquitectura de RoBERTa

La arquitectura de RoBERTa es similar a la de BERT, con una pila de transformadores apilados. Cada transformador consta de una capa de atención multi-cabeza y una red neuronal de feed-forward. La entrada a RoBERTa también consiste en una secuencia de tokens, con tokens especiales añadidos para denotar el inicio y el final del texto.

RoBERTa se entrena utilizando una estrategia de preentrenamiento sin discriminación y luego se ajusta a tareas específicas de PLN a través del ajuste fino. Esta técnica de ajuste fino permite que RoBERTa se adapte a tareas específicas mediante el entrenamiento con un conjunto de datos etiquetados más pequeño.

2.2.5.4. Ejemplos de uso

RoBERTa ha demostrado un rendimiento sobresaliente en una amplia variedad de tareas de PLN, incluyendo el procesamiento de lenguaje natural, la traducción automática, el análisis de sentimientos [5], la generación de texto y muchas otras. Su capacidad para capturar información contextual y su enfoque en el preentrenamiento sin discriminación han llevado a mejoras significativas en el rendimiento en comparación con los modelos anteriores.

Algunos ejemplos de uso de RoBERTa incluyen:

En la traducción automática, RoBERTa puede aprender representaciones más generales del lenguaje que mejoran la calidad de las traducciones. En el análisis de sentimientos, RoBERTa puede capturar las sutilezas y matices en el lenguaje para determinar las emociones expresadas en un texto. En la generación de texto, RoBERTa puede generar texto coherente y natural que se ajusta al estilo y al contexto de entrada.

Los resultados de RoBERTa han demostrado mejoras significativas en comparación con los modelos anteriores, estableciendo nuevos estándares de rendimiento en una amplia gama de tareas de PLN.

2.3. PREPROCESAMIENTO DE TEXTO

El preprocesamiento de texto es un paso crucial en cualquier proyecto de procesamiento de lenguaje natural. Asegura que el texto esté en un formato que los modelos puedan utilizar de manera eficaz. Aunque los pasos específicos pueden variar según el proyecto y el lenguaje de los datos, algunas tareas comunes de preprocesamiento de texto incluyen la eliminación

[4] https://www.researchgate.net/figure/RoBERTa-masked-language-modeling-with-the-input-sentence-The-cat-is-eating-some-food_fig1_358563215

de la puntuación, la conversión de todas las letras a minúsculas, la eliminación de las palabras vacías (stop words), la lematización y la tokenización. Además, en este trabajo se ha abordado el desafío del desequilibrio en la distribución de clases en el conjunto de datos.

2.3.1. Datos Desbalanceados

En muchos problemas de clasificación, la distribución de las clases en el conjunto de datos está desequilibrada, es decir, hay muchas más instancias de algunas clases que de otras. Este desequilibrio puede afectar al rendimiento de los modelos de aprendizaje automático, ya que tienden a estar sesgados hacia las clases mayoritarias y a tener un rendimiento deficiente en las clases minoritarias.

2.3.1.1. Undersampling

Una forma común de abordar el desequilibrio de clases es a través del undersampling, que consiste en reducir el número de instancias de las clases mayoritarias para igualarlas con las de las clases minoritarias (véase Figura 3.2). Aunque esta técnica puede ayudar a mejorar el rendimiento en las clases minoritarias, también puede llevar a la pérdida de información importante, ya que se eliminan instancias de las clases mayoritarias.

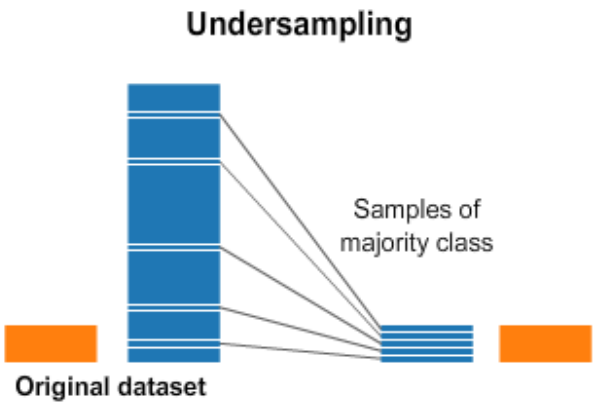

Figura 2.5: Ejemplo de visualización de dataset tras Undersampling[5]

2.3.1.2. Data Augmentation con Backtranslation

Otra forma de abordar el desequilibrio de clases es a través de técnicas de aumento de datos, que generan nuevas instancias sintéticas de las clases minoritarias. En particular, el aumento de datos con backtranslation es una técnica común en el procesamiento del lenguaje natural que implica traducir el texto a otro idioma y luego volver a traducirlo al idioma original (véase Figura2.6). Esta técnica puede generar variantes del texto original que conservan su significado pero tienen una forma ligeramente diferente, lo que puede ayudar a aumentar la diversidad y robustez del conjunto de datos.

En resumen, el preprocesamiento de texto y la gestión del desequilibrio de clases son dos aspectos fundamentales en la construcción de modelos eficaces para la detección de valores humanos en textos. En este trabajo, hemos utilizado técnicas de preprocesamiento de texto estándar, así como estrategias de undersampling y aumento de datos con backtranslation, para preparar los datos para el entrenamiento de los modelos.

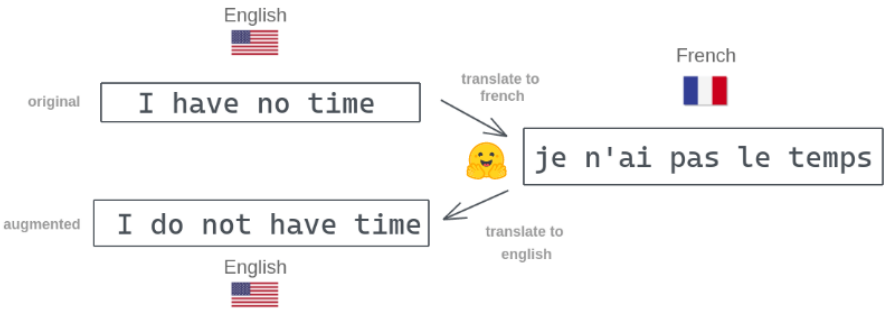

Figura 2.6: Ejemplo de Back Translation[6]

[5] https://towardsdatascience.com/5-techniques-to-work-with-imbalanced-data-in-machine-learning-80836d45d30c

2.3.2. Eliminación de Stopwords

Las *stopwords*[9] son palabras comunes en un lenguaje que generalmente se consideran de poco valor informativo en el procesamiento de texto. Estas incluyen palabras como ”y”, ”o”, ”la”, ”el”, entre otras en el caso del español, o ”and”, ”or”, ”the”, ”is” en inglés, como podemos apreciar en la Figura 2.7. Dado que estas palabras aparecen con alta frecuencia y suelen aportar poco o ningún valor distintivo en el análisis de texto, a menudo se eliminan durante el preprocesamiento.

Figura 2.7: Ejemplo de Stop words[7]

La eliminación de stopwords tiene varias ventajas:

- **Eficiencia**: Reducir el número de palabras no informativas puede acelerar significativamente las operaciones de procesamiento, especialmente en grandes conjuntos de datos.
- **Mejora en la precisión**: Al eliminar las palabras que no son relevantes para el significado de un texto, se pueden mejorar los resultados en tareas como la clasificación de textos o la agrupación (clustering).
- **Reducción de ruido**: En tareas como el análisis de sentimientos, las stopwords pueden introducir ruido que distraiga al modelo de las palabras verdaderamente relevantes para determinar el sentimiento.

Sin embargo, es crucial considerar el contexto en el que se está trabajando. En algunas tareas, como el análisis semántico o la traducción automática, eliminar las stopwords podría ser perjudicial, ya que estas palabras podrían contener información contextual valiosa.

En la práctica, la eliminación de stopwords suele realizarse utilizando listas predefinidas para un lenguaje en particular, aunque también se pueden adaptar según las necesidades específicas de un proyecto. Algunas bibliotecas de procesamiento de lenguaje natural, como NLTK[8] o spaCy[9], ofrecen listas de stopwords integradas que pueden utilizarse para este fin.

2.4. CONCLUSION TEÓRICA

La clasificación automática de textos, como la detección de valores humanos en textos, requiere de modelos capaces de entender y representar el lenguaje natural de manera efectiva. En este sentido, hemos explorado diferentes conceptos y técnicas en el marco teórico para abordar esta tarea.

En primer lugar, hemos discutido el uso de modelos de Machine Learning, como las redes bayesianas, para la clasificación de textos. Estos modelos han demostrado su eficacia en tareas de clasificación, pero presentan limitaciones en el modelado del lenguaje natural debido a su dependencia de características manuales y su incapacidad para capturar representaciones contextuales complejas.

Posteriormente, nos adentramos en el Deep Learning y su capacidad para aprender representaciones más ricas y complejas del texto. Con el surgimiento de los transformadores, representados por modelos como BERT y RoBERTa, se ha logrado un gran avance en el entendimiento del lenguaje natural. Estos modelos son capaces de capturar dependencias a largo plazo entre palabras y aprenden representaciones bidireccionales del texto, lo que los convierte en herramientas poderosas para la detección de valores humanos en textos.

Además, hemos explorado el preprocesamiento de texto, una etapa crucial para garantizar que los datos estén en un formato adecuado para el análisis. En particular, hemos abordado el desafío del desequilibrio en la distribución de clases en el conjunto de datos y hemos discutido técnicas como el undersampling y el data augmentation con backtranslation, así como la eliminación de *stopwords*, de esta forma conseguimos mitigar este desequilibrio y mejorar la calidad del conjunto de entrenamiento.

En conclusión, el marco teórico presentado nos ha proporcionado una comprensión sólida de los conceptos y técnicas fundamentales necesarios para abordar la detección de valores humanos en textos. La combinación de modelos de Machine Learning, como las redes bayesianas, con enfoques más avanzados basados en el Deep Learning y los transformers, como BERT y RoBERTa, nos permite mejorar la capacidad de los modelos para comprender y clasificar textos de manera más

[6] https://amitness.com/back-translation/
[7] https://www.youtube.com/watch?v=0D7ae7OaaHQ
[8] https://www.nltk.org/
[9] https://spacy.io/

precisa y efectiva. El preprocesamiento de texto también desempeña un papel crucial al garantizar que los datos estén en un formato adecuado para el análisis y al abordar desafíos como el desequilibrio en la distribución de clases.

En las siguientes secciones, exploraremos cómo se han aplicado estos conceptos y técnicas en el desarrollo de este trabajo, detallando el enfoque y los resultados obtenidos en la detección automática de valores humanos en textos.

CAPÍTULO 3

Competición - *ValueEval*

3.1. DESCRIPCIÓN

Para abordar este proyecto, se ha participado en la tarea *"ValueEval: Identification of Human Values behind Arguments"* de detección automática de valores humanos en textos. El objetivo de la tarea es clasificar si un argumento textual hace referencia o no a una categoría de valores humanos específica. Se utiliza un conjunto de 20 categorías de valores definidas gracias a las ciencias sociales.

Los argumentos se presentan en forma de texto de premisa, texto de conclusión y postura binaria de la premisa hacia la conclusión ("a favor de" o "en contra"). Por ejemplo, se presentan argumentos relacionados con la creatividad, la curiosidad, la libertad de pensamiento, las tradiciones, entre otros valores.

El desafío radica en desarrollar enfoques y sistemas que puedan detectar y clasificar de manera precisa los valores humanos presentes en los argumentos. Se invita a los participantes a presentar enfoques que detecten una o varias de estas categorías de valores en los argumentos.

La detección automática de valores humanos en textos es una tarea importante, ya que nos permite comprender mejor cómo se expresan y se manifiestan los valores en el lenguaje natural. Además, este conocimiento puede tener aplicaciones en diversos campos, como la comprensión de las motivaciones y creencias de las personas, la detección de sesgos en el discurso y la mejora de la interacción entre individuos y sistemas de inteligencia artificial.

3.2. ANÁLISIS DEL CONJUNTO DE DATOS

El conjunto de datos utilizado [8] en este estudio está compuesto por 5393 argumentos únicos. Estos fueron recopilados de diversas fuentes, como textos religiosos, discusiones políticas, editoriales de periódicos y plataformas de democracia en línea. Cada argumento se representa de la siguiente manera:

- **ID del Argumento:** Es un identificador único para cada argumento.
- **Conclusión:** Representa la conclusión del argumento.
- **Postura:** Indica si el argumento está a favor o en contra, estableciendo la relación entre la conclusión y su premisa.
- **Premisa:** Es la premisa del argumento.
- **Valores:** Los 20 valores representan categorías de valores asociadas a cada argumento, como Autodeterminación: pensamiento, Autodeterminación: acción, Estimulación, Hedonismo, Logro, Poder: dominancia, Poder: recursos, Rostro, Seguridad: personal, Seguridad: societal, Tradición, Conformidad: reglas, Conformidad: interpersonal, Humildad, Benevolencia: cuidado, Benevolencia: confiabilidad, Universalismo: preocupación, Universalismo: naturaleza, Universalismo: tolerancia y Universalismo: objetividad. Cada uno de estos valores es una característica binaria que indica si está presente (1) o ausente (0) en el argumento.

¿Cómo se identifican los valores?

Autodeterminación: pensamiento: Es bueno tener ideas e intereses propios, lo que implica la capacidad de reflexionar, cuestionar y analizar de manera independiente. Este valor se manifiesta en una serie de argumentos que fomentan diferentes aspectos del pensamiento crítico como la creatividad, la curiosidad o tener libertad de pensamiento.

Autodeterminación: acción: Es esencial que las personas determinen sus propias acciones. Este valor aboga por la elección individual de metas, promoviendo la independencia y la capacidad de actuar sin requerir consentimiento externo. También subraya la importancia de la libertad de acción y la preservación de la privacidad, defendiendo espacios

personales y el control sobre la información personal. Todo ello resalta la necesidad de autonomía personal en las decisiones y acciones diarias.

Estimulación: Es beneficioso experimentar emociones intensas, novedad y cambio. Este valor se refleja en argumentos que promueven una vida emocionante, variedad en las actividades y la audacia para asumir riesgos. Estos ejemplos ilustran cómo la estimulación se manifiesta en la búsqueda de emociones intensas, la apertura a la novedad y la disposición a experimentar cambios. La comprensión de estos argumentos nos permite reconocer la importancia de la estimulación como un valor humano y su impacto en la forma en que las personas perciben e interactúan con el mundo.

Hedonismo: Es esencial experimentar placer y satisfacción sensorial. Este valor se manifiesta en argumentos que promueven hacer de la vida una experiencia disfrutable, proporcionando momentos de ocio, oportunidades para divertirse y experimentar gratificación sensorial. Este enfoque resalta la importancia de buscar el placer y el goce en las actividades diarias, subrayando la relevancia del bienestar personal y la celebración de los pequeños placeres de la vida.

Logro: Es valioso ser exitoso conforme a las normas sociales. Este valor se refleja en argumentos que enfatizan la ambición, la búsqueda del éxito y el reconocimiento de logros. Subraya la importancia de ser capaz y competente en diversas tareas, de adquirir habilidades cognitivas elevadas y de actuar con valentía. Destaca la relevancia de la perseverancia, la reflexión y la valentía para mantenerse firme en las convicciones propias.

Poder: dominancia: Es valioso estar en posiciones de control sobre otros. Este valor se manifiesta en argumentos que subrayan la importancia de tener influencia y la capacidad de moldear eventos. Se enfatiza la relevancia de tener el derecho de dar órdenes, colocando a expertos al mando y estableciendo jerarquías claras. Resalta la necesidad de un liderazgo adecuado y de asegurar que las personas adecuadas tomen decisiones críticas.

Poder: recursos: Es valioso tener posesiones materiales y recursos sociales. Este valor se refleja en argumentos que promueven la acumulación y muestra de riqueza, y cómo esta puede ser una herramienta de control. Destaca la importancia del bienestar financiero y cómo la prosperidad material puede influir en la posición social y el poder de un individuo.

Reputación: Es esencial mantener una imagen pública positiva. Este valor se manifiesta en argumentos que enfatizan la importancia de ganar respeto y reconocimiento social, así como de evitar la humillación. Se destaca la necesidad de construir y proteger una buena reputación, garantizando que la imagen pública se mantenga intacta y favorable. La gestión de esta imagen es crucial para la percepción social y el estatus de un individuo.

Seguridad Personal: Es esencial tener un entorno inmediato seguro. Este valor se refleja en argumentos que subrayan la importancia de establecer un sentido de pertenencia, permitiendo que las personas formen, se unan y permanezcan en grupos, mostrando su afiliación grupal y su cuidado mutuo. También abarca la necesidad de mantener una buena salud, evitando enfermedades y promoviendo el bienestar físico y mental. La estabilidad financiera, simbolizada por la ausencia de deudas y preocupaciones monetarias, es otro aspecto fundamental, junto con la importancia de llevar una vida ordenada, limpia y cómoda, lo que resulta en una mayor felicidad general.

Seguridad Societal: Es fundamental contar con una sociedad amplia segura y estable. Este valor se refleja en argumentos que subrayan la necesidad de un país seguro, donde el Estado pueda actuar eficazmente contra los delitos, defender y cuidar a sus ciudadanos, promoviendo la fortaleza estatal en general. Además, se destaca la importancia de una sociedad estable, respaldando la estructura social existente y evitando el caos y desorden a nivel societal.

Tradición: Es esencial mantener las tradiciones culturales, familiares o religiosas. Este valor se refleja en argumentos que enfatizan la importancia de respetar las tradiciones, permitiendo seguir las costumbres familiares, honrar prácticas tradicionales y mantener valores y formas de pensar arraigados. Además, se subraya la relevancia de mantener la fe religiosa, permitiendo la práctica de costumbres religiosas y dedicando la vida a la fe, promoviendo así la piedad y la difusión de la religión propia.

Conformidad: reglas: Es fundamental cumplir con las reglas, leyes y obligaciones formales. Este valor se refleja en argumentos que destacan la importancia de ser cumplido, resaltando la adherencia a leyes y normativas y la promoción del reconocimiento a aquellos que cumplen con sus obligaciones. También se pone énfasis en la autodisciplina, impulsando la restricción propia, el seguimiento de reglas incluso cuando no hay supervisión y la autodeterminación de reglas personales. Además, se resalta la importancia de comportarse adecuadamente, evitando la violación de reglas informales o convenciones sociales y promoviendo buenas maneras.

Conformidad: interpersonal: Es crucial evitar molestar o dañar a los demás. Este valor se ve reflejado en argumentos que enfatizan la importancia de ser cortés, evitando ofender a otras personas, considerando a los demás y minimizando molestias hacia otros. También se subraya la relevancia de honrar a los mayores, siguiendo las recomendaciones de los padres y demostrando fe y respeto hacia los ancianos.

Humildad: Es esencial reconocer la propia insignificancia en el gran esquema de las cosas. Este valor se refleja en argumentos que enfatizan la importancia de ser humilde, reduciendo la arrogancia y la jactancia, y evitando pensar

que uno merece más que los demás. También destaca la importancia de priorizar el éxito del grupo por encima de las personas individuales y devolver a la sociedad. Además, se subraya la relevancia de aceptar la vida tal como es, aceptando el propio destino, sometiéndose a las circunstancias de la vida y estando satisfecho con lo que uno tiene.

Benevolencia: cuidado: Es fundamental trabajar para el bienestar de los miembros del propio grupo. Este valor se manifiesta en argumentos que enfatizan la importancia de ser útil, ayudando a las personas de su grupo y promoviendo el trabajo por el bienestar de los demás en ese grupo. También resalta la relevancia de la honestidad y el reconocimiento a las personas por su sinceridad. Además, subraya la necesidad de ser perdonador, permitiendo que las personas se perdonen mutuamente y dándoles una segunda oportunidad, así como mostrando misericordia. También se destaca la importancia de asegurar y cuidar a la propia familia, así como de fomentar relaciones estrechas, priorizando el bienestar de los demás por encima del propio y permitiendo mostrar afecto, compasión y simpatía.

Benevolencia: confiabilidad: Es esencial ser un miembro confiable y digno de confianza en el grupo al que se pertenece. Este valor se refleja en argumentos que destacan la importancia de ser responsable, promoviendo responsabilidades claras, fomentando la confianza y destacando la fiabilidad. Además, subraya la relevancia de tener lealtad hacia los amigos, siendo un amigo en quien se puede confiar, leal y digno de confianza, así como brindar un apoyo incondicional a los amigos.

Universalismo: preocupación: Es esencial luchar por la igualdad, justicia y protección de todas las personas. Este valor destaca la importancia de la igualdad, reflejada en argumentos que fomentan la elevación de individuos de estatus social inferior, ayudar a las regiones más desfavorecidas del mundo, proporcionar igualdad de oportunidades para todos y aspirar a un mundo donde el éxito no esté determinado por el origen. La justicia es otro pilar, enfatizando la imparcialidad y protección de los más vulnerables, promoviendo un mundo con menos discriminación basada en raza, género y otros factores. Además, se subraya la relevancia de un mundo en paz, promoviendo la cesación de conflictos, evitando guerras y valorando la paz como un bien precioso para toda la humanidad.

Universalismo: naturaleza: Es esencial preservar el entorno natural. Este valor subraya la importancia de proteger el medio ambiente, reflejada en argumentos que promueven la prevención de la contaminación, el cuidado de la naturaleza y la implementación de programas de restauración ambiental. Asimismo, se enfatiza la armonía con la naturaleza, abogando por evitar el uso de productos químicos y organismos genéticamente modificados, tratando a los animales y plantas con respeto y considerando sus seres sensibles. Esto resulta en una vida en equilibrio con la naturaleza, donde las personas reflexionan sobre las consecuencias de sus acciones en el entorno. Además, se valora la belleza natural y artística, permitiendo a las personas apreciar el arte y la majestuosidad de la naturaleza, fomentando una mayor conexión con el mundo natural y las expresiones artísticas.

Universalismo: tolerancia: Es esencial aceptar y tratar de comprender a aquellos que son diferentes a uno mismo. Este valor se centra en la amplitud de mente, reflejándose en argumentos que abogan por el fomento del diálogo entre distintos grupos, el esfuerzo por superar prejuicios, la importancia de escuchar a aquellos con diferentes perspectivas y la promoción de experiencias de vida en ambientes diferentes al propio. Estos argumentos subrayan la importancia de la tolerancia entre diversas personas y colectivos. Además, se valora la sabiduría en la aceptación de los demás, enfatizando la capacidad de aceptar desacuerdos y opiniones contrarias, y la promoción de un entendimiento maduro y equilibrado de las distintas opiniones, evitando caer en el partidismo o el fanatismo.

Universalismo: objetividad: Es esencial buscar la verdad y pensar de manera racional y sin prejuicios. Este valor resalta la importancia de la lógica, manifestándose en argumentos que priorizan el uso de datos objetivos en lugar de simples intuiciones, enfatizando un pensamiento racional, centrado y consistente, y la aplicación del método científico. Asimismo, se destaca la relevancia de tener una perspectiva objetiva, en la cual se busca la verdad, se adopta una postura neutral y se forma una opinión imparcial evaluando detenidamente todos los pros y contras. Estos argumentos subrayan la necesidad de brindar a las personas las herramientas necesarias para tomar decisiones informadas.

En relación a la distribución de los datos, se puede observar que los valores más frecuentes en los argumentos son *Universalismo: preocupación y Seguridad: personal*. Por otro lado, los valores menos frecuentes son *Hedonismo y Conformidad: interpersonal* (ver Figura 3.1).

Este conjunto de datos proporciona una rica fuente de información para el análisis de argumentos y la detección de valores debido a su estructura que incluye la conclusión, la postura y la premisa de cada argumento. Esta información brinda un contexto completo para comprender el contenido y su intención. La conclusión representa la afirmación o punto principal del argumento, mientras que la postura indica si el argumento está a favor o en contra de la premisa y esta, proporciona el respaldo o evidencia para la conclusión.

Al tener acceso a estas tres partes clave del argumento, los modelos de detección de valores pueden analizar y relacionar la forma en que estos se expresan. Esto permite una mejor comprensión de cómo los valores se manifiestan en el razonamiento y cómo influyen en la posición y perspectiva del argumento. El uso de este dataset, permite desarrollar y

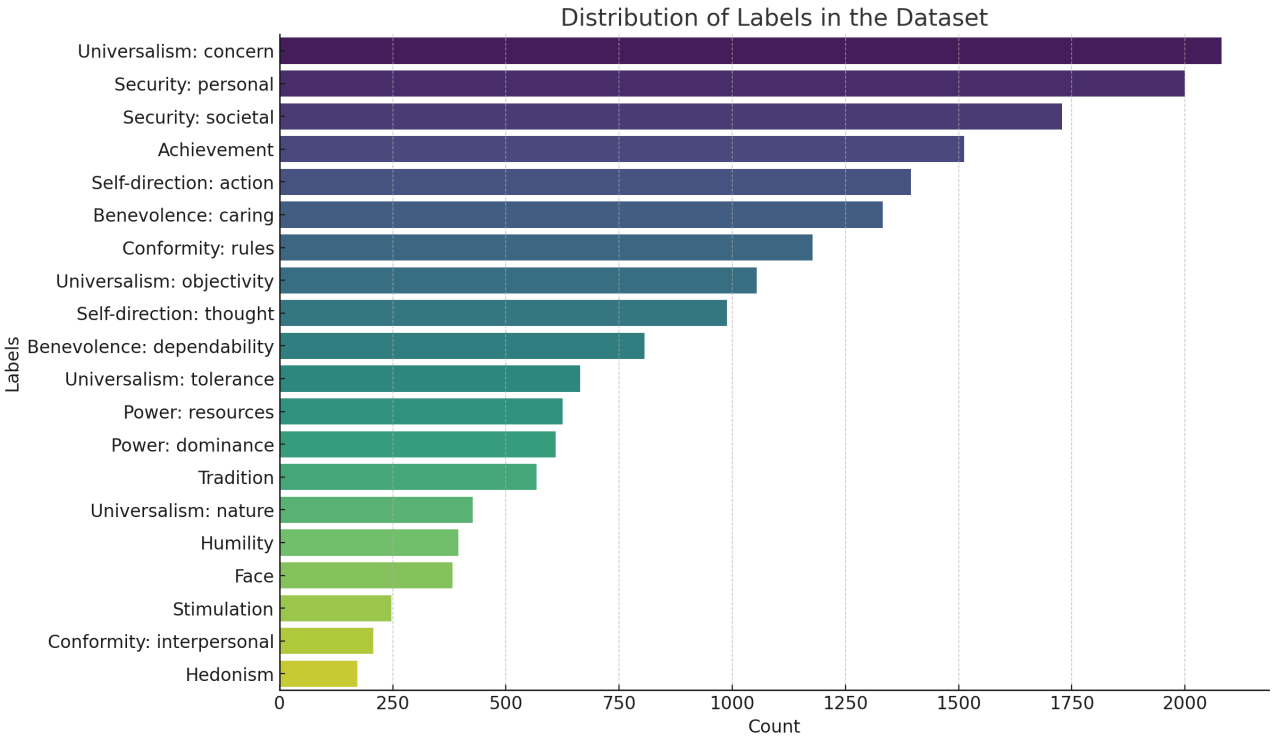

Figura 3.1: Distribución de los valores en el dataset

evaluar de manera más precisa modelos y enfoques para identificar y clasificar los valores humanos presentes en los textos argumentativos.

3.3. IMPLEMENTACIÓN

Para el desarrollo del modelo de detección automática de valores humanos en textos argumentativos, se ha optado por utilizar Python como lenguaje de programación debido a su versatilidad y amplias posibilidades en el Procesamiento del Lenguaje Natural, Machine Learning y Deep Learning.

Para implementar el modelo, se utilizaron las siguientes librerías:

- **PyTorch y PyTorch Lightning**[1]**:** Se utilizó la librería PyTorch junto con PyTorch Lightning para construir y entrenar el modelo. PyTorch es una biblioteca popular en el campo del aprendizaje profundo, que permite crear modelos de manera flexible y realizar operaciones con tensores de forma eficiente. PyTorch Lightning proporciona una interfaz de alto nivel que simplifica el entrenamiento y la evaluación de modelos en PyTorch, lo que facilita la implementación y experimentación.

- **NumPy**[2]**:** Se empleó la librería NumPy para realizar operaciones numéricas y trabajar con matrices de forma eficiente. NumPy es ampliamente utilizado en el análisis de datos y el aprendizaje automático debido a su capacidad para realizar cálculos matemáticos de manera rápida y eficiente.

- **Transformers y Hugging Face**[3]**:** Se utilizó la librería Transformers de Hugging Face para trabajar con modelos preentrenados de BERT y RoBERTa. Esta librería proporciona una interfaz sencilla para cargar modelos preentrenados y utilizarlos para tareas específicas, lo que nos permitió aprovechar el poder de los transformers en nuestro modelo.

- **Datasets:** Se empleó la librería Datasets de Hugging Face para cargar y gestionar el conjunto de datos. Esta librería proporciona una amplia variedad de conjuntos de datos preprocesados y listos para su uso, lo que facilitó la carga y manipulación de los datos para nuestro modelo.

[1] https://pytorch.org/
[2] https://numpy.org/
[3] https://huggingface.co/

- **sklearn**[4]**:** La librería sklearn se utilizó para calcular métricas de evaluación y realizar la división del conjunto de datos. Esta biblioteca proporciona una amplia gama de funciones para evaluar el rendimiento de los modelos y facilita la preparación de los datos para el entrenamiento.
- **SentencePiece y Contractions:** Se emplearon las librerías SentencePiece y Contractions para el preprocesamiento del texto. SentencePiece es una librería que se utiliza para realizar la tokenización de texto en modelos de lenguaje, mientras que Contractions es una librería que permite expandir las contracciones en el texto, lo que contribuyó a mejorar la calidad del preprocesamiento de los datos.
- **TextBlob**[5]**:** Se utilizó la librería TextBlob para realizar análisis de sentimiento en los textos. Esta librería proporciona una interfaz fácil de usar para realizar tareas de procesamiento del lenguaje natural, como análisis de sentimiento, extracción de frases clave, traducción y más.
- **Google Colab y GPU**[6]**:** Todo el desarrollo del modelo se realizó en el entorno de Google Colab, lo que nos permitió aprovechar el poder de las GPUs disponibles en esta plataforma para acelerar el entrenamiento del modelo y mejorar el rendimiento del mismo.

El uso de estas librerías facilitó la implementación y entrenamiento del modelo, permitiéndonos aprovechar el poder de los transformers para abordar la tarea de detección automática de valores humanos en textos argumentativos de manera eficiente y efectiva.

3.4. SOLUCIONES PROPUESTAS

3.4.1. Métricas de Evaluación

Para evaluar el rendimiento de los modelos en la detección de valores humanos en los argumentos, se utilizan las siguientes métricas:

3.4.1.1. Precisión

La precisión es una medida de la exactitud del modelo en la clasificación de las instancias positivas. Se calcula dividiendo el número de verdaderos positivos (TP) entre la suma de verdaderos positivos y falsos positivos (FP). En otras palabras, la precisión nos dice cuántas de las predicciones positivas realizadas por el modelo son realmente correctas en comparación con todas las predicciones positivas. Una alta precisión indica que el modelo tiene una baja tasa de falsos positivos, lo que significa que es cuidadoso en la clasificación de las instancias como positivas.

La fórmula de la precisión es:

$$\text{Precisión} = \frac{TP}{TP + FP}$$

3.4.1.2. Recall (Exhaustividad)

El recall, también conocido como exhaustividad, mide la capacidad del modelo para identificar todas las instancias positivas. Se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos (FN). En otras palabras, el recall nos dice cuántas de las instancias positivas reales en el conjunto de datos fueron identificadas correctamente por el modelo. Un recall alto indica que el modelo puede encontrar con precisión la mayoría de las instancias positivas.

La fórmula del recall es:

$$\text{Recall} = \frac{TP}{TP + FN}$$

3.4.1.3. F1-Score

El F1-score es una medida que combina la precisión y el recall en una sola métrica. Es útil cuando las clases están desequilibradas en el conjunto de datos y queremos tener una medida equilibrada del rendimiento del modelo. Se calcula como la media armónica de la precisión y el recall. El F1-score alcanza su valor máximo de 1 cuando tanto la precisión como el recall son perfectos. Es especialmente útil cuando estamos interesados en el equilibrio entre la precisión y el recall.

[4] https://scikit-learn.org/stable/
[5] https://textblob.readthedocs.io/en/dev/
[6] https://colab.research.google.com/

La fórmula del F1-score es:

$$F1 = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

3.4.1.4. Matriz de Confusión

La matriz de confusión es una tabla que muestra el desempeño del modelo en la clasificación de las instancias en cada clase. Es una herramienta útil para visualizar y analizar el rendimiento del modelo. La matriz de confusión tiene cuatro entradas:

- Verdaderos positivos (TP): El número de instancias positivas que fueron correctamente clasificadas como positivas.
- Falsos positivos (FP): El número de instancias negativas que fueron incorrectamente clasificadas como positivas.
- Verdaderos negativos (TN): El número de instancias negativas que fueron correctamente clasificadas como negativas.
- Falsos negativos (FN): El número de instancias positivas que fueron incorrectamente clasificadas como negativas.

La matriz de confusión se presenta de la siguiente manera:

Clase/Realidad	**Positivo (+)**	**Negativo (-)**
Predicción Positiva (+)	Verdaderos positivos (TP)	Falsos positivos (FP)
Predicción Negativa (-)	Falsos negativos (FN)	Verdaderos negativos (TN)

Para la competición, se medía el rendimiento del modelo utilizando el F1-score de la clase minoritaria, es decir, el valor de interés que se buscaba detectar en los argumentos. Esto es especialmente relevante cuando se tienen clases desequilibradas, ya que nos permite evaluar cómo se comporta el modelo en la detección de instancias de la clase menos frecuente. Un F1-score alto para la clase minoritaria indica que el modelo es capaz de identificar correctamente los argumentos que contienen ese valor específico.

3.4.2. Balanceo de Datos mediante Undersampling

En el conjunto de datos utilizado para la detección de valores humanos en argumentos, se enfrentó un desafío de desequilibrio en la distribución de clases, donde algunas categorías de valores tenían muchas más instancias que otras. Este desequilibrio podía afectar el rendimiento de los modelos, ya que tienden a estar sesgados hacia las clases mayoritarias y a tener un rendimiento deficiente en las clases minoritarias.

Para abordar este problema, se implementó una técnica de undersampling, que consiste en reducir el número de instancias de las clases mayoritarias para igualarlas con las de las clases minoritarias. El objetivo es crear un conjunto de datos más equilibrado que permita que el modelo aprenda de manera más efectiva de las clases minoritarias.

El undersampling se aplicó de la siguiente manera: primero, se contó el número de instancias correspondientes a la clase minoritaria (representada por el valor '1' en la etiqueta, llamada `ntrain`). Luego, se estableció un factor multiplicador, denotado por "p", para seleccionar una cantidad apropiada de instancias de la clase mayoritaria.

Valor de ntrain	**Factor "p"**
$ntrain > 450$	1.6
$290 < ntrain \leq 450$	2.8
$150 < ntrain \leq 290$	5
$ntrain \leq 150$	8

Tabla 3.1: Cálculo del factor "p" basado en el número de instancias de la clase minoritaria "ntrain".

Este enfoque permitió mantener un equilibrio adecuado entre las clases y reducir el impacto del desequilibrio en el rendimiento del modelo. De tal forma que reducimos las instancias de las clases negativas de la siguiente forma:

Al aplicar el undersampling, se logró crear un conjunto de datos más balanceado, lo que permitió a los modelos aprender de manera más efectiva de todas las clases y mejorar el rendimiento en la detección de valores humanos en argumentos (Figura 3.2), especialmente en las categorías minoritarias (Figura 3.3). En estas figuras comparamos el impacto producido por nuestro factor multiplicador `p` con los resultados obtenidos con el dataset original y también con el resultado obtenido al igualar el numero de instancias positivas y negativas.

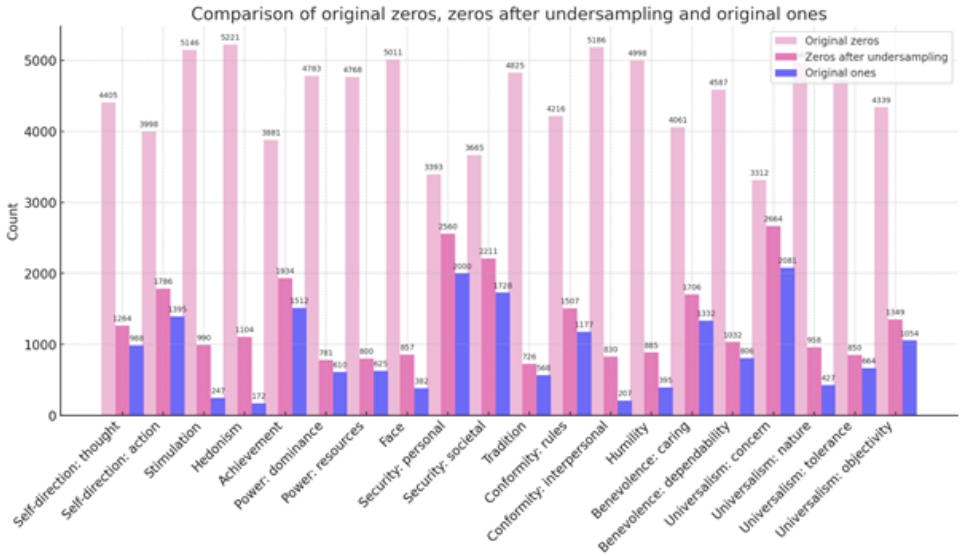

Figura 3.2: Distribución de clases negativas tras *undersampling*

Dataset	Class 1	Class 0
Original	0.51	0.85
Equally balanced	0.58	0.82
With undersampling	0.60	0.86

Tabla 3.2: Impact on F1 Scores of label *"Self-direction: action"*

Dataset	Class 1	Class 0
Original	0.16	0.98
Equally balanced	0.15	0.78
With undersampling	0.30	0.94

Tabla 3.3: Impact on F1 Scores of label *"Stimulation"*

3.4.3. Preprocesado de Texto

El preprocesado de texto es una etapa fundamental en el procesamiento de lenguaje natural que ayuda a preparar los datos para su posterior análisis y modelado. En este trabajo, se aplicó un preprocesado básico que consta de dos pasos:

- **Minúsculas:** Para asegurar la uniformidad y facilitar el procesamiento del texto, todas las letras del corpus se convirtieron a minúsculas. Esto ayuda a evitar problemas de duplicación y facilita la comparación de palabras en minúsculas con las palabras en mayúsculas en el texto. Además, el uso de minúsculas es útil para evitar errores en el procesamiento del lenguaje natural relacionados con las diferencias de mayúsculas y minúsculas.
- **Extensión de Contracciones:** Otro paso importante del preprocesado fue la extensión de contracciones. Las contracciones son combinaciones de dos palabras en una sola palabra, donde se omite una o más letras, generalmente con un apóstrofe. Por ejemplo, "don't.es una contracción de "do not", y çan't.es una contracción de çannot". Al extender las contracciones, se reemplazan con las palabras completas. Este proceso ayuda a reducir la ambigüedad y a asegurar que las palabras sean reconocidas correctamente durante el análisis del texto.

El preprocesado de texto realizado en este trabajo contribuyó a la creación de un conjunto de datos limpio y homogéneo, lo que facilitó el entrenamiento y evaluación de los modelos de detección de valores humanos en argumentos. Con un corpus preprocesado de manera consistente, los modelos pudieron aprender y capturar mejor las características relevantes del lenguaje natural, lo que resultó en un rendimiento mejorado en la tarea de clasificación de valores en los argumentos.

3.4.4. Entrenamiento del Modelo

Para llegar a este punto en la competición, se realizaron diversas etapas y tareas que se resumen a continuación:

1. **Recopilación del Conjunto de Datos:** Se obtuvo el Touché23-ValueEval Dataset, que consta de 5394 argumentos etiquetados con 20 valores humanos. Este conjunto de datos fue recopilado a partir de diversas fuentes, incluyendo textos religiosos, discusiones políticas, editoriales de periódicos y plataformas de democracia en línea.
2. **Preprocesado de Texto:** Se realizó un preprocesado básico del texto, que incluyó la conversión de todas las letras a minúsculas y la extensión de contracciones. Esto aseguró que el corpus estuviera limpio y homogéneo, lo que facilitó el análisis y modelado del lenguaje natural.
3. **Balanceo de Datos:** Dado que el conjunto de datos tenía un desequilibrio en la distribución de clases, se aplicó undersampling para reducir este desequilibrio. Se utilizó un factor multiplicador para determinar el número de instancias de la clase mayoritaria que se reducirían, asegurando así un conjunto de datos equilibrado para el entrenamiento del modelo.
4. **Selección de Modelo y Tokenizador:** Se seleccionaron los modelos de lenguaje BERT y RoBERTa para la clasificación.
5. **Configuración de Parámetros de Entrenamiento:** Durante el proceso de entrenamiento del modelo, se utilizaron los siguientes hiperparámetros: MAX_LEN (tamaño máximo de un argumento después de la tokenización) se estableció en 64 tokens, batch_S (tamaño del lote de argumentos) fue de 32, se entrenó el modelo durante 5 épocas

y se aplicó Early Stopping con una paciencia de 3 épocas. Estos hiperparámetros fueron seleccionados mediante experimentación y ajuste fino para lograr el mejor rendimiento del modelo en la tarea de detección de valores humanos en argumentos.

6. **Entrenamiento y Validación:** Se realizó el entrenamiento del modelo utilizando el conjunto de datos equilibrado y se validó su rendimiento utilizando un conjunto de validación. Se utilizó la métrica F1 de la clase minoritaria como medida principal de rendimiento para la competición.

A continuación, se muestra un diagrama que representa el proceso general de entrenamiento y evaluación del modelo:

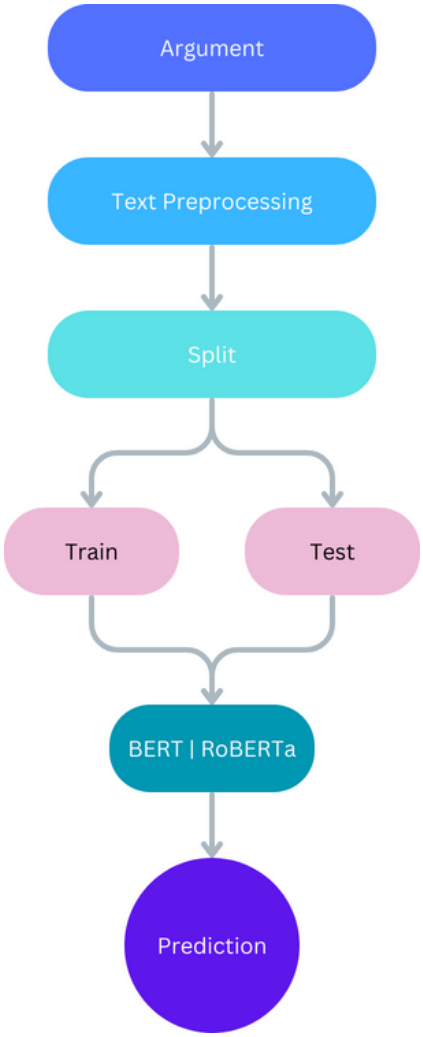

Figura 3.3: Diagrama del Proceso de Entrenamiento y Evaluación del Modelo

3.5. RESULTADOS

Los resultados obtenidos fueron óptimos (véase Tabla 3.4), medido a través de las métricas oficiales, hemos obtenido un mejor resultado en general, sin embargo, podemos apreciar una mejoría notable en aquellas etiquetas más desequilibradas gracias a la técnica de undersampling aplicada (Figura 3.4).

Nuestro mejor propuesta nos otorgó un *f1-score* de 0.46, obteniendo así la vigésimo cuarta posición en la competición.

Test set / Approach	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Main																					
Best per category	.59	.61	.71	.39	.39	.66	.50	.57	.39	.80	.68	.65	.61	.69	.39	.60	.43	.78	.87	.46	.58
Best approach	.56	.57	.71	.32	.25	.66	.47	.53	.38	.76	.64	.63	.60	.65	.32	.57	.43	.73	.82	.46	.52
BERT	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
BERT (1st run)	.46	.43	.60	.25	.26	.53	.30	.44	.29	.71	.57	.47	.52	.27	.22	.50	.28	.68	.70	.40	.41
roBERTa (2nd run)	.41	.32	.58	.29	.18	.39	.26	.49	.14	.71	.58	.50	.48	.00	.07	.46	.23	.68	.73	.32	.52
BERT (3rd run)	.45	.42	.59	.23	.25	.58	.33	.46	.28	.70	.59	.43	.50	.28	.20	.49	.27	.70	.68	.38	.47

Tabla 3.4: F_1-score del equipo Marquis-de-Sade para todas las categorias. Las propuestas en gris se muestran para comparación con el modelo BERT del organizador y el baseline, así como la mejor propuesta y mejor resultado por categoría.

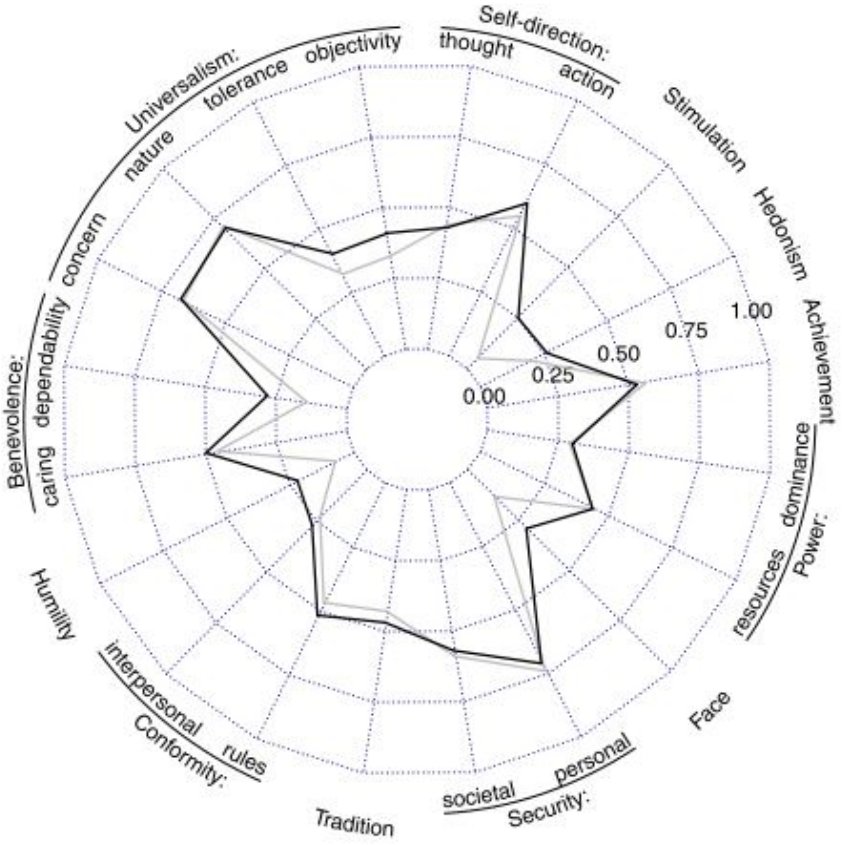

Figura 3.4: BERT baseline (gris) en comparación con nuestra mejor propuesta

CAPÍTULO 4

Propuestas de Mejora

Una vez finalizada la competición y con los datos de prueba etiquetados a nuestra disposición, tuvimos la oportunidad de seguir experimentando con nuestros modelos de una forma más dinámica. Nuestro objetivo era mejorar los resultados obtenidos en la competición. Para ello, seguimos las estrategias que se muestran a continuación.

4.1. OPTIMIZACIÓN DE HIPERPARÁMETROS

Para mejorar el rendimiento de nuestros modelos, adoptamos los hiperparámetros utilizados por el ganador de la competición[1], ya que demostraron obtener resultados significativos[10]. Los hiperparámetros que utilizamos son los siguientes:

- Tamaño del lote (BATCH_SIZE): 8
- Número de épocas de entrenamiento (NUM_TRAIN_EPOCHS): 10
- Tasa de aprendizaje (LEARNING_RATE): 5e-5
- Longitud máxima de la secuencia (MAX_LENGTH): 128
- Decaimiento del peso (WEIGHT_DECAY): 0.01

4.2. MODELOS ENTRENADOS

Utilizamos estos hiperparámetros para entrenar lo siguientes modelos: BERT (v1) y DeBERTa (v1).

BERT fue nuestra elección inicial por una razón de peso: su probada capacidad en diversas tareas lingüísticas. Este modelo, que utiliza representaciones bidireccionales, ha establecido nuevos estándares en el entendimiento contextual de las palabras, considerando la información a su izquierda y derecha simultáneamente. Su éxito en la competición inclinó la balanza hacia el uso este modelo.

Sin embargo, no nos detuvimos ahí. Introdujimos DeBERTa en nuestra propuesta, convencidos de las mejoras sustanciales que aporta en relación a BERT. DeBERTa[4] redefine la auto-atención al separar la atención basada en contenido de la basada en posición relativa. Adicionalmente, al introducir capas adicionales que refuerzan las interacciones entre palabras, DeBERTa promete una precisión aún mayor.

Al fusionar BERT y DeBERTa, buscamos consolidar una propuesta que no solo garantice precisión sino aprovechar las fortalezas complementarias de ambos modelos.

4.2.1. Primera Propuesta

Primero, nos enfocamos en entrenar los modelos mencionados anteriormente. Después de entrenar los dos modelos con los nuevos hiperparámetros, decidimos combinarlos en un modelo en conjunto (Ensemble v1) con nuestra mejor propuesta en la competición (BERT 1st run). Esto se hizo con la esperanza de que este modelo pudiera agregar alguna diversidad al conjunto, ya que fue entrenado con diferentes hiperparámetros. La idea detrás de este enfoque es que diferentes modelos pueden ser buenos en diferentes aspectos de la tarea y, al combinarlos, podemos beneficiarnos de las fortalezas de todos ellos. Como podemos apreciar en la Tabla 4.1, hemos utilizado un sistema de votación blanda usando los siguientes pesos:

[1] https://github.com/danielschroter/human_value_detector

Ejecución	Peso
BERT (1st run)	0.50
BERT (v1)	0.25
DeBERTa (v1)	0.25

Tabla 4.1: Pesos utilizados para el soft-voting

La decisión de darle más valor a nuestra primera propuesta es la siguiente: Al haber entrenado una a una cada una de las etiquetas y, haberles aplicado undersampling obteniendo de esta forma tan buenos resultados en las etiquetas más desequilibradas, obtenemos un mejor resultado (véase Tabla 4.2) pues los modelos multietiquetas otorgan peores resultados en las etiquetas más desbalanceadas. Se observan unos resultados constantes y ligeramente superiores, sin embargo cabe destacar que algunas categorías como *Universalism: Objectivity* han recibido una mejoría del *34 %*.

Test set / Approach	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Main																					
BERT	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
BERT (1st run)	.46	.43	.60	.25	.26	.53	.30	.44	.29	.71	.57	.47	.52	.27	.22	.50	.28	.68	.70	.40	.41
BERT (v1)	.46	.37	.61	.13	.28	.60	.33	.49	.06	.74	.59	.49	.50	.15	.05	.47	.19	.68	.73	.41	.53
deBERTa (v1)	.44	.41	.58	.07	.14	.56	.33	.50	.20	.69	.59	.52	.49	.21	.06	.45	.20	.65	.74	.41	.57
Ensemble (v1)	.47	.44	.61	.25	.27	.59	.33	.45	.30	.73	.59	.49	.53	.29	.21	.48	.29	.69	.70	.43	.55

Tabla 4.2: F_1-score para todas las categorias. Las propuestas en gris se muestran para comparación con el modelo BERT del organizador y el baseline.

4.2.2. Segunda Propuesta

4.2.2.1. Aumento de datos con *back-translation*

Finalmente, con el objetivo de mejorar aún más el rendimiento de nuestro modelo de conjunto, decidimos explorar técnicas de aumento de datos. En particular, utilizamos la técnica de retro-traducción para aumentar nuestro conjunto de datos. El aumento de datos mediante retro-traducción implica traducir un texto a un idioma diferente y luego volver a traducirlo al idioma original. Esto a menudo resulta en una reescritura del texto original que conserva el mismo significado pero utiliza diferentes palabras y estructuras sintácticas. Concretamente hemos realizado la traducción apoyándonos en el español.

Después de aumentar nuestro conjunto de datos con *back-translation*, volvimos a entrenar nuestros modelos y los combinamos en un nuevo *Ensemble v2* realizado a través de un sistema de votación dura (otorgando el mismo peso a cada predicción).

Como podemos apreciar en la Tabla 4.3, el resultado obtenido ha mejorado bastante hasta *0.48*, aún así hemos obtenido una peor puntuación en algunas categorías como **Stimulation, Humility o Universalism: Objectivity**, por ello hemos realizado otro ensemble el cual, para estas etiquetas, tenga en cuenta en mayor medida la predicción de nuestro primer ensemble *Ensemble(v1)*, otorgándole a estas un mayor peso.

Test set / Approach	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Main																					
Best approach	.56	.57	.71	.32	.25	.66	.47	.53	.38	.76	.64	.63	.60	.65	.32	.57	.43	.73	.82	.46	.52
BERT	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
BERT (1st run)	.46	.43	.60	.25	.26	.53	.30	.44	.29	.71	.57	.47	.52	.27	.22	.50	.28	.68	.70	.40	.41
BERT (v1)	.46	.37	.61	.13	.28	.60	.33	.49	.06	.74	.59	.49	.50	.15	.05	.47	.19	.68	.73	.41	.53
deBERTa (v1)	.44	.41	.58	.07	.14	.56	.33	.50	.20	.69	.59	.52	.49	.21	.06	.45	.20	.65	.74	.41	.57
Ensemble (v1)	.47	.44	.61	.25	.27	.59	.33	.45	.30	.73	.59	.49	.53	.29	.21	.48	.29	.69	.70	.43	.55
BERT (v2)	.46	.47	.59	.18	.26	.59	.32	.53	.22	.73	.58	.48	.53	.40	.14	.50	.30	.69	.74	.36	.47
deBERTa (v2)	.46	.45	.66	.13	.23	.60	.36	.46	.24	.73	.61	.51	.53	.23	.14	.52	.32	.69	.76	.40	.50
Ensemble (v2)	.48	.47	.65	.20	.25	.60	.36	.52	.28	.75	.62	.51	.56	.32	.16	.54	.32	.71	.77	.42	.45

Tabla 4.3: F_1-score para todas las categorias. Las propuestas en gris se muestran para comparación con el modelo BERT del organizador, el baseline y la propuesta ganadora de la competición junto con las propuestas anteriores.

De esta forma, como se muestra en la Tabla 4.4, se ha conseguido obtener un *f1-score* de *0.49*, lo que supone un incremento del 7 % del resultado obtenido en la competención.

Test set / Approach	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Main																					
Best approach	.56	.57	.71	.32	.25	.66	.47	.53	.38	.76	.64	.63	.60	.65	.32	.57	.43	.73	.82	.46	.52
BERT	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
BERT (1st run)	.46	.43	.60	.25	.26	.53	.30	.44	.29	.71	.57	.47	.52	.27	.22	.50	.28	.68	.70	.40	.41
BERT (v1)	.46	.37	.61	.13	.28	.60	.33	.49	.06	.74	.59	.49	.50	.15	.05	.47	.19	.68	.73	.41	.53
deBERTa (v1)	.44	.41	.58	.07	.14	.56	.33	.50	.20	.69	.59	.52	.49	.21	.06	.45	.20	.65	.74	.41	.57
Ensemble (v1)	.47	.44	.61	.25	.27	.59	.33	.45	.30	.73	.59	.49	.53	.29	.21	.48	.29	.69	.70	.43	.55
BERT (v2)	.46	.47	.59	.18	.26	.59	.32	.53	.22	.73	.58	.48	.53	.40	.14	.50	.30	.69	.74	.36	.47
deBERTa (v2)	.46	.45	.66	.13	.23	.60	.36	.46	.24	.73	.61	.51	.53	.23	.14	.52	.32	.69	.76	.40	.50
Ensemble (v2)	.48	.47	.65	.20	.25	.60	.36	.52	.28	.75	.62	.51	.56	.32	.16	.54	.32	.71	.77	.42	.45
Ensemble (v3)	.49	.48	.66	.25	.27	.61	.37	.52	.30	.76	.61	.52	.56	.31	.21	.52	.32	.71	.77	.43	.51

Tabla 4.4: F_1-score para todas las categorias. Las propuestas en gris se muestran para comparación con el modelo BERT del organizador, el baseline y la propuesta ganadora de la competición junto con las propuestas anteriores.

CAPÍTULO 5
Conclusiones

5.1. CONCLUSIONES Y TRABAJO FUTURO

5.1.1. Conclusiones

En el transcurso de este trabajo de fin de grado, abordamos el desafiante ámbito del procesamiento del lenguaje natural, con un enfoque específico en la tarea de identificar valores humanos en argumentos. Iniciamos nuestro proceso alistándonos en la competición *ValueEval*[1],en la cual nos enfocamos en usar undersampling y un entrenamiento individual para cada etiqueta, lo que sentó una base sólida para nuestras futuras exploraciones y adaptaciones.

Inspirados por el éxito de otros competidores[10], adoptamos hiperparámetros que demostraron ser eficaces y emprendimos una serie de experimentaciones meticulosas para refinar nuestro modelo. En el epicentro de nuestras mejoras estuvieron los modelos BERT y deBERTa, que se convirtieron en herramientas esenciales para nuestra investigación. A través de un enfoque de ensamblaje, fusionamos estos modelos con nuestra primera propuesta, permitiendo que estos se complementaran, conduciendo a una mejora notable en el rendimiento.

Para añadir aún más robustez a nuestros modelos, nos aventuramos en el terreno del aumento de datos, utilizando específicamente la técnica de retrotraducción. Esta metodología, que generó variaciones en nuestro dataset manteniendo intacto el significado original, realzó la robustez de nuestros modelos en términos de interpretación y clasificación precisa de argumentos.

Pasando de nuestro modelo BERT inicial al último ensemble (v3), los resultados, presentados en la Tabla 4.4, delinean claramente nuestro avance y logros en este proyecto. Es particularmente interesante observar cómo, incluso con las notables mejoras generales, ciertas categorías, como Stimulation, Humility y Universalism: Objectivity, aún ofrecen espacio para mejoras adicionales. Además, las estrategias de ensamblaje implementadas y las técnicas de aumento de datos con ”backtranslation” no sólo solidificaron nuestro modelo general, sino que también dirigieron nuestra atención hacia una mejora en la precisión en las etiquetas más desbalanceadas, validando así nuestra hipótesis inicial sobre la eficacia de abordar conjuntos de datos desequilibrados y el uso de múltiples modelos para maximizar el rendimiento. Cabe destacar también que todos estos experimentos han sido realizados a través de *google collab básico*, el cuál no nos daba pié a usar modelos más avanzados los cuales si fueron usados por otros competidores y que, aún así, tras estas mejoras hemos conseguido avanzar hasta el puesto numero *11*.

Al reflexionar sobre este trabajo, queda claro que, con la combinación adecuada de técnicas de preprocesamiento, modelos avanzados y métodos de ensamblaje, es posible avanzar significativamente en tareas complejas de procesamiento del lenguaje natural y abrir caminos prometedores para investigaciones futuras.

5.1.2. Trabajo futuro

A medida que avanzamos en nuestra investigación sobre la identificación de valores humanos mediante el procesamiento del lenguaje natural, hay múltiples caminos y estrategias prometedoras a considerar. Estos son algunos potenciales enfoques para mejorar aún más nuestros modelos:

1. **Técnicas de Balanceo de Datos:** Aunque ya hemos explorado el ”undersampling” y el ”oversampling” podrían aplicarse combinaciones de ambos que podrían aportar soluciones más efectivas al desafío de conjuntos de datos desequilibrados.
2. **Diversificación del Aumento de Datos:** La retrotraducción ha demostrado ser efectiva, pero existen otras técnicas como la paráfrasis automática o la generación de texto[1] que podrían introducir aún más variabilidad y enriquecimiento en nuestro conjunto de datos.

[1] https://touche.webis.de/semeval23/touche23-web/index.html

3. **Ensamble de nuevos Modelos:** En el epicentro de nuestra experimentación estuvieron BERT y deBERTa. Sin embargo, podríamos incorporar modelos como RoBERTa, ALBERT.
4. **Optimización Avanzada de Hiperparámetros:** Aunque hemos adoptado hiperparámetros basados en buenos resultados obtenidos en la competición, herramientas más avanzadas como WanDB[2] podrían otorgar unos mejores, perfeccionando aún más nuestros modelos.

En conclusión, este trabajo ha sido un gran paso en el ámbito del procesamiento del lenguaje natural para identificar valores humanos. Sin embargo, el campo es muy amplio y se encuentra en constante evolución por lo que confiamos en que las técnicas y enfoques futuros continuarán avanzando en gran medida obteniendo así avances mas significativos en este dominio.

[2] https://wandb.ai/site

Bibliografía

[1] Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. *International journal of machine learning and cybernetics*, 14(1):135–150, 2023.

[2] Nordin El Balima Cordero, Jacinto Mata Vázquez, Victoria Pachón Álvarez, and Abel Pichardo Estevez. I2C Huelva at SemEval-2023 task 4: A resampling and transformers approach to identify human values behind arguments. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1382–1387, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Yihong Han, Yoko Nishihara, and Junjie Shan. Human values estimation on news articles through bert-extracted opinion expressions. In *2022 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 95–100. IEEE, 2022.

[4] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[5] Maciej Hercog, Piotr Jaroński, Jan Kolanowski, Paweł Mieczyński, Dawid Wiśniewski, and Jedrzej Potoniec. Sarcastic roberta: A roberta-based deep neural network detecting sarcasm on twitter. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 46–52. Springer, 2022.

[6] MV Koroteev. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021.

[7] Benjamin Lindemann, Benjamin Maschler, Nada Sahlab, and Michael Weyrich. A survey on anomaly detection for technical systems using lstm networks. *Computers in Industry*, 131:103498, 2021.

[8] Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, et al. The touch\'e23-valueeval dataset for identifying human values behind arguments. *arXiv preprint arXiv:2301.13771*, 2023.

[9] Serhad Sarica and Jianxi Luo. Stopwords in technical language processing. *Plos one*, 16(8):e0254937, 2021.

[10] Daniel Schroter, Daryna Dementieva, and Georg Groh. Adam-smith at SemEval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 532–541, Toronto, Canada, July 2023. Association for Computational Linguistics.

[11] Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. Bert for stock market sentiment analysis. In *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)*, pages 1597–1601. IEEE, 2019.

[12] Ankit Thakkar, Dhara Mungra, Anjali Agrawal, and Kinjal Chaudhari. Improving the performance of sentiment analysis using enhanced preprocessing technique and artificial neural network. *IEEE Transactions on Affective Computing*, 13(4):1771–1782, 2022.

[13] Amit Kumar Tyagi and Ajith Abraham. Recurrent neural networks: Concepts and applications. 2022.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

ANEXOS

ANEXO A

Paper Científico

I2C Huelva at SemEval-2023 Task 4: A Resampling and Transformers Approach to Identify Human Values behind Arguments

Nordin El Balima Cordero, Jacinto Mata Vázquez,
Victoria Pachón Álvarez, Abel Pichardo Estévez
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
nordin.elbalima531@alu.uhu.es, mata@dti.uhu.es,
vpachon@dti.uhu.es, abel.pichardo107@alu.uhu.es

Abstract

This paper presents the approaches proposed for I2C Group to address the SemEval-2023 Task 4: Identification of Human Values behind Arguments (ValueEval)" (Kiesel et al., 2023), whose goal is to classify 20 different categories of human values given a textual argument. The dataset of this task consists of one argument per line, including its unique argument ID, conclusion, stance of the premise towards the conclusion and the premise text. To indicate whether the argument draws or not on that category a binary indication (1 or 0) is included. Participants can submit approaches that detect one, multiple, or all of these values in arguments. The task provides an opportunity for researchers to explore the use of automated techniques to identify human values in text and has potential applications in various domains such as social science, politics, and marketing. To deal with the imbalanced class distribution given, our approach undersamples the data. Additionally, the three components of the argument *(conclusion, stance and premise)* are used for training. The system outperformed the BERT baseline according to official evaluation metrics, achieving a *f1 score* of 0.46.

1 Introduction

Human values refers to the beliefs, principles and standards that individuals or groups hold to be important and worthwhile. These values guide people´s attitudes and behaviors; they can vary across cultures (Civitillo et al., 2019), communities, and individuals. According to a study of *Global Values Survey* (White et al., 2020), the most widely held values across the world are a sense of community, a sense of national pride, and a desire for social order. Other commonly held values include equality, respect for others, and a desire for a peaceful world.

In this context, the classification of human values in textual arguments is an important task in the field of *Natural Language Processing*. Understanding the values that underlie an argument could provide valuable insights into people´s beliefs, attitudes, and motivations. It could also be useful in various applications such as opinion analysis (Hemmatian and Sohrabi, 2019), argumentation mining, emotion recognition, and persuasive technology, among others.

Despite its importance, the automatic classification of human values in arguments remains a challenging problem. The task requires the ability to identify values in a text, understand the argument structure, and make a binary judgment about the presence of a value in the argument. It aims to advance the state of the art in human value classification in textual arguments , these textual arguments are compiled from the social science literature and described in detail in the accompanying ACL paper (Kiesel et al., 2022).

ValueEval had the advantage of bringing together 40 teams, taking in 112 different runs *(including the competition organizers)*. They provided a set of labelled data to solve the task. This dataset was highly imbalanced over the 20 categories, with a large number of instances belonging to the negative class. Undersampling (Arefeen et al., 2020) is a common technique used in machine learning to balance the class distribution in imbalanced datasets. To tackle this problem, we employed an undersampling strategy to decrease the number of instances within the negative class. On the other hand, we conducted experiments using various combinations of the information provided for the arguments.

Finally, our approach uses the premise, conclusion and stance of the argument as input features. These three components provide important information about the argument to help identify the values. In this work, we utilized transfer learning techniques by fine-tuning state-of-the-art pre-trained language models, using the transformers library. This approach allowed us to leverage the knowl-

edge learned from large-scale datasets and apply it to our specific task of argumentative text classification.

The results of our experiments show that our approach is effective in classifying arguments based on human values. The combination of undersampling and the use of all the data given leads to improved performance compared to other methods. Our findings contribute to a better understanding of how human values can be identified in arguments and have implications for a range of applications, including opinion analysis and argumentation (Lawrence and Reed, 2020). However, as is common in such tasks, some categories have a low number of positive instances, i.e. instances where the argument draws on that category. This low number of positive instances can pose a challenge for machine learning algorithms, as they may not have enough examples to learn from. This can result in overfitting, where the algorithm memorizes the training data instead of generalizing to new data, therefore the model may not be able to accurately predict the outcome for new instances.

2 Background

The input data for the study consists of two tab-separated value files, "arguments-training.tsv" and "labels-training.tsv". Both contains 5,394 rows, on arguments each row represents a unique argument with its ID, conclusion, stance and premise, Figure 1 shows an example of an argument from the training dataset. . For the label each row corresponds to the unique argument ID, and one column for each of the 20 value categories, indicating whether the argument aligns with the particular value category (1) or not (0). The dataset used in this paper was extracted from the descriptions of articles in arXiv, as described in the source (Mirzakhmedova et al., 2023).

[Argument ID] *A01010*
[Conclusion] *We should prohibit school prayer*
[Stance] *against*
[Premise] *it should be allowed if the student wants to pray as long as it is not interfering with his classes.*

Figure 1: Example of argument from training dataset

The identification of human values behind arguments is an important aspect of argument mining. Some work have been researched in this area, which aims to extract natural language arguments and their relations from text (Cabrio and Villata, 2018), there are a lot of use cases like (Passon et al., 2018) predicting the usefulness of online reviews based solely on the amount of argumentative text that they contain, or finding relevant evidence (on argument premises) in the study of adjudication decisions about veteran´s claims for disability (Walker et al., 2018)

3 System Overview

The task of classifying textual arguments based on human values categories is challenging due to the subjective nature of human values. However, the system was able to address this challenge by using advanced deep learning algorithms such as BERT and RoBERTa.

3.1 Implemented Models

For our argument classification task, we employed the BERT and RoBERTa models. Both of these models are based on the transformer architecture and have been pre-trained on massive amounts of text data. BERT, short for Bidirectional Encoder Representations from Transformers, was introduced by (Devlin et al., 2019) and has achieved state-of-the-art performance on various natural language processing tasks. RoBERTa, a variant of BERT, was introduced by (Liu et al., 2019) and further improved the pre-training process by optimizing hyperparameters and using larger batch sizes. We used the Hugging Face[1] library to fine-tune the pre-trained BERT and RoBERTa models for our task. We employed the common hyperparameters, including batch size, learning rate, and weight decay, and used early stopping with a maximum of five epochs. We did not tune any specific hyperparameters.Our choice of these models was based on their proven success in various natural language processing tasks and their pre-training on large amounts of text data. We fine-tuned the models on our argument classification task to leverage their ability to understand and extract meaningful information from natural language text. Recent studies have demonstrated the effectiveness of these models in various tasks, such as sentiment analysis (Khan and Fu, 2021), question answering (Ju et al., 2019), and document classification (Liu et al., 2021).

[1] https://huggingface.co/

Class	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
0	1264	1786	990	1104	1934	781	800	857	2560	2211	726	1507	830	885	1706	1032	2664	958	850	1349
1	790	1116	198	138	1209	488	500	306	1600	1382	454	942	166	316	1066	645	1665	342	531	843

Table 1: Number of examples of each class after performing *Undersampling*.

3.2 Selected Inputs

We experimented with different input formats for our models, including using just the conclusion or just the premise as input. However, we found that using the conclusion stance and premise together as a single input yielded the best results in our experiments. This is because the conclusion stance provides important context for the premise, allowing the model to better understand the argument being made. To illustrate the impact of using different input formats, we show the F1 scores for each format in Table 2. As can be seen, the model using the combination of conclusion stance and premise as input achieved the highest F1 score.[2]

C+S+P	Premise	Conclusion
0.73	0.71	0.56

Table 2: Impact of Input Format on F1 Scores in minor class of label *"Security: Personal"*

3.3 Preprocessing Text

We have applied text preprocessing to clean and simplify the text:

- Conversion of all characters to lowercase: All the characters in the text were converted to lowercase for consistency and ease of processing.
- Expansion of all possible contractions in English: Contractions such as "don't" were expanded to "do not" and "can't" was expanded to "cannot" to ensure that the model could understand the text properly.
- Removal of special characters: Special characters such as punctuation marks and symbols were removed from the text. This helped to simplify the text and remove any unnecessary noise.
- Removal of multiple spaces between characters: Multiple spaces between characters were reduced to a single space. This was done to ensure that there was consistency in the text and that all the spaces were uniform.

By applying these text processing techniques, we were able to simplify and clean the text, making it easier to process and analyze.

3.4 Undersampling Techniques

In order to address the issue of class imbalance in the provided training dataset, we implemented an undersampling technique that reduces the size of the majority class to increase the representation of the minority class. Specifically, we used a multiplier on the size of the majority class to determine the number of samples to keep for each label, with the multiplier being determined by the size of the minority class. To ensure the models had sufficient data to train on, we trained at least with 1000 arguments. Table 1 shows the distribution of argument categories after performing undersampling. Overall, these preprocessing steps helped improve (see Table 3) the models' performance on the data.

Original dataset	With Undersampling
0.53	0.57

Table 3: Impact of Undersampling on F1 Scores in minor class of label *"Security: societal"*

[2]Note C+S+P stands for Conclusion+Stance+Premise.

4 Experimental Setup

For our experimental setup, we used a train-validation-test split to evaluate the performance of our models. We split the dataset into 80% for training, 10% for validation, and 10% for testing. This allowed us to train our models on a sufficiently large amount of data while still having a separate set of data to test the models' generalization ability. We chose not to use the validation dataset during the training process because its arguments were substantially different from those in the test dataset, which could have affected the generalization ability of our models. Therefore, we solely relied on the training and test datasets for all the experiments. We used the PyTorch library and the transformers package to implement our models. We used the Hugging Face Transformers library to fine-tune pre-trained transformer models for the argument classification task. TIRA (Fröbe et al., 2023) is the platform used for the shared task that we submitted our system to.

For the experiment, the models were trained with 5 epochs, 32 batch size, 64 token length. Early stopping was used to avoid overfitting while training. Labels like *"Stimulation"* only left us with 198 positive examples and 4116 negatives for our training dataset. This made the identification of values a more difficult task.

By balancing it to an equal number of *0s* and *1s*, we have observed worse results compared to the original one. Then, to improve our approach which is identifying the human value, the dataset has been adjusted to bring at least 1000 arguments for training, which has shown better results as we can see in Table 4, our model gave us a *56%* improvement over the original one.

Dataset	Class 1	Class 0
Original	0.16	0.98
Equally balanced	0.15	0.78
With undersampling	0.30	0.94

Table 4: Impact on F1 Scores of label *"Stimulation"*

In addition to that, we have more balanced labels like *"Self-direction: action"* with 1116 positive examples and 3198 negatives. With our undersampling approach (see Table 5) we improved the result by *18%*.

Dataset	Class 1	Class 0
Original	0.51	0.85
Equally balanced	0.58	0.82
With undersampling	0.60	0.86

Table 5: Impact on F1 Scores of label *"Self-direction: action"*

5 Results

Our system achieved performance above the BERT baseline (shown in Figure 2), as measured by official evaluation metrics, we got better results overall. But the biggest gap is in more imbalanced classes like *"Stimulation", "Humility" and "Face"* where we nearly doubled the performance. Table 6 shows the overall results. In the competition, there were 39 participating teams. Our best-performing system was a *BERT* pretrained model with resampling. This approach achieved a *f1 score* of 0.46, obtaining the 24th position in the final leaderboard, demonstrating its effectiveness.

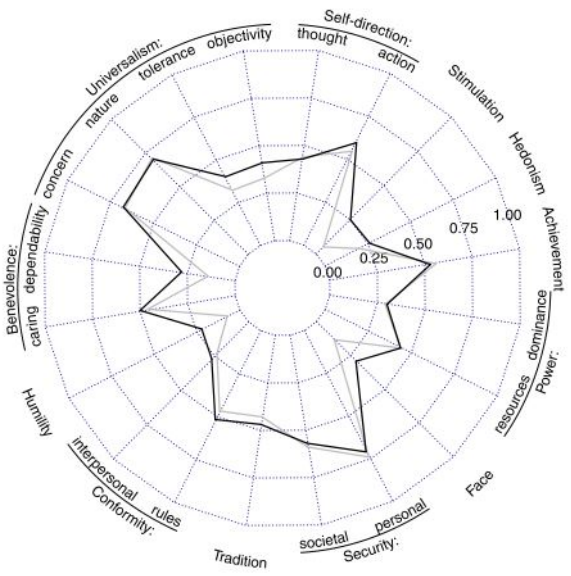

Figure 2: BERT *(grey)* baseline comparison with our best approach *(black)*

6 Conclusion

To conclude, our system for detection of human values behind arguments achieved competitive results, as it presents better results than the baseline BERT model. Our approach of utilizing undersampling to balance the dataset and focusing on the positive class proved to be effective, especially given

Test set / Approach	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Main																					
Best per category	.59	.61	.71	.39	.39	.66	.50	.57	.39	.80	.68	.65	.61	.69	.39	.60	.43	.78	.87	.46	.58
Best approach	.56	.57	.71	.32	.25	.66	.47	.53	.38	.76	.64	.63	.60	.65	.32	.57	.43	.73	.82	.46	.52
BERT	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
BERT (1st run)	.46	.43	.60	.25	.26	.53	.30	.44	.29	.71	.57	.47	.52	.27	.22	.50	.28	.68	.70	.40	.41
roBERTa (2nd run)	.41	.32	.58	.29	.18	.39	.26	.49	.14	.71	.58	.50	.48	.00	.07	.46	.23	.68	.73	.32	.52
BERT (3rd run)	.45	.42	.59	.23	.25	.58	.33	.46	.28	.70	.59	.43	.50	.28	.20	.49	.27	.70	.68	.38	.47

Table 6: Achieved F_1-score of team marquis-de-sade per test dataset, from macro-precision and macro-recall (All) and for each of the 20 value categories. Approaches in gray are shown for comparison: an ensemble using the best participant approach for each individual category; the best participant approach; and the organizer's BERT and 1-Baseline.

the nature of the competition where identifying it is more crucial than not identifying it. However, there is still room for improvement. One potential direction for future work is to explore data augmentation techniques to increase the size of our training arguments and potentially improve model performance. There is also the possibility of extending our system to other languages, which would require additional preprocessing and potentially the development of language-specific models. Finally, investigating the performance of other transformer-based models and their variants on this task could also lead to better results in future researches.

7 Acknowledgments

This paper is part of the I+D+i Project titled *"Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]", PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"*.

References

Md Adnan Arefeen, Sumaiya Tabassum Nimi, and M Sohel Rahman. 2020. Neural network-based undersampling techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(2):1111–1120.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Sauro Civitillo, Linda P Juang, Marcel Badra, and Maja K Schachner. 2019. The interplay between culturally responsive teaching, cultural diversity beliefs, and self-reflection: A multiple case study. *Teaching and Teacher Education*, 77:341–351.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Fatemeh Hemmatian and Mohammad Karim Sohrabi. 2019. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial intelligence review*, 52(3):1495–1545.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *arXiv preprint arXiv:1909.10772*.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3034–3042.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. Semeval-2023 task 4: Identification of human values behind arguments. In *17th International Workshop on Semantic Evaluation (SemEval-2023)*.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. *CoRR*, abs/2301.13771.

Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the usefulness of amazon reviews using off-the-shelf argumentation mining. *arXiv preprint arXiv:1809.08145*.

Vern R. Walker, Dina Foerster, Julia Monica Ponce, and Matthew Rosen. 2018. Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment. In *Proceedings of the 5th Workshop on Argument Mining*, pages 68–78, Brussels, Belgium. Association for Computational Linguistics.

Cindel White, Michael Muthukrishna, and Ara Norenzayan. 2020. Worldwide evidence of cultural similarity among co-religionists within and across countries using the world values survey. *PsyArXiv. https://psyarxiv. com/uetg6*.