

Detección de patrones en conversaciones de WhatsApp mediante técnicas de Procesamiento del Lenguaje Natural



Máster Universitario en Ciencia de Datos

Trabajo Fin de Máster

Autor:

Manuel Alberola Torres

Tutor/es:

José García Rodríguez

David Tomás Díaz

Septiembre 2023



Universitat d'Alacant
Universidad de Alicante

Detección de patrones en conversaciones de WhatsApp mediante técnicas de Procesamiento del Lenguaje Natural

Un sistema de análisis de conversaciones de texto empleando modelos de Machine Learning preentrenados

Autor

Manuel Alberola Torres

Tutor/es

José García Rodríguez

David Tomás Díaz

*Departamento de Tecnología Informática y Computación,
Departamento de Lenguajes y Sistemas Informáticos*



Máster Universitario en Ciencia de Datos



Escuela Politécnica Superior



Universitat d'Alacant
Universidad de Alicante

RESUMEN

Este trabajo desarrolla y experimenta un sistema de análisis de conversaciones de WhatsApp. El sistema es capaz de leer conversaciones, interpretarlas, analizarlas y reproducir los resultados del análisis gráficamente con el objetivo de facilitar al usuario la detección de patrones en la conversación.

Para ello se emplean técnicas de minería de textos que permiten la explotación de la información y, de esta manera, medir automáticamente aspectos que sólo podrían medirse manualmente pero que sería inviable por el gran volumen de datos con el que se trabaja.

Las métricas implicadas en el análisis miden la tasa de participación de los integrantes de la conversación, el flujo de mensajes entre usuarios y la cantidad de palabras, así como la longitud de los mensajes o el sentimiento que estos transmiten. Al representar gráficamente estas métricas el usuario es capaz de detectar patrones de conducta individuales y grupales que, de no ser por el análisis, no sería capaz de detectar.

Algunos ejemplos de patrones que podrían detectarse son, la velocidad de cambio de temática en la conversación, la afinidad entre los integrantes del grupo, fluctuaciones de actividad distribuidas temporalmente y figuras de liderazgo, entre muchos otros.

Para probar la efectividad del sistema se realiza una experimentación con un grupo de WhatsApp real donde aparecen varias personas manteniendo diferentes conversaciones durante un largo periodo de tiempo. Se analiza la conversación y se comenta cada resultado justificando cada patrón detectado.

Índice

RESUMEN	4
AGRADECIMIENTOS	6
PREÁMBULO	7
CONTEXTO	9
1. INTRODUCCIÓN	10
1.1 ANÁLISIS DE DATOS.....	10
1.2 ¿POR QUÉ ES IMPORTANTE EL ANÁLISIS DE DATOS?	11
1.3 FASES DE UN PROYECTO DE ANÁLISIS DE DATOS	12
1.4 MINERÍA DE TEXTOS	14
2. ESTADO DE LA CUESTIÓN	16
2.1 APROXIMACIONES AL ANÁLISIS DE SENTIMIENTO	16
2.2 HERRAMIENTAS Y PLATAFORMAS DE MINERÍA DE TEXTOS	28
2.3 TENDENCIAS ACTUALES EN MINERÍA DE TEXTOS.....	31
2.4 FUTURAS DIRECCIONES DE INVESTIGACIÓN.....	34
3. PROPUESTA	37
3.1 ESCENARIO	37
3.2 OBJETIVOS.....	38
3.3 ALCANCE.....	39
3.4 METODOLOGÍA.....	40
4. DESARROLLO DE LA PROPUESTA	43
4.1 SOLUCIÓN CONCEPTUAL.....	43
4.2 TECNOLOGÍAS SELECCIONADAS	44
4.3 DESARROLLO.....	47
6. EXPERIMENTACIÓN.....	52
7. CONCLUSIONES Y LÍNEAS FUTURAS.....	62
7.1 CONCLUSIONES.....	62
7.2 LÍNEAS FUTURAS	64
8. BIBLIOGRAFÍA Y REFERENCIAS	65
9. LISTA DE ACRÓNIMOS Y ABREVIACIONES	67

AGRADECIMIENTOS

Quisiera dar gracias a todos aquellos que han hecho posible tanto este trabajo como mi crecimiento profesional. A todos los profesores que, a lo largo de tantos años, han tratado de enseñar y ofrecer sus conocimientos. A mi familia y amigos, por haber hecho este proceso mucho más ameno. A la Universidad de Alicante, por haber permitido que fuese posible y, en definitiva, a toda la comunidad docente.

Quisiera, además, agradecer especialmente a mi compañero Álvaro Muñoz Delgado, quien ha estado a mi lado durante todo el camino. Juntos hemos enfrentado y superado todos los retos. A mis tutores José García Rodríguez y David Tomás Díaz, por haberme guiado escuchando mis propuestas y aportando su profesionalidad.

PREÁMBULO

Hasta hace unos años, la información era escasa, cara y difícil de conseguir, por lo que se pretendía que fuese de buena calidad con el fin de que lo que se dedujera a partir de ella resultara válido. Pero todo eso ha cambiado: en la actualidad, la recopilación y almacenamiento de datos masivos se ha simplificado llegando a un punto en el que la cantidad de información crece exponencialmente cada año. Con esas cantidades ingentes de datos, resulta más interesante encontrar el **qué** que el **porqué** de las cosas, lo que supone un cambio drástico de enfoque en la forma de ver el mundo.

Numerosos autores sostienen que nos encontramos al inicio de la “*era de los datos*”, donde el valor de la información ya no reside en los datos concretos sino en la forma de correlacionarlos para descubrir patrones que ni siquiera se habían imaginado, ni por supuesto buscado. En este nuevo paradigma merece la pena tolerar la imprecisión y aceptar el desorden natural del universo, si a cambio se obtiene un sentido más completo de la realidad. Sobre cómo alcanzar ese nuevo sentido desarrollará este trabajo.

¿Por qué el análisis masivo de datos supone un punto de inflexión? ¿Acaso no es equivalente a lo que se hacía en tiempos anteriores, pero a mayor escala? La diferencia se encuentra en el resultado obtenido del análisis. Si bien antes éramos capaces de obtener información a partir de los datos, ahora se pretende transformar dicha información en conocimiento. Porque no, no es lo mismo hablar de datos, que de información, ni mucho menos conocimiento.

Decir que la temperatura es de 23º es dar un dato. Por sí solo no aporta ningún mensaje. ¿Qué son 23º? ¿Es mucho? ¿Es poco? La respuesta depende del contexto. Mientras que 23º en verano es interpretado como una temperatura baja, en invierno es alta. Lo mismo ocurre al contextualizar el dato en el día o la noche, o la ubicación geográfica donde ha sido medido, la altitud o las condiciones meteorológicas concretas. El contexto del dato es imprescindible para poder interpretarlo y obtener información de él. Pero si lo que se pretende, por el contrario, es obtener conocimiento entonces hay que ir más allá. Se requiere comparar con registros de años anteriores y de otras ubicaciones, observar

tendencias respecto a la media, medir la desviación anual... En definitiva, para comprender la realidad que rodea al dato hay que analizarlo junto al resto de componentes y no de forma aislada.

Entramos entonces en un campo muy amplio en el que se combinan técnicas de recogida y almacenamiento de datos, modelos analíticos, minería de datos, componentes estadísticos y herramientas de representación de información. A este conjunto de estrategias de obtención de conocimiento a partir de conjuntos masivos de datos se les conoce como Análisis de Datos.

A lo largo de este trabajo se profundizará en la explotación de información, concretamente información textual. Para ello se emplearán técnicas de Minería de Textos y Procesamiento del Lenguaje Natural (PLN). Se desarrollará un sistema de análisis de conversaciones de WhatsApp capaz de leer e interpretar los mensajes con el objetivo de detectar patrones en las interacciones entre usuarios, porque... ¿Somos realmente impredecibles? ¿Podemos considerarnos arbitrarios? ¿Nuestro comportamiento define patrones o los patrones definen nuestro comportamiento?

Estas preguntas no son el foco de este trabajo. Sin embargo, aquellos que deseen abordarlas necesitarán un sistema que facilite la detección de dichos patrones, y el desarrollo de ese sistema sí es el objetivo principal de esta investigación.

Adicionalmente, esta puesta en práctica resulta especialmente interesante porque es el usuario quien genera sus propios datos, que posteriormente serán analizados desvelando patrones que en la mayoría de casos él mismo podrá confirmar. Es una demostración empírica de que la *“era de los datos”* ha llegado a nuestra realidad más cercana.

CONTEXTO

El presente trabajo se encuentra en sintonía con la visión presentada por Andrew McAfee y Erik Brynjolfsson en su artículo 'Big Data: The Management Revolution', publicado en Harvard Business Review. En dicho artículo, los autores argumentan que la abundancia de datos en la era de la información no solo radica en su volumen, sino en la capacidad de las organizaciones para extraer conocimiento y tomar decisiones fundamentales a partir de estos datos. (McAfee, A., & Brynjolfsson, E. 2012).

El enfoque de este trabajo se basa en la idea central de que los datos por sí solos no son suficientes; es esencial utilizar técnicas avanzadas de análisis, como la minería de textos y el Procesamiento del Lenguaje Natural, para convertir datos en información valiosa y conocimiento significativo.

Al desarrollar un sistema de análisis de conversaciones de WhatsApp que detecta patrones en las interacciones de los usuarios, se busca demostrar la aplicabilidad práctica de las teorías presentadas por McAfee y Brynjolfsson, respaldando la noción de que la Ciencia de Datos es un motor de la transformación empresarial y la toma de decisiones estratégicas en la actualidad.

Así mismo, esta aproximación también se alinea con el enfoque presentado en el libro 'Natural Language Processing in Action' de Lane, Howard y Hapke. En dicho libro, los autores destacan cómo el PLN se ha convertido en una herramienta fundamental para el análisis de datos textuales, permitiendo la extracción de información valiosa y la detección de patrones en el lenguaje humano. (Lane H, Howard C & Hapke H. 2019)

De igual forma, al desarrollar un sistema de análisis de conversaciones de WhatsApp, se busca demostrar la aplicabilidad práctica de las técnicas de PLN en la obtención de conocimiento a partir de datos textuales, respaldando así la noción de que la Ciencia de Datos no se limita a la cantidad de datos, sino a la capacidad de transformarlos en información valiosa y conocimiento significativo a través de análisis avanzados.

1. INTRODUCCIÓN

En el transcurso de este capítulo, se abordará la comprensión del proceso de análisis de datos, su importancia en el panorama actual y cómo se puede llevar a la práctica. Se explorarán las fases que son comunes a cualquier proyecto de análisis, así como los desafíos y dificultades inherentes a cada una de ellas. Además, se presentará el concepto de Procesamiento del Lenguaje Natural y se analizará cómo este conjunto de técnicas puede contribuir de manera significativa a la explotación de la información contenida en los datos.

1.1 ANÁLISIS DE DATOS

Al consultar la bibliografía se encuentran diversas definiciones, pero una de las más completas y ampliamente aceptadas es la que expresa:

“Es el conjunto de estrategias, tecnologías y metodologías que nos van a ayudar a convertir los datos en información de calidad, y dicha información en el conocimiento que nos permita una toma de decisiones más acertadas y nos ayude así a mejorar nuestra competitividad” (Ramos, S. 2017)

Esta definición abarca de manera integral el alcance del concepto porque recoge todas las partes esenciales que lo conforman y subraya el fin al que se aspira.

Es frecuente encontrar fuentes donde se habla de análisis de datos como el conjunto de tecnologías que facilitan el análisis, pero no es del todo correcto. El análisis de datos es mucho más que eso, va más allá de la tecnología, aunque ésta sea necesaria.

Es imprescindible seguir una serie de estrategias que garanticen que el modelo analítico es funcional. Así como unas metodologías a través de las cuales se alcance un **sistema válido**. El éxito del estudio dependerá principalmente de la validez del sistema.

La importancia de que el sistema analítico sea válido no puede ser subestimada. ¿Pero qué significa exactamente que un sistema sea válido? Al diseñar un proyecto de análisis de datos se vuelve crucial reconocer que el sistema analítico desarrollado servirá como referente principal para la toma de decisiones futuras. Aquí entra en juego un factor determinante para el éxito del estudio: la **confianza** en la precisión de los resultados generados por el sistema. En consecuencia, los datos deben ser tratados con la máxima consideración; deben ser procesados y estructurados de manera que se garantice la coherencia y precisión, eliminando cualquier posibilidad de inconsistencias.

Por lo tanto, hablar de analítica de datos va más allá de la tecnología en sí. Requiere un enfoque holístico que incluya estrategias y metodologías para asegurar la validez y, en última instancia, la confianza en los resultados proporcionados por el sistema.

1.2 ¿POR QUÉ ES IMPORTANTE EL ANÁLISIS DE DATOS?

Es imperativo dedicar una sección de este capítulo a explorar la significativa importancia del análisis de datos, especialmente en el contexto empresarial. Esto tiene como propósito justificar y esclarecer los beneficios sustanciales que un sistema analítico puede aportar a empresas y organizaciones en general.

Es crucial tener en cuenta que el análisis y la obtención de conocimiento son posibles en prácticamente cualquier conjunto de datos, independientemente de su origen. En consecuencia, el alcance de un proyecto analítico puede ser aplicado en una amplia variedad de áreas.

A través de la observación de un análisis de datos, es posible responder a preguntas fundamentales como "*¿qué está ocurriendo en un proceso de negocio específico?*". Sin embargo, la verdadera riqueza se manifiesta en su capacidad para responder a preguntas más profundas, como "*¿por qué está ocurriendo esto?*". Esto se logra mediante el entendimiento y el análisis cruzado de datos. En última instancia, permite anticipar

comportamientos futuros, lo que permite responder a cuestionamientos cruciales, como *"¿qué ocurrirá en el futuro?"*. Este conocimiento adquirido capacita a las empresas para tomar decisiones informadas y estratégicas con la máxima precisión.

La disponibilidad de este conocimiento en tiempo real ofrece ventajas considerables, ya que permite una planificación más efectiva y la capacidad de tomar medidas inmediatas para mitigar el impacto de cualquier adversidad, lo que, a su vez, reduce los costos asociados. Además, un aspecto que a menudo pasa desapercibido, pero que es de vital importancia, es que proporciona una sólida base para justificar las decisiones tomadas. En entornos empresariales de gran envergadura, esta capacidad de justificación transmite una sensación de control esencial para la toma de decisiones eficaces.

En resumen, el análisis de datos no solo brinda información valiosa, sino que también empodera a las empresas con la capacidad de comprender, anticipar y actuar en función de las tendencias y patrones emergentes en tiempo real. Esta perspicacia se traduce en una planificación más efectiva, toma de decisiones precisas y, en última instancia, en una ventaja competitiva destacada en un mercado en constante cambio.

1.3 FASES DE UN PROYECTO DE ANÁLISIS DE DATOS

Un proyecto de análisis de datos sigue un proceso estructurado que consta de tres fases fundamentales, las cuales son esenciales para lograr resultados exitosos. Estas fases son comunes a cualquier desarrollo y están intrínsecamente relacionadas con las metodologías sólidas mencionadas anteriormente.

Al seguir estas fases, se garantiza la consistencia de la información y, de este modo, se construye la confianza de los usuarios en el sistema. Esta confianza es un factor crítico para el éxito del proyecto, ya que influye en la adopción y la utilización efectiva de las soluciones de análisis de datos.

1.3.1 FASE DE INTEGRACIÓN DE DATOS

La primera fase del proyecto se centra en el proceso de integración de datos. En esta etapa se adquieren datos de diversas fuentes y se los prepara para su posterior análisis.

Esto implica la recopilación, limpieza y transformación de datos de múltiples orígenes, que pueden incluir bases de datos, registros de transacciones, sensores, datos en tiempo real y mucho más.

La calidad y la integridad de los datos adquiridos son esenciales para garantizar que los resultados del análisis sean precisos y confiables.

1.3.2 FASE DE ANÁLISIS DE DATOS

La segunda fase es el proceso de análisis de datos es una etapa crucial en la que los datos se enriquecen con la lógica empresarial y se extrae información valiosa mediante técnicas de minería de datos.

Durante esta fase, se aplican algoritmos y modelos avanzados para descubrir patrones, tendencias y relaciones ocultas en los datos. Esta es la etapa en la que se transforma la información en conocimiento accionable.

El análisis de datos puede abordar una variedad de objetivos, desde la identificación de oportunidades de mercado hasta la detección de anomalías y la optimización de procesos empresariales.

1.3.3 FASE DE INFORMACIÓN Y VISUALIZACIÓN

La tercera y última fase se enfoca en informar y presentar los resultados de manera que se facilite la obtención de conocimiento a partir de ellos. En esta etapa, la información procesada y los conocimientos extraídos se presentan a los usuarios de manera efectiva. Esto implica la visualización de datos utilizando gráficos, tablas, informes interactivos y otras herramientas visuales.

La visualización de datos desempeña un papel esencial en la comunicación de información compleja de manera comprensible y accesible.

En este trabajo, el enfoque se centrará principalmente en las dos últimas fases: la explotación de datos y la visualización de los resultados obtenidos. Se utilizarán técnicas de minería de datos, en particular la minería de textos, para explorar y analizar datos de texto con el objetivo de descubrir patrones, tendencias y conocimientos relevantes en las conversaciones de WhatsApp.

Este enfoque permitirá demostrar cómo las técnicas de análisis de datos pueden aplicarse de manera efectiva para obtener información valiosa en un contexto práctico.

1.4 MINERÍA DE TEXTOS

La minería de textos emerge como un campo de estudio revolucionario e interdisciplinar que abarca un conjunto de técnicas y herramientas diseñadas para extraer información de grandes colecciones de texto, revelando tanto información explícita como implícita.

Estas técnicas son fundamentales para abordar la abrumadora cantidad de información textual que se genera a diario y que sería imposible de analizar eficientemente por un ser humano.

La información se encuentra codificada en el lenguaje humano, y para comprenderla y procesarla, la minería de textos debe interpretar los textos en cuestión. Esto implica una comprensión, aunque sea parcial, del contenido de los textos para extraer información relevante.

La disciplina computacional que idea y desarrolla sistemas para la interpretación de textos es el Procesamiento del Lenguaje Natural, disciplina híbrida entre la lingüística y la computación, la cual se centra en el procesamiento de las comunicaciones humanas, dividiéndolas en partes e identificando los elementos más relevantes del mensaje.

El campo del PLN ofrece una amplia gama de aplicaciones, que van desde la traducción automática hasta el reconocimiento y la síntesis del habla. En este proyecto, el enfoque se centrará específicamente en una de sus aplicaciones más innovadoras: el análisis de sentimientos y emociones en textos. (Mohammad, S.M. 2016)

1.4.1 ANÁLISIS DE SENTIMIENTO

El análisis de sentimiento es una rama del PLN que se centra en identificar, extraer y comprender los sentimientos expresados en el lenguaje escrito. Su objetivo principal es determinar si un fragmento de texto (como una revisión de un producto, un comentario en redes sociales o una respuesta de encuesta) transmite una connotación positiva, negativa o neutra, así como comprender la intensidad y el contexto de esos sentimientos.

La detección de sentimientos es una técnica cada vez más relevante, utilizada por empresas y profesionales del marketing para comprender cómo se sienten los usuarios sobre una marca, producto o servicio. En este caso, el objetivo será comprender los sentimientos de los usuarios en conversaciones de WhatsApp a través del análisis de sus mensajes.

Desde la perspectiva de la minería de textos, el análisis de sentimientos implica una clasificación automática de documentos en función de la connotación del lenguaje utilizado. En términos generales, busca determinar la **actitud** de un interlocutor hacia un tema específico o la polaridad general contextual de un documento.

La **actitud** puede comprenderse como el **estado afectivo**, es decir, el estado emocional del autor al escribir el mensaje o la **intención comunicativa emocional**, es decir, el efecto emocional que el autor pretende transmitir al lector.

Esta capacidad de analizar sentimientos en el texto es esencial para comprender las opiniones y percepciones de los usuarios y puede tener aplicaciones valiosas en campos como la gestión de la reputación en línea, la atención al cliente y la toma de decisiones estratégicas en marketing y branding.

2. ESTADO DE LA CUESTIÓN

En este capítulo se realizará un estudio de distintas tecnologías y herramientas de minería de textos desarrolladas para el análisis de sentimiento. Se expondrán los factores más relevantes a la hora de seleccionar una u otra dependiendo del proyecto que se va a abordar.

2.1 APROXIMACIONES AL ANÁLISIS DE SENTIMIENTO

El análisis de sentimiento es un campo multidisciplinario que ofrece varias aproximaciones para comprender las emociones y opiniones expresadas en el lenguaje escrito. Inicialmente estas aproximaciones se dividían principalmente en dos categorías: supervisadas y no supervisadas. (Pang, B. & Lee, L. 2008)

A medida que fue madurando el conocimiento fueron apareciendo nuevas aproximaciones como la híbrida y el Deep Learning.

Cada enfoque tiene sus propias ventajas y desafíos, y se decide cuál emplear en cada caso en función de la disponibilidad de datos etiquetados, el dominio de aplicación y los objetivos del proyecto. A continuación, se exponen las distintas aproximaciones:

2.1.1 APROXIMACIÓN SUPERVISADA

Las aproximaciones supervisadas se basan en el uso de conjuntos de datos etiquetados. Estos conjuntos contienen fragmentos de texto previamente clasificados en categorías de sentimiento, como positivo, negativo o neutro. Los algoritmos de aprendizaje automático se entrenan con estos datos etiquetados para aprender patrones y reglas que les permitan clasificar automáticamente nuevos textos en las categorías apropiadas. (H. A. Patrick, P. G. J, M. H. Sharief and U. Mukherje 2023)

Algunos de los métodos de aprendizaje supervisado incluyen el uso de algoritmos de clasificación como SVM (Máquinas de Soporte Vectorial), Naive Bayes y redes neuronales.

Las ventajas de las aproximaciones supervisadas incluyen su capacidad para lograr un alto rendimiento en dominios específicos, donde se dispone de suficientes datos etiquetados para el entrenamiento. Sin embargo, su desafío radica en la necesidad de conjuntos de datos etiquetados de calidad y en su falta de flexibilidad para adaptarse a nuevos dominios o lenguajes sin etiquetar.

El objetivo de esta aproximación es conseguir un modelo entrenado a partir de un conjunto de datos etiquetado. La estrategia para obtenerlo es la siguiente:

En primer lugar, se recopila un conjunto de datos etiquetados con categorías de sentimiento, como positivo, negativo o neutro. Estos datos se utilizan para entrenar un modelo inicial. Estos datos deben preprocesarse antes de utilizarlos para entrenar un modelo. Este proceso incluye:

- Tokenización: Dividir el texto en unidades más pequeñas, como palabras o frases.
- Eliminación de Stop Words: Eliminar palabras comunes que no aportan información significativa (por ejemplo, "y", "el", "en").
- Lematización o Stemming: Reducir las palabras a su forma base para simplificar el análisis (por ejemplo, "corriendo" se reduce a "correr").

A continuación, se extraen las características relevantes de los textos preprocesados. Las características pueden incluir:

- Frecuencia de Palabras: Contar la frecuencia de aparición de cada palabra en el texto.
- TF-IDF (Term Frequency-Inverse Document Frequency): Calcular una puntuación que destaque las palabras importantes en el contexto del documento y en comparación con otros documentos.

- Embeddings de Palabras: Representar las palabras en un espacio vectorial para capturar similitudes semánticas.

En este paso el texto se convierte en un vector de características. Este puede ser de diversas características como, por ejemplo, la frecuencia de las palabras en el texto, o la frecuencia adjetivos, palabras de opinión, metadatos como fechas, categorías ... (Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. & Potts, C. 2011)

Cada palabra distinta del texto ocupa una posición del vector de características.

Ejemplo:

Frase 1 - "Este momento es maravilloso" – Connotación positiva

Frase 2 - "Este libro es aburrido" – Connotación negativa

Si la palabra pertenece a la frase asignamos un 1, si no un 0:

Tabla 1. Ejemplo aproximación supervisada

FRASE	ESTE	MOMENTO	ES	MARAVILLOSO	LIBRO	ABURRIDO	POLARIDAD
Frase 1	1	1	1	1	1	0	POS
Frase 2	1	0	1	0	1	1	NEG

A partir de dichos vectores se crean los modelos computacionales mediante aprendizaje automático. En la *Tabla 1* el modelo podría aprender que la palabra "ESTE" tiene tanto connotaciones positivas como negativas mientras que "MARAVILLOSO" sólo positivas o "ABURRIDO" sólo negativas.

Si se dispone de un conjunto de datos lo suficientemente grande y correctamente etiquetado, el vector de características resultante es muy eficaz.

2.1.2 APROXIMACIÓN NO SUPERVISADA

Las aproximaciones no supervisadas se aplican cuando no se dispone de un conjunto de datos etiquetados para el análisis de sentimientos. En lugar de depender de datos etiquetados, estas técnicas utilizan recursos como diccionarios, lexicones y ontologías para identificar y cuantificar las emociones y opiniones en el texto. Los algoritmos de agrupamiento, como el Análisis de Componentes Principales (PCA) y la agrupación k-means, se utilizan a menudo en enfoques no supervisados. (Karabelly, J, 2020)

Una de las ventajas de las aproximaciones no supervisadas es su aplicabilidad en dominios generales donde la disponibilidad de datos etiquetados es limitada. Sin embargo, pueden ser menos precisas en comparación con las técnicas supervisadas, ya que no se basan en ejemplos etiquetados para aprender patrones de sentimiento.

En esta aproximación no se dispone de un conjunto de datos etiquetados, únicamente de un recurso previamente preparado, generalmente un lexicón que no es más que un diccionario con las palabras previamente etiquetadas. (Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. 2011)

Algunos lexicones conocidos son:

- MPQA¹: Asociaciones de palabras con su polaridad
- SentiWordNet²: Indica el grado de positividad/negatividad de cada palabra (Baccianella et al., LREC 2010)
- Senticon: Indica el grado de positividad/negatividad de cada palabra en español (Cruz, Fermín L et al., 2014).
- EmoLex³: Asociaciones de palabras con su polaridad y emoción

A partir del lexicón se puede captar tanto el grado de polaridad como la pertenencia a una emoción. El procedimiento es el siguiente:

¹ MPQA: <https://mpqa.cs.pitt.edu/>

² SentiWordNet: <https://github.com/aesuli/SentiWordNet>

³ EmoLex: <https://lidilem.univ-grenoble-alpes.fr/ressources/corpus/emolex>

En primer lugar, al igual que en la aproximación supervisada, deben preprocesarse los datos antes de utilizarlos para entrenar un modelo. El procedimiento es el previamente comentado.

A continuación, se extraen las características relevantes de los textos preprocesados. A diferencia de la aproximación supervisada, donde se utilizan características como la frecuencia de palabras o el TF-IDF⁴, en este enfoque no supervisado, las características pueden incluir:

- **Análisis de Tópicos:** Identificación de temas clave en el texto mediante técnicas de modelado de tópicos, como Latent Dirichlet Allocation (LDA). (Ossai, C. I., & Wickramasinghe, N. 2023)
- **Agrupación (Clustering):** Agrupación de documentos o textos similares en clústeres o grupos basados en la similitud de contenido.
- **Análisis de Coocurrencia:** Identificación de palabras que suelen aparecer juntas en el texto, lo que puede revelar relaciones semánticas.

Posteriormente se aplica un modelo de análisis de sentimientos no supervisado para asignar polaridades o tonos emocionales a los textos. Los métodos comunes incluyen:

- **Análisis de Polaridad Lexical:** Utilización de diccionarios de palabras con etiquetas de sentimiento (positivo, negativo o neutro) para asignar polaridades a las palabras en el texto y calcular una polaridad global para el documento.
- **Agrupación de Sentimientos:** Aplicación de algoritmos de agrupación (clustering) para agrupar textos similares en categorías de sentimiento sin etiquetas previas.
- **Análisis de Coocurrencia de Palabras:** Identificación de patrones de coocurrencia que indican tonos emocionales en el texto.

⁴ TF-IDF: <https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558>

Finalmente se divide el texto en palabras y se buscan en el lexicón obteniendo su valor asociado a cada categoría. Se agregan todos los valores agrupados por categoría (sumados o contados) y se obtiene la categoría con más peso, es decir, la más probable.

Ejemplo:

Genial – POS (0,9) | NEG (0,1)

Genial – Felicidad (0,8) | Disgusto (0,0) | Miedo (0,1) | Tristeza (0,1) | Enfado (0,0)

La palabra “Genial” pertenece al 90% al sentimiento positivo y al 80% a la emoción felicidad.

2.1.3 APROXIMACIÓN HÍBRIDA

Las aproximaciones híbridas buscan combinar lo mejor de ambos mundos: las ventajas de las técnicas supervisadas y no supervisadas. Estos enfoques pueden utilizar técnicas supervisadas para etiquetar automáticamente datos no etiquetados, lo que amplía su conjunto de datos y mejora su rendimiento. También pueden integrar diccionarios y ontologías para refinar la comprensión de las emociones en el texto. (Lee, V.L., Gan, K.H., Tan, T., & Abdullah, R. 2019)

A continuación, se detallan los pasos de cómo funciona una aproximación híbrida:

1. Fase Supervisada

Al igual que en la aproximación supervisada, se recopila un conjunto de datos etiquetados con categorías de sentimiento, como positivo, negativo o neutro. Estos datos se utilizan para entrenar un modelo inicial.

A continuación, se entrena el Modelo Supervisado: Se utiliza un algoritmo de aprendizaje supervisado, como máquinas de soporte vectorial o redes neuronales, para entrenar un modelo inicial en función de los datos etiquetados. El modelo aprende a clasificar textos en categorías de sentimiento basándose en las etiquetas de entrenamiento.

2. Fase No Supervisada

Se recopila un conjunto adicional de datos de texto no etiquetados en el que se desea realizar el análisis de sentimientos y se le realiza el preprocesamiento de datos, que incluye la tokenización, eliminación de stop words y lematización, al igual que en la aproximación supervisada.

A continuación, se extraen características relevantes de los textos preprocesados. Esto puede incluir características como la frecuencia de palabras o la representación TF-IDF. Finalmente se aplica un enfoque no supervisado, como el análisis de tópicos o la agrupación de documentos, para identificar patrones de sentimiento en los datos no etiquetados. Esto puede resultar en la creación de grupos o clústeres de textos con sentimientos similares.

3. Fase de Fusión de Resultados

En esta etapa, se combinan los resultados de ambas fases (supervisada y no supervisada) para obtener una evaluación conjunta de los textos en términos de sentimiento. Esto se puede lograr de diversas formas:

Los resultados de la fase supervisada y no supervisada se ponderan según su confiabilidad o precisión individual. Por ejemplo, si el modelo supervisado tiene una alta precisión en ciertos casos, se puede dar más peso a sus predicciones en esos casos.

Se pueden fusionar categorías de sentimiento de ambas fases para obtener una categorización más completa. Por ejemplo, si la fase supervisada identifica un sentimiento positivo y la fase no supervisada encuentra un sentimiento de satisfacción, estos se pueden fusionar en una categoría "positiva".

En algunos casos, los resultados de ambas fases pueden requerir afinación manual por parte de un experto en el dominio para mejorar la precisión y la coherencia.

Ejemplo:

Frase 1 - "Este momento es maravilloso"

Frase 2 - "Este libro es aburrido"

Se cuenta la cantidad de palabras que pertenecen al diccionario de palabras positivas:

Tabla 2. Ejemplo aproximación híbrida

FRASE	PALABRAS POSITIVAS	PALABRAS NEGATIVAS	POLARIDAD
Frase 1	1	0	POS
Frase 2	0	1	NEG

A partir de dichos vectores se crean los modelos computacionales mediante aprendizaje automático. Tal y como se refleja en la *Tabla 2*, el modelo podría aprender que las frases con más palabras positivas que negativas suelen ser positivas.

2.1.4 APROXIMACIÓN DE APRENDIZAJE PROFUNDO

La aproximación de Aprendizaje Profundo es un caso particular de aproximación supervisada donde se emplean redes neuronales profundas, como las redes neuronales recurrentes (RNN), las redes neuronales convolucionales (CNN) y las redes neuronales de atención (Transformers), para analizar el lenguaje natural y extraer información de texto no estructurado. Estas redes tienen la capacidad de aprender automáticamente características y patrones complejos en el texto, lo que las hace adecuadas para tareas de análisis de sentimientos.

Redes Neuronales Recurrentes: Estas redes son adecuadas para el análisis de sentimientos en secuencias de texto, como opiniones en redes sociales o revisiones de productos. Las RNN tienen conexiones recurrentes que les permiten mantener una memoria a largo plazo de las secuencias de texto y capturar dependencias temporales.

Redes Neuronales Convolucionales: Aunque se utilizan comúnmente en el procesamiento de imágenes, las CNN también se aplican al procesamiento de texto.

Pueden extraer características relevantes de las secuencias de palabras y se utilizan para tareas de análisis de sentimientos basadas en el contexto local.

Redes Neuronales de Atención: Las redes Transformer, en particular el modelo BERT (Bidirectional Encoder Representations from Transformers), han revolucionado el procesamiento del lenguaje natural. Son capaces de capturar relaciones de largo alcance en el texto y han demostrado un rendimiento sobresaliente en tareas de análisis de sentimientos y comprensión del lenguaje. (Devlin J, Chang M, Lee K, Toutanova K. 2018)

El Aprendizaje Profundo ha demostrado ser altamente efectivo en el análisis de sentimientos debido a su capacidad para aprender automáticamente representaciones de texto y comprender la semántica y el contexto. Sin embargo, a menudo requiere grandes conjuntos de datos etiquetados y recursos computacionales significativos para el entrenamiento.

Al igual que en las aproximaciones supervisadas, se requiere un conjunto de datos etiquetados que contenga ejemplos de texto junto con etiquetas de sentimiento (positivo, negativo o neutro). Estos datos se utilizan para entrenar y evaluar el modelo de aprendizaje profundo.

A continuación, se preprocesan los datos de manera similar a otras aproximaciones. Esto incluye tokenización, eliminación de Stop Words y lematización o Stemming.

Posteriormente se extraen características, pero en lugar de utilizar características específicas como enfoques anteriores (por ejemplo, frecuencia de palabras o TF-IDF), el aprendizaje profundo utiliza representaciones de texto basadas en vectores de palabras pre-entrenados o embeddings de palabras. Estos vectores capturan la semántica de las palabras y permiten que la red neuronal capture relaciones semánticas y contextuales entre palabras.

Se procede con el diseño de la arquitectura de la red, la cual es la parte principal de esta aproximación. Puede ser una red CNN, RNN o una combinación de ambas. La elección de la arquitectura depende del problema específico y del tipo de datos.

Las CNN son efectivas para capturar patrones locales en secuencias de texto. Se utilizan capas convolucionales para detectar características relevantes en fragmentos de texto. Mientras que las RNN son adecuadas para capturar dependencias a largo plazo en el texto. Las RNN procesan secuencias de texto de manera recursiva, teniendo en cuenta la información anterior para comprender el contexto.

Finalmente se entrena el Modelo utilizando el conjunto de datos etiquetados para entrenar la red neuronal. Durante el entrenamiento, la red aprende a mapear las representaciones de texto a las etiquetas de sentimiento correspondientes mediante la adaptación de los pesos de las capas de la red. La retropropagación del error es una técnica clave para ajustar los pesos y minimizar el error de predicción.

2.1.5 CONCLUSIONES

La elección entre estas cuatro aproximaciones dependerá de la complejidad de la tarea de análisis de sentimientos, la disponibilidad de datos y los recursos disponibles. El Aprendizaje Profundo se considera una de las opciones más avanzadas y poderosas, pero también puede ser más exigente en términos de recursos y experiencia en el modelado de redes neuronales.

Supervisada

- Ventajas
 - Precisión en dominios específicos: Cuando se dispone de suficientes datos etiquetados y representativos del dominio de interés, las aproximaciones supervisadas pueden lograr una alta precisión en la clasificación de sentimientos.
 - Flexibilidad para abordar tareas específicas: Se pueden entrenar modelos supervisados para analizar sentimientos en tareas específicas, como la

detección de opiniones en reseñas de productos o la clasificación de sentimientos en redes sociales.

- Desafíos
 - Dependencia de datos etiquetados: Esta aproximación requiere un conjunto de datos etiquetados de alta calidad, lo que puede ser costoso y requerir mucho tiempo de etiquetado manual.
 - Limitaciones en dominios no etiquetados: Puede tener dificultades para adaptarse a dominios o idiomas sin datos etiquetados disponibles.

La aproximación supervisada es efectiva cuando se dispone de datos etiquetados adecuados y se aplica en un dominio específico. Puede proporcionar resultados precisos y es ampliamente utilizada en aplicaciones comerciales como la evaluación de comentarios de clientes, la monitorización de redes sociales y la clasificación de opiniones en reseñas en línea.

No supervisada

- Ventajas:
 - No requiere datos etiquetados: No depende de la disponibilidad de datos etiquetados, lo que la hace adecuada para dominios en los que la etiquetación manual es difícil o costosa.
 - Flexibilidad: Puede aplicarse a una amplia variedad de dominios y tipos de texto.
- Desafíos:
 - Menos precisión: Puede ser menos preciso que los enfoques supervisados, ya que no se basa en ejemplos etiquetados.
 - Requiere afinación y validación: Los resultados de sentimiento pueden requerir afinación y validación manual para garantizar la precisión.
 - Sensible a ruido en los datos: Puede ser sensible a ruido en los datos de entrada, lo que puede afectar la calidad de los resultados.

La aproximación no supervisada es útil cuando se carece de etiquetas de sentimiento y se necesita una solución flexible para el análisis de sentimientos en textos no

estructurados. Es especialmente valiosa en la exploración inicial de datos y la identificación de patrones ocultos en grandes conjuntos de datos textuales.

Híbrida

- Ventajas:
 - Mayor precisión: La combinación de enfoques supervisados y no supervisados tiende a producir resultados más precisos y adaptables en una variedad de dominios y tipos de texto.
 - Flexibilidad: Permite adaptarse a la disponibilidad de datos etiquetados y no etiquetados, lo que la hace adecuada para una amplia gama de aplicaciones.
- Desafíos:
 - Requiere conocimientos expertos: La fusión de resultados y la afinación manual pueden requerir conocimientos expertos en el dominio y en técnicas de procesamiento de lenguaje natural.
 - Mayor complejidad: La implementación de una aproximación híbrida puede ser más compleja que los enfoques puramente supervisados o no supervisados.

La aproximación híbrida es especialmente útil cuando se buscan resultados de análisis de sentimientos precisos y adaptables en entornos en los que se dispone de una combinación de datos etiquetados y no etiquetados. Combina la fuerza de ambos enfoques para obtener una evaluación más completa de los sentimientos en el texto.

Aprendizaje profundo

- Ventajas:
 - Captura de características complejas: Las redes neuronales profundas pueden capturar características complejas y relaciones semánticas en el texto, lo que las hace adecuadas para el análisis de sentimientos.
 - Adaptabilidad: Son altamente adaptables a una variedad de dominios y tipos de texto.

- Desafíos:
 - Requiere grandes conjuntos de datos: El entrenamiento eficaz de modelos de aprendizaje profundo a menudo requiere grandes conjuntos de datos etiquetados y recursos computacionales.
 - Afinación de hiperparámetros: La elección adecuada de la arquitectura de la red y los hiperparámetros es esencial y puede requerir experimentación.

La aproximación de aprendizaje profundo es altamente efectiva para el análisis de sentimientos y se ha convertido en un enfoque líder en esta área debido a su capacidad para aprender representaciones de texto de alta calidad.

Teniendo una imagen general de las distintas aproximaciones y habiendo comprendido el contexto del problema a solucionar, se considera que la mejor aproximación para desarrollar el motor de análisis es la aproximación de aprendizaje profundo.

2.2 HERRAMIENTAS Y PLATAFORMAS DE MINERÍA DE TEXTOS

En esta sección, se explorarán algunas de las herramientas y plataformas más utilizadas en el campo de la minería de textos y el análisis de sentimientos. Estas herramientas son esenciales para llevar a cabo proyectos de procesamiento de lenguaje natural y pueden facilitar en gran medida la tarea de procesar, analizar y visualizar grandes volúmenes de datos de texto. A continuación, se presentan algunas de las más destacadas:

2.2.1 PYTHON CON BIBLIOTECAS PLN

Python es un lenguaje de programación ampliamente utilizado en el campo de la minería de textos debido a su versatilidad y a la disponibilidad de numerosas bibliotecas de procesamiento de lenguaje natural. Algunas de las bibliotecas más populares incluyen:

- NLTK⁵(Natural Language Toolkit): Es una biblioteca de Python que proporciona herramientas y recursos para trabajar con texto y lenguaje humano. Incluye módulos para tokenización, lematización, análisis de sentimientos y más.
- spaCy⁶: Es una biblioteca PLN de alto rendimiento diseñada específicamente para el procesamiento eficiente de texto en varios idiomas. Ofrece capacidades de procesamiento de texto avanzadas, como el análisis de dependencias y la extracción de entidades.
- TextBlob⁷: Es una biblioteca sencilla de usar que proporciona una API simple para realizar tareas comunes de procesamiento de lenguaje natural, como la clasificación de texto y el análisis de sentimientos.

2.2.2 VADER (VALENCE AWARE DICTIONARY AND SENTIMENT REASONER)

VADER⁸ es una herramienta de análisis de sentimientos diseñada específicamente para el procesamiento de texto en redes sociales. Utiliza un diccionario de palabras y reglas gramaticales para determinar el sentimiento de un texto, incluyendo la polaridad (positiva, negativa o neutra) y la intensidad del sentimiento. VADER es ampliamente utilizado para analizar comentarios y publicaciones en redes sociales debido a su capacidad para manejar lenguaje coloquial y emoticones.

2.2.3 GATE (GENERAL ARCHITECTURE FOR TEXT ENGINEERING)

GATE⁹ es una plataforma de código abierto que proporciona un conjunto completo de herramientas y recursos para el procesamiento de texto y la minería de textos. Ofrece

⁵ NLTK: <https://www.nltk.org/>

⁶ SpaCy: <https://spacy.io/>

⁷ TextBlob: <https://textblob.readthedocs.io/en/dev/>

⁸ VADER: <https://vadersentiment.readthedocs.io/en/latest/>

⁹ GATE: <https://gate.ac.uk/sentiment/>

una amplia gama de capacidades, como tokenización, análisis morfológico, extracción de información y análisis de sentimientos. GATE es altamente personalizable y se utiliza en aplicaciones académicas y empresariales.

2.2.4 IBM WATSON NATURAL LANGUAGE UNDERSTANDING¹⁰

Esta plataforma de análisis de texto basada en la nube de IBM utiliza técnicas de procesamiento de lenguaje natural para extraer información significativa de documentos de texto. Ofrece capacidades avanzadas de análisis de sentimientos, detección de emociones, extracción de entidades y más. Es utilizado por muchas organizaciones para analizar grandes volúmenes de datos de texto de manera eficiente.

2.2.5 TENSORFLOW Y PYTORCH

TensorFlow¹¹ y PyTorch¹² son bibliotecas populares de aprendizaje automático y aprendizaje profundo que se utilizan en la creación y entrenamiento de modelos de análisis de sentimientos avanzados. Estas bibliotecas ofrecen herramientas para construir redes neuronales profundas y aplicar técnicas de procesamiento de lenguaje natural de vanguardia.

Estas son solo algunas de las herramientas y plataformas disponibles en el campo de la minería de textos y el análisis de sentimientos. La elección de la herramienta adecuada dependerá de los requisitos específicos del proyecto, el conjunto de datos y las capacidades deseadas. La combinación de estas herramientas con enfoques adecuados de procesamiento de texto y análisis de sentimientos puede conducir a resultados significativos en una amplia variedad de aplicaciones.

¹⁰ Watson Natural Language Understanding: <https://www.ibm.com/products/natural-language-understanding>

¹¹ TensorFlow <https://www.tensorflow.org/?hl=es-419>

¹² PyTorch: <https://pytorch.org/>

2.3 TENDENCIAS ACTUALES EN MINERÍA DE TEXTOS

La minería de textos y el análisis de sentimientos son campos dinámicos que evolucionan constantemente. Aquí se presentan algunas de las tendencias actuales más destacadas en estas áreas:

2.3.1 MODELOS DE LENGUAJE PRE-ENTRENADOS

Una de las tendencias más significativas en la minería de textos es la adopción de modelos de lenguaje pre-entrenados. Modelos como BERT (Bidirectional Encoder Representations from Transformers) y GPT (Generative Pre-trained Transformer) han demostrado un rendimiento sobresaliente en una variedad de tareas de procesamiento de lenguaje natural, incluido el análisis de sentimientos. (Zhu, Q., & Luo, J. 2022)

Estos modelos pre-entrenados capturan el conocimiento lingüístico de grandes cantidades de texto y pueden ser afinados para tareas específicas, lo que ha revolucionado la capacidad de comprender y analizar el lenguaje humano en profundidad.

Este avance supone una revolución en el campo del procesamiento de lenguaje natural porque capturan patrones lingüísticos complejos y conocimiento general del lenguaje que se puede aplicar a diversas tareas, como análisis de sentimientos, traducción automática, resumen de texto, entre otros. Aquí hay más información sobre los modelos de lenguaje pre-entrenados:

- Entrenamiento previo: Estos modelos se entrenan primero en grandes cantidades de texto no etiquetado, como corpus de texto web, libros, artículos y más. Durante el entrenamiento, los modelos aprenden a predecir la siguiente palabra en una oración o a llenar espacios en blanco en un fragmento de texto. Este proceso les permite adquirir un conocimiento profundo de la gramática, el contexto y la semántica del lenguaje.
- Transformers: La mayoría de los modelos de lenguaje pre-entrenados se basan en arquitecturas de redes neuronales llamadas "Transformers". Los Transformers

son capaces de capturar relaciones a largo plazo en el texto y han demostrado ser especialmente efectivos en tareas de procesamiento de lenguaje natural.

- **Afinamiento para tareas específicas:** Después del entrenamiento previo en datos masivos, estos modelos se pueden afinar o ajustar para tareas específicas. Por ejemplo, en el análisis de sentimientos, el modelo pre-entrenado se ajustaría para clasificar texto en categorías como positivo, negativo o neutral.
- **Transferencia de conocimiento:** La ventaja clave de estos modelos es su capacidad para transferir conocimientos generales del lenguaje a tareas específicas con relativamente poca información de entrenamiento. Esto los hace eficientes y efectivos para una variedad de aplicaciones.

Algunos de los modelos de lenguaje pre-entrenados más conocidos incluyen BERT (Bidirectional Encoder Representations from Transformers), GPT-3 (Generative Pre-trained Transformer 3) y RoBERTa ¹³(A Robustly Optimized BERT Pretraining Approach). Estos modelos han demostrado un rendimiento excepcional en una amplia variedad de tareas de procesamiento de lenguaje natural. (Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. 2019)

En el contexto del proyecto, los modelos pre-entrenados pueden ser particularmente útiles para el análisis de sentimientos y emociones en los mensajes de WhatsApp. Se pueden aprovechar modelos como **beto-sentiment-analysis**¹⁴ y **beto-emotion-analysis** ¹⁵para realizar estas tareas de manera efectiva, ya que estos modelos ya tienen conocimiento sobre cómo interpretar el lenguaje y los matices emocionales en él.

¹³ RoBERTa: <https://huggingface.co/roberta-base>

¹⁴ beto-sentiment-analysis: <https://huggingface.co/finiteautomata/beto-sentiment-analysis>

¹⁵ beto-emotion-analysis: <https://huggingface.co/finiteautomata/beto-emotion-analysis>

2.3.2 APRENDIZAJE AUTOMÁTICO EXPLICATIVO

Los enfoques de aprendizaje automático explicativo, como LIME ¹⁶ (Local Interpretable Model-agnostic Explanations) y SHAP¹⁷ (SHapley Additive exPlanations), son métodos que ayudan a explicar cómo un modelo llega a una decisión, lo que es esencial en aplicaciones críticas como la toma de decisiones médicas o legales.

2.3.3 MULTILINGÜISMO Y DETECCIÓN DE SENTIMIENTOS EN MÚLTIPLES IDIOMAS

Con la creciente globalización, la detección de sentimientos en múltiples idiomas se ha convertido en una tendencia importante. Los investigadores y las empresas están desarrollando modelos y técnicas que pueden aplicarse a textos en una variedad de idiomas, lo que facilita el análisis de opiniones y emociones a nivel internacional.

2.3.4 ANÁLISIS DE EMOCIONES MÁS COMPLEJAS

A medida que el análisis de sentimientos madura, las investigaciones se centran en la detección de emociones más complejas que simplemente positivas, negativas o neutras. Esto incluye la identificación de emociones como la tristeza, la felicidad, el miedo y la sorpresa. Los modelos están siendo entrenados para comprender y etiquetar estas emociones en textos de manera más precisa.

2.3.5 ÉNFASIS EN LA ÉTICA Y LA PRIVACIDAD

A medida que la minería de textos se utiliza en una variedad de aplicaciones, la ética y la privacidad se han convertido en preocupaciones críticas. Los investigadores están trabajando en el desarrollo de técnicas que minimicen el sesgo en los resultados y protejan la privacidad de los usuarios al tiempo que se realizan análisis de sentimientos y minería de textos.

¹⁶ LIME: <https://homes.cs.washington.edu/~marcotcr/blog/lime/>

¹⁷ SHAP: <https://shap.readthedocs.io/en/latest/>

2.3.6 APLICACIONES EN TIEMPO REAL Y A ESCALA

La capacidad de realizar análisis de sentimientos en tiempo real y a gran escala es esencial en aplicaciones como las redes sociales y la atención al cliente. Las soluciones de procesamiento de lenguaje natural en tiempo real se están volviendo más accesibles y eficientes, lo que permite a las organizaciones tomar decisiones basadas en datos en tiempo real y gestionar la interacción con los clientes de manera más efectiva.

Estas tendencias reflejan el crecimiento continuo y la importancia de la minería de textos y el análisis de sentimientos en una amplia variedad de industrias y aplicaciones. A medida que la tecnología y la investigación avanzan, se espera que estas tendencias sigan dando forma al campo y generando oportunidades para la innovación y el desarrollo.

2.4 FUTURAS DIRECCIONES DE INVESTIGACIÓN

El campo de la minería de textos y el análisis de sentimientos está en constante evolución y presenta muchas oportunidades emocionantes para la investigación futura. A medida que se exploran nuevas fronteras y se enfrentan a desafíos emergentes, se vislumbran varias direcciones prometedoras:

2.4.1 INTERPRETACIÓN DE EMOCIONES COMPLEJAS

Una de las áreas de investigación más emocionantes es la interpretación de emociones complejas en el texto. Si bien los enfoques actuales se centran en la detección de emociones básicas como la felicidad o la tristeza, el futuro podría traer avances en la identificación y comprensión de emociones más sutiles y complejas. Esto podría incluir emociones mixtas o estados emocionales cambiantes en el texto.

2.4.2 DETECCIÓN DE SENTIMIENTOS EN CONTEXTOS ESPECÍFICOS

La detección de sentimientos en contextos específicos es una dirección de investigación prometedora. Esto implica entrenar modelos para comprender mejor los matices del

lenguaje en entornos particulares, como el análisis de sentimientos en reseñas de películas, comentarios de productos o conversaciones médicas. Los modelos personalizados podrían adaptarse a las necesidades de industrias específicas.

2.4.3 ANÁLISIS MULTIMODAL

La combinación de texto con otros tipos de datos, como imágenes o audio, abre nuevas posibilidades. La investigación futura podría centrarse en la creación de modelos que analicen y comprendan el contexto en múltiples modalidades, lo que permitiría un análisis de sentimientos más enriquecido y preciso.

2.4.4 MINERÍA DE TEXTOS EN LENGUAS MENOS COMUNES

A medida que la tecnología se expande a nivel global, la minería de textos en idiomas menos comunes y regionales se vuelve esencial. Investigaciones futuras podrían enfocarse en el desarrollo de modelos y recursos para idiomas menos representados, lo que ampliaría la aplicabilidad de estas técnicas en un contexto global.

2.4.5 ÉNFASIS EN LA ÉTICA Y LA PRIVACIDAD

La ética y la privacidad seguirán siendo áreas críticas de investigación en la minería de textos y el análisis de sentimientos. Los investigadores podrían trabajar en el desarrollo de métodos y técnicas que garanticen un análisis de sentimientos más ético y respetuoso de la privacidad, al tiempo que se mantienen altos estándares de calidad y precisión.

2.4.6 APLICACIONES EN DOMINIOS ESPECÍFICOS

El análisis de sentimientos y la minería de textos encontrarán una amplia gama de aplicaciones en industrias específicas, como la atención médica, la educación y la política. Investigar cómo adaptar estas técnicas a dominios particulares será fundamental para impulsar la adopción y la utilidad en sectores diversos.

2.4.7 DESARROLLO DE FRAMEWORKS DE CÓDIGO ABIERTO

El desarrollo de frameworks de código abierto para la minería de textos y el análisis de sentimientos permitirá a la comunidad de investigadores y profesionales colaborar y avanzar más rápidamente en el campo. La creación de herramientas y recursos accesibles para todos será una dirección importante de investigación.

Estas futuras direcciones de investigación prometen ampliar aún más la comprensión y el alcance de la minería de textos y el análisis de sentimientos. La colaboración entre investigadores, empresas y académicos será fundamental para abordar estos desafíos y aprovechar las oportunidades que se presentan en el campo en constante evolución.

3. PROPUESTA

Tras comprender qué es el análisis de datos, su importancia, las fases que debe seguir un proyecto y conocer las técnicas de minería de textos y análisis de sentimiento más populares, se procede a desarrollar un proyecto desde cero. En este capítulo se detallará el escenario, los objetivos y el alcance del proyecto para brindar una comprensión completa de su propósito y valor.

3.1 ESCENARIO

WhatsApp, como una de las aplicaciones de mensajería instantánea más utilizadas en todo el mundo, ofrece un escenario rico en datos para la exploración y el análisis. Con más de 2,000 millones de usuarios en más de 180 países y más de 1,000 millones de usuarios activos diarios. Es una fuente inmensa de información generada por usuarios. Con un promedio de 100,000 millones de mensajes enviados diariamente, la plataforma alberga una cantidad significativa de datos valiosos. Esta abundancia de datos es especialmente relevante en el contexto de la minería de textos y el análisis de sentimientos. (Kenton P. et al., 2014)

El 53% de los usuarios de WhatsApp en Estados Unidos utilizan la aplicación al menos una vez al día. Los usuarios pasan un promedio de 28 minutos diarios en la aplicación, lo que supone casi un 2% del día. Un usuario medio abre la aplicación entre 23 y 25 veces al día.

La característica de los chats grupales, que permite la participación de hasta 256 usuarios simultáneamente, genera un flujo constante de datos en cada conversación grupal activa. Estos datos son principalmente en forma de mensajes de texto, y su análisis manual es impracticable debido al volumen. Por lo tanto, se requiere la implementación de técnicas de minería de texto avanzadas para automatizar la extracción de información valiosa de estas conversaciones.

3.2 OBJETIVOS

El objetivo principal de este proyecto es desarrollar un sistema de análisis de mensajes de WhatsApp que permita a los usuarios extraer información valiosa y detectar patrones en sus conversaciones exportadas. Los objetivos específicos incluyen:

- Implementación de procesos de minería de texto: Desarrollar algoritmos y procesos de minería de texto que puedan procesar los mensajes de WhatsApp de manera eficiente y efectiva.
- Interpretación de mensajes: Crear un sistema que pueda interpretar los mensajes y comprender su contexto y significado, incluida la detección de emociones y sentimientos expresados en el texto.
- Transformación de datos en conocimiento: Transformar los datos brutos de los mensajes en información estructurada y útil para el usuario.
- Representación gráfica de la información: Presentar la información de manera visual a través de gráficas y visualizaciones que permitan a los usuarios comprender y analizar mejor los patrones en sus conversaciones.
- Facilitar la obtención de conocimiento: Facilitar que los usuarios obtengan conocimiento a partir de la información proporcionada, lo que les permitirá tomar decisiones informadas y mejorar su comprensión de las conversaciones de WhatsApp.

Este proyecto busca cerrar la brecha entre los datos generados por las conversaciones de WhatsApp y el conocimiento que se puede extraer de ellos. Al automatizar el análisis de textos y proporcionar una representación gráfica de la información, se espera que los usuarios puedan aprovechar al máximo los datos de sus conversaciones para una variedad de aplicaciones, desde mejorar la comunicación hasta tomar decisiones más informadas en contextos personales o profesionales.

El usuario será capaz de interpretar y detectar los patrones visualizando una serie de gráficas. De esta forma se está realizando la transformación de datos a información y de información a conocimiento.

Los datos harían referencia al fichero exportado que genera WhatsApp, el sistema toma como entrada ese fichero y realiza una serie de procesos de minería de datos que dan lugar a la transformación de datos en información. Finalmente, el sistema representa esa información gráficamente para que el usuario la interprete y pueda obtener conocimiento a partir de ella.

3.3 ALCANCE

Es esencial definir claramente el alcance del proyecto para garantizar una ejecución eficiente y exitosa. Aunque el objetivo principal es la detección de patrones en conversaciones de WhatsApp, el alcance se establecerá de manera precisa para delimitar el enfoque del proyecto:

- Tipo de datos analizados: El proyecto se centrará exclusivamente en el análisis de mensajes de texto. Aunque WhatsApp permite el intercambio de mensajes de voz, imágenes, videos, reenvío de mensajes, listas de difusión y llamadas, estos elementos no se incluirán en el alcance del proyecto en esta etapa inicial.
- Métricas de actividad: Se desarrollarán algoritmos y procesos para medir métricas de actividad en las conversaciones de WhatsApp. Esto incluirá la cantidad de mensajes enviados y recibidos, la frecuencia de actividad en el chat y la distribución temporal de los mensajes.
- Temas principales: El sistema se centrará en identificar los temas principales de las conversaciones. Esto implica la extracción de palabras clave y la clasificación de los mensajes en categorías o temas relevantes.
- Análisis de Sentimientos/Emociones: Se implementarán técnicas de análisis de sentimientos para detectar las emociones y sentimientos expresados en los mensajes de texto. Esto permitirá etiquetar los mensajes como positivos, negativos o neutros, así como identificar emociones específicas como la felicidad, la tristeza, la ira, etc.
- Conversaciones Grupales: El proyecto incluirá el análisis de conversaciones grupales de WhatsApp. Esto significa que el sistema será capaz de procesar y analizar interacciones en chats que involucran a múltiples usuarios.

El alcance definido se ajusta al enfoque inicial del proyecto y permitirá el desarrollo de un sistema de análisis de mensajes de WhatsApp que pueda cumplir con los objetivos establecidos. En futuras etapas, existe la posibilidad de expandir el alcance para incluir otros tipos de contenido, como imágenes y mensajes de voz, así como mejorar y afinar aún más las capacidades de análisis.

3.4 METODOLOGÍA

Para alcanzar los objetivos propuestos se ha establecido una metodología que ayudará a garantizar el éxito del proyecto. Será comentada a continuación.

3.4.1 PLAN DE TRABAJO

Disponer de un plan de trabajo bien definido es clave de cara a afrontar un proyecto. En él se detallan las fases del desarrollo, los tiempos estimados de dedicación a cada tarea y el orden en el que deben realizarse.

A efectos personales, considero que disponer de una distribución temporal de las tareas pendientes es esencial. No tanto por ir cumpliendo los plazos establecidos sino por ser consciente de cuándo se están produciendo desviaciones respecto al plan inicial. Me parece una reflexión interesante el hecho de necesitar una herramienta de análisis de trabajo para realizar un trabajo de análisis. Me reafirmo más aún en lo necesario que es analizar los datos para comprender la realidad que rodea a una situación cualquiera.

Desde el inicio del proyecto se fue consciente de esta necesidad y se propuso un plan de trabajo inicial. En él se fijó como fecha límite para finalizar el desarrollo el 10 de septiembre del 2023. Teniendo en cuenta esta premisa y los objetivos del desarrollo, se fijaron una serie de hitos y se configuró el siguiente plan de trabajo, que abarca desde el 1 de julio hasta la fecha de fin:

- **Semana 1-2.** Definición de requisitos y alcance: Identificar los requisitos clave del sistema de análisis de mensajes de WhatsApp. Establecer el alcance del proyecto y definir las funcionalidades principales.
- **Semana 3-4.** Investigación y recopilación de datos: Investigar las técnicas de minería de textos y análisis de sentimientos más adecuadas. Recopilar conjuntos de datos de conversaciones de WhatsApp para su análisis.
- **Semana 5-6-7.** Desarrollo de algoritmos: Desarrollar algoritmos para el análisis de mensajes de WhatsApp. Implementar la detección de métricas de actividad y temas principales.
- **Semana 8.** Implementación de análisis de sentimientos: Implementar técnicas de análisis de sentimientos en el sistema. Realizar pruebas y ajustes para mejorar la precisión del análisis de emociones.
- **Semana 9.** Diseño de interfaz de usuario: Diseñar la interfaz de usuario para la visualización de resultados. Integrar la representación gráfica de la información.
- **Semana 10.** Pruebas finales y evaluación: Realizar pruebas exhaustivas del sistema en diferentes conjuntos de datos. Evaluar el rendimiento del sistema y ajustar cualquier problema identificado.
- **Semana 11.** Revisión y entrega final: Finalizar la documentación del proyecto, incluyendo informes técnicos y manuales de usuario. Preparar la entrega final y presentar el proyecto según lo programado.

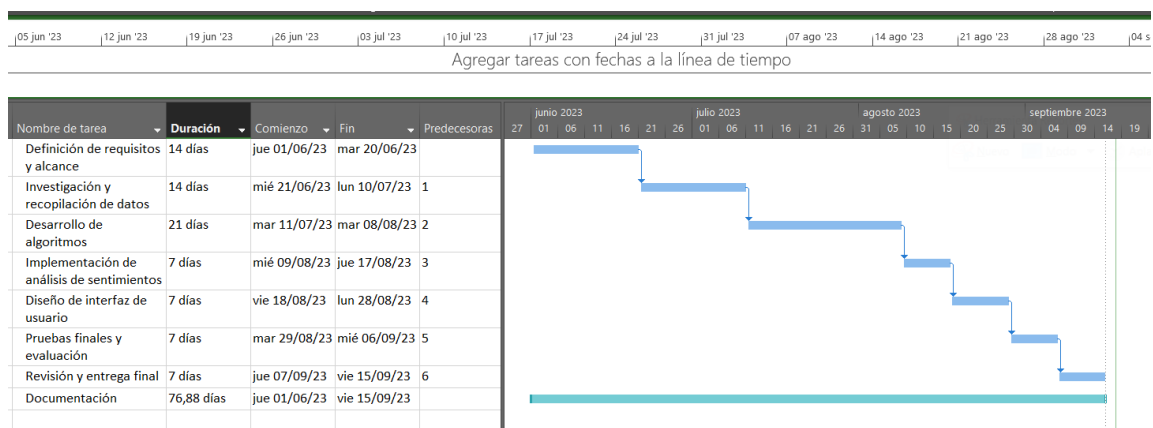


Ilustración 1. Plan de trabajo

La *Ilustración 1* representa gráficamente el plan de trabajo, el cual proporciona una estructura clara para el desarrollo del proyecto y permite un seguimiento efectivo del progreso. Además, establece fechas límite para cada fase y tarea, lo que facilita la detección temprana de desviaciones en caso de que se produzcan. La planificación es esencial para asegurar la finalización exitosa del proyecto y garantizar que se cumplan los objetivos establecidos.

Aparecen las fases del proyecto en el orden en el que se deben implementar cada una. Estas tareas no se pueden paralelizar puesto que son dependientes unas de otras.

3.4.2 REUNIONES DE SEGUIMIENTO

Otro factor clave a la hora de afrontar un proyecto es la comunicación. En este caso, se acordó desde el inicio realizar reuniones de con una frecuencia determinada. Estas servían para resolver dudas, mostrar avances y fijar siguientes pasos. Además de esta forma de comunicación también se contaba con otras más ágiles como el correo electrónico.

Este modelo de desarrollo encaja perfectamente con modelos ágiles de desarrollo software por su carácter iterativo e incremental. (Salvador, C., Nakasone, A. & Pow-Sang, J. A. 2014)

4. DESARROLLO DE LA PROPUESTA

En este capítulo, se aborda la propuesta enunciada en el capítulo 3 y se proporciona una síntesis de todo lo discutido hasta el momento.

4.1 SOLUCIÓN CONCEPTUAL

En primera instancia se expondrá la propuesta de solución o solución conceptual. En ella se introducirán los elementos que conformarán la arquitectura y la forma en que interactuarán entre ellos. La siguiente imagen resume el esquema propuesto:

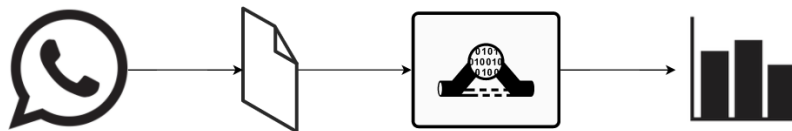


Figura 1. Esquema conceptual de la solución

La solución conceptual del proyecto se basa en una arquitectura sencilla pero efectiva que permite el análisis de conversaciones de WhatsApp y la generación de informes visuales. Se compone de dos componentes principales:

Generación del Fichero de Conversaciones: El usuario debe generar un archivo de exportación de la conversación de WhatsApp que desea analizar. Este archivo contendrá los mensajes de texto y servirá como entrada para el motor de análisis.

Motor de Análisis: El motor de análisis es la parte central del sistema. Toma el archivo de conversación como entrada y realiza un análisis completo de los mensajes. Utiliza técnicas de minería de texto y análisis de sentimientos para extraer información relevante. La salida de este motor consiste en informes y gráficas que representan los patrones, métricas de actividad y emociones detectadas en la conversación.

Es importante destacar que la introducción manual del archivo de conversación en el motor de análisis se eligió deliberadamente por motivos de seguridad y control. Aunque

existen alternativas como acceder a las conversaciones desde copias de seguridad en la nube, esta implementación garantiza que el usuario sea plenamente consciente de qué conversación y qué mensajes se están analizando en detalle.

Esta solución conceptual proporciona una base sólida para el desarrollo del proyecto y la implementación de las funcionalidades previamente definidas. Permite que el usuario obtenga información valiosa a partir de sus conversaciones de WhatsApp de manera controlada y eficiente.

4.2 TECNOLOGÍAS SELECCIONADAS

Para llevar a cabo este proyecto, se han seleccionado cuidadosamente varias tecnologías y herramientas que ofrecen la flexibilidad y la capacidad computacional necesarias para el análisis de conversaciones de WhatsApp.

4.2.1 ENTORNO DE DESARROLLO

Se ha elegido Google Colab¹⁸ como el entorno de desarrollo principal para este proyecto. Colab es una herramienta de Google que permite ejecutar notebooks en formato Jupyter Notebook en la nube, directamente desde Google Drive. Ofrece varias ventajas clave:

- Acceso gratuito a GPUs: Proporciona acceso gratuito a unidades de procesamiento gráfico (GPUs), lo que permite el procesamiento rápido y eficiente de grandes volúmenes de datos, esencial para la minería de textos y análisis de sentimientos.
- Facilidad de compartir: Permite compartir los notebooks de manera similar a otros documentos de Google Drive, lo que facilita la colaboración con otros miembros del equipo y la interacción con el tutor.
- Formato Markdown y narrativa: El soporte para formato markdown permite integrar código con celdas explicativas, lo que brinda una narrativa que mejora la interpretación del código.

¹⁸ Google Colab: <https://colab.google/>

- Sin necesidad de instalación: No se requiere instalación previa, lo que simplifica la configuración del entorno de desarrollo.

Estas características hacen que Google Colab sea una elección sólida para el desarrollo de este proyecto.

4.2.2 MODELOS PRE-ENTRENADOS

Se ha optado por utilizar modelos pre-entrenados. La elección esta estrategia se basa en que no se dispone de conjuntos de datos etiquetados y la mayor efectividad en el análisis de texto. El modelo seleccionado es BETO¹⁹.

BETO es un modelo de lenguaje pre-entrenado desarrollado específicamente para el procesamiento de lenguaje natural en español. El nombre "BETO" es una abreviatura de "BERT para Español" (Bidirectional Encoder Representations from Transformers).

La arquitectura de BETO se basa en la arquitectura de BERT la cual es conocida por su capacidad para capturar el contexto y las relaciones entre palabras en un texto de manera bidireccional, lo que lo hace efectivo para tareas de procesamiento de lenguaje natural como el análisis de sentimientos, la traducción automática y la generación de texto.

A diferencia de los modelos de lenguaje pre-entrenados más ampliamente conocidos, como el modelo original de BERT de Google, que se entrenó principalmente en inglés, BETO se entrenó específicamente en una gran cantidad de texto en español. Esto lo hace altamente competente en el procesamiento de lenguaje español y lo convierte en una opción poderosa para tareas de PLN en este idioma.

La principal ventaja de BETO es su capacidad para transferir conocimientos del lenguaje español a diversas tareas de procesamiento de lenguaje natural. Esto significa que, aunque se haya entrenado en una tarea genérica (por ejemplo, predecir la siguiente

¹⁹ BETO: <https://github.com/dccuchile/beto>

palabra en una oración), puede afinarse o ajustarse para tareas específicas en español, como el análisis de sentimientos en textos en español.

BETO se encuentra disponible públicamente para su uso en la comunidad de investigación y desarrollo en español. Esto significa que los desarrolladores y científicos pueden aprovechar este modelo para crear aplicaciones y soluciones en español relacionadas con el procesamiento de lenguaje natural.

Se puede acceder a modelos pre-entrenados de BETO a través de la biblioteca "Transformers" de Hugging Face, que proporciona una interfaz sencilla para cargar, ajustar y utilizar modelos de lenguaje pre-entrenados en una variedad de tareas de procesamiento de lenguaje natural.

En resumen, BETO es un modelo de lenguaje pre-entrenado altamente efectivo y específico para el procesamiento de lenguaje natural en español. Su capacidad para capturar el contexto y el significado en el idioma español lo convierte en una herramienta valiosa para tareas como el análisis de sentimientos y emociones en textos en español.

4.2.3 LIBRERÍA PYSENTIMIENTO

Para el análisis de sentimientos y emociones en las conversaciones de WhatsApp, se ha seleccionado la librería Pysentimiento²⁰. Esta librería ofrece modelos preentrenados basados en Transformers y se especializa en el análisis de sentimientos y emociones en el ámbito social. Pysentimiento es capaz de realizar análisis de sentimientos y emociones en español e inglés, lo que es esencial para este proyecto.

Los modelos preentrenados utilizados son beto-sentiment-analysis y beto-emotion-analysis, que son parte de la librería Pysentimiento.

²⁰ Pysentimiento: <https://pypi.org/project/pysentimiento/0.5.2rc4/>

Estas tecnologías y herramientas se han elegido cuidadosamente para garantizar un enfoque efectivo en el análisis de conversaciones de WhatsApp y la detección de patrones, sentimientos y emociones en el texto. Permiten un desarrollo eficiente y ofrecen las capacidades necesarias para abordar los objetivos del proyecto de manera exitosa.

4.3 DESARROLLO

En este apartado se detallará el desarrollo de las partes más relevantes del sistema de análisis, las cuales son la lectura de mensajes y la explotación de la información. Cada parte hace referencia a las fases de un proyecto de análisis de datos, en concreto a la fase de análisis y representación de la información.

4.3.1 LECTURA DE MENSAJES

En primer lugar, se aborda la parte fundamental del sistema de análisis de conversaciones de WhatsApp: la lectura y preparación de los mensajes para su análisis posterior. Esta fase sigue las buenas prácticas de procesamiento de datos para garantizar la calidad y relevancia de la información obtenida.

Los archivos de conversación de WhatsApp siguen una estructura de fecha y hora, nombre de usuario y mensaje. La comprensión de esta estructura es esencial para la extracción efectiva de datos.

Ejemplo:

```
d/m/y h:m - user : message /n
10/9/23 00:00 - Manuel Alberola: Este es un ejemplo de mensaje de WhatsApp
```

Al leerlo se obtiene un dataframe en el que cada fila representa un mensaje de la conversación, las columnas que contiene son: datetime, user y message. Una vez se ha leído la conversación se realizan una serie de transformaciones que limpian y estructuran el conjunto de datos:

1. Eliminación de mensajes multimedia: Los mensajes multimedia se tratan de manera especial debido a su naturaleza no textual. Estos mensajes se identifican mediante la etiqueta "*<Multimedia omitido>*" y se excluyen del análisis de sentimientos. Sin embargo, se consideran en las métricas de actividad de usuario para reflejar su participación en la conversación.

Ejemplo de mensaje multimedia:

```
14/9/23 00:00 - Manuel Alberola: <Multimedia omitido>
```

2. Manejo de mensajes eliminados. Cuando un usuario elimina un mensaje de la conversación éste aparece en el fichero de la siguiente forma:

```
10/9/23 00:00 - Manuel Alberola: Eliminaste este mensaje
```

```
10/9/23 00:00 - Otro user: Se eliminó este mensaje
```

En este caso, estos mensajes también se deben tratar de forma especial porque, además de formar parte de las tasas de actividad, también forman parte de otras métricas relacionadas con la tasa de mensaje eliminados.

3. Separación de Fecha y Hora: La columna "datetime" se divide en "date" y "time" para permitir el análisis de mensajes a nivel de fecha y hora. Esto facilita la identificación de patrones temporales en la conversación ya que se reduce la granularidad del dato.
4. Cálculo del flujo de mensajes: El flujo de mensajes se refiere a la dirección a la que está dirigido un mensaje. Para calcularlo, se utilizan las citas explícitas, es decir, cuando se usa el símbolo "@" para mencionar a alguien o la respuesta directa al mensaje. Esto permite rastrear la interacción entre usuarios y comprender quién está interactuando con quién en la conversación.
5. Limpieza de caracteres especiales: Los mensajes de WhatsApp pueden contener emojis, enlaces, números de teléfono y otros caracteres especiales que pueden no

ser relevantes para el análisis de sentimientos o minería de textos. Debe realizarse una limpieza adicional para eliminar estos caracteres y dejar solo el texto legible.

Estas transformaciones son cruciales para preparar los datos de manera adecuada antes de aplicar técnicas de minería de textos y análisis de sentimientos. La eliminación de mensajes multimedia y mensajes eliminados evita la inclusión de datos irrelevantes en el análisis de sentimientos, mientras que el cálculo del flujo de mensajes ofrece información valiosa sobre las interacciones entre los participantes de la conversación.

En resumen, la lectura y preparación de mensajes es una fase esencial en el proceso de análisis de conversaciones de WhatsApp, ya que asegura que los datos estén limpios, estructurados y listos para ser procesados por las técnicas de minería de textos y análisis de sentimientos. Esto sienta las bases para obtener información valiosa y patrones significativos de la conversación.

4.3.2 MINERÍA DE TEXTOS

Una vez leídos los mensajes se procede con la fase de explotación de la información en la que se emplean técnicas de minería de textos. En esta fase se analiza el sentimiento de cada mensaje, la emoción y se tokenizan y lematizan los mensajes.

Tokenizar un texto consiste en dividir el texto en las unidades que lo conforman, entendiendo por unidad el elemento más sencillo con significado propio para el análisis en cuestión, en este caso, las palabras. Este proceso será relevante para posteriormente detectar las temáticas de la conversación.

Lematizar un texto, por el contrario, es el proceso mediante el cual las palabras que pertenecen a un mismo paradigma flexivo o derivativo son llevadas a una forma normal que representa a toda la clase. Es decir, a partir de una forma flexionada (plural, femenino, conjugada ...) se obtiene la raíz de la palabra.

Previo al proceso de tokenización y lematización, y únicamente para este proceso, se debe realizar una limpieza de datos, que consiste principalmente en la eliminación de caracteres especiales y palabras de un solo carácter. La sustitución de varios espacios en blanco consecutivos por solo uno y la transformación de todo el texto en minúsculas.

Para el posterior diseño gráfico se ha decidido crear un segundo dataset y se han eliminado de él aquellas palabras que aportan poco valor al análisis, como determinantes, pronombres y demás. Estas palabras aparecen contenidas en un fichero denominado “stopwords”. Estas no han sido eliminadas del dataset sobre el que se evalúa el sentimiento puesto que los modelos preentrenados han sido entrenados con textos completos.

Por tanto, al dataframe creado anteriormente se le añaden las siguientes columnas:

- **Sentiment:** Valor categórico generado por el analizador de sentimiento.
 - o Puede ser: POS, NEU ó NEG
- **Sentiment_Value:** Valor numérico generado por el analizador de sentimiento
- **Emotion:** Valor categórico generado por el analizador de emociones.
 - o Puede ser: Others, Joy, Sadness, Anger, Surprise, Disgust, Fear
- **Emotion_Value:** Valor numérico generado por el analizador de emociones
- **Tokens:** Contenido del mensaje tokenizado
- **Stems:** Contenido del mensaje lematizado

A continuación, se muestra un ejemplo del dataframe resultante. Por motivos de privacidad los nombres de usuarios se han sustituido por user1 y user2:

date	username	message	Sentiment	Sentiment_Value	Emotion	Emotion_Value	Tokens	Stems
2020-04-06 11:36:00	user2	Daba problemas	NEG	0.9979901313781738	fear	0.5590141415596008	[daba, problemas]	[dab, problem]
2020-04-06 11:37:00	user2	Porque ya no me acuerdo	NEG	0.9814265370368958	sadness	0.8809124827384949	[acuerdo]	[acuerd]
2020-04-06 11:53:00	user2	Es que veo que me vas a querer hacer el lio	NEG	0.9990339279174805	anger	0.610905110836029	[veo, vas, querer, hacer, lio]	[veo, vas, quer, hac, lio]
2020-04-06 12:21:00	user2	Vaya tela	POS	0.9900845289230347	surprise	0.7363753318786621	[vaya, tela]	[vay, tel]
2020-04-06 12:35:00	user1	Pues chungo va a estar	POS	0.9899493455886841	surprise	0.6996089220046997	[pues, chungo, va]	[pues, chung, va]

Figura 2. Dataframe resultante de la lectura y explotación de la información

En este dataframe aparecen tres mensajes negativos y dos positivos pertenecientes a las emociones: fear, sadness, ager y surprise. Se observa que en las columnas tokens y stems aparecen las palabras filtradas.

Una vez se ha obtenido el dataframe final, el siguiente paso consiste en representar y visualizar la información de manera efectiva. Para esta tarea, se han aprovechado las capacidades de librerías como Plotly²¹, Matplotlib²² y Seaborn²³. Estas herramientas facilitan la creación de gráficos y visualizaciones a partir de los datos contenidos en listas, arrays o dataframes.

Esta fase de visualización desempeña un papel esencial en la conversión de datos en conocimiento. Sin embargo, en lugar de profundizar en ella en este momento, se abordará de manera más detallada en el próximo apartado, donde se vinculará directamente con los resultados de la experimentación.

²¹ Plotly: <https://plotly.com/>

²² Matplotlib: <https://matplotlib.org/>

²³ Seaborn: <https://seaborn.pydata.org/>

6. EXPERIMENTACIÓN

A continuación, se presenta un caso de estudio del sistema analítico desarrollado en este trabajo. Se comentarán los resultados obtenidos tras analizar un grupo de WhatsApp de amigos. Está formado por 6 integrantes que aparecerán de forma anónima en el conjunto de datos. La conversación cubre una franja temporal de un año y cuatro meses que van desde mayo del 2019 hasta septiembre del 2021.

Los usuarios que en él aparecen son:

```
['user1', 'user2', 'user3', 'user4', 'user5', 'user6']
```

El número de mensajes por usuario es:

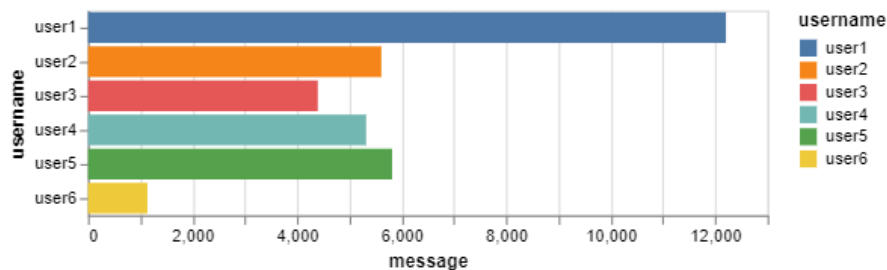


Figura 1. Número de mensajes por usuario

Se observa que el usuario más activo es user1, está bastante igualado entre los usuarios 2, 3, 4 y 5 y por último el usuario 6 que tiene una menor tasa de actividad en el grupo.

Del total de mensajes enviado por usuario, la proporción de multimedia (foto, video, audio, stickers, ...) es:

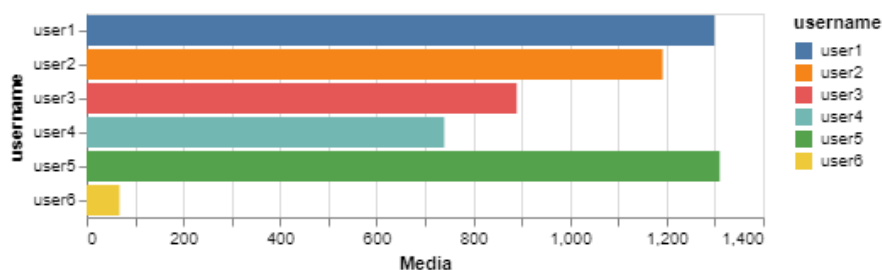


Figura 2. Número de mensajes multimedia por usuario

Al comparar el número total de mensajes con el número de mensajes multimedia se observa que user1 apenas envía un 10% de mensajes multimedia mientras que user2 y user5 alcanzan una tasa cercana al 20%.

Si se analizan los mensajes eliminados por usuario:

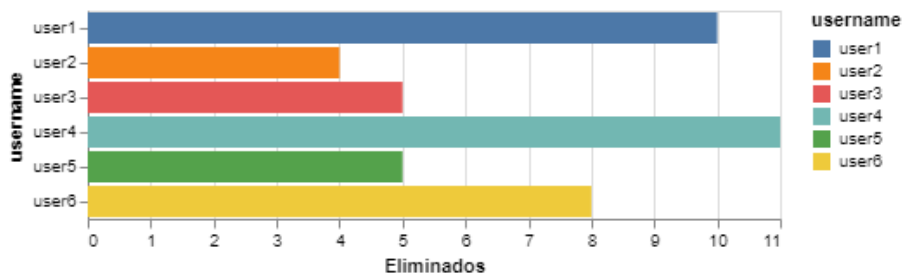


Figura 3. Número de mensajes eliminados por usuario

Se observa que el usuario que más mensajes elimina es user4, pero no supera una tasa del 0.18%.

La distribución de los mensajes en el tiempo es la siguiente:

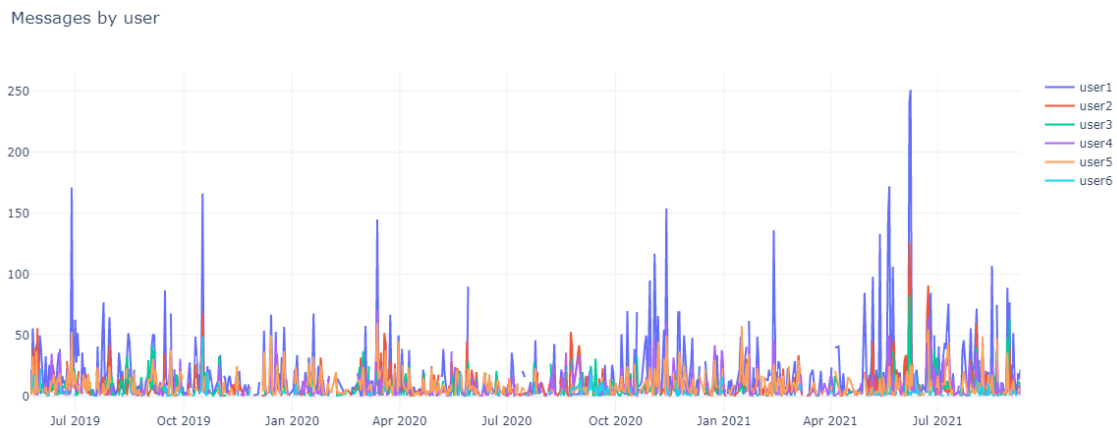


Figura 4. Número de mensajes por usuario y fecha

Se observan picos muy pronunciados que superan los 150 mensajes por día, llegando a alcanzar un máximo de 250 mensajes. Estos picos vienen acompañados por varios usuarios. Esto es evidente porque en las conversaciones siempre participan más de un usuario. Cada pico se corresponde con una festividad como cumpleaños o celebraciones, o periodos vacacionales donde el grupo presenta una mayor actividad social.

Al analizar estos picos en profundidad se pueden apreciar qué usuarios participaron más en esas conversaciones. El siguiente gráfico muestra como de los 10 primeros días de agosto, los días más activos fueron el 3 y el 8 debido a festividades locales, pero en cada día participaron unos usuarios distintos. Esto es debido a que el 8 de agosto es fin de semana y el 3 no por lo que la disponibilidad de cada usuario es distinta.

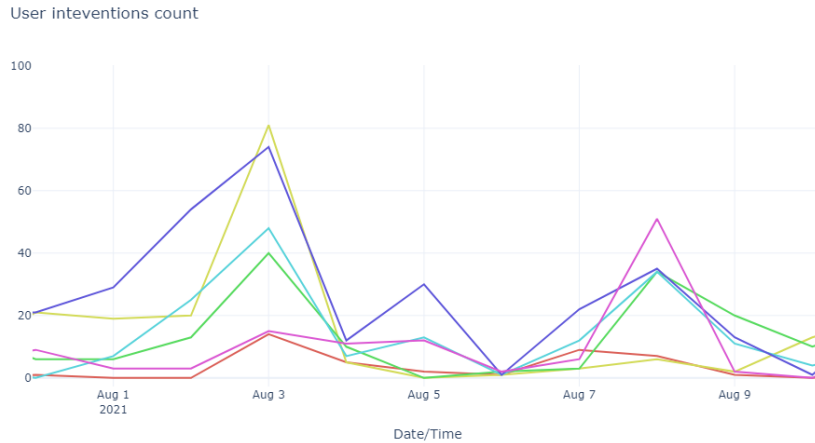


Figura 5. Número de mensajes por usuario y fecha en un intervalo de 10 días

Este mismo análisis se puede realizar por cada usuario, obteniendo su actividad:

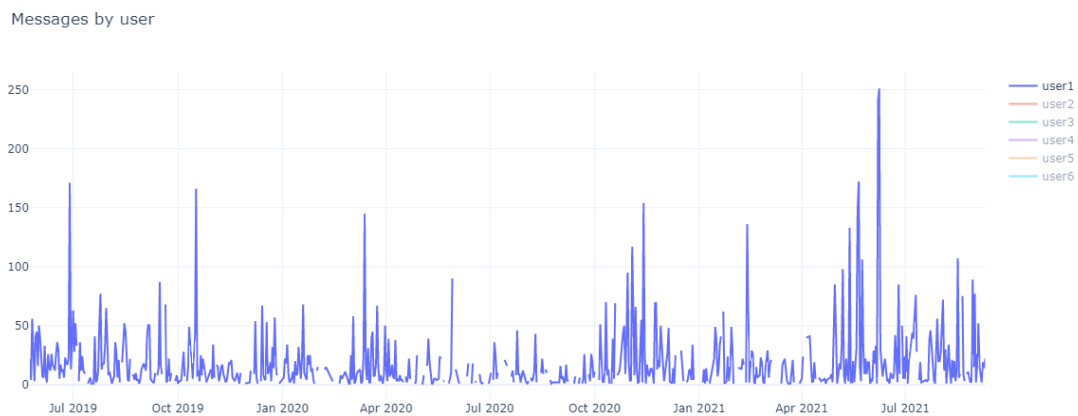


Figura 6. Número de mensajes de user1 por fecha

Se observa que la línea no es continua porque este usuario no ha enviado mensajes todos los días, pero sí se pueden detectar que la tendencia tiene la misma forma o similar a la del grupo, por lo que se puede confirmar que este usuario habría sido responsable, en parte o en gran medida de los días de mayor actividad.

Al obtener la tasa de mensajes acumulada por usuario se observa que desde el primer momento user1 fue quien más participación tuvo en el grupo y, además, durante los últimos meses ha aumentado su participación.

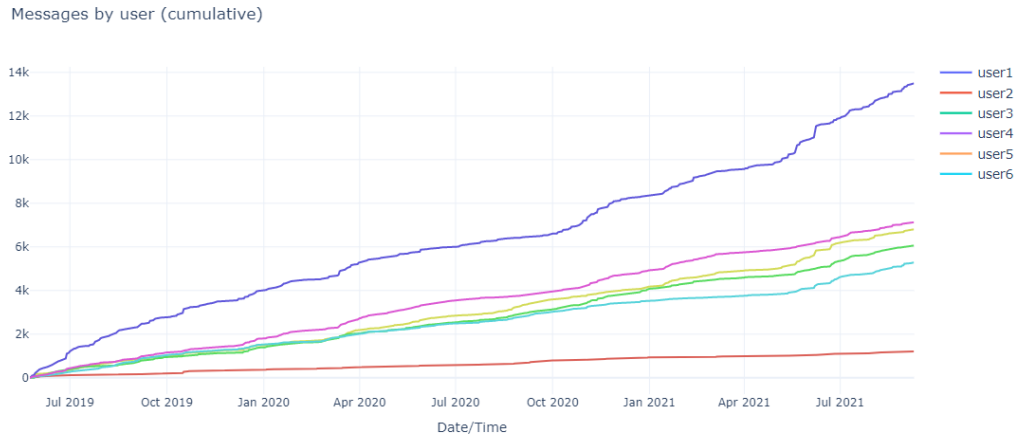


Figura 7. Número de mensajes por usuario acumulados

Si se analizan los mensajes por cantidad de palabras se observa que user1 es quien más mensajes envía pero que la mayoría de sus mensajes son de una sola palabra. Posiblemente debido a respuestas breves y cortas que corroboran su alta participación en el grupo:

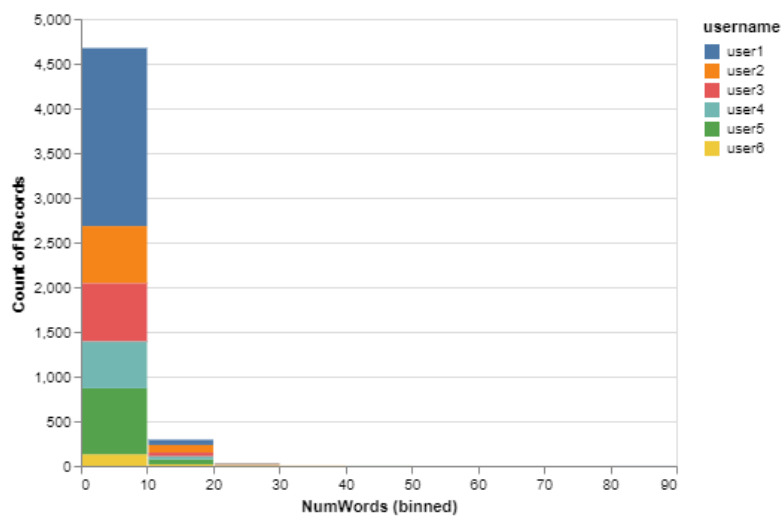


Figura 8. Número de palabras por mensaje y usuario

Si se analiza la longitud en caracteres de los mensajes enviados se observa que rara vez los mensajes superan los 250 caracteres, aunque alguna vez se han enviado mensajes superiores a 2000 caracteres tal y como muestra la dispersión:

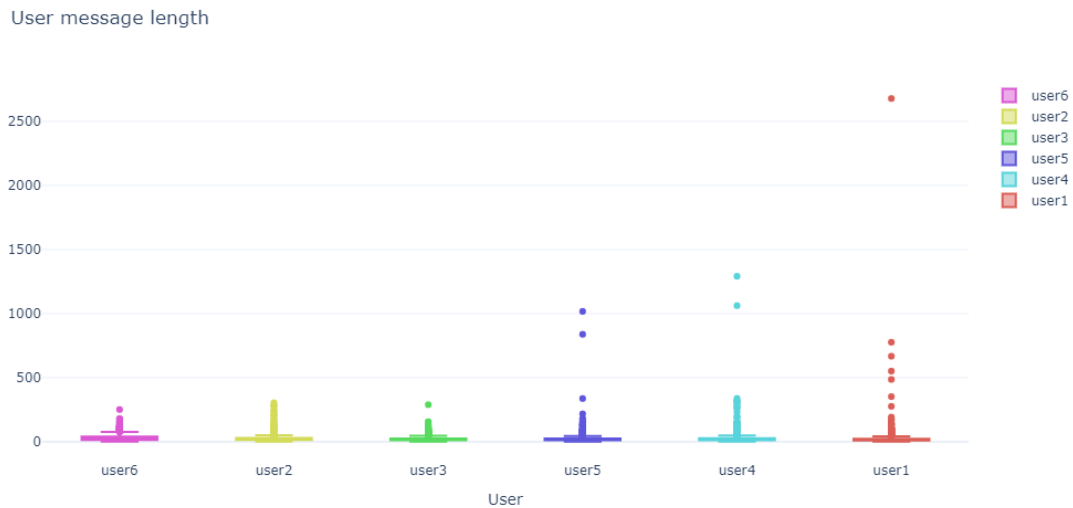


Figura 9. Longitud de los mensajes en caracteres por usuario

Observando la *Figura 9* se aprecia que los usuarios 5, 4 y 1 son más propensos a reenviar mensajes de mayor longitud.

El siguiente gráfico muestra cómo fluyen los mensajes dentro del grupo:

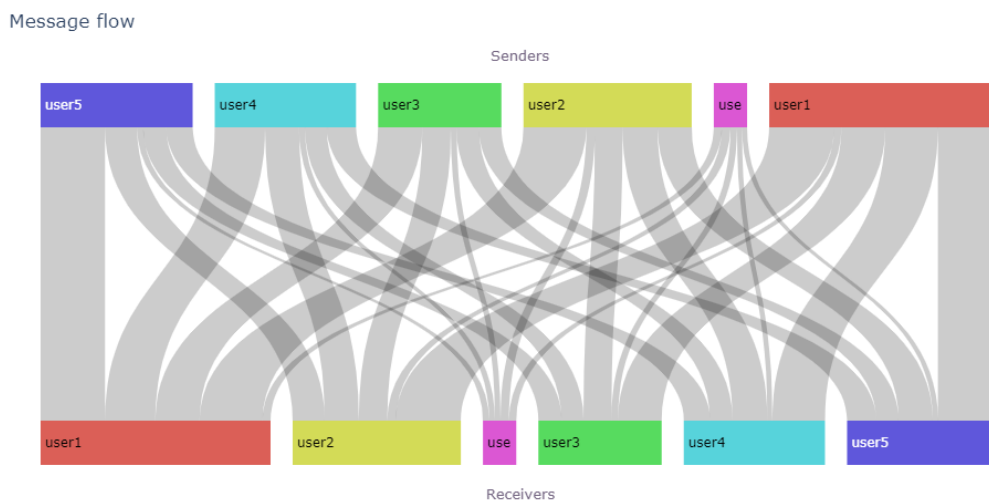


Figura 10. Flujo de los mensajes entre usuarios

Se puede observar que todos los usuarios interactúan mayoritariamente con user1, esto está relacionado con que es el usuario más activo del grupo, pero además puede denotar cierto liderazgo. También se puede inferir que la tasa de actividad está directamente relacionada a las interacciones que tienen contigo el resto de usuarios. Por ejemplo, user6 es el menos activo del grupo y, por tanto, además de ser quien menos mensajes envía también es quien menos recibe.

En este caso se observa que aparecen nuevos nombres como “Azahara” o “Galvañ” y nuevos planes como “Pádel” y “Campo”.

Si se analiza el sentimiento de cada mensaje por usuario se observa que la gran mayoría de mensajes son neutrales. Esta medida es proporcional al número de mensajes enviados. También se observa que la cantidad de mensajes positivos y negativos es similar predominando por muy poco los negativos.

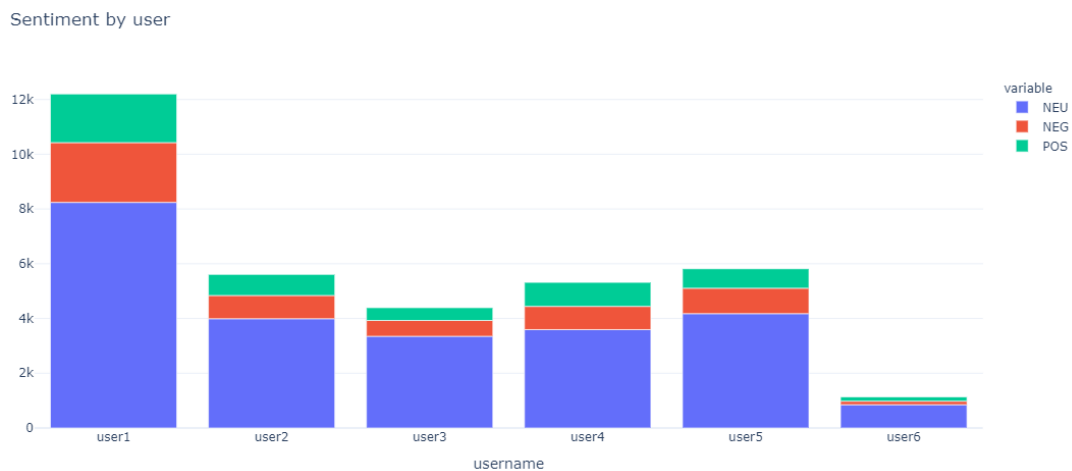


Figura 13. Sentimiento de los mensajes por usuario

Si se omiten del estudio los mensajes neutrales para obtener más detalle se observa que únicamente user4 y user6 tienen una mayor tasa de mensajes positivos que negativos:

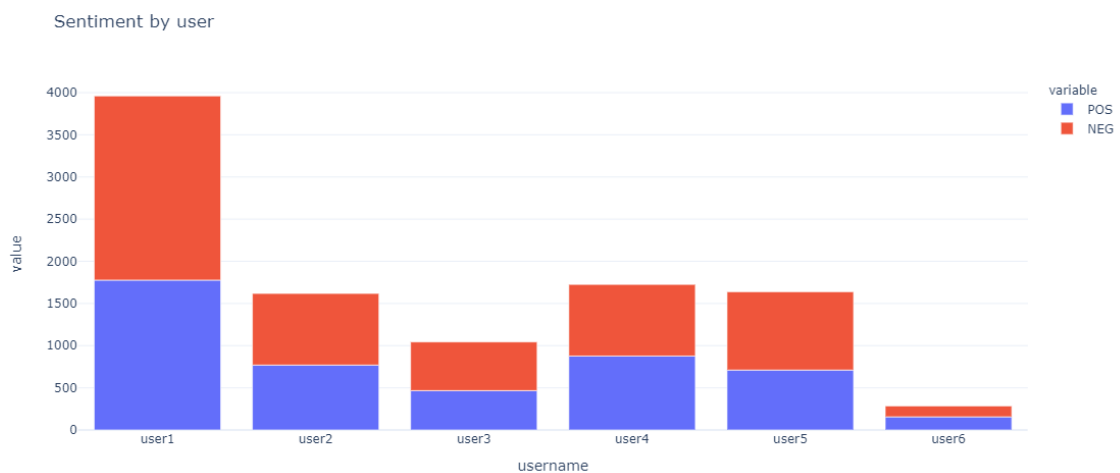


Figura 14. Sentimiento POS/NEG de los mensajes por usuario

Al medir la distribución de estos mensajes en el tiempo también se observa que están distribuidos equitativamente, lo cual indica que no se deben necesariamente a discusiones:

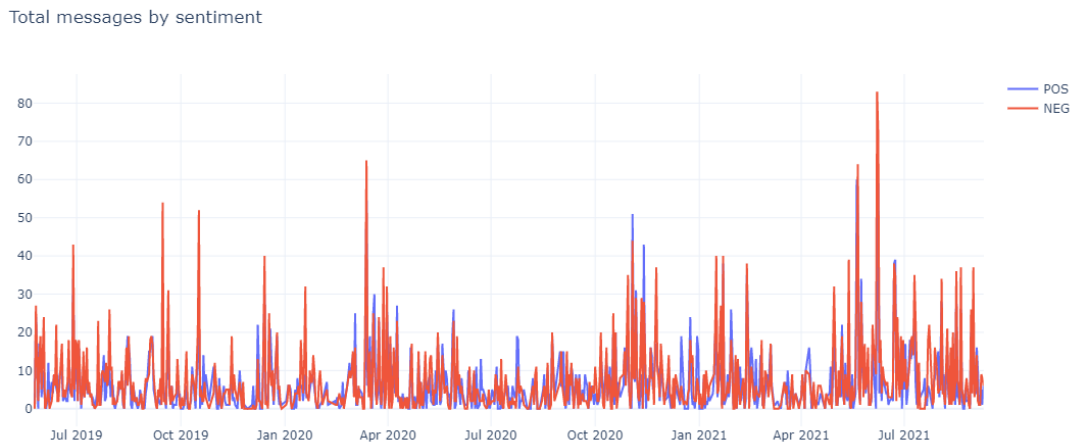


Figura 15. Sentimiento POS/NEG de los mensajes por usuario y fecha

Si se mide la correlación entre mensajes positivos y negativos se observa que prácticamente por cada mensaje positivo hay uno negativo.

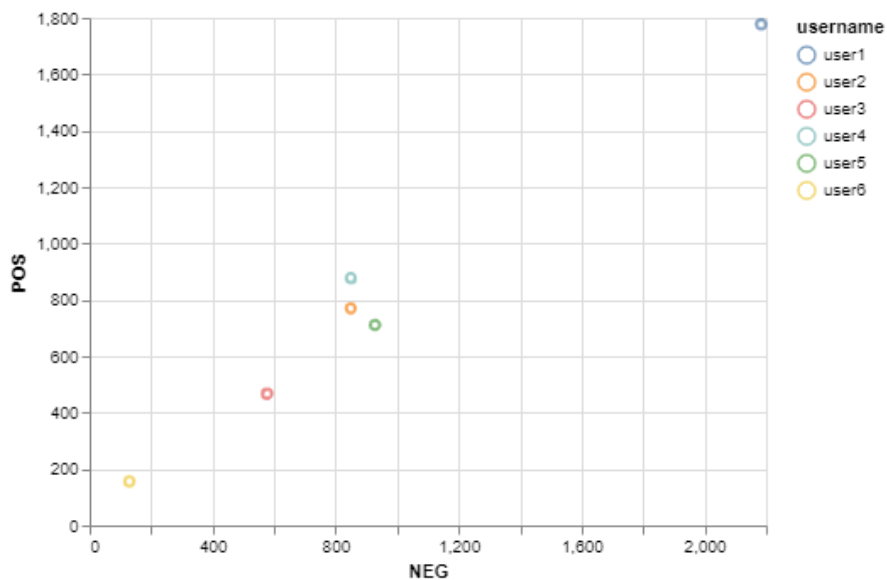


Figura 16. Correlación entre sentimiento POS y NEG de los mensajes por usuario

En cuanto a las emociones, se observa que oscilan entre la sorpresa y el enfado, lo cual encaja perfectamente con la temática y el contexto del grupo en cuestión:

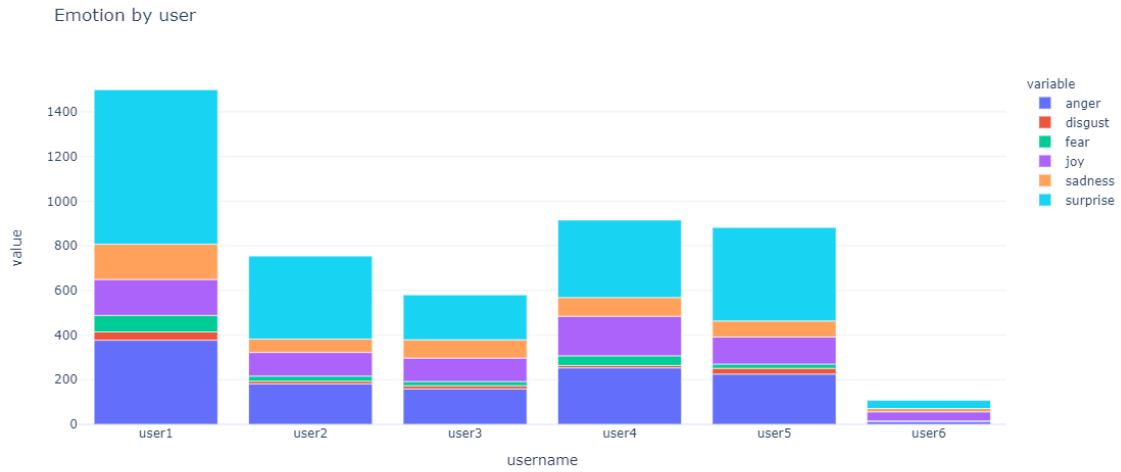


Figura 17. Emoción de los mensajes por usuario

Si se analizan las emociones en el tiempo se pueden detectar algunos patrones:

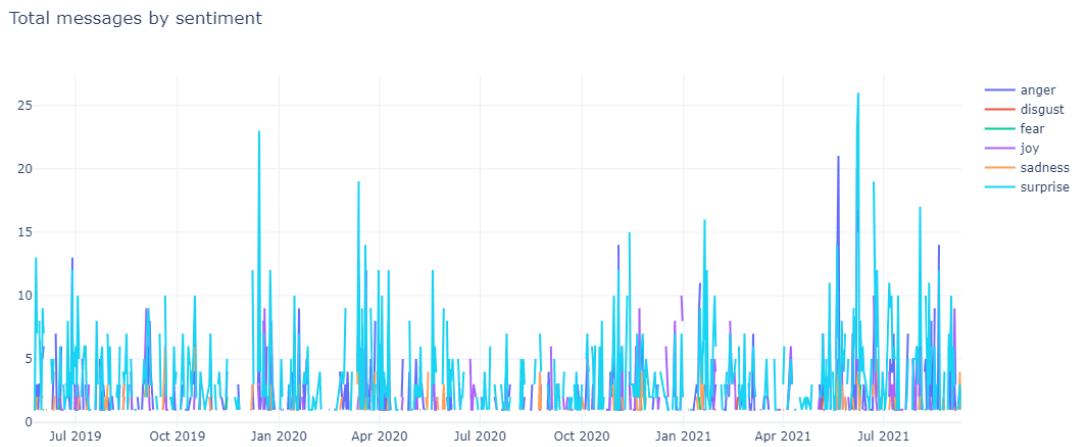


Figura 18. Emoción de los mensajes por usuario y fecha

Aunque para ello es necesario estudiar las emociones por separado:

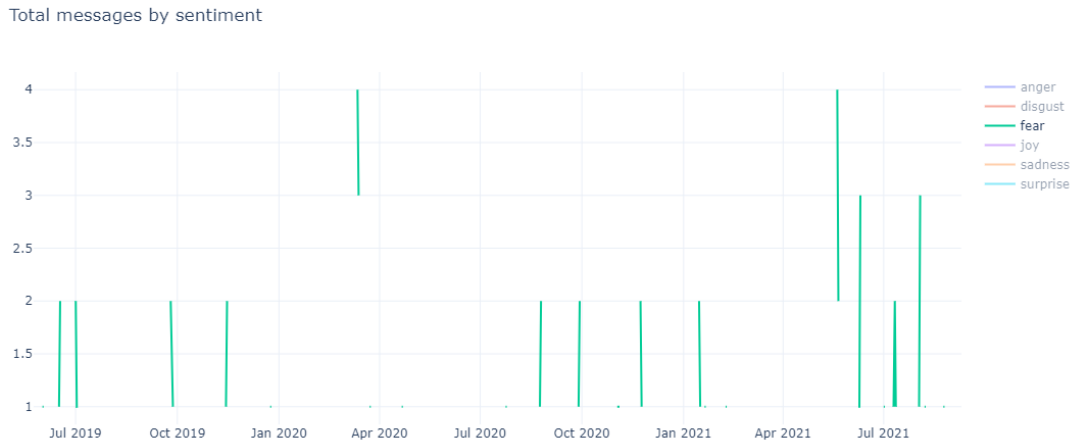


Figura 19. Mensajes de miedo por fecha

Por ejemplo, este gráfico demuestra que los mensajes de miedo aparecen puntualmente y por lo general en espacios temporales breves y cercanos, llegando incluso a poder detectarse tres clusters. Uno a finales del 2019, otro a finales del 2020 y otro a mediados del 2021 que coinciden con distintos sucesos que afectan a miembros del grupo como el fallecimiento de un familiar, la cancelación de un viaje grupal o el positivo en coronavirus de un miembro.

7. CONCLUSIONES Y LÍNEAS FUTURAS

La experimentación se ha realizado sobre un grupo de WhatsApp muy conocido por mí con el objetivo de poder detectar fácilmente los patrones que representan los gráficos generados. A continuación, se extraerán algunas conclusiones sobre la experimentación y se plantearán líneas futuras hacia donde podría avanzar el desarrollo realizado.

7.1 CONCLUSIONES

Sobre las tasas de participación de los usuarios, y conociendo bien a los integrantes de grupo, puedo confirmar que los resultados se ajustan perfectamente a la realidad. La “pandilla” está formada por 6 miembros de los cuales uno es bastante discreto y suele pasar desapercibido y otro miembro que se hace notar más que el resto. Las diferencias entre los otros cuatro miembros del grupo son menores.

Respecto a la tasa de participación analizada por fecha, ha sido complicado detectar algún patrón claro que determine las fechas en las que más se usa el grupo, pero precisamente ese es el patrón, y es que se trata de un grupo de amigos íntimos en el que se habla prácticamente a diario. De hecho, es la única forma de comunicación que tenemos además de la presencial, por lo que cuadra perfectamente que no haya grandes diferencias más allá de cosas puntuales como cumpleaños, celebraciones o acontecimientos especiales.

Al analizar la cantidad de palabras por mensaje se observó que el 95% de los mensajes tienen menos de 20 palabras. Esto se debe a dos motivos, el primero es intrínseco al uso de WhatsApp y es que muchos textos se suelen dividir en varios mensajes. Y el segundo motivo es que suelen ser conversaciones muy participativas en las que se realizan preguntas o propuestas que requieren una respuesta. Esta perspectiva tan contraria a un monólogo cuadra perfectamente con el carácter del grupo.

En cuanto al flujo de mensajes entre usuarios, también cuadran los resultados obtenidos con la realidad puesto que uno de los miembros del grupo, además de ser activo en el

grupo, también lo es socialmente. Quiere hacer planes constantemente por lo que hace muchas propuestas haciendo que aumente el flujo de mensajes de él a el resto y en consecuencia del resto a él.

En cuanto a la detección de temas, tanto grupales como individuales me ha sorprendido la cantidad de información que revela. Tras varias pruebas he sido capaz de reconocer a cada usuario a partir de su nube de palabras. Esto indica que, a pesar de ser un grupo, cada miembro tiene su individualidad y la pone en común con el resto.

Por último, sobre el análisis de sentimiento y emociones, considero que el resultado es bueno, pero no es completamente válido para el análisis porque hay muchos mensajes irónicos que han sido interpretados como negativos cuando en realidad no lo son. Lo mismo sucede con las emociones. Aún así, a grandes rasgos es cierto que los usuarios que menos palabras negativas dicen son los menos agresivos.

En definitiva, considero que este análisis ha sido muy interesante porque ha demostrado de forma empírica muchas cosas que ya conocía o intuía, pero en cualquier otro caso, con un grupo desconocido, además de interesante sería útil porque ayudaría a detectar muchos patrones que serían imposibles de detectar en grupos que no son tu grupo de confianza.

Por todo esto, y teniendo en cuenta todo lo anterior, considero que el resultado es altamente satisfactorio, se han alcanzado los objetivos propuestos y me muestro muy satisfecho con el desarrollo realizado.

7.2 LÍNEAS FUTURAS

Si bien los resultados obtenidos son satisfactorios todavía faltaría mucho por hacer, a continuación, se plantean distintas líneas futuras por las que se podría continuar el desarrollo.

Sería indispensable añadir el contenido multimedia, en estos mensajes (sobre todo en notas de voz) va muchísima información que ahora mismo no se está teniendo en cuenta. También sería necesario filtrar de alguna forma los mensajes irónicos puesto que distorsionan mucho los resultados de emoción y sentimiento.

Para la trazabilidad del flujo de mensajes se están empleando las citas explícitas en las conversaciones como “@user1 mensaje”. Sería interesante, además de usar estas citas emplear también los vocativos encontrados en el texto.

8. BIBLIOGRAFÍA Y REFERENCIAS

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2, 1--135. doi: 10.1561/1500000011

Cruz, F. L., Troyano, J. A., Pontes, B. & Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels.. *Expert Syst. Appl.*, 41, 5984-5994.

Devlin J, Chang M, Lee K, Toutanova K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. DOI:<https://doi.org/10.48550/arXiv.1810.04805>

H. A. Patrick, P. G. J, M. H. Sharief and U. Mukherje. (2023). Sentiment Analysis Perspective using Supervised Machine Learning Method. DOI: 10.1109/ICECCT56650.2023.10179807.

Karabelly, Jozef. Non-Supervised Sentiment Analysis. Brno, 2020. Bachelor's Thesis. Brno University of Technology, Faculty of Information Technology. 2020-07-10. DOI:<https://www.fit.vut.cz/study/thesis/22391/>

Kenton P. O'Hara, Michael Massimi, Richard Harper, Simon Rubens, and Jessica Morris. 2014. Everyday dwelling with WhatsApp. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14). Association for Computing Machinery, New York, NY, USA, 1131–1143. DOI:<https://doi.org/10.1145/2531602.2531679>

Lane H, Howard C & Hapke H. (2019). *Natural Language Processing in Action*. Manning

Lee, V.L., Gan, K.H., Tan, T., & Abdullah, R. (2019). Semi-supervised Learning for Sentiment Classification using Small Number of Labeled Data. *Procedia Computer Science*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. & Potts, C. (2011). Learning word vectors for sentiment analysis. , .

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*, 90(10), 60–128.

Mohammad, S.M. (2016). Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. *Emotion Measurement*, abs/2005.11882, 201–237.

Ossai, C. I., & Wickramasinghe, N. (2023). Sentiments prediction and thematic analysis for diabetes mobile apps using Embedded Deep Neural Networks and Latent Dirichlet Allocation. *Artificial intelligence in medicine*, 138, 102509. <https://doi.org/10.1016/j.artmed.2023.102509>

Ramos, S. (2017). ¿Qué es Business Intelligence? Disponible en: <https://blogvisionarios.com/e-learning/articulos-data/que-es-business-intelligence-introduccion-01/>

Salvador, C., Nakasone, A. & Pow-Sang, J. A. (2014). A systematic review of usability techniques in agile methodologies.. *EATIS* (p./pp. 17:1-17:6), : ACM. ISBN: 978-1-4503-2435-9

SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining (Baccianella et al., LREC 2010)

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37, 267-307.

Xuefeng Yang, Kezhi Mao, Xuefeng Yang, and Kezhi Mao. 2017. Task Independent Fine Tuning for Word Embeddings. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 25, 4 (April 2017), 885–894. DOI:<https://doi.org/10.1109/TASLP.2016.2644863>

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for twitter sentiment analysis.

Zhu, Q., & Luo, J. (2022). Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society*, 2, 1825-1834. doi:10.1017/pds.2022.185

9. LISTA DE ACRÓNIMOS Y ABREVIACIONES

BD	Big Data.
PLN	Procesamiento del Lenguaje Natural.
MT	Minería de Textos.
MD	Minería de Datos.
DL	Deep Learning.
PCA	Análisis de Componentes Principales.
RNN	Redes Neuronales Recurrentes
CNN	Redes Neuronales Convolucionales
BERT	Bidirectional Encoder Representations from Transformers
LDA	Latent Dirichlet Allocation
NLKT	Neural Language Toolkit