UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN



GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DISEÑO Y DESARROLLO DE UN SISTEMA PARA EL ANÁLISIS Y DETECCIÓN DE CAMPAÑAS DE DESINFORMACIÓN EN REDES SOCIALES

MARCOS PÉREZ RUIZ

2026

GRADO EN INGENIERIA DE TECNOLGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Título: DISEÑO Y DESARROLLO DE UN SISTEMA PARA EL ANÁLISIS Y DETECCIÓN DE CAMPAÑAS DE DESINFORMACIÓN EN REDES SOCIALES

Autor: D. Marcos Pérez Ruiz

Tutor: D. Sonia Solera Cotanilla

Departamento: Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL

Presidente: D. ……

Vocal: D. ……..

Secretario: D. ……..

Suplente: D. ……..

Los miembros del tribunal arriba nombrados acuerdan otorgar la calificación de: ……

Madrid, a de de 20…

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN



GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DISEÑO Y DESARROLLO DE UN SISTEMA PARA EL ANÁLISIS Y DETECCIÓN DE CAMPAÑAS DE DESINFORMACIÓN EN REDES SOCIALES

MARCOS PÉREZ RUIZ

2026

RESUMEN

En el presente Trabajo de Fin de Grado se propone una métrica para evaluar el riesgo de desinformación en publicaciones de redes sociales, así como un prototipo funcional que implementa dicha métrica para la red social X. El objetivo es proporcionar una forma homogénea de cuantificar la desinformación, considerando no solo la veracidad del contenido, sino también otros factores relacionados con el contexto de la publicación, el tono empleado y determinados patrones lingüísticos asociados a este fenómeno.

La creciente influencia de la desinformación en ámbitos como la política, la ciencia o la opinión pública pone de manifiesto la necesidad de desarrollar herramientas capaces de identificar contenidos potencialmente problemáticos de forma temprana. Con este propósito, se realiza una revisión del estado del arte en materia de desinformación, analizando las principales estrategias, métricas y herramientas utilizadas actualmente para combatirla, así como sus limitaciones.

A partir de este análisis se define una metodología de evaluación compuesta por distintas fases, cada una orientada al estudio de un aspecto concreto de la publicación. Esta metodología combina indicadores relacionados con la veracidad de las afirmaciones, el contexto de la cuenta emisora, el tono del mensaje y la presencia de patrones frecuentemente asociados a contenidos desinformadores. Mediante la integración de estos elementos se obtiene una métrica capaz de proporcionar una valoración más completa y generalizable que aquellas basadas exclusivamente en la verificación de contenidos.

La propuesta se concreta en una implementación en Python de la herramienta expuesta, utilizando técnicas de inteligencia artificial y procesamiento del lenguaje natural. El prototipo integra diversos modelos especializados para el análisis automático de publicaciones y permite obtener una puntuación de riesgo de desinformación de forma automatizada.

Finalmente, se valida el comportamiento del sistema mediante distintos casos de estudio basados en publicaciones reales de naturaleza periodística, satírica y desinformadora. Los resultados obtenidos prueban que el sistema es capaz de diferenciar entre estos tipos de contenido de forma coherente, identificando distintos niveles de riesgo asociados a cada uno de ellos. Asimismo, ponen de manifiesto que la incorporación de factores adicionales a la veracidad contribuye a distinguir con mayor precisión entre información falsa y desinformación. Estos resultados respaldan la viabilidad del enfoque propuesto y establecen una base sobre la que desarrollar futuras investigaciones y mejoras.

SUMMARY

This Bachelor's Thesis proposes a metric for assessing the risk of disinformation in social media posts, together with a functional prototype that implements this metric for the social network X. The objective is to provide a standardized method for quantifying disinformation by considering not only the veracity of the content, but also additional factors related to the context of the publication, the tone employed, and specific linguistic patterns associated with this phenomenon.

The growing influence of disinformation in areas such as politics, science, and public opinion highlights the need for tools capable of identifying potentially problematic content at an early stage. To address this need, a review of the state of the art in disinformation research is conducted, examining the principal strategies, metrics, and tools currently used to combat it, as well as their limitations.

Based on this analysis, an evaluation methodology composed of several phases is defined, each focused on the assessment of a specific aspect of a publication. This methodology combines indicators related to the veracity of claims, the context of the originating account, the tone of the message, and the presence of patterns frequently associated with disinformation. By integrating these elements, the proposed metric provides a more comprehensive and generalizable assessment than approaches based solely on content verification.

The proposed approach is implemented in Python, employing Artificial Intelligence and Natural Language Processing techniques. The prototype integrates several specialized models for the automated analysis of social media posts and enables the automatic generation of a disinformation risk score.

Finally, the system is validated through a series of case studies based on real journalistic, satirical, and disinformation-related publications. The obtained results demonstrate that the system is capable of consistently distinguishing between these categories of content and identifying different levels of associated risk. Furthermore, they show that incorporating factors beyond content veracity contributes to a more accurate distinction between false information and disinformation. These findings support the feasibility of the proposed approach and establish a foundation for future research and further improvements.

PALABRAS CLAVE

Análisis de sentimiento, Desinformación, Inteligencia artificial, Métricas de riesgo, Procesamiento del lenguaje natural

KEYWORDS

Artificial intelligence, Disinformation, Natural language processing, Risk metrics, Sentiment analysis

ÍNDICE DEL CONTENIDO

1. INTRODUCCIÓN Y OBJETIVOS 1

1.1. Introducción 1

1.2. Objetivos 1

1.3. Estructura de la memoria 2

2. ESTADO DEL ARTE 4

2.1. Fact-checking 4

2.1.1. Verificación manual 4

2.1.2. Verificación comunitaria 5

2.1.3. Verificación automática 5

2.2. Análisis de Emociones 6

2.3. Métricas existentes 7

3. ARQUITECTURA DEL SISTEMA 9

3.1. Propuesta de clasificación 12

3.2. Fase 0 - Evaluación del contexto 12

3.3. Fase 1 - Verificación de la publicación 14

3.4. Fase 2 - Análisis de tono 14

3.5. Fase 3 - Identificación de patrones 15

3.6. Pesos globales 16

4. DESARROLLO DEL SISTEMA 18

4.1. Preprocesado del mensaje 19

4.2. Evaluación de contexto 19

4.2.1. Foto de perfil 19

4.2.2. Nombre de usuario 20

4.2.3. Descripción 20

4.2.4. Amigos 20

4.2.5. Últimos 5 Posts 21

4.2.6. Hecho reciente 21

4.3. Verificación de la publicación 21

4.4. Análisis de tono 23

4.5. Análisis de patrones 23

4.6. Integración de los módulos. 25

4.7. Integración en Python 26

4.8. Limitaciones actuales y del sistema 28

5. RESULTADOS Y VALIDACIÓN 30
5.1. Metodología 30
5.2. CASO 1: Publicación Periodística 30
5.3. CASO 2: Publicación Desinformadora 33
5.4. CASO 3: Publicación Satírica 35
5.5. Comparativa de resultados 37
6. CONCLUSIONES Y LÍNEAS FUTURAS 39
6.1. Conclusiones 39
6.2. Líneas futuras 39
BIBLIOGRAFÍA 41
ANEXO A: ASPECTOS ÉTICOS, ECONÓMICOS, SOCIALES Y AMBIENTALES 45
A.1 Introducción 45
A.2 Descripción de impactos relevantes relacionados con el proyecto 45
A.3 Análisis detallado del impacto social 46
A.4 Conclusiones 47
ANEXO B: PRESUPUESTO ECONÓMICO 48

FIGURAS

Figura 3.1 Definición de desinformación para este trabajo 9
Figura 3.2 Propuesta de diseño compuesta de 4 fases 12
Figura 4.1 Flujo del modelo desarrollado 18
Figura 4.2 Pantalla de inicio 27
Figura 4.3 Pantalla de carga 27
Figura 4.4 Pantalla de resultados 28
Figura 5.1 Publicación periodística elegida 30
Figura 5.2 Resultados del análisis de la publicación periodística 31
Figura 5.3 Desglose de “context” y “patterns” 31
Figura 5.4 Publicación desinformadora elegida 33
Figura 5.5 Resultados del análisis de la publicación desinformadora 34
Figura 5.6 Desglose de “context” y “patterns” 34
Figura 5.7 Publicación satírica elegida 35
Figura 5.8 Resultados del análisis de la publicación satírica 36
Figura 5.9 Desglose de “context” y “patterns” 36
Figura 5.10 Gráfico comparativo de los resultados 37

TABLAS

Tabla 3.1 Mecanismos de manipulación del receptor 10
Tabla 3.2 Mecanismos de manipulación del contenido 11
Tabla 3.3 Pesos de los rasgos medidos en la fase 0 14
Tabla 3.4 Pesos de los rasgos medidos en la fase 3 16
Tabla 3.5 Pesos de las 4 fases 17
Tabla 4.1 Puntuaciones asignadas a los resultados de ClaimeAI 23
Tabla 4.2 Modelos elegidos para la implementación de la fase 3 24
Tabla 4.3 Estrategias de normalización de las salidas de los modelos 24
Tabla 5.1 Tabla comparativa de los resultados 37

1. INTRODUCCIÓN Y OBJETIVOS

1.1. INTRODUCCIÓN

La aparición de las redes sociales ha transformado profundamente la forma en la que la sociedad consume y comparte información. Plataformas como X, Facebook, Instagram o TikTok permiten la difusión inmediata de contenidos a millones de usuarios, facilitando el acceso a la publicación de información, pero también favoreciendo la propagación de noticias falsas y contenidos manipulativos [1]. La desinformación se ha convertido así en uno de los principales problemas asociados al ecosistema digital moderno, especialmente debido a su capacidad para influir sobre la opinión pública, aumentar la polarización social y generar desconfianza hacia instituciones y medios de comunicación [2].

El funcionamiento de los algoritmos de recomendación de las plataformas sociales contribuye además a amplificar este fenómeno. Estos sistemas suelen priorizar contenidos capaces de generar altos niveles de interacción, favoreciendo la difusión de publicaciones sensacionalistas o emocionalmente intensas [3]. Diversos estudios han demostrado que emociones como miedo, ira o indignación aumentan la probabilidad de compartición de contenidos, facilitando la propagación de campañas desinformativas [4]. En este contexto, la Inteligencia Artificial (IA) desempeña un papel dual. Por un lado, los avances recientes en modelos generativos han facilitado enormemente la creación automática de textos, imágenes y vídeos falsos, incluyendo tecnologías como los *deepfakes* [5]. Por otro lado, la IA también constituye una de las principales herramientas para combatir la desinformación, mediante sistemas automáticos basados en Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés), aprendizaje profundo y análisis emocional [6].

Sin embargo, la detección automática de desinformación continúa siendo un problema abierto. La dependencia del contexto, la complejidad del lenguaje humano y la presencia de fenómenos como ironía o sarcasmo dificultan el desarrollo de sistemas completamente fiables [7]. Además, muchas soluciones actuales se centran únicamente en clasificar contenidos como verdaderos o falsos, sin considerar otros factores relevantes como emocionalidad, toxicidad o polarización.

A partir de esta problemática surge el presente Trabajo Fin de Grado (TFG). Se propone investigar la aplicación de técnicas de IA y NLP al análisis de desinformación en redes sociales, así como diseñar un modelo teórico orientado a la generación de métricas relacionadas con riesgo de desinformación y desarrollar un prototipo funcional basado en dicho modelo. Para ello, se plantea una arquitectura modular compuesta por distintos modelos especializados capaces de analizar características lingüísticas, emocionales y contextuales de las publicaciones, permitiendo construir métricas híbridas orientadas a estimar el posible riesgo asociado a un contenido determinado.

1.2. OBJETIVOS

El objetivo principal de este TFG es diseñar y evaluar un sistema capaz de estimar el riesgo de desinformación asociado a publicaciones en redes sociales mediante el uso de técnicas de IA y NLP. Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos:

1. Investigar el estado actual de la desinformación y de las principales técnicas utilizadas para combatirla. Se realiza una revisión bibliográfica orientada al estudio de los métodos actuales de *fact-checking*, análisis emocional y detección automática de contenidos engañosos. Además,

se analizarán las principales limitaciones de las soluciones existentes y las tendencias actuales dentro del ámbito de investigación.

2. Analizar las métricas y enfoques existentes utilizados para medir fenómenos relacionados con la desinformación. Se estudiarán distintos métodos utilizados para cuantificar aspectos como veracidad, polarización o interacción de los usuarios, evaluando sus ventajas y limitaciones dentro del contexto de redes sociales.
3. Diseñar un modelo teórico orientado a la generación de métricas relacionadas con riesgo de desinformación. El modelo buscará combinar múltiples indicadores parciales obtenidos mediante técnicas de NLP, permitiendo construir una aproximación más contextual al problema de la desinformación. Asimismo, se justificará la selección de las variables y parámetros utilizados dentro de la propuesta.
4. Diseñar una arquitectura modular basada en distintos modelos especializados de análisis lingüístico y emocional. El sistema propuesto estará compuesto por diferentes módulos independientes encargados de analizar características concretas de las publicaciones, como emociones, toxicidad, falacias o patrones lingüísticos asociados a contenidos manipulativos.
5. Desarrollar un prototipo funcional basado en el modelo diseñado. Se implementará una solución capaz de procesar publicaciones simuladas o reales y generar automáticamente distintas métricas relacionadas con el posible riesgo de desinformación asociado al contenido analizado.
6. Evaluar el comportamiento del prototipo mediante un conjunto de casos de estudio representativos. Finalmente, se analizarán los resultados generados por el sistema sobre distintos ejemplos de publicaciones, evaluando la coherencia de las métricas obtenidas y las principales limitaciones observadas durante el desarrollo del trabajo.

1.3. ESTRUCTURA DE LA MEMORIA

La presente memoria se ha estructurado con el objetivo de guiar al lector desde el contexto general del problema de la desinformación hasta el diseño e implementación de la propuesta. Para ello, el documento avanza progresivamente desde la contextualización teórica del problema hasta la descripción técnica del sistema propuesto. Cada capítulo responde a un propósito específico y se encuentra relacionado con los objetivos planteados anteriormente.

- Capítulo 1 - Introducción. Presenta el contexto general del problema de la desinformación en redes sociales, el papel de la IA tanto en la generación como en la detección de contenidos manipulativos, los objetivos del trabajo y la organización general de la memoria.
- Capítulo 2 - Estado del arte. Revisa las principales líneas de investigación relacionadas con la detección de desinformación. En este capítulo se analizan las técnicas actuales de *fact-checking*, los métodos de análisis emocional aplicados a redes sociales y las métricas utilizadas para evaluar contenidos desinformativos y su propagación.
- Capítulo 3 - Arquitectura. Describe la arquitectura general del sistema propuesto, así como el modelo conceptual y matemático utilizado para combinar los distintos indicadores parciales generados por los módulos de análisis. Además, se justifican las decisiones de diseño adoptadas y el enfoque modular utilizado durante el desarrollo del sistema.
- Capítulo 4 - Desarrollo. Detalla la implementación práctica del prototipo desarrollado, incluyendo los modelos empleados, las herramientas utilizadas y el funcionamiento de los distintos módulos de análisis. También se presentan ejemplos de funcionamiento, pruebas realizadas y una evaluación general de las capacidades y limitaciones del sistema desarrollado.

- Capítulo 5 – Resultados. Recoge los resultados obtenidos durante la fase de validación del prototipo. Para ello, se aplica la métrica desarrollada sobre diferentes categorías de publicaciones reales y se comparan las puntuaciones generadas, evaluando la capacidad del sistema para identificar patrones asociados a la desinformación y analizar la consistencia de los resultados obtenidos.
- Capítulo 6 – Conclusiones y líneas futuras. Resume las principales aportaciones y resultados del trabajo, valorando la viabilidad del enfoque propuesto para el análisis de desinformación en redes sociales. Además, se presentan diversas líneas futuras destinadas a mejorar la precisión, escalabilidad y aplicabilidad del sistema desarrollado.

2. ESTADO DEL ARTE

La desinformación se ha consolidado en los últimos años como uno de los principales problemas asociados al crecimiento de las redes sociales. Estas plataformas se han convertido en los principales canales de difusión de información, pero el hecho de que cualquier usuario pueda publicar contenido ha llevado a que estos contenidos no siempre sean fiables. Cada vez más vemos como campañas de desinformación se utilizan para la desestabilización de la sociedad durante momentos críticos, como en la reciente guerra de Irán, o en las últimas elecciones de EE.UU.

Ante esta importancia creciente, también ha aumentado la investigación en relación con como frenar esta nueva amenaza. Entre las principales líneas de actuación destacan el uso de distintas formas de verificación de hechos, el uso de algoritmos de aprendizaje automático para la detección de patrones, o el uso de mecanismos de filtrado de contenidos. Estas son solo algunas de las medidas que han empezado a tomar las diferentes plataformas e instituciones para hacer frente a la desinformación, y buscan abordar el problema desde diferentes perspectivas. Sin embargo, dada la complejidad de este, ninguna de ellas resulta completamente efectiva por si sola.

A continuación, se presenta una revisión del estado actual de estas estrategias, analizando su funcionamiento, su relevancia en el contexto actual y las principales limitaciones que presentan.

2.1. FACT-CHECKING

Pese a no ser totalmente imprescindible, uno de los rasgos más característicos de la desinformación es que la información difundida sea parcial o totalmente falsa. Por este motivo, la verificación de contenidos se ha consolidado como una de las principales herramientas empleadas para combatirla.

Históricamente, esta labor recaía principalmente en periodistas y medios de comunicación, que contaban con departamentos específicos encargados de verificar la información antes de su publicación. Sin embargo, el crecimiento exponencial de las redes sociales y la velocidad de difusión de contenidos han dificultado la aplicación exclusiva de estos métodos tradicionales. Como consecuencia, han surgido nuevas estrategias de verificación apoyadas tanto en la colaboración de los usuarios como en sistemas automatizados basados en IA.

En la actualidad, los principales mecanismos de verificación de información pueden clasificarse en tres grandes categorías: verificación manual, verificación comunitaria y verificación automática.

2.1.1. VERIFICACIÓN MANUAL

Este es el tipo de verificación más antiguo, y era el usado por los medios de comunicación convencionales en sus publicaciones. Esta verificación puede ser realizada por los propios medios de comunicación, como en el caso de la revista *Time*, que cuenta con una unidad específica encargada de la verificación de las noticias de los periódicos antes de que estos sean impresos [8], o bien por otras plataformas independientes y dedicadas únicamente a la verificación de la información, como Newtral.

Estas plataformas se caracterizan por el uso principal de fuentes gubernamentales y de medios de comunicación, y aunque no en todos los casos, las más modernas sí que permiten cierta implicación por parte de los usuarios, permitiendo el intercambio de mensajes, o publicando sus resoluciones en redes sociales [9]. Pese a que este método sigue siendo popular hoy en día, su limitación principal es el tiempo requerido: para poder verificar una información es necesario buscar fuentes, contrastarlas, y relacionarlas con la información original, y esto lleva una cantidad de tiempo y esfuerzo mucho mayor que el que lleva crear una nueva sentencia, por lo que a largo plazo siempre se irá con retraso. [10]

2.1.2. VERIFICACIÓN COMUNITARIA

Ante la necesidad de reaccionar contra el volumen de información que se maneja en redes sociales con suficiente rapidez, surgió la primera comunidad de verificación con BIRDWATCH de la plataforma Twitter [11]. A esta le siguieron otras como Meta [12] o TikTok [13]. El funcionamiento más habitual de estas herramientas es otorgar la capacidad de añadir comentarios a publicaciones en forma de notas a determinados usuarios. Si estas notas son aprobadas por un número significativo de personas, podrán ser mostradas en la plataforma como pie de la publicación. Mientras tanto, solo otros usuarios que estén en el programa de verificación podrán leer las notas que estén publicadas.

Estas herramientas han demostrado tener una eficacia cercana a la de las páginas especializadas, lo que sumado a su capacidad de reaccionar con rapidez ante desinformación que se propague por redes sociales, algo que la verificación tradicional no es capaz de conseguir, demuestran que pueden tener mucha utilidad de cara a combatir la desinformación [14]. Además, algunos estudios como [15] han demostrado que, aunque estén hechas por usuarios anónimos, siguen siendo consideradas fiables para los usuarios.

Sin embargo, estas herramientas presentan una desventaja, originada en el hecho de que no son trabajadores profesionales, sino usuarios de una red social. Principalmente a la hora de confirmar o desmentir publicaciones de carácter político, se ha comprobado que los usuarios de estas plataformas tienen una tendencia a defender las publicaciones que defiendan sus ideales, y a atacar a aquellas que sean contrarias a estos, con independencia del contenido de las publicaciones en sí [16].

Por lo tanto, las herramientas de verificación comunitaria son bastante más rápidas que las herramientas tradicionales, pero son menos precisas, especialmente en el ámbito político, debido a los sesgos previos de los participantes.

2.1.3. VERIFICACIÓN AUTOMÁTICA

La creciente cantidad de contenido publicado diariamente en redes sociales y medios digitales ha impulsado el desarrollo de sistemas automáticos de verificación de información. A diferencia de la verificación manual o comunitaria, estos sistemas buscan automatizar parcial o totalmente el proceso de detección y validación de afirmaciones mediante técnicas de NLP, recuperación de información y aprendizaje automático.

El objetivo principal de estos sistemas es reducir el tiempo y el esfuerzo necesarios para verificar una publicación, permitiendo analizar grandes volúmenes de información de forma escalable. Para ello, los sistemas automáticos suelen dividir el proceso de *fact-checking* en diferentes etapas, como la detección de afirmaciones verificables, la búsqueda de evidencias y la validación final de la afirmación [16].

Los primeros enfoques automáticos se basaban principalmente en métodos clásicos de NLP, utilizando representaciones estadísticas del texto como *Term Frecuency-Inverse Document Frecuency* (TF-IDF) o técnicas de similitud textual. Posteriormente, comenzaron a incorporarse modelos de aprendizaje profundo capaces de representar relaciones semánticas más complejas entre las palabras y las frases [17].

En la actualidad, gran parte de la investigación se centra en el uso de arquitecturas *Transformer* y modelos de lenguaje de gran tamaño, como BERT o T5, debido a su capacidad para comprender el contexto semántico de una afirmación [17]. Estos modelos permiten comparar publicaciones con bases de conocimiento, artículos periodísticos o verificaciones previas, y estimar automáticamente el grado

de veracidad de una afirmación. Además, algunos trabajos recientes utilizan estos modelos para generar resúmenes automáticos de verificaciones complejas, facilitando la comprensión de los resultados por parte de los usuarios [17].

Otro aspecto importante en los sistemas modernos de verificación automática es el uso de recuperación de información y grafos de conocimiento. Algunos modelos incorporan información procedente de bases de datos estructuradas, verificaciones anteriores o búsquedas web en tiempo real para contextualizar las afirmaciones analizadas [18]. Este enfoque permite mejorar la capacidad del sistema para trabajar con afirmaciones ambiguas, incompletas o relacionadas con eventos recientes.

Sin embargo, la verificación automática sigue presentando importantes limitaciones. Uno de los principales problemas es la dependencia del contexto: muchas afirmaciones no pueden clasificarse correctamente sin información externa adicional o sin comprender el contexto político, social o temporal en el que fueron publicadas [18]. Además, los modelos automáticos pueden verse afectados por sesgos presentes en los datos de entrenamiento, dificultades para interpretar ironía o sarcasmo, y problemas para adaptarse a eventos nuevos o información cambiante.

Otro problema relevante es la interpretabilidad de los modelos. Aunque algunos sistemas alcanzan resultados precisos en tareas de clasificación, en muchos casos resulta difícil justificar de manera transparente por qué una publicación ha sido clasificada como falsa o engañosa. Por este motivo, algunas investigaciones recientes como [18] proponen sistemas híbridos en los que la IA actúa como apoyo a verificadores humanos, en lugar de sustituir completamente el proceso manual.

En consecuencia, aunque la verificación automática representa una herramienta fundamental para combatir la desinformación a gran escala, las investigaciones actuales coinciden en que todavía resulta necesario combinar distintos enfoques y mantener supervisión humana en aquellos casos donde exista ambigüedad o falta de contexto.

2.2. ANÁLISIS DE EMOCIONES

Otra de las líneas de investigación más relevantes en la lucha contra la desinformación es el análisis emocional del lenguaje. Numerosos estudios como [19] han demostrado que los contenidos difundidos en redes sociales suelen apelar a emociones intensas, especialmente miedo, ira, indignación o ansiedad, con el objetivo de aumentar la interacción de los usuarios y favorecer la difusión del mensaje. Por este motivo, el análisis automático de emociones se ha convertido en una herramienta de gran interés para detectar contenidos potencialmente manipulativos o engañosos.

Las primeras aproximaciones al análisis emocional se basaban principalmente en diccionarios léxicos y técnicas estadísticas simples, donde determinadas palabras eran asociadas a emociones concretas. Sin embargo, estos enfoques presentaban limitaciones importantes debido a la ambigüedad del lenguaje y a la dificultad para interpretar correctamente el contexto de las publicaciones [20].

Con el avance del aprendizaje profundo y del NLP, comenzaron a utilizarse modelos neuronales capaces de capturar relaciones semánticas complejas dentro del texto. Actualmente, gran parte de los sistemas modernos emplean arquitecturas *Transformer*, como BERT, junto con técnicas de *word embeddings* y *transfer learning* para identificar emociones en publicaciones de redes sociales [21, 22]. Estos modelos permiten analizar no solo palabras individuales, sino también relaciones contextuales, tono y estructura lingüística.

Además del análisis textual tradicional, algunos trabajos recientes han explorado enfoques multimodales que combinan texto, imágenes, audio o expresiones faciales para detectar emociones de manera más precisa [23]. Este tipo de sistemas resulta especialmente útil en redes sociales, donde gran parte del contenido desinformativo se presenta acompañado de imágenes, vídeos o memes diseñados para generar reacciones emocionales intensas.

Diversas investigaciones como [24] han mostrado también una relación directa entre determinadas emociones y la propagación de noticias falsas. En particular, los mensajes que generan miedo, enfado o polarización tienden a difundirse más rápidamente que aquellos con un tono neutro o informativo [25, 26]. Por ello, algunos modelos de detección de *fake news* incorporan el análisis emocional como una característica adicional para mejorar la clasificación automática de contenidos sospechosos.

Otro aspecto relevante es el análisis emocional de las respuestas de los usuarios. Estudios como [27] combinan el análisis del contenido original con el estudio de comentarios y reacciones generadas en redes sociales, permitiendo detectar patrones colectivos de comportamiento asociados a campañas de desinformación. Este enfoque resulta especialmente útil para identificar contenidos altamente polarizantes o diseñados para provocar respuestas emocionales masivas.

Sin embargo, el análisis automático de emociones sigue presentando limitaciones importantes. El lenguaje utilizado en redes sociales suele incluir ironía, sarcasmo, dobles sentidos o referencias culturales difíciles de interpretar para los modelos automáticos. Además, las emociones humanas son complejas y no siempre pueden clasificarse de forma precisa dentro de categorías discretas [28].

A esto se suma el problema de los sesgos presentes en los datos de entrenamiento. Muchos modelos son entrenados con conjuntos de datos limitados o centrados en determinados idiomas y contextos culturales, lo que dificulta su generalización a otros entornos [28]. Asimismo, aunque los modelos actuales alcanzan resultados muy elevados en tareas específicas, todavía existen dificultades para interpretar de manera transparente qué características lingüísticas han llevado al sistema a asociar un mensaje con una emoción concreta.

En consecuencia, aunque el análisis emocional constituye una herramienta prometedora para la detección de desinformación y contenidos manipulativos, las investigaciones actuales apuntan a la necesidad de combinarlo con otros mecanismos de análisis contextual y verificación de información para obtener resultados más fiables y robustos.

2.3. MÉTRICAS EXISTENTES

Además de la detección y verificación de contenidos, una parte importante de la investigación sobre desinformación se centra en cómo medir su presencia, difusión e impacto. Esta cuestión resulta especialmente relevante porque la desinformación no solo puede analizarse como un problema de veracidad, sino también como un fenómeno social relacionado con la propagación de contenidos, la interacción de los usuarios y la eficacia de las intervenciones aplicadas para reducir sus efectos.

Una primera línea de trabajo se basa en el uso de métricas de interacción en redes sociales. Estas métricas incluyen variables como el número de compartidos, comentarios, respuestas, visualizaciones o “me gusta”, y permiten estimar el grado de difusión o alcance de un contenido potencialmente desinformativo. Algunos estudios [29, 30] han aplicado este enfoque al análisis de brotes concretos de desinformación, como el caso de la viruela del mono en Twitter, o al estudio de desinformación científica y sanitaria. Otros trabajos muestran que las métricas de acceso sencillo disponibles en

plataformas sociales pueden estar asociadas con la actividad de compartición de contenidos desinformativos [31].

Otra aproximación consiste en estudiar el papel de los algoritmos de recomendación y posicionamiento de contenidos. En este caso, la métrica no se limita al contenido en sí, sino al modo en que las plataformas amplifican determinadas publicaciones. Algunos estudios señalan que los sistemas optimizados para maximizar la interacción pueden favorecer indirectamente la difusión de contenidos polarizantes o desinformativos [32]. En esta línea, también se ha analizado la contribución de los algoritmos de recomendación a la exposición de los usuarios a contenidos problemáticos [33].

También existen métricas orientadas a evaluar la capacidad de los usuarios para identificar información falsa o engañosa. Estas métricas se centran en variables como la susceptibilidad a la desinformación, la capacidad de discernir entre noticias verdaderas y falsas, o la resiliencia del usuario ante contenidos manipulativos. En este ámbito se han propuesto instrumentos psicométricos específicos para medir la vulnerabilidad frente a la desinformación, como el "Misinformation Susceptibility Test", así como escalas orientadas a evaluar la resiliencia frente a la desinformación en redes sociales [34, 35].

Otra línea relevante se centra en la evaluación de intervenciones contra la desinformación. En lugar de medir únicamente si una publicación es verdadera o falsa, estos trabajos analizan la eficacia de estrategias como advertencias, correcciones, etiquetas de verificación o intervenciones educativas. Algunas investigaciones como [36, 37] estudian los retos de diseñar métricas adecuadas para evaluar este tipo de intervenciones, especialmente cuando se busca medir no solo el cambio inmediato en la creencia del usuario, sino también efectos posteriores sobre el comportamiento de compartición o la confianza en la información.

Desde el punto de vista computacional, las métricas más habituales siguen siendo las propias de los modelos de clasificación, como *accuracy*, *precision*, *recall* y *F1-score*. Estas métricas permiten evaluar el rendimiento de sistemas automáticos de detección de desinformación, aunque presentan limitaciones cuando los conjuntos de datos están desbalanceados o cuando las etiquetas utilizadas no capturan adecuadamente la complejidad del fenómeno. Por ello, algunas revisiones recientes [38, 39] destacan la necesidad de combinar métricas técnicas con medidas sociales, cognitivas y contextuales.

En conjunto, la literatura muestra que no existe una única métrica universal para medir la desinformación. Dependiendo del objetivo del estudio, puede ser necesario medir la veracidad del contenido, su difusión, la reacción emocional de los usuarios, la susceptibilidad individual, el efecto de las intervenciones o el rendimiento de modelos automáticos. Esta diversidad de enfoques evidencia la necesidad de desarrollar métricas más integradoras, capaces de combinar varios indicadores para ofrecer una evaluación más completa del riesgo asociado a un contenido.

3. ARQUITECTURA DEL SISTEMA

Este capítulo recoge la arquitectura propuesta para el presente trabajo. En primer lugar, debido a la inexistencia de una definición universalmente aceptada, se ofrece una definición operativa de desinformación. Además, en la literatura especializada aparecen conceptos relacionados que conviene diferenciar:

- *Misinformation*: información falsa, errónea o inexacta difundida sin intención de engañar o manipular. Generalmente, es consecuencia de errores, negligencias o sesgos inconscientes.
- *Malinformation*: utilización de información verídica, o parcialmente basada en hechos reales, con el propósito de causar daño a una persona, colectivo, organización, institución o país.

Tomando como referencia el "Trabajo del foro contra las campañas de desinformación", publicado en 2024 [40], se identifican inicialmente dos grandes grupos de conceptos asociados a la desinformación. A partir de dicho trabajo, se propone la siguiente definición operativa de desinformación, representada en la Figura 3.1:

«Difusión intencionada de información falsa, engañosa, descontextualizada o parcial cuyo objetivo es confundir, persuadir o manipular a las personas.»

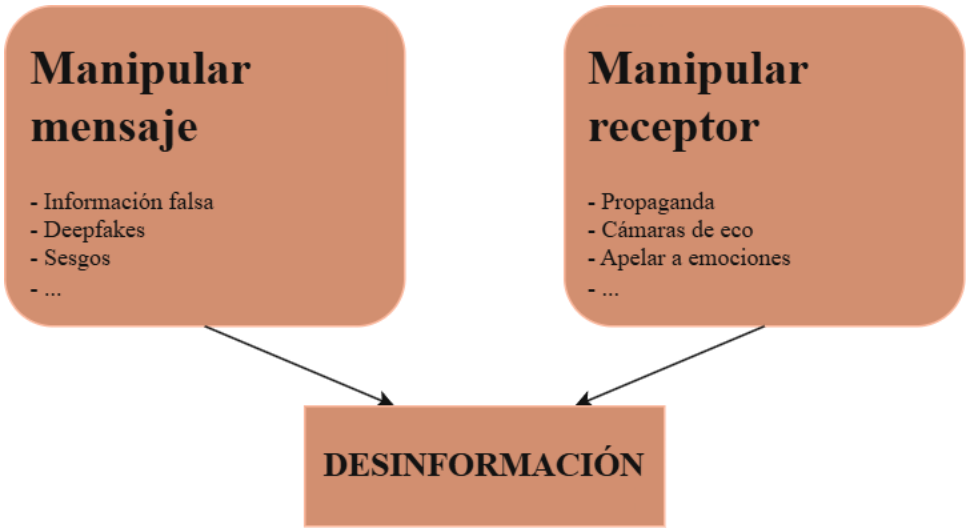


Figura 3.1 Definición de desinformación para este trabajo

Por tanto, para el desarrollo de la métrica es necesario considerar tanto la veracidad de la información como la finalidad que persigue su difusión. Este planteamiento introduce dos desafíos principales. En primer lugar, resulta necesario establecer qué criterios permiten determinar la veracidad de una publicación. En segundo lugar, debe abordarse la medición de un concepto tan abstracto como la intención comunicativa subyacente al mensaje.

La definición adoptada pone de manifiesto la existencia de dos dimensiones diferenciadas de la desinformación. Por un lado, se encuentran los mecanismos orientados a influir en el receptor. Por otro, aquellos centrados en la manipulación del contenido del mensaje. Esta distinción sirve como base para la clasificación de los conceptos asociados a la desinformación.

Para influir en el receptor, pueden identificarse los siguientes mecanismos, incluidos en la Tabla 3.1:

Mecanismo	Definición
Humor	Utilización de memes, sátiras o contenidos aparentemente humorísticos para difundir propaganda o información engañosa.
Apelación a las emociones	Empleo de mensajes diseñados para provocar respuestas emocionales intensas y reducir el análisis racional de la información.
Propaganda	Difusión de mensajes orientados a influir en las opiniones, actitudes o comportamientos del receptor.
Polarización	Generación o refuerzo de divisiones ideológicas, sociales o políticas para favorecer la confrontación entre grupos.
Exageración	Sobredimensionamiento de hechos o problemas con el fin de alterar la percepción del receptor.
Sensacionalismo	Presentación exagerada o alarmista de acontecimientos para maximizar el impacto emocional.
Coordinación	Difusión sincronizada de un mismo mensaje por múltiples cuentas para aumentar artificialmente su relevancia percibida.
Relativismo	Minimización de la importancia de determinados hechos o acontecimientos para reducir su impacto social.
Falsa equivalencia	Establecimiento de comparaciones desproporcionadas o engañosas entre situaciones distintas para justificar una determinada narrativa
Negacionismo	Rechazo sistemático de evidencias o hechos contrastados con el objetivo de sostener una posición concreta.
Cámaras de eco	Entornos sociales o digitales en los que predominan exclusivamente opiniones afines a las creencias de sus participantes.
Priorización de la cantidad sobre la veracidad	Difusión masiva de contenidos sin considerar su exactitud, con el objetivo de maximizar el alcance o la interacción.
Clickbait	Técnica basada en titulares, imágenes o descripciones diseñadas para atraer clics mediante la exageración o la omisión de información relevante.
Discursos de odio	Difusión de mensajes que atacan, degradan o deshumanizan a personas o colectivos por características inherentes como la raza, la religión o el sexo.

Tabla 3.1 Mecanismos de manipulación del receptor

Por otra parte, en la Tabla 3.2 se incluyen los mecanismos centrados en la manipulación del contenido.

Mecanismo	Definición
Información falsa	Información presentada como verdadera pese a no corresponderse con la realidad.

Falacias	Razonamientos aparentemente válidos que contienen errores lógicos o argumentativos.
Ambigüedad intencionada	Utilización deliberada de mensajes imprecisos o ambiguos para inducir interpretaciones erróneas.
Divulgación selectiva	Difusión únicamente de aquellos elementos de una información que favorecen una narrativa concreta.
Tergiversación de la realidad	Distorsión o reinterpretación forzada de hechos con el fin de obtener un beneficio comunicativo o ideológico.
Verdades mal utilizadas	Empleo de información veraz para justificar conclusiones incorrectas o razonamientos falaces.
Versiones sesgadas	Presentación parcial o selectiva de información con el objetivo de respaldar una narrativa determinada.
Pseudomedios	Plataformas que imitan la apariencia de medios de comunicación legítimos sin respetar estándares periodísticos o deontológicos
Conspiraciones	Explicaciones alternativas que atribuyen acontecimientos complejos a planes secretos promovidos por actores poderosos.
Suplantación	Utilización no autorizada de la identidad, nombre o imagen de una persona para actuar en su nombre.
Deepfakes	Imágenes, vídeos o audios generados o modificados mediante IA para representar hechos o personas de manera ficticia.
Manipulación de material	Alteración, edición o fabricación de contenidos con el fin de aparentar autenticidad y transmitir información engañosa.
Información falsa accidental	Información errónea difundida sin conocimiento de su falsedad.
Imprecisiones	Elementos incorrectos o inexactos presentes en una información que no es completamente falsa.

Tabla 3.2 Mecanismos de manipulación del contenido

Esta investigación preliminar permite comprender mejor la complejidad del fenómeno estudiado y proporciona una base conceptual sobre la que comparar y analizar las distintas publicaciones.

A partir del análisis realizado, pueden identificarse dos desafíos fundamentales que este TFG trata de abordar. El primero está relacionado con la intencionalidad del mensaje. Se trata de un concepto difícil de medir de forma objetiva, ya que una publicación puede haberse difundido con la intención de persuadir o manipular, pero también puede proceder de usuarios que comparten información falsa creyendo que es veraz.

El segundo desafío consiste en determinar la veracidad de la información. En la práctica, no todas las publicaciones pueden clasificarse de forma binaria como verdaderas o falsas. Algunos mensajes contienen información parcialmente correcta, mientras que otros no disponen de evidencias suficientes que permitan verificar su contenido.

Las posibles estrategias adoptadas para abordar ambos problemas se presentan en los apartados posteriores.

3.1. PROPUESTA DE CLASIFICACIÓN

El análisis se fundamenta en dos dimensiones principales: el contenido y el contexto del mensaje. El contenido hace referencia a la información transmitida por la publicación y a los elementos que la componen. Por su parte, el contexto engloba las características asociadas a su difusión, incluyendo quién publica el mensaje y las circunstancias en las que lo hace.

Debido a la complejidad del problema, se adopta una arquitectura dividida en cuatro fases claramente diferenciadas. Cada fase se centra en el análisis de una característica específica del mensaje. Posteriormente, los resultados obtenidos se integran para generar una evaluación global que considere de forma conjunta los distintos factores asociados a la desinformación.

A partir de estas consideraciones, se propone el modelo mostrado en la Figura 3.2.

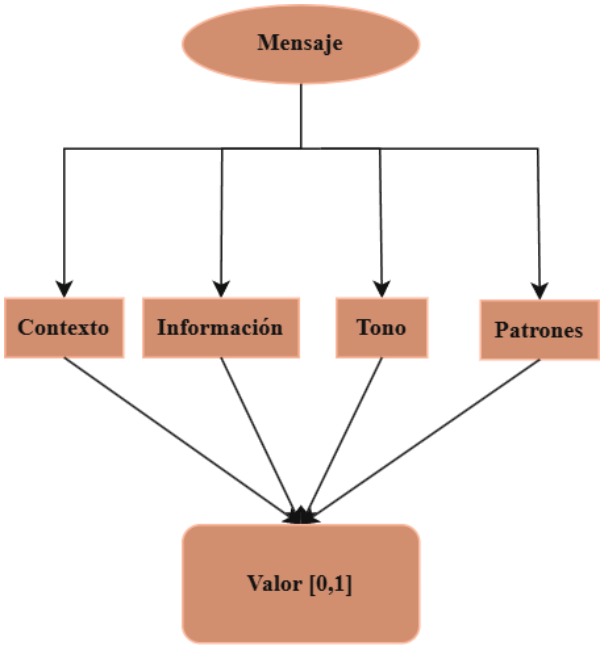


Figura 3.2 Propuesta de diseño compuesta de 4 fases

3.2. FASE 0 - EVALUACIÓN DEL CONTEXTO

La evaluación del contenido de una publicación resulta insuficiente si se analiza de forma aislada. También es necesario considerar el contexto en el que dicho contenido ha sido difundido. Como se ha indicado anteriormente, la intención con la que se publica un mensaje no puede medirse de forma directa. Sin embargo, determinadas características asociadas a la cuenta emisora y al entorno de publicación pueden aportar indicios sobre dicha intención.

Factores como la presencia de comportamientos automatizados, la participación en conversaciones especialmente relevantes en un momento determinado o la existencia de patrones de difusión coordinada pueden proporcionar información adicional para interpretar el mensaje analizado. No

obstante, esta aproximación debe aplicarse con cautela. Las redes sociales albergan una gran diversidad de perfiles, incluyendo cuentas humorísticas, satíricas o irónicas, cuyos contenidos pueden presentar características similares a las observadas en determinados casos de desinformación. Un diseño inadecuado de esta fase podría incrementar significativamente el número de falsos positivos.

Con el objetivo de caracterizar el contexto de la publicación, se analizan los siguientes elementos:

- Foto de perfil: se evalúa la presencia de elementos visuales potencialmente asociados a posicionamientos ideológicos, movimientos políticos o narrativas concretas.
- Usuario y nombre: se analiza la existencia de palabras con significado en distintos idiomas, principalmente español e inglés, así como la posible presencia de términos relacionados con ámbitos políticos o ideológicos.
- Descripción: se evalúa la presencia de lenguaje político, expresiones ofensivas y otros elementos relevantes para determinar el tono general del perfil.
- Amigos: se analiza la red de contactos de la cuenta con el objetivo de identificar posibles patrones de coordinación o estructuras compatibles con cámaras de eco.
- Últimos cinco posts: se estudia el historial reciente de publicaciones para detectar comportamientos recurrentes o patrones consistentes de difusión.
- Hecho reciente: se evalúa si la publicación hace referencia a acontecimientos relevantes ocurridos durante los treinta días anteriores a su publicación.

Dado que no todos los elementos analizados aportan la misma cantidad de información, se asigna una ponderación específica a cada uno de ellos. En la Tabla 3.3 se presentan los pesos utilizados y la justificación de cada decisión de diseño.

Rasgo	Peso	Justificación
Foto de perfil	5%	La fotografía de perfil puede aportar información contextual, pero su utilidad es limitada. Muchas cuentas automatizadas carecen de fotografía o utilizan imágenes genéricas. Por ello, este rasgo recibe una ponderación reducida frente a otros indicadores más representativos.
Usuario/Nombre	5%	El nombre de usuario puede aportar información contextual relevante, pero su capacidad discriminatoria es limitada. Por ello, recibe una ponderación reducida.
Descripción	20%	La descripción del perfil constituye un rasgo más relevante, ya que puede aportar información útil sobre la naturaleza de la cuenta. Además de facilitar la identificación de perfiles satíricos o humorísticos, contribuye a reducir posibles falsos positivos. Asimismo, la presencia de lenguaje político o ideológico suele proporcionar indicios sobre el tipo de contenido difundido por la cuenta.
Amigos	30%	Este es uno de los rasgos con mayor peso dentro de la evaluación contextual. El análisis de la red de contactos permite identificar posibles cámaras de eco, entornos altamente politizados o patrones de interacción compatibles con la difusión coordinada de contenido. Además, este criterio penaliza principalmente a las cuentas integradas en dichas estructuras, reduciendo el impacto sobre usuarios ordinarios o cuentas automatizadas sin actividad sospechosa.

Últimos 5 posts	30%	Este es el rasgo con mayor peso dentro de la evaluación contextual. El análisis de las publicaciones recientes permite caracterizar el comportamiento de la cuenta e identificar patrones consistentes en el contenido difundido. Además, se trata de un criterio especialmente robusto, ya que reduce la probabilidad de falsos positivos al basarse exclusivamente en la actividad de la propia cuenta.
Hecho reciente	10%	La relación de una publicación con acontecimientos recientes puede aportar información contextual relevante, ya que las campañas de desinformación suelen concentrarse en torno a temas de actualidad. Sin embargo, este criterio recibe una ponderación moderada, puesto que los eventos de gran repercusión generan un elevado volumen de publicaciones legítimas. Una ponderación excesiva podría incrementar significativamente el número de falsos positivos.

Tabla 3.3 Pesos de los rasgos medidos en la fase 0

Como puede observarse, los dos elementos con mayor peso dentro de esta fase son el entorno cercano de la cuenta y su comportamiento reciente. El objetivo de este análisis es obtener una estimación razonable del contexto en el que se difunde la publicación y de los patrones de comportamiento asociados a la cuenta emisora.

Esta información se complementa posteriormente con los resultados de las fases 1, 2 y 3, centradas en el análisis directo del contenido del mensaje.

3.3. FASE 1 - VERIFICACIÓN DE LA PUBLICACIÓN

Una de las características más representativas de la desinformación es la difusión de información falsa o engañosa. Por este motivo, resulta imprescindible incorporar una fase dedicada exclusivamente a la evaluación de la veracidad del contenido.

La fase 1 tiene como objetivo analizar la veracidad de las afirmaciones presentes en el mensaje, centrándose únicamente en aquellas que pueden contrastarse mediante evidencia externa. Para ello, se introduce el concepto de sentencias *checkworthy*, entendidas como afirmaciones que describen hechos verificables a través de fuentes de información disponibles.

El análisis se realiza a nivel de sentencia, ya que una misma publicación puede combinar afirmaciones verificables, opiniones y otros elementos que no pueden contrastarse objetivamente. Este enfoque permite obtener una evaluación más granular y precisa del contenido del mensaje.

Como resultado, esta fase distingue entre contenido veraz, contenido falso y contenido no verificable. Esta diferenciación resulta fundamental para evaluar la fiabilidad de la información difundida y complementar el análisis realizado en el resto de las fases del sistema.

3.4. FASE 2 - ANÁLISIS DE TONO

Otro rasgo frecuentemente asociado a la desinformación es la utilización de mensajes con una elevada carga negativa. Dado que este tipo de contenido suele buscar una respuesta emocional por parte de los usuarios, es habitual recurrir a formulaciones que transmitan enfado, miedo, indignación u otras emociones similares. Además, diversos estudios han señalado que los algoritmos de recomendación de las redes sociales tienden a favorecer contenidos capaces de generar un elevado nivel de interacción.

El análisis de tono tiene como objetivo identificar la carga emocional general del mensaje, clasificándolo como positivo, neutro o negativo. Aunque un tono negativo no implica necesariamente la existencia de desinformación, sí constituye un indicador relevante debido a su frecuente presencia en este tipo de contenidos.

Esta fase permite caracterizar el tono global de la publicación y complementa el análisis más específico realizado en la fase 3. De este modo, se distingue entre la orientación general del mensaje y la presencia de emociones concretas, reduciendo el riesgo de penalizar en exceso publicaciones críticas o negativas que no contengan elementos desinformadores. Asimismo, los resultados obtenidos contribuyen a completar la caracterización contextual iniciada en la fase 0 y aportan información adicional sobre los patrones comunicativos presentes en el mensaje.

3.5. FASE 3 - IDENTIFICACIÓN DE PATRONES

Como se ha expuesto anteriormente, el análisis de la desinformación constituye un problema complejo. No existe una definición universalmente aceptada del fenómeno, la intención detrás de una publicación no puede medirse de forma directa y numerosos mensajes se sitúan en zonas grises difíciles de clasificar. Sin embargo, es posible identificar determinados patrones que aparecen con frecuencia en contenidos asociados a la desinformación, independientemente de la cuenta que los publique o del grado de veracidad de la información difundida.

Con base en esta idea, se ha decidido analizar los siguientes patrones:

- Sesgo: permite identificar presentaciones parciales o desequilibradas de la información. Aunque los hechos descritos puedan ser ciertos, su selección o interpretación puede favorecer una narrativa concreta.
- Emociones: las emociones negativas suelen estar presentes en publicaciones que buscan maximizar su impacto sobre el receptor. A diferencia del análisis de tono, este rasgo se centra en emociones específicas como el miedo, la ira o el desprecio.
- Discursos de odio: permiten detectar mensajes dirigidos contra personas o colectivos mediante expresiones ofensivas, discriminatorias o deshumanizadoras.
- Falacias: identifican razonamientos incorrectos o engañosos que pueden inducir al receptor a conclusiones erróneas.
- Propaganda: permiten detectar estrategias discursivas orientadas a influir en opiniones, actitudes o comportamientos mediante recursos persuasivos.
- Violencia: identifican la presencia de lenguaje agresivo, amenazante o que promueva conductas violentas.

Para el análisis de estos rasgos se adopta un enfoque modular, en el que cada uno de ellos es evaluado de forma independiente para posteriormente combinar sus resultados en una medida conjunta.

Dado que no todos estos patrones contribuyen en igual medida a la identificación de contenidos potencialmente desinformadores, se ha asignado una ponderación específica a cada uno de ellos. La Tabla 3.4 recoge los pesos utilizados y la justificación de cada uno.

Rasgo	Peso	Justificación
Sesgo	25%	Una parte significativa de la desinformación no se basa en información completamente falsa, sino en la presentación sesgada de hechos con el objetivo de reforzar una narrativa o posición ideológica determinada. Por este motivo, la detección de sesgos se considera uno de los indicadores más relevantes.

Propaganda	30%	La propaganda constituye uno de los mecanismos más frecuentes en la difusión de narrativas desinformadoras. Aunque no toda la propaganda implica desinformación, una parte significativa de los contenidos desinformadores presenta elementos propagandísticos. Por este motivo, se asigna a este rasgo la mayor ponderación dentro de la fase.
Falacias	20%	Al igual que el sesgo, las falacias permiten transmitir mensajes aparentemente válidos sin necesidad de recurrir a información completamente falsa. Mediante razonamientos incorrectos o engañosos es posible inducir al receptor a conclusiones erróneas, por lo que este rasgo recibe una ponderación elevada.
Violencia y Discursos de odio	10%	Aunque la violencia constituye un rasgo frecuente en determinados contenidos desinformadores, su presencia no implica necesariamente la existencia de desinformación. En el contexto de las redes sociales, este tipo de mensajes suele generar elevados niveles de interacción, por lo que puede aparecer también en publicaciones legítimas. Por este motivo, se le asigna una ponderación reducida, suficiente para tener en cuenta este factor sin incrementar significativamente el riesgo de falsos positivos.
Emociones negativas	15%	Al igual que la violencia, la presencia de emociones negativas es un rasgo frecuente en numerosos contenidos publicados en redes sociales. Sin embargo, las emociones específicas analizadas en esta fase, como el miedo, la ira o el desprecio, presentan una relación más estrecha con los mensajes potencialmente desinformadores. Por este motivo, se le asigna una ponderación moderada, superior a la de la violencia, pero suficientemente contenida para evitar un aumento significativo de los falsos positivos.

Tabla 3.4 Pesos de los rasgos medidos en la fase 3

Con esto finaliza el proceso de análisis del mensaje. La combinación de las cuatro fases permite obtener una evaluación conjunta de la veracidad de la información difundida y de los patrones contextuales y lingüísticos asociados a la publicación.

A partir de estos resultados, es posible realizar una estimación razonable del potencial desinformador del mensaje, considerando tanto su contenido como el contexto en el que ha sido difundido.

3.6. PESOS GLOBALES

Dado que las distintas fases aportan información de relevancia desigual, se ha asignado una ponderación específica a cada una de ellas. La Tabla 3.5 recoge los pesos utilizados en el cálculo de la métrica global.

Fase	Peso	Justificación
Contexto	10%	Aunque las características de la cuenta emisora aportan información relevante, el elemento principal de análisis debe ser la propia publicación. Por este motivo, el contexto actúa como un factor complementario que puede modificar ligeramente el resultado sin condicionar de forma excesiva la evaluación final.
Verificación	30%	La veracidad de la información constituye uno de los elementos centrales de la desinformación. Sin embargo, no todas las afirmaciones falsas tienen la

		misma relevancia y numerosos mensajes contienen información parcialmente verificable. Por este motivo, esta fase recibe una ponderación elevada, aunque no dominante.
Tono	20%	El tono del mensaje aporta información relevante sobre su orientación comunicativa. Los contenidos desinformadores suelen presentar una carga negativa superior a la media, aunque este rasgo también está presente en publicaciones legítimas de carácter crítico. Por ello, se le asigna una ponderación moderada.
Patrones	40%	Esta fase evalúa la presencia de patrones frecuentemente asociados a la desinformación, como el sesgo, las falacias o la propaganda. Dado que estos elementos reflejan características estructurales del mensaje más allá de su veracidad inmediata, constituyen la principal fuente de información del modelo y reciben la mayor ponderación.

Tabla 3.5 Pesos de las 4 fases

El resultado final del modelo es un valor comprendido entre 0 y 1. Un valor de 0 representa una publicación completamente verídica, mientras que un valor de 1 representa una publicación con el máximo nivel de desinformación según los criterios definidos.

Esta métrica podría emplearse posteriormente para la clasificación y comparación de publicaciones, o como base para el desarrollo de indicadores agregados a nivel de usuario. No obstante, dichas aplicaciones quedan fuera del alcance de este trabajo.

4. DESARROLLO DEL SISTEMA

Para el desarrollo del prototipo del sistema, se ha optado por un enfoque modular basado en el uso de múltiples modelos de NLP, cada uno especializado en una tarea concreta. Esta decisión se justifica en la complejidad y tiempo requeridos en el entrenamiento de un modelo y desarrollo de un *dataset* con el que afinar un modelo generalista. El uso de modelos ya existentes permite el desarrollo de un prototipo funcional, aprovechando los avances existentes en tareas como el análisis de tono o la identificación de falacias.

Desde el punto de vista teórico, este enfoque representa una ventaja clara frente a la utilización de un único modelo generalista. Al estar empleando modelos especializados, cada uno en una tarea concreta, se busca conseguir una mayor precisión en la detección de los rasgos característicos de la desinformación. En lugar de analizar de manera global, se realizan numerosos análisis específicos, consiguiendo dividir el problema principal en tareas más manejables.

Sin embargo, en la práctica, esta ventaja teórica no siempre va a materializarse plenamente. Las condiciones en las que se realiza este prototipo, especialmente la falta de control sobre el entrenamiento de los modelos utilizados, y la necesidad de integrarlos de manera coherente, hacen que el rendimiento del sistema completo no necesariamente sea mucho mejor que el que podría obtener un modelo generalista bien entrenado.

Este planteamiento presenta varias ventajas importantes. En primer lugar, la modularidad facilita mucho la escalabilidad y mantenibilidad del sistema. Cada componente puede ser sustituido por otro modelo más avanzado sin necesidad de rediseñar el sistema completo. Esto resulta especialmente útil en un campo tan volátil como el de NLP, y también en concreto en el análisis de la desinformación, ya que esta también es muy cambiante.

Por otro lado, el principal inconveniente de este planteamiento es la falta de control sobre los modelos utilizados. Al estar trabajando con modelos entrenados con *datasets* ajenos, es posible que incorporen sesgos inherentes a estos datos, lo que es especialmente problemático en un asunto tan subjetivo como la detección de desinformación. Además, requiere de una normalización previa de los outputs de los modelos, lo que también puede afectar al rendimiento al no estarse usando necesariamente de la manera óptima.

A continuación, en la Figura 4.1 se presenta el flujo de funcionamiento del modelo planteado. Este diagrama será utilizado como base para el desarrollo de un prototipo funcional.

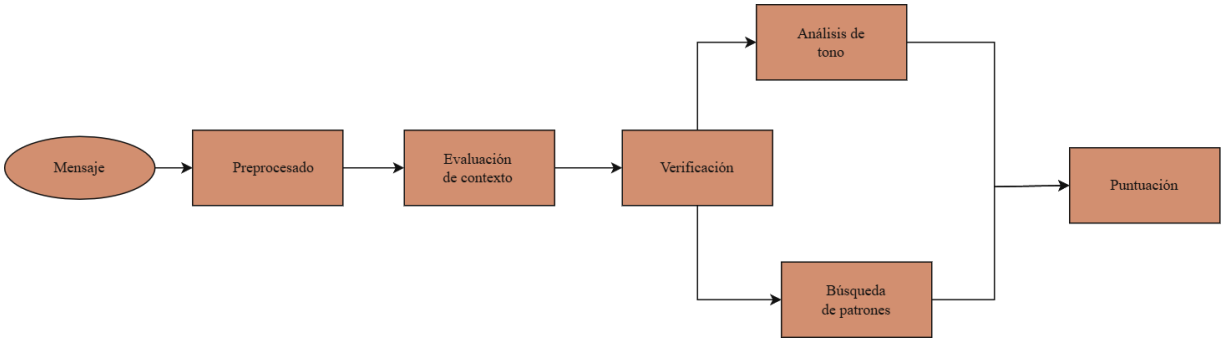


Figura 4.1 Flujo del modelo desarrollado

4.1. PREPROCESADO DEL MENSAJE

Debido a que en gran parte del sistema se utilizan modelos de NLP que han sido entrenados principalmente para el inglés, se ha decidido añadir una fase de preprocesado del mensaje. En esta fase primero se estandariza el texto, eliminando "@" y enlaces presentes en el mensaje que no aportan información a nivel semántico y pueden incluso suponer ruido para los modelos que se van a encargar de procesar el mensaje.

Debido al uso simultáneo de múltiples modelos diferentes y ya entrenados, se puede plantear un preprocesado más avanzado que ocupe hashtags, emoticonos y expresiones complejas, como palabras en clave o con letras duplicadas. Posteriormente se realiza una traducción al inglés del texto procesado usando un modelo específico, obtenido de Hugging Face: "Helsinki-NLP/opus-mt-es-en" [41]. Este es un modelo que, aunque fue entrenado en 2020, muestra bastante precisión y sigue siendo muy utilizado hoy en día.

Esta decisión puede afectar a la precisión del modelo en algunas situaciones, puesto que en el proceso de traducción se pueden perder características específicas del idioma y el mensaje, algo que este modelo acusa especialmente debido a que se analizan rasgos del texto abstractos como el tono, la presencia de falacias, etc. Igualmente, se ha considerado adecuado implementarlo en formato bilingüe a pesar de enfrentarse a un posible decremento residual de la precisión del modelo. Esta decisión está fundada en la necesidad de cubrir los 2 lenguajes más populares de X, la red social para la que este prototipo ha sido diseñado.

4.2. EVALUACIÓN DE CONTEXTO

La evaluación del contexto se realiza a partir de distintos elementos asociados a la cuenta y al contenido publicado por el usuario. Cada uno de estos elementos se analiza mediante submódulos específicos, obteniéndose una puntuación parcial normalizada entre 0 y 1. El resultado de este módulo es el parámetro "context".

4.2.1. FOTO DE PERFIL

Para la evaluación de la foto de perfil se emplea un modelo de análisis de imágenes capaz de determinar si una imagen contiene elementos relacionados con contenido político. Debido a la ausencia de modelos específicamente entrenados para esta tarea concreta, se ha optado por utilizar un modelo visual generalista especializado mediante *prompt engineering*.

El modelo clasifica cada imagen en una de dos categorías, política o no política, devolviendo un valor binario de 0 o 1, donde 0 representa ausencia de contenido político y 1 presencia de contenido político. El *prompt* diseñado considera como contenido político elementos como políticos reconocibles, símbolos partidistas, propaganda electoral, pancartas de protesta, banderas utilizadas en un contexto político o imágenes de manifestaciones y actos institucionales. Además, el *prompt* fuerza un comportamiento conservador, de forma que únicamente se asigna un valor positivo cuando la evidencia visual resulta suficientemente clara, para evitar falsos positivos.

El *prompt* utilizado es el siguiente:

> "*You inspect images and decide whether they contain political imagery. Political imagery includes politicians, political symbols, campaign signs, party logos, protest placards, flags used in a political context, election materials, government officials, political rallies, or other clearly political visuals. Return only JSON with one key: has_political_imagery (0 or 1). Use*

1 only when the image clearly contains political imagery. If the evidence is weak or ambiguous, return 0."

4.2.2. NOMBRE DE USUARIO

Para el análisis del nombre de usuario se utiliza el modelo de NLP "Political_DEBATE_large_v1.0" [42], especializado en la clasificación binaria de textos en las categorías "political" y "non-political". Además de la categoría asignada, el modelo devuelve una puntuación de confianza comprendida entre 0 y 1.

Con el objetivo de normalizar la salida del modelo y mantener una escala homogénea en el sistema, cuando el resultado obtenido corresponde a la categoría "political" se utiliza directamente el valor de confianza proporcionado por el modelo. Por el contrario, cuando la categoría asignada es "non-political", la puntuación utilizada se calcula como el complemento de la confianza respecto a 1. De esta forma, valores altos representan una mayor asociación política independientemente de la categoría devuelta inicialmente.

4.2.3. DESCRIPCIÓN

La descripción del perfil se analiza utilizando el mismo modelo empleado para el nombre de usuario. Sin embargo, en este caso el análisis se complementa mediante el cálculo de las métricas de patrones y tono desarrolladas previamente en el sistema. La puntuación final correspondiente a la descripción se obtiene mediante una combinación ponderada de estos tres factores:

$$Description = politicalness \cdot 0.5 + patterns \cdot 0.25 + tone \cdot 0.25$$

Esta ponderación permite otorgar un mayor peso al carácter político explícito de la descripción, manteniendo al mismo tiempo la influencia de elementos relacionados con el tono discursivo y la presencia de patrones lingüísticos asociados a desinformación.

4.2.4. AMIGOS

Con el objetivo de analizar las relaciones del usuario se seleccionan aleatoriamente cinco cuentas pertenecientes a las listas de seguidores y seguidos de la cuenta bajo estudio. El uso de cuentas aleatorias permite obtener una representación aproximada del entorno social del usuario evitando introducir sesgos derivados de otros criterios de selección, como la popularidad o el número de seguidores.

Sobre cada una de estas cuentas se realiza un análisis simplificado limitado a la fotografía de perfil, el nombre de usuario y la descripción, siguiendo la fórmula siguiente:

$$Score = Profile_picture \cdot 0.25 + Username \cdot 0.25 + Description \cdot 0.5$$

Cada uno de estos elementos se evalúa siguiendo la metodología descrita anteriormente. La puntuación de esta sección se obtiene de hacer la media de las puntuaciones obtenidas por cada perfil

La decisión de utilizar un análisis reducido responde principalmente a criterios de eficiencia computacional. Realizar un análisis completo de múltiples cuentas relacionadas incrementaría significativamente el coste temporal y económico del sistema, mientras que el impacto relativo de este

apartado dentro de la puntuación global es reducido. Por ello, se considera aceptable sacrificar parcialmente la precisión a cambio de mejorar la escalabilidad del prototipo.

4.2.5. ÚLTIMOS 5 POSTS

El análisis de las últimas publicaciones consiste en aplicar el módulo de análisis textual sobre los cinco posts más recientes del usuario. En este caso se utilizan pesos distintos a los empleados en el análisis principal, con el objetivo de reducir la influencia del contexto y centrarse en las características intrínsecas del contenido publicado. La puntuación de cada publicación se calcula mediante la siguiente expresión:

$$Last_Posts = Patterns \cdot 0.5 + Tone \cdot 0.2 + Fakeness \cdot 0.3$$

La puntuación final de este apartado corresponde a la media aritmética de las puntuaciones obtenidas para todas las publicaciones analizadas.

4.2.6. HECHO RECIENTE

Para determinar si el contenido hace referencia a un hecho reciente se emplea un submódulo basado en dos APIs externas: Tavily API y OpenAI API. Tavily API se utiliza para recopilar fuentes e información relacionada con el evento descrito en el texto, mientras que OpenAI API se emplea posteriormente para analizar dicha información y extraer una conclusión temporal. El sistema evalúa si el hecho referido ocurrió dentro de un intervalo temporal reciente respecto a la fecha de publicación. Para ello, el modelo analiza las evidencias encontradas, identifica posibles referencias temporales y estima la fecha aproximada del evento.

Como resultado, el submódulo devuelve una clasificación binaria indicando si el evento es reciente o no, acompañada de una puntuación de confianza y de información auxiliar relacionada con las evidencias encontradas. El *prompt* utilizado fuerza nuevamente un comportamiento conservador, priorizando evitar falsos positivos cuando las fuentes disponibles no permiten alcanzar una conclusión suficientemente sólida.

El *prompt* utilizado para evaluar las fuentes recopiladas es:

> *"You determine whether a described event is recent relative to today's date. Use the web evidence carefully and be conservative when evidence is weak. If the event happened within the last 30 days, or is clearly being reported as current, mark it recent. If the event is historical or older than 30 days, mark it not recent. If the evidence does not support a confident judgment, mark it not recent and lower the confidence. Return only JSON with these keys: is_recent (boolean), confidence (number 0 to 1), event_date (string or null), evidence_summary (string), reasoning (string), signals (array of strings)."*

4.3. VERIFICACIÓN DE LA PUBLICACIÓN

Para la verificación de la información se emplea una herramienta denominada “ClaimeAI” [43], diseñada para la extracción automática de afirmaciones verificables a partir de un texto, la búsqueda de información relacionada y el contraste de dichas afirmaciones con las fuentes encontradas.

El funcionamiento de la herramienta se apoya en el uso combinado de dos APIs externas: OpenAI API y Tavily API. En primer lugar, a partir del texto de entrada, se utiliza un *prompt* específico orientado a identificar aquellas sentencias consideradas *checkworthy*, es decir, afirmaciones susceptibles de ser verificadas objetivamente. Posteriormente, un segundo *prompt* se encarga de generar consultas de búsqueda y recuperar información relevante mediante Tavily API. Finalmente, un tercer *prompt* analiza la consistencia entre las afirmaciones extraídas y la información obtenida, determinando si estas quedan verificadas, desmentidas o si no existe evidencia suficiente para alcanzar una conclusión clara.

La herramienta original ha sido adaptada para ajustarse a los requisitos específicos de este TFG. Entre las modificaciones realizadas destacan la adaptación parcial de los *prompts* utilizados y la modificación del formato de salida, con el objetivo de integrarlo dentro de la arquitectura general del sistema desarrollado. El *prompt* que hubo que modificar fue el encargado de seleccionar las sentencias. Este estaba configurado para discriminar sentencias que no fueran afirmaciones científicas, por lo que, para hacerlo más adecuado al tipo de publicaciones que cubre este trabajo, se añadió lo siguiente:

> *"You are an assistant to a fact-checker that is focused on battling misinformation...*
> *A sentence DOES contain a specific and verifiable proposition if it asserts, implies, or accuses a person, organization, or group of having committed a concrete action, especially:*
>
> - *crimes*
> - *corruption*
> - *theft*
> - *fraud*
> - *abuse*
> - *illegal acts*
> - *deception*
> - *misconduct*
> - *participation in conspiracies*
> - *harmful real-world actions*
>
> *This applies even if:*
>
> - *the sentence is emotionally charged,*
> - *insulting,*
> - *politically biased,*
> - *hyperbolic,*
> - *defamatory,*
> - *sarcastic,*
> - *or expressed as rhetoric…"*

A cada afirmación identificada como *checkworthy* se le asigna una puntuación en función del resultado del proceso de verificación. En la Tabla 4.1 se especifican las puntuaciones asignadas, así como la justificación detrás de cada decisión. Para finalizar, el sistema calcula la media de las puntuaciones obtenidas para todas las afirmaciones analizadas en el texto. El resultado de este módulo es el parámetro "fakeness".

Resultado	**Puntuación**	**Justificación**
Verificada	0.0	Si la sentencia es cierta no debe aportar nada al cálculo final

Desmentida	1.0	Al estar la variable de verificación entre 0 y 1, siendo 0 "cierto" y 1 "falso", a una sentencia falsa deberá asignársele la máxima puntuación
Sin conclusión clara	0.4	Se escoge este valor de manera que un texto con una sentencia marcada como cierta y una como falsa puntúe más alto que un texto con únicamente una sentencia ambigua, provocando que lo que más penalice en el cálculo sea la presencia de información verificable como falsa

Tabla 4.1 Puntuaciones asignadas a los resultados de ClaimeAI

4.4. ANÁLISIS DE TONO

Para el análisis del tono, se hace uso de un modelo específico para analizar el tono de los tweets. Este es "Twitter-roBERTa-base for Sentiment Analysis" [44], un modelo entrenado con más de 124 millones de tweets. A pesar de ser del 2022, los resultados que arroja demuestran ser más fiables que otros modelos más actuales.

Este modelo fue entrenado principalmente con datos en inglés, siendo mucho más preciso en este idioma. Esto justifica la decisión de añadir el preprocesado del mensaje a este modelo en concreto, pues además no es el único caso.

El modelo devuelve una clasificación del texto en tres categorías (positivo, neutral o negativo), junto con una puntuación asociada a la categoría. Estas puntuaciones, en línea con el resto de los módulos, pueden tomar un valor entre 0 y 1. Para el modelo utilizaremos el valor de confianza asociado a la etiqueta negativo. Se ha decidido el uso de únicamente esta etiqueta puesto que es la más relevante para el sistema en general. Las otras 2 etiquetas que ofrece el modelo como salida no se consideran lo suficientemente relevantes en el contexto de este trabajo como para ser utilizadas. Por tanto, el resultado de este módulo es el parámetro "tone", cuyo valor corresponderá con aquel asociado a la etiqueta elegida.

Este modelo es el más afectado por el problema de los falsos positivos, como ya se explicó en el apartado de diseño, puesto que pueden existir tanto textos irónicos como críticas genuinas que queden marcadas. Igualmente, debido al bajo peso del apartado, no se considera muy relevante, ya que en el caso de ser falsos positivos se verán compensados por el resto de medidas.

4.5. ANÁLISIS DE PATRONES

Para el análisis de los distintos patrones definidos en la fase de diseño se emplean modelos de NLP especializados en la detección de cada uno de los rasgos. Dado que no existe un único modelo capaz de identificar todos ellos de manera conjunta, se adopta un enfoque modular en el que cada modelo procesa el texto de forma independiente.

Al estar utilizando tantos modelos diferentes, ha sido necesario establecer una normalización de los *outputs* de los mismos, de manera que se pudiera trabajar con todos ellos de la misma forma. En línea con el resto del trabajo, los outputs de todos los modelos serán un número acotado entre 0 y 1.

A continuación, se presentan en la Tabla 4.2 los modelos utilizados para cada uno de los rasgos que se medirán en la fase 3. En posteriores subsecciones se especifica de qué manera se utiliza cada uno y las

adaptaciones que se realizan para cumplir con la estandarización de los *outputs*. El resultado de este módulo es por tanto el valor del parámetro “patterns”.

Rasgo	Modelo
Sesgo	“Bias-detector”. [45]
Propaganda	En este caso no se utiliza un modelo directamente, sino que se hacen llamadas a una API externa: Tanbih API. [46]
Falacias	“Distilbert-base-fallacy-classification”. [47]
Violencia	“Small-guard-32m-en-prompt-violence-binary-moderation”. [48]
Discursos de odio	“Pysentimiento”. [49]
Emociones negativas	“Pysentimiento”. [49]

Tabla 4.2 Modelos elegidos para la implementación de la fase 3

En la Tabla 4.3 se presentan las adaptaciones realizadas para homogeneizar las salidas generadas por los distintos modelos.

Rasgo	Salida original	Estrategia de normalización	Rango final
Sesgo	Etiqueta ("bias" o "non-bias") y confianza.	Si la etiqueta es "bias", se utiliza directamente la confianza. Si es "non-bias", se utiliza 1 menos la confianza.	[0,1]
Emociones negativas	Puntuaciones para siete emociones.	Media aritmética de las puntuaciones asociadas a "disgust", "anger", "sadness" y "fear".	[0,1]
Discurso de odio	Puntuaciones para "hateful", "targeted" y "aggressive".	Media aritmética de las tres puntuaciones.	[0,1]
Falacias	Puntuaciones para trece falacias y una categoría miscelánea.	Diferencia entre la puntuación de la falacia más probable y la puntuación de la categoría miscelánea. Los valores negativos se sustituyen por 0.	[0,0.99]
Propaganda	Etiqueta ("propagandistic" o "not-propagandistic") y confianza.	Si la etiqueta es "propagandistic", se utiliza la confianza normalizada. Si es "not-propagandistic", se utiliza 1 menos la confianza normalizada.	[0,1]
Violencia	Puntuación asociada a la etiqueta "violent".	No se realiza ninguna transformación.	[0,1]

Tabla 4.3 Estrategias de normalización de las salidas de los modelos

La normalización aplicada al detector de falacias requiere una consideración adicional. Dado que el modelo distribuye una probabilidad total de 1 entre todas las categorías posibles, una puntuación

elevada para una falacia concreta no implica necesariamente una alta confianza en su presencia. Por este motivo, se descuenta la puntuación de la categoría miscelánea, obteniendo una medida más representativa de la presencia efectiva de razonamientos falaces.

4.6. INTEGRACIÓN DE LOS MÓDULOS.

Una vez desarrollados los distintos submódulos del sistema, ha sido necesario definir un mecanismo de integración que permita combinar sus resultados de manera coherente dentro de una única medida global. Debido a que cada uno de los módulos empleados utiliza modelos, arquitecturas y metodologías diferentes, uno de los principales retos durante el desarrollo ha sido conseguir que todos ellos produzcan resultados comparables y compatibles entre sí.

Para resolver este problema, se ha adoptado una arquitectura modular basada en puntuaciones normalizadas. Cada fase del sistema devuelve un valor comprendido entre 0 y 1, donde los valores próximos a 0 representan contenido con bajo riesgo de desinformación y los valores cercanos a 1 representan contenido potencialmente desinformador. Esta normalización permite integrar modelos heterogéneos dentro de una misma arquitectura sin necesidad de realizar transformaciones complejas posteriores.

El flujo general de funcionamiento del sistema puede dividirse en las siguientes etapas:

1. Obtención y preprocesado de la publicación y de la información contextual asociada.
2. Ejecución independiente de las cuatro fases principales del sistema.
3. Normalización de las salidas generadas por cada módulo.
4. Agregación ponderada de las puntuaciones parciales.
5. Generación de la puntuación final de riesgo de desinformación.

La independencia entre módulos constituye una de las principales características del sistema desarrollado. Cada fase opera de manera autónoma sobre el texto o sobre la información contextual obtenida, sin depender directamente de los resultados internos del resto de módulos. Esto facilita tanto el mantenimiento como la escalabilidad del prototipo, permitiendo sustituir modelos concretos o añadir nuevos módulos sin necesidad de rediseñar completamente el sistema.

Además, esta arquitectura modular mejora la interpretabilidad de los resultados obtenidos. A diferencia de modelos monolíticos de tipo *end-to-end*, donde resulta difícil justificar el motivo de una clasificación concreta, el sistema desarrollado permite identificar qué fases o características han contribuido en mayor medida a la puntuación final obtenida. De esta forma, es posible analizar por separado factores como la veracidad del contenido, el tono emocional o la presencia de patrones asociados a desinformación.

Una vez calculadas las puntuaciones parciales de cada fase, se aplica la combinación ponderada definida durante la etapa de diseño. La puntuación final del sistema se obtiene mediante la siguiente expresión:

$$Score = Context \cdot 0.10 + Verification \cdot 0.30 + Tone \cdot 0.20 + Patterns \cdot 0.40$$

Durante la integración también ha sido necesario contemplar posibles errores derivados de la utilización de múltiples modelos y APIs externas. Debido a la dependencia de servicios de terceros, pueden producirse situaciones en las que determinados módulos no consigan completar correctamente el análisis, ya sea por fallos de conexión, límites de uso de APIs o errores durante el procesamiento de la

información. En el estado actual del prototipo, se ha optado por priorizar la capacidad de depuración frente a la tolerancia automática a fallos. Por ello, cuando se produce un error durante la ejecución de alguno de los módulos, el sistema detiene el proceso y muestra información detallada sobre el origen del fallo. Este comportamiento facilita identificar problemas durante la fase de desarrollo y evaluar con precisión qué componentes presentan mayores limitaciones o inestabilidad.

Aunque esta aproximación reduce la robustez del sistema desde el punto de vista de producción, resulta adecuada dentro del contexto de un prototipo experimental orientado principalmente a validar la arquitectura propuesta y analizar el comportamiento de los distintos módulos de análisis.

4.7. INTEGRACIÓN EN PYTHON

La implementación del prototipo se ha realizado íntegramente en Python, utilizando una estructura modular alineada con las distintas fases descritas en el diseño del sistema. Esta organización permite separar claramente la lógica correspondiente a cada tipo de análisis, facilitando tanto el mantenimiento como la ampliación futura del proyecto.

La estructura general del repositorio se divide en distintos directorios asociados a las fases principales del sistema. Cada uno de estos directorios contiene los módulos necesarios para realizar una tarea concreta, como el análisis contextual, la verificación de información, el análisis de tono o la detección de patrones lingüísticos. Además, existe un módulo central encargado de coordinar la ejecución del sistema y combinar los resultados obtenidos por cada una de las fases.

El cálculo principal de la métrica se realiza desde el módulo central, que recibe el texto de la publicación y la información contextual asociada. Este módulo ejecuta secuencialmente las distintas fases de análisis, recoge las puntuaciones generadas por cada una de ellas y aplica posteriormente la fórmula de agregación definida durante la etapa de diseño.

Antes de ejecutar los modelos de análisis textual, el sistema aplica una fase común de traducción y preprocesado. Esta fase permite transformar el texto de entrada a inglés y eliminar elementos que pueden introducir ruido en el análisis, como menciones o enlaces. De esta forma, los modelos especializados reciben una entrada homogénea y compatible con las condiciones para las que fueron entrenados originalmente.

La implementación hace uso tanto de modelos descargados desde Hugging Face como de APIs externas accesibles mediante peticiones HTTP. Esto ha requerido integrar distintos tipos de interacción dentro del mismo flujo de ejecución. Mientras que algunos modelos se ejecutan localmente utilizando librerías de inferencia NLP, otros requieren llamadas externas y procesamiento posterior de respuestas en formato JSON.

Además del núcleo de cálculo, se ha desarrollado una aplicación web basada en Flask con el objetivo de facilitar el uso del prototipo. Esta aplicación permite introducir la URL de una publicación de X y ejecutar automáticamente el análisis completo del mensaje. La interfaz se encarga de extraer el contenido de la publicación, recopilar la información contextual asociada y mostrar finalmente tanto la puntuación global como las puntuaciones parciales obtenidas en cada fase.

Dadas las limitaciones actuales de acceso a la API oficial de X, especialmente tras la introducción de restricciones de pago para determinadas operaciones de extracción de datos, se ha optado por utilizar herramientas externas de código abierto para automatizar la recopilación de información necesaria para

el análisis. En concreto, se emplea la librería “Twscrape” [50] para obtener el contenido de las publicaciones a partir de su URL, mientras que la librería “Scweet” [51] se utiliza para recopilar información contextual asociada a la cuenta emisora, como publicaciones recientes, descripción del perfil o relaciones de seguimiento. El uso de estas herramientas permite automatizar completamente el flujo de análisis sin depender de servicios propietarios de pago, aunque introduce ciertas limitaciones relacionadas con estabilidad, mantenimiento y posibles cambios futuros en la plataforma.

En conjunto, la implementación en Python permite trasladar la arquitectura teórica planteada durante el diseño a un prototipo funcional, manteniendo una estructura modular, interpretable y fácilmente extensible para futuras versiones del sistema.

A continuación, se presentan las Figuras 4.2, 4.3 y 4.4, capturas de la versión final del prototipo, bautizado como “Bulometer”. El código completo del sistema se encuentra alojado en el repositorio del TFG[1].

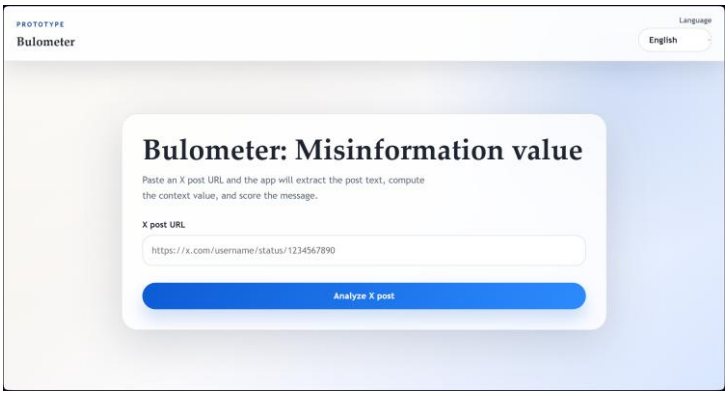


Figura 4.2 Pantalla de inicio

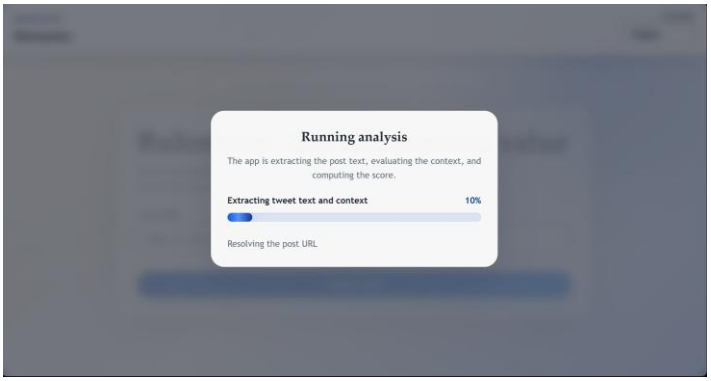


Figura 4.3 Pantalla de carga

[1] https://github.com/marcosistoocommon/misinfo_metric_tool

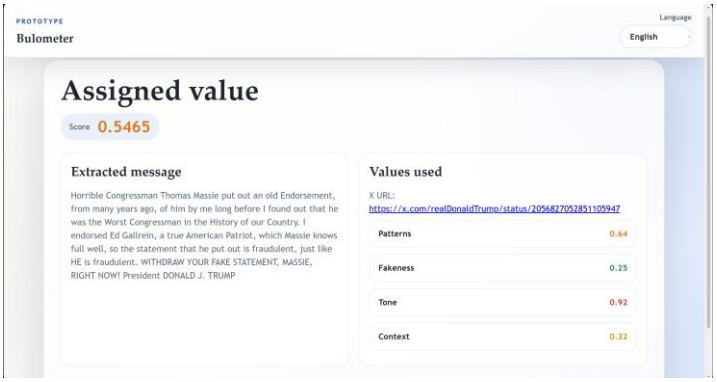


Figura 4.4 Pantalla de resultados

4.8. LIMITACIONES ACTUALES Y DEL SISTEMA

Aunque el prototipo desarrollado permite analizar publicaciones de redes sociales y generar métricas relacionadas con riesgo de desinformación, el sistema presenta limitaciones derivadas tanto de las características del problema abordado como de las decisiones técnicas adoptadas durante el desarrollo.

La principal limitación del sistema reside en la utilización de modelos preentrenados desarrollados por terceros. Debido a las restricciones temporales y de recursos asociadas al trabajo, no ha sido posible entrenar modelos propios ni realizar procesos avanzados de *fine-tuning* sobre *datasets* específicamente diseñados para este proyecto. Como consecuencia, el comportamiento de algunos módulos depende directamente de modelos entrenados para tareas similares, pero no necesariamente optimizados para la detección de desinformación en el contexto concreto planteado en este trabajo. Esto resulta especialmente relevante en tareas complejas y subjetivas como la detección de sesgo, propaganda, falacias o discursos de odio. Los modelos utilizados pueden incorporar sesgos presentes en sus *datasets* de entrenamiento originales, así como limitaciones relacionadas con idioma, contexto cultural o tipo de contenido analizado.

Otra limitación que tener en consideración es la dependencia de APIs y servicios externos. Parte del sistema requiere realizar llamadas a plataformas de terceros para obtener información contextual, buscar evidencias o ejecutar determinadas tareas de análisis. Esto introduce posibles problemas relacionados con latencia, estabilidad, límites de uso, costes económicos o cambios futuros en las políticas de acceso de dichos servicios. Esta dependencia externa afecta especialmente a los módulos de verificación de información y recuperación de contexto, ya que ambos necesitan acceso a búsquedas web o datos asociados a publicaciones reales. Además, determinadas funcionalidades de extracción de datos de X han pasado a estar limitadas mediante modelos de pago, obligando a utilizar herramientas alternativas de *scraping* cuya estabilidad puede verse afectada por cambios futuros en la plataforma.

Por otro lado, existen múltiples casos ambiguos que resultan especialmente complejos de analizar automáticamente, como mensajes irónicos, humorísticos, satíricos o altamente dependientes de contexto cultural y político. Este tipo de contenido puede provocar falsos positivos o clasificaciones inconsistentes, especialmente en módulos relacionados con emociones, propaganda o tono negativo. También debe tenerse en cuenta que el sistema ha sido diseñado principalmente para publicaciones en español e inglés. Aunque la fase de traducción permite ampliar parcialmente la compatibilidad lingüística, el rendimiento del sistema podría degradarse significativamente en otros idiomas o en publicaciones con expresiones altamente coloquiales, jerga específica o combinaciones de varios idiomas dentro del mismo mensaje.

El objetivo principal del trabajo ha sido validar una arquitectura modular orientada a la generación de métricas híbridas relacionadas con riesgo de desinformación, más que desarrollar un clasificador optimizado para producción. Por este motivo, los resultados obtenidos deben interpretarse como una aproximación experimental al problema y no como un sistema definitivo de detección automática de desinformación.

5. RESULTADOS Y VALIDACIÓN

5.1. METODOLOGÍA

Para evaluar el comportamiento del prototipo desarrollado, se han seleccionado tres casos de estudio representativos de distintos tipos de publicaciones presentes en redes sociales: una publicación periodística, un bulo actualmente desmentido y una publicación de carácter satírico o irónico.

La selección de estos casos permite analizar la respuesta del sistema ante escenarios con características diferenciadas en términos de veracidad, intención comunicativa y estilo discursivo. Para cada publicación se estudian los resultados obtenidos en las distintas fases de análisis descritas en capítulos anteriores, así como la puntuación final asignada por la herramienta.

Posteriormente, se realiza una comparación conjunta de los resultados obtenidos con el objetivo de identificar las capacidades actuales del prototipo, evaluar su coherencia frente a distintos tipos de contenido y detectar posibles limitaciones o aspectos susceptibles de mejora.

5.2. CASO 1: PUBLICACIÓN PERIODÍSTICA

En la Figura 5.1 se muestra el contenido seleccionado para este primer caso de estudio. Se ha optado por una noticia publicada por un medio de comunicación, cuyo titular presenta un carácter predominantemente informativo y neutro, con el objetivo de analizar la puntuación asignada por la herramienta a un contenido que, en principio, no presenta indicios de desinformación.

Este caso sirve como referencia para evaluar el comportamiento del sistema ante información legítima y contrastarla posteriormente con los resultados obtenidos en publicaciones desinformadoras y satíricas.



Figura 5.1 Publicación periodística elegida

Los resultados obtenidos tras el procesamiento de la noticia mediante el prototipo desarrollado se muestran en la Figura 5.2. En ella se recogen las puntuaciones asignadas en cada una de las fases de análisis, así como la puntuación final calculada por el sistema.

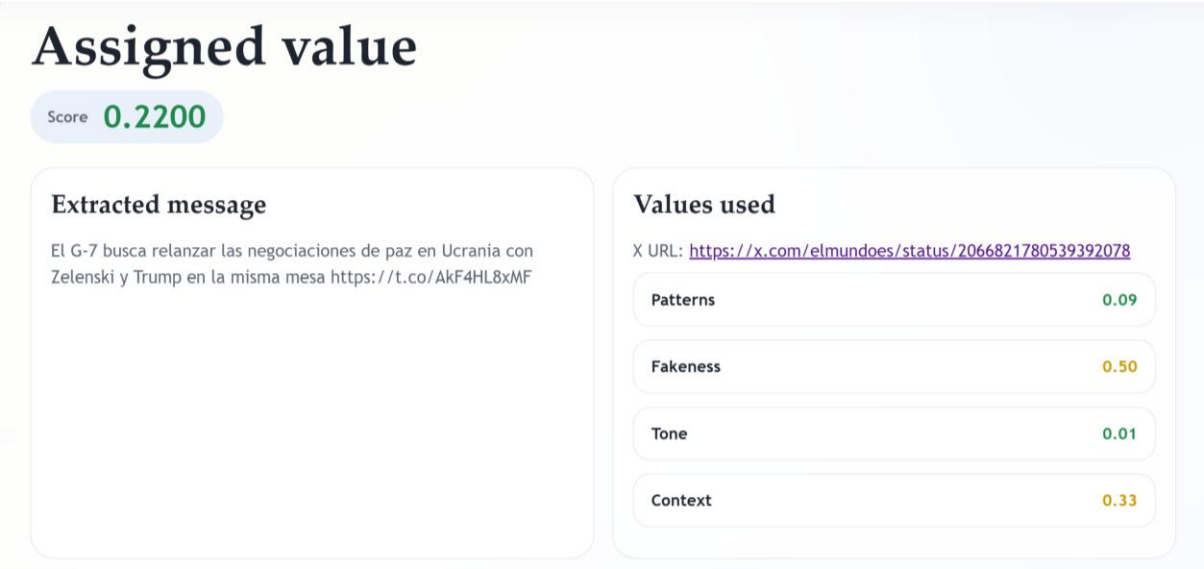


Figura 5.2 Resultados del análisis de la publicación periodística

Las Figura 5.3 muestra los desgloses de las variables "context" y "patterns" respectivamente.

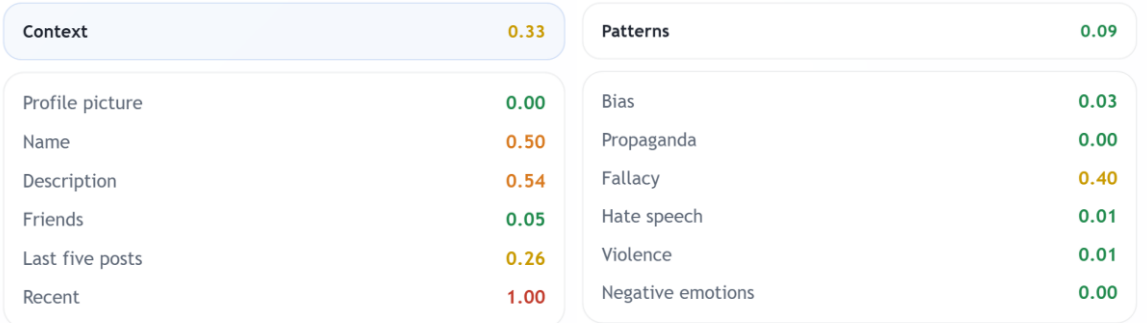


Figura 5.3 Desglose de "context" y "patterns"

La publicación obtiene 0.22/1, un resultado bastante bajo y que entra dentro de lo esperable, al tratarse de un texto periodístico. Se observa que, en línea con los criterios de selección, ha obtenido 0.09 puntos en la variable "patterns" y 0.01 puntos en la variable "tone". Esto es muy razonable puesto que se trata de un texto corto y con un lenguaje muy neutro.

La puntuación en el apartado de contexto resulta significativa, y es conveniente analizarla. Vemos que el nombre obtiene 0.5 puntos, lo cual tiene sentido según los criterios establecidos en el desarrollo del prototipo, pero en este caso se podría catalogar como falso positivo. Aún así, el peso que tiene en la puntuación final ese rasgo en concreto es muy bajo, 0.003, por lo que se consigue que no tenga una influencia significativa en el resultado final. Algo similar sucede con la descripción, pero al igual que en el caso anterior, el efecto sobre el total sigue siendo bajo: 0.01. El hecho es correctamente identificado como reciente, y vemos que las últimas 5 publicaciones han obtenido una puntuación de 0.26, bastante cercana a la puntuación obtenida por el post bajo análisis, lo que también es algo esperable.

La mayor anomalía aparece con la etiqueta “fakeness”. En esta se obtiene un 0.5, cuando lo esperable, al tratarse de un medio de comunicación, sería un valor muy cercano a 0. Estudiando los registros del prototipo se observa lo siguiente:

```
2026-06-16 12:15:11,347 - INFO - Extracted a total of 4 potential claims

…

2026-06-16 12:15:24,208 - INFO - Verdict 'VerificationResult.SUPPORTED' for 'The Group of Seven (G-7) seeks to relaunch peace negotiations in Ukraine'…

2026-06-16 12:15:35,032 - INFO - Verdict 'VerificationResult.SUPPORTED' for 'Volodymyr Zelensky is involved in the peace negotiations in Ukraine'…

2026-06-16 12:15:45,507 - INFO - Verdict 'VerificationResult.REFUTED' for 'Donald Trump is involved in the peace negotiations in Ukraine'…

2026-06-16 12:16:03,296 - INFO - Verdict 'VerificationResult.REFUTED' for 'The G-7 seeks to have Volodymyr Zelensky and Donald Trump at the same table for the peace negotiations in Ukraine'…
```

Se identifican 4 sentencias como *checkworthy*, pero solo 2 de ellas son marcadas como “SUPPORTED”. Respecto a las otras 2, marcadas como falsas, la justificación que otorga el prototipo es la siguiente:

```
2026-06-16 12:16:03,296 - INFO - Verdict 'VerificationResult.REFUTED' for 'The G-7 seeks to have Volodymyr Zelensky and Donald Trump at the same table for the peace negotiations in Ukraine': The evidence confirms that G7 leaders, including Trump and Zelensky, are meeting at the G7 summit to discuss Ukraine, but there is no indication that the G7 seeks to have both leaders at the same table specifically for peace negotiations. The sources describe a G7 session with Zelensky and Trump present, but do not state that the G7 is organizing or seeking a direct peace negotiation between Zelensky and Trump. The focus is on a unified G7 response and diplomatic discussions, not a bilateral peace negotiation table. (2 sources, 2 influential)

2026-06-16 12:15:45,507 - INFO - Verdict 'VerificationResult.REFUTED' for 'Donald Trump is involved in the peace negotiations in Ukraine': The evidence from both Politico and Kyiv Post indicates that while Donald Trumphas expressed interest in settling the Ukraine war and has discussed theissue with other leaders, he is not directly involved in the peace negotiations. In fact, Trump has publicly supported direct talks between Ukraine and Russia without initial US mediation, and there is no evidence in the sources that he is actively participating in or mediating the negotiations himself. (2 sources, 2 influential)
```

Analizando las justificaciones generadas por el sistema, se observa que ambas verificaciones negativas no se deben a la detección de información objetivamente falsa, sino a una discrepancia entre la afirmación extraída y la evidencia encontrada durante el proceso de búsqueda.

En ambos casos, la herramienta interpreta las afirmaciones de forma especialmente restrictiva. La participación de Donald Trump en las conversaciones sobre la guerra de Ucrania es reinterpretada como una participación directa en las negociaciones de paz, mientras que la afirmación relativa a la reunión entre Trump y Zelensky es evaluada exigiendo una evidencia explícita de que dicha reunión tenga como finalidad exclusiva una negociación de paz. Ante la ausencia de pruebas que confirmen exactamente esas formulaciones, el sistema opta por clasificarlas como falsas.

Este comportamiento pone de manifiesto una de las principales limitaciones de los sistemas automáticos de verificación basados en recuperación de información y modelos de lenguaje: la dificultad para diferenciar entre afirmaciones falsas y afirmaciones insuficientemente respaldadas por las fuentes recuperadas. En este caso, el error no parece derivar de la noticia original, sino de una interpretación excesivamente estricta de las afirmaciones generadas durante el proceso de extracción y verificación.

En conjunto, los resultados obtenidos pueden considerarse satisfactorios. A pesar de los falsos positivos observados en las fases de contexto y verificación, la publicación recibe una puntuación final reducida (0.22), coherente con su carácter informativo. Además, el análisis realizado permite identificar con

claridad los factores responsables de la desviación observada, lo que facilita la detección de posibles líneas de mejora para futuras versiones del prototipo.

Este primer caso de estudio sugiere que la herramienta es capaz de identificar correctamente publicaciones periodísticas de tono neutro, aunque sigue mostrando sensibilidad a errores derivados de la extracción automática de afirmaciones y de las limitaciones inherentes a los mecanismos de verificación empleados.

5.3. CASO 2: PUBLICACIÓN DESINFORMADORA

En la Figura 5.4 se muestra el contenido seleccionado para este segundo caso de estudio. Se ha optado por una noticia falsa previamente desmentida por organismos de verificación, cuyo titular presenta un carácter sensacionalista e informal. La elección de este caso permite evaluar el comportamiento del prototipo ante un ejemplo representativo de contenido desinformador.

A diferencia del caso anterior, se espera que la publicación obtenga una puntuación significativamente superior, especialmente en las fases de verificación y detección de patrones, debido tanto a la falsedad de las afirmaciones contenidas en el mensaje como a la presencia de elementos discursivos frecuentemente asociados a la desinformación.

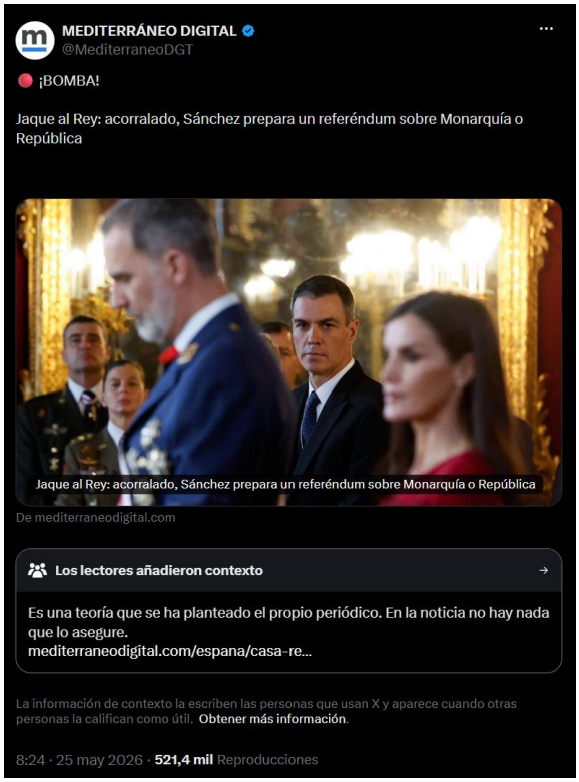


Figura 5.4 Publicación desinformadora elegida

Los resultados obtenidos tras el procesamiento de la noticia mediante el prototipo desarrollado se muestran en la Figura 5.5. En ella se recogen las puntuaciones asignadas en cada una de las fases de análisis, así como la puntuación final calculada por el sistema.

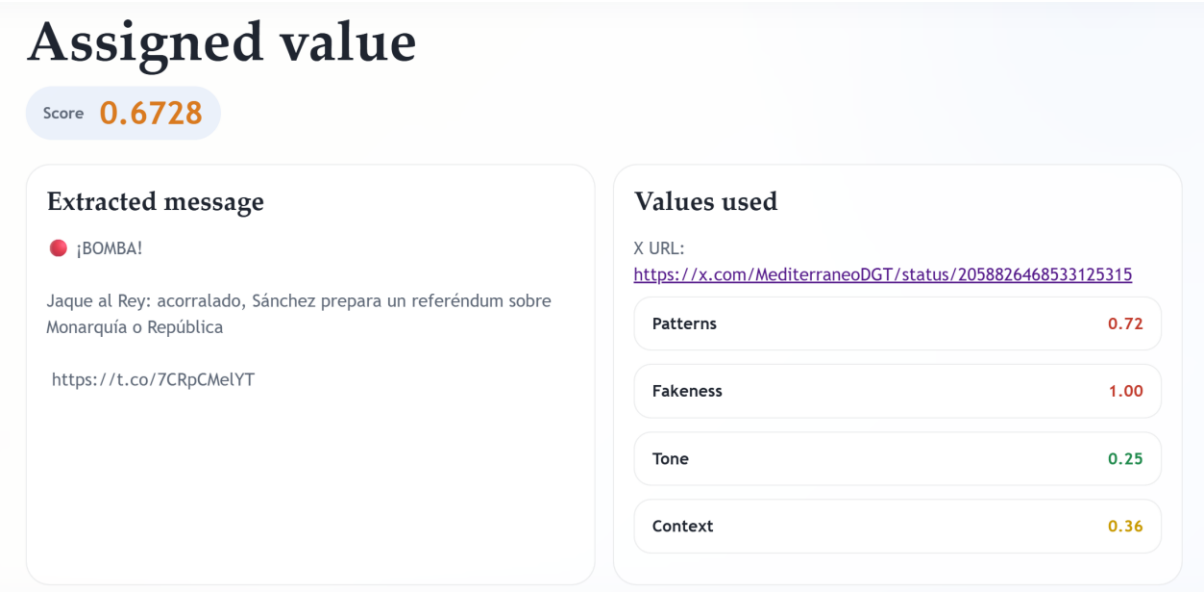


Figura 5.5 Resultados del análisis de la publicación desinformadora

La Figura 5.6 muestra los desgloses de las variables "context" y "patterns" respectivamente.

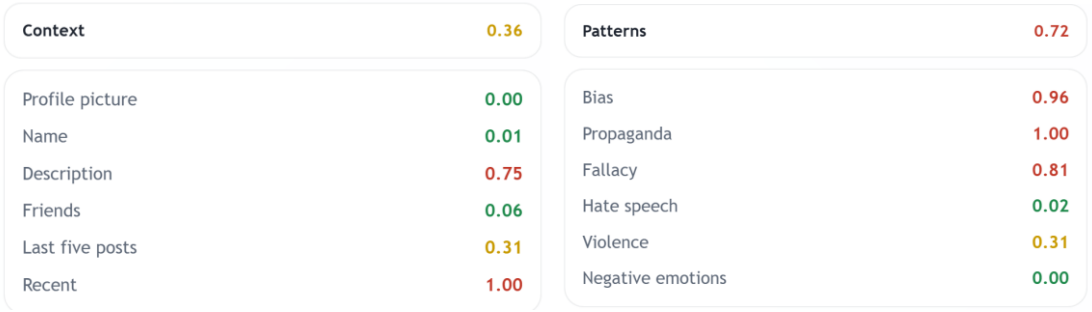


Figura 5.6 Desglose de "context" y "patterns"

La publicación en este caso obtiene un 0.67/1, un resultado alto que parece acorde al tipo de publicación estudiada. Como es esperable, en este caso "fakeness" tiene su valor máximo, puesto que se trata de una noticia falsa. También se observa que "patterns" tiene un valor muy elevado, de 0.72, a pesar de no ser un texto muy violento. Este resultado sugiere que la distribución de pesos adoptada resulta adecuada para priorizar rasgos habitualmente asociados a la desinformación frente a otros indicadores menos representativos, como la presencia de lenguaje agresivo o violento.

En el caso del contexto, se aprecia que los mayores valores son "Hecho reciente", "Descripción" y "5 últimos posts". La diferencia con el caso anterior es que aquí la puntuación de los últimos posts es menos de la mitad de la puntuación obtenida en la publicación bajo análisis. Este comportamiento resulta especialmente interesante, ya que pone de manifiesto la utilidad de incorporar información contextual al análisis. Mientras que la publicación concreta presenta numerosos indicadores asociados a la desinformación, la actividad reciente de la cuenta muestra valores considerablemente más bajos. Esto permite distinguir entre una cuenta cuya actividad habitual está dominada por contenidos potencialmente problemáticos y una cuenta que, de forma puntual, ha difundido una publicación de estas características.

Por último, se aprecia que la puntuación del tono es bastante baja. Esto se debe a que, aunque el lenguaje sea informal, no es especialmente negativo ni violento. Este valor bajo contribuye a moderar la puntuación final: con un valor de "tone" de 0.72 (valor obtenido en patterns), la puntuación final subiría casi 1 décima, llegando a 0.768.

En conjunto, los resultados obtenidos son coherentes con la naturaleza de la publicación analizada. La herramienta asigna una puntuación elevada (0.67), impulsada principalmente por la detección de afirmaciones falsas y por la presencia de patrones lingüísticos asociados a la desinformación. Al mismo tiempo, la baja puntuación obtenida en la fase de tono evita una sobreestimación del riesgo basada únicamente en el carácter informal o sensacionalista del mensaje. Este caso de estudio sugiere que el prototipo es capaz de identificar correctamente contenidos desinformadores incluso cuando estos no presentan niveles especialmente elevados de agresividad o lenguaje negativo.

5.4. CASO 3: PUBLICACIÓN SATÍRICA

La Figura 5.7 muestra la publicación elegida para este último caso de estudio. Se ha seleccionado una publicación de carácter irónico, procedente de una cuenta dedicada a la difusión de contenido humorístico con formato y estilo periodístico. Con este caso se pretende evaluar cómo reacciona el prototipo ante contenidos satíricos, determinando si es capaz de distinguirlos de la desinformación o si, por el contrario, los identifica erróneamente como publicaciones engañosas.

Al tratarse de un contenido deliberadamente falso, pero cuya finalidad no es informar ni manipular al lector, se espera un comportamiento intermedio entre los dos casos anteriores. Este escenario resulta especialmente relevante, ya que constituye uno de los principales desafíos para los sistemas automáticos de detección de desinformación.



Figura 5.7 Publicación satírica elegida

Los resultados obtenidos tras el procesamiento de la publicación mediante el prototipo desarrollado se muestran en la Figura 5.8. En ella se recogen las puntuaciones asignadas en cada una de las fases de análisis, así como la puntuación final calculada por el sistema.

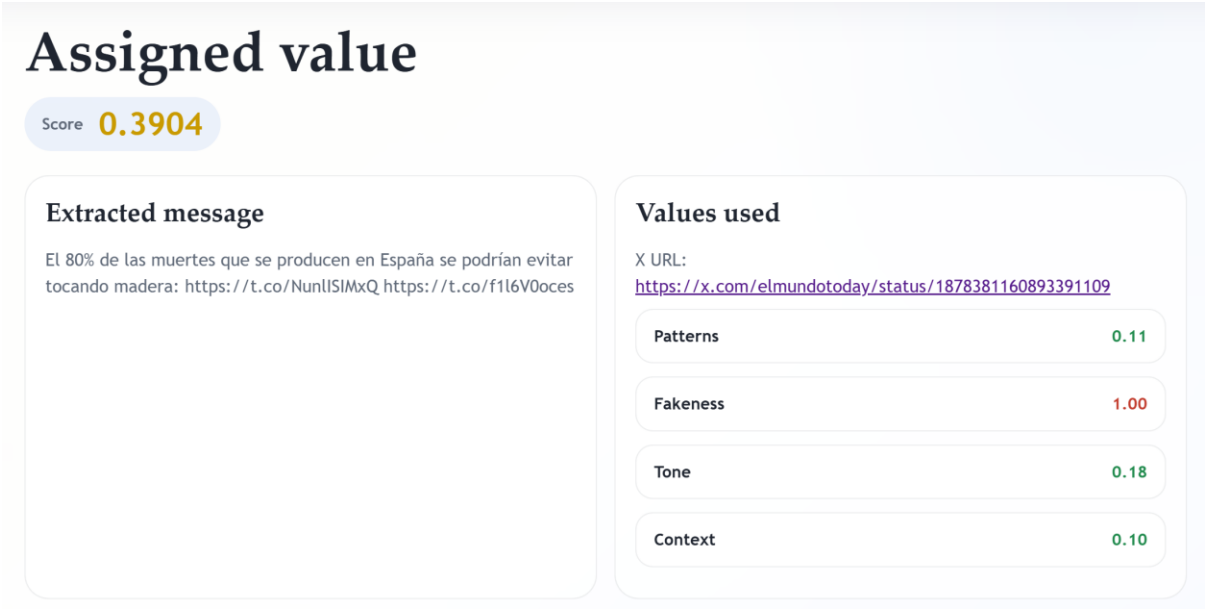


Figura 5.8 Resultados del análisis de la publicación satírica

La Figura 5.9 muestra los desgloses de las variables "context" y "patterns" respectivamente.

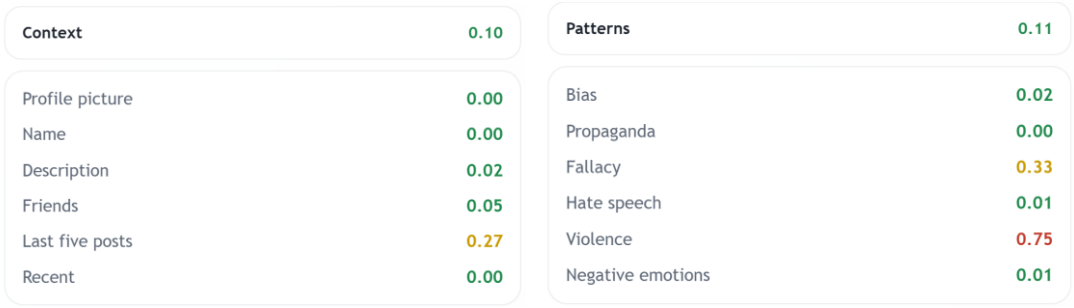


Figura 5.9 Desglose de "context" y "patterns"

La publicación obtiene una puntuación de 0.39/1. Este resultado puede considerarse coherente con la naturaleza del contenido analizado. Aunque no se trata de una publicación desinformadora, ya que presenta un carácter claramente humorístico y satírico, sí contiene información objetivamente falsa, lo que justifica una puntuación superior a la obtenida en el caso de la publicación periodística.

Resulta especialmente relevante observar que 0.3 de los 0.39 puntos obtenidos (aproximadamente un 77 % de la puntuación final) proceden de la variable "fakeness". Este resultado pone de manifiesto la importancia de no limitar el análisis únicamente a la veracidad del contenido, sino de incorporar factores adicionales que permitan diferenciar entre información falsa difundida con intención humorística y publicaciones diseñadas para desinformar.

En el resto de variables se observan puntuaciones reducidas. La puntuación obtenida en la fase de "context" sugiere que la actividad reciente de la cuenta no presenta niveles elevados de riesgo según los criterios definidos en el modelo. Del mismo modo, la puntuación obtenida en los cinco últimos mensajes es considerablemente inferior a la obtenida por la publicación analizada, lo que indica que este tipo de contenido no constituye necesariamente la totalidad de la actividad desarrollada por la cuenta.

El valor obtenido en “patterns” resulta también significativo. La mayoría de los subparámetros presentan puntuaciones próximas a cero, un resultado coherente con las características del texto analizado. La mayor puntuación corresponde a la variable “violence”, aunque su contribución representa únicamente alrededor del 4 % de la puntuación final. Este resultado parece estar motivado por la presencia de términos asociados a situaciones violentas, como la palabra “muertes”, lo que evidencia que los modelos de NLP, pese a ofrecer un rendimiento notable, siguen siendo susceptibles a falsos positivos derivados del contexto semántico de determinadas expresiones.

En conjunto, los resultados obtenidos pueden considerarse satisfactorios. La publicación recibe una puntuación intermedia, superior a la asignada al contenido periodístico pero claramente inferior a la obtenida por la publicación desinformadora. Este comportamiento sugiere que el enfoque multifactorial adoptado permite distinguir entre la mera falsedad de un contenido y la presencia de características habitualmente asociadas a la desinformación. Asimismo, este caso de estudio pone de manifiesto tanto la utilidad de combinar distintas métricas de análisis como las limitaciones actuales de los modelos empleados para interpretar contenidos satíricos y humorísticos.

5.5. COMPARATIVA DE RESULTADOS

En la Tabla 5.1 y la Figura 5.10 se expone una comparación de los resultados obtenidos en estos 3 casos de estudio.

Publicación	Patterns	Fakeness	Tone	Context	Score
Publicación periodística	0.09	0.5	0.01	0.33	0.22
Publicación satírica	0.11	1.0	0.18	0.1	0.39
Publicación desinformadora	0.72	1.0	0.25	0.36	0.67

Tabla 5.1 Tabla comparativa de los resultados

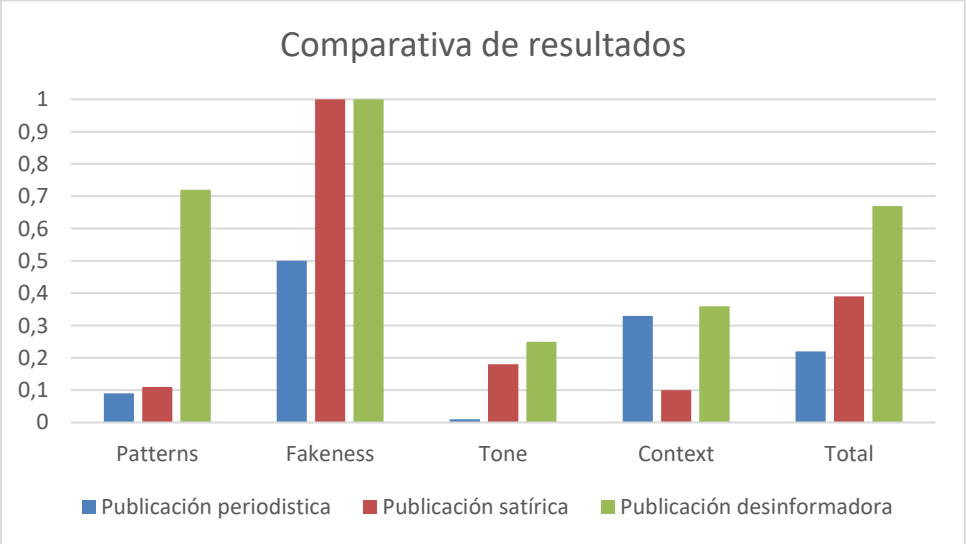


Figura 5.10 Gráfico comparativo de los resultados

El sistema tiene un comportamiento que entra dentro de lo esperado. La publicación periodística es la que menos puntuación obtiene, seguida de la publicación satírica y, con la mayor puntuación de las 3, se encuentra la publicación desinformadora.

Aunque tanto la publicación satírica como la desinformadora contienen afirmaciones falsas, la diferencia entre sus puntuaciones finales indica que el prototipo no basa su evaluación exclusivamente en la veracidad del contenido. La incorporación de variables relacionadas con el tono, los patrones lingüísticos y el contexto permite obtener una valoración más completa del riesgo asociado a cada publicación.

Analizando individualmente la puntuación obtenida en cada parámetro, se aprecia la importancia del parámetro “patterns” a la hora de discernir entre información meramente falsa y desinformación. Esta es la mayor diferencia que existe entre las 3 publicaciones, y constituye el principal factor diferenciador entre la publicación satírica y la desinformadora. Resulta llamativo en estudio del parámetro “fakeness”, pues las 3 publicaciones han obtenido una puntuación mayor de 0.5. Esto, como ya se explicó antes, se debe en el caso de la publicación periodística a las imprecisiones que van asociadas al uso de verificación automatizada. En las otras 2 publicaciones se obtiene un valor esperable en este parámetro.

Respecto a “context” y “tone”, se puede comprobar lo que ya se razonó en el capítulo 3: son rasgos que pueden aportar más información, pero no son los más determinantes. Ninguna puntuación supera los 0.3 puntos en “tone”, y en el caso del parámetro “context”, resulta curioso que el medio periodístico obtenga una puntuación mayor que la cuenta satírica. Esto es debido principalmente a que la publicación hace referencia a un hecho reciente, y la cuenta del medio de comunicación contiene referencias políticas en su descripción. Igualmente, la mayor puntuación la obtiene la cuenta autora de la publicación desinformadora, y esta no supera los 0.4.

Por lo tanto, se puede concluir que el planteamiento teórico y su implementación han proporcionado resultados razonables. La separación obtenida entre las tres categorías analizadas sugiere que el sistema es capaz de diferenciar entre información legítima, contenido satírico y publicaciones desinformadoras, validando así el interés de combinar métricas de verificación, contexto, tono y patrones lingüísticos.

Si bien persisten limitaciones asociadas a los modelos empleados y a los mecanismos automáticos de verificación, los resultados obtenidos respaldan la viabilidad de la arquitectura propuesta como herramienta de apoyo para el análisis de desinformación en redes sociales.

6. CONCLUSIONES Y LÍNEAS FUTURAS

6.1. CONCLUSIONES

El presente TFG tiene como principal objetivo investigar la aplicación de la IA y de las técnicas de NLP al análisis de desinformación en redes sociales, así como definir una métrica de riesgo de desinformación basada en múltiples factores y desarrollar un prototipo funcional capaz de implementarla.

Para alcanzar este objetivo, se realizó un estudio de la desinformación, sus principales características y las técnicas actualmente empleadas para su detección y mitigación. Este análisis permitió obtener una visión global del problema e identificar aquellos factores que pueden resultar relevantes para la caracterización de la desinformación. Posteriormente, dichos factores fueron analizados de forma individual para determinar su relevancia y su posible contribución dentro de un modelo de evaluación conjunto.

Como resultado de este proceso se definió el modelo teórico propuesto en este TFG, basado en la combinación de diferentes dimensiones de análisis relacionadas con el contexto de la publicación, la verificación factual, el tono empleado y la presencia de patrones lingüísticos asociados a la desinformación. A partir de estas dimensiones se estableció una métrica capaz de integrar la información procedente de distintos indicadores en una única puntuación de riesgo. Con el fin de validar este planteamiento, se desarrolló una implementación modular que integra diferentes herramientas y modelos especializados para el análisis de cada uno de los parámetros considerados. La evaluación realizada sobre publicaciones periodísticas, satíricas y desinformadoras permitió comprobar que el sistema genera resultados coherentes con la naturaleza de los contenidos analizados y que la incorporación de factores adicionales a la veracidad contribuye a diferenciar con mayor precisión entre información falsa y desinformación. El prototipo resultante es capaz de analizar publicaciones de la red social X y asignar una puntuación de riesgo coherente con el contenido evaluado.

Los resultados obtenidos sugieren que la veracidad de una publicación, aunque constituye un elemento fundamental, no resulta suficiente por sí sola para caracterizar adecuadamente la desinformación. La incorporación de factores adicionales relacionados con la intención comunicativa, el lenguaje empleado y los mecanismos de manipulación permite obtener una evaluación más completa del riesgo asociado a cada publicación y diferenciar con mayor precisión entre contenidos falsos, satíricos o potencialmente desinformadores.

La principal aportación de este trabajo reside en la propuesta de una métrica híbrida capaz de combinar múltiples indicadores procedentes de diferentes técnicas de análisis. Este enfoque permite superar algunas de las limitaciones de los sistemas centrados exclusivamente en la verificación factual y ofrece una aproximación más amplia al problema de la desinformación en redes sociales.

Aunque el sistema desarrollado constituye un prototipo y no está concebido para sustituir la supervisión humana ni para su despliegue inmediato a gran escala, los resultados obtenidos demuestran la viabilidad del enfoque propuesto y su potencial como herramienta de apoyo para la detección temprana de desinformación.

6.2. LÍNEAS FUTURAS

Si bien los resultados obtenidos permiten validar la viabilidad del enfoque propuesto, existen diversas líneas de investigación y desarrollo que podrían contribuir a mejorar tanto la precisión del modelo teórico como las capacidades del prototipo implementado:

- Desarrollo y entrenamiento de modelos propios. Aunque se han seleccionado modelos especializados para cada una de las tareas de análisis, todos ellos han sido desarrollados por terceros. Esto implica una dependencia de criterios de entrenamiento ajenos al objetivo específico de este trabajo, pudiendo introducir sesgos o limitaciones que afecten al rendimiento global del sistema. El desarrollo de modelos entrenados específicamente para la detección de desinformación permitiría obtener soluciones más adaptadas al problema abordado.
- Ajuste automático de pesos mediante técnicas de aprendizaje automático. Los pesos asignados a las distintas fases y parámetros del modelo han sido definidos a partir de criterios teóricos derivados del estudio realizado. Sin embargo, estos valores podrían no representar la configuración óptima. La utilización de técnicas de aprendizaje automático sobre conjuntos de datos etiquetados permitiría optimizar dichos pesos de forma objetiva y mejorar la capacidad predictiva del sistema.
- Incorporación de análisis multimodal. El modelo desarrollado se centra exclusivamente en publicaciones textuales. Sin embargo, una parte importante de la desinformación actual se difunde mediante imágenes, vídeos o combinaciones de distintos formatos. La incorporación de mecanismos capaces de analizar y evaluar este tipo de contenidos permitiría ampliar significativamente el alcance del sistema.
- Desarrollo de una solución escalable. El prototipo implementado presenta limitaciones derivadas de la utilización de APIs externas y de la necesidad de emplear técnicas de *scraping* para la obtención de información. Una futura implementación orientada a entornos de producción debería reducir estas dependencias, optimizar el rendimiento y garantizar la capacidad de procesar grandes volúmenes de publicaciones de forma eficiente.
- Ampliación a otras plataformas. Aunque el trabajo se ha centrado en la red social X, muchos de los mecanismos de desinformación identificados también están presentes en otras plataformas. La adaptación del modelo a entornos como Instagram, TikTok o Facebook permitiría analizar su capacidad de generalización y ampliar considerablemente sus ámbitos de aplicación.

En conjunto, las líneas futuras planteadas evidencian el potencial de evolución del modelo propuesto. La mejora de los mecanismos de análisis, la ampliación de su ámbito de aplicación y la reducción de las limitaciones actuales permitirían incrementar significativamente la utilidad práctica del sistema y su capacidad para afrontar escenarios reales de detección de desinformación.

BIBLIOGRAFÍA

[1] T. Scantamburlo and F. Zollo, "A Scoping Review of Misinformation Research: Definitions, Conceptualizations, and Measurement Approaches," *ACM Transactions on the Web*, Feb. 2026, doi: 10.1145/3796546.

[2] G. Murphy *et al.*, "What do we study when we study misinformation? A scoping review of experimental research (2016-2022)," *Harvard Kennedy School Misinformation Review*, vol. 4, no. 6, Nov. 2023, doi: 10.37016/mr-2020-130.

[3] E. Denniss and R. Lindberg, "Social Media and the Spread of misinformation: Infectious and a Threat to Public Health," *Health Promotion International*, vol. 40, no. 2, pp. 1–10, Mar. 2025, doi: 10.1093/heapro/daaf023.

[4] J. Lühring, A. Shetty, C. Koschmieder, D. Garcia, A. Waldherr, and H. Metzler, "Emotions in misinformation studies: distinguishing affective state from emotional response and misinformation recognition from acceptance," *Cognitive Research Principles and Implications*, vol. 9, no. 1, Dec. 2024, doi: 10.1186/s41235-024-00607-0.

[5] C. Vaccari and A. Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society*, vol. 6, no. 1, Feb. 2020, doi: 10.1177/2056305120903408.

[6] A. Iqbal, K. Shahzad, Shakeel Ahmad Khan, and Muhammad Shahzad Chaudhry, "The relationship of artificial intelligence (AI) with fake news detection (FND): a systematic literature review," *Global knowledge, memory and communication*, Oct. 2023, doi: 10.1108/gkmc-07-2023-0264.

[7] S. Abdali, S. Shaham, and Bhaskar Krishnamachari, "Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–29, Oct. 2024, doi: 10.1145/3697349.

[8] P. Juneja and T. Mitra, "Human and Technological Infrastructures of Fact-checking," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–36, Nov. 2022, doi: 10.1145/3555143.

[9] M.-Á. Chaparro-Domínguez, M. Mauri-de-los-Ríos, and R. Rodríguez-Martínez, "Plataformas de fact-checking y confianza: análisis de sus relaciones con las fuentes y audiencias," *index.comunicación*, vol. 14, no. 2, pp. 165–189, Jul. 2024, doi: 10.62008/ixc/14/02plataf.

[10] P. VAJDEČKA. *Abstractive summarization of fact check reports with pre-trained transformer tuning on extractive summaries*. Online. Master's thesis. Praha: University of Economics, Prague. 2022. Available from: https://theses.cz/id/5d9sbi/.

[11] K. Coleman, "Introducing Birdwatch, a community-based approach to misinformation," *blog.x.com*, Jan. 25, 2021. https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

[12] Meta, "Testing Begins for Community Notes on Facebook, Instagram and Threads," *Meta*, Mar. 13, 2025. https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/.

[13] E. Roth, "TikTok videos are about to get crowdsourced fact checks on them," *The Verge*, Jul. 30, 2025. https://www.theverge.com/news/715798/tiktok-videos-footnotes-crowdsourced-fact-checks-launch.

[14] M. Saeed, N. Traub, M. Nicolas, G. Demartini, and P. Papotti, "Crowdsourced Fact-Checking at Twitter," *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, Oct. 2022, doi: 10.1145/3511808.3557279.

[15] F. Panizza, P. Ronzani, T. Morisseau, S. Mattavelli, and C. Martini, "How do online users respond to crowdsourced fact-checking?," *Humanities and Social Sciences Communications*, vol. 10, no. 1, Nov. 2023, doi: 10.1057/s41599-023-02329-y.

[16] D. J. Flynn, B. Nyhan, and J. Reifler, "The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs about Politics," *Political Psychology*, vol. 38, no. S1, pp. 127–150, Jan. 2017, doi: 10.1111/pops.12394.

[17] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A Survey on Automated Fact-Checking," *arXiv.org*, Jun. 06, 2022. https://arxiv.org/abs/2108.11896.

[18] W. Wang, "Fact-checking as a multi-step process: from ambiguity resolution to claim validation," *Digital Commons @ NJIT*, 2025. https://digitalcommons.njit.edu/dissertations/1840.

[19] Z. Liu, T. Zhang, K. Yang, P. Thompson, Z. Yu, and S. Ananiadou, "Emotion detection for misinformation: A review," *Information Fusion*, vol. 107, pp. 102300–102300, Feb. 2024, doi: 10.1016/j.inffus.2024.102300.

[20] S. Peng *et al.*, "A survey on deep learning for textual emotion analysis in social networks," *Digital Communications and Networks*, Oct. 2021, doi: 10.1016/j.dcan.2021.10.003.

[21] A. Alslaity & R. Orji (2022). *Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions*. Behaviour & Information Technology. https://doi.org/10.1080/0144929X.2022.2156387.

[22] T. Widmann and M. Wich, "Creating and Comparing Dictionary, Word Embedding, and Transformer-based Models to Measure Discrete Emotions in German Political Text," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4127133.

[23] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face," *Entropy*, vol. 25, no. 10, p. 1440, Oct. 2023, doi: 10.3390/e25101440.

[24] A. S. Hosseini and S. Staab, "Emotional Framing in the Spreading of False and True Claims," *Proceedings of the 15th ACM Web Science Conference 2023*, pp. 96–106, Apr. 2023, doi: 10.1145/3578503.3583611.

[25] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.

[26] F. Monti *et al.*, "Fake News Detection on Social Media using Geometric Deep Learning," 2019. Available: https://arxiv.org/pdf/1902.06673

[27] S. Kh. Hamed, M. J. A. Ab Aziz, and M. R. Yaakub, “Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users’ Comments,” *Sensors*, vol. 23, no. 4, p. 1748, Feb. 2023, doi: 10.3390/s23041748.

[28] A. A. Maruf, F. Khanam, M. M. Haque, Z. M. Jiyad, M. F. Mridha and Z. Aung, "Challenges and Opportunities of Text-Based Emotion Detection: A Survey," in *IEEE Access*, vol. 12, pp. 18416-18450, 2024, doi: 10.1109/ACCESS.2024.3356357.

[29] Y. Ortiz-Martínez, J. Sarmiento, D. K. Bonilla-Aldana, and A. J. Rodríguez-Morales, “Monkeypox goes viral: measuring the misinformation outbreak on Twitter,” *The Journal of Infection in Developing Countries*, vol. 16, no. 07, pp. 1218–1220, Jul. 2022, doi: 10.3855/jidc.16907.

[30] R. Chen, G. Chen, L. Zhang, R. Xie, and R. Chen, “An analysis of the factors influencing engagement metrics within the dissemination of health science misinformation,” *Frontiers in Public Health*, vol. 13, Jun. 2025, doi: 10.3389/fpubh.2025.1571210.

[31] J. Számely, A. Galeazzi, J. Koltai, and E. Omodei, “Easy-access online social media metrics are associated with misinformation sharing activity,” *Scientific Reports*, vol. 15, no. 1, Nov. 2025, doi: 10.1038/s41598-025-25049-6.

[32] F. Germano, V. Gómez, and F. Sobbrio, “Ranking for engagement: How social media algorithms fuel misinformation and polarization,” *Journal of Public Economics*, vol. 255, p. 105589, Mar. 2026, doi: 10.1016/j.jpubeco.2026.105589.

[33] R. S. Pathak, F. Spezzano, and Maria Soledad Pera, “Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks,” *ACM Transactions on the Web*, vol. 17, no. 4, Aug. 2023, doi: 10.1145/3616088.

[34] R. Maertens *et al.*, “The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment,” *Behavior Research Methods*, vol. 56, Jun. 2023, doi: 10.3758/s13428-023-02124-2.

[35] R. Rosário *et al.*, “Development of an Instrument to Measure Resilience to Misinformation on Social Media: Measurement Properties and Validation,” *Journal of Medical Internet Research*, vol. 27, pp. e72449–e72449, Sep. 2025, doi: 10.2196/72449.

[36] Y. Green, A. Gully, A. Roy, Y. Roth, J. Tucker, and A. Wanless, “Evidence-Based Misinformation Interventions: Challenges and Opportunities for Measurement and Collaboration.” Available: https://www.gisf.ngo/wp-content/uploads/2023/01/202212-Wanless_et_al_Misinfo_Interventions2.pdf

[37] H. Zubairu, A. Abdou, and A. Matrawy, “Evaluation Metrics for Misinformation Warning Interventions: Challenges and Prospects,” *arXiv.org*, 2025. https://arxiv.org/abs/2505.09526

[38] T. Scantamburlo and F. Zollo, “A Scoping Review of Misinformation Research: Definitions, Conceptualizations, and Measurement Approaches,” *ACM Transactions on the Web*, Feb. 2026, doi: 10.1145/3796546.

[39] B. G. SOUTHWELL, S. BABWAH, R. PAQUIN, V. BOUDEWYNS, and J. ZENG, “Defining and Measuring Scientific Misinformation,” *The Annals of the American Academy of Political and Social Science*, vol. 700, pp. 98–111, 2022, doi: 10.2307/27253358.

[40] "Portal Transparencia | Trabajos del foro contra las campañas de desinformación. Iniciativas 2024," *Transparencia.gob.es*, 2024. https://transparencia.gob.es/masinformacion/informes-de-interes/seguridad/campa-asdesinformacion.

[41] Helsinki-NLP, "opus-mt-es-en," Hugging Face, [Online]. Available: https://huggingface.co/Helsinki-NLP/opus-mt-es-en.

[42] M. Burnham, K. Kahn, R. Y. Wang, and R. X. Peng, "Political DEBATE: Efficient Zero-shot and Few-shot Classifiers for Political Text," *arXiv:2409.02078*, Sep. 2024.

[43] BharathxD, "GitHub - BharathxD/ClaimeAI" *GitHub*, 2025. https://github.com/BharathxD/ClaimeAI .

[44] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, 2020, pp. 1644–1650.

[45] H. Ghosh, A. Mosharafa, and G. Groh, "To Bias or Not to Bias: Detecting Bias in News with bias-detector," *arXiv preprint arXiv:2505.13010*, 2025.

[46] "Tanbih news project," *Tanbih.org*, 2019. https://apihub.tanbih.org/.

[47] Z. Jin *et al.*, "Logical Fallacy Detection," *arXiv:2202.13758 [cs]*, May 2022, Available: https://arxiv.org/abs/2202.13758

[48] S. Tulkens and T. van Dongen, *Model2Vec: Fast State-of-the-Art Static Embeddings*. Zenodo, 2024. [Online]. Available: https://github.com/MinishLab/model2vec. doi: 10.5281/zenodo.17270888.

[49] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, and M. V. Martínez, "pysentimiento: A Python Toolkit for Opinion Mining and Social NLP Tasks," *arXiv preprint arXiv:2106.09462*, 2023.

[50] Vladkens, *twscrape*. GitHub repository. [Online]. Available: https://github.com/vladkens/twscrape.

[51] Altimis, *Scweet*. GitHub repository. [Online]. Available: https://github.com/Altimis/Scweet.

ANEXO A: ASPECTOS ÉTICOS, ECONÓMICOS, SOCIALES Y AMBIENTALES

A.1 INTRODUCCIÓN

El presente TFG aborda el problema de la desinformación en redes sociales mediante el desarrollo de una métrica orientada a estimar el riesgo de desinformación asociado a una publicación, así como de un prototipo funcional que implementa dicha métrica para la red social X.

El objetivo principal del proyecto es investigar la aplicación de técnicas de IA y NLP al análisis de desinformación, diseñar un modelo teórico para la obtención de una métrica de riesgo y desarrollar un prototipo capaz de calcular dicha métrica de forma automática. Para ello, se emplea una arquitectura modular basada en distintos modelos especializados encargados de analizar aspectos relacionados con la veracidad, el contexto, el tono y determinados patrones lingüísticos presentes en las publicaciones.

Desde una perspectiva social, el proyecto pretende contribuir al desarrollo de herramientas que faciliten la identificación de contenidos potencialmente desinformadores. Desde el punto de vista económico, puede ayudar a optimizar los recursos destinados a tareas de verificación y moderación de contenidos. Asimismo, el uso de sistemas automatizados para analizar información plantea importantes cuestiones éticas relacionadas con la transparencia, la justificación de los resultados y la posible presencia de sesgos en los modelos utilizados. Finalmente, aunque el impacto ambiental directo es reducido, el uso de modelos de IA implica un consumo de recursos computacionales que debe ser considerado.

A.2 DESCRIPCIÓN DE IMPACTOS RELEVANTES RELACIONADOS CON EL PROYECTO

Durante el desarrollo del proyecto se identificaron diversos impactos potenciales asociados a la utilización de herramientas automáticas para el análisis de desinformación. Los aspectos considerados más relevantes se resumen a continuación:

- Impacto social (alto, positivo). El sistema desarrollado puede servir como herramienta de apoyo para la detección temprana de publicaciones potencialmente desinformadoras. Su utilización podría facilitar el trabajo de verificadores, investigadores y plataformas digitales, contribuyendo a mejorar la calidad de la información disponible para los usuarios.
- Impacto económico (medio, positivo). La automatización parcial de determinadas tareas de análisis y priorización puede reducir el tiempo y los recursos necesarios para revisar grandes volúmenes de contenido. Este aspecto resulta especialmente relevante para organizaciones dedicadas a la verificación de información o a la moderación de contenidos en línea.
- Impacto ético (alto, mixto). El empleo de modelos de IA introduce riesgos asociados a posibles errores de clasificación, falta de justificación o presencia de sesgos en los datos de entrenamiento. Por este motivo, los resultados generados por el sistema deben interpretarse como una ayuda a la toma de decisiones y no como una evaluación definitiva de la veracidad de una publicación.
- Impacto ambiental (bajo). El proyecto no requiere infraestructuras físicas específicas ni genera residuos materiales. No obstante, la ejecución de modelos de IA implica un consumo energético asociado al uso de recursos computacionales, especialmente cuando intervienen modelos de gran tamaño o servicios externos.

Los principales grupos de interés identificados son los usuarios de redes sociales, las organizaciones dedicadas a la verificación de información, las plataformas digitales, la comunidad investigadora y las entidades responsables de la regulación del entorno digital.

A.3 ANÁLISIS DETALLADO DEL IMPACTO SOCIAL

El impacto social constituye el principal ámbito de interés del presente proyecto, ya que la desinformación es un fenómeno con consecuencias directas sobre la calidad de la información consumida por la ciudadanía. A continuación, se analiza de qué forma la métrica propuesta puede contribuir a la identificación de publicaciones potencialmente desinformadoras y qué factores condicionan su utilidad práctica.

Se identifican varios mecanismos mediante los cuales el sistema podría generar un impacto positivo:

- Apoyo a la identificación temprana de contenidos potencialmente desinformadores. La métrica desarrollada permite combinar distintos indicadores relacionados con la veracidad, el contexto y el lenguaje utilizado en una publicación, proporcionando una valoración homogénea del riesgo asociado a ella. Además, el hecho de que se haya conseguido automatizar el proceso de cálculo permite poder medir el riesgo de un post prácticamente desde el momento en el que se publica.
- Estandarización de los procesos de evaluación. El uso de una métrica común facilita la comparación entre publicaciones diferentes utilizando los mismos criterios de análisis, reduciendo la variabilidad derivada de evaluaciones exclusivamente subjetivas.
- Apoyo a la investigación sobre desinformación. La generación de puntuaciones cuantitativas permite realizar estudios comparativos, analizar tendencias y evaluar el comportamiento de distintas categorías de publicaciones de forma sistemática.

Asimismo, pueden identificarse diversos grupos de interés:

- Investigadores: disponen de una herramienta que facilita el análisis cuantitativo de contenidos relacionados con la desinformación.
- Organizaciones dedicadas a la verificación de información: pueden emplear la métrica como mecanismo complementario para priorizar publicaciones susceptibles de requerir un análisis más detallado.
- Plataformas digitales: podrían utilizar este tipo de métricas como apoyo en procesos de monitorización o análisis de contenidos.
- Usuarios de redes sociales: se beneficiarían indirectamente de la existencia de herramientas orientadas a mejorar la detección de contenidos potencialmente engañosos.

El beneficio potencial del sistema depende de varias condiciones:

- Transparencia metodológica. La utilidad de una herramienta de este tipo depende en gran medida de la claridad con la que se definan los criterios de evaluación utilizados. La publicación de los parámetros analizados, las estrategias de normalización empleadas y los pesos asignados a cada componente facilita la comprensión de los resultados obtenidos y permite identificar posibles limitaciones del sistema.
- Actualización periódica de los modelos utilizados. El fenómeno de la desinformación evoluciona constantemente, incorporando nuevas estrategias comunicativas y patrones lingüísticos. Como consecuencia, los modelos empleados deben ser revisados y actualizados periódicamente para garantizar que continúan siendo capaces de identificar adecuadamente las características asociadas a este tipo de contenidos.
- Evaluación continua de posibles sesgos. Dado que el sistema integra modelos entrenados sobre conjuntos de datos externos, resulta conveniente analizar periódicamente su comportamiento

para detectar posibles sesgos ideológicos, culturales o lingüísticos que puedan afectar a las puntuaciones generadas. Esta evaluación permitiría mejorar progresivamente la robustez y equidad del sistema.

- Validación sobre conjuntos de datos amplios y representativos. La capacidad de generalización de la métrica depende de que sea evaluada utilizando publicaciones de naturaleza diversa, procedentes de distintos contextos, temáticas e idiomas. Una validación más extensa permitiría determinar con mayor precisión el alcance real del sistema y sus posibles limitaciones.

En conjunto, el proyecto contribuye al estudio de la desinformación mediante una aproximación cuantitativa basada en IA y NLP. Aunque el prototipo desarrollado no pretende sustituir la evaluación humana, demuestra la viabilidad de construir métricas capaces de integrar múltiples factores relacionados con este fenómeno.

A.4 CONCLUSIONES

El trabajo desarrollado presenta un impacto social potencialmente positivo al proporcionar una herramienta capaz de apoyar la identificación temprana de contenidos que podrían estar relacionados con procesos de desinformación. Aunque el prototipo no sustituye el análisis humano ni garantiza la detección completa de este fenómeno, sí demuestra la viabilidad de combinar diferentes técnicas de IA para obtener una evaluación homogénea del riesgo asociado a una publicación.

Desde el punto de vista económico, la automatización de determinadas tareas de análisis puede contribuir a optimizar los recursos destinados a actividades de verificación y moderación de contenidos, especialmente en entornos donde se generan grandes volúmenes de información.

En el ámbito ambiental, el impacto directo del proyecto es reducido y se limita principalmente al consumo energético asociado a los recursos computacionales empleados durante la ejecución de los modelos utilizados.

La dimensión ética constituye un aspecto relevante del sistema desarrollado. La utilización de modelos de IA requiere considerar posibles sesgos, limitaciones de interpretabilidad y riesgos asociados a una utilización inadecuada de los resultados. Por ello, la supervisión humana y la transparencia metodológica deben considerarse elementos esenciales para un uso responsable de la herramienta.

En conjunto, la incorporación de criterios éticos, sociales, económicos y ambientales ha permitido analizar el proyecto desde una perspectiva más amplia que la puramente técnica. Este análisis aporta valor añadido al desarrollo realizado, al facilitar la identificación de beneficios potenciales, limitaciones y riesgos que deberán ser considerados en futuras evoluciones del sistema.

ANEXO B: PRESUPUESTO ECONÓMICO

Se presenta a continuación la estimación económica de los recursos empleados durante el desarrollo del proyecto. Esta estimación incluye los principales costes de mano de obra, materiales y otros gastos asociados al TFG.

COSTE DE MANO DE OBRA (coste directo)	**Horas**	**Precio/hora**	**Total**
	300	15 €	**4.500 €**

COSTE DE RECURSOS MATERIALES (coste directo)	**Precio de compra**	**Uso en meses**	**Amortización (en años)**	**Total**
Portátil Acer Swift SF14-11	800€	6	5	80,00 €
Tokens de OpenAI API	10 €	-	-	10,00 €

COSTE TOTAL DE RECURSOS MATERIALES	**90,00 €**

GASTOS GENERALES (costes indirectos)	15%	sobre CD	**688,50 €**
BENEFICIO INDUSTRIAL	6%	sobre CD+CI	**316,71 €**

SUBTOTAL PRESUPUESTO		**5.595,21 €**
IVA APLICABLE	21%	**1.174,99 €**
TOTAL PRESUPUESTO		**6.770,20 €**