Trabajo Fin de Grado

Grado en Filología Hispánica

Más allá del texto: análisis multimodal de sentimientos.

Autora: María Rebollo Sevilla

Tutora: Raquel Benítez Burraco

UNIVERSIDAD DE SEVILLA: FACULTAD DE FILOLOGÍA HISPÁNICA
AÑO 2023/24

Índice

1. INTRODUCCIÓN

En el área del procesamiento del lenguaje natural y la inteligencia artificial el análisis de sentimientos ha recibido mucha atención en las últimas décadas; sin embargo, estos análisis se centran únicamente en el procesamiento de textos. En este trabajo, se explora una perspectiva innovadora e interdisciplinaria: el análisis de emociones multimodal, que comprende los campos de estudio de la lingüística computacional, el procesamiento lingüístico de la inteligencia artificial, la inteligencia artificial multimodal y la manifestación lingüística de las emociones. La finalidad de la creación de un modelo multimodal para el análisis de sentimientos en redes sociales es integrar información textual (que puede estar o no codificada con el lenguaje figurado), información visual y contexto, ya sea mediante imagen, tipografía, audio, vídeo, etc.; todo ello para que sea más sencillo para la computadora asimilar la complejidad del contenido emocional del mensaje y descubrir nuevas formas de comprender y representar expresiones o eventos emocionales, algo que el ser humano hace de manera automática. La metodología incluye una muestra de etiquetado de corpus, que puede ayudar a superar las limitaciones actuales descodificando elementos emocionales metafóricos o metonímicos; la presentación de un modelo estadístico para el análisis de sentimientos y el análisis de los resultados obtenidos del funcionamiento de modelos multimodales.

2. OBJETIVOS

Es necesario establecer metas claras y específicas para orientar la investigación de manera efectiva. Por tanto, los objetivos de la investigación son los siguientes:

- Explorar la relación entre la lingüística computacional y la inteligencia artificial multimodal.
- Desarrollar un modelo multimodal de análisis para predecir y analizar emociones.
- Evaluar la efectividad del modelo propuesto.
- Comparar la efectividad del modelo multimodal con enfoque unimodal.
- Explorar el papel del filólogo en este ámbito.

3. MARCO TEÓRICO

3.1 ¿Qué es la lingüística computacional[1]?

Para Grishman, la Lingüística Computacional (a partir de ahora, LC) se puede definir como ‘‘el estudio de sistemas computacionales utilizados para la comprensión y evolución de lenguas naturales’’ (2021, p. 128).

Desde, aproximadamente, los años cuarenta, ha sido difícil encasillar esta disciplina dentro de una rama científica concreta. Lo que es innegable es que esta nueva realidad no puede pertenecer a un solo campo. De esta forma, hay definiciones que la encasillan como parte de la Lingüística general, tanto si es la Lingüística Teórica[2] como si es Lingüística Aplicada. Otras definiciones prefieren enmarcarla en el campo de la Informática, más concretamente en la Inteligencia Artificial[3] (para ver definición, véase Anexo A).

En los estudios[4] que se han hecho sobre la Lingüística Computacional a lo largo de las décadas se han descubierto dos métodos de afrontar los retos y objetivos propuestos, el de la gramática y el de la estadística. Así lo ilustra la siguiente cita:

> La LC se ocupa de la creación de programas informáticos que simulan parcialmente el comportamiento verbal humano. Es un campo multidisciplinar que fusiona conceptos y métodos de la lingüística, la computación, la lógica, la psicología y la estadística. (Russell y Norvig, 2010, p.16)

En cuanto a la pugna sobre si forma parte de la Lingüística Teórica o la Lingüística Aplicada[5], esto depende de la perspectiva desde la que se vaya a abordar: si consideramos

---

1 La RAE define la lingüística computacional como ‘‘aplicación de los métodos de inteligencia artificial al tratamiento de cuestiones lingüísticas’’.

2 “Desde el punto de vista de su vinculación a la lingüística, la lingüística computacional puede ser considerada una subdisciplina de la lingüística teórica, en tanto que uno de sus objetivos es la elaboración de modelos formales (e implementables informáticamente) del lenguaje humano” (Gómez Guinovart, 1998, p. 135).

3 ‘‘La lingüística computacional está considerada como una rama de la inteligencia artificial (IA). Como todos los campos dentro de la IA, se ocupa de la investigación y sistematización de una capacidad cognitiva. En el caso de la lingüística computacional, el objetivo central es la capacidad lingüística’’ (Halvorsen, 1991, p. 252).

4 Algunos de ellos son *Syntactic Structures* (1957) y *Aspects of the Theory of Syntax* (1965) de Chomsky, que han influido en la gramática generativa y en la LC. Kenneth Church y Frederick Jelinek han sido reconocidos por sus contribuciones con investigaciones sobre el modelado de lenguaje estadístico y la traducción automática.

5 ‘‘Computational linguistics can be seen as a branch of applied linguistics, dealing with computer processing of human language. Automatic translation between natural languages, text processing and communication between people and computers are among its central concerns. Speech recognition and understanding and speech synthesis allow people to communicate with computers using spoken language. Computational grammars with top-down and bottom-up processing capabilities have been developed in this connection. Computer-assisted language learning programs are among numerous applications of the

un enfoque científico, el objetivo es comprender el conocimiento lingüístico humano y reproducirlo a través de simulaciones por computadora. Por otro lado, está el enfoque aplicado, donde se busca abordar problemas de comunicación lingüística como la traducción automática o el reconocimiento de voz.

> En LC se entiende que el lenguaje es un proceso comunicativo donde emisor y receptor procesan determinada información en función de un conocimiento lingüístico y del mundo (pragmático) compartidos. La perspectiva individual e introspectiva no tiene cabida en un sistema PLN[6], cuyo objeto es resolver una interacción comunicativa, ya sea proporcionar una traducción, establecer un diálogo con un interlocutor humano o extraer la información solicitada. (Moreno Sandoval, 2015, p.204)

Para ese objeto es necesario disponer de un modelo de lenguaje matemático que incluya factores extralingüísticos. Existen dos modelos[7]: el modelo simbólico y el modelo estadístico; y ambos tienen sus ventajas e inconvenientes[8].

El modelo simbólico está destinado a la elaboración de sistemas de almacenaje de los actos lingüísticos en relación con diferentes módulos lingüísticos: la fonética o fonología, la morfología, la sintaxis, el discurso, etc.

En cambio, según Sandoval (2015):

> Los sistemas que utilizan modelo estadístico, desarrollado a partir de la teoría de la información, tratan las lenguas como un conjunto de sucesos que presentan una determinada frecuencia: cada fonema, cada palabra, cada categoría sintáctica, cada significado o cada traducción posible tiene una cierta probabilidad de aparecer en un contexto determinado. Conociendo esa información, se puede

new technology. Computerized corpora of written and spoken texts facilitate research on usage using concordances.'' (Johnson & Johnson, 1998)

[6] Procesamiento del Lenguaje Natural.

[7] Los modelos de lenguaje y los métodos de procesamiento de lenguaje natural son dos cosas diferentes e independientes entre sí. Un modelo es una representación computacional de cómo funciona el lenguaje humano, mientras que el método es la técnica que se utiliza para resolver una tarea en concreto en el PLN, para ello puede utilizar los modelos de lenguaje además de otras herramientas.

[8] Las ventajas de ambos modelos son las que aparecen mencionadas. Las limitaciones del modelo simbólico incluyen la dificultad para resolver la ambigüedad y la incapacidad para proporcionar una representación adecuada para todas las estructuras gramaticales. Los inconvenientes del modelo estadístico son la dependencia de grandes cantidades de datos de entrenamiento para ''aprender'' patrones lingüísticos y la dificultad para generalizar los mismos, la difícil interpretación de la toma de decisiones del propio modelo afectando la transparencia y la sensibilidad a datos que contienen sesgos sociales, culturales o lingüísticos, de forma que puede que asimile y reproduzca esta orientación en sus predicciones.

> predecir cuál es la siguiente palabra en la oración, sin necesidad de recurrir a reglas gramaticales explícitas (p.20).

Este modelo, a diferencia del modelo simbólico[9], que tiene dificultad para resolver ambigüedades, es esencial para que la Inteligencia Artificial ‘‘aprenda’’ a partir de grandes cantidades de datos, lo que se conoce como método de aprendizaje mediante corpus (véase apartado 1.4).

Para desarrollar un modelo[10] hay que tener claro el tipo de enfoque de modelado dependiendo del uso de etiquetado que se lleve a cabo durante el entrenamiento.

El método que interesa en este estudio es el semi-supervisado[11], donde se parte de un corpus ya etiquetado y se añaden más datos de forma manual para crear un corpus de mayor tamaño. Este corpus va a necesitar ser supervisado para asegurar la transparencia y determinación del etiquetado.

> El objetivo final sería construir un sistema que analice texto no restringido cumpliendo los requisitos de robustez (al menos un análisis por oración), desambiguación (solo un análisis por oración) y precisión (que el análisis sea correcto). La robustez y la desambiguación se consiguen con el método estadístico y la precisión con un buen modelo gramatical. (Moreno Sandoval, 2015, p.209)

3.2 Procesamiento lingüístico de la Inteligencia Artificial. Definición y uso de vectores.

La base de un sistema de PLN (para ver definición, consultar Anexo A), es asignar un significado a cada oración teniendo en cuenta el contexto en el que se está dando.

[9] Además del modelo estadístico y el modelo simbólico existe el modelo híbrido, que combina ambos para superar las limitaciones que tienen cada uno de ellos individualmente.

[10] El modelo de lenguaje más conocido es *BERT* (*Bidirectional Encoder Representations from Transformers* o Representación de Codificador Bidireccional de Transformadores), que fue creado por Google para el PLN mediante la técnica de redes neuronales (definición en el Anexo A), un enfoque de entrenamiento diferente a lo que se conocía hasta la fecha (2018). A partir de este modelo se han creado otros como fue el caso de *RoBERTa* (*Robustly optimized BERT approach*), un modelo de lenguaje desarrollado por *Facebook AI* en 2019. Es una variante de *BERT* basada en su arquitectura e ideas subyacentes, pero introduciendo mejoras y ajustes que lo hacen más efectivo en varias tareas de PLN. *RoBERTa* difiere de su proceso de preentrenamiento con respecto a *BERT*, además de utilizar técnicas como el entrenamiento dinámico de máscaras o el *batch size*. Liu et al. (2019) presentan *RoBERTa* y describen los detalles de su arquitectura, el proceso de preentrenamiento optimizado y su rendimiento en una variedad de tareas de procesamiento del lenguaje natural.

[11] Los modelos que se entrenan con datos etiquetados con una respuesta conocida son los llamados supervisados. Los modelos que se entrenan con datos no etiquetados y buscan descubrir nuevos patrones o estructuras inherentes son los no supervisados. El semi-supervisado combina elementos de ambos enfoques.

En su estudio sobre los sistemas de procesamiento del lenguaje natural, Contreras y Dávila (2001) señalan la existencia de dos módulos distintos: uno gramatical, que se centra en aspectos morfológicos y sintácticos; y otro interpretativo, que se encarga de aspectos semánticos y pragmáticos. En el nivel interpretativo, que contiene la semántica y la pragmática, se busca expresar con técnicas y redes semánticas el conocimiento del mundo del hablante: ‘‘Este conocimiento del mundo interviene en la interpretación oracional y del discurso, pues sobre inicio dicho conocimiento se maneja el conocimiento implícito de los hablantes, que permiten resolver las ambigüedades’’ (Contreras & Dávila, 2001). Para esto se utilizan ontologías (véase Anexo A) que son la fuente principal que refleja el conocimiento del mundo implícito, donde se relacionan las palabras que el hablante conoce con el sentido que les otorgamos, de modo que resultan ser estructuras conceptuales o mentales.

En cuanto al módulo gramatical, está estructurado por niveles: fonológico, morfológico, sintáctico, semántico y pragmático, que facilitan la clasificación del contenido para que la computadora pueda interpretarlo.

> La arquitectura del sistema NLP[12] da a conocer la interpretación de la computadora y examina las oraciones proporcionadas. El manejo de forma sencilla de este sistema es el siguiente: primero el usuario le habla a la computadora expresando lo que desea realizar; segundo la computadora analiza el comando de voz proporcionado, en el sentido morfológico y sintáctico, es decir, analiza si las oraciones contienen palabras compuestas por morfemas y su conformación es correcta y completa; después se analizan las oraciones semánticamente, es decir, conocer cada oración y consignar el significado de estas a expresiones lógicas (verdadero / falso); de seguida se realiza el análisis pragmático de la instrucción, y una vez analizadas las oraciones, estas se analizan juntas, analizando la situación de cada oración, finalizando estos pasos la computadora ya sabe que función realizar; cuando se obtiene la expresión u orden final, ya ejecutado brinda al usuario lo solicitado o pedido. (Vásquez, Quispe, & Huayna, 2009)

[12] El orden de las siglas se debe a que es una cita y se ha utilizado el término en inglés *Natural Language Processing*.

A pesar de los grandes avances que hay actualmente en cuanto a los sistemas informáticos y la Inteligencia Artificial, siguen existiendo algunos límites a la hora de conseguir un Procesamiento del Lenguaje Natural lo más real y afinado posible, que sería muy deseable, puesto que ‘‘computers would become vastly more accessible if people could use really natural language to communicate with them – that is, talk to them as they would talk to another person, and have responses of the kind another person would give’’ (Machine Learning for Multimodal Interaction, 2004, p.308).

Hay dos campos en los que es necesario avanzar si se quiere conseguir lograr un proceso lo más fiel posible al de la mente humana. El primero de ellos es el análisis de las emociones y el segundo un área que se viene desarrollando en los últimos meses, el modelo multimodal (para ver definición, consultar Anexo A). Aunque no lo parezca, ambas están estrechamente relacionadas, pues las emociones son puramente multimodales.

> Emotion is profoundly multi-modal. It is reflected in facial expressions, gestures, body language, and actions; in the propositions expressed, the words and syntax chosen to express them, and the way they are spoken; in involuntary visceral changes, and in blood flow and electrical activity in the brain. (Machine Learning for Multimodal Interaction, 2004, p.311).

Como se explica en la cita, la emoción es intrínsecamente multimodal. Además, no hay ninguna medida que defina con exactitud cuál es el verdadero estado emocional de una persona. Debido a esta complejidad, encontrar signos que identifiquen o sinteticen una emoción es un desafío sustancial.

La forma más común de describir las emociones es de manera categórica; es decir, las emociones se identifican mediante la asignación de etiquetas verbales que se obtienen del lenguaje cotidiano o a partir de él[13]. Es una forma de reflejar mejor la diversidad y complejidad de las experiencias emocionales humanas.

Una vez que se han asimilado ambos módulos, tanto el gramatical como el interpretativo, la computadora[14] comienza un nuevo proceso llamado vectorización, que utiliza un

[13] ‘‘The most familiar form of description is categorical. Emotions are identified by identifying verbal labels, which are either drawn directly from everyday language, or adapted from it’’ (Machine Learning for Multimodal Interaction, 2004, p.312).

[14] El término computadora hace referencia a un servidor especializado, sin pantalla y sin disco duro, que no hace las funciones genéricas de un ordenador.

lenguaje numérico. La vectorización consiste en traducir las palabras a un lenguaje que las computadoras entienden. En esencia, convierte el texto en un conjunto de números (llamados vectores) que les dan sentido computacional a las oraciones según el siguiente proceso:

En primer lugar, se le asigna un número específico a cada palabra única.

En segundo lugar, está la representación como vectores. Cada palabra (ahora representada por un número) se transforma en un vector. Un vector es simplemente una lista de números que no son aleatorios, sino que representan diferentes características de la palabra, como su significado, uso o su relación con otras palabras. Estos vectores colocan, efectivamente, las palabras en un espacio numérico en el que palabras con significados o usos similares estarán más cerca unas de otras. Por ejemplo, "gato" y "perro" podrían estar más cerca entre sí que "gato" y "avión".

La vectorización es fundamental en PLN porque, al convertir las palabras en vectores, las computadoras pueden realizar operaciones matemáticas con ellas, lo que les permite entender la similitud entre palabras, analizar el sentimiento de un texto (por ejemplo, si es positivo o negativo), traducir el texto en diferentes idiomas, responder a cuestiones planteadas por el usuario o crear resúmenes de textos largos.

En resumen, la vectorización es como enseñarle a una computadora a entender y procesar el lenguaje humano mediante un nuevo idioma, el idioma de los números. Es una técnica fundamental que permite a las computadoras realizar una amplia variedad de tareas complejas relacionadas con el lenguaje. [15]

3.3 IA Multimodal y su aplicación al análisis de sentimientos.

Ya se ha hecho especial énfasis en el hecho de que las emociones son multimodales. Como ya se ha mencionado anteriormente, las IA multimodales existen desde hace solo unos meses para el público, por lo que encontrar fuentes de información para la investigación ha sido un trabajo complejo.

Cuando los textos consisten en imagen y escritura surgen formas específicas de cohesión y coherencia textual y se requieren nociones teóricas para darles sentido[16]; por ejemplo,

[15] Para la información de este apartado (vectorización) se ha utilizado a Manuel del Toro Mateos como fuente de comunicación personal el 25 de enero de 2024.
[16] Skim AI. (s.f.). *¿Qué es la AI Multimodal? Casos de uso de la AI Multimodal*. https://skimai.com/es/que-es-la-ai-multimodal-casos-de-uso-de-la-ai-multimodal/

para una IA multimodal se requiere un acercamiento retórico a la creación del texto, además de la cuidadosa consideración del propósito, el público, el contexto y las estrategias persuasivas para elaborar un mensaje efectivo y convincente. La creación de un texto es un acto semiótico en el cual el significado es relevante en todos los aspectos, debido a que es también un acto social con consecuencias sociales.

Se ha debatido mucho sobre el efecto que tiene la multimodalidad en la escritura. Algunos investigadores consideran que la imagen o el texto tienen el papel principal[17]; pero hay otros que consideran que esto va a depender de las funciones que desempeñe cada uno y las formas de escritura del texto.

Cuando se intenta ‘‘traducir’’ a un lenguaje computacional un complejo de imagen, escritura, movimiento y sonido se pueden utilizar diferentes recursos como léxico, representación o sintaxis. La Inteligencia Artificial Multimodal tiene varios niveles o cuadros explicativos del proceso. [18]

En el Box[19] I se encuentra la base de datos que la computadora utiliza para inferir grandes cantidades de información. A esta base de datos se le conoce técnicamente como corpus. Un corpus debe ser lo suficientemente grande y accesible para realizar investigaciones relevantes para el análisis lingüístico y la mejora de los modelos de IA.

[17]Tanto Gunther Kress como David Machin destacan la importancia de considerar la función de cada elemento en la comunicación multimodal sin relegar ninguno de ellos a un segundo plano. Michael Halliday tiene una perspectiva que podría considerarse contraria en el sentido de que da más énfasis a la escritura sobre la imagen en ciertos contextos afirmando que es un modelo de comunicación altamente sofisticado y versátil, y que puede transmitir conceptos de manera precisa y detallada. Algunos teóricos del cine como Sergei Eisenstein y André Bazin defienden que las imágenes tienen el poder de evocar respuestas emocionales profundas y transmitir mensajes complejos de manera más directa que el texto escrito.

[18] En la fuente a la que se ha recurrido (Machine Learning for Multimodal Interaction, 2004) a estos niveles se les llama *Box* por lo que se va a utilizar la misma nomenclatura.

[19] Se define como un nivel o componente específico dentro de un modelo de aprendizaje automático. Son utilizados para organizar y estructurar el análisis multimodal y procesar los datos provenientes de fuentes diversas con una apropiada integración.

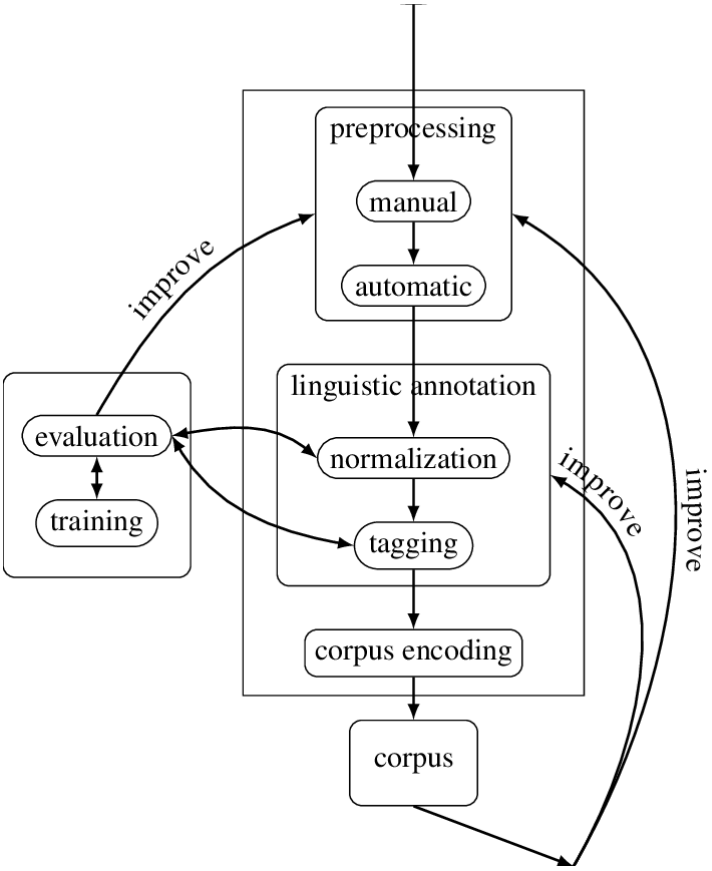

Ilustración 1 (Machine Learning for Multimodal Interaction, 2004)

El Box II consiste en la anotación por capas[20].

La anotación directa de eventos se centra en describir lo que se ve de manera factual y sin interpretación. La capa 1 es la visual, la capa 2 es la temporal y la capa 3 es el audio. La interpretación a nivel semántico, de forma o de función, agregaría un nivel de comprensión más profundo sobre lo que está sucediendo en el evento y por qué. Por tanto, la capa 4 es la capa semántica[21] y la capa 5 es la funcional.

Capa 1	Capa visual	Dos personas en una entrevista.
Capa 2	Capa temporal	La entrevista tiene una duración de diez minutos.
Capa 3	Capa auditiva	Se escuchan voces y sonidos de fondo.

[20] Esta especialización de las capas de diferentes tipos de datos tiene que ver con la arquitectura de red neuronal.
[21] Se enfoca en la semántica léxica y oracional, es decir, la descripción del evento.

Capa 4	Capa semántica	Interpretación de entusiasmo o pasión dado que el entrevistador utiliza un tono de voz más enérgico.
Capa 5	Capa funcional	La entrevista sigue un formato de pregunta-respuesta, lo que indica un intercambio de información entre los participantes.

La división por capas ayuda a diferenciar entre la anotación directa de eventos objetivos y la interpretación de estos eventos a nivel semántico, o forma y función[22].

El Box III contiene los modelos y la semántica[23] desde una perspectiva más compleja e integradora. Para estas anotaciones se necesitan los modelos de interacción humana, ya que, además de esto, el corpus puede proporcionar información sobre la intencionalidad de los hablantes y el análisis de patrones que puedan existir en su comportamiento.

El Box IV consiste en las herramientas de simulación, que se usan para probar hipótesis sobre la interacción humana y validar modelos. Por ejemplo, se simulan patrones de conversación para comparar con la realidad.

El Box V[24] sigue en relación con el Box III, el comportamiento humano. Se resalta la psicología social, que proporciona información sobre los patrones de conducta y, a su vez, sirve como base para el análisis automático de esta interacción.

Para el modelo de entrenamiento que se propone en este trabajo se usarán de forma simultánea ambos modelos citados, tanto el corpus como la anotación por capas. El corpus ayuda a descubrir irregularidades en los patrones y construye modelos hipotéticos y las anotaciones prueban y evalúan estos modelos mediante simulaciones.

[22] ‘‘This division reflects the difference between direct annotation of objective events and interpretation of these events on a semantic level, or form and function’’ (Machine Learning for Multimodal Interaction, 2004, p.24).
[23] Hace referencia a la semántico-pragmática y busca aspectos verbales y no verbales.
[24] En cuanto al Box IV no se ha extraído información que se considere relevante para la investigación.

Los modelos multimodales más comunes que existen actualmente son *GPT 4 – Turbo*[25], *Gemini 1.5*[26], *Llama 2*[27]; este último es de código abierto para el público y se utilizará en la metodología de este trabajo.

3.4 Corpus etiquetado.

El corpus y el anotado o etiquetado son fundamentales, puesto que van a proporcionar información clave para la elaboración de una propuesta metodológica.

> En particular, en el ámbito de la Lingüística Computacional, se han constituido en punto de partida imprescindible para la elaboración de léxicos y gramáticas, y representan una línea de investigación transversal en lo que al tratamiento del lenguaje con medios informáticos se refiere, al ser indispensables para el desarrollo de aplicaciones basadas tanto en el texto como en el habla. (Moure y Llisterri, 1996)

Efectivamente, los corpus ayudan a tomar muestras de datos reales y usos lingüísticos para constituir un material esencial antes de comenzar a desarrollar un modelo de entrenamiento.

La ‘‘lingüística de corpus’’ tal y como la conocemos en la actualidad no va a surgir hasta principios de los años ochenta[28], aunque fueron

> los trabajos de antropólogos, etnógrafos y, sobre todo, de los lingüistas estructurales norteamericanos –F. Boas, E. Sapir, L. Bloomfield, Ch. Fries…– los que, durante la primera mitad del siglo XX, contribuyeron a sentar las bases de la lingüística de corpus como metodología empírica basada en la observación de datos. (Llamazares, 2019, p.296)

Algunos de estos estudiosos llegaron a considerar que el corpus era el método más efectivo para el estudio de las lenguas antes de la llegada de la propuesta de Chomsky[29].

---

[25] Para obtener más información sobre GPT 4 – Turbo, consultar: https://help.openai.com/en/articles/8555510-gpt-4-turbo-in-the-openai-api

[26] Para obtener más información sobre Gemini 1.5, consultar: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note

[27] Para obtener más información sobre Llama 2, consultar: https://llama.meta.com/llama2/

[28] Las primeras pinceladas dentro de este ámbito se dan en las primeras décadas del siglo XX.

[29] Se caracterizaba por ser un conjunto de grabaciones y transcripciones en papel, especialmente enfocado hacia lenguas vivas no previamente documentadas por escrito, como las lenguas amerindias. Este método era esencial debido a la falta de registros escritos previos. Se centraba en aspectos fonéticos y (morfo)fonológicos y carecía de representatividad debido a la limitación en el manejo de datos, lo que generó críticas sobre su parcialidad en la descripción de la realidad.

En las décadas de los sesenta y setenta, el corpus estructuralista se enfrentaba a un entorno poco favorable. Aunque los primeros trabajos en lingüística de corpus involucraban computadoras, estaban al margen de la corriente lingüística predominante y algunos corpus aún no estaban listos para ser digitalizados. En los ochenta, el uso de computadoras para desarrollar corpus se generalizó debido a sus ventajas y flexibilidad, especialmente para construir modelos de procesamiento del lenguaje natural y formular hipótesis lingüísticas. La lingüística de corpus moderna comenzó a tomar forma en 1984 (con la publicación del libro *Corpus Linguistics I: Recent Developments in the Use of Computer Corpora*, de J. Aarts y W. Meijs) y se define como el estudio del lenguaje basado en colecciones de textos disponibles en formato legible por máquina.

En el ámbito español cabe hacer mención del Corpus de Referencia del Español Actual (CREA) y el Corpus Diacrónico del Español (CORDE), creados en los años noventa por la Real Academia Española.

En la creación de un corpus, se consideran criterios internos (lingüísticos) y externos (situacionales). Puede ayudar a esclarecer a cuál es la temática del texto partiendo de su contenido, y suele ser una mecánica muy utilizada en modelos de lenguaje largos (*LLM*) (para ver definición, consultar Anexo A).

La finalidad del corpus guía las decisiones. Luego, se lleva a cabo la anotación, que implica la introducción de etiquetas para aspectos lingüísticos y extralingüísticos. Las anotaciones pueden ser automáticas, semiautomáticas o manuales, y abarcan diversos niveles del lenguaje.

En la actualidad es innegable el impacto que han tenido los corpus en diferentes áreas como la literatura, la lexicografía, la enseñanza de lenguas, el análisis del discurso o la lingüística forense; proporcionando datos a los diccionarios, extrayendo patrones o tendencias o ayudando a diagnosticar casos de plagio.

3.5 Manifestación lingüística de las emociones.

Ya se ha explicado anteriormente que las emociones son intrínsecamente multimodales.

> Klein (1986) distingue dos tipos de información, la puramente lingüística y la información paralela (signos visuales como gestos, mímica, etc.). En el dominio de la información paralela, se ha prestado mucha atención a la descripción de los gestos faciales y los cambios fisiológicos suscitados por los estados emocionales.

> […] Pese a su carácter no lingüístico, el conocimiento de estos aspectos fisiológicos y gestuales resulta sumamente importante para el estudio lingüístico emocional. (Gómez, 2012, p.59)

Como ya se sabe, hay diferentes formas de expresar una misma emoción como la alegría, el asco o el enfado, que varían según la lengua. Esto va a depender de la situación lingüística, del artificio del lenguaje o del contexto. Según Mitchell (2005), las imágenes tienen una vida y deseos propios, influyendo en nuestra percepción del mundo de maneras complejas.

Hay dos formas de representar lingüísticamente las emociones: el lenguaje literal y el figurado. El lenguaje literal se basa en la semántica o la sintaxis para expresar un evento emocional; por ejemplo, ‘‘Estoy contenta porque es mi cumple’’ se basa en la semántica porque la palabra ‘‘contenta’’ está cargada de significado y se vincula a una emoción específica. En la oración ‘‘Sonreía cuando le daban el regalo de su cumpleaños’’ la sintaxis de la oración transmite la emoción de alegría, aunque la palabra ‘‘felicidad’’ no aparezca explícita. Esto se consigue a través de la descripción de las acciones y los detalles. Se puede hablar de una ‘‘gramática de las emociones’’ en el sentido de que se combinan entre ellas para crear eventos complejos a partir de un ‘‘lexicón’’ específico[30]. El lenguaje figurado codifica la semántica del evento emocional mediante metáforas[31] y metonimias[32].

La finalidad de la creación de un modelo multimodal para el análisis de sentimientos[33] en redes sociales es saber interpretar una emoción estando o no codificada con lenguaje figurado aportando contexto, ya sea mediante imagen, tipografía, audio, vídeo, etc. Es decir, que la computadora no solo tenga en cuenta los parámetros pertenecientes al lenguaje literal con referencias léxicas marcadas, sino también aquellas emociones que

[30] Ejemplo: siento preocupación al tener un examen para el que no he estudiado, me pongo a estudiar por la noche, aprendo lo importante e improviso el resto y siento inseguridad. Las chicas de mi clase antes del examen me ponen nerviosa haciéndome preguntas. Cuando hago el examen me doy cuenta de que era más fácil de lo que yo pensaba y estoy tranquila. Cuando me dan la nota y apruebo estoy contenta.

[31] Ejemplo de metáfora: ‘‘Le sale humo por las orejas’’. Se utiliza para expresar enfado o frustración. La metáfora evoca la imagen de alguien que está tan furioso que emite humo por las orejas, aunque en realidad es una emoción figurada para enfatizar la intensidad de la emoción.

[32] Ejemplo de metonimia: ‘‘Aquí tienes mi hombro si necesitas desahogarte’’. El término ‘‘hombro’’ se utiliza para representar la idea más amplia de consuelo o apoyo emocional. Se asocia el objeto físico al concepto abstracto.

[33] Hay mucha información y estudios previos referidos al análisis de las emociones, sobre todo, en relación con la pertenencia a grupos religiosos (Pew Research Center, 2016), publicidad comercial (Bagozzi, Gopinath, & Nyer, 1999) o intencionalidad política (López-López, Cuadrado, & Navas, 2009).

están encubiertas por el lenguaje y que surgen de su relación con la imagen. Es una mecánica que nuestro cerebro está acostumbrado a procesar de forma automática, pero para el que es preciso entrenar a la computadora.

La creación de un corpus y el entrenamiento posterior en un modelo multimodal puede ayudar a solventar las limitaciones existentes actualmente para la descodificación de elementos emocionales metonímicos o metafóricos.

4. METODOLOGÍA

Es fundamental para el desarrollo de un modelo saber la definición de *machine learning* o aprendizaje automático (véase Anexo A). Es importante aclarar también que el desarrollo del modelo se ha llevado a cabo en inglés, aunque también podría haberse realizado en español. A través de un modelo estadístico se intenta predecir, mediante el lenguaje Python, si los mensajes son positivos, negativos o neutros; lo que se conoce como *sentiment analysis*.

Hay que distinguir en esta parte del trabajo tres objetivos diferentes: la creación de un corpus a través del etiquetado, el desarrollo de un modelo estadístico de análisis de sentimientos y el análisis de los resultados del modelo.

4.1 Etiquetado de corpus.

No es común crear un corpus desde el inicio, ya que el etiquetado es un trabajo costoso tanto económicamente como en lo que a esfuerzo se refiere y solo grandes empresas con una fuerte solvencia económica pueden permitírselo. Por esto, por lo general, los desarrolladores de inteligencia artificial que trabajan en empresas pequeñas o de forma autónoma parten de un corpus ya etiquetado que extraen de repositorios en línea (como *Google Dataset Search*), proyectos de código abierto (como *GitHub*, el utilizado por *Hugging Face*[34], que también es un repositorio en línea), instituciones académicas como universidades o grandes empresas. Aun así, es importante e interesante conocer cómo funciona un proceso de etiquetado. Además, para la creación de un corpus hace falta que haya un gran número datos anotados y que pueda servir como recurso para otros modelos. Se pueden ampliar con participación de la comunidad de desarrolladores y los hay de temáticas muy diversas.

[34] Es una comunidad para el desarrollo y el intercambio de modelos de aprendizaje automático y recursos PLN. Es conocido por su amplia variedad de modelos preentrenados, aunque también se encuentran datos etiquetados o pipelines (para la definición de 'pipelines', véase Anexo A).

La imagen que se muestra es una captura de un etiquetado de *tweets* para un modelo de análisis de sentimientos. No se ha utilizado en el modelo de aprendizaje creado, sino en una de las *demos* utilizadas en el apartado 2.3. Ha servido como ejemplo para explicar detalladamente cómo funciona la creación de un corpus.


```
{"tweetid":1956818303164211 20,"ds_language":"en","language":"lt","label":"neutral"}
{"tweetid":1988864201259 0912,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2181128274859868 16,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2333108799895142 40,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2350143107478364 16,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2398115659378974 72,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2439045493692907 52,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2465673723858247 68,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2540478897168138 24,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2554035120815513 60,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2554498988104949 76,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2556643608385372 16,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2609165019163074 56,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2609171171535749 12,"ds_language":"en","language":"tl","label":"positive"}
{"tweetid":2612837135754403 85,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2616510204117278 72,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2621835608138915 84,"ds_language":"en","language":"en","label":"neutral"}
{"tweetid":2623757824023715 84,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2624427060878376 97,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2627592561498726 42,"ds_language":"en","language":"et","label":"neutral"}
{"tweetid":2629861916814991 39,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2632979893705891 85,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2633544993709793 29,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2634137514269818 88,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2637866615178608 65,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2638327625643663 37,"ds_language":"en","language":"en","label":"neutral"}
{"tweetid":2638513941408768 01,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2639650097412259 84,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2640383689528647 68,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2640736758028083 21,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2641918385655767 04,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2641980179497779 20,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2641224352704880 64,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2642417329489920 00,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2500200937555353 60,"ds_language":"en","language":"en","label":"negative"}
{"tweetid":2583472338068029 44,"ds_language":"en","language":"en","label":"neutral"}
{"tweetid":2636103954267054 08,"ds_language":"en","language":"en","label":"neutral"}
{"tweetid":2636727061033984 00,"ds_language":"en","language":"fi","label":"neutral"}
{"tweetid":2637687026448015 36,"ds_language":"en","language":"en","label":"positive"}
{"tweetid":2637735094933422 08,"ds_language":"en","language":"en","label":"negative"}
{"tweetid":2639479804978913 28,"ds_language":"en","language":"de","label":"positive"}
{"tweetid":1113445996996935 68,"ds_language":"en","language":"en","label":"neutral"}
{"tweetid":1113449772122112 01,"ds_language":"en","language":"en","label":"neutral"}
{"tweetid":2090471588533411 84,"ds_language":"en","language":"en","label":"neutral"}
```


[35]

Ilustración 2

Las etiquetas están en formato *json*[36], de forma que si hay algún error se encuentre de forma automática y se pueda corregir. Los **objetos** aparecen entre llaves ({}) y son pares de clave-valor (identificador único para un valor asociado), los ***arrays*** (para su definición véase Anexo A) aparecen entre corchetes ([]) normalmente (no en este caso, son las numeraciones del margen izquierdo) y contienen una secuencia de valores que sirven para organizar y manipular datos de manera eficiente[37], y los **valores** pueden ser cadenas de texto, números, objetos, *arrays*, etc.

[35] Este etiquetado pertenece al usuario de *Hugging Face* [*@cardiffnlp*] para la creación de un modelo basado en la extracción de datos de Twitter que utiliza como referencia a Camacho-Collados et al. (2022).
[36] (*JavaScript Object Notation*) es un formato de intercambio de datos ligero y de fácil lectura que sirve para almacenar datos estructurados.
[37] Importante remarcar que comienza desde la posición 0.

Todos los códigos que aparecen en la imagen son *tweets* siendo etiquetados. Cada *tweet* tiene un ID (identificador) único. Para encontrar la ID solo hay que fijarse en la serie numérica que aparece al final del enlace del *tweet*. Por ejemplo, en la Ilustración 3 el ID sería ‘‘1793719295630856628’’.

Ilustración 3

Todos los valores del objeto van separados entre comas; el siguiente valor, en orden, es el lenguaje, introducido por el código `‘‘ds_language’’`. En este caso, se puede observar que todos los *tweets* están en inglés (`‘‘en’’`), aunque su idioma original se muestra en el siguiente valor (`‘’language’’`) y se pueden encontrar en italiano (`‘’it’’`), inglés (`‘’en’’`), etc. Esto se explica porque este modelo de aprendizaje no es multilingüe, sino que detecta el idioma del texto y lo traduce utilizando el modelo de traducción desarrollado por *Facebook AI Research*. Por último, encontramos la clave de ‘label’ o etiqueta con el valor de ‘positivo’, ‘negativo’ o ‘neutro’. Esta anotación de una gran cantidad de datos permite que el modelo mediante el aprendizaje de inferencia generalice patrones para hacer predicciones sobre datos.[38]

Una vez que ya se han etiquetado todos los *tweets* que se quieren utilizar para el corpus comienza el siguiente proceso: el entrenamiento (para la definición, véase Anexo A); pero, antes, hace falta seguir una serie de pasos[39] para prepararlo. Hay que preparar el entorno con las librerías (para su definición, véase Anexo A) que se vayan a necesitar, cargar los datos en el formato adecuado para que puedan ser utilizados, definir la arquitectura de la red neuronal[40] del modelo que se va a entrenar y ajustar los parámetros

[38] Las etiquetas de ‘positivo’, ‘negativo’ y ‘neutro’ se han utilizado porque son las más comunes de encontrar en conjuntos de datos ya etiquetados. Estas etiquetas podrían editarse añadiendo más variables (es decir, más emociones en este caso), aunque añadirían dificultad y tiempo a este proceso.
[39] Cabe mencionar que algunos programas permiten un ajuste automático de estos valores, como *PyTorch*.
[40] Las neuronas son equivalentes a las del cerebro humano. Reciben una entrada, realizan una función matemática que ayuda a procesarla y pasa a la siguiente capa. Cada capa tiene un número determinado de neuronas y está especializada en procesar diferentes tipos de datos (visión, auditiva, etc.). La cantidad de neuronas y la disposición de las capas definen la capacidad del modelo para aprender de los datos.

para minimizar la pérdida (*loss*)[41], el optimizador[42], las métricas[43] para evaluar el modelo y los datos de validación[44].

El entrenamiento es una fase muy importante; pero, aparte de observar que no haya ningún error y, en caso de que lo haya, resolverlo, el trabajo en su mayoría lo realiza la computadora. Puede durar minutos, horas o incluso días dependiendo de varios factores, por lo que es importante realizar pruebas piloto y ajustar bien los parámetros para encontrar equilibrio entre la precisión del modelo y el tiempo disponible. Según el entorno de desarrollo que se utilice puede haber herramientas de aprendizaje automático y un botón de *play* (véase Anexo B, Ilustración 17) que inicia el proceso. Este botón está asociado a una función específica definida por los datos y la configuración establecida previamente en el código. Sin embargo, añadir el parámetro '`fit ()`' al código es lo que indica al modelo que inicie el entrenamiento si no hay ningún tipo de herramienta.

Puede ocurrir que el entrenamiento falle o dé error[45] en algún punto de la ejecución. Si esto ocurre se empieza desde uno de los puntos de *checkpoint* o restauración que se guardan automáticamente.

4.2 Desarrollo del modelo estadístico (*Sentiment_analyzer*)

Existe un procedimiento oculto que realiza una serie de predicciones. El primer paso dentro de este procedimiento está en el texto base, que tiene que pasar por un filtro de limpieza donde se normaliza: se eliminan las mayúsculas, las vocales acentuadas, y los caracteres complicados, se normaliza el uso en los espacios, etc. El siguiente objetivo es poder llevar este texto a una representación numérica donde cada mensaje y cada columna representa una palabra. Para continuar el proceso se tiene que seguir un criterio de transformación en el que se hayan dos matrices iniciales: entrada y salida. La matriz de entrada es aquella que numera las veces que aparece una palabra en cada uno de los mensajes; la matriz de salida tiene una sola columna, en la que, por cada uno de los mensajes, se va a decir si es positivo, negativo o neutro. La predicción de la matriz de

[41] La pérdida o *loss* representa si el modelo está funcionando correctamente a la hora de hacer predicciones. El objetivo es mejorar la precisión del modelo y para ello la pérdida debe reducirse.
[42] Ayuda a reducir la pérdida ajustando parámetros de la red neuronal.
[43] En términos de precisión, aptitud, capacidad de generalizar, adaptación y desempeño en datos de prueba.
[44] Es un proceso donde se detecta si el archivo está bien escrito acorde a la función que va a realizar.
[45] Puede ocurrir por caída de servidor, porque haya algún parámetro mal escrito (no debería pasar si se utiliza el formato *json*), que se apague la computadora en la que se está realizando el entrenamiento o que una actividad en segundo plano lo interrumpa. Que el error aparezca reflejado es algo positivo, ya que suele ir acompañado de una explicación de por qué está ocurriendo y cómo se puede solucionar (consultar Anexo B: Ilustración 18 para ver ejemplo de error).

salida a partir de la entrada se hace con alguno de los algoritmos (véase Anexo A) de *machine learning*. Existe una amplia variedad de modelos estadísticos y matemáticos, cada uno con ventajas y desventajas, que son los que realizan los cálculos y tareas especializadas para poder llevar a cabo lo que se conoce como ‘‘fase de predicción’’[46]. Este aprendizaje del modelo (los cálculos que se llevan a cabo para convertir un texto en numeración legible para la computadora) se utiliza para predecir nuevas entradas.

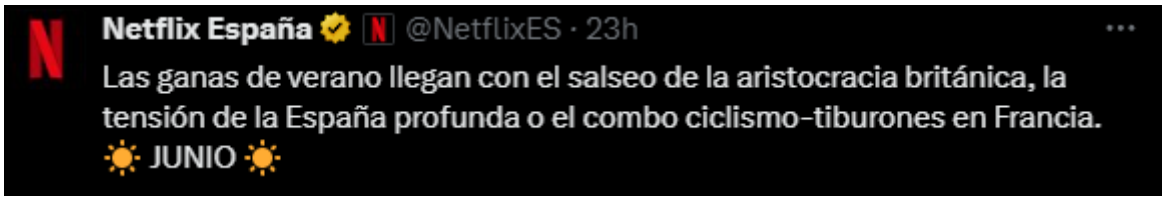

Ilustración 4

```
las ganas de verano llegan con el salseo de la aristocracia britanica
la tensión de la españa profunda o el combo ciclismo tiburones en francia
junio

ENTRADA la el las verano salseo con tiburones tensión junio    SALIDA
msje1   1   1  1   1       1     1    0          0       0      pos (0,6)
msje2   2   1  0   0       0     0    1          1       0      pos (0,5)
msje3   0   0  0   0       0     0    0          0       1      neu (0,1)
```

Ilustración 5

Se ha llevado a cabo un prototipo de modelo basado en reglas semánticas. Se ha utilizado el entorno de desarrollo *DataSpell 2024.1.1*. En un archivo nuevo, lo primero que hay que hacer es importar una librería ya preparada con un procedimiento para hacer un análisis de sentimientos y lista para usar; en este caso es la librería *nltk_vader*[47]. Lo siguiente es escribir tres variables de referencia (*y*, *z* y *x*) que van a representar los valores de positivo, negativo y neutro. Al ejecutar los resultados se encuentran unos valores cuyo cómputo total se acerca a 1, 0 o -1. Si el resultado se acerca a 1 es una variable positiva, si se acerca a 0 es una variable neutra y si se acerca a -1 es una variable negativa.

[46] Anteriormente, se han mencionado otras fases como la fase de inferencia o la de entrenamiento.
[47] Esta librería es muy utilizada para el análisis de mensajes cortos como *tweets*.

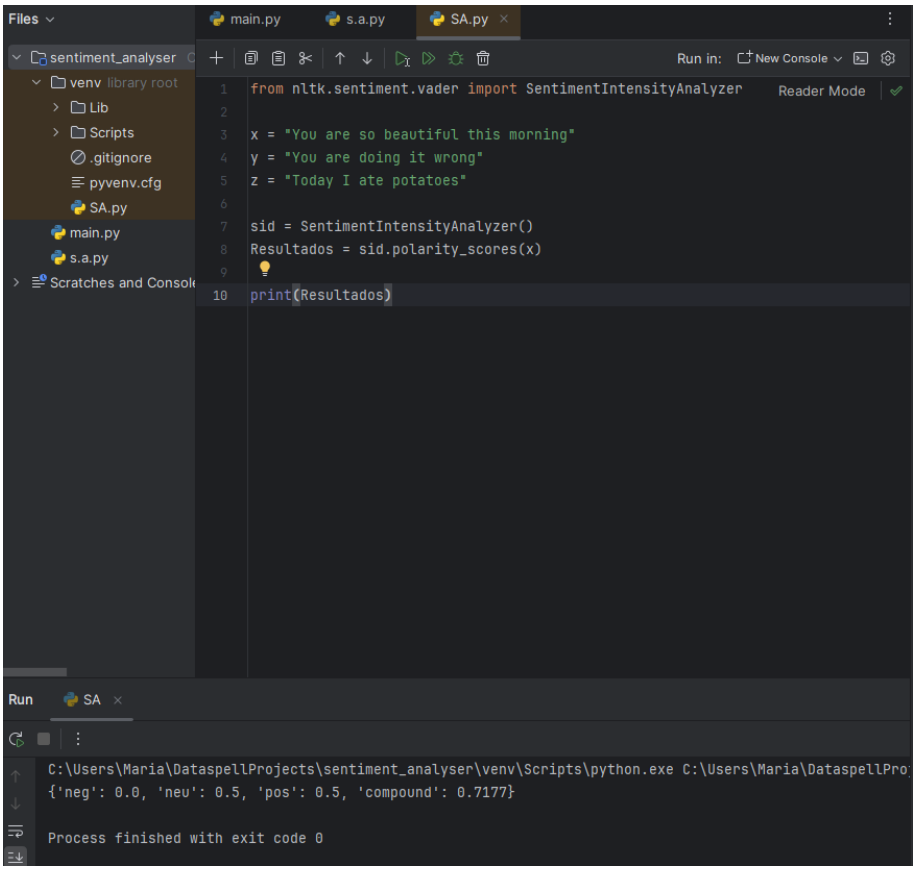

Ilustración 6

Como se puede observar en la Ilustración 6 el resultado de la variable *x* es positivo, ya que tiene un valor de 0.7177; es decir, más cerca de 1 que de 0[48].

Para comprobar aún mejor el funcionamiento del modelo de análisis de sentimientos se inserta un archivo que contiene más de tres variables, cada una en un *array* diferente y en columnas. Se utiliza la librería *pandas*, que es muy buena para el procesamiento de información de manera ordenada.

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import pandas as pd

sid = SentimentIntensityAnalyzer()

df = pd.read_csv("a.csv")
df["sentimiento"] = df["mensaje"].apply(lambda i: sid.polarity_scores(i)['compound'])
df.to_csv("mensajes_con_sentimientos.csv")
```

Ilustración 7

[48] Consultar Anexo B: Ilustraciones 19 y 20 para comprobar los resultados de los valores *y* y *z*.

Al insertar el código que aparece en la Ilustración 7 se va a crear un nuevo documento en formato CSV[49] en el que van a aparecer cada una de las oraciones ponderadas del -1 al 1 según la polaridad que la computadora prediga que tiene el mensaje.

	<anonymous>	mensaje	sentimiento
1	0	He's very annoying.	-0.4576
2	1	Chicaco is very different from Boston.	0.0
3	2	I don't want to bother you.	0.2057
4	3	I feel good.	0.4404
5	4	Please, take me to this address.	0.3182
6	5	That's not right.	0.0
7	6	That smells bad.	-0.5423
8	7	Thank you very much.	0.3612
9	8	You're beautiful	0.5994
10	9	Your things are all here	0.0
11	10	Try it.	0.0
12	11	This doesn't work.	0.0

Ilustración 8

Se puede observar que las oraciones 1,5,9,10 y 11 han sido categorizadas como neutras con valor 0; mientras que, por ejemplo, las variables 0 y 6 son consideradas negativas. El grupo del compendio que ha obtenido una ponderación entre 0 y 1 no pueden considerarse en su totalidad como positivas, de hecho, algunas de ellas están más próximas a ser neutras que positivas.

Se puede llegar a la conclusión de que hay un método con un alto porcentaje de éxito con el entrenamiento adecuado, es decir, se pueden crear modelos propios para hacer predicciones de sentimientos. Si no queremos hacer nuestro propio modelo, no somos expertos en *machine learning* y ya existen modelos hechos y probados, la mejor solución es utilizarlos siempre recordando que todos los modelos tienen limitaciones a la hora de ser utilizados y no siempre los resultados son siempre 100% correctos.

4.2.1 ¿Cómo se puede implementar la multimodalidad?

Se puede entrenar a la IA para convertirla en multimodal. Se sigue el procedimiento básico visto en este proyecto que consiste en el etiquetado. Al no estar manejando una cadena de texto se le tiene que asignar a cada elemento un valor equitativo que nosotros podamos manejar con facilidad. Para esto existe un sistema de codificación llamado Unicode, que permite que, cuando se combinan números, letras y caracteres especiales,

[49] (*Comma-Separated Values*) es un tipo de archivo que se utiliza para almacenar datos donde los valores están separados por comas.

al ser leído por la máquina se nos traduzca en formato de icono, lo que se conoce actualmente como emoji.

Para entrenar a una máquina con estos caracteres especiales se debe crear un archivo CSV, donde en cada *array* hay que asignar un número a un emoticono (véase Anexo B: Ilustración 21). Hay que asociar a estos emoticonos uno de los valores entre positivo, negativo o neutro (o bien, como ya se ha explicado, recurrir a un etiquetado ya realizado y compartido por otros desarrolladores). Existen modelos que tienen variables mucho más profundas que estos tres valores, con números reales, pero significaría una labor de etiquetado inmensa que ya ha sido previamente hecha y es usada para entrenar otros modelos.

Para la incorporación del análisis de sentimientos en imágenes se usa una IA entrenada con patrones, que divide la imagen en casillas (también *arrays*) y evalúan entre 1, 0 o -1 (igual que se ha visto anteriormente) cada casilla. Una vez que todas han sido puntuadas se hace una media y se delibera cuál es el sentimiento que predomina en esa imagen. Esto se puede observar mejor en el apartado 2.3, en el que aparece reflejado mediante un gráfico de porciones cuál es la emoción predominante.

En este caso se ha realizado sobre iconografía. Cualquier modelo podría ampliarse siendo entrenado con imágenes, figuras, vídeos o audios; pero es un proceso más complejo y requiere más tiempo de procesado y preparación. Además, existen actualmente modelos más desarrollados que pueden llevar a cabo esa tarea; algunos son públicos, pero, los más potentes suelen ser privados para compañías como *Google*, *Facebook* o *X*.

4.2.2 Análisis de resultados de modelos multimodales.

Para el desarrollo de este apartado no se ha utilizado el modelo creado[50], sino que se ha recurrido a dos modelos multimodales preentrenados e implementados en una interfaz gráfica. Los modelos son de *Hugging Face* de los desarrolladores *[@FFZG-cleopatra*][51] y [*@Cardiffnlp*][52].

[50] El modelo creado *Sentiment_analyzer* se encuentra en una fase muy temprana de desarrollo, por lo que no es posible realizar un análisis riguroso de sus resultados. La creación de un modelo robusto requiere de un proceso repetitivo y un desarrollo a largo plazo para que sea lo más preciso y afinado posible.

[51] Hugging Face. (s/f). M2SA-demo-multimodal. URL: https://huggingface.co/spaces/FFZG-cleopatra/M2SA-demo-multimodal/tree/main

[52] Cardiff NLP. (s/f). Twitter-roberta-base-sentiment. URL:https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Comenzando por el modelo *M2SA*, se observa que utiliza los tres valores ya mencionados (positivo, negativo y neutro) para clasificar.

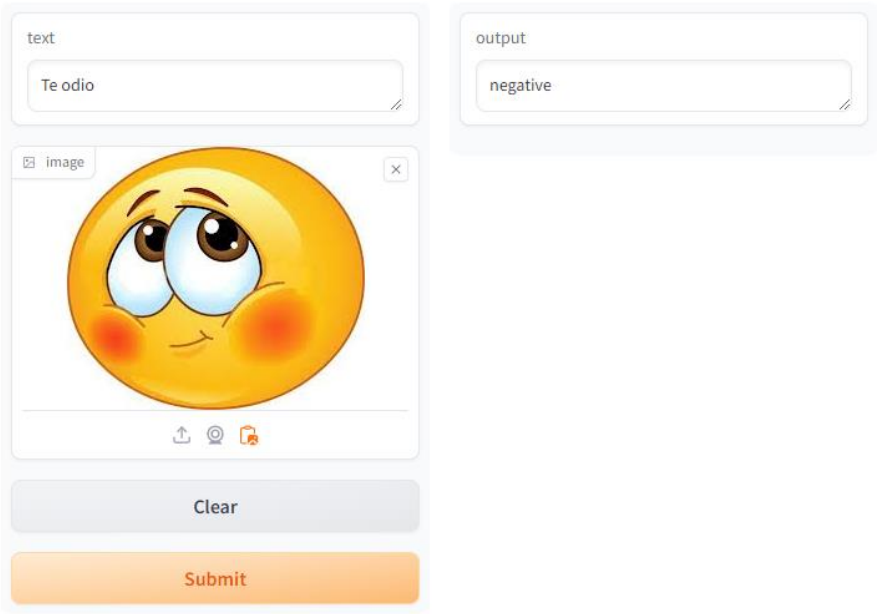

Ilustración 9

En la Ilustración 9 el texto expresa un mensaje negativo, mientras que la imagen debería matizar el texto ya que expresa una expresión positiva. Esta discrepancia entre el texto y la imagen genera confusión sobre la intención real del mensaje y el modelo no es capaz de interpretarlo correctamente.

Además de no interpretar correctamente la imagen y los textos también se ha encontrado dificultad para que funcione correctamente, ya que en muchos casos daba error como es el caso de las Ilustraciones 10 y 11.

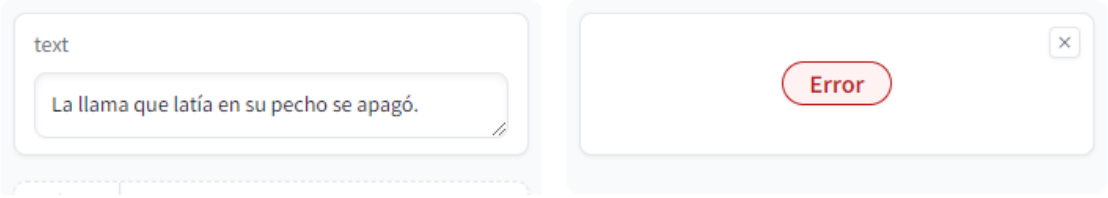

Ilustración 10

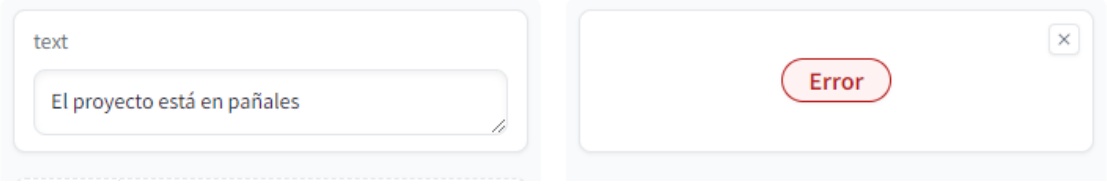

Ilustración 11

Ambas imágenes muestran haberse utilizado metáforas para ver si el modelo es capaz de interpretar ambigüedades, pero no ha dado resultado.

De este resultado con este modelo se puede deducir que no es fácil implementar un modelo tan específico al uso cotidiano, ya que da error con muchas oraciones o imágenes. Hay diferentes motivos por los que puede ocurrir esto, como que perciba el mensaje con ambigüedad en el lenguaje o en el contexto, por la propia multimodalidad (puede que sus características no sean suficientes para determinar con precisión el sentimiento), por carencias en la fase de entrenamiento (puede ser por la falta de diversidad en la selección de datos o por errores en el etiquetado) o porque, al estar utilizando lenguaje coloquial y metáforas, pueda tener dificultades en textos más complejos o contradictorios. Se puede concluir con que este modelo no sirve para analizar la eficacia de la IA en textos que usan figuras retóricas y multimodalidad en conjunto.

En cuanto al modelo *twitter-roberta-base-sentiment* se encuentra en la interfaz[53] una demo más compleja, ya que no solo posee análisis de sentimientos, sino que también hay predicción de palabras, generador de pregunta/respuesta o análisis de *hashtag* (véase Anexo A). Se puede inferir que hay, detrás del desarrollo de este modelo, mucho tiempo de desarrollo y evaluación de las predicciones. Se ha utilizado un *tweet* de *Netflix España* para corroborar y analizar el funcionamiento del modelo.

Ilustración 12

[53] TweetNLP. (s/f). *Demo de TweetNLP*. URL: https://tweetnlp.org/demo/?q

En este *tweet* de la Ilustración 12 aparece una combinación de imagen y texto y ambas serían negativas, puesto que la imagen ayudaría a matizar el texto, pero no cambiaría la polaridad negativa.

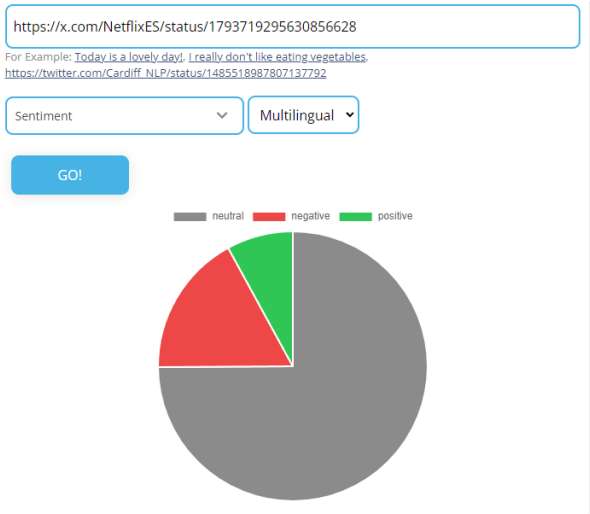

Ilustración 13

Como se puede observar, hay un gráfico de porciones que muestra esos resultados que, usualmente, aparecen ocultos de los cálculos que se realizan sobre la matriz de entrada y la de salida. En la Ilustración 13 se puede ver que hay un porcentaje muy alto de neutralidad al que no hay que prestar atención a no ser que se trate de una porción completa o casi completa. El *tweet* es considerablemente más negativo que positivo, lo que concuerda con el análisis que puede realizar un humano.

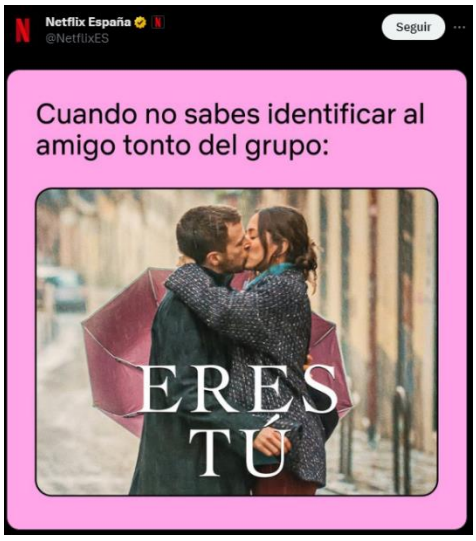

Ilustración 14

En un siguiente ejemplo en la Ilustración 14 encontramos un *tweet* que carece de texto escrito en la caja de texto: el único texto que se puede encontrar es el que aparece en la imagen; por tanto, previsiblemente, si el modelo funcionara correctamente, solo por la imagen extraería que se trata de un *tweet* negativo ya que está llamando ‘‘tonto’’ al propio lector.

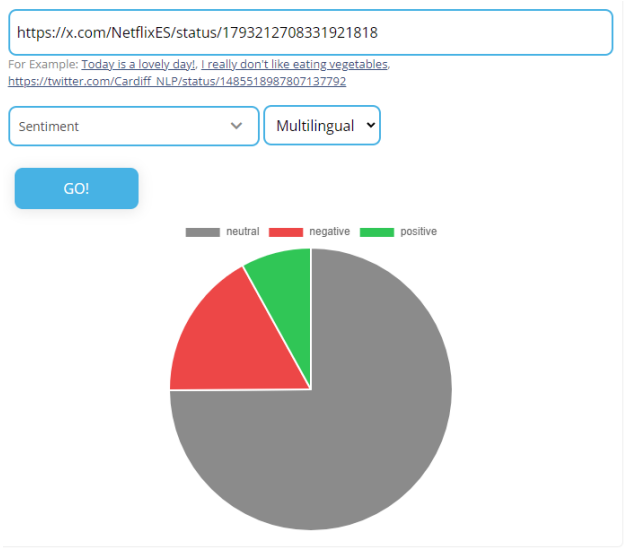

Ilustración 15

Efectivamente, en la Ilustración 15 se puede observar que el resultado es mayoritariamente negativo, por lo que se trata de un modelo que trata la multimodalidad y la multilingualidad de manera efectiva.

Ilustración 16

En la Ilustración 16 se utiliza uno de los ejemplos (‘‘La llama que latía en su pecho se apagó.’’) que se ha utilizado también en el modelo *M2SA* para comprobar la diferencia de funcionamiento entre modelos, la capacidad de interpretar metáforas y la diferencia entre un *tweet* y un texto solo sin imagen. Las diferencias son considerables: a diferencia del modelo *M2SA*, el modelo *twitter-roberta-base-sentiment* obtiene un buen funcionamiento, una buena interpretación de la metáfora y una diferencia considerable en el gráfico de porciones. Esta diferencia en el gráfico ocurre porque, para la IA, es más fácil interpretar un texto único como los utilizados para su entrenamiento, mientras que un enlace a una web externa a la que tiene, primero, que acceder y después, en la que debe analizar tanto texto como contenido, puede que no consiga un resultado preciso. Por esto, como ya se ha explicado, se va a encontrar que en la mayoría de *tweets* predomina la neutralidad, no porque sean neutrales, sino por la falta de precisión del modelo entrenado.

En conclusión, el primer modelo (*M2SA*) tiene muchas carencias y falta pulir multimodalidad, aunque tampoco entiende el uso de ironías, metáforas y otras figuras retóricas. El segundo modelo (*twitter-roberta-base-sentiment*) está muy bien implementado y tiene un buen reconocimiento y análisis de las emociones, pero ha requerido un mayor tiempo de dedicación en cada una de las fases que se han explicado en la metodología y hacer un modelo comparable para un trabajo de investigación sería inviable. Esto resalta la dificultad para comprender completamente el significado de un mensaje y la importancia de considerar múltiples perspectivas a la hora de realizar un análisis de texto e imagen. Incluso los modelos de IA multimodal, al estar aun en desarrollo, están limitados cuando se trata de comprender las complejidades del lenguaje humano y el contexto.

5. CONCLUSIÓN

He buscado resolver los objetivos propuestos para la realización de este proyecto, como investigar cómo los enfoques tradicionales de la lingüística computacional pueden combinarse con técnicas de procesamiento de datos para mejorar el análisis de sentimientos. Estos objetivos han sido alcanzados, si bien sería posible profundizar aun más en este tema, explorando áreas concretas de interés o aplicando otras perspectivas de trabajo para un análisis más completo. Además, he realizado el esbozo de un modelo para el análisis de sentimientos desde el inicio que puede predecir, aunque en un reducido

abanico, las emociones. Al no haber podido terminar de modelarlo no se ha podido evaluar su efectividad, pero se ha podido valorar el rendimiento de algunos modelos escogidos. También se ha podido comprobar la diferencia de funcionamiento entre la multimodalidad y la unimodalidad (solo texto).

A pesar de los resultados obtenidos, siguen quedando áreas que también requieren atención. Futuras investigaciones podrían enfocarse en la integración de modalidades sensoriales adicionales (como el audio) y en la exploración de diferentes contextos culturales para comprobar su inclusividad en los resultados. También se podrían ampliar los datos utilizados para evaluar el modelo, incluyendo una mayor variedad de redes sociales (como *Instagram* o *Facebook*). No se descarta la investigación en colaboración con expertos en psicología, sociología u otras disciplinas relacionadas para obtener una comprensión más completa de las emociones y su expresión en las redes. Estas direcciones no solo ampliarán el conocimiento existente, sino que también podrían tener aplicaciones prácticas significativas. Algunos de estos usos prácticos podrían ser: publicidad o *marketing*, adaptando las campañas de *marketing* conforme a las emociones predominantes en los grupos de usuarios, política, analizando la opinión pública para saber las preocupaciones de los votantes o detectando la propagación de bulos, en salud mental monitoreando el bienestar de los usuarios en plataformas y proporcionando alertas o detectando contenido sensible y permitiendo a los usuarios proteger su privacidad.

Los lingüistas y filólogos son perfectamente capaces en el desarrollo y mejora de estos modelos de diferentes maneras: contribuyendo al proceso de etiqueta de corpus o desarrollando una categorización emocional que refleje la diversidad de expresiones en diferentes idiomas y contextos, ayudando en la evaluación de los resultados de los modelos y analizando las predicciones desde una perspectiva lingüística y cultural añadiendo precisión a los contextos; o, para finalizar, pueden ofrecer consultoría y asesoramiento lingüístico a equipos de desarrollo de IA.

REFERENCIAS[54]

- *Análisis del Estado Actual de Procesamiento de Lenguaje Natural*. (2021). RISTI, N. º E42, 126-136.

- Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The role of emotions in marketing. *Journal of the Academy of Marketing Science*, 27(2), 184-206. URL: https://doi.org/10.1177/0092070399272005

- Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. En *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 258-266). Marseille, France: European Language Resources Association.URL: https://aclanthology.org/2022.lrec-1.27

- Cardiff NLP. (s/f). Twitter-roberta-base-sentiment. URL:https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

- Contreras, H., & Dávila, J. (2001). *Procesamiento del lenguaje natural basado en una "gramática de estilos" para el idioma español*. Centro de Investigación y Proyectos en Simulación y Modelos, Postgrado en Computación. Universidad de los Andes. Venezuela. URL: http://www.saber.ula.ve/bitstream/handle/123456789/15961/CLEI_2001-a218.pdf?isAllowed=y&sequence=1

- Gómez Guinovart, J. (1998). *Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones*. En Baró i Queralt, P. & Cid Leal, P. (Eds.), Anuario SOCADI de Documentación e Información (pp. 135-146). Barcelona: Societat Catalana de Documentació i Informació. URL: http://www.raco.cat/index.php/Bibliodoc/article/viewFile/56629/66051

- Gómez, L. (2012). *Conceptualización y expresión lingüística del evento emocional en español (L1/L2) y francés: un enfoque cognitivo: análisis lingüístico y proposición didáctica* [Tesis doctoral, Universidad de Grenoble; Universidad de Granada]. URL: https://theses.hal.science/tel-01152554

[54] No se encuentran referencias a libros debido a la contemporaneidad del tema trabajado en este estudio y la necesidad de actualidad de las fuentes.

- Google. (2024, febrero). *Google Gemini: The next generation model*. URL: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note
- Halvorsen, P.-K. (1991). Las aplicaciones informáticas de la teoría lingüística. En Newmeyer, F. J. (Ed.), *Panorama de la Lingüística Moderna de la Universidad de Cambridge, vol. II: Teoría lingüística: Extensiones e Implicaciones* (pp. 247-271). Madrid: Visor. (Traducción de J. Gómez Guinovart y A. Tusón Valls. Edición supervisada por L. Eguren).
- Hugging Face. (s/f). *M2SA-demo-multimodal*. URL: https://huggingface.co/spaces/FFZG-cleopatra/M2SA-demo-multimodal/tree/main
- Johnson, K., & Johnson, H. (Eds.). (1998). *Encyclopedic Dictionary of Applied Linguistics: A Handbook for Language Teaching*. Oxford: Blackwell.
- Liu et al. (2019). RoBERTa: *A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692
- Llamazares, M. V. (2019). *Aproximación a la lingüística computacional* [Tesis doctoral, Universidad de León]. URL: https://doi.org/10.18002/10612/903
- López-López, P., Cuadrado, I., & Navas, M. (2009). *El impacto emocional de la campaña electoral en las elecciones generales españolas de 2008*. Revista de Psicología Social, 24(3), 281-293. URL: https://doi.org/10.1174/021347409788142146
- *Machine Learning for Multimodal Interaction*, First International Workshop, MLMI 2004, Martigny, Switzerland, June 2004, Revised Selected Papers. (2004).
- Meta. (s.f.). *Llama 2*. URL: https://llama.meta.com/llama2/
- Mitchell, W. J. T. (2005). *What Do Pictures Want?: The Lives and Loves of Images*. University of Chicago Press.
- Moreno Sandoval, A. (2015). Antonio Moreno Sandoval. En Gutiérrez Rexach, J. (Ed.), *Enciclopedia de Lingüística Hispánica* (Vol. 1, pp. 204-215). ISBN 978-0-415-84086-6.
- Moure, T. y Llisterri, J. (1996). Lenguaje y nuevas tecnologías: el campo de la lingüística computacional. En M. Fernández Pérez (coord.), *Avances en Lingüística aplicada* (pp. 147-227). Universidade de Santiago de Compostela: Servicio de Publicacións e Intercambio Científico. URL: http://liceu.uab.es/~joaquim/publicacions/llisterri_moure_96.html

- Pew Research Center. (2016). *Religion and emotion regulation: An experimental study of the effects of prayer and scripture reading on anger and aggression*. URL: https://www.pewresearch.org
- OpenAI. (s.f.). *GPT-4 Turbo in the OpenAI API*. URL: https://help.openai.com/en/articles/8555510-gpt-4-turbo-in-the-openai-api
- Russell, S. y Norvig, P. (2010). *Artificial intelligence: A modern approach*, (3.ª ed.), Upper Sadler River, NJ: Prentice-Hall.
- Skim AI. (s.f.). *¿Qué es la AI Multimodal? Casos de uso de la AI Multimodal*. https://skimai.com/es/que-es-la-ai-multimodal-casos-de-uso-de-la-ai-multimodal/
- TweetNLP. (s/f). *Demo de TweetNLP*. URL: https://tweetnlp.org/demo/?q
- Vásquez, A. C., Quispe, J., & Huayna, A. (2009). Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, 6(2), 45-54.

ANEXO A: GLOSARIO.

- Ajuste Fino (*Fine-Tuning*): Después del entrenamiento inicial, a menudo se hace un ajuste fino para especializar el modelo en tareas o tipos de lenguaje específicos.
- Algoritmo: Es un conjunto de instrucciones que le dice a la IA cómo aprender de los datos. Puedes pensar en él como una receta que sigue la computadora.
- Aprendizaje Automático (*Machine Learning*): Una rama de la IA donde las máquinas aprenden a realizar tareas sin ser programadas explícitamente para cada una. Aprenden de los ejemplos y experiencias, como los humanos.
- Aprendizaje Profundo (*Deep Learning*): Una técnica dentro del aprendizaje automático que utiliza redes neuronales (inspiradas en el cerebro humano) para aprender de grandes cantidades de datos.
- Datos de Entrenamiento: Son ejemplos (como textos) que se utilizan para enseñar al modelo. Cuantos más datos de calidad tenga, mejor aprenderá el modelo.
- Entorno de desarrollo: es un programa que facilita la tarea de la programación al desarrollador. Puede incluir herramientas de aprendizaje automático como análisis sintáctico o búsqueda de errores en el código.
- Entrenamiento del Modelo: Es el proceso de enseñar al modelo a entender y generar lenguaje. Esto se hace alimentándolo con grandes cantidades de texto y ajustando sus parámetros para mejorar su rendimiento.
- *Hashtag*: es una manera de etiquetar una palabra o conjunto de palabras (sin espacio) en redes sociales añadiendo el símbolo numeral #. Permiten a los usuarios encontrar todas las publicaciones que posean esa etiqueta y ayuda a popularizarlas. Ejemplo: es muy famoso el *hashtag* #lentejas en *TikTok España* para viralizar el contenido. Otro ejemplo puede ser cuando sucede un evento importante y todo el mundo habla de ello en Twitter: #Eurovision2024.
- Inteligencia Artificial (IA): Es la tecnología detrás de estos modelos. La IA intenta imitar la forma en que los humanos piensan y aprenden.
- Librerías: en PLN son un conjunto de códigos ya desarrollados por una o varias personas que añaden funcionalidades nuevas (funciones aritméticas complejas, análisis sintáctico, el uso de una interfaz, tokenización, etiquetado gramatical, lematización o análisis de sentimientos) que se pueden implementar en nuestro código tan solo utilizando la palabra reservada ‘‘*import*’’ junto al nombre de la librería que se quiera añadir en las primeras líneas del código.

- Modelo multimodal en el contexto de la inteligencia artificial se refiere a un sistema que puede comprender, interpretar y generar información a través de múltiples tipos de datos o modalidades. Estas modalidades pueden incluir texto, imagen, sonido, vídeo, y a veces señales táctiles o de otro tipo.
- Modelo de Lenguaje de Gran Escala (*LLM*): Es un programa informático diseñado para entender y generar lenguaje humano. Imagínalo como un asistente muy avanzado que puede escribir, responder preguntas y hasta componer textos creativos.
- Ontología: es una representación minuciosa y bien estructurada de un conjunto de conceptos, además de las relaciones entre ellos, dentro de un área específica de conocimiento. Es utilizada en el ámbito de la IA para estructurar información para que sea accesible tanto para humanos como para computadoras.
- *Pipeline*: es una serie de pasos que se aplican de manera secuencial para la realización de una tarea específica. Estos pasos se organizan en una estructura de flujo que permite procesar los datos de manera eficiente y sistemática.
- Procesamiento del Lenguaje Natural (PLN): Es un campo de la IA que se centra en cómo las computadoras pueden entender e interactuar con el lenguaje humano.
- Redes Neuronales: Son estructuras de software que imitan la forma en que el cerebro humano procesa la información. Son fundamentales en el aprendizaje profundo.

ANEXO B: IMÁGENES.

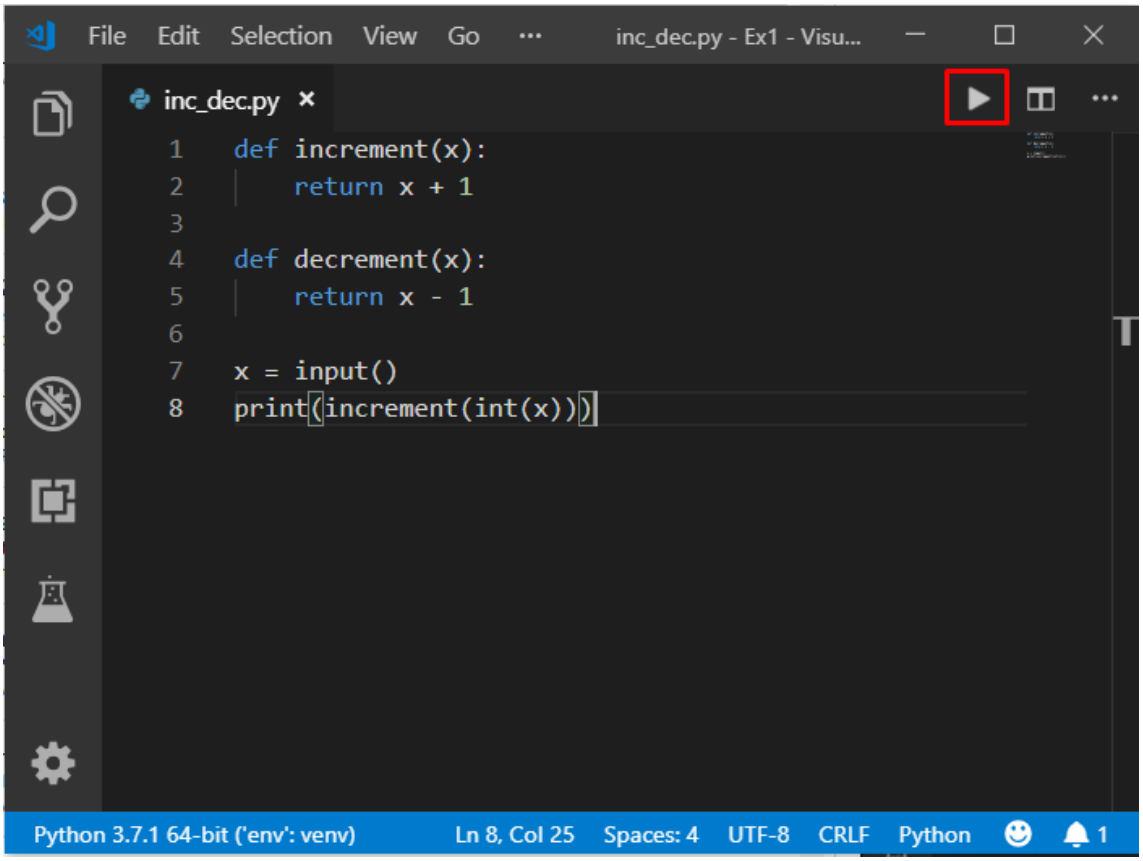

Ilustración 17

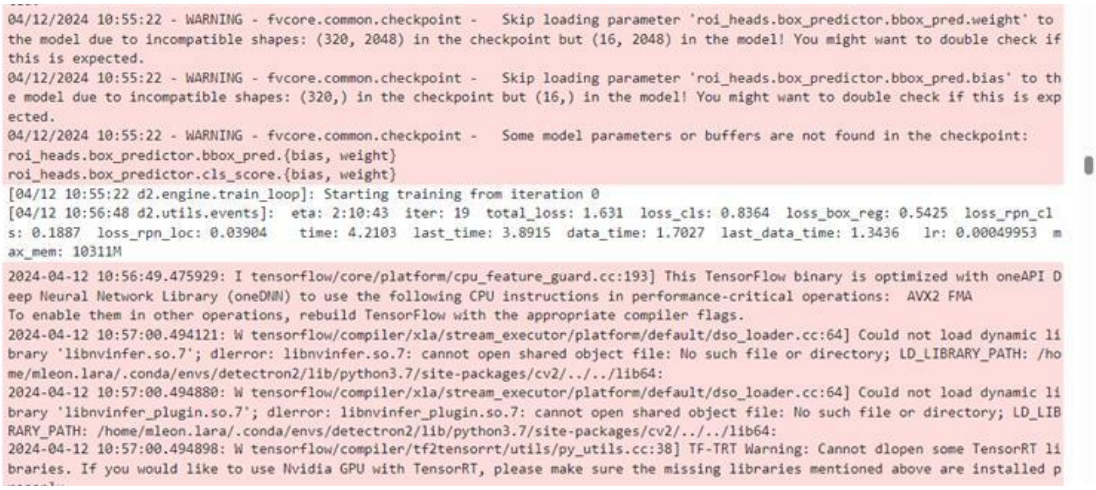

Ilustración 18

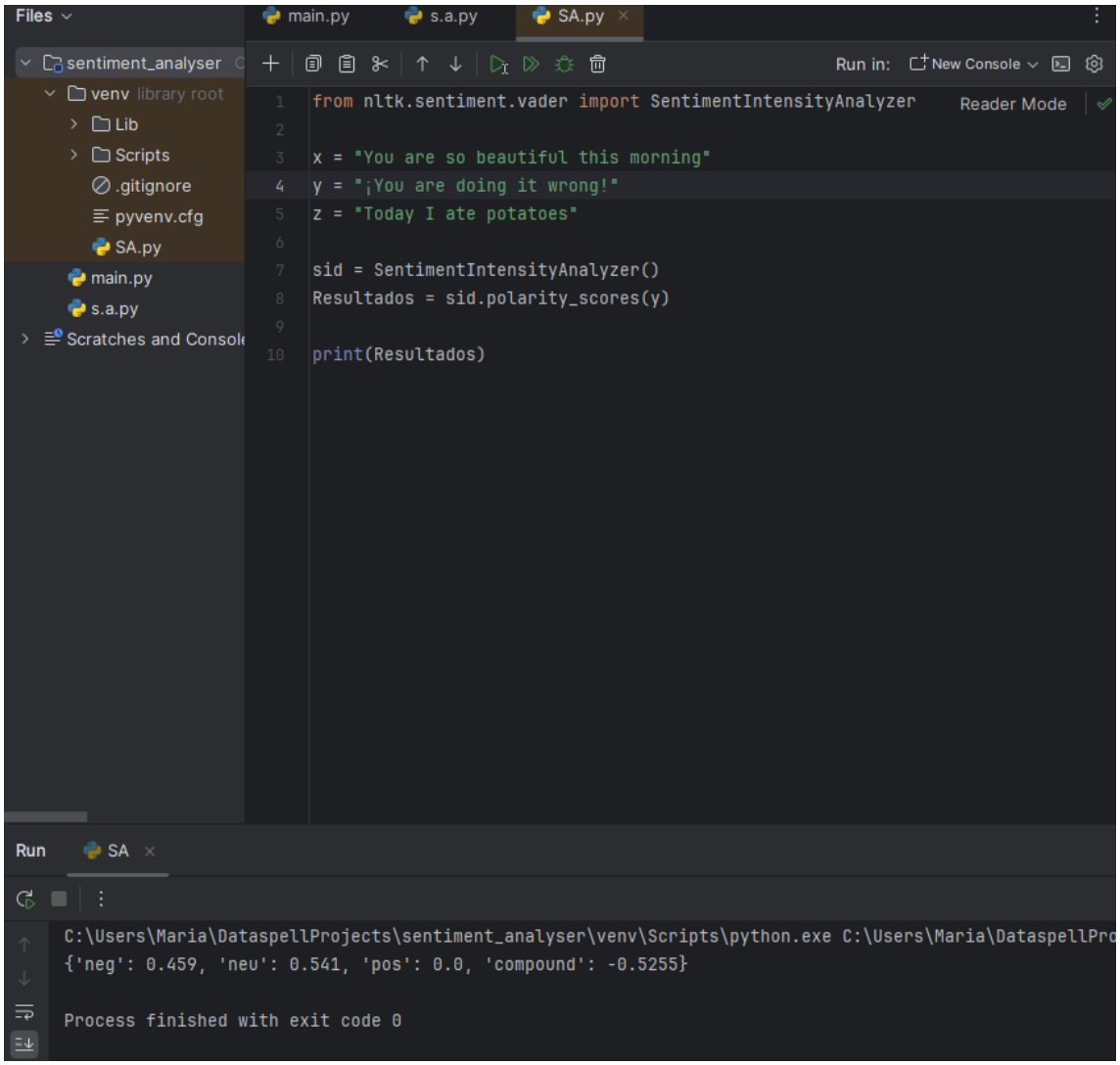

Ilustración 19

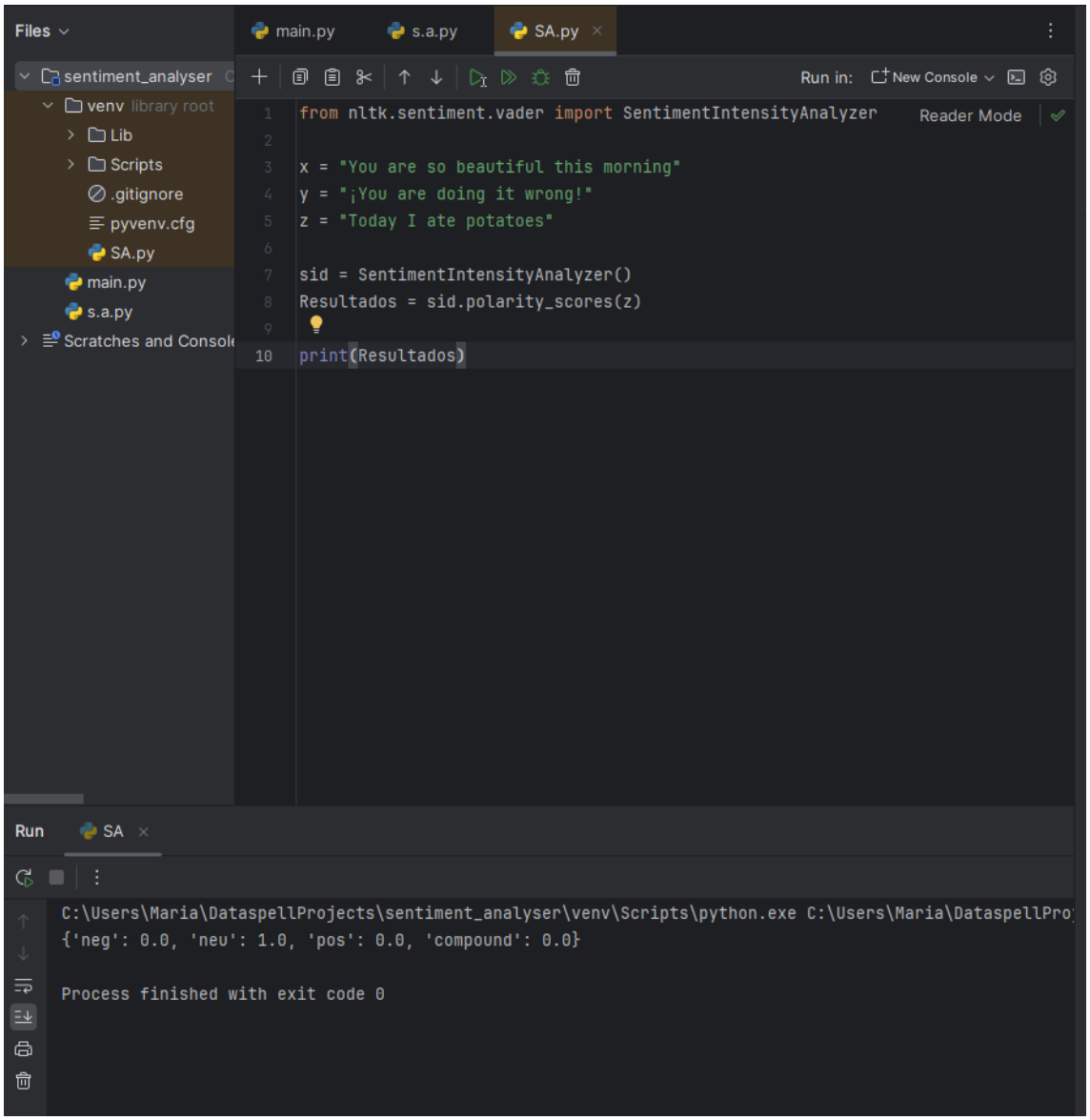

Ilustración 20

0: ❤️

1: 😍

2: 😂

3: 💔

Ilustración 21