Trabajo Fin de Máster: Detección y alerta de noticias falsas en procesos electorales con una metodología basada en el estructuralismo narrativo

Trabajo Fin de Máster

Pedro de Alzaga Fraguas

Trabajo de investigación para el

Máster en Tecnologías del Lenguaje

Universidad Nacional de Educación a Distancia

Dirigido por:

Prof. Dr. D. Álvaro Rodrigo Yuste

Prof. Dr. D. Roberto Centeno Sánchez

Septiembre 2023

Agradecimientos

A Álvaro y a Roberto, por su criterio, generosidad y paciencia.

Resumen

La desinformación no es un problema nuevo, pero ha adquirido dimensiones preocupantes en la última década con la polarización de la política y la expansión de las redes sociales. También ha demostrado ser un fenómeno especialmente grave cuando actúa en procesos electorales, desvirtuando la voluntad popular y arrebatando a los ciudadanos el derecho a estar informados cuando acuden a las urnas; el derecho a elegir libremente, en definitiva. Los algoritmos de detección de noticias falsas y las agencias de verificación intentan mitigar este problema, pero en la mayoría de los casos solo consiguen contrarrestarlo cuando el bulo ya se ha extendido demasiado. El presente trabajo propone un modelo de detección y alerta temprana de noticias falsas, basado en el estructuralismo narrativo iniciado por Vladimir Propp. Este antropólogo y lingüista ruso analizó más de un centenar de cuentos y leyendas rusos para encontrar las estructuras subyacentes que los hacían similares pese a contar historias aparentemente distintas. A diferencia de otros sistemas, que analizan el contenido, la forma o el contexto de las noticias falsas, esta propuesta estudia la estructura que sujeta los mensajes informativos, esos *cuentos* y *leyendas* que se repiten en los procesos electorales de casi todo el mundo. El modelo propuesto ha conseguido una mejora de más del 6% con respecto al modelo preentrenado base. Se propone también un sistema que jerarquiza los bulos de acuerdo con su peligrosidad, para que se priorice su comprobación antes de que se extiendan demasiado. Este trabajo parte de una colección de casi 3.400 tuiteos recopilados durante las Elecciones generales de Brasil 2022, que incluye 77 mensajes desinformativos anotados manualmente.

Abstract

Disinformation is not a new issue, but it has taken on alarming dimensions over the past decade, fueled by political polarization and the proliferation of social media platforms. It has proven to be an especially grave phenomenon when it influences electoral processes, distorting the popular will and depriving citizens of their right to be informed when casting their votes; essentially, their right to make a free choice. Algorithms designed to detect fake news and fact-checking agencies attempt to mitigate this issue, yet in most cases they only manage to counteract it after the falsehood has already spread widely. The present study proposes a model for early detection and alert of fake news, grounded in the narrative structuralism initiated by Vladimir Propp. This Russian anthropologist and linguist analyzed over a hundred Russian tales and legends to uncover underlying structures that made them similar despite narrating apparently different stories. Unlike other systems that analyze the content, form, or context of fake news, this proposal examines the structure that underpins news stories—those tales and legends that recur in electoral processes worldwide. The proposed model has achieved an improvement of more than 6% compared to the baseline pre-trained model. Additionally, a system is proposed that ranks falsehoods according to their level of potential damage, so that their verification is prioritized before they spread too extensively. This study is based on a collection of nearly 3,400 tweets gathered during the General Elections in Brazil in 2022, which includes 77 manually annotated disinformation messages.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

En los países democráticos, el pueblo soberano acude a las urnas para elegir a sus representantes y mandatarios. Es un proceso por lo general breve, pero acompañado de una gran tensión social y política, en el que los electores se ven sometidos a un bombardeo constante de estímulos ideológicos, informativos y persuasivos. En las últimas dos décadas, esta tensión se ha agravado con un nuevo problema: una oleada de desinformación creciente que ha polarizado el debate político y ha tenido en las redes sociales uno de sus principales canales de difusión (Vosoughi et al., 2018). La expansión y agravamiento de este problema ha creado confusión e incertidumbre en diversos procesos electorales de todo el mundo, arrebatando a los ciudadanos el derecho a estar informados y, por tanto, a elegir libremente. En algunos casos, esta desinformación ha desembocado en episodios de violencia, por no haber podido atajar a tiempo las noticias falsas que circularon por las redes y llamaron a la protesta en las calles (Galarraga, 2023).

En casi todos los ámbitos de la comunicación actual, la desinformación está provocando una gran preocupación en amplios sectores sociales por su capacidad para influir en la opinión pública y manipular procesos de muy diversa índole. En el ámbito político, este problema es casi tan antiguo como la política misma. Un estudio de Polo (2019) sitúa en la antigua Roma el uso de las primeras estrategias desinformativas por parte de los gobernantes, quienes instrumentalizaban la información que enviaban al pueblo al servicio de los intereses propios o del Imperio. Desde entonces, el uso de la propaganda y la manipulación informativa se ha ido sofisticando hasta llegar a un punto en que el desarrollo tecnológico hace muy difícil detectarla y mucho más difícil controlar su propagación a través de los canales digitales de comunicación masiva.

Los efectos de la desinformación son especialmente perniciosos en el caso de los procesos electorales, que son el mecanismo por el que las sociedades democráticas eligen sus representantes, evolucionan como Estados y, en definitiva, determinan su

destino. Se han documentado casos de interferencia desinformativa en numerosas elecciones, pero aún resuenan con especial fuerza los comicios de Estados Unidos en 2016, tras los que salió elegido presidente el hoy procesado Donald Trump (Rogers y Bromwich, 2016), y el referéndum para la salida del Reino Unido de la Unión Europea, conocido popularmente como *Brexit* (Bastos y Mercea, 2017) y de consecuencias aún hoy muy lamentables para los británicos.

En esta batalla entre la información y la desinformación, entre el dato y el relato, entre la verdad y la *postverdad*, la tecnología ha jugado un papel muy importante, y no siempre beneficioso para la sociedad, con un denominador común en casi todos los países: las redes sociales. Por este motivo, este trabajo se centrará en la detección y evaluación de la desinformación que se distribuye a través de estas plataformas durante los procesos electorales.

Desde hace décadas, las misiones de observación electoral se despliegan en los países para evaluar las elecciones a la luz de las legislaciones nacionales y los acuerdos internacionales suscritos por los países. Esta evaluación incluye el análisis político, legal, electoral, mediático y, desde hace unos años, de las redes sociales. Las misiones deben determinar hasta qué punto en estos ámbitos se cumplen con los estándares electorales adoptados por el país que está siendo observado. Para realizar esta evaluación las misiones disponen de expertos, de observadores de corto y largo plazo desplegados sobre el terreno y, en el caso de el entorno mediático y de redes sociales, de equipos de monitoreo que realizan una observación somera de la cobertura y el debate que sucede en los medios y las plataformas digitales.

En el caso de los medios y las redes sociales, los estándares internacionales en materia electoral establecen que el pueblo necesita disponer de una información completa y plural para ejercer el derecho al voto. Un ciudadano no puede tomar una decisión informada, si no dispone de información adecuada sobre los candidatos y sus programas. Por su parte, los candidatos tienen derecho a recibir una cobertura ecuánime y completa que haga llegar a los ciudadanos sus propuestas, sin desvirtuarlas ni tergiversarlas.

Una campaña desinformativa coordinada en redes sociales puede arrebatar al pueblo la información que necesita para ejercer su derecho al voto y desequilibrar las oportunidades de los candidatos para obtener el apoyo popular. Aun peor, una campaña desinformativa puede desacreditar el proceso electoral, total o parcialmente, con la intención de provocar su suspensión o repetición por motivos

ilegítimos, poniendo en peligro los cimientos mismos de los sistemas democráticos y provocando en algunos casos episodios de violencia y rebeldía.

Los equipos de monitoreo de las misiones de observación electoral revisan diariamente una muestra de la información que se publica en las redes sociales y suelen trabajar con los grupos de verificación locales para localizar esta información y evaluar su alcance y peligro potencial. Por lo general, las misiones disponen de algoritmos para revisar las redes sociales en busca de discurso de odio -racista y xenófobo, machista o xenófobo- pero no para encontrar mensajes desinformativos ni mucho menos para valorar su peligro potencial.

Los sistemas para la detección de noticias falsas propuestos hasta ahora adolecen de dos problemas. Por un lado, son capaces de detectar con mayor o menor éxito las noticias falsas, pero no siempre reconocen su temática ni las clasifican según su contenido. Por otro lado, estos sistemas tratan todo el contenido desinformativo como si tuviera la misma gravedad, por lo que un bulo sobre un candidato, sobre un fallo en las cabinas electorales y sobre una encuesta manipulada tienen exactamente el mismo *peso* que mensajes manipulados sobre el asalto a un edificio oficial o sobre un supuesto atentado en un colegio electoral. No existe hasta el momento un sistema que jerarquice estos mensajes de acuerdo con la eventual gravedad de sus efectos.

Ambas cosas serían beneficiosas para el trabajo de un equipo de monitoreo o de verificación, que podrían analizar mejor cada asunto, priorizar la investigación de los mensajes con un impacto potencialmente más grave sobre el proceso electoral o, incluso, aquellos capaces de generar violencia. Por otro lado, hay escenarios y situaciones en las que no es posible verificar la información de manera inmediata, ya sea por una sobrecarga de trabajo del equipo de verificación, o por la imposibilidad de contrastar la información a corto o medio plazo. En estos casos, un sistema capaz de agrupar los mensajes que contengan la misma narrativa permitiría a los verificadores aportar, al menos, una información general provisional que fuera de utilidad para los receptores de esos mensajes.

En definitiva, para las misiones de observación electoral sería útil disponer de un modelo de detección de señales; concretamente, para la detección de la desinformación y su gravedad durante los procesos electorales, que permitiera responder con la celeridad que merezca cada caso.

1.1. Propuesta y objetivos

Este trabajo tiene como objetivo principal proponer un nuevo modelo para la detección y alerta de mensajes desinformativos en redes sociales, así como su clasificación de acuerdo con la gravedad de sus posibles efectos en los procesos electorales. La propuesta se basa en analizar las narrativas de acuerdo con su estructura, un trabajo inspirado en uno de los padres del estructuralismo, el antropólogo y lingüista ruso Vladimir Propp.

Partiendo del marco teórico propuesto por Propp, y que se explicará con detalle más adelante, se plantea la siguiente hipótesis:

- Identificar en mensajes de redes sociales algunos personajes del proceso electoral (personajes, según Propp), separar las acciones que realizan (funciones) y suministrar esta información a un modelo de detección mensajes desinformativos puede mejorar su rendimiento.

El modelo de detección mencionado podría clasificarse en el grupo de los que analizan el contenido de los mensajes, pero además de mirar sus elementos o estilo, buscaría las estructuras subyacentes que comparten los mensajes desinformativos. La posibilidad de agrupar las noticias de acuerdo con su temática y gravedad facilitaría su revisión, valoración y priorización y permitiría dar una respuesta temprana a la viralidad de las campañas desinformativas.

Para corroborar esta hipótesis, se proponen los siguientes objetivos específicos:

- Recopilar un conjunto de datos de redes sociales que contenga mensajes legítimos y noticias falsas, y crear un esquema de anotación semiautomatizado que ayude a los modelos de aprendizaje automático a reconocer los protagonistas y las narrativas de la desinformación electoral en estos mensajes.

- Desarrollar un modelo que, a partir de la información previamente seleccionada y anotada, permita reconocer las estructuras subyacentes de la desinformación para facilitar su detección.

- Probar un modelo de clasificación que reconozca la temática y gravedad del mensaje desinformativo, para que los equipos de verificación que trabajen con el modelo puedan actuar en consecuencia.

En definitiva, este modelo no puede ni pretende identificar todas los mensajes desinformativos que circulan por las redes sociales, una tarea que como se verá más adelante resulta muy difícil en el actual estado del arte. Tampoco puede diferenciar los diferentes tipos de mensajes desinformativos tal y como se reconocen en la literatura científica en inglés, que diferencia entre *disinformation* y *misinformation*, según la intención, deliberada o no, respectivamente, de quien elabora, manipula o difunde una información falsa (Stahl, 2006).

El trabajo propuesto pretende detectar las narrativas más habituales en las que se esconde la desinformación electoral, plantear un esquema de clasificación de estas narrativas y establecer un sistema de alertas de acuerdo con su posible gravedad que ayude a mitigar sus posibles efectos perniciosos en un proceso democrático.

1.2. Estructura del documento

Este trabajo ha sido redactado con la siguiente estructura:

Capítulo 1. Introducción. Este capítulo explica el problema creciente de la desinformación en todos los ámbitos de la sociedad y en especial en el contexto político y electoral, donde sus efectos han producido gran convulsión en elecciones de todo el mundo. En este apartado se apunta también la necesidad un sistema que pueda dar respuesta a este problema más allá de lo propuesto hasta ahora.

Capítulo 2. Estado del arte. Los sistemas de detección de noticias falsas se han centrado fundamentalmente en analizar los mensajes desinformativos desde tres puntos de vista: el contenido de estos mensajes, su temática; la forma, estilo y complementos con que se presentan; y el contexto en que aparecen y son difundidos. Este capítulo muestra las características y fortalezas de cada una de estos sistemas, así como sus debilidades.

Capítulo 3. Metodología y narrativas propuestas. El trabajo avanza explicando el trabajo de Vladimir Propp y su análisis de la narrativa popular rusa partiendo de los personajes y funciones que aparecen en estos relatos. Análogamente, se explican las narrativas que pueden verse más frecuentemente durante los procesos electorales y, al estilo de Propp, se proponen cuatro tipos de protagonistas y cuatro grupos o temáticas de desinformación electoral. Asimismo, se explica como procesar la información original para obtener estas estructuras análogas.

Capítulo 4. Evaluación. En este capítulo se describen los experimentos realizados y se publican sus resultados, ordenados de acuerdo con su rendimiento para facilitar su interpretación posterior.

Capítulo 5. Discusión. Los resultados son analizados e interpretados, hasta donde permiten los datos, para conocer el rendimiento del modelo según la información elegida para entrenarlo y evaluarlo.

Capítulo 6. Conclusiones y trabajo futuro. Por último, se hace una interpretación final del funcionamiento del sistema propuesto y se proponen nuevas vías para un desarrollarlo futuro.

Capítulo 2

Estado del arte

Para la detección y evaluación de noticias falsas, los equipos de monitoreo y verificación apenas disponen de herramientas tecnológicas. Estos profesionales acceden al contenido de las redes sociales por dos vías: a través de los interfaces de programación de aplicaciones (API, en sus siglas en inglés) de las redes sociales, cuyo acceso no siempre está abierto a todas las misiones, o por medio de herramientas de gestión de cuentas de redes sociales, que facilitan su visualización individual y conjunta, así como la selección por diversos criterios: cronológicos, temáticos, por autoría... El uso de algoritmos de aprendizaje automático permite a estos equipos detectar el discurso de odio, pero no existen modelos análogos para la detección de desinfomación que se usen de manera generalizada. Solo el filtrado del contenido mediante palabras clave, o las denuncias recibidas por parte de terceros, les permite localizar mensajes susceptibles de contener desinformación.

El resultado es que los equipos de verificación no siempre pueden publicar con la celeridad necesaria los desmentidos y las campañas de información que contrarresten los efectos de los mensajes desinformativos. En la mayoría de los casos, esta respuesta llega cuando el mensaje se ha viralizado por diversos canales.

2.1. Detección de noticias falsas

Hasta la fecha se han probado varias técnicas para la detección de noticias falsas que usan aprendizaje automático, redes neuronales (Sastrawan et al., 2022), grafos y sistemas híbridos con otras tecnologías (Okunoye y Ibor, 2022), para analizar el lenguaje (Choudhary y Arora, 2021), la temática del mensaje (Bharadwaj y Shao, 2019), su propagación o viralidad (Liu y Wu, 2018) e incluso las intenciones de quienes distribuyen este contenido (Zhou et al., 2022).

Estas técnicas de detección pueden agruparse según el enfoque empleado para la detección de noticias en:

- Detección basada en el **contenido**, que analiza las palabras desde un punto de vista semántico para determinar los asuntos tratados en el mensaje y establecer eventuales relaciones de estos temas con los mensajes desinformativos.
- Detección basada en la **forma** del mensaje, que revisa su estructura sintáctica para encontrar relaciones significativas entre su orden, disposición o incluso el material audiovisual con que se acompaña.
- Detección basada en el **contexto**, que examina asuntos tangenciales, como la viralidad del mensaje, las características de los autores que lo propagan o el mismo entorno en donde se publica.

Se incluyen en este estado del arte algunos estudios que, al margen del sistema empleado para analizar los mensajes desinformativos, han enfocado su trabajo en el estudio de la desinformación que puede encontrarse en el ámbito político, y, más concretamente, en el electoral.

2.2. Detección basada en el contenido

El sistema óptimo para detectar e identificar una noticia falsa debería ser aquel que pudiera contrastar los datos y afirmaciones que contiene un mensaje. Esta comprobación podría hacerse de manera manual, mediante verificadores humanos que contrastaran la noticia recurriendo a las fuentes originales; o de manera automática, con algoritmos que cotejaran la información con datos de inteligencia de fuentes abiertas (OSINT), por ejemplo.

Es el caso de MedOSINT (Martinez Monterrubio et al., 2021), un sistema que recopila información de salud de fuentes públicas para contrastar las noticias que aparecen sobre COVID-19. De la información obtenida, por lo general de boletines médicos, se extraen los datos clave con los que se crean nuevas características para alimentar un modelo de aprendizaje automático de razonamiento basado en casos (CBR, en sus siglas en inglés). Este paradigma del aprendizaje automático se sirve de la memoria de los casos anteriores para resolver nuevos casos de forma análoga. La capacidad del sistema para extraer información actualizada de fuentes abiertas de

manera constante lo hace especialmente recomendable para luchar contra la desinformación relacionada con el coronavirus y su enfermedad, que abundó durante las distintas fases de la pandemia.

Otros enfoques utilizados por los sistemas automáticos de análisis de contenido van desde el simple análisis de la frecuencia de ciertas palabras (Setiawan et al., 2022) hasta la extracción de información semántica del texto que pueda aportar nuevas variables para el proceso de detección (Jadhav y Thepade, 2019), pasando por el análisis de la novedad y el elemento de sorpresa incluidos en el enunciado del mensaje (Kumari et al., 2022).

El primer estudio convierte las palabras en vectores TF-IDF, que luego procesa de manera individual, en palabras; atomizada, en caracteres; o de forma conjunta, con otras palabras (N-gramas), para crear características que intenten capturar la importancia relativa de cada palabra dentro del texto. En esta misma línea de utilizar TF-IDF y N-gramas trabajan Setiawan et al. (2022) aunque amplían las características con un sistema de vectores GloVe capaz de capturar información semántica y las relaciones entre palabras del mensaje.

El segundo utiliza Deep Semantic Structured Model (DSSM), una técnica de redes neuronales profundas que permite modelar la similitud semántica entre cadenas de texto. Jadhav y Thepade (2019) usan esta similitud para crear nuevas características de un modelo que permita reconocer mejor la desinformación en un mensaje. Por último, Kumari et al. (2022) utilizan una aproximación al problema basada en lo que denominan como "novedad textual", o capacidad de un texto para enfatizar lo nuevo de un mensaje, así como la emoción o sorpresa que provoca. Ambas características son, siempre según estos autores, tan importantes para entender la viralidad de las noticias falsas como para detectarlas.

Algunos autores no se limitan a analizar el contenido falso que aparece regularmente en redes sociales sino que estudian también el publicado en medios tradicionales, supuestamente libres de esta lacra. Medios tan prestigiosos como *The New York Times*, el diario estadounidense en el que trabajaba Jayson Blair, un periodista despedido por redactar información falsa, plagiada o directamente inventada en al menos 36 artículos publicados en apenas tres años de carrera en este rotativo generalista. Grieve y Woodfield (2023) proponen un marco de análisis gramatical y sintáctico con el que analizar las noticias falsas y toman como muestra 64 artículos de Blair, para concluir que aquellos que son ciertos albergan más

información y destilan un tono de convicción más notable que no se ve en los artículos que son falsos.

Cuando se destapó el caso de Blair, el diario puso a un equipo a comprobar hasta dónde llegaban las mentiras publicadas y a verificar de forma manual todos los artículos redactados por el joven reportero. Y es en este punto donde los sistemas automáticos se topan con el límite que impone la virtualidad: algunas informaciones no pueden verificarse sin recurrir al diálogo con las personas implicadas o la visita a algún lugar mencionado en el mensaje supuestamente desinformativo, extremos ambos que de momento no están al alcance de un programa informático.

Los verificadores humanos sí pueden entrevistar a las personas implicadas y visitar los escenarios, pero el volumen de información y la rapidez a la que circula por las redes sociales hacen que la verificación manual completa sea poco operativa en un sistema de alerta temprana. Por otra parte, el tamaño del equipo de verificación necesario para una tarea así, además de su formación, requeriría un presupuesto del que no siempre disponen unas instituciones de verificación, que suelen recibir sus fondos de subvenciones públicas y donantes privados (Castellet et al., 2023).

2.3. Detección basada en la forma

Al margen del contraste del contenido, algunos de los sistemas conocidos hasta la fecha analizan en el lenguaje utilizado en la redacción del texto desinformativo, para detectar patrones de estilo que permitan identificar una noticia falsa (Tsai, 2023). Este método prescinde de la temática del mensaje y se centra en la forma en que se redacta, se publica y se presenta, normalmente acompañado de elementos multimedia que apoyan la confusión (Uppada et al., 2022).

En este punto juega un papel relevante la estilometría, una rama de la lingüística que estudia los estilos de escritura, por lo general por medios matemáticos y estadísticos, para identificar características únicas en el uso del lenguaje (Daelemans, 2013). Esta disciplina analiza la frecuencia de uso de las palabras, la gramática, la sintaxis, el uso de la puntuación, la longitud de las frases y la diversidad de léxico, entre otras variables, para determinar la autenticidad de un texto atribuido a alguien, conocer las escuelas o géneros a los que pertenecen los textos

clásicos, detectar el plago o, incluso, como herramienta forense, para certificar la autoría de los documentos que se usan como prueba en los tribunales.

En su estudio, Tsai (2023) adopta un enfoque estilométrico y parte de la premisa de que un mensaje cierto y uno falso deberían tener necesariamente diferencias de estilo. Es lo que se conoce como la Hipótesis Undeutsch (Undeutsch, 1967), que sostiene que las descripciones de eventos que han sucedido en realidad difieren cualitativamente de las descripciones de aquellos sucesos que solo se han imaginado.

El autor utiliza un modelo preentrenado sobre esta premisa para analizar varios conjuntos de datos públicos de noticias falsas. Previamente crea un banco de entidades nombradas (NER) en las noticias reales, y parte de la base de que este banco es un campo finito de entidades que no siempre se mencionan en la noticias falsas, para detectar estas últimas. En definitiva, el estudio concluye que si un artículo de noticias contiene una "alta frecuencia de entidades nombradas relevantes para un tema y coherentes con otras fuentes fiables, es más probable que sea digno de confianza". Por el contrario, si una noticia contiene una "alta frecuencia de entidades nombradas que son irrelevantes para el tema o incoherentes con otras fuentes, es más probable que sea sospechoso".

Otros autores prefieren estudiar el lenguaje de los usuarios, para reconocer en él las fuentes más probables de creación o difusión de noticias falsas. Manna et al. (2020) estudian un conjunto de perfiles de Twitter y analizan sus textos en busca de características estilométricas y léxicas: el número y tipo de emojis utilizados, así como los URL, palabras, espacios y signos de puntuación, por un lado, y las opiniones y palabras persuasivas o propias de titulares cebo, por otro, para calcular la probabilidad de que un usuario así sea un difusor de noticias falsas.

El uso de contenido multimedia para armar o acompañar el mensaje desinformativo también es objeto de estudio. Según Uppada et al. (2022) este contenido, cada vez más habitual en las redes sociales, suele utilizarse para reforzar la respuesta emocional al mensaje desinformativo. Los autores proponen un sistema multimodal y multiarquitectura – texto y vídeo- para analizar de manera independiente no solo la imagen – su polaridad y la posibilidad de que haya sido manipulada- , sino también el mensaje que la acompaña.

2.4. Detección basada en el contexto

Por último, las técnicas basadas en el contexto suelen enfocarse en el análisis de la topología de las redes que se forman entre los agentes que difunden los mensajes desinformativos, en las pautas temporales con que se publican las noticias falsas o en la credibilidad de los agentes propagadores, entre otras características.

En los análisis de la topología de las redes de difusión de noticias falsas, destaca el estudio de Pierri et al. (2020) . Los investigadores recopilaron un amplio conjunto de datos sobre cómo se comparten las noticias en Twitter y aplicaron técnicas de aprendizaje automático para identificar patrones distintivos entre las redes difusión de las noticias reales y las redes de difusión de las falsas. En esta comparación descubrieron diferencias sutiles pero sistemáticas que se manifestaron en el número de nodos y las distribución de las medidas de centralidad, indicios que según los autores podrían servir para detectar las noticias falsas durante su difusión.

Siguiendo esta misma línea de fijarse en la distribución de las noticias, sobresale el estudio de Murayama et al. (2020), quienes no solo analizaron la topología de estas redes sino que pusieron el foco en cómo cambiaba la atención que reciben las noticias en el tiempo. Analizando datos lingüísticos, de usuario y temporales, crearon un indice de contagiosidad para determinar la probabilidad de que una noticia sea compartida. Los autores descubrieron que la difusión de las noticias falsas se realiza de forma más atropellada, con ráfagas periódicas de difusión, que creen que podrían deberse a los usuarios que cuestionan las noticias y despiertan la atención del resto de la comunidad sobre ellas. Las noticias reales, mientras tanto, suelen tener una difusión más uniforme que tiende a la irrelevancia de manera más progresiva.

Por último, Yuan et al. (2020) se sirven de los grafos para analizar el comportamiento de las noticias falsas en su paso por las redes sociales. Los autores proponen un sistema híbrido basado en tres ejes de investigación: la credibilidad de la entidad o persona que publica la noticia, la credibilidad de los usuarios que la difunden y un índice creado a partir de las anteriores y el propio contenido de la noticia. La mayor parte de la información es obtenida a partir de la estructura de los grafos que muestran la relación entre el publicador y los difusores de la noticia, pero sin despreciar las características lingüísticas del propio mensaje. En definitiva, el

estudio busca emular algo que los lectores de noticias hacen de manera inconsciente: valorar la credibilidad de las noticias de acuerdo con la credibilidad de su emisor, un extremo que aunque los autores admiten que no es totalmente fiable sí contribuye notablemente a la detección del contenido falso.

2.5. Estudios desde el punto de vista político

Al margen de los enfoques estrictamente tecnológicos, algunos autores se centran en el ecosistema mediático en el que se produce la propagación de las noticias falsas y ponen en duda la idea misma que separa la *disinformation* y la *misinformation*. Estas denominaciones son agrupadas por Giglietto et al. (2019) en una categoría superior como información engañosa (*misleading information*), en la medida en que puede variar a lo largo de su proceso de propagación: la persona que introduce la noticia falsa en la red puede creer que es cierta (misinformation), pero un propagador puede difundirla a sabiendas de que es falsa (disinformation). Y viceversa, el autor puede saber que es falsa (disinformation), pero el propagador no (misinformation).

Otros estudios inciden en los efectos perniciosos de la desinformación en la sociedad, que pueden llegar al descrédito democrático incluso entre los sectores de la población teóricamente no expuestos a esta lacra. Nisbet et al. (2021) abordan este problema en un estudio realizado en Estados Unidos (EEUU) en el que introducen un nuevo concepto: la Influencia Presunta de la Desinformación (PIM, en sus siglas en inglés), que se refiere a cuánto creen las personas que la desinformación influye en las acciones de los demás, y cómo esta percepción influye en sus propias acciones. En definitiva, el impacto que la desinformación tenga en nosotros no nos preocupa tanto como el que creemos que pueda tener en los demás y, después, indirectamente, en nosotros. Es lo que se como el Efecto tercera persona (Davison, 1983), en el que las personas tienden a sobrestimar el efecto dañino de los mensajes difundidos por los medios en los demás en comparación con ellos mismos. En cualquier caso este impacto se traduce según los autores en una insatisfacción con la democracia electoral estadounidense que es transversal a la ideología política y puede tener efectos duraderos.

2.6. Ventajas e inconvenientes de los modelos de detección

Los sistemas de detección de noticias falsas cuentan con ventajas e inconvenientes que los hacen más propensos a ser utilizados en diferentes dominios y bajo distintas circunstancias, ya sea de forma individual o conjunta. A medida que transcurre el tiempo y aumenta el número de trabajos de investigación, puede detectarse un uso cada vez mayor de sistemas híbridos, que utilizan enfoques de contenido, de forma y contextuales indistintamente (Ying et al., 2021).

Como ya se ha apuntado, la mayor dificultad a la que se enfrentan los modelos de detección basados en el contenido tiene que ver con un su incapacidad para contrastar la información por medio de entrevistas o *in situ*. También afrontan otras dificultades derivadas de factores culturales y sociales, como el idioma o las costumbres, que hacen difícil su uso fuera de un determinado ámbito cultural.

Por su parte, los modelos que analizan la forma de los mensajes se topan con la capacidad creciente de la desinformación para adaptarse e imitar el lenguaje y el estilo de los informativos, haciendo muy difícil su distinción. Además, la obligada brevedad de estos mensajes en redes sociales, el uso de abreviaturas y las erratas en su redacción agravan la dificultad para reconocer patrones estilísticos.

Por último, la principal limitación que deben afrontar los modelos de detección de noticias falsas que se basan en el contexto es que suelen requerir un periodo de tiempo y una cierta evaluación previa, manual o automática, hasta tener información que analizar. Es necesario que la noticia se difunda hasta cierto punto para tener información temporal o topológica que los algoritmos puedan analizar, lo que dificulta su uso en sistemas de alerta temprana.

El sistema propuesto en este trabajo puede clasificarse dentro del primer grupo, el de los modelos que analizan el contenido, aunque no comparte la mayoría de las dificultades de los otros modelos. Por ejemplo, no es tan dependiente de las peculiaridades del idioma, porque no analiza tanto las palabras como las estructuras que subyacen en el mensaje; por explicarlo más llanamente, intenta descubrir los *metacuentos* y *metaleyendas* que pueden leerse entre líneas de los mensajes desinformativos que suelen verse durante los procesos electorales. Una vez detectado el mensaje, puede clasificarlo además según su gravedad, para que se tomen las medidas oportunas por parte de los equipos de monitoreo y verificación.

Capítulo 3

Detección de noticias electorales falsas basada en narrativas

El sistema propuesto en este trabajo se inspira en el trabajo de Vladimir Propp, un antropólogo y lingüista ruso que a principios del siglo XX analizó los cuentos y leyendas de su país en busca de los componentes más básicos y comunes que unían a sus protagonistas y sus temáticas.

En su obra más conocida, *Morfología del cuento* (1928), Propp analizó más de un centenar de cuentos y leyendas populares rusas, y redujo sus argumentos a un grupo de funciones básicas realizadas por un conjunto de protagonistas comunes. La idea fundamental del lingüista era demostrar que, a pesar de la aparente variedad de temas y personajes que podían encontrarse en la narrativa popular rusa, todos ellos podían reducirse a apenas 31 funciones y siete personajes.

Su trabajo abrió la puerta al análisis narrativo, al tiempo que ponía sobre la mesa una verdad incómoda en aquella época: a pesar de que la literatura humana parecía fruto de la inspiración, era original y hasta cierto punto aleatoria, contaba con elementos estructurales comunes que era posible analizar y hasta cuantificar para conocer cómo se componen las historias y se relacionan los personajes. En definitiva, a pesar de la variedad y originalidad de los cuentos y leyendas rusas analizados, todos ellos podían reducirse a un puñado de protagonistas que realizaban unas cuantas acciones.

3.1. Marco teórico: Personajes y funciones en *Morfología del cuento*

Después de un detenido análisis, Propp *destiló* siete personajes que aparecían en todas las obras con una forma y denominación parecida o análoga, y descubrió que estos personajes, pese a la aparente complejidad de las obras, no llegaban a realizar más de 31 acciones o funciones, con las que se componía la trama.

Tanto los personajes como estas funciones no son exclusivas y pueden utilizarse de manera individual o combinada: es decir, por un lado un donante puede actuar también como auxiliar, y por otro un agresor puede participar en un interrogatorio y en un engaño dentro de la misma historia.

3.1.1. Personajes

1. El **agresor** o malvado, autor de la fechoría, el combate y otras formas de lucha contra el héroe.
2. El **donante** o proveedor, que da al héroe el objeto mágico.
3. El **auxiliar**, que traslada o presta socorro al héroe.
4. La **princesa**, el personaje buscado, **y su padre**, quien por lo general propone las tareas difíciles.
5. El **mandatario** u ordenante, que envía al héroe.
6. El **héroe**, protagonista de la historia.
7. El **falso héroe** o antagonista.

3.1.2. Funciones

1. **Alejamiento**
   Uno de los miembros de la familia se aleja de la casa

Desde una simple visita, a al muerte de algún familiar pasando por la marcha a la guerra o al trabajo.

2. **Prohibición**
   Recae sobre el protagonista una prohibición
   También puede tratarse de una orden, una proposición o un consejo.

3. **Transgresión**
   Se transgrede la prohibición
   Esta transgresión, que no siempre requiere una prohibición expresa previa, suele provocar la entrada en escena del agresor del protagonista.

4. **Interrogatorio**
   El agresor intenta obtener noticias
   Casi siempre sobre la víctima, el bien a proteger o algún objeto precioso.

5. **Información**
   El agresor recibe informaciones sobre su víctima
   Consigue una respuesta a sus preguntas. A menudo se produce un diálogo.

6. **Engaño**
   El agresor intenta engañar a su víctima para apoderarse de ella o de sus bienes
   Utiliza la persuasión o medios mágicos, engañosos o violentos.

7. **Complicidad**
   La víctima se deja engañar y ayuda así a su enemigo a su pesar
   Acepta y ejecuta las proposiciones engañosas

8. **Fechoría**
   El agresor daña a uno de los miembros de la familia o le causa perjuicios
   Una función muy importante que da al cuento movimiento. Puede que algo le falte a uno de los miembros de la familia; o uno de los miembros de la familia quiere poseer algo.

9. **Mediación, momento de transición**
   Se divulga la noticia de la fechoría o de la carencia, se dirigen al héroe con una pregunta o una orden, se le llama o se le hace partir
   Esta función hace aparecer en escena al héroe, que puede ser el buscador de una persona raptada (héroe-buscador) o la propia persona raptada o expulsada (héroe-víctima).

10. **Principio de la acción contraria**
    El héroe-buscador acepta o decide actuar
    Solo existe en los cuentos en los que el héroe parte para efectuar una búsqueda.

11. **Partida**
    El héroe se va de su casa
    El héroe comienza su búsqueda y aparece el donante o proveedor, un nuevo personaje que le ofrece un objeto mágico que le será de ayuda pero que solo obtendrá tras algunas pruebas.

12. **Primera función del donante**
    El héroe sufre una prueba, un cuestionario, un ataque, etc., que le preparan para la recepción de un objeto o de un auxiliar mágico
    El nuevo personaje pide una acción al héroe para ponerle a prueba y así recibir el objeto mágico. Puede ser una prueba en sí, o un interrogatorio, una petición, o un ataque.

13. **Reacción del héroe**
    El héroe reacciona ante las acciones del futuro donante
    Esta reacción puede ser positiva o negativa, y puede llevarla a superar la prueba o no.

14. **Recepción del objeto mágico**
    El objeto mágico pasa a disposición del héroe
    Puede ser un animal, un arma, un objeto, una virtud, un conjuro... Lo obtiene como recompensa, tras una lucha, por azar o incluso por medio del robo.

15. **Desplazamiento**
    El héroe es transportado, conducido o llevado cerca del lugar donde se halla el objeto de su búsqueda
    Este objeto se encuentra en otro reino que puede estar muy lejos, en lo alto o incluso bajo tierra.

16. **Combate**
    El héroe y el agresor se enfrentan en un combate
    Puede ser una lucha, una competición, un juego o algún tipo de acertijo.

17. **Marca**
    El héroe recibe una marca

Suele ser una herida sufrida en el combate, que puede ser vendada con un pañuelo de la princesa.

18. **Victoria**
    El agresor es vencido
    Puede darse el caso de que el agresor no se derrotado directamente sino que huya del combate.

19. **Reparación**
    La fechoría inicial es reparada o la carencia colmada
    "En este punto, el cuento alcanza su culminación", dice el autor. El héroe consigue su objetivo y la fuerza antagonista es derrotada. Señala Propp que en este punto el cuento alcanza su culminación. Es el clímax de la trama, cuando la fuerza antagonista ha sido vencida y el héroe consigue el objeto de su deseo.

20. **Regreso**
    El héroe regresa
    A veces este retorno toma la forma de una huída.

21. **Persecución**
    El héroe es perseguido
    El perseguidor intenta acabar on el héroe incluso adoptando distintas formas.

22. **Socorro**
    El héroe es auxiliado o escapa de su perseguidor
    A veces huye, se transforma o se oculta. Muchos cuentos tienen en esta función su final, aunque no siempre. Puede ser el principio del fin del cuento, o el comienzo de nuevas peripecias.

23. **Llegada de incógnito**
    El héroe llega de incógnito a su casa o a otra comarca
    Se enrola como aprendiz en un pueblo o cocinero en el palacio para no llamar la atención.

24. **Pretensiones engañosas**
    Un falso héroe reivindica para sí pretensiones engañosas
    Alguien reclama la gesta como propia y esgrimen el objeto que sirve de prueba.

25. **Tarea difícil**
    Se propone al héroe una tarea difícil
    Pueden ser pruebas de fuerza, de habilidad, de valor, de ingenio, mágicas...

26. **Tarea cumplida**
    La tarea es realizada
    A veces incluso antes de que sean encomendadas.

27. **Reconocimiento**
    El héroe es reconocido
    Gracias al objeto conseguido, la marca o herida recibida en combate, o la realización de una tare difícil.

28. **Descubrimiento**
    El falso héroe o el agresor, el malvado, queda desenmascarado
    A veces porque fracasa en una tarea encomendada o porque se delata él mismo.

29. **Transfiguración**
    El héroe recibe una nueva apariencia
    Su aspecto de aprendiz pasa a ser de cortesano con lujosos vestidos.

30. **Castigo**
    El falso héroe o el agresor es castigado
    Se le castiga duramente, disparándole, dándole caza o atándolo a la cola de un caballo. A veces se suicida y otras encuentra el perdón.

31. **Matrimonio**
    El héroe se casa y asciende al trono
    La recompensa final en la que el hombre recibe a la mujer, el reino o ambas cosas. El cuento se termina en este punto.

Como puede comprobarse, la mayoría de las funciones de Propp se expresan con una estructura gramatical simple, como sujeto y predicado. Casi todas en voz activa y con un nombre, un verbo y un objeto directo o indirecto. Son los personajes y esta estructura simple de las funciones lo que permite reconocer el mismo suceso en cuentos distintos. El autor sugiere, a fin de cuentas, que hay una estructura general que está por encima de las historias particulares de cada cuento y leyenda.

Precisamente esta idea de que las historias cuentan con estructuras reconocibles y reproducibles en otras historias supuso la base del estructuralismo narrativo y uno de los fundamentos del estructuralismo, una corriente teórica que pretende conocer las leyes y principios universales que marcan los fenómenos sociales y culturales, por medio del análisis conjunto de sus estructuras subyacentes.

3.2. Narrativas más frecuentes en los procesos electorales

Tras participar en varias misiones internacionales de observación electoral con el Centro Carter y la Unión Europea, el autor ha podido constatar cómo las narrativas utilizadas en las campañas locales de desinformación tenían muchas similitudes, incluso entre diferentes países. Las campañas desinformativas sobre los partidos y candidatos pueden variar más, en la medida en que los candidatos son distintos, pero las noticias falsas sobre el proceso electoral son muy parecidas en casi todos los sitios donde se celebran elecciones democráticas (The Carter Center, 2021).

En casi todos ellos aparecen, al menos, una vez:

- imágenes de urnas volcadas en la calle, con los votos esparcidos por el pavimento y un mensaje lamentando dónde acaba realmente la voluntad popular.
- vídeos de urnas electrónicas que no recogen adecuadamente el voto que se introduce en ellas.
- Actas con información falsa, visiblemente alterada o difícil de creer.
- Imágenes de miembros de la institución que organiza las elecciones cenando con candidatos, dando a entender una supuesta parcialidad del regulador.
- Protestas en las calles, el ejército tomando las plazas….

Estas y otras noticias manipuladas, descontextualizadas o directamente falsas suelen repetirse en casi todos los procesos electorales, y en casi todas ellas subyace una narrativa, una intención última. A fin de cuentas, si desinformar es difundir información manipulada al servicio de ciertos fines, los fines más habituales en un proceso electoral pueden resumirse en los siguientes :

- Afectar la **reputación** de un actor electoral (candidato, partido, regulador...)
- Cambiar la **intención de voto** o fomentar la **abstención.**
- Sembrar **dudas sobre el proceso** para rechazar su resultado e incluso reclamar su suspensión o repetición

- Otros: generar **incertidumbre**, asustar o **llamar a la violencia.**

Al igual que las historias analizadas por Propp podían descomponerse y clasificarse en protagonistas y funciones limitadas, las narrativas subyacentes en la desinformación electoral cuentan con algunos patrones y protagonistas que se repiten en elecciones de todo el mundo y que pueden servir para su identificación y detección. De esta premisa parte este trabajo que, al igual que hizo Propp hace un siglo, pretende reducir y descomponer las narrativas de la desinformación para entender su estructura. Una pretensión mucho más modesta que la del genio ruso.

Los protagonistas y las funciones de la desinformación electoral se han agrupado en cuatro bloques, respectivamente, y se han circunscrito a las instituciones y materias encargadas de organizar las elecciones, que son el elemento más crítico. Las narrativas -análogas a las funciones de Propp- que subyacen en los mensajes desinformativos y que pretenden conseguir los fines antes citados pueden clasificarse en cuatro grupos:

- Un candidato o partido no está preparado o no es honesto y por eso no debería gobernar.
- El regulador no está preparado para organizar el proceso, por falta de competencia u honestidad.
- La maquinaria electoral no funciona adecuadamente y puede influir de manera ilegítima en el resultado.
- Otros: injerencia externa, llamadas a la violencia, sucesos extraños relacionados con el proceso, etcétera.

3.3. Propuesta de personajes y narrativas electorales

Los principales actores que participan en un proceso electoral pueden clasificarse en cuatro grupos generales:

- Los **partidos y candidatos** que concurren a las elecciones.
- El **electorado** que elige con su voto a esos partidos y candidatos.
- El **regulador**, compuesto por una o más instituciones que organizan y gestionan el proceso electoral.
- Y los **influyentes**, aquellas personas que, a pesar de no participar directamente en el proceso, tienen una gran influencia sobre él.

Actualmente, las campañas desinformativas más peligrosas, y en las que se centrará este trabajo, son las que intentan únicamente desacreditar al regulador y al proceso electoral, al margen de candidatos y partidos. No son ni mucho menos las más numerosas, porque las narrativas sobre la honestidad o idoneidad de los candidatos y partidos son más habituales. Sin embargo, el efecto que puede tener el descrédito de los organismos reguladores suele ser mucho más pernicioso para el proceso electoral que lo que se diga de los candidatos.

Las narrativas sobre el proceso electoral suelen referirse a cuatro asuntos principales:

- La **parcialidad** del regulador, que intenta favorecer o perjudicar a determinados candidatos o partidos.
- El **proceso o la maquinaria electoral**, que funciona mal y no está recogiendo adecuadamente el voto de los electores o lo está dificultando y por tanto está interfiriendo en el proceso democrático, favoreciendo o perjudicando a determinados candidatos o partidos.
- El **recuento**, que está siendo manipulado para alterar el resultado electoral reflejo de la voluntad popular y favorecer o perjudicar a determinados candidatos o partidos.

- **Otros**: fuerzas externas o incluso extranjeras que están interfiriendo en el proceso electoral; y llamadas a la abstención o a la violencia, por medio de la rebelión o la intervención militar, etcétera.

El sistema propuesto en este trabajo propone identificar primero los protagonistas de un mensaje, para comprobar si encajan en la clasificación anteriormente enunciada. Estos protagonistas serian etiquetados como:

- **[candidato]** (para candidatos y partidos)
- **[regulador]** (para las instituciones que gestionan el proceso electoral)
- **[electorado]** el pueblo con capacidad de voto.
- **[influyente]** toda aquella institución, empresa o persona que pueda influir sobre el proceso sin participar directamente en él.

Esta labor de identificación es relativamente fácil, una vez se entrena al sistema en las diferentes denominaciones que pueden recibir los agentes de un proceso electoral. Más difícil es encontrar las funciones realizadas por estos actores electorales, que podrían clasificarse con las siguientes etiquetas:

- **Parcialidad**
- **Maquinaria**
- **Recuento**
- **Otros**

Cada uno de estos grupos contiene un número elevado de frases que podrían expresar la narrativa desinformativa. Clasificar el mensaje en uno de estos grupos es relativamente sencillo, pero encontrar las funciones dentro de la frase en las que la desinformación se mueve es algo más complejo.

¿Cómo reconocer entonces estas funciones en los mensajes de redes sociales? Las que propuso Propp suelen tener una estructura simple de sujeto, verbo y predicado: *alguien hace algo*., o más concretamente: *personaje hace función*.

Como se explicará con detalle más adelante, este sistema propone descomponer los mensajes en una estructura similar, encontrar el personaje o personajes implicados en el mensaje y aproximarse a las acciones descomponiendo el sujeto y el predicado del mensaje, separando sus nombres, verbos y complementos directos, entre otras características.

3.4. La colección de datos

El principal problema a la hora de crear un modelo de detección de noticias falsas en el ámbito electoral es que no existe un conjunto de datos adecuado para entrenarlo. En las misiones en las que ha participado el autor, ha necesitado recopilar una gran cantidad de entradas proveniente de diversas redes sociales. Así ha sido en Bolivia, Colombia, Honduras, Venezuela, Colombia, Ecuador y Brasil, en donde ha podido reunir decenas de millones de tuiteos y entradas de Facebook e Instagram.

Para este trabajo se ha decidido utilizar un conjunto de tuiteos del último país nombrado, Brasil, publicados durante las elecciones presidenciales y legislativas que tuvieron lugar el 2 y el 30 octubre de 2022. El hasta entonces presidente Jair Bolsonaro y el expresidente Luiz Inácio Lula da Silva concurrieron a unos comicios que ganó este último en un contexto de extrema polarización y desinformación que condujo, solo dos meses después, a un episodio de rebelión ciudadana. Los edificios oficiales de la Plaza de los Tres Poderes de Brasilia, sede de las principales instituciones del Estado brasileño, fueron asaltados por varios miles de ciudadanos que habían sido convocados mayoritariamente a través de las redes sociales. Por este motivo, se ha incluido también una muestra de interacciones desde el final de las elecciones hasta el 8 de enero de 2023, para abarcar también la desinformación previa e inmediatamente posterior a estos sucesos.

Este conjunto de datos se propone para la realización de este trabajo, como base de las transformaciones que se realizaran para encontrar en ellos los personajes y funciones, que al igual que los descritos por Propp, componen los relatos de la desinformación electoral.

Se trata de un *dataset* para el que se han seleccionado 3.376 entradas en el que se han incluido 77 mensajes desinformativos provenientes de las siguientes fuentes:

- 60 fueron recopilados de las principales organizaciones de verificación brasileñas - Aos Fatos, Boatos y Agência Lupa-, en las principales redes sociales que operan en Brasil. Fundamentalmente, Twitter, Facebook, Instagram y WhatsApp.

- 17 fueron localizados y contrastados a posteriori por el propio autor de este trabajo en una base de datos de más de 23 millones de interacciones sucedidas en Twitter durante el proceso electoral.

Algunos de estos mensajes fueron extraídos de los textos incluidos en las imágenes que acompañaban la entrada en redes sociales. Otros solo reflejan el texto que acompaña a una imagen o un vídeo con más información, aunque estos últimos no están incluidos.

Las variables de este primer conjunto de datos son:

- **texto_br**, con el texto en portugués brasileño
- **bulo**, con un valor de 1 ó 0, según sea o no un mensaje desinformativo, respectivamente.
- **tipo**, etiqueta numérica sobre el tipo de desinformación que contiene el mensaje:
    1. Parcialidad
    2. Maquinaria
    3. Recuento
    4. Otros

Todos los textos desinformativos incluidos en el dataset pertenecen a al menos alguna de las categorías que fueron descritas anteriormente, aunque de momento estas categorías no serán tenidas en cuenta para la labor de detección.

Las 3.299 entradas marcadas como no bulo (0) fueron seleccionadas aleatoriamente del conjunto de tuiteos general y revisadas una por una para asegurarse de que no contenían mensajes desinformativos. Como precaución adicional se comprobó que los mensajes seguían activos y se descartaron todos aquellos que hubieran sido borrados posteriormente o hubieran sido enviados por una cuenta suspendida después de la

publicación. Ambas cosas suelen ser indicio de una infracción de las normas que hasta hace pocos meses tenía Twitter sobre la publicación de información falsa.

Por su parte, el número de bulos correspondiente a cada categoría se reparte como se ve en la figura 1:

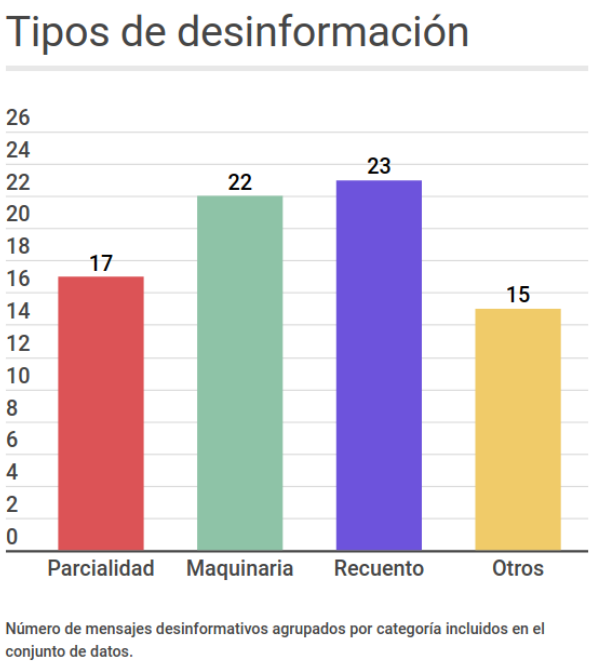

Figura1. Número de mensajes desinformativos del conjunto de datos, agrupados por categoría.

Se trata de un conjunto de datos de partida desequilibrado en el número de mensajes desinformativos, pero razonablemente equilibrado en el tipo de desinformación utilizada, y cuya frecuencia coincide aproximadamente con la que puede verse en las elecciones: los mensajes que denuncian fallos en el recuento son los más frecuentes y los de la categoría *Otros* los menos habituales.

3.5. Metodología

Como se indicó en la introducción, este sistema por sí solo no puede ni pretende reconocer de manera infalible todas y cada una de las frases con que un mensaje desinformativo introduce en el debate público -las idea de parcialidad del regulador, de fallos en la maquinaria o de manipulación del recuento-, pero sí encontrar un grupo de elementos gramaticales que sean comunes a cada una de estas narrativas y hallar así las consiguientes funciones, al estilo de las de Propp.

Se propone así utilizar un modelo clasificador para detectar estas funciones y otro para clasificar las narrativas dentro de una escala que permita regular una respuesta más o menos temprana según la gravedad que pudieran alcanzar los efectos del mensaje desinformativo.

El principal objetivo, para hacer analogía de la propuesta de Propp, es descubrir quiénes serían los protagonistas y cuáles las funciones en estos cuentos y leyendas que son los mensajes desinformativos.

En una primera aproximación a este problema vamos a partir de las siguientes premisas:

- el sujeto o los nombres son equivalentes a los protagonistas de Propp.
- el predicado de los mensajes, o incluso solo el verbo, es el equivalente a la función de Propp.

La estructura de los mensajes en redes sociales complica este análisis, por varios motivos:

- Propp partía de relatos de varias páginas, algunos muy detallados. Pero la escasa longitud de los mensajes de redes sociales no permite extraer un tema o función claros.
- En el mismo tuiteo puede haber más de un mensaje. Y cada mensaje puede representar distintas narrativas.
- A menudo se trata de frases que no disponen de un sujeto y un predicado. A veces contienen una sola palabra, una interjección, un insulto…

- En los casos en que sí existe sujeto y predicado, puede haber varios:
    - Si hay varios sujetos, puede inferirse que hay varios protagonistas.
    - Pero cuando hay varios predicados, hay que elegir la función principal del relato, o inferirla de las anteriores.
- Algunos tuiteos ni siquiera disponen de texto, solo perfiles de Twitter mencionados y enlaces a imágenes u otros recursos de red, que sí pueden contener algún mensaje desinformativo. O tal vez una o dos palabras con interjecciones, o una secuencia de emojis... Se han mantenido por rigor estadístico, aunque será difícil que sean detectados correctamente.

En ocasiones, será casi imposible extraer esta información de unos mensajes desinformativos indistinguibles de los legítimos. Un ejemplo de esto se aprecia en el tuiteo que muestra la figura 2, marcado como bulo en el conjunto de datos, que siembra la duda en el proceso, informando de que en algunas ciudades hay más votos que habitantes.

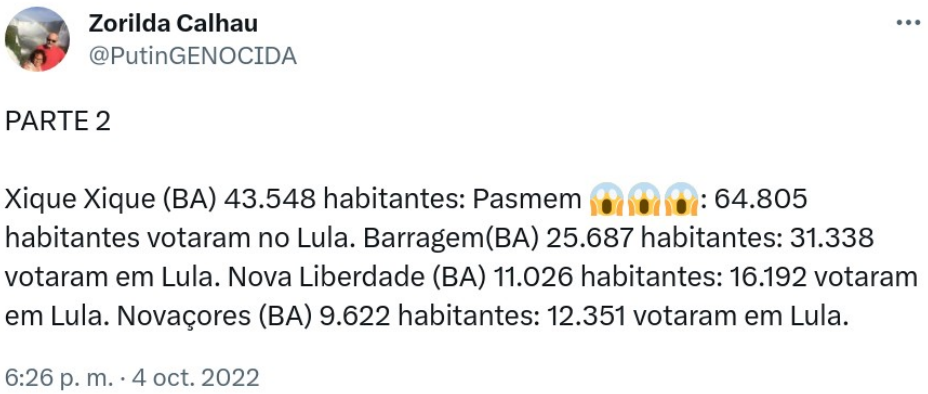

Figura 2. Tuiteo extraído del conjunto de datos que pone en duda el proceso electoral con una supuesta denuncia de inconsistencias entre el padrón electoral y el recuento de votos en varias ciudades.

Desde el punto de vista gramatical, este tuiteo es apenas distinguible de un avance sobre datos de participación por ciudades. La estructura también es complicada: Un nombre de ciudad, un guarismo y la palabra "habitantes". Y otra frase con un guarismo que hace las veces de sujeto, y un predicado ("*votaram em Lula*", votaron por Lula).

Esta estructura se repite varias veces, y solo se rompe con la palabra "*Pasmem*" (asombrado) y algunos emojis de sorpresa.

¿Quienes son los personajes de Propp en este caso? En principio podríamos pensar que los habitantes, representados por números. ¿Y las funciones? Los predicados: que "votaron por Lula". Pero incluso si pudiéramos descomponer correctamente esta información, ¿la reconocería el algoritmo? Este tipo de mensajes, a pesar de la dificultad que entrañan, se han mantenido en el conjunto de datos.

En definitiva, el objetivo último es asignar unos protagonistas y una función a partir de unos mensajes de texto que por lo general son breves y no siempre claros.

3.6. Transformación y creación de variables

El proceso de transformación consiste en la creación de nuevas variables, nuevas columnas en el conjunto de datos, a partir del tuiteo original. Estas variables pretenderán acercarse lo más posible a las ideas de *Personaje* y de *Función* formuladas por Propp y deberían coincidir lo más posible con el sujeto y el predicado de una frase.

Para el preprocesado, transformación y creación de nuevas variables se han utilizado las siguientes librerías de Python:

- Spacy [1], una biblioteca que proporciona varias funcionalidades para el procesamiento del lenguaje natural, como la *tokenización* de texto y el etiquetado gramatical.
- Textacy [2], construida sobre la anterior, esta biblioteca ofrece algunas funciones más adecuadas para este proyecto, como el reconocimiento de palabras clave, así como el sujeto y el predicado de oraciones.
- Bert Summarizer [3], una herramienta que utiliza la biblioteca de *transformers* Pytorch de HuggingFace para realizar resúmenes extractivos basados en un modelo Bert.

Las variables creadas a partir del tuiteo original **texto_br** en portugués brasileño fueron:

- **texto_anotado.** En el tuiteo original (**texto_br)** se sustituyeron los nombres de los diferentes actores electorales con los nombres comunes entre corchetes. Ejemplo:
    - Texto original:
        - Bolsonaro e Lula vão pedir recontagem de votos ao TSE para respeitar a vontade dos eleitores.

1 Spacy.io: Industrial-Strength Natural Language Processing. https://spacy.io/

2 textacy: NLP, before and after spaCy. https://textacy.readthedocs.io/en/latest/

3 bert-extractive-summarizer 0.10.1. https://pypi.org/project/bert-extractive-summarizer/

(Bolsonaro y Lula pedirán al TSE que se ha haga un recuento para respetar la voluntad de los electores).

- Texto anotado:
  - *[candidato] e [candidato] vão pedir recontagem de votos ao [regulador] para respeitar a vontade dos [electorado]*

    ([Candidato] y [Candidato] pedirán al [regulador] que se ha haga un recuento para respetar la voluntad de los [electorado])

Estas anotaciones serían análogas a los personajes de Propp.

A partir de estos dos campos se han creado diez más, siguiendo los criterios enumerados a continuación y aplicados por partida doble al texto original en lenguaje brasileño (b) y al texto anotado (a):

- **sujetos_b** y **sujetos_a**. Los sujetos que se encontraron en el tuiteo, separados por una barra vertical.
- **predicados_b** y **predicados_a**. Los predicados que se encontraron en el tuiteo, separados por una barra vertical.
- **svp_b** y **svp_a**. El sujeto el verbo y el predicado que se encontraron en el tuiteo, separados por una barra vertical.
- **nombres_b** y **nombres_a**, con los nombres propios o comunes que se encontraron en el tuiteo, lematizados.
- **verbos_b** y **verbos_a**, con los verbos que se encontraron en el tuiteo, lematizados.

Por último, el conjunto de datos incluye la etiqueta binaria del clasificador:

- **bulo**, con un valor de 1 ó 0, según fuera o no un mensaje desinformativo.

Al final del proceso, no en todos los tuiteos fue posible encontrar todos los campos. Por los motivos explicados anteriormente, no todos los mensajes tenían un sujeto o un predicado. En algunos casos, el modelo preentrenado en portugúes brasileño no fue capaz de identificar correctamente estructuras gramaticales sencillas. En otros, simplemente no existía esta información. En estos casos se mantuvieron los tuiteos originales por rigor estadístico, pues así era la muestra tomada del conjunto original de tuiteos de la campaña brasileña.

Las condiciones de las que parte el modelo no son fáciles, pero aunque en algunas casos la información complementaria sea escasa debería aportar algo al rendimiento del modelo, como se comprobará en el capítulo de Evaluación.

3.7. El modelo clasificador y su entrenamiento

Para crear el clasificador se ha utilizado un modelo de aprendizaje automático que parte de un modelo preentrenado BERT (Bidirectional Encoder Representations from Transformers) como base para realizar tareas de clasificación de texto (Devlin et al., 2018). Este sistema captura relaciones semánticas y contextuales complejas en un texto, y se preentrena utilizando un gran corpus general de procesamiento de lenguaje natural. Para esta tarea se ha utilizado el modelo base en lenguaje portugués brasileño con distinción de mayúsculas (bert-base-portuguese-cased).

El 70% del conjunto de datos fue dedicado al entrenamiento, mientras que un 10% se reservó para la validación y otro 20% para la prueba. El modelo fue entrenado con un tamaño de batch de 32 y un número de épocas no inferior a cuatro y no superior a ocho, con interrupción temprana (*early stopping*) cuando la perdida de validación superara la pérdida de entrenamiento.

Para entrenar el modelo, se han seleccionado arbitrariamente los campos a utilizar en grupos de uno, dos, tres, cuatro, cinco, seis o todas las variables, dando prioridad, por este orden, a:

- el texto original en portugués brasileño (**texto_br**).
- el texto anotado (**texto_anotado**) con los cuatro personajes propuestos (candidato, regulador, influyente y electorado).
- Los campos que recogieran el predicado o predicados o al menos los verbos (como posibles funciones: **verbos**, **predicados**, **svp**).
- Los campos que recogieran forma individual el sujeto o sujetos (como posibles personajes: **nombres** y **sujetos**)

La propuesta detrás de esta jerarquía es que para la tarea de detección debería ser más importante la función que los personajes. Es decir, los candidatos, el regulador, los influyentes y los electores pueden aparecer en numerosos tuiteos relacionados con temas muy diversos, pero es cuando se asocian a determinadas acciones (funciones de Propp análogas) cuando sube la probabilidad de que nos encontremos ante un mensaje desinformativo.

Capítulo 4

Evaluación

El modelo ha sido evaluado en dos grupos de pruebas, en cada uno de los cuales se han suministrado distintas variables para el entrenamiento del modelo. Ante la imposibilidad de hacer un barrido de todas las combinaciones de variables del modelo, se han elegido de manera arbitraria aquellas que fueran más análogas a la propuesta de Propp, es decir, aquellas que representen mejor los protagonistas (nombres y sujetos de las oraciones incluidos en el mensaje) y las funciones (verbos y predicados) en sus distintas variantes.

En los dos grupos de pruebas se ha incluido el tuiteo original y el tuiteo anotado, y se ha priorizado:

- en el primer grupo, el uso de nombres y sujetos (para comprobar la influencia de los elementos más similares a los personajes).
- En el segundo, el uso de nombres, sujetos, verbos y predicados (para comprobar la influencia en el rendimiento de todos los elementos: personajes y funciones).

El objetivo es corroborar la hipótesis de que aportar la información anotada de los personajes del mensaje y de las acciones que realizan (lo más parecido a funciones), puede mejorar el rendimiento del modelo.

4.1 Métricas

Como el conjunto de datos contiene muy pocos mensajes desinformativos y no está equilibrado (3.299 no bulos frente a solo 77 bulos) se ha descartado la precisión

(*Accuracy* en inglés) para evaluar el modelo. En su lugar se ha escogido la Puntuación F1 (*F1 Score*, en inglés), y sus submétricas de Precisión (*Precision*) y Exhaustividad (*Recall*) como métricas más informativas, sobre todo en conjuntos de datos desequilibrados como el que se está utilizando.

Y precisamente por tratarse de un conjunto de datos muy desequilibrado, los resultados del entrenamiento y evaluación del modelo tienen mucha variabilidad. En los conjuntos de prueba, suele haber entre 10 y 20 mensajes desinformativos, de características muy diversas, como ya se apuntó en el capítulo del conjunto de datos, lo que hace que la puntuación F1 fluctúe demasiado, entre los valores 0,5 y 1.

Para tener una mejor medida del funcionamiento del sistema, cada uno de los conjuntos de datos con cada una de las selecciones de variables se han entrenado 20 veces, siempre reordenados de manera aleatoria, y se han promediado los resultados obtenidos en estos entrenamientos. De este modo se han conseguido valores más constantes que permiten comparar mejor el impacto de incluir o excluir las variables.

Por último, mencionar que por tratarse de un sistema de alerta temprana, cuyos resultados serán comprobados posteriormente por un equipo de verificación, debería priorizarse la detección de bulos frente a la detección de no bulos. La puntuación F1 es una media armónica de la precisión y la exhaustividad (recall), y convendría que estas métricas fueran ambas altas y más o menos equilibradas, pero con algo de prevalencia de la exhaustividad. En definitiva, si el sistema llega a estar en funcionamiento, no importará tanto que al equipo lleguen varios no bulos calificados como bulos erróneamente, pues serán posteriormente descartados, como que varios bulos reales no lleguen al equipo de verificación por no haber sido detectados por el sistema. Si esos mensajes no detectados alcanzan luego mucha viralidad, el sistema propuesto en este trabajo perdería sentido como sistema de detección de alerta temprana.

4.2 Resultados

Las tablas 1 y 2 muestran las puntuaciones obtenidas al evaluar el modelo de detección seleccionando distintas variables en cada una de los grupos. Algunas consideraciones:

- Cada selección de variables se ha entrenado y evaluado 20 veces, para obtener una puntuación más estable.
- En ambas tablas, se muestra marcada en color azul la puntuación base, que es la conseguida al entrenar el modelo con el tuiteo original, sin procesar.
- Los resultados han sido ordenados inversamente según su puntuación F1, exhaustividad y precisión, respectivamente.

Grupo 1: Texto original, texto anotado, nombres y sujetos.

Variables	Precisión	Exhaustividad	**Puntuación F1**
tbr_tan_suja	0,95	0,69	0,79
tbr_tan_nomb	0,92	0,69	0,78
tbr_tan	0,95	0,68	0,78
tbr_tan_nomb_noma	0,94	0,66	0,77
tbr_tan_nomb_noma_sujb_suja	0,97	0,64	0,76
tbr	0,95	0,64	0,76
tbr_tan_noma	0,92	0,64	0,75
tbr_tan_sujb	0,96	0,63	0,75
tbr_tan_sujb_suja	0,94	0,61	0,73
tan	0,9	0,49	0,63

Tabla 1: Resultados de la evaluación del grupo 1.

Leyenda:
tbr: Texto del tuiteo original en portugués brasileño.
tan: Texto anotado con los nombres de los protagonistas.
nomb y **noma:** Nombres extraidos del tuiteo original y del anotado, respectivamente.
sujb y **suja:** Sujetos del tuiteo original y del anotado, respectivamente.

Grupo 2: Texto original, texto anotado, nombres, sujetos, verbos y predicados.

Variables	Precisión	Exhaustividad	Puntuación F1
tbr_tan_suja_vera	0,93	0,73	0,81
tbr_tan_noma_prea_svpb	0,93	0,72	0,8
tbr_tan_preb	0,92	0,72	0,8
tbr_tan_verb_prea_svpb	0,93	0,71	0,8
tbr_tan_preb_svpb	0,93	0,71	0,79
tbr_tan_suja_verb_preb	0,91	0,71	0,79
tbr_tan_suja_prea_svpb	0,94	0,7	0,79
tbr_tan_noma_verb_prea	0,91	0,7	0,79
tbr_tan_vera	0,96	0,69	0,79
tbr_tan_prea_svpa	0,95	0,69	0,79
tbr_tan_suja	0,95	0,69	0,79
tbr_tan_noma_verb_prea_svpb	0,97	0,67	0,79
tbr_tan_preb_svpa	0,88	0,71	0,78
tbr_tan_prea_svpb	0,93	0,69	0,78
tbr_tan_nomb	0,92	0,69	0,78
tbr_tan_nomb_noma_preb_svpa_prea_	0,9	0,69	0,78
tbr_tan	0,95	0,68	0,78
tbr_tan_noma_preb_svpa	0,93	0,68	0,78
tbr_tan_noma_preb	0,97	0,66	0,78
tbr_tan_verb_preb_svpb	0,9	0,7	0,77
tbr_tan_sujb_verb	0,9	0,69	0,77
tbr_tan_sujb_verb_prea	0,93	0,68	0,77
tbr_tan_noma_verb_preb_svpa	0,91	0,68	0,77
tbr_tan_nomb_noma_verb_prea_svpb	0,9	0,68	0,77
tbr_tan_suja_verb	0,94	0,67	0,77
tbr_tan_verb_preb_svpa	0,93	0,67	0,77
tbr_tan_sujb_preb	0,93	0,67	0,77
tbr_tan_suja_preb_svpb	0,93	0,67	0,77
tbr_tan_noma_prea	0,94	0,66	0,77
tbr_tan_nomb_noma	0,94	0,66	0,77
tbr_tan_noma_preb_svpb	0,93	0,66	0,76
tbr_tan_noma_verb	0,93	0,66	0,76
tbr_tan_verb	0,92	0,66	0,76
tbr_tan_suja_preb_svpa	0,91	0,66	0,76
tbr_tan_suja_prea	0,94	0,65	0,76
tbr_tan_noma_verb_preb_svpb	0,96	0,64	0,76
tbr	0,95	0,64	0,76
tbr_tan_noma_verb_prea_svpa	0,89	0,67	0,75
tbr_tan_verb_prea	0,9	0,66	0,75
tbr_tan_prea	0,93	0,65	0,75
tbr_tan_noma_verb_preb	0,93	0,65	0,75
tbr_tan_suja_preb	0,95	0,64	0,75
tbr_tan_sujb_vera	0,95	0,64	0,75
tbr_tan_verb_prea_svpa	0,93	0,64	0,75
tbr_tan_noma_prea_svpa	0,93	0,64	0,75
tbr_tan_noma	0,92	0,64	0,75
tbr_tan_sujb	0,96	0,63	0,75
tbr_tan_nomb_noma_sujb_suja	0,95	0,63	0,75
tbr_tan_nomb_noma_preb_svpa	0,89	0,65	0,74
tbr_tan_nomb_noma_verb_preb_prea_svpa	0,92	0,64	0,74
tbr_tan_sumb_suma_nomb_noma	0,92	0,64	0,74
tbr_tan_verb_preb	0,92	0,63	0,74
tbr_tan_sujb_prea	0,96	0,62	0,74
tbr_tan_nomb_noma_verb_prea_svpa	0,92	0,62	0,73
tbr_tan_sujb_suja	0,94	0,61	0,73
tbr_tan_suja_prea_svpa	0,94	0,59	0,72
tbr_tan_noma_vera	0,98	0,58	0,72
tan	0,9	0,49	0,63

Tabla 2: Resultados de la evaluación del grupo 1.

Leyenda:

tbr: Texto del tuiteo original en portugués brasileño.

tan: Texto anotado con los nombres de los protagonistas.

verb y **vera:** Verbos del tuiteo original y del anotado, respectivamente.

nomb y **noma:** Nombres del tuiteo original y del anotado, respectivamente.

sujb y **suja:** Sujetos del tuiteo original y del anotado, respectivamente.

preb y **prea**: Predicados del tuiteo original y del anotado, respectivamente.

svpb y **svpa**: Sujeto, verbo y predicado del tuiteo original y del anotado, respectivamente.

Por su parte el modelo para clasificar el tipo y gravedad del mensaje se entrenó únicamente con el texto original de los 77 mensajes desinformativos en portugués brasileño, etiquetados en los cuatro grupos (parcialidad, maquinaria, recuento y otros) y con un valor de entre 0 y 1 para representar su gravedad (poco grave y grave, respectivamente) . Este modelo obtuvo una puntuación F1 de 0,6.

Capítulo 5

Discusión

A continuación se interpretan los resultados obtenidos de acuerdo con la información utilizada para entrenar y evaluar el modelo en cada uno de los supuestos mostrados en las tablas del capítulo anterior:

- El rendimiento del modelo base, en el que solo se utiliza el tuiteo original, es bastante bueno, con una puntuación F1 de 0,76. Esto habla muy bien del modelo preentrenado BERT y del conjunto de datos de entrenamiento. También pone el listón muy alto a la hora de mejorar este resultado. Aun así, el sistema mejora este rendimiento en 5 décimas, hasta conseguir una puntuación F1 de 0,81, que es un valor muy notable para un sistema de detección de noticias falsas.

- En un principio cabría pensar que sustituir el tuiteo original por un texto anotado con los nombres de los personajes podría mejorar el rendimiento del modelo de detección, pero no es así. Quitar el tuiteo original hace que el modelo pierda información útil para esta tarea, lo que podría sugerir que el modelo estaba capturando algunos detalles del texto original que se perdieron en el anotado. Su puntuación F1 se reduce hasta 0,63, lo que supone 1,3 puntos menos que la puntuación base del tuiteo original.

- ¿Qué información puede haber perdido el modelo al anotar el tuiteo original y sustituir los nombres propios por nombres comunes? Para entenderlo puede utilizarse un ejemplo ficticio (un poco extremo): Si en lugar de decir "Moraes quiere limitar la campaña de Bolsonaro, ayudar a Lula, apelar al Tribunal Supremo y acorralar al Ejército", introducimos "[Regulador] quiere limitar la campaña de [Candidato], ayudar a [Candidato], apelar a [Influyente] y acorralar a [Influyente]", puede comprobarse de manera intuitiva que mucha información importante queda por el camino.

- Sin embargo, añadir este texto anotado al tuiteo original produce una mejora de dos décimas en el rendimiento, hasta 0,78. Esta variable sí aporta información al tuiteo original y supone una mejora del modelo de detección, pero solo de manera complementaria, no exclusiva. Todos los casos en los que el tuiteo original ha sido retirado del entrenamiento y sustituido por el tuiteo anotado han obtenido una puntuación por debajo de la puntuación base. O dicho de otro modo, retirar o sustituir el tuiteo original no solo no ha aportado nada al rendimiento, sino que lo ha deteriorado.

- Los resultados del grupo 1 demuestran que simplemente aportando la información de los personajes mejora el rendimiento del modelo hasta tres décimas. Pero los resultados del grupo 2, que incluyen también los verbos y predicados, aumentan en dos décimas más este resultado, hasta 0,81 de F1, con una exhaustividad de 0,73.

- La mejor combinación es aquella en la que simplemente se aporta el tuiteo original, el tuiteo anotado, y el sujeto y los verbos del texto anotado, que entre todos compondrían la forma más simple de representación de los personajes (sujeto) y función (verbos).

- En las mejores combinaciones siguientes aparece el tuiteo original, el tuiteo anotado, los verbos, los predicados y los sujetos, verbos y predicados unidos en una variable. Todas estas combinaciones superan la puntuación base.

- Incluir únicamente el predicado original aporta información útil al modelo. De hecho, la combinación sencilla del tuiteo original, el tuiteo anotado (personajes) y el predicado del tuiteo original (funciones) ha sido la que ha obtenido la tercera mejor puntuación. La combinación que incluye los verbos (parte de la función), el predicado anotado (con el complemento indirecto sustituido por el personaje) y el sujeto, verbo y predicado del tuiteo original, obtiene la misma puntuación F1, pero con una exhaustividad una décima por debajo.

- Por su parte, el modelo clasificador para la detección de la gravedad de los bulos, con una puntuación f1 de 0,6, no es muy eficaz, pero habla bien del modelo preentrenado BERT, si tenemos en cuenta que solo disponemos de

un conjunto de datos de entrenamiento y evaluación muy limitado (apenas 77 tuiteos etiquetados).

- Por último, y como se ha apuntado antes, cada una de las combinaciones que aparecen en la tabla han sido entrenadas y evaluadas 20 veces para obtener una media que palíe la variabilidad que produce un conjunto de datos tan limitado y desequilibrado. Este número de veces se ha elegido como compromiso entre lo posible y lo factible, pues un mayor número de rondas de entrenamiento y evaluación podría haber conseguido resultados más precisos, que cambiaran en una o, tal vez, dos décimas los actuales y que pudieran alterar ligeramente el orden de las combinaciones. Pero el tiempo necesario para hacer estos experimentos se habría extendido varias semanas o incluso un mes sobre el tiempo previsto.

Capítulo 6

Conclusiones y trabajo futuro

Después de intentar hacer analogía de la propuesta de Vladimir Propp y probar un modelo de clasificación que reconozca la desinformación en mensajes de redes sociales, pueden extraerse las siguientes conclusiones.

6.1. Conclusiones

Con el método propuesto se consigue mejorar la puntuación F1 base en cinco décimas, lo que supone un 6,6% de mejora. Esto confirma la hipótesis de que identificar en mensajes de redes sociales algunos personajes del proceso electoral (personajes, según Propp), separar las acciones que realizan (funciones) y suministrar esta información a un modelo de detección de mensajes desinformativos puede mejorar su rendimiento.

El conjunto de datos y el sistema de anotado, a pesar de los problemas mencionados anteriormente, han demostrado ser una base muy sólida para entrenar un modelo de detección de mensajes desinformativos. No obstante, sería conveniente ampliar el conjunto de datos para incorporar más mensajes desinformativos y reducir el desequilibrio con los legítimos. Por sí sola, esta medida debería mejorar aún más el rendimiento del modelo y la variabilidad de las evaluaciones. Además, serviría para mejorar el rendimiento del modelo de clasificación de la temática y la gravedad del mensaje desinformativo, que dispone de muy pocos datos de entrenamiento.

6.2. Trabajo futuro

Varias ideas han surgido durante el desarrollo de este trabajo que animan a probar nuevas líneas de investigación:

- Mejorar el sistema de anotación. Es relativamente sencillo sustituir los nombres propios que aparecen en una elección, que suelen reducirse a los candidatos, los partidos, las instituciones y funcionarios encargados gestionar el proceso y un grupo de influyentes más o menos extenso pero manejable con los recursos adecuados. Anotar las funciones o narrativas es más complicado porque el uso de palabras, verbos y complementos directos o indirectos posibles hace crecer exponencialmente las posibles anotaciones. Incluso así podría llegar a manejarse con los recursos necesarios, al tratarse de un dominio concreto y limitado como es el de los procesos electorales. Otra cosa sería intentar anotar las narrativas detrás de una campaña desinformativa sobre una persona, una institución o una empresa, sin un dominio con concreto, donde las posibilidades parecen casi ilimitadas.

- Ampliar el número de variables, para descomponer el sujeto y el predicado de cada una de las frases incluidas en los tuiteos. Y buscar nuevas variables que se aproximen a las funciones de Propp.

- Probar el modelo con un barrido completo de todas estas variables y ampliar el número de iteraciones para mejorar la precisión de los resultados finales, aunque esto signifique emplear muchas semanas de cálculo.

- Probar este sistema en otros idiomas, para comprobar si es tan independiente del lenguaje como parece. En principio, no debería haber obstáculos para que pueda aplicarse en otros idiomas con una gramática similar, porque el anotado inicial y la selección de variables se basa en la estructura, en la gramática, y no en el contenido, la forma o el contexto.

Bibliografía

Bastos, M. T., & Mercea, D. (2017). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, *37*(1), 38-54. https://doi.org/10.1177/0894439317734

Bharadwaj, P., & Shao, Z. (2019). Fake News Detection with Semantic Features and Text Mining. *International Journal on Natural Language Computing (IJNLC)*, *8*(3). https://ssrn.com/abstract=3425828

Carsten Stahl, B. (2006). On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective. *Informing Science: The International Journal of an Emerging Transdiscipline*, *9*, 083-096. https://doi.org/10.28945/473

Castellet, A., Varona, D., & Álvarez García, S. (2023). Verificadores en España: Una visión de su lógica de negocio. *Espejo de Monografías de Comunicación Social*, *13*, 119-136. https://doi.org/10.52495/c6.emcs.13.p99

Choudhary, A., & Arora, A. (2021). Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, *169*, 114171. https://doi.org/10.1016/j.eswa.2020.114171

Daelemans, W. (2013). Explanation in Computational Stylometry. En A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 7817, pp. 451-462). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37256-8_37

Davison, W. P. (1983). The Third-Person Effect in Communication. *Public Opinion Quarterly*, *47*(1), 1. https://doi.org/10.1086/268763

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/10.48550/ARXIV.1810.04805

Giglietto, F., Iannelli, L., Valeriani, A., & Rossi, L. (2019). 'Fake news' is the invention of a liar: How false information circulates within the hybrid news system. *Current Sociology*, *67*(4), 625-642. https://doi.org/10.1177/0011392119837536

Grieve, J., & Woodfield, H. (2023). *The Language of Fake News* (1.ª ed.). Cambridge University Press. https://doi.org/10.1017/9781009349161

Jadhav, S. S., & Thepade, S. D. (2019). Fake News Identification and Classification Using DSSM and Improved Recurrent Neural Network Classifier. *Applied Artificial Intelligence*, *33*(12), 1058-1068. https://doi.org/10.1080/08839514.2019.1661579

Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2022). What the fake? Probing misinformation detection standing on the shoulder of novelty and emotion. *Information Processing & Management*, *59*(1), 102740. https://doi.org/10.1016/j.ipm.2021.102740

Liu, Y., & Wu, Y.-F. (2018). Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11268

Manna, R., Pascucci, A., & Monti, J. (2020). *Profiling Fake News spreaders through Stylometry and Lexical Features. UniOR NLP @PAN2020.* Conference and Labs of the Evaluation Forum.

Martinez Monterrubio, S. M., Noain-Sánchez, A., Verdú Pérez, E., & González Crespo, R. (2021). Coronavirus fake news detection via MedOSINT check in health care official bulletins with CBR explanation: The way to find the real information source through OSINT, the verifier tool for official journals. *Information Sciences*, *574*, 210-237. https://doi.org/10.1016/j.ins.2021.05.074

Murayama, T., Wakamiya, S., & Aramaki, E. (2020). *Fake News Detection using Temporal Features Extracted via Point Process*. https://doi.org/10.48550/ARXIV.2007.14013

Naiara Galarraga. (2023, enero 8). Miles de partidarios de Bolsonaro asaltan el Congreso, la Presidencia y el Supremo de Brasil. *El País*. https://elpais.com/internacional/2023-01-08/cientos-de-partidarios-de-bolsonaro-invaden-el-congreso-de-brasil.html

Nisbet, E. C., Mortenson, C., & Li, Q. (2021). The presumed influence of election misinformation on others reduces our own satisfaction with democracy. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-59

Okunoye, O. B., & Ibor, A. E. (2022). Hybrid fake news detection technique with genetic search and deep learning. *Computers and Electrical Engineering*, *103*, 108344. https://doi.org/10.1016/j.compeleceng.2022.108344

Pierri, F., Piccardi, C., & Ceri, S. (2020). Topology comparison of Twitter diffusion networks effectively reveals misleading information. *Scientific Reports*, *10*(1), 1372. https://doi.org/10.1038/s41598-020-58166-5

Polo, F. P. (2019). Noticias falsas, desinformación y opinion pública en la Roma republicana. *Le Monnier Università*. https://www.academia.edu/42736691/Noticias_falsas_desinformaci%C3%B3n_y_opinion_p%C3%BAblica_en_la_Roma_republicana

Rogers, K., & Bromwich, J. E. (2016, noviembre 8). The Hoaxes, Fake News and Misinformation We Saw on Election Day. *The New York Times*. https://www.nytimes.com/2016/11/09/us/politics/debunk-fake-news-election-day.html

Sastrawan, I. K., Bayupati, I. P. A., & Arsa, D. M. S. (2022). Detection of fake news using deep learning CNN–RNN based methods. *ICT Express*, *8*(3), 396-408. https://doi.org/10.1016/j.icte.2021.10.003

Setiawan, R., Ponnam, V. S., Sengan, S., Anam, M., Subbiah, C., Phasinam, K., Vairaven, M., & Ponnusamy, S. (2022). Certain Investigation of Fake News Detection from Facebook and Twitter Using Artificial Intelligence Approach. *Wireless Personal Communications*, *127*(2), 1737-1762. https://doi.org/10.1007/s11277-021-08720-9

The Carter Center: Analyzing Bolivia's 2020 General Elections. Final Report. (2021). The Carter Center. https://www.cartercenter.org/resources/pdfs/news/peace_publications/election_reports/bolivia-2020-final-report.pdf

Tsai, C.-M. (2023). Stylometric Fake News Detection Based on Natural Language Processing Using Named Entity Recognition: In-Domain and Cross-Domain Analysis. *Electronics*, *12*(17), 3676. https://doi.org/10.3390/electronics12173676

Undeutsch, U. (1967). *Beurteilung der glaubhaftigkeit von aussagen. Handbuch der psychologie.*

Uppada, S. K., Patel, P., & B., S. (2022). An image and text-based multimodal model for detecting fake news in OSN's. *Journal of Intelligent Information Systems*. https://doi.org/10.1007/s10844-022-00764-y

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Ying, L., Yu, H., Wang, J., Ji, Y., & Qian, S. (2021). Multi-Level Multi-Modal Cross-Attention Network for Fake News Detection. *IEEE Access*, *9*, 132363-132373. https://doi.org/10.1109/ACCESS.2021.3114093

Yuan, C., Ma, Q., Zhou, W., Han, J., & Hu, S. (2020). Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users based on Weakly Supervised Learning. *Proceedings of the 28th International Conference on Computational Linguistics*, 5444-5454. https://doi.org/10.18653/v1/2020.coling-main.475

Zhou, X., Shu, K., Phoha, V. V., Liu, H., & Zafarani, R. (2022). "This is Fake! Shared it by Mistake":Assessing the Intent of Fake News Spreaders. *Proceedings of the ACM Web Conference 2022*, 3685-3694. https://doi.org/10.1145/3485447.3512264