Medición de Credibilidad en Plataformas de Redes Sociales

Tesis de Maestría

Sebastián García Parra

Programa de Posgrado en Ingeniería en Computación
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
Diciembre de 2024

Medición de Credibilidad en Plataformas de Redes Sociales

Tesis de Maestría

Sebastián García Parra

Tesis de Maestría presentada al Programa de Posgrado en Ingeniería en Computación, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magíster en informática

Director:
Dra. Adriana Marotta

Director académico:
Dra. Adriana Marotta

Montevideo – Uruguay
Diciembre de 2024

García Parra, Sebastián

Medición de Credibilidad en Plataformas de Redes Sociales / Sebastián García Parra. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2024.

XIII, 157 p.: il.; 29, 7cm.

Director:

Adriana Marotta

Director académico:

Adriana Marotta

Tesis de Maestría – Universidad de la República, Programa en Ingeniería en Computación, 2024.

Referencias bibliográficas: p. 127 – 135.

1. Calidad de Datos, 2. Redes Sociales, 3. Desinformación. I. Marotta, Adriana, . II. Universidad de la República, Programa de Posgrado en Ingeniería en Computación. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Dr. Alejandro Vaisman

Dra. Libertad Tansini

Dr. Carlos Luna

Montevideo – Uruguay
Diciembre de 2024

Agradecimientos

Este seguramente sea siempre el último capítulo a completar luego de un trabajo tan demandante como una tesis, cuando finalmente podemos mirar hacia atrás y reflexionar sobre todo el apoyo recibido a lo largo de este tiempo.

En este largo camino, lleno de altos y bajos, he contado con apoyos incondicionales que merecen ser plasmados en blanco y negro.

En primer lugar, a mi Directora Académica, Adriana. No es fácil describir en palabras la energía y dedicación que ha volcado para que este trabajo llegara a buen puerto. Su compromiso y entrega fueron mucho más allá de lo que su rol demanda. Su acompañamiento humano fue fundamental en los momentos de incertidumbre. De esto último, también, me llevo un gran aprendizaje.

Mi agradecimiento también a PEDECIBA por el apoyo brindado desde el inicio hasta la conclusión de esta tesis, sin el cual este trabajo no habría sido posible.

Sin lugar a dudas, mi más profundo agradecimiento es para los incondicionales de siempre. A mis hijos, Bruno y Julieta, y a mi esposa, Beatriz. Han resignado tiempo de su padre y esposo para que este trabajo se hiciera realidad. Este logro es tan mío como de ustedes.

Finalmente, agradezco a todas aquellas personas que, de una forma u otra, contribuyeron a que este proyecto viera la luz.

A todos ustedes, gracias.

Resumen

Esta tesis aborda la medición de la credibilidad en plataformas de redes sociales, un problema relevante en un contexto donde la desinformación impacta significativamente en la toma de decisiones de los usuarios. Las redes sociales han evolucionado hasta convertirse en fuentes primarias de información para millones de personas en todo el mundo.

La investigación se enfoca en el desarrollo de un modelo de calidad de datos que permite evaluar la credibilidad de las publicaciones mediante un análisis multidimensional que incorpora la confianza (*trustworthiness*), la reputación (*reputation*) y la verificabilidad (*verifiability*).

El análisis del *provenance* también desempeña un rol importante en este modelo, ya que permite rastrear el origen de la información, los cambios que ha sufrido durante su ciclo de vida y las personas o fuentes que la han compartido. Para medir el impacto del *provenance* en la credibilidad de una publicación, se ha desarrollado una nueva métrica denominada *Trustworthiness Path Stability*.

Como parte de este trabajo, se diseñó un flujo de procesamiento genérico, modular y adaptable, con la capacidad de integrarse con diferentes plataformas de redes sociales y de incorporar métricas de calidad definidas por terceros. Este enfoque permite ampliar y personalizar el análisis de credibilidad según las necesidades específicas del dominio o las características de las plataformas evaluadas.

Este flujo de procesamiento fue implementado y evaluado en un caso de uso concreto relacionado con publicaciones sobre el uso de estatinas en el tratamiento del colesterol. Los resultados obtenidos fueron validados por expertos en el dominio, mostrando una alta concordancia con sus evaluaciones y demostrando la efectividad del modelo propuesto.

Lista de figuras

Lista de tablas

Tabla de contenidos

Capítulo 1

Introducción

En los últimos años, las redes sociales han tenido un gran crecimiento, tanto en el número de usuarios como en la cantidad de contenido que generan. Las plataformas de redes sociales están aumentando, tomando diferentes formas y usos. Se pueden clasificar como weblogs, microblogs, redes sociales basadas en la ubicación, foros de discusión, wikis, plataformas para compartir imágenes y videos, comunidades de calificación y revisión y sitios de marcadores sociales. Estas plataformas pueden ser independientes o ser parte de otro sitio. Como ejemplo, los sitios de comercio electrónico tienen sus propias plataformas de redes sociales, para que los usuarios puedan intercambiar experiencias y evaluar productos para la compra final.

Las redes sociales se han posicionado fuertemente como un componente crítico del ecosistema de la información. Estas plataformas y aplicaciones están obteniendo una adopción generalizada con un alcance sin precedentes para usuarios, consumidores, votantes, empresas, gobiernos, academia y organizaciones sin fines de lucro.

El contenido que circula por las redes sociales es materia prima para la toma de decisiones, tanto para personas como para organizaciones. Los usuarios consultan las redes sociales antes de tomar una decisión de compra de un bien o servicio, optar que posición tomar ante una situación de salud, sanitaria o catástrofe natural, incluso como elemento de decisión para decidir el voto a un político. Las organizaciones y gobiernos se nutren de las redes sociales para medir su reputación, medir el impacto de campañas o de la aplicación de nuevas políticas públicas.

Los actores toman decisiones basadas en datos cuya calidad es dudosa, por

lo cual, el nivel de calidad de la información inferida no puede ser clasificada como confiable. Éste es un aspecto central en la sociedad del siglo XXI, en donde las fuentes de información se han comenzado a desplazar desde medios de prensa tradicional a las plataformas abiertas, alimentadas por usuarios y organizaciones independientes.

Los fenómenos asociados a la desinformación han generado graves incidentes a nivel mundial, provocando situaciones de pánico social Dinh y Parulian, 2020, poniendo en jaque sistemas democráticos Kušen y Strembeck, 2018 Grinberg et al. 2019 e incluso cobrándose vidas humanas Al-Rakhami y Al-Amri, 2020. Según Atske, 2021 el 48 % de los estadounidenses se informan a través de redes sociales. Un tercio de la población (31 %) declara informarse por medio de *Facebook*, un 22 % dice informarse regularmente por medio de *Youtube*. *X* e *Instagram* son fuentes regulares de información para el 13 % y 11 % de los norteamericanos, respectivamente. De acuerdo a Atske, 2021, estos datos evidencian la penetración de las redes sociales en el proceso de toma de decisión de la población.

En el Apéndice A se presenta una revisión detallada de trabajos que estudian el uso de los datos provenientes de redes sociales, en situaciones de emergencia, farmacovigilancia y en inspección de la salud pública.

A diferencia de los medios tradicionales (TV, radio, diarios) las redes sociales permiten que usuarios sin formación periodística puedan publicar y eventualmente masificar noticias y opiniones sobre hechos de la realidad, lo cual sin dudas es un gran avance en términos de libertad de expresión. Sin embargo, cuando el público no logra discernir sobre la credibilidad de las publicaciones, puede volverse un juego muy peligroso.

Datos de un estudio del año 2024 "Social media as a news source worldwide 2024", s.f., a nivel global, muestran que más del 60 % de los adultos que usan redes sociales en Chile, Colombia, México y Perú lo utilizan como fuente de noticias. De acuerdo al mismo estudio, la preocupación por las noticias falsas y la propaganda en las redes sociales no ha impedido que millones de usuarios accedan a sus redes favoritas diariamente. La mayoría de los *Millennials* en Estados Unidos usan redes sociales para informarse todos los días, y los consumidores más jóvenes en países europeos son mucho más propensos a utilizar redes sociales para noticias políticas nacionales que sus pares mayores. Leer noticias en redes sociales está convirtiéndose rápidamente en la norma para las generaciones más jóvenes, y esta forma de consumo de noticias probable-

mente aumentará aún más, independientemente de si los consumidores confían completamente en su red elegida o no.

Por otro lado, el mecanismo de compartir es una funcionalidad de absoluta importancia para las redes sociales, tanto sea para compartir dentro de la propia red como para compartir con otras redes. El modelo de negocio de estas plataformas Brown, 2021 depende fuertemente de la cantidad de contenido generado y compartido. Un escenario muy habitual es compartir una fotografía recibida en *Whatsapp* a *Instagram*, o viceversa. Estas casuísticas, que incluso pueden llegar a ser mucho más complejas, causan que para el usuario consumidor sea cada vez más complejo determinar la credibilidad del dato. En este contexto, se hace imperioso que el usuario consumidor disponga de herramientas que le permiten construir un criterio de confianza hacia una publicación generada por un tercero, cosa que permita que su accionar *online* y *offline* esté mediado por datos objetivos. El accionar *online* de un usuario implica todas aquellas acciones que éste puede disparar, por ejemplo, compartir la publicación con todos los contactos de su red. Potencialmente, esto último puede implicar difundir una noticia falsa de forma involuntaria. Por otra parte, el accionar *offline* se refiere a las acciones que el usuario tome por fuera de la red, incluso por fuera de los sistemas informáticos. Ejemplos de esto incluyen sustituir la toma de un medicamento en pro de un tratamiento alternativo no probado Swire-Thompson y Lazer, 2020, ejercer actos de violencia Gabel et al. 2022 o emitir un voto producto de información falsa Ross y Rivers, 2018.

La posibilidad de que los usuarios dispongan de medidas de calidad asociadas a los datos sobre los cuales basan sus decisiones se convierte en un requisito fundamental. Este es un campo muy poco explorado a nivel académico, que requiere de abordajes multidisciplinarios, tales como, Ciencias de la Computación, Sociología y Ciencias de la Comunicación.

1.1. Problema Planteado

Es fundamental que los actores que consumen datos producidos en redes sociales cuenten con información que les permita determinar el nivel de credibilidad del dato. Algunas redes sociales han incorporado, como metadatos de las publicaciones producidas por sus usuarios, indicadores que les pueden permitir a los consumidores inferir indicios sobre la veracidad del dato a consumir. A modo de ejemplo, las cuentas verificadas de X o el indicador de "mensaje

enviado muchas veces"de *Whatsapp* permiten a los usuarios tomar la decisión de cómo utilizar el contenido con el cual están interactuando de una manera más informada. En general, con estos mecanismos, las redes sociales persiguen limitar la circulación de noticias falsas o *fake news*.

Algunos estudios Chuai et al. 2023 han concluido que los mecanismos existentes son insuficientes, pero además, es importante notar que los usuarios suelen utilizar más de una red social. De acuerdo a Ruby, 2022, al momento de escribir este trabajo, los usuarios interactúan en promedio con 8.3 redes sociales. El número más destacado se da en India con 11.4 redes sociales por usuario, siendo 7.1 en Estados Unidos y 3.8 en Japón.

Dado que no existen estándares para el intercambio de información entre las diferentes plataformas de redes sociales, los metadatos mencionados anteriormente se pierden cuando el contenido es compartido de una red social a otra.

El problema planteado en esta tesis es la falta de mecanismos que, dada una publicación en una red social, permitan obtener información acerca de su credibilidad. Esta carencia dificulta la evaluación objetiva de la información, lo que puede derivar en la propagación de desinformación o en la toma de decisiones basada en datos no confiables.

1.2. Objetivos

El objetivo general de este trabajo es que los usuarios consumidores de publicaciones realizadas en redes sociales dispongan de información veraz que les permita inferir el nivel de credibilidad de las mismas. De esta forma, contar con la mayor cantidad de información posible antes de tomar una decisión que afecte su contexto online y/o offline

Los objetivos específicos son los siguientes:

- Proponer métricas claras y objetivas que permitan a los usuarios evaluar el nivel de credibilidad de las publicaciones.

- Diseñar un método para obtener el linaje (o *provenance*) de las publicaciones, es decir rastrear el origen y las transformaciones de éstas, incluso cuando son compartidas entre diferentes plataformas.

- Diseñar un flujo de procesamiento que permita la evaluación de la credibilidad de una publicación.
- Validar la propuesta mediante su aplicación a un conjunto de publicaciones previamente analizadas por expertos.

1.3. Contribuciones

Este trabajo presenta un modelo de calidad de datos enfocado en evaluar la credibilidad de publicaciones en redes sociales. El modelo incorpora diferentes dimensiones de calidad que permiten determinar, con la mayor exhaustividad posible, el nivel de credibilidad de una publicación. Las principales contribuciones son:

- Modelo de Calidad de Datos: Se definen dimensiones y factores de calidad para determinar la credibilidad de publicaciones. Es agnóstico de la red social, diferenciándose de trabajos previos enfocados únicamente en *X*.
- Reconstrucción de Linaje o *Provenance*: Se diseñan mecanismos para reconstruir el linaje de una publicación, permitiendo evaluar su credibilidad en distintas etapas del ciclo de vida, considerando el *cross-sharing* entre plataformas.
- Flujo de Procesamiento: Se propone un flujo de procesamiento modular que permite implementar el modelo de calidad y la reconstrucción del linaje en entornos productivos.

1.4. Organización del documento

El presente trabajo se organiza en los siguientes capítulos:

- **Capítulo 1: Introducción.** Se presenta el contexto general del trabajo, se define el problema abordado y se describen los objetivos generales y específicos.
- **Capítulo 2: Marco Teórico.** Se presentan los conceptos fundamentales que sustentan este trabajo.

- **Capítulo 3: Trabajos Relacionados.** Se realiza un análisis detallado de las investigaciones previas, relacionadas con la evaluación de credibilidad en redes sociales.
- **Capítulo 4: Propuesta del Modelo de Calidad.** Se introduce y discute el modelo de calidad propuesto. Se describen las dimensiones, factores y métricas empleadas para evaluar la credibilidad de las publicaciones en redes sociales.
- **Capítulo 5: Reconstrucción de Provenance.** Se describe el método desarrollado para reconstruir el *provenance* de las publicaciones, abordando los desafíos asociados al *cross-sharing* entre diferentes plataformas.
- **Capítulo 6: Implementación del Flujo de Procesamiento.** Se presenta y discute el diseño del flujo de procesamiento necesario para medir la credibilidad de una publicación.
- **Capítulo 7: Experimentación.** Se implementa una instancia del flujo de procesamiento. Se aplica el modelo propuesto a un conjunto de publicaciones clasificadas por expertos del dominio y se discuten los resultados obtenidos para validar su efectividad.
- **Capítulo 8: Conclusiones.** Se resumen las principales contribuciones del trabajo, se destacan los resultados alcanzados y se proponen posibles líneas de investigación futuras.

Capítulo 2

Marco Teórico

En este capítulo se presentan algunos conceptos básicos que dan el marco teórico para comprender el resto del documento. Los conceptos globales fundamentales sobre los que se construye esta tesis son: calidad de datos y linaje o *provenance* de los datos.

2.1. Tipos de información

Según Batini y Scannapieco, 2016 los tipos de información pueden clasificarse de acuerdo a su carácter perceptivo sensorial y por otro lado, por su carácter lingüístico. En cuanto al carácter sensorial, se pueden encontrar dibujos, mapas, imágenes, sonidos y videos. Respecto al carácter lingüístico, se pueden distinguir varios tipos de información, tales como:

- **Información estructurada**, esto es, información representada en términos de un conjunto de instancias y un esquema, los cuales están altamente acoplados e integrados. El esquema define la interpretación semántica y las propiedades de las instancias, tales como, tipos y restricciones de integridad. Las bases de datos relacionales son un claro ejemplo de este tipo de información.
- **Información semi-estructurada** se trata de información parcialmente estructurada, o en donde, existe un esquema descriptivo en lugar de prescriptivo. Los registros de los archivos XML o JSON son un claro ejemplo de información semi-estructurada. El esquema solo define estructuralmente una parte del registro, otros componentes del mismo pueden estar compuestos por datos sin estructura alguna.

- **Información no estructurada** incluye cualquier secuencia de símbolos, siendo un caso particular el lenguaje natural, en donde la semántica y la estructura no está restringida por ningún esquema.

La Web es un extraordinario motor de producción, difusión e intercambio de información. Según Batini y Scannapieco, 2016, esta información se puede clasificar de acuerdo a su estructura, uso o localización en la Web, lo cual resulta en las siguientes categorías:

- **Open data:** Datos accesibles al público de manera gratuita y sin restricciones de uso.
- **Linked open data:** Datos abiertos interconectados mediante estándares web, facilitando su integración y consulta semántica.
- **Deep Web:** Información no indexada por los motores de búsqueda convencionales, accesible solo a través de métodos específicos.
- **Surface web:** Parte de la Web accesible mediante motores de búsqueda tradicionales.
- **Social data:** Datos generados y compartidos en plataformas de redes sociales.
- **Big data:** Conjuntos de datos extremadamente grandes y complejos, que requieren tecnologías avanzadas para su almacenamiento, procesamiento y análisis.

Cabe destacar que estas categorías no son excluyentes, una misma pieza de información puede pertenecer a varias de estas categorías al mismo tiempo. A modo de ejemplo, cuando en nuestro trabajo hagamos referencia a una publicación en redes sociales, las categorías que aplican serán Surface Web, Social Data y Big Data.

Un término también relevante para nuestro trabajo es el *User-Generated Content* (UGC por sus siglas en inglés). Éste se refiere al contenido creado y compartido por usuarios finales en plataformas digitales. Este tipo de contenido puede tomar diversas formas, como texto, imágenes, videos, o combinaciones de estos, y se caracteriza por ser generado de manera descentralizada y sin una curaduría editorial formal. A este tipo de contenido, en el contexto de nuestro trabajo, lo denominaremos clip.

2.2. Calidad de Datos

Existen diferentes versiones del término Calidad de Datos, en nuestro trabajo usaremos la establecida por el estándar ISO/IEC 25012:2008 14:00-17:00, s.f. "el grado en que las características de los datos satisfacen las necesidades declaradas e implícitas, cuando se utilizan en condiciones específicas". Para poder medir este grado de adherencia respecto a las necesidades que se requieren de los datos, tanto el estándar como una amplia literatura asociada a la disciplina, definen lo que se conocen como dimensiones de calidad. La idea detrás del concepto de dimensión de calidad, es que puedan capturarse aspectos específicos, que en conjunto, definan la calidad del dato. Si comenzamos a analizar este concepto sobre datos estructurados, la calidad no solo aplica sobre las instancias de los datos, sino también sobre el esquema. Los datos de "baja calidad" impactan en la calidad de los procesos empresariales, mientras que un esquema de "baja calidad", por ejemplo con problemas de diseño que redunden en niveles insuficientes de normalización, da lugar a posibles redundancias y anomalías durante el ciclo del vida del uso de los datos.

2.2.1. Dimensiones de Calidad

Las dimensiones de calidad pueden referirse tanto a los valores que toman los datos, como a su intensión, esto es, a su esquema. Tanto las dimensiones de los datos como de los esquemas, suelen definirse en términos cualitativos, refiriéndose a propiedades generales de los datos y los esquemas. Más concretamente, las definiciones de las dimensiones no proporcionan medidas cuantitativas, por lo que una o varias métricas deben asociarse a las dimensiones como propiedades separadas y distintas. Para cada métrica, se deben de proporcionar uno o más métodos que permitan efectuar la medición, donde se definan que datos se incluyen, el mecanismo de medición y la escala en la que se expresan los resultados. Más adelante, en nuestro trabajo, haremos referencia explícita a las dimensiones y a las respectivas métricas asociadas.

En Batini et al. 2012 se definen clusters o grupos de dimensiones de calidad, dentro de cada cluster se incluyen aquellas dimensiones de calidad que se consideran similares. Los clusters definidos más relevantes son *Accuracy*, *Completeness*, *Accessibility*, *Consistency* y *Trustworthiness*. En otros trabajos de investigación, no se consideran los clusters, definiéndose directamente la dimensión de calidad como el máximo nivel de agregación posible. Sin embargo,

a los efectos de expresar con mayor claridad las características que describen las dimensiones de calidad, vamos a apoyarnos en las definiciones de Batini et al. 2012.

En esta sección nos centraremos solamente en comentar las dimensiones más directamente relacionadas con nuestro trabajo, que se encuentran en la literatura.

Credibilidad

Según Batini y Scannapieco, 2016 la credibilidad o *credibility* puede ser definida como la medida subjetiva de la creencia de un usuario de que un dato es verdadero. Por otra parte, para Fogg y Tseng, 1999 es una cualidad percibida, no reside en un objeto, una persona o una pieza de información. La percepción de *credibility* resulta de la evaluación de múltiples *subdimensiones* simultáneamente, haciendo especial énfasis en *Expertise* y *Trustworthiness*. Estas dos dimensiones se conjugan para formar la percepción general de *credibility* de una fuente de información. Una fuente puede ser vista como experta en un campo, pero si no se considera confiable, su credibilidad puede ser cuestionada, y viceversa. Al mismo tiempo en Simmhan et al. 2005 se afirma que el linaje puede utilizarse para estimar la calidad de los datos y la fiabilidad de los mismos basándose en los datos de origen y las transformaciones. Dicho en otros términos, la credibilidad está fuertemente asociada al linaje o *provenance* de los datos. De aquí en adelante, en nuestro trabajo, haremos referencia al término en inglés.

Según Fogg y Tseng, 1999 *credibility* se define simplemente como la capacidad de ser creíble. Personas creíbles son aquellas que son vistas como creíbles; información creíble es aquella que es vista como creíble. A lo largo de su investigación, los autores han encontrado que "*believability*" es un buen sinónimo de *credibility* en casi todos los casos. Aunque la literatura académica sobre *credibility* ofrece una visión más sofisticada, el significado esencial es similar al propuesto por ellos.

En lo que resta de este trabajo utilizaremos el término credibilidad para referirnos al concepto de que algo sea creíble, y el término *credibility* para referirnos a la dimensión de calidad específicamente.

Trustworthiness

De acuerdo a Batini y Scannapieco, 2016 la confianza es un nivel de probabilidad subjetiva y local con la cual un agente evalúa que otro agente realizará una acción particular. La confiabilidad es la probabilidad objetiva de que el fiduciario realice una acción particular en la cual dependen los intereses del confiador. En otras términos, la confiabilidad es la garantía de que un sistema funcionará como se espera. Aunque la confianza y la confiabilidad son dos conceptos distintos, al tratar con técnicas para evaluarlos, los dos conceptos a menudo desempeñan un único papel. En la medida de que aumentan la cantidad de sistemas de información, la confiabilidad comienza a ser un concepto relevante. La búsqueda de datos confiables está motivada por la fuerte necesidad de resolver conflictos entre información contradictoria de múltiples fuentes, en particular en el escenario de los datos obtenidos de la web. A modo de ejemplo, si se consulta la altura del Monte Everest en tres fuentes diferentes, ChatGPT, Google y Bing, se obtienen respectivamente los siguientes resultados 29.029 pies, 29.030 pies y 29.028 pies. Lograr inferir la fuente más confiable, a los efectos de obtener el resultado correcto, es el objetivo que persigue esta dimensión de calidad. La caracterización de la confiabilidad está vinculada con dos dimensiones, que serán profundamente estudiadas en las próximos capítulos de este trabajo: *Verifiability* y *Reputation*. A continuación, se presentan las primeras nociones conceptuales de estas dimensiones.

Reputation

La reputación es un juicio realizado por un usuario para determinar la integridad de una fuente Batini y Scannapieco, 2016 Gil y Artz, 2007. Puede estar asociada con una persona, organización, grupo de personas, o puede ser una característica de un conjunto de datos. Se basa en experiencias pasadas, tanto directas como indirectas (a través de recomendaciones). La reputación se puede medir a través de calificaciones o evaluaciones proporcionadas por los usuarios, que reflejan sus experiencias y la satisfacción general con la fuente de datos.

Verifiability

De acuerdo a Bizer, 2007 la *verifiability* se describe como el grado y facilidad con los que se puede verificar la corrección de la información. Por otra

parte en "Quality-Driven Query Answering for Integrated Information Systems", 2002 se refieren a *verifiability* como el grado en que un consumidor de datos puede evaluar la corrección de un conjunto de datos. Se describe como el "grado y la facilidad con la que se puede verificar la corrección de la información". De manera similar, Flemming, 2010 se refiere a la dimensión *verifiability* como el medio que se proporciona a un consumidor para examinar los datos en busca de corrección. Sin tales medios, la garantía de la corrección de los datos provendría de la confianza del consumidor en esa fuente. Aquí se puede observar que, por un lado, "Quality-Driven Query Answering for Integrated Information Systems", 2002 proporcionan una definición formal, mientras que Flemming, 2010 describe al factor, proporcionando sus ventajas y métricas. En particular, propone que la *verifiability* se mida (i) en base a información de procedencia, (ii) por la presencia de una firma digital, o (iii) por un tercero imparcial, si el propio conjunto de datos apunta a la fuente. La *verifiability* es una dimensión importante cuando un conjunto de datos incluye fuentes con baja *credibility* o *reputation* Batini y Scannapieco, 2016. Esta dimensión permite a los consumidores de datos decidir si aceptar o no la información proporcionada.

Expertise

Según Bourne et al. 2014 *expertise* se define como un nivel de desempeño elite, excepcionalmente alto, en una tarea específica o dentro de un dominio dado. Los individuos que alcanzan este estatus, denominados expertos, son reconocidos por su habilidad superior. La definición enfatiza que estos términos no solo identifican a alguien cuyo rendimiento es sobresaliente sino que también implican una causa subyacente, como el arduo trabajo y la capacitación prolongada.

De acuerdo a Kuutila et al. 2024 el *expertise* del autor de una publicación en redes sociales, se resalta como un factor fundamental que afecta las evaluaciones de *credibility*. En el contexto del estudio abordado por los autores de este trabajo, los clips con autores que poseen una mayor *expertise* (identificados por su formación y logros profesionales) son juzgados como más creíbles, independientemente de si la información es precisa o no. Esto subraya que la percepción de la *expertise* influye significativamente en la confianza que los lectores depositan en la información, especialmente en las plataformas de redes

sociales, donde la información varía ampliamente en precisión y calidad.

Este abordaje de *expertise* resalta su importancia como un factor de calidad que no solo se basa en la capacidad técnica o el conocimiento de un individuo, sino también en cómo estos atributos mejoran la percepción de la *credibility* entre el público general.

2.2.2. Modelos de Calidad

En Serra et al. 2022 se define el concepto de modelo de calidad, como una estructura que permite ordenar e instanciar los principales conceptos asociados a la medición de la calidad de los datos. En este trabajo se describen los cinco conceptos principales que conforman un modelo de calidad:

- **Dimensión de calidad:** Captura una faceta de alto nivel de la calidad. La calidad de los datos se pueden caracterizar utilizando múltiples dimensiones, sin embargo no dodas las dimensiones son de interés para una realidad concreta. En un caso de uso concreto y un conjunto de datos específico, es necesario seleccionar las dimensiones de calidad relevantes

- **Factor de calidad:** Representa un aspecto particular de una dimensión de calidad. A modo de ejemplo, la consistencia puede hacer referencia a la consistencia inter-relación o a la consistencia intra-relación. Pueden haber varios factores de calidad para la misma dimensión, ya que cada factor de calidad se adapta mejor a un problema o tipo de sistema particular.

- **Métrica de calidad:** es un instrumento para medición de un determinado factor de calidad. Por ejemplo, la proporción de datos nulos en un tabla es una métrica de calidad para el factor de calidad densidad. Pueden existir varias métricas de calidad para el mismo factor de calidad. Cada métrica tiene una granuralidad asociada, a modo de ejemplo, en los sistemas relacionales, la granuralidad puede ser una tabla, un atributo o una tupla.

- **Métrica de calidad aplicada:** Las métricas se pueden aplicar para evaluar la calidad de diferentes conjuntos de datos, en cada caso proporcionando nombres y parámetros específicos. Por ejemplo, la métrica de consistencia que verifica el cumplimiento de una dependencia funcional,

aplicada a datos geográficos, podría verificar que el código de la ciudad determine el estado.

- **Método de calidad aplicado:** Los métodos pueden aplicarse para evaluar la calidad de diferentes conjuntos de datos, proporcionando nombres, parámetros y algoritmos adecuados. Además, se definen dos tipos de métodos de calidad aplicados: (i) métodos de medición de calidad, que calculan la calidad de un objeto midiendo directamente (por ejemplo, contando el número de valores nulos en una tupla), y (ii) métodos de agregación de calidad, que calculan la calidad de un objeto compuesto. Por ejemplo, calcular la precisión de una tabla promediando la precisión de sus tuplas.

A modo de ejemplo, en la tabla 2.1 se define un modelo de calidad para medir la dimensión *accuracy*, considerando solo el factor *syntactic accuracy*. Este ejemplo fue obtenido de Serra et al. 2022.

2.3. Provenance

Según Buneman et al. 2001, el término *Provenance*, a veces denominado linaje, se refiere al origen de un dato y al proceso mediante el cual llegó a formar parte de una base de datos. Este concepto permite entender qué datos de entrada contribuyeron a generar un dato y de dónde proviene dentro de los datos de origen.

Intuitivamente, puede definirse como el proceso inverso de la difusión de información. La manera común de rastrear la procedencia es utilizando anotaciones que representan información sobre los datos, como comentarios u otros tipos de metadatos. Tradicionalmente, la procedencia de un elemento de datos informa su origen, derechos y propiedad, información que puede ayudar a los usuarios a formar un juicio sobre cuándo confiar, o no, en el contenido de un clip. Según Simmhan et al. 2005 la información de *provenance* o linaje, también puede mejorar la confianza y la verificabilidad de los datos. Los autores consideran que el *provenance* de los datos es un aspecto fundamental para evaluar su calidad, origen y los procesos a través de los cuales se han transformado y derivado, lo que a su vez influye directamente en su *credibility*.

Los autores de Buneman et al. 2001 distinguen dos tipos principales de *provenance*:

Tabla 2.1: Métricas de calidad y métodos de calidad para medir la exactitud sintáctica

Métrica de calidad	
Nombre de la métrica	SynAcc.dictionary.check
Descripción	Evalúa si un elemento de datos es sintácticamente correcto, comparándolo contra un diccionario
Granularidad	Valor
Dominio del resultado	{0,1}
Métrica de calidad aplicada	
Valores de los parámetros	citizenName, atributo de la tabla Population
Método de calidad	
Nombre del método	Check_value
Descripción	Evalúa la exactitud sintáctica de un string chequeando contra un conjunto de valores correctos
Tipos de parámetros	string <data>, atributo de una tabla relacional
Método de calidad aplicado	
Valores de los parámetros	Datos del atributo citizenName, atributo de la tabla Population
Algoritmo	`Check_value (data, attribute):` `  Return isInCollection(data, attribute)`

- **Why-Provenance:** Identifica qué datos de entrada influyeron en la existencia de un dato en el resultado de una consulta.
- **Where-Provenance:** Indica la ubicación exacta en los datos de origen de donde fue extraído un valor.

Si bien el *provenance* no se considera tradicionalmente una dimensión de calidad de datos, en el contexto de nuestro trabajo será tratado como tal debido a su relevancia para determinar la *credibility*.

2.4. El fenómeno de la desinformación

Internet se ha convertido en una fuente primaria de información para usuarios de todo el mundo. A pesar de existir una vasta cantidad de información útil y confiable, la creación y diseminación de contenido incorrecto o falso no para de crecer Leu et al. 2013. Este fenómeno está siendo profundamente estudiado. Según Wang et al. 2022 entre los años 2002 y 2021 se han publicado 5666 artículos asociados a esta materia, con una participación de 17.661 autores. Los términos en inglés asociados son *disinformation*, *misinformation* e *information pollution*. De acuerdo a los mismos autores, es posible agrupar los artículos más relevantes en cuatro grupos: heterogeneidad grupal de la desinformación en la memoria, mecanismos de desinformación en redes sociales, salud pública asociada al COVID-19 y aplicación de tecnología de *big data* en la generación de noticias falsas,

Los usuarios más inexpertos, tal como los niños en edad escolar, suelen correr un riesgo mayor de ser engañados por este tipo de contenido, por este motivo, la evaluación de la credibilidad es considerada un tema de alta prioridad por parte de los educadores. De acuerdo a Jenkins et al. s.f. la búsqueda web es una parte esencial de las "alfabetizaciones de los nuevos medios", considerada clave para los estudiantes. El autor identifica el juicio, esto es, "la capacidad de evaluar la confiabilidad y credibilidad de diferentes fuentes de información" como un componente clave del proceso. En Leu et al. 2011 se señala que la lectura en línea ha modificado el significado de la alfabetización, incorporando nuevas habilidades relacionadas con la Web, en especial discernir la credibilidad. Según Hargittai et al. 2010 este desafío se extiende más allá de la infancia, los adultos con educación universitaria tienden a visitar páginas web de baja credibilidad cuando realizan búsquedas, debido a la tendencia de

combinar fuentes, con clasificación alta, obtenidas de las listas de los resultados de búsqueda.

El fenómeno de la desinformación no solo se encapsula en la población jóven, es muy más amplio y profundo. La desinformación en el campo de política [Kušen y Strembeck, 2018, Atske, 2021, Wang et al. 2022, Grinberg et al. 2019] y la salud pública [Al-Rakhami y Al-Amri, 2020, Dinh y Parulian, 2020, Ramos, s.f., "Muerte y Discapacidad por Desinformación sobre las estatinas – Fundación Hipercolesterolemia Familiar", s.f.] son ejemplos detectados y muy bien estudiados, que están poniendo en jaque a instituciones democráticas y a la salud de la población mundial. A modo de ejemplo, el 22 de mayo de 2023 una falta noticia de un explosión en el Pentágono Oremus et al. 2023 generó un importante impacto social y económico. Esta noticia fue viralizada en la red X por cuentas verificadas.

Capítulo 3

Trabajos Relacionados

Como parte fundamental de este trabajo se analiza una amplia variedad de artículos relacionados con calidad de datos, en particular, aplicada a los datos generados por humanos (UGC): redes sociales, sitios de comercio electrónico y prensa digital a modo de ejemplo. De acuerdo a lo visto en el capítulo 2, estos datos pueden clasificarse como semi-estructurados y en cuanto a su carácter sensorial se puede encontrar un amplio abanico de opciones, como textos, imágenes, audios y videos, lo cual se contempla también en el análisis.

En este capítulo abordamos las dimensiones de calidad asociadas con credibilidad. Una característica interesante de las dimensiones de calidad que trataremos en este capítulo, es que se presentan diferentes definiciones para una misma dimensión, en algunos casos, presentando sensibles diferencias conceptuales. A diferencia de las dimensiones de calidad clásicas mencionadas en el capítulo 2 (como *accuracy*, *completeness*, etc.), las asociadas a *credibility* están teniendo un desarrollo muy importante en los últimos años.

3.1. La Dimensión de Calidad *Credibility*

Según Amintoosi y Kanhere, 2014 el término confianza representa el nivel de certeza sobre la confiabilidad de un usuario. En otras palabras, la confianza es un concepto de pares, que define la confianza de un usuario en la credibilidad, la capacidad o la solidez de otro usuario. Es importante enfatizar que la confianza es una medida subjetiva. La confiabilidad es una medida objetiva de la probabilidad de que el usuario emisor realice una acción particular de la cual dependen los intereses de los usuarios receptores Batini et al. 2015. Aunque

confiabilidad y credibilidad son conceptualmente diferentes, en particular, una se refiere a una medida subjetiva y la otra a una objetiva. Sin embargo, desde el punto de vista de las técnicas de evaluación, ambas se refieren al mismo objetivo. Por esta razón, es que en la literatura se suelen usar indistintamente, refiriéndose a un mismo concepto.

El campo de la *credibility* aplicada al ámbito online ha cobrado especial atención de investigadores, a partir de que las plataformas de redes sociales se han popularizado. Uno de los primeros estudios de revisión sobre *credibility* enfocados en entornos *online* es el de Shah et al. 2015. El objetivo principal de este trabajo es proporcionar una comprensión profunda de los factores que afectan la credibilidad web y explorar las técnicas utilizadas para su evaluación. Estos factores son, por ejemplo, el diseño y la estructura del sitio, la calidad del contenido, la experiencia de la fuente, los comentarios de los usuarios y la reputación del sitio. Se destaca cómo estos factores pueden influir en la percepción de credibilidad de los usuarios y se subraya la importancia de considerarlos en el proceso de evaluación. Este trabajo explora además las diferentes técnicas utilizadas para evaluar la credibilidad web. Estas técnicas pueden incluir encuestas a los usuarios, evaluaciones de expertos, análisis de contenido, análisis de reputación del sitio y aplicación de técnicas de aprendizaje automático. Como aspecto fundamental, se resalta la necesidad de utilizar un enfoque multidimensional para capturar la complejidad de la evaluación de la credibilidad web.

Posteriormente, en Choi y Stvilia, 2015 el objetivo principal es analizar de manera exhaustiva los aspectos clave relacionados con la evaluación de la credibilidad web, incluyendo su conceptualización, operacionalización, la variabilidad y los modelos utilizados. En cuanto a la conceptualización de la credibilidad web, es decir, cómo se entiende y define este concepto en el contexto de los sitios web, se discuten diferentes perspectivas y enfoques teóricos. Para esto, los autores se apoyan en definiciones clásicas de la credibilidad en la comunicación interpersonal Gaziano y McGrath, 1986, Hovland et al. 1953, Whitehead, 1968 y McCroskey y Teven, 1999, concluyendo que las dimensiones clave de la *credibility* son *Trustworthiness* y *Expertise*. Respecto a la operacionalización de la credibilidad web, es decir, cómo se puede medir y evaluar de manera práctica, se examinan diferentes métodos y técnicas utilizados, como escalas de evaluación, análisis de contenido y métodos experimentales. En lo que respecta a plataformas de redes sociales, el trabajo se enfoca en las ca-

racterísticas del perfil del usuario, especialmente en toda aquella información que permita validar su identidad *online*. Esto es, si hace referencia a perfiles de otras redes (ej: *LinkedIn*), blogs o sitios web personales.

Es relevante mencionar que, en Choi y Stvilia, 2015, se presentan seis modelos teóricos, todos ellos recogidos de trabajos anteriores. Estos modelos son utilizados para comprender y explicar la credibilidad web, proporcionando marcos conceptuales que ayudan a comprender los procesos cognitivos y sociales involucrados en la evaluación de la credibilidad. El principal aporte de este trabajo, es la comparación de estos modelos, utilizando para esto la definición de cuatros facetas: (1) contexto, (2) características del usuario, (3) características de la información y (4) procesos. Estas facetas son un aporte relevante para otros trabajos relacionados y para nuestro trabajo. Conceptualmente, la dimensión *credibility* se presenta como un concepto multidimensional.

Castillo et al. Castillo et al. 2011 fueron los primeros en trabajar en la credibilidad de X, estudiando la previsibilidad de los temas creíbles y de interés periodístico en esta red social. El método presentado en este artículo escala las ideas propuestas por los autores para encontrar temas que sean creíbles independientemente del tema de discusión. El resultado de su investigación, arroja que los siguientes factores de la red X son relevantes para determinar el nivel de credibilidad:

- **Reacciones y emociones de los usuarios:** Se analiza si los usuarios utilizan expresiones de opinión que representan sentimientos positivos o negativos hacia el tema en discusión. Esto permite evaluar la respuesta emocional generada por ciertos temas.

- **Nivel de certeza de los usuarios:** Se evalúa si los usuarios cuestionan la información que se les presenta o si la aceptan sin dudar. Esto proporciona información sobre la confianza que los usuarios tienen en la información compartida.

- **Fuentes externas citadas:** Se verifica si los usuarios citan fuentes externas específicas al compartir información. Se considera la credibilidad de estas fuentes, como si son dominios populares o si tienen una reputación establecida.

- **Características de los usuarios:** Se considera el número de seguidores que tiene cada usuario en la plataforma. Se reconoce que los usuarios

con una mayor cantidad de seguidores pueden tener un impacto más significativo en la difusión de información, lo que puede influir en su credibilidad.

Al mismo tiempo, los autores definen grupos de características para cuatros niveles de granularidad en cuanto a la información de X:

- **Características basadas en mensajes:** Consideran las características de los mensajes, como la longitud del mensaje, la presencia de signos de exclamación o interrogación, y el número de palabras de sentimiento positivo/negativo en el mensaje.
- **Características basadas en usuarios:** Consideran las características de los usuarios que publican los mensajes, como la antigüedad de registro, el número de seguidores, el número de personas que sigue y el número de tweets que han publicado anteriormente.
- **Características basadas en temas:** Son agregados calculados a partir de las características anteriores, como la fracción de tweets que contienen URLs, la fracción de tweets con hashtags y la fracción de sentimientos positivos/negativos en un conjunto.
- **Características basadas en propagación:** Consideran características relacionadas con el árbol de propagación construido a partir de los retweets de un mensaje, como la profundidad del árbol de retweets o el número de tweets iniciales de un tema.

Como se mencionó en este mismo capítulo, X es la red por excelencia seleccionada para la gran mayoría de los trabajos relevados. A pesar de que X es una red cuyo origen estuvo fuertemente asociado a compartir información en formato de texto, algunos trabajos abordaron la problemática de la credibilidad en otros formatos de información. Gupta et al. Gupta et al. 2013 se enfocaron en el análisis de la difusión de imágenes falsas en X durante el huracán Sandy en 2012 "Huracán Sandy", 2023. Los autores investigaron las características de estas imágenes falsas y propusieron métodos para identificarlas. Para llevar a cabo el estudio, se recopilaron millones de tweets relacionados con el huracán Sandy y se filtraron las imágenes asociadas. A través de un análisis exhaustivo, se identificaron las imágenes que eran falsas y se analizaron las características comunes que presentaban. Los resultados revelaron

que las imágenes falsas se compartían ampliamente y rápidamente durante el evento del huracán. Además, se encontró que estas imágenes falsas a menudo tenían atributos visuales y contextuales engañosos, como superposiciones de texto, imágenes manipuladas y fuentes no confiables. Los autores propusieron métodos para la detección automática de imágenes falsas en tiempo real. Estos métodos se basaron en el análisis de características visuales y el uso de técnicas de aprendizaje automático.

En un trabajo posterior, los mismos autores Gupta y Kumaraguru, 2012 se enfocaron en el desarrollo de un método para clasificar la credibilidad de los tweets durante eventos de alto impacto. Los autores destacan que durante estos eventos, la información que circula en X puede ser desafiante de verificar y puede haber una propagación masiva de rumores y noticias falsas. Para abordar este problema, el trabajo propone un enfoque basado en el análisis de características y aprendizaje automático para determinar la credibilidad de los tweets. Recopilaron un conjunto de datos de tweets relacionados con eventos de alto impacto y etiquetaron manualmente cada tweet con una puntuación de credibilidad.

A continuación, extrajeron características de los tweets, como la estructura del lenguaje, la veracidad de la información, la confiabilidad del usuario y la viralidad del tweet. Estas características se utilizaron para entrenar un modelo de aprendizaje automático (SVM) que clasifica los tweets en función de su credibilidad. Los resultados mostraron que el enfoque propuesto fue efectivo para clasificar la credibilidad de los tweets durante eventos de alto impacto. Se encontró que las características relacionadas con la veracidad de la información y la confiabilidad del usuario eran especialmente relevantes para determinar la credibilidad de un tweet. Este aspecto se estudia en profundidad en la propuesta de esta tesis.

Canini et al. Canini et al. 2011 investigó métodos para predecir qué usuarios de X serían considerados influyentes o expertos en temas particulares. Si un usuario de una red social está interesado en recibir información sobre un tema en particular, una tarea relevante es decidir a qué actualizaciones de otros usuarios suscribirse para maximizar la relevancia, credibilidad y calidad de la información recibida. Los autores llevaron a cabo un experimento para medir cómo diferentes factores de X afectan los juicios explícitos e implícitos de credibilidad.

Para llevar a cabo el experimento, los investigadores reclutaron participan-

tes de *Amazon Mechanical Turk* "Amazon Mechanical Turk", s.f., quienes eran usuarios activos y actuales de la red social *X*. Se les pidió a los participantes que juzgaran el valor de automóviles usados antes y después de ver las evaluaciones de varios usuarios terceros de *X*. Los evaluadores terceros variaban en su nivel de experiencia y sesgo, lo que permitió a los investigadores medir la influencia de estos factores en los juicios de credibilidad de los participantes. Al pedir a los participantes que evalúen el valor de los automóviles usados tanto antes como después de ver las evaluaciones de terceros, se pudo medir el efecto de la opinión de cada usuario de *X* en los juicios de los participantes, lo que proporcionó una calificación implícita de la credibilidad y experiencia percibida de los usuarios de *X*.

Basándose en los resultados del experimento, los investigadores desarrollaron un nuevo método para identificar y clasificar automáticamente a los usuarios de redes sociales según su relevancia y experiencia en un tema determinado. Este método involucraba combinar una búsqueda estándar de *X* con una técnica de clasificación social y un análisis de modelado de temas. El objetivo es generar una lista clasificada de usuarios creíbles para cualquier tema dado.

A pesar de que este trabajo no presenta arquitecturas ni marcos de referencia, expone resultados empíricos de gran valor para el desarrollo de nuestro trabajo, mostrando cómo la reputación de los usuarios tiene un peso preponderante en la credibilidad.

Los trabajos más relevantes de credibilidad en redes sociales, posteriores al año 2015, se enfocan en la diseminación de noticias y de eventos del mundo real o físico, tales como terremotos, erupción de volcanes o salud pública. Esto es, eventos que promueven fuertemente la interacción entre usuarios. Por otra parte, la red social utilizada para estos estudios es *X*. Que sea *X* la red seleccionada para estos estudios responde a dos motivos muy relevantes. El primero es que esta red es el sitio de *microblogging* más popular del mundo, registrando en el año 2023 una cantidad de 550 millones de usuarios Kolodny, 2023. Se ha convertido en una plataforma para la difusión de noticias, en la cual tienen presencia los medios de prensa y personalidades más influyentes del mundo. X, además de ser considerada como una red social, puede ser definida como una plataforma de noticias Sen y Morozov, s.f.

El segundo motivo, y para nada menor, es que se trata de una red que durante varios años ha desplegado una API pública y gratuita, a través de

la cual, académicos de todo el mundo han tenido al posibilidad de extraer millones de tweets para sus investigaciones. Esta política se mantuvo hasta febrero de 2023, desde esta fecha todo acceso a la API tiene un costo asociado.

En Arolfo et al. 2020 los autores abordan el problema de la calidad de los datos provenientes de *X*, desde una perspectiva de *Data Quality* tradicional (DQ), abordaje similar al que presentaremos más adelante en nuestro trabajo. Un aporte fundamental del artículo es la redefinición de las dimensiones de calidad clásicas, proponiendo cuatro dimensiones específicas: *Readability*, *Completeness*, *Usefulness* y *Trustworthiness*. Los autores realizan un análisis exhaustivo de la calidad de los datos, evaluando flujos de tweets en distintos escenarios: datos no filtrados, datos filtrados mediante palabras clave y datos clasificados según temáticas específicas. Se observa que los flujos filtrados muestran una mejora en las métricas de calidad respecto a los flujos no filtrados. Además, los autores destacan que los tweets generados por usuarios verificados por *X* presentan mayores niveles de *Trustworthiness*. Es importante destacar, que al momento de que este trabajo fue realizado, el mecanismo de verificación de cuentas era gratuito y con criterios de selección diferentes "Política de verificación antigua en X", s.f. a los vigentes al momento de escribir nuestro trabajo. Otro aporte relevante es el análisis de la calidad de los datos en función de los temas tratados. Por ejemplo, los tweets relacionados con política presentaron una calidad superior en comparación con aquellos asociados a deportes. Que el dominio tenga impacto en la calidad del dato, es un argumento que profundizaremos más adelante en nuestro trabajo.

En Alrubaian et al. 2019 se afirma que es muy complejo determinar la confiabilidad de los miembros de las redes sociales y del contenido que generan, utilizando para esto un enfoque específico en la red *X*. Los autores clasifican en dos categorías los fenómenos de desinformación: intencionada y no intencionada. En el caso de datos falsos creados accidentalmente, muchos usuarios promocionan y comparten noticias importantes sin verificarlas. Por lo tanto, un rumor puede transformarse rápidamente en contenido nuevo y de aspecto oficial que pretende ser una noticia real. El trabajo resume cuatro niveles de granularidad, sobre los cuales pueden ser aplicadas las definiciones y propiedades de la credibilidad:

- **credibilidad de una publicación individual (tweet):** Se refiere a la creencia de que una publicación individual (tweet) es creíble y confiable.

Esto implica que el mensaje incluye información relevante y precisa sobre un tema específico.

- **credibilidad del miembro:** Se refiere a la confiabilidad de una cuenta de usuario, medida mediante un puntaje calculado para cada usuario de la red. Cuanto menos confiable sea un miembro, menos probable es que los mensajes provenientes de esa cuenta sean confiables.

- **Credibilidad a nivel de tema:** Se refiere a la creencia, confiabilidad y aceptación de un tema o evento en particular, calculados como un puntaje numérico para cada tweet relacionado con ese tema o evento. Este concepto permite evaluar la credibilidad de la discusión en torno a un tema específico.

- **Credibilidad social:** Se refiere a la creencia esperada en la veracidad de un usuario en función de su estatus en X en un dominio temático específico, considerando todos los metadatos disponibles.

Se puede observar que en este trabajo se comparten los niveles de granularidad presentados en Castillo et al. 2011.

De acuerdo a Alrubaian et al. 2019 la evaluación de la credibilidad en redes sociales, puede ser abordado utilizando enfoques automáticos, basados en humanos o híbridos. Los estudios más recientes asociados a credibilidad en X, utilizan técnicas automáticas o semi-automáticas, incluyendo algoritmos de aprendizaje automático supervisados y no supervisados, así como también métodos basados en grafos.

En AlMansour y Communication, 2014 se propone un enfoque basado en aprendizaje automático, aplicado específicamente a la red X en países árabes. El aspecto más interesante de este trabajo, es el análisis del contexto como una propiedad de peso para la determinación de la credibilidad, tal como se analizó en Arolfo et al. 2020. En primer lugar, la cultura influye en cómo los individuos perciben y evalúan la información. Diferentes culturas tienen normas, valores y creencias distintos, lo cual puede afectar su confianza en ciertos tipos de información. Por ejemplo, lo que puede considerarse creíble en una cultura puede no ser percibido diferente en otra. Al incorporar el contexto cultural en la evaluación de la *credibility* de la información, los autores pueden obtener una comprensión más profunda de los factores que moldean los juicios de *credibility* de los usuarios.

En segundo lugar, la situación o contexto en el que se consume la información también afecta cómo se evalúa su credibilidad. Los usuarios pueden tener diferentes expectativas y requisitos de credibilidad según el contexto en el que encuentren la información. Por ejemplo, la información relacionada con emergencias o eventos políticos puede someterse a un mayor escrutinio en comparación con las noticias generales o los temas informales. Tomar en cuenta el contexto puede ayudar a desarrollar modelos más precisos para evaluar la credibilidad.

Además, el tema o dominio de la información compartida puede afectar su credibilidad percibida. Los usuarios pueden tener diferentes niveles de experiencia e interés en ciertos temas, lo cual puede influir en su evaluación de la información. Por ejemplo, las personas con experiencia en un campo particular, como la medicina, pueden ser más críticas y exigentes al evaluar información sobre ese tema. Comprender esta evaluación de credibilidad específica del tema puede ayudar a diseñar modelos de credibilidad más efectivos.

También, el lenguaje juega un papel crucial en la evaluación de la credibilidad, ya que afecta cómo se comunica y se entiende la información. La elección del lenguaje puede afectar la claridad, precisión y confiabilidad del mensaje. Además, la competencia lingüística puede influir en la capacidad de los usuarios para evaluar críticamente la información. Según este trabajo, considerar el factor del lenguaje es también un factor esencial para evaluar con precisión la credibilidad en diferentes contextos lingüísticos.

En Kuutila et al. 2024 se presenta una muy reciente investigación sobre como los usuarios de redes sociales infieren la credibilidad de un clip. Los autores reclutaron 844 participantes, utilizando para esto dos plataformas de *crowdsourcing*, *Prolific* "Prolific — Quickly find research participants you can trust", s.f. y *Toloka* "Powering AI with human insight - Toloka AI", 2022. Se solicitó a los participantes que evaluaran la credibilidad de diez publicaciones en redes sociales, resultando en un total de 8380 evaluaciones. El estudio tuvo como objetivo obtener una muestra geográfica y culturalmente amplia mediante el uso de estas dos plataformas de *crowdsourcing*, ya que los usuarios de *Prolific* residen principalmente en el Reino Unido y Estados Unidos, mientras que los usuarios de *Toloka* residen principalmente en Rusia y otros países ex soviéticos. Ya pudo observarse, según AlMansour y Communication, 2014, que este es un factor relevante en la determinación de la credibilidad. Las publicaciones evaluadas por los usuarios variaron en características de fuente, exactitud de

las afirmaciones y calidad de la evidencia. También se evaluaron sus creencias previas sobre los temas de las publicaciones antes de evaluar la credibilidad. Las publicaciones evaluadas se centraron en temas asociados a la salubridad de los alimentos, la carne roja procesada, la seguridad de las vacunas y la vitamina D. A pesar de que el trabajo se centra en clips con temáticas de salud, las conclusiones son extrapolables a otros dominios. Según los resultados de la investigación, factores como la experiencia de la fuente y las creencias previas de los usuarios influyen significativamente en la credibilidad percibida de publicaciones precisas e inexactas en redes sociales. En otros términos, los usuarios perciben que los clips generados por fuentes que se presentan como expertas tienen mayor credibilidad que los generados por fuentes inexpertas. Al mismo tiempo, la calidad de las evidencias que respaldan a la publicación tienen un impacto relativamente pequeño en las evaluaciones de credibilidad. Por esto, los usuarios tienden a juzgar la información consistente con sus creencias como más creíble que la información inconsistente con ellas, por lo que tienden a ignorar, hasta cierto punto, la calidad de la evidencia en su evaluación de credibilidad. Esto puede deberse a la resistencia a revisar las propias creencias, la incapacidad para diferenciar la calidad de diferentes tipos de evidencia, o el no valorar la investigación o la evidencia de expertos. Así, los usuarios pueden llegar a ser más vulnerables a la desinformación publicada por fuentes que se muestran o presentan como expertos. Esta conclusión será de carácter fundamental para nuestro trabajo, pues implicará darle un peso significativo a la evaluación objetiva del *expertise* de la fuente que genera el clip.

3.2. Provenance

En términos de redes sociales, la procedencia puede informar a los usuarios sobre la fuente de un clip. Las fuentes se refieren a los nodos que publican la información por primera vez Barbier et al. 2013. Este es un elemento clave que puede ayudar al usuario a hacer un juicio sobre la credibilidad del contenido del clip. La fuente podría ser una persona, una empresa, una organización o un proveedor de noticias; para cada tipo de *originador*, necesitamos definir e implementar atributos específicos que caractericen a la entidad fuente. Según Barbier et al. 2013, la pregunta principal es "¿Qué tipo de metadatos sobre la información recibida en redes sociales es útil para que un receptor identifique la procedencia de la información?". Para cada tipo de entidad, es necesario definir

atributos generales y específicos. Por ejemplo, para una cuenta de usuario, es importante tener el nombre de usuario, nombre, ocupación o edad como atributos generales. En dominios específicos, también se requieren atributos específicos, como la afiliación política del usuario. Una declaración política publicada por un candidato político o partidario puede ser evaluada con algún sesgo más que otras, esta información es relevante para el usuario final.

3.2.1. Medición de provenance

A los efectos de medir el *provenance* en el contexto de las redes sociales, resulta interesante partir del enfoque de Batini y Scannapieco, 2016. Los autores definen dimensiones de calidad asociadas a *provenance*, considerando tres perspectivas diferentes:

- ***Provenance* centrada en el agente**: es decir, qué personas u organizaciones participaron en la generación o manipulación de un recurso. Por ejemplo, en la *provenance* de una fotografía en un artículo de noticias, es posible identificar al fotógrafo que la tomó, a la persona que la editó y al periódico que la publicó.

- ***Provenance* centrada en el objeto**: rastreando los orígenes de partes de una entidad, ya sea un objeto o un recurso, hasta otras entidades.

- ***Provenance* centrada en el proceso**: capturando las actividades y los pasos realizados para generar un recurso.

A los efectos de nuestro trabajo, resulta relevante considerar las perspectivas centradas en el agente y la centrada en el objeto, siendo la primera el usuario de la red social y la segunda el clip. La perspectiva centrada en el proceso, queda por fuera del alcance de nuestro trabajo.

Según los mismos autores, las dimensiones clave relacionadas con *provenance* son contenido, gestión y uso. En el cuadro 3.1 se detallan las dimensiones que definen.

Tabla 3.1: Dimensiones del *provenance* de datos web

Categoría	Dimensión	Descripción
Contenido		
	Attribution	*Provenance* como las fuentes o entidades que fueron usadas para crear un nuevo resultado. Incluye detalles sobre responsabilidad y origen del contenido.
	Process	El *provenance* como el proceso que generó un artefacto, considerando su reproducibilidad y el acceso a los datos.
	Evolution and versioning	Cómo el contenido ha sido republicado y actualizado con el tiempo.
	Justification for decisions	Las razones o argumentos subyacentes en las decisiones tomadas respecto al contenido.
	Entailment	Los axiomas o tuplas que llevaron a los resultados específicos obtenidos.
Gestión		
	Publication	Disponibilidad de la información de *provenance* para su divulgación.
	Access	Acceso y consulta de la información de *provenance*.
	Dissemination control	Políticas de diseminación controlada especificadas por el creador del artefacto.
	Scale	Gestión de grandes volúmenes de información de *provenance*.
Uso		
	Understanding	Comprensión del *provenance* por parte del usuario final.
	Accountability	Verificación de *provenance* y rendición de cuentas.
	Interoperability	Combinación de *provenance* generada por diferentes sistemas.

Continúa en la siguiente página

Tabla 3.1 – *Continuación de la página anterior*

Categoría	**Dimensión**	**Descripción**
	Comparison	Comparación para encontrar lo común en el *provenance* de diferentes entidades.
	Trust	Evaluación de la confianza basada en el *provenance*, considerando la calidad de la información y la reputación.
	Imperfections	Manejo del *provenance* incompleto, inciertaoo erróneo.
	Debugging	Uso del *provenance* para la detección de errores o fallos en los procesos.

En cuanto a Contenido, la medición de las dimensiones como *Attribution*, *Process*, *Evolution and versioning*, *Justification for decisions* y *Entailment* aportan a la evaluación y comprensión de la credibilidad de una publicación. A continuación se analiza cada una de ellas:

Attribution: Conocer quiénes son las fuentes o entidades detrás de un contenido permite verificar la autoridad y la fiabilidad de la información. Si el contenido proviene de una fuente respetada y con autoridad en el tema, esto puede aumentar la credibilidad de la publicación.

Process: Comprender cómo se generó la información, incluyendo los métodos, procedimientos y datos subyacentes, brinda transparencia y permite a otros reproducir los resultados. Una metodología sólida y transparente respalda la fiabilidad de los resultados publicados.

Evolution and versioning: Saber cómo ha evolucionado un contenido a lo largo del tiempo y cómo se ha re-publicado puede señalar la vigencia y relevancia de la información. Las actualizaciones frecuentes y basadas en datos recientes pueden aumentar la confianza en su actualidad y precisión.

Justification for decisions: Proporcionar las justificaciones o el razonamiento detrás de las decisiones editoriales o metodológicas puede fortalecer la confianza en la integridad editorial y en la selección de contenidos de la publicación.

Entailment: Identificar los axiomas o las bases lógicas que llevaron a los resultados específicos agrega un nivel de rigurosidad científica. Esto permite a los usuarios entender el razonamiento detrás de los hallazgos y determinar si la lógica subyacente es sólida.

Las dimensiones de la categoría Gestión, conjuntamente, buscan garantizar que la información de *provenance* sea accesible, manejable y utilizable, respetando al mismo tiempo los derechos y *restricciones* asociados con ella. Revisamos cada una de ellas:

Publication: Se refiere a la disponibilidad de la información de *provenance* y cómo esta se pone a disposición para su uso y distribución. Esto incluye la exposición de la *provenance* a los interesados y la forma en que se distribuye entre los usuarios y sistemas.

Access: Trata sobre la capacidad de localizar y acceder a la información de *provenance*. Implica encontrar y recuperar datos de *provenance* relevantes, lo cual es fundamental para poder utilizar la información de manera efectiva.

Dissemination Control: Se ocupa de las políticas y restricciones que el creador de un recurso pone en cuanto a cómo y cuándo se puede utilizar la *provenance*. Esto asegura que se respeten los derechos de autor y otras consideraciones legales y éticas relacionadas con el uso de la información de *provenance*.

Scale: Aborda los desafíos asociados con el manejo de grandes cantidades de información de *provenance*. La capacidad de procesar y manejar datos de *provenance* a gran escala es crucial, especialmente con el crecimiento exponencial de datos en entornos web y de *big data*.

Todos estos factores de calidad están asociados a la propia estructura de *provenance*. Dado que el enfoque de nuestro trabajo se focaliza en la calidad de un clip, está por fuera del alcance la calidad de la metadata generada. Sin embargo, en la sección 6, en la que se introduce la gestión del *provenance*, se analizan y definen algunos de estos aspectos, pero sin una mirada de calidad. A modo de ejemplo, se asegura que la publicación del *provenance* generado sea simple, utilizando para estos estándares mundialmente conocidos y adoptados,

sin embargo no se hará hincapié en su calidad. Por lo tanto, todos estos factores de calidad quedan por fuera del alcance de nuestro trabajo como parte de modelo de calidad de *credibility*. Sin perjuicio de esto, nos resulta importante su estudio de cara a las posibles líneas futuras de investigación.

La categoría Uso refiere a cómo los usuarios finales interactúan con y se benefician de la información de *provenance*. Esta categoría aborda distintos aspectos que determinan la aplicabilidad práctica del *provenance*:

Understanding: Esta dimensión se centra en la facilidad con la que los usuarios pueden comprender la información de *provenance*. Busca garantizar que los datos sean presentados de manera que los usuarios no especializados puedan interpretar correctamente el historial y el contexto de la información o del recurso en cuestión.

Accountability: Se relaciona con la capacidad de los usuarios para verificar la *provenance* y por ende, la autenticidad y la integridad de los datos. Esto es esencial para que los usuarios puedan responsabilizar a las entidades correspondientes por la información proporcionada.

Interoperability: Enfatiza la importancia de que la información de *provenance* pueda integrarse y utilizarse entre diferentes sistemas y plataformas, lo que es fundamental en un entorno web diverso y en aplicaciones de *big data*.

Comparison: Se refiere a la capacidad de analizar y encontrar similitudes en la *provenance* de diferentes entidades o conjuntos de datos. Esto es crucial cuando se quieren comparar y contrastar la calidad y la fiabilidad de distintas fuentes de información.

Trust: Esta dimensión aborda la construcción de la confianza en la información basada en su *provenance*, teniendo en cuenta la reputación y la fiabilidad de las fuentes.

Imperfections: Reconoce que la información de *provenance* puede ser incompleta, incierta o incluso errónea, y destaca la necesidad de herramientas y métodos para gestionar estas imperfecciones.

Debugging: Se relaciona con la utilización de *provenance* para identificar y corregir errores en el proceso de generación y manejo de datos, lo cual es un aspecto crítico en la garantía de calidad y en la depuración de datos.

En esta categoría aparecen dimensiones que evalúan exclusivamente la usabilidad del *provenance*. Algunas de estas dimensiones son estudiadas en el capítulo 6, pero sin considerar métricas de calidad asociadas. Tal como se destacó en la categoría gestión, en el capítulo 6 se aborda el uso de estándares mundiales para la gestión del *provenance*, por lo cual, los factores *Understanding*, *Interoperability* y *Comparison* son parte de los objetivos del diseño del estándar, siendo este un aspecto heredado por nuestro trabajo. De todas formas, no se consideran dimensiones de calidad para medir la *performance* de estos aspectos, por lo que estrictamente estas dimensiones están por fuera del alcance de nuestro trabajo. En cuanto a *Debugging* e *Imperfections* no serán abordados en nuestro trabajo, también, por estar fuera del alcance de nuestra investigación.

En cuanto a la dimensión *Trust*, hace referencia a como los usuarios finales pueden utilizar la información de *provenance* para inferir los niveles de confianza sobre una pieza de información. Ya hemos discutido profundamente la relación entre *provenance* y la confianza en los datos, pero no hemos abordado aún la necesidad de que el *provenance* sea expresado de una forma clara hacia los usuarios finales. En el capítulo 6 nos apoyaremos para esto en las recomendaciones de los estándares, pero este punto es tan relevante que requerirá de un serio análisis posterior. Cabe aclarar que la definición de métricas sobre el *provenance* no es suficiente para cumplir el requisito planteado sobre la dimensión *Trust*.

3.2.2. El modelo PROV-DM

Una referencia fundamental dentro de *provenance*, es el trabajo llevado adelante por la W3C (World Wide Web Consortium), con la definición del estándar PROV-DM.

El modelo PROV-DM, abreviatura de *Provenance Data Model*, es un marco de representación y descripción del *provenance* de los datos en diferentes contextos. Proporciona un enfoque estructurado y semántico para capturar y representar la información de *provenance*.

El objetivo principal de PROV-DM es establecer un marco común y estandarizado para el intercambio, integración e interpretación de la información de *provenance* en diferentes aplicaciones. Al adoptar PROV-DM, se facilita la captura y comunicación efectiva del *provenance* de los datos, lo cual contribuye

a la reproducibilidad, validación e interpretación de los resultados.

El modelo se basa en registros y relaciones, permitiendo representar el *provenance* en términos de entidades, actividades y agentes involucrados en la generación, manipulación y uso de los datos. Incorpora además, conceptos como roles, atributos y relaciones.

En la figura 3.1 se observa un esquema con las entidades fundamentales de PROV-DM.

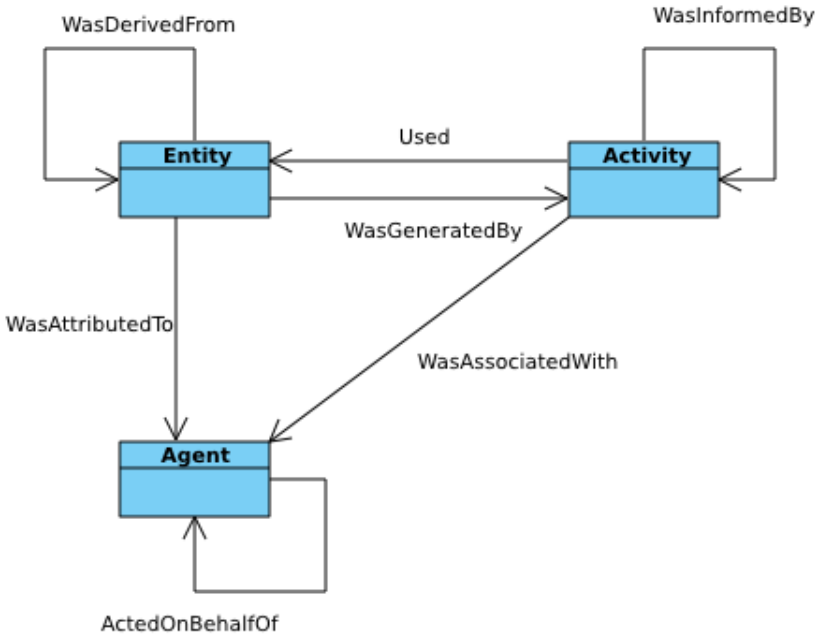

Figura 3.1: Conceptos fundamentales de PROV-DM

- **Entidades**

En PROV, las *cosas* cuya procedencia se quieren describir se denominan entidades y tienen algunos aspectos fijos. El término "cosas" abarca una amplia diversidad de nociones, incluidos objetos digitales como un archivo o una página web, cosas físicas como una montaña, un edificio, un libro impreso o un auto, así como conceptos e ideas abstractas.

- **Actividades**

Una actividad es algo que ocurre durante un periodo de tiempo y actúa sobre o con entidades; puede incluir consumir, procesar, transformar, modificar, reubicar, utilizar o generar entidades. Así como las entidades abarcan una amplia

variedad de nociones, lo mismo sucede con las actividades. Las actividades de tratamiento de la información pueden, por ejemplo, mover, copiar o duplicar entidades digitales; las actividades físicas pueden incluir conducir un auto entre dos lugares o imprimir un libro. Las actividades y las entidades se asocian entre sí de dos maneras diferentes: las actividades utilizan entidades y las actividades producen entidades. El acto de utilizar o producir una entidad puede tener una duración. El término "generación" se refiere a la finalización del acto de producir. Del mismo modo, el término "utilización" se refiere al comienzo del acto de utilizar entidades.

- **Agentes**

Para muchos propósitos, una consideración clave para decidir si algo es fiable y/o digno de confianza es saber quién o qué ha sido responsable de su producción. Una afirmación de un científico conocido con una trayectoria establecida puede ser más creíble que la de un estudiante. Un agente es una entidad que asume algún tipo de responsabilidad por una actividad que tiene lugar, por la existencia de una entidad o por la actividad de otro agente. Un agente puede ser un tipo concreto de entidad o actividad. Esto significa que el modelo puede utilizarse para expresar la procedencia de los propios agentes.

3.2.3. La extensión PROV-SAID

En Taxidou et al. 2015 se propone un modelo para la difusión de información y el *provenance* en las redes sociales. Los autores justifican su trabajo en el hecho de que las redes sociales no proporcionan mecanismos adecuados para el *provenance* de la información, lo cual es importante para juzgar su relevancia y credibilidad.

La difusión de información se refiere al proceso mediante el cual una pieza de información se propaga a través de las redes sociales y otros medios de comunicación digitales. Este concepto se basa en la capacidad de las personas para compartir y transmitir información a través de redes sociales, lo que permite que la información se extienda rápidamente a un gran número de personas. Dado que en las redes sociales existe una variedad de opiniones y múltiples fuentes de información, es fundamental evaluar la credibilidad de dicha información. El conocimiento sobre cómo se propaga una pieza de información en las redes sociales brinda un contexto adicional que incluye la fuente

y sus características, los intermediarios involucrados y las modificaciones que ha sufrido dicha pieza. Los usuarios de redes sociales pueden aprovechar este contexto para evaluar la credibilidad de la información. De acuerdo a Kwon et al. 2013 la detección de rumores es posible no solo mediante la identificación de las fuentes, sino también mediante el análisis de las características del proceso de difusión y los pasos intermedios. Partiendo de la definición de difusión de información como referencia, es posible definir el *provenance* como el proceso inverso. Esto es, el mecanismo que permite rastrear los caminos hasta las fuentes originales.

En Taxidou et al. 2015 Taxidou et al. 2018 Taxidou, 2018 se propone un modelo para la representación del *provenance*, llamado PROV-SAID (**P**rovenance of **S**ocial medi**A** **I**nformation **D**iffusion). Este modelo extiende el modelo PROV de la W3C para mejorar su expresividad y cubrir diferentes casos de uso en la difusión de información en las redes sociales. PROV-SAID incluye conceptos como mensajes originales, mensajes copiados y mensajes revisados, los cuales representan diferentes formas en las que la información puede propagarse en las redes sociales. Además, se introduce el concepto de influencia, el cual juega un papel clave en la difusión de información en las redes sociales. PROV-SAID define diferentes tipos de influencia, como las relaciones de seguimiento entre usuarios y las interacciones en las que un usuario modifica o comparte los mensajes de otro usuario.

Es de especial interés para nuestro trabajo, apoyarse en este tipo de estándar, para conseguir un formato web nativo e interoperable, reduciendo los esfuerzos de integración posteriores con otros sistemas, por ejemplo, los navegadores web.

Este modelo es aplicable a diversas redes sociales, capturando tanto la influencia de las conexiones sociales como la influencia externa, que puede ser significativa en algunas plataformas. En el caso de X, la influencia externa es aquella que no se desprende explícitamente de un retweet. De acuerdo a un trabajo de los mismos autores Taxidou y Fischer, 2014 se estima que en X existe aproximadamente un 20 % de influencia externa. No hemos encontrado mediciones para otras redes sociales.

PROV-SAID considera actividades y relaciones relacionadas con la difusión de información, como el intercambio de mensajes, la búsqueda de la fuente de difusión y la expresión de los cambios que experimenta el mensaje a medida que se propaga. La influencia de los usuarios juega un papel fundamental en esta

difusión, y aunque el concepto de influencia puede variar según el contexto, se define y amplía en este modelo. Se presenta una visión general del modelo PROV-SAID en la Figura 3.2, donde las extensiones propuestas por los autores al estándar PROV están resaltadas en azul.

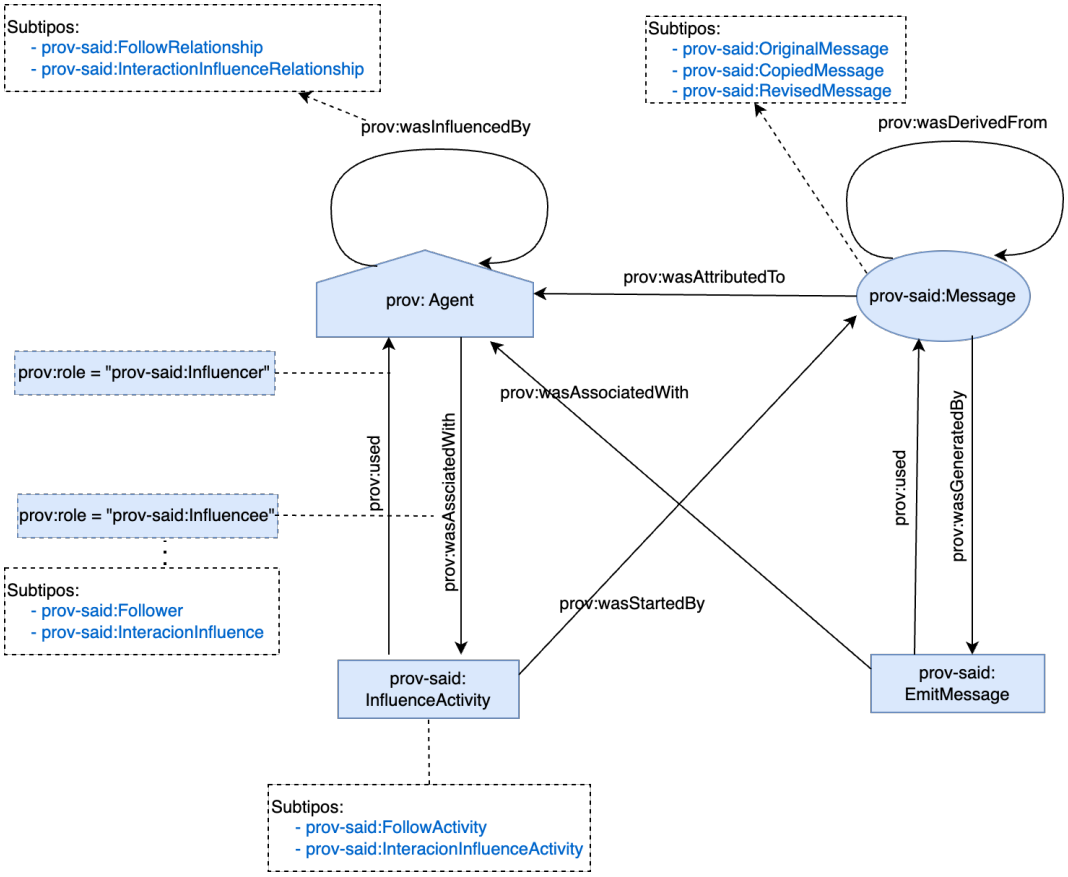

Figura 3.2: Modelo Prov-SAID

A los efectos de diferenciar el concepto de difusión de información y *provenance*, Taxidou et al. 2015 propone el ejemplo de la figura 3.3. En este escenario, tres usuarios de X están compartiendo un mensaje similar: Alice es la fuente de la difusión de la información, ya que ella emite un mensaje original. Más adelante, el usuario Bob modifica el mensaje original y luego el usuario Carol lo copia y reenvía (retuitea). En este proceso, es importante comprender cómo se modificó y transmitió el mensaje. El usuario Carol fue influenciado indirectamente por el usuario Alice, ya que su mensaje se derivó indirectamente de la fuente original. Esto significa que la credibilidad de los tres usuarios involucrados debe ser evaluada, ya que todos participaron en la difusión y modificación de este mensaje.

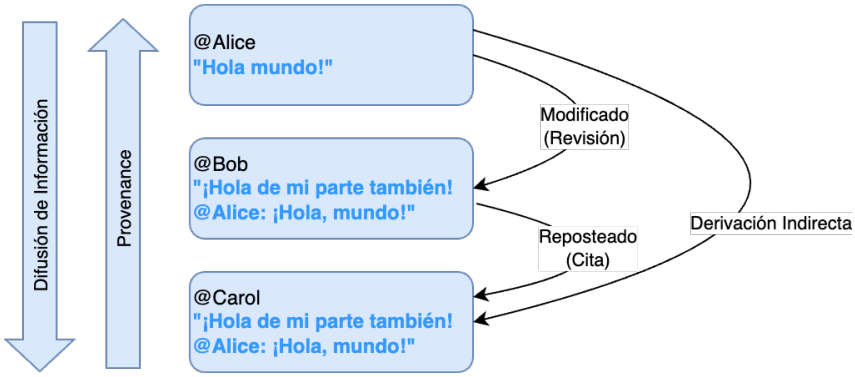

Figura 3.3: Difusión de información y *provenance* según Taxidou et al. 2015

En Taxidou et al. 2015 se modelan dos extensiones fundamentales en PROV, que son claves para su aplicación en datos generados sobre redes sociales. Estos son mensajes e influencia.

Para representar los mensajes emitidos por los usuarios, los autores proponen las siguientes extensiones como subtipos de prov:Entity:

- **prov-said:Message:** se refiere a la categoría general de mensajes. En el contexto de las redes sociales, los mensajes pueden clasificarse como originales, copiados o revisados. Tal como se enumera a continuación, los autores definen un conjunto de categorías como subtipos de prov-said:Message.
    - **prov-said:OriginalMessage:** representa un mensaje inicial que no se deriva de ningún otro mensaje, y el usuario que lo emite es el iniciador de la propagación de información para ese mensaje en particular.
    - **prov-said:CopiedMessage:** se refiere a un mensaje que se basa en otro mensaje previamente publicado y se reenvía como una copia exacta. Los usuarios que emiten mensajes copiados respetan íntegramente el contenido y las opiniones del mensaje original. Por ejemplo, en plataformas como *X*, los usuarios pueden realizar retweets, lo que les permite reenviar fácilmente copias de mensajes emitidos por otros usuarios.
    - **prov-said:RevisedMessage:** representa un mensaje que se modifica a partir de un mensaje existente. Esto implica que el usuario que emite este tipo de mensaje puede o no estar de acuerdo

con la opinión expresada en el mensaje original. Es posible que se realicen alteraciones en la información contenida en el mensaje original.

En cuanto a la influencia, los autores introducen una relación denominada **prov-said:InfluenceRelationship** para representar la influencia general entre agentes en el contexto de las redes sociales. Esta relación es un subtipo de prov:Influence. Además, proponen dos subtipos para especificar las diferentes formas en que se puede manifestar dicha influencia:

- **prov-said:FollowRelationship:** indica que un agente ha sido influenciado por otro agente mediante una relación unidireccional de seguimiento. En el ámbito de las redes sociales, esto implica estar expuesto a los mensajes emitidos por el agente seguido. Por ejemplo, en plataformas como *X*, el seguimiento es la principal forma de conexión entre usuarios, mientras que en *Facebook*, además de las amistades bidireccionales, también se puede establecer una conexión unidireccional mediante la suscripción a los mensajes de otros usuarios. En este contexto, se asume que una vez que un agente comienza a seguir a otro, queda expuesto tanto a los mensajes previos como a los futuros del agente seguido.

- **prov-said:InteractionInfluenceRelationship:** se trata de otro subtipo de prov-said:InfluenceRelationship. Esta relación indica que un agente ha sido influenciado por otro mediante la cita o revisión de los mensajes de este último. Esta relación puede detectarse mediante el análisis de la similitud entre los mensajes del primer y segundo agente.

Resulta también, de especial interés, capturar la forma en que Taxidou et al. 2015 modelan las actividades de influencia. A pesar de que en PROV-DM ya existe un tipo *Influence* es necesario especificar subtipos para ofrecer una mayor expresividad, al proporcionar más información sobre su hora de inicio y fin, qué las ha provocado, entre otros atributos propios de las redes sociales.

- **prov-said:FollowActivity**: denota la actividad de un agente para establecer una conexión unidireccional con otro. Una vez iniciada dicha actividad, el primer agente queda expuesto a las emisiones de mensajes (futuros y pasados) del segundo. Esta actividad tiene una hora de inicio

que denota el momento de establecer la conexión y una hora de finalización opcional en caso de que el agente elimine la conexión con respecto al otro agente.

- **prov-said:InteractionInfluenceActivity**: denota la actividad de un agente para influir en otro, de modo que este último interactúa reenviando los mensajes del primero. Es importante destacar que este tipo de influencia es instantánea y, por tanto, tiene el mismo tiempo de inicio y fin. De este modo, es posible modelar múltiples interacciones de agentes generando múltiples instancias de prov-said:InteractionInfluenceActivity.

Capítulo 4

Modelo de Calidad de Credibilidad

En este capítulo se presenta un caso de estudio que será utilizado a lo largo de este trabajo. Posteriormente, se introduce el modelo de calidad propuesto para la dimensión *credibility*.

4.1. Caso de estudio: Las estatinas

El caso de estudio presentado en esta sección nos permitirá ir introduciendo los nuevos conceptos con ejemplificaciones claras que le permitan al lector incorporarlos de la forma más amena posible.

De acuerdo a diferentes estudios científicos, las estatinas son fuertemente recomendadas por cardiólogos a aquellos pacientes que tienen un alto riesgo de padecer ataques cardíacos, tales como infartos. Las estatinas son una droga fundamental para reducir el colesterol-LDL, el principal factor de riesgo causal para la enfermedad cardiovascular aterosclerótica. Las enfermedades cardiovasculares son la primera causa de muerte a nivel mundial, tanto en hombres como en mujeres. A pesar de que los beneficios de esta droga superan ampliamente los muy estudiados efectos secundarios, sociedades cardiológicas de varios países del mundo han detectado que los pacientes rechazan el tratamiento. En el octogésimo noveno Congreso Europeo de Ateroesclerosis "89th EAS Congress, Helsinki 2021 – EAS", s.f., celebrado en Helsinki en el año 2021, se identificó a la desinformación en redes sociales como un factor que contribuye a la interrupción temprana del tratamiento, a la inercia clínica en

el seguimiento y al impacto en el resultado final. Se hizo especial hincapié en que es responsabilidad de todos los científicos contrarrestar rigurosamente la desinformación médica que es relevante para la salud pública. Se están posicionando a las sociedades médicas con un papel clave en el desarrollo de un sistema que permita la evaluación rápida y la refutación de la información engañosa producida en redes sociales y en la propia literatura médica. De los 56 millones de ciudadanos estadounidenses que podrían beneficiarse de la terapia con estatinas, solo la mitad toma estos medicamentos para reducir sus niveles de colesterol en sangre y prevenir enfermedades cardiovasculares. A pesar de que esta droga es efectiva y segura en la mayoría de las personas, una tasa significativa de pacientes que reciben una prescripción de estatinas abandona el tratamiento. Esta tasa de abandono puede oscilar entre el 40 % y el 60 %.

De acuerdo a un artículo publicado en la Revista Uruguaya de Cardiología en el año 2019 Ramos, s.f., los efectos adversos de las estatinas son de conocimiento popular, pero las redes sociales están magnificando los aspectos negativos sobre los importantes beneficios que superan ampliamente los riesgos. Las dudas de los pacientes sobre determinado tratamiento existen desde mucho antes de la aparición de las redes sociales, pero actualmente este aspecto está acentuado por la facilidad con que se difunden los conocimientos y las noticias (sin importar los fundamentos) y también por las experiencias personales. Sobre este último punto, es importante destacar que los grupos de pacientes en internet son una fuente de constante intercambio de información entre pares. El artículo destaca que tampoco se puede pasar por alto que existe, en gran parte de la población, una desconfianza creciente hacia el médico y la industria farmacéutica. Los tiempos acotados en policlínica suelen impedir el diálogo necesario para despejar las dudas del paciente y que el médico pueda explicar las decisiones tomadas. El paciente con una búsqueda del tipo "efectos secundarios de estatinas" se encuentra con millones de resultados, incluyendo experiencias anecdóticas, poniendo en jaque el tratamiento, y de fondo, la confianza en la relación médico-paciente.

4.2. Modelo de Calidad para Credibilidad

En nuestro trabajo, definimos el cluster *Credibility* como un paraguas que engloba las principales dimensiones relacionadas con los aspectos clave de calidad vinculados a los clips. El objetivo de este cluster es obtener una medida

de calidad genérica que permita adaptarse a los contenidos generados en distintas plataformas de redes sociales. Se busca que el concepto de Credibilidad abarque todas las dimensiones necesarias para evaluar, de manera objetiva, el nivel de confianza que un usuario podría depositar en un clip.

En la Figura 4.1 se presenta una vista gráfica de cómo se compone el cluster *Credibility* y en el Cuadro 4.1 se presenta el modelo de calidad que será utilizado a lo largo de nuestro trabajo. En las siguientes secciones se definen y discuten cada una de las dimensiones, factores y métricas introducidas.

Dimensión de Calidad	**Factor**	**Métrica**
Nombre: Trustworthiness **Descripción:** Percepción de un usuario sobre la medida esperada en que un contenido en redes sociales cumple con una expectativa	**Nombre: Reputation** **Descripción:** Evalúa la percepción general que tienen otros usuarios sobre la fiabilidad del autor o de la fuente del clip	**Nombre: Engagement** **Descripción:** Cantidad de veces que los clips de un usuario son compartidos o mencionados por otros usuarios. Por ejemplo, se puede considerar el número de interacciones sobre los últimos N clips publicados o un conjunto específico de clips en función de un rango temporal. **Granularidad:** Conjunto de cClips **Dominio resultado:** [0..1]
		Nombre: Confidence **Descripción:** Se refiere a la cantidad y calidad de usuarios que siguen a otro. **Granularidad:** Usuario **Dominio resultado:** [0..1]

Dimensión de Calidad	Factor	Métrica
	Nombre: Verifiability **Descripción:** Mide la facilidad con la que se puede corroborar la información contenida en un clip	**Nombre: Access to Source Data** **Descripción:** Indica si el clip contiene enlaces a fuentes originales o documentos que respalden las afirmaciones hechas **Granularidad:** Clip **Dominio resultado:** [0..1]
		Nombre: Modification Transparency **Descripción:** Hace referencia a la presencia en el clip de un Registro de cómo el contenido ha sido alterado desde su publicación inicial **Granularidad:** Clip **Dominio de resultado:** [0..1]
		Nombre: Access to Verification Metadata **Descripción:** Hace referencia a la presencia en el clip de metadata para verificación del contenido **Granularidad:** Clip **Dominio resultado:** [0..1]
	Nombre: Expertise **Descripción:** Representa el nivel de conocimiento y experiencia del autor en el dominio asociado al clip	**Nombre: Certification** **Descripción:** Indica si el clip contiene referencias al grado de conocimiento certificable que tiene el usuario autor **Granularidad:** Clip **Dominio resultado:** [0..1]

Dimensión de Calidad	Factor	Métrica
		Nombre: Production **Descripción:** Indica si el clip contiene referencias a las investigaciones o publicaciones realizadas por el usuario autor **Granularidad:** Clip **Dominio resultado:** [0..1]
		Nombre: Influence **Descripción:** Indica la cantidad de usuarios relacionados con el tema, que lo avalan. **Granularidad:** Clip **Dominio resultado:** [0..1]
		Nombre: Experience **Descripción:** Indica la cantidad de años que el usuario ha dedicado a formarse (nivel académico y laboral) dentro del dominio asociado al contenido del clip **Granularidad:** Clip **Dominio resultado:** [0..1]
Nombre: Provenance **Descripción:** Se enfoca en el origen y los procesos a través de los cuales se ha transformado y derivado un clip	**Nombre: Attribution** **Descripción:** Se refiere al conocimiento de quiénes son las fuentes o entidades autoras de un clip	**Nombre: AttributionMetadata** **Descripción:** Indica si existe en el clip metadata que permita determinar de forma precisa la o las fuentes del clip **Granularidad:** Clip **Dominio resultado:** [0..1]

Dimensión de Calidad	Factor	Métrica
	Nombre: Process **Descripción:** Indica si se conoce cómo se generó el clip, incluyendo los métodos, procedimientos y datos subyacentes	**Nombre: Edit Count** **Descripción:** Mide la cantidad de ediciones en relación a la media de ediciones recibidos por los clips de la plataforma de red social **Granularidad:** Clip **Dominio resultado:** [0..1]
		Nombre: Edit History **Descripción:** Valora cuánto ha cambiado el contenido de un clip en relación con su longitud original, normalizado para que valores cercanos a 1 indiquen cambios sustanciales **Granularidad:** Clip **Dominio resultado:** [0..1]
		Nombre: User Contributions **Descripción:** Mide la diversidad de contribuyentes en relación al total de interacciones. Valores más altos indican una mayor diversidad de opiniones en el proceso de creación de contenido **Granularidad:** Clip **Dominio resultado:** [0..1]

Dimensión de Calidad	Factor	Métrica
	Nombre: **Trustworthiness Path** **Descripción:** Indica cómo la confianza en la información cambia a lo largo de su distribución	**Nombre: Trustworthiness Path Stability** **Descripción:** Mide la estabilidad del trustworthiness a lo largo del provenance. Valores cercanos a 1 indican que la confianza se ha mantenido estable hacia las últimas versiones. **Granularidad:** Clip **Dominio resultado:** [0..1]

Tabla 4.1: Modelo de calidad para el cluster *Credibility*

4.2.1. Descripción de la dimensión *Trustworthiness*

Tal como se presentó en el Capítulo 3 existen diversos trabajos y abordajes sobre la dimensión *trustworthiness*. Desde el punto de vista de la definición, existe consenso, centrando la misma en la percepción de un usuario sobre la medida en que puede esperar que un contenido en redes sociales cumpla con una expectativa, sobre una transacción que implica un riesgo. Desde la perspectiva del *trustor* (quien confía), *trustworthiness* define la cantidad de confianza asociada con el *trustee* (en quien se confía).

En el contexto de la evaluación de la calidad del contenido en plataformas de redes sociales, hemos optado por utilizar la estrategia propuesta por Batini y Scannapieco, 2016, quienes conceptualizan *trustworthiness* como un cluster de dimensiones en lugar de una única dimensión. Esta perspectiva permite una mayor flexibilidad y profundidad en la evaluación de la calidad, al desglosar la confiabilidad en múltiples facetas interrelacionadas.

A pesar de que Batini y Scannapieco, 2016 no definen explícitamente a *trustworthiness* como una dimensión, para fines prácticos en nuestro análisis, adoptamos este concepto como una dimensión global compuesta por tres factores de calidad específicos: *verifiability*, *expertise* y *reputation*. *Verifiability* se refiere a la capacidad de verificar la exactitud del contenido mediante fuentes externas confiables o evidencia empírica, lo que permite a los usuarios confirmar la autenticidad de la información presentada. *Reputation*, evalúa la

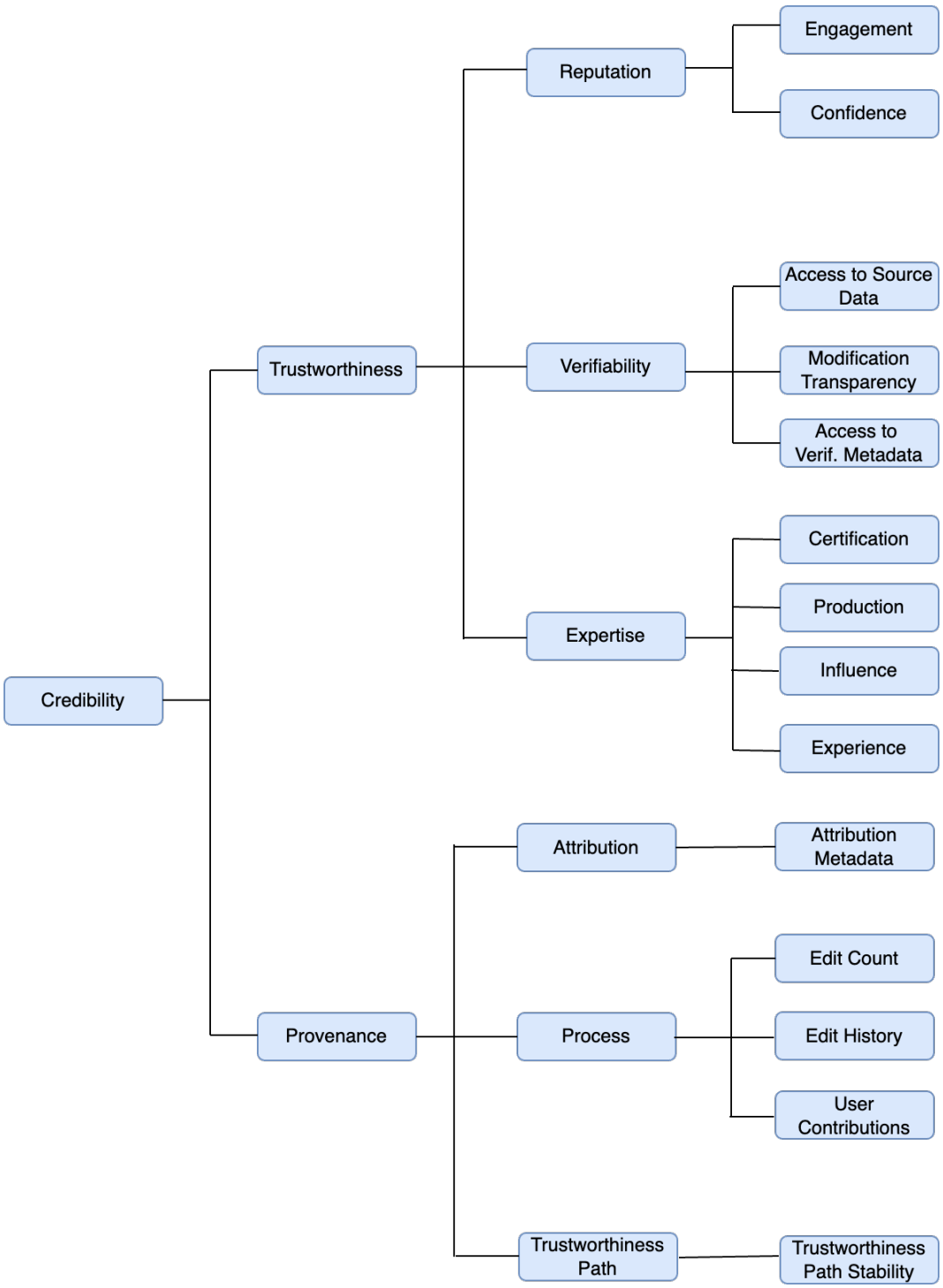

Figura 4.1: Cluster Credibility

percepción pública de la fuente del contenido, incluyendo su fiabilidad histórica y la consistencia en la entrega de información precisa y bien fundamentada. Estas dos características son complementarias y presentan una perspectiva amplia de *trustworthiness*, ya que ofrece indicadores de confianza relevantes para el trustor, describiendo con precisión las características más destacables del *trustee* en el contexto de las redes sociales. Sin embargo, para ampliar aún más la capacidad descriptiva de *trustworthiness*, hemos incorporado el factor *expertise*, con el objetivo de ponderar también la experiencia comprobable del usuario creador de una publicación en un dominio específico.

Factor *Reputation* De acuerdo a las definiciones de Reputation presentadas previamente e incorporando el contexto de las redes sociales, la *reputation* se puede expresar al menos por dos métricas: *Confidence* y *Engagement*.

En términos de las acciones en las redes sociales, *Confidence* hace referencia a la cantidad y calidad de usuarios que siguen a otro.

Engagement permite medir el impacto que los clips de un usuario generan en la red social, siendo éste otro factor relevante que describe parcialmente su *reputation*. Esta métrica se refiere no solo a la frecuencia con que el contenido de un usuario es visto, sino también a cómo y cuánto interactúan otros usuarios con dicho contenido. Medidas de *Engagement* incluyen la cantidad de veces que los clips de un usuario son compartidos, comentados, o reaccionados. Estas acciones indican un nivel de influencia y visibilidad del usuario, aspectos que contribuyen a su percepción de *reputation* dentro de la red social. Una variable relevante para esta métrica es la forma en que se define el conjunto de clips considerados para el cálculo. Este conjunto puede estar compuesto por los últimos N clips publicados, los clips correspondientes a un período de tiempo determinado o aquellos relacionados con un tema específico.

Es importante destacar que *reputation*, aunque significativo, no es un factor que por sí solo caracterice la *credibility* del contenido generado por un usuario. La *reputation* proporciona un contexto útil sobre la fiabilidad de la fuente, pero no es suficiente. Por ejemplo, un usuario con alta *reputation* puede, en ocasiones, compartir contenido que es inexacto o sesgado, lo que podría no ser evidente solo basándose en su nivel de *reputation*, sobre todo fuera de su dominio de *expertise*.

En el Cuadro 4.2 se resumen las métricas definidas para *reputation*.

Métrica	Descripción
Engagement	Se refiere a la cantidad de veces que un clip de un usuario es compartido o mencionado por otros usuarios. Es un indicador de la influencia y la extensión de la presencia de un usuario en la red social.
Confidence	Se refiere a la cantidad y calidad de usuarios que siguen a otro.

Tabla 4.2: Métricas del factor *Reputation*

Factor *Verifiability* De acuerdo a los presentado en el Capítulo 2, en Flemming, 2010 se describe al factor *verifiability*, proporcionando sus ventajas y métricas. En particular, propone que la *verifiability* se mida (i) en base a información de procedencia, (ii) por la presencia de una firma digital, o (iii) por un tercero imparcial.

Aquí la discusión que se debe abordar es cómo estas tres métricas propuestas por Flemming, 2010 en el año 2010, pueden trasladarse a las plataformas de redes sociales, de forma que puedan implementarse métricas adecuadas para los clips.

La herramienta *Community Notes de X* "Community Notes", s.f. ilustra un enfoque práctico para implementar estos principios en el entorno dinámico de las redes sociales. *Community Notes* permite a los usuarios añadir anotaciones verificativas a los clips, proporcionando así un mecanismo directo para mejorar la *verifiability* del contenido mediante la colaboración comunitaria, tal como se puede observar en la Figura 4.2. Según Chuai et al. 2023, *Community Notes* no ha llevado a una reducción significativa del compromiso con tweets engañosos. Esto, debido a que los tiempos de respuesta de las notas de la comunidad, pueden no ser lo suficientemente rápidos como para intervenir eficazmente en las etapas tempranas de la difusión de desinformación. Sin embargo, para el propósito de nuestro trabajo, es una herramienta conceptualmente interesante.

En el Cuadro 4.3 se presentan las métricas de calidad para el factor *verifiability*, adaptados al contexto de las redes sociales. Aunque las investigaciones anteriores no han especificado métricas que se ajusten funcionalmente a las plataformas de redes sociales, hemos identificado tres métricas que no solo se alinean con la definición tradicional de *verifiability*, sino que también pueden medirse utilizando los servicios y datos proporcionados por las propias plataformas de redes sociales.

Métrica	Descripción
Access to Source Data	Indica si el clip contiene enlaces a fuentes originales o documentos que respalden las afirmaciones hechas
Modification Transparency	Hace referencia a la presencia en el clip de un Registro de cómo el contenido ha sido alterado desde su publicación inicial, quién ha hecho los cambios, y por qué razón. Esto es especialmente relevante en plataformas que permiten la edición del contenido después de su publicación.
Access to Verification Metadata	Indica la presencia en el clip de metadata como etiquetas para artículos revisados por *fact-checkers*, alertas sobre contenido disputado, o vinculaciones a verificaciones de hechos externas.

Tabla 4.3: Métricas del factor *Verifiability*

Por otra parte, aunque las redes sociales típicamente no utilizan firmas digitales en el sentido tradicional empleado en los documentos, la verificación de la autenticidad de las cuentas mediante insignias verificadas puede considerarse un paralelismo funcional. Este método certifica que las cuentas de alto perfil o de interés público son genuinas y no imitaciones, proporcionando así una forma de "firma digital" que aumenta la confianza en los contenidos publicados por estas cuentas.

Factor *Expertise* De acuerdo a lo que presentamos en el Capítulo 2, se desprende que las métricas de calidad asociadas al factor *expertise*, parten de la base de que los usuarios, de acuerdo a Kuutila et al. 2024, evalúan la credibilidad basándose en la biografía proporcionada por el autor. Por ejemplo, los autores que se presentan como profesores de medicina o profesionales de enfermería, identificados por su formación académica y rol profesional, son percibidos como más creíbles. En la Tabla 4.4 se definen las métricas de calidad identificadas.

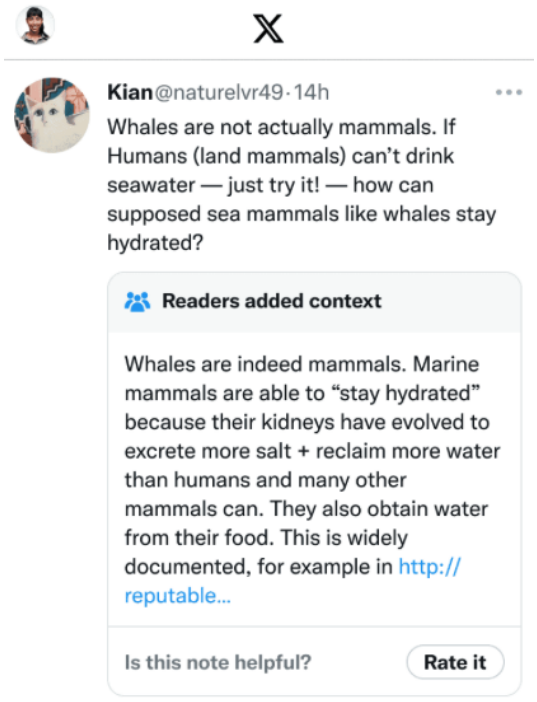

Figura 4.2: Ejemplo de clip de X con *Community Notes*

Factor	Descripción
Certification	Se refiere al grado de conocimiento certificable que tiene un usuario, por ejemplo, título académico, certificaciones realizadas, reconocimientos obtenidos dentro de su dominio.
Production	Indica las investigaciones o publicaciones realizadas por el usuario de un tema específico.
Influence	Indica la cantidad de usuarios relacionados con el tema que lo avalan.
Experience	Se refiere a la cantidad de años que el usuario ha dedicado a formarse (nivel académico y laboral) dentro de su dominio.

Tabla 4.4: Factores de calidad de *Expertise*

Ejemplo de *Trustworthiness* A modo de ejemplo, dentro de nuestro caso de estudio, supongamos que la *American Heart Association (AHA)* publica en su página oficial de *Facebook* un clip sobre los beneficios de las estatinas para prevenir enfermedades cardíacas. El artículo está escrito por un reconocido cardiólogo, lo que aporta un alto nivel de *expertise*.

El clip proporciona enlaces directos a estudios clínicos publicados en revis-

tas médicas prestigiosas. La capacidad de los usuarios para acceder y verificar esta información fortalece significativamente la credibilidad del clip.

Al mismo tiempo, la *reputation* de la AHA como una entidad confiable y respetada en el campo médico es bien conocida, reforzando aún más la confiabilidad.

Este enfoque integrado, que combina un experto altamente calificado, la capacidad de verificación directa a través de revistas de prestigio, y una fuente con una reputación sólida, conduce a un alto *trustworthiness* del clip.

Por otra parte, consideremos que una cuenta de X personal, sin afiliación reconocida, publica un clip sobre los "peligros ocultos" de las estatinas. El autor del artículo es un autoproclamado "experto en salud natural", sin credenciales médicas verificables. El clip contiene afirmaciones sobre los efectos secundarios severos de las estatinas, pero no proporciona enlaces a estudios de investigación que respalden estas afirmaciones. La falta de capacidad para verificar la información, la ausencia de referencias a publicaciones médicas confiables y la falta de elementos que acrediten *expertise*, influyen negativamente en el *trustworthiness* del contenido.

4.2.2. Descripción de la dimensión *Provenance*

De las dimensiones de *provenance* presentadas en el Capítulo 3 debemos seleccionar aquellas que permitan evaluar el clip de forma que aporten a la composición de la *credibility*. Dado que en nuestro enfoque, *provenance* es una dimensión del cluster *credibility*, consideraremos las dimensiones presentadas como factores.

En el Cuadro 4.5 se resumen los factores de la dimensión *provenance*. A los factores *Attribution* y *Process*, ya presentes en la el estado de arte de *Data Quality*, se agrega un tercer factor: *Trusthworthiness Path*. Este incorpora las características de *provenance* desde el punto de vista del *trusthworthiness*. Incorpora una faceta que entendemos fundamental, esto es, la perspectiva de la evolución de la confianza a lo largo del tiempo. Esto es, desde que un clip es generado en la red social, hasta que un usuario final consume la información. Notar que la información será consumida potencialmente en otro clip, producto de múltiples reposteos y comentarios, incluso en una red social diferente a cual fue producido el clip original.

Tal como se puede observar en la figura 3.3, los autores de Taxidou, 2018

marcan la estrecha relación entre el *provenance* y la difusión de información en redes sociales. Tres usuarios de *X* están emitiendo un mensaje similar: Alice es la fuente de difusión de información, ya que emite un mensaje original. Posteriormente, el usuario Bob modifica el mensaje original y luego la usuaria Carol copia y reenvía (retuitea) el mensaje de Bob. En este proceso, es crucial entender cómo se modificó y reenvió el mensaje. Carol fue influenciada indirectamente por Alice, ya que su mensaje se derivó indirectamente de la fuente (procedimiento en dos pasos). Esto implica que el *trusthworthiness* de los tres clips debe ser evaluado, dado que participan en la difusión y modificación de este mensaje.

Factor	**Descripción**
Attribution	Se refiere al conocimiento de quiénes son las fuentes o entidades autoras de un clip
Process	Indica si se conoce cómo se generó el clip, incluyendo los métodos, procedimientos y datos subyacentes
Trusthworthiness Path	Indica la evolución del clip en términos de las medidas de trustworthiness calculado en cada uno de los pasos del *provenance*.

Tabla 4.5: Factores de calidad de *provenance*

A modo de ejemplo, supongamos que la cuenta oficial en *X* de la *Clinic Mayo*, publica un clip detallando los beneficios y riesgos de las estatinas.

El factor de *Trustworthiness Path* destaca cómo la confianza en la información cambia a lo largo de su distribución. Aunque el artículo comienza con un alto grado de *trustworthiness*, debido a su origen en la *Clinic Mayo*, cada alteración a lo largo del camino puede potencialmente disminuir esta métrica, especialmente si los cambios fueron realizados por usuarios con bajo *trustworthiness*.

La Figura 4.3 muestra un ejemplo de cómo se comporta el factor *Trustworthiness Path*. Partiendo de la base de un clip de entrada, el factor debe considerar la mayor cantidad posible de clips anteriores, en términos temporales, hasta llegar al clip original. Para cada uno de estos clips anteriores, se observa cómo la métrica de *Trustworthiness* sufre alteraciones en el transcurso del flujo de información. Se puede observar que el clip original tiene un *trustworthiness* inicial de 0.9, lo que refleja una alta confianza en la fuente original. El clip 2, es una modificación del clip original, efectuado por un usuario menos confiable, con un *trustworthiness* de 0.6. El clip 3 también ha sido modificado

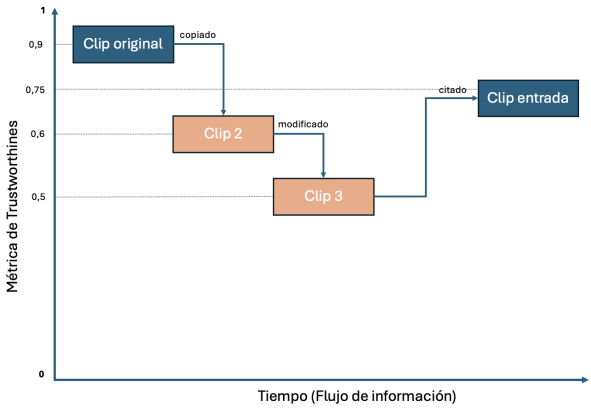

Figura 4.3: Ejemplo de *Trustworthiness Path*

respecto al anterior, lo que provoca una nueva disminución del *trustworthiness* a 0.5. El clip de entrada, ha sido citado desde el clip 3, pero su *trustworthiness* aumenta a 0.75. Esto puede reflejar que, aunque el contenido se ha propagado, la fuente que lo citó es relativamente más confiable, lo que mejora la percepción de confianza en este último clip en comparación con los anteriores.

En términos generales, el *Trustworthiness Path*, busca capturar cómo el *Trustworthiness* en un clip cambia conforme es copiado, modificado o citado. Para este factor, es posible desarrollar diferentes métricas, de acuerdo a lo que se requiera reflejar. En nuestro caso, hemos desarrollado la métrica *Trustworthiness Path Stability*. Lo que se persigue es medir qué tan constante se mantiene el valor de *trustworthiness* a lo largo del flujo de información. Un posible método de medición puede ser la aplicación de la desviación estándar. De esta forma, si los valores de *trustworthiness* están muy dispersos, la desviación estándar será alta, lo que indicaría que la confianza ha fluctuado. Por el contrario, si los valores son similares, la desviación estándar será baja, lo que indicaría que la confianza ha sido estable. En nuestro trabajo optaremos por darle prioridad al *Trustworthiness Path Stabilit*y, sin embargo, dependiendo del dominio es posible incorporar nuevas métricas para el factor *Trustworthiness Path* o incluso nuevos métodos de medición para la métrica *Trustworthiness Path Stability*. En los capítulos 6 y 7 se discuten en profundidad estrategias para la implementación de esta métrica.

4.2.3. Discusión sobre Credibilidad

En esta sección se presentan las bases conceptuales sobre las que se construyó el modelo de calidad para *Credibility*. Es fundamental discutir cuáles son las facetas o componentes que intervienen en la composición de una noción de credibilidad. De acuerdo a lo visto en el capítulo 3, en concreto Fogg y Tseng, 1999, se define *credibility* como la capacidad de ser creíble. Esto es una calidad percibida, que no reside inherentemente en un objeto o información, sino que es atribuida por los usuarios.

También en Fogg y Tseng, 1999 se profundiza en diferentes facetas de la *credibility*, que los autores denominan credibilidad presumida, credibilidad reputada y credibilidad superficial. A continuación se analiza cada una de estas miradas de la *credibility* de cara a los objetivos específicos de nuestro trabajo.

La credibilidad presumida describe el grado de confianza que una persona deposita en algo o alguien basándose en suposiciones generales derivadas de su conocimiento previo del mundo. Por ejemplo, se presume que la mayoría de los médicos dicen la verdad, mientras que también se podría suponer que las personas que practican medicinas alternativas a alto costo no son completamente honestas. Este tipo de *credibility* se fundamenta en supuestos y estereotipos culturales, y refleja un componente exclusivamente subjetivo, influenciado por las creencias y experiencias previas del usuario que consume la información. Como presentamos en el Capítulo 3, en Kuutila et al. 2024, se evidencia que los usuarios tienden a considerar más creíble la información que es consistente con sus creencias, en comparación con la información que no lo es. Este fenómeno se suele denominar efecto de refuerzo de creencias. Esta tendencia lleva a los usuarios a ignorar, hasta cierto punto, la calidad de la evidencia al evaluar la *credibility*. Este enfoque, respaldado por investigaciones científicas en redes sociales, nos permite afirmar que explicitar una medición objetiva del *expertise* y *reputation* de la fuente es muy relevante. Esto permite reducir la brecha, entre la potencial subjetividad del usuario y el nivel de confianza objetivo, que merece la fuente en un determinado dominio de conocimiento.

La credibilidad reputada se refiere al grado de confianza que una persona deposita en algo o alguien basándose en lo que terceros han informado. Por ejemplo, una sociedad cardiológica podría realizar estudios que demuestren que un medicamento es altamente efectivo para controlar los niveles de colesterol, y además, tiene pocos efectos adversos. Este informe de una entidad externa

otorgaría un alto nivel de credibilidad reputada al producto del laboratorio. Esta perspectiva se relaciona con lo que presentamos en el Capítulo 2 como *reputation*, es decir, un juicio realizado por un usuario para determinar la integridad de una fuente, basado en calificaciones o evaluaciones proporcionadas por otros usuarios.

La credibilidad superficial es aquella que se asigna basada en una inspección superficial de una publicación. Por ejemplo, una empresa farmacéutica lanza un nuevo medicamento y decide publicar un clip en la red X para anunciarlo. El clip incluye una imagen de alta calidad del producto, una interfaz visualmente atractiva y un enlace a una página web profesional que describe el medicamento en detalle, conteniendo además testimonios de pacientes. Todos estos elementos, de carácter estético y superficial, inciden en que los usuarios lo perciban como creíble. En este contexto, la *verifiability* funciona como filtro que permite confirmar la percepción inicial de confianza en una credibilidad fundamentada, o, por el contrario, debilitar esa confianza si se demuestra que los indicios superficiales no están respaldados por hechos verificables.

En el Cuadro 4.6 se presenta un resumen del análisis vinculando los tipos de credibilidad con las dimensiones que, para nuestro trabajo, entendemos como fundamentales para ayudar a evaluar la calidad de un clip.

En la Figura 4.4 se puede observar el flujo de interacción entre el productor y el consumidor de un clip, destacando cómo se construye y percibe la credibilidad.

Tipo de credibilidad	**Conceptos Clave**	**Factor de Calidad Asociado**
Reputada	Reputación, Evaluaciones externas, Calificaciones	*Reputation* *Expertise*
Presumida	Estereotipos culturales, Supuestos personales, Experiencias previas	*Reputation*
Superficial	Apariencia visual, Percepción inicial, Popularidad	*Verifiability*

Tabla 4.6: Resumen de Tipos de Credibilidad, Conceptos Clave y Factores de Calidad Asociadas

Por otra parte, es fundamental comprender cuáles son los tipos de errores en que los usuarios suelen incurrir al enfrentarse a contenido sobre el cual de-

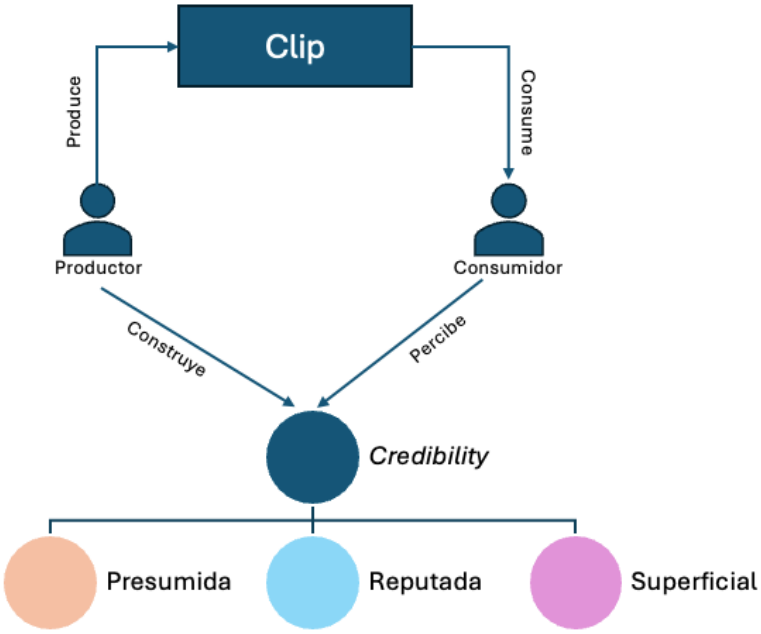

Figura 4.4: Composición de *credibility*

ben determinar su credibilidad. En Fogg y Tseng, 1999 se presentan dos tipos de errores, *Gullibility Error* o Error de Credulidad e *Incredulity Error* o Error de Incredulidad.

El *Gullibility Error* refiere a la situación en la que, a pesar de que una fuente o producto no sea creíble, los usuarios perciben que sí lo es. Este error es particularmente significativo en el contexto de la sobrecarga de información en Internet, donde los usuarios pueden no tener el tiempo o los recursos para verificar adecuadamente la credibilidad.

En el contexto de nuestro caso de estudio, un ejemplo de *Gullibility Error* podría surgir en la decisión de pacientes de rechazar el tratamiento con estatinas. A pesar de la evidencia científica y las recomendaciones de los médicos, algunos pacientes pueden desconfiar y optar por no seguir el tratamiento. Esta desconfianza puede ser alimentada por el consumo de fuentes en línea, que exageran potenciales efectos secundarios o cuestionan la integridad de la investigación farmacéutica. Esto induce al usuario a un *Gullibility Error*. En las Figuras 4.5 y 4.6 se presenta una publicación de la red social *Reddit* y su

Figura 4.5: Ejemplo de publicación en *Reddit*

respectiva respuesta, en un *subredit* específico de Colesterol [1].

En concreto, los usuarios incurren en el *Gullibility Error* cuando no disponen de mecanismos para verificar adecuadamente la fuente, cuando la *reputation* de la fuente es aceptada (sin una evaluación objetiva) o cuando no evalúan adecuadamente si una fuente tiene la *expertise* necesaria en el dominio que está tratando.

Incredulity Error ocurre cuando, aunque la fuente sea creíble, los usuarios perciben que no lo es. Este error puede llevar a los usuarios a rechazar información válida y útil, lo que puede resultar en decisiones mal informadas.

Entendemos que explicitar las métricas de las dimensiones ya presentadas, colabora en reducir este error. El *Incredulity Error* suele ser producto de des-

[1] https://www.reddit.com/r/Cholesterol/

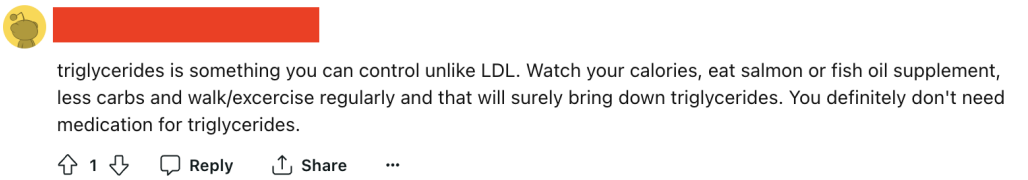

Figura 4.6: Ejemplo de publicación en *Reddit*, en respuesta a la publicación de la figura 4.5

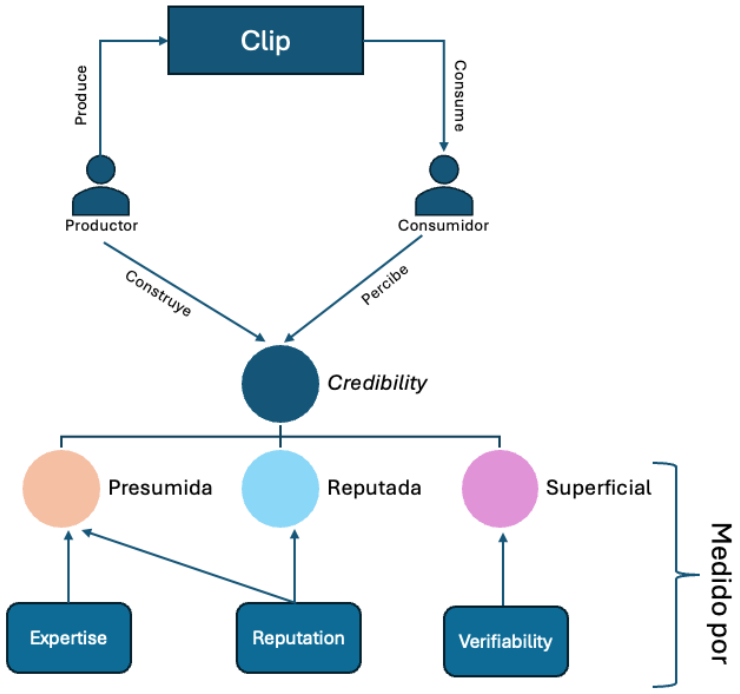

Figura 4.7: Composición de credibilidad

estimar información crítica porque los usuarios no perciben, o no comprenden, el nivel de conocimiento o autoridad de la fuente. En otros términos, a pesar de que una fuente disponga de *expertise* y *reputation*, los usuarios pueden no valorarla adecuadamente.

En la Figura 4.7 se resume la composición de credibilidad en términos de los factores de calidad analizados previamente. Sin embargo, es necesario incorporar la dimensión de *provenance* para obtener una perspectiva más amplia de la credibilidad.

Tal como presentamos en el Capítulo 3, la información de *provenance* sobre

un clip, tiene un aporte significativo a la percepción de credibilidad por parte del usuario consumidor. Por esto, a los aspectos ya expuestos, es fundamental incorporar la visión de otros autores Simmhan et al. 2005. La información de *provenance*, también puede mejorar la confianza. Los autores consideran que el *provenance* de los datos son un aspecto fundamental para evaluar su calidad, origen y los procesos a través de los cuales se han transformado y derivado. Esto impacta directamente en la credibilidad percibida por los usuarios consumidores.

Volviendo a nuestro caso de estudio, consideremos el escenario en que varios usuarios de redes sociales comienzan a compartir clips sobre un nuevo tratamiento para el colesterol. Lo que se plantea en Simmhan et al. 2005 es que la documentación y verificación del *provenance* son fundamentales para la evaluación de la credibilidad de estas publicaciones. En nuestro trabajo, entendemos importante destacar las siguientes etapas:

- **Verificación de la Fuente Original:** Mediante el *provenance* de la publicación, los usuarios pueden identificar si la información proviene de un médico acreditado, un centro de investigación reconocido, o si es simplemente un rumor iniciado por una fuente no verificable. La claridad sobre el origen del clip ayuda a establecer su credibilidad.

- **Evaluación del Contenido Modificado:** A menudo, la información se modifica a medida que se comparte, lo que puede llevar a interpretaciones erróneas o a la alteración del clip original. El registro del *provenance* permite a los usuarios revisar cómo ha cambiado el clip y si las modificaciones han alterado el significado inicial.

- **Resolución de Información Contradictoria:** Si diferentes clips sobre el mismo tema muestran afirmaciones contradictorias, el *provenance* de cada clip puede ayudar a resolver cuál de ellas proviene de una fuente con mayor credibilidad.

De acuerdo a lo analizado hasta el momento, podemos afirmar que la credibilidad es una vista de calidad multifacética, que debe considerar al menos los siguientes aspectos clave: *Verifiability*, *Expertise*, *Reputation* y *Provenance*. A partir de esta afirmación, construimos una propuesta del cluster *Credibility*, utilizando herramientas tradicionales de *Data Quality*.

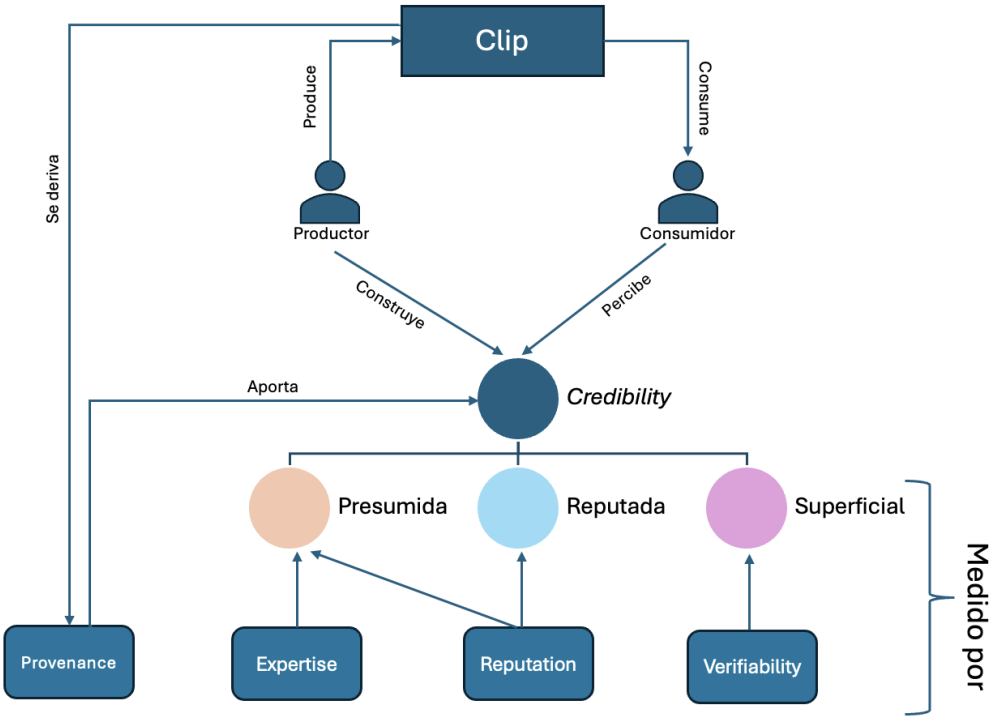

Figura 4.8: Composición de *credibility* y dimensiones de calidad asociadas

La figura 4.8 muestra la composición completa de la credibilidad. Esta estructura sienta las bases para el diseño de nuestro modelo de calidad.

4.3. Cálculo del valor de *Credibility*

Siendo T como *Trustworthiness* y P como *Provenance*, cada una de las cuales toma valores en el rango de [0,1], definimos la fórmula de *Credibility* C como se representa en (4.1), donde t y p son los ponderadores respectivos para cada una de las dimensiones. Los valores que toman estas variables también están en el rango [0,1].

Del mismo modo, T se calcula según la ecuación (4.2) siendo E como *Expertise*, V como *Verifiability* y R como *Reputation*.

$$C = \frac{T * t + P * p}{2} \tag{4.1}$$

$$T = \frac{V * v + R * r + E * e}{3} \tag{4.2}$$

Un problema interesante es determinar los valores de cada uno de los ponderadores, lo cual podemos entenderlo como el conocimiento aportado por el usuario experto del dominio. Este aspecto es decisivo para la obtención de métricas de *credibility* que tengan adherencia con la realidad. La selección puede ser manual, o inferida mediante reglas o técnicas de aprendizaje automático.

La inferencia mediante reglas puede adoptar formas triviales, como asignar el valor 0 a aquellos ponderadores de dimensión que no pueden obtenerse para el clip analizado. Esto puede ocurrir, por ejemplo, cuando el clip no contiene datos necesarios para el cálculo de las métricas.

La inferencia mediante técnicas de aprendizaje automático permite ajustar los ponderadores a partir de la experiencia del usuario experto. Dado un clip C, sea M $(m_1, \ldots, m_n)$ el vector con los valores de las métricas, P_e $(pe_1, \ldots, pe_n)$ el vector con los ponderadores sugeridos por el usuario experto para cada una de las métricas y V_c el valor de la *credibility* calculada para el clip C. Se busca inferir valores de M que se ajusten mejor a V_e, siendo V_e el valor de *credibility* otorgada por un usuario experto que evalúa el clip C.

Lo que se busca con esta estrategia es asistir a los usuarios expertos a determinar el valor de los ponderadores P a partir del análisis de la *credibility* del clip. Este proceso, busca alcanzar el mayor nivel de automatización posible en el proceso del cálculo de V_c, partiendo de una estrategia de ensayo y error en el ajuste de M $(m_1, \ldots, m_n)$.

4.3.1. Estrategia basada en aprendizaje automático

Se propone el siguiente proceso de trabajo.

1. **Recopilación de datos**: Se recopila un conjunto de datos que incluya clips (C), métricas (M) asociadas a cada clip, los ponderadores sugeridos por usuarios expertos (P_e), y los valores de credibilidad otorgados por usuarios expertos (V_e). En el Capítulo 7 se detalla como se aborda esta etapa con un doctor en medicina como experto de dominio.

2. **Selección del modelo de aprendizaje automático**: Se debe seleccionar un modelo de aprendizaje automático apropiado para abordar el problema. Más adelante se sugerirán modelos adecuados para este problema.

3. **Preparación de los datos de entrenamiento**: Los datos de entrenamiento están compuestos por los valores de las métricas (M), los ponderadores (P) y los valores de credibilidad determinadas por el usuario experto (Ve).

4. **Entrenamiento del modelo**: Utilizando el conjunto de datos de entrenamiento, se entrenará el modelo de aprendizaje automático para inferir la relación entre las métricas (M) y los valores de credibilidad (V_e). El modelo deberá de determinar los mejores valores de ajuste entre (M) y (V_e), siendo esto un nuevo conjunto de ponderadores que denominaremos (P_m), o ponderadores inferidos por el modelo M.

5. **Evaluación del modelo**: Una vez que el modelo está entrenado, se evaluará su rendimiento utilizando un conjunto de clips de prueba para verificar su capacidad para predecir los valores de credibilidad (V_c), utilizando los ponderadores inferidos por el modelo (P_m). El error del modelo estará dado por el error cuadrático medio (RME por sus siglas en inglés) entre Vc y Ve para los clips del conjunto de prueba.

6. **Ajuste y optimización del modelo**: Se podrían realizar ajustes en el modelo, como la selección de diferentes algoritmos de aprendizaje automático o la optimización de hiperparámetros.

En la Figura 4.9 se resumen los pasos para el entrenamiento del modelo predictivo.

Para ilustrar el proceso propuesto, consideremos un conjunto de clips (C), con sus métricas asociadas (M), ponderadores sugeridos por usuarios expertos (P_e) y valores de credibilidad otorgados también por usuarios expertos (V_e). Supongamos que se dispone de un clip C_1 con métricas $M_1 = (0.8, 0.6, 0.7)$, ponderadores sugeridos $P_{e1} = (0.3, 0.5, 0.2)$ y un valor de credibilidad otorgado por un usuario experto de $V_{e1} = 0.7$. Utilizando un modelo de aprendizaje automático entrenado previamente, se ajustan los ponderadores (P_m) para el clip C_1 como $P_{m1} = (0.4, 0.4, 0.2)$. Luego, se evalúa el modelo utilizando el valor de *credibility* V_{e1} como referencia, y se obtiene un valor de *credibility* calculado $V_{c1} = 0.68$. Puede observarse en este ejemplo, la capacidad del proceso para inferir ponderadores que se ajusten mejor a los valores de *credibility* otorgados por usuarios expertos. Esto contribuye a la automatización del proceso de evaluación de la *credibility* de los clips.

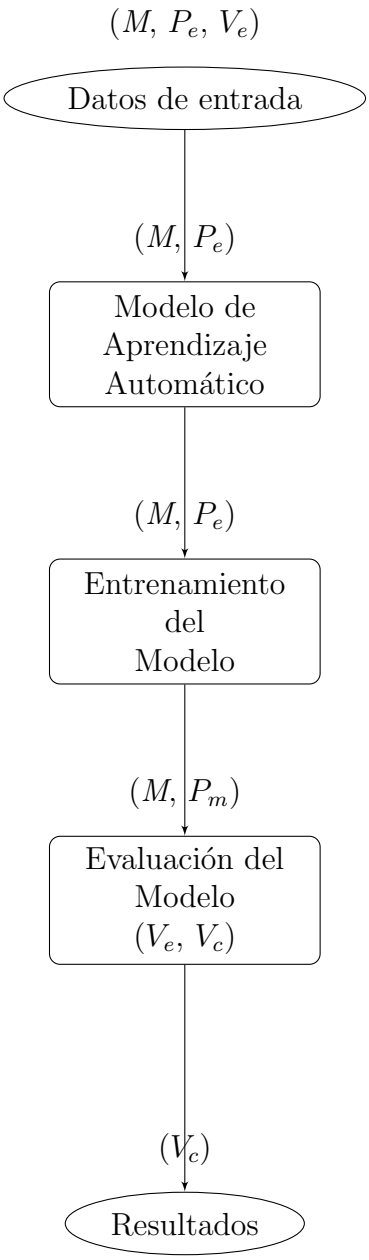

Figura 4.9: Diagrama del proceso de entrenamiento y evaluación del modelo

Se abre una discusión interesante cuando debe de seleccionarse el modelo que mejor se adapte a la necesidades planteadas. Uno de los enfoques más simples y ampliamente utilizados es la regresión lineal, que asume una relación lineal entre las métricas de calidad del clip y su valor de *credibility*. Este modelo presenta la ventaja de su simplicidad y facilidad de interpretación, lo que permite comprender cómo cada característica contribuye al resultado final, aspecto que para este trabajo entendemos como muy relevante. Sin embargo, este enfoque puede ser limitado en su capacidad para capturar relaciones no lineales entre las métricas y el valor de credibilidad.

Una alternativa son los árboles de regresión. Estos modelos dividen el espacio de características en regiones más simples y predicen un valor de *credibility* para cada región. Los árboles de regresión son especialmente útiles cuando las relaciones entre las características y el valor de *credibility* no son lineales. Al mismo tiempo, generan estructuras que son fáciles de interpretar por los usuarios.

Existen muchas otras alternativas en el ámbito de los modelos de aprendizaje automático, como Support Vector Regression (SVR), XGBoost o redes neuronales, entre otras. Sin embargo, consideramos que dichas herramientas exceden los requerimientos específicos de nuestro caso. Además, priorizamos el uso de modelos que sean interpretables para los usuarios.

En este sentido, los árboles de regresión son opciones destacadas, ya que dividen el espacio de características en regiones fácilmente interpretables y proporcionan una visualización intuitiva de las decisiones del modelo. Además, estos modelos permiten identificar las métricas más influyentes en la predicción de la *credibility*, lo que facilita la comprensión y validación por parte de los usuarios expertos.

En el Apéndice B se presenta un ejemplo de inferencia de ponderadores utilizando un árbol de regresión.

Capítulo 5

Flujo de procesamiento

En esta sección presentamos un flujo de procesamiento de clips, diseñado para satisfacer las necesidades de los usuarios finales, esto es, determinar el nivel de *credibility* de un clip específico.

Nuestra decisión de adoptar un modelo basado en un flujo de procesamiento responde a la necesidad de manejar de manera estructurada el cálculo de las métricas de *credibility*. Este enfoque permite descomponer nuestro problema en etapas independientes, donde cada una de ellas se especializa en una tarea concreta, como la normalización de datos, la generación de atributos o la evaluación de métricas de calidad. Nuestro diseño modular habilita que cada componente sea reutilizable y escalable. Esto facilita la incorporación de nuevos módulos o el ajuste de los existentes, permitiendo que sean desarrollados por equipos distintos y en momentos diferentes.

Hemos definido seis fases en el procesamiento de un clip, tal como se presenta en la figura 5.1. Estas etapas incluyen desde la normalización del formato del clip hasta el cálculo de las métricas de calidad relevantes. Todas estas etapas serán presentadas en términos conceptuales, delegando los detalles técnicos de implementación para los diseñadores de solución. En el capítulo 7 presentaremos una implementación de referencia.

En este capítulo nos enfocaremos en el flujo general de procesamiento. En el capítulo 6 presentaremos los detalles específicos del procesamiento del *Provenance*.

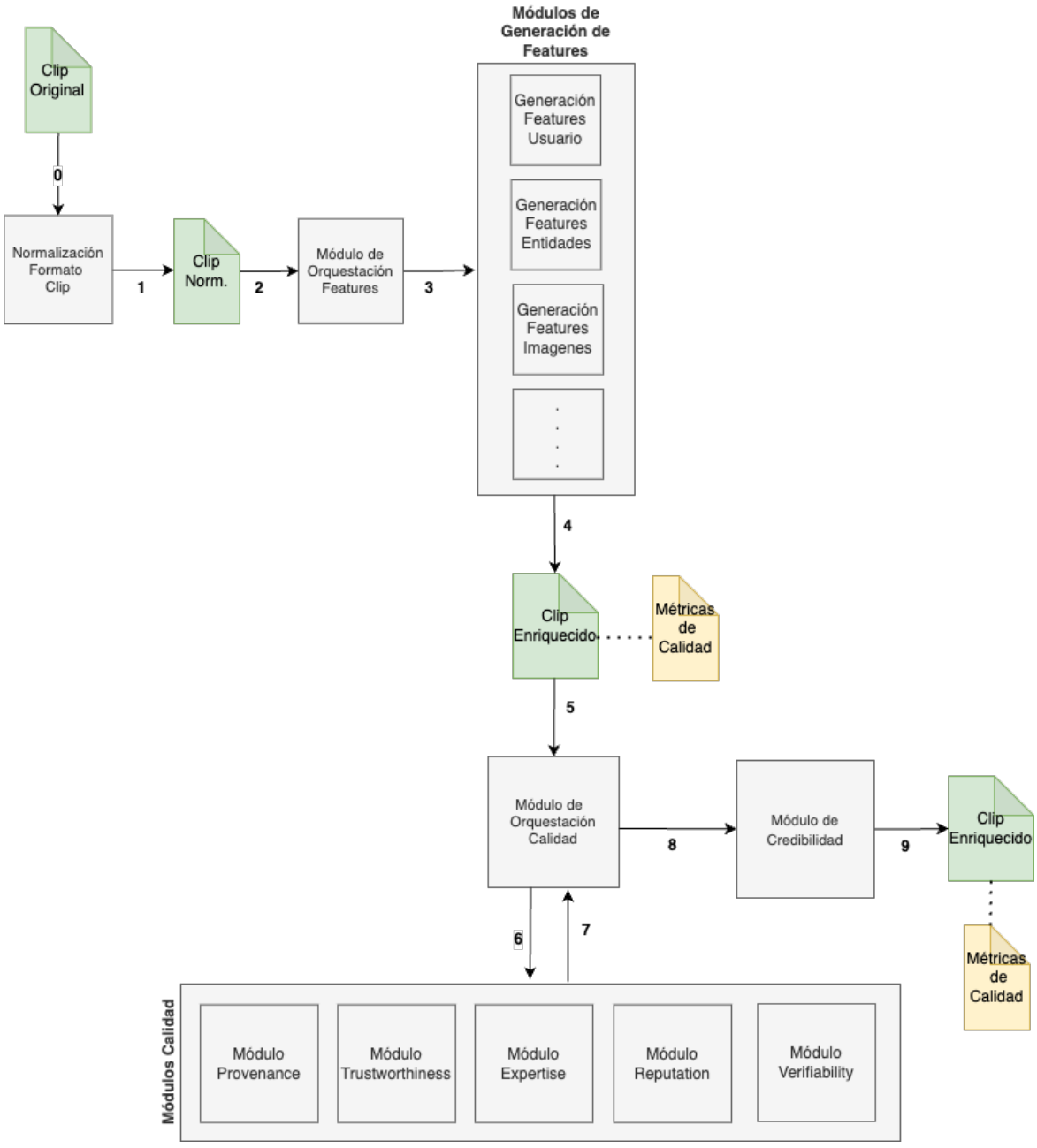

Figura 5.1: Flujo de procesamiento del módulo Credibility

5.1. Usuarios

En esta sección definimos los diferentes tipos de usuarios que interactuarán con la solución propuesta.

5.1.1. Usuario final

El usuario final representa a las personas o entidades que consumen directamente las métricas generadas por el sistema.

Los usuarios finales pueden incluir:

- Consumidores de contenido en redes sociales que requieran verificar la *credibility* de clips antes de interactuar con ellos o compartirlos.
- Profesionales, como periodistas o investigadores, que necesiten validar la veracidad de un clip para incluirlo en sus trabajos.
- Organizaciones que basan decisiones estratégicas en el análisis de tendencias y datos provenientes de redes sociales.

5.1.2. Experto en Calidad de Datos

El experto en calidad de datos tiene como objetivo la configuración y monitoreo de los módulos del flujo. Este usuario es responsable de garantizar que las métricas cumplan con los requisitos establecidos para un determinado caso de uso.

Las principales tareas de este usuario incluyen:

- Configurar los módulos de calidad, a través del respectivo módulo de orquestación, asegurando de que se adapten a los requerimientos específicos del caso de uso.
- Monitorear la ejecución de los procesos de normalización, enriquecimiento y evaluación de *credibility*.
- Definir el modelo de calidad, ajustar y extender las métricas de calidad, incorporando nuevas dimensiones o características según sea necesario.

No se descarta que, en algunos escenarios particulares, sea necesario que el Experto en Calidad de Datos requiera del apoyo de un rol de experto en

sistemas de información. A modo de ejemplo, la implementación de módulos de generación de *features* y de módulos de calidad requieren de conocimientos técnicos que, eventualmente, pueden ir más allá del expertise del Experto en Calidad de Datos.

5.1.3. Experto de Dominio

El experto de dominio aporta conocimiento especializado sobre el contexto específico del contenido que se está evaluando. Este usuario no interactúa directamente con los módulos técnicos, pero colabora con el experto en calidad de datos para asegurar que las métricas de *credibility* reflejen las particularidades del dominio.

Entre las funciones del experto de dominio destacan:

- Asesorar sobre las dimensiones y factores de calidad relevantes para un caso de uso específico. En el contexto de nuestro trabajo, fue necesario hacer partícipes a usuarios expertos en medicina.
- Evaluar si las características y metadatos extraídos del contenido enriquecido son representativos y útiles para su contexto.
- Proveer ejemplos y escenarios reales que permitan entrenar o ajustar los ponderadores de evaluación de *credibility*.
- Proveer medidas y umbrales para métricas particulares.
- Determinar las redes sociales que mejor se adaptan al caso de uso definido.

5.2. Módulo de *Credibility*

La imagen 5.1 muestra el flujo de procesamiento del módulo de *credibility*. En esta sección se describirán cada uno de los pasos y componentes del flujo.

5.2.1. Normalización del Formato del Clip

La normalización del formato del clip tiene como objetivo unificar la estructura de los datos provenientes de diversas plataformas de redes sociales.

Dado que cada red social utiliza formatos y nomenclaturas diferentes para representar la estructura y los datos de un clip, este paso del flujo permite que todos los clips sean procesados bajo un esquema común. Esto asegura que las siguientes etapas del flujo de procesamiento operen de manera consistente, sin depender de las particularidades de cada plataforma.

La estructura de normalización se define utilizando el formato *JSON*. La selección de este formato facilita tanto la integración con las *APIs* de las plataformas de redes sociales como la interoperabilidad entre los diferentes módulos del flujo de procesamiento. Por otra parte, permite extender el modelo para incluir nuevos atributos o adaptarse a requisitos específicos sin alterar la compatibilidad con los módulos existentes en el flujo.

En las tablas 5.1 y 5.2 se presentan los atributos principales del clip normalizado. Estos atributos han sido seleccionados para permitir que todos los datos relevantes de clip puedan ser capturados, independientemente de la plataforma de origen. La construcción del formato común se basa en el análisis de las características específicas de cada red social considerada en este trabajo: *X*, *Instagram*, *Facebook*, *TikTok*, *WhatsApp*, *Reddit* y *YouTube*.

Por tanto, la principal función de este módulo es el mapeo de los atributos del clip original y los atributos del clip normalizado.

Categoría	**Atributo y Descripción**
Identificación	**clip_id**: Identificador único del clip.
	platform: URI de la plataforma de origen del clip (e.g., `https://www.youtube.com/`).
	clip_url: URL del clip.
Contenido	**message**: Contenido textual del mensaje asociado al clip.
	message_type: Tipo de mensaje (e.g., texto, imagen, video).
Fechas	**creation_date**: Fecha y hora de creación del clip.
	update_time: Fecha y hora de la última actualización del clip.
Geolocalización	**has_geo_data**: Indica si el clip contiene datos geográficos.
	coordinates.longitude: Longitud geográfica asociada al clip.

Categoría	Atributo y Descripción
	coordinates.latitude: Latitud geográfica asociada al clip.
Interacción	**hashtags**: Lista de hashtags incluidos en el clip.
	links: Lista de enlaces compartidos en el clip.
	mentions: Lista de usuarios mencionados en el contenido.
	reactions_breakdown: Desglose de reacciones específicas (ej., like, love, angry).
Multimedia	**media_id**: Identificador único del contenido multimedia.
	duration: Duración del contenido multimedia (en segundos).
	audio_track: Información sobre el audio asociado al contenido (ej., nombre de la pista).
Comentarios	**comments**: Lista de comentarios hechos en el clip.
	user_id: Identificador único del usuario que realizó el comentario.
	text: Contenido textual del comentario.
	comment_date: Fecha y hora del comentario.
Métricas de Rend.	**view_count**: Número de visualizaciones del contenido.
	impression_count: Número de impresiones del contenido (veces que apareció en la pantalla).
Metadatos	**extraction_metadata.extracted_at**: Fecha y hora de extracción de los datos.
	extraction_metadata.processing_time_ms: Tiempo que tomó el procesamiento (en milisegundos).
	extraction_metadata.source_details: Información sobre la fuente utilizada para extraer los datos.

Tabla 5.1: Campos principales del clip normalizado.

Categoría	Atributo y Descripción
Identificación	**user_id**: Identificador único del usuario creador del clip.
	user_info.name: Nombre del usuario.
	user_info.full_name: Nombre completo del usuario.

Información Personal

Categoría	Atributo y Descripción
	user_info.is_private: Indica si la cuenta del usuario es privada.
	user_info.location: Ubicación geográfica asociada al usuario.
	user_info.audience_size: Tamaño de la audiencia del usuario (e.g., seguidores, amigos).
Contacto	**user_info.business_contact_method**: Método de contacto comercial del usuario.
	user_info.public_email: Correo electrónico público del usuario.
Biografía	**user_info.biography**: Breve descripción del usuario (biografía).
Expe. y Educación	**user_info.experience**: Lista de experiencias laborales asociadas al usuario.
	user_info.education: Lista de antecedentes educativos del usuario.

Tabla 5.2: Datos del usuario creador del clip.

5.2.2. Modulo de Orquestación de Features

El Modulo de Modulo de Orquestación de Features recibe como entrada un clip en formato normalizado y debe determinar cuáles son los Módulos de Generación de Features que requiere que sean ejecutados. A su vez, debe inferir en que orden hacerlo. Para cumplir con esta tarea, este módulo debe de considerar un conjunto de criterios, los cuales pueden ser clasificados de la siguiente forma:

Características del clip El tipo de contenido del clip (texto, imagen, video o audio) define qué módulos son relevantes. A modo de ejemplo, si el clip contiene solo texto, se ejecutan módulos como análisis de palabras clave o reconocimiento de entidades con nombre. Si además incluye imágenes, se pueden ejecutar módulos de *computer vision*. La plataforma donde se originó el clip también puede tener influencia. Por ejemplo, para un clip de X se pueden activar módulos de análisis de *hashtags* o menciones, mientras que para uno de

YouTube se puede ejecutar módulos de análisis de metadatos de video.

Campos disponibles en el clip El módulo de orquestación revisa los campos presentes en los datos del clip. A modo de ejemplo si el clip no tiene datos de ubicación, no se ejecuta el módulo de análisis de *geolocalización*. Esto asegura que no se intenten ejecutar módulos que dependan de datos no presentes en el clip.

Configuración del flujo El flujo puede ser preconfigurado para ejecutar solo los módulos necesarios en función de los objetivos del análisis. Esta actividad es generalmente realizada por el usuario experto en calidad de datos. Por ejemplo, en un caso de uso enfocado en medir la *credibility* de artículos científicos, se pueden priorizar módulos de extracción de datos de fuentes académicas.

Recursos y prioridades Si el flujo está diseñado para ser eficiente, el módulo de orquestación puede decidir no ejecutar módulos con alto costo computacional si no son críticos. Un ejemplo de ésto, puede ser evitar la ejecución de análisis de imágenes en clips donde el texto es suficiente para inferir la *credibility*.

Factor	**Métrica**	**Módulo de Generación de Features**
Reputation	**Engagement**	**Módulo de Análisis de Engagement**: Evalúa la interacción del clip en redes sociales (likes, menciones, compartidos).
	Confidence	**Módulo de Análisis de Seguidores**: Mide la cantidad y calidad de seguidores del autor del clip.
Verifiability	**Access to Source Data**	**Módulo de Extracción de URLs**: Detecta y valida enlaces a fuentes originales presentes en el clip.

Factor	Métrica	Módulo de Generación de Features
	Modification Transparency	**Módulo de Registro de Modificaciones**: Documenta los cambios realizados al contenido desde su creación.
	Access to Verification Metadata	**Módulo de Identificación de Fact-Checkers**: Encuentra etiquetas o metadatos de verificación asociados al clip.
Expertise	**Certification**	**Módulo de Análisis de Certificaciones**: Evalúa las referencias a certificaciones del usuario autor.
	Production	**Módulo de Análisis de Producción**: Detecta referencias a investigaciones o publicaciones del usuario autor.
	Influence	**Módulo de Análisis de Influencia**: Determina la cantidad de usuarios relacionados que avalan al autor.
	Experience	**Módulo de Análisis de Experiencia**: Calcula los años de formación académica o laboral del usuario en el dominio asociado al caso de estudio.

Tabla 5.3: Ejemplos de módulos de generación de features

5.2.3. Módulos de Generación de Features

Cuando un clip es ingresado por un usuario final para su análisis, debe ser enriquecido con metadatos o características. Estos metadatos son requeridos por las tareas posteriores dentro del flujo de procesamiento, presentado en la Figura 6.13. Estos módulos, pueden actuar sobre diversos componentes del clip, como su contenido (texto, imagen, video o audio), los datos del usuario

que lo generó y los metadatos asociados a la red social en la que fue publicado.

Además de generar características, cada módulo debe proporcionar métricas de calidad que permitan evaluar la confiabilidad de los resultados obtenidos. Por ejemplo, si un módulo encargado de extraer la información académica de un usuario creador de un clip, no logra verificar alguno de los atributos (como la validez de una URL asociada a su universidad), debe de reflejarse en los metadatos de calidad generados. Los módulos de calidad, pueden utilizar esta información para calcular las métricas finales de *credibility* de manera más precisa. En la Tabla 5.3 se describen, a modo de ejemplo, algunos módulos de generación de features. Como puede observarse, cada feature está asociado a una métrica de calidad. El experto del dominio junto con el experto en calidad datos, determinan qué características del clip o inferidas de él, son relevantes para el cálculo de las métricas de calidad. En otros términos, la generación de features está fuertemente acoplada a la instancia del modelo de calidad que el caso de uso requiera.

5.2.4. Módulos de calidad

Cada métrica de calidad es gestionada por un módulo específico. Nuestro diseño incluye implementaciones para cada uno de los módulos de calidad requeridos en el contexto del caso de uso presentado en este trabajo, el cual es abordado en el capítulo 7. Estos módulos de calidad permiten implementar las métricas descritas en el modelo de calidad presentado en el capítulo 4 Sin embargo, el valor de nuestra propuesta de flujo de procesamiento radica en su flexibilidad, ya que permite la incorporación y configuración de nuevos módulos de calidad. Esto garantiza que el cálculo de la métrica de *credibility* pueda evolucionar y adaptarse a las necesidades emergentes o a escenarios cambiantes.

Estos módulos de calidad, podrán ser desarrollados y adaptados por los usuarios expertos en Calidad de Datos.

Cada uno de esos módulos recibirá como entrada el clip enriquecido por parte del Módulo de Orquestación de Calidad y generará como salida la metadata de calidad correspondiente. En la tabla 5.4 se presentan los campos que el Módulo de Orquestación espera como salida.

Campo	Descripción	Tipo de Dato
clip_id	Identificador único del clip al que pertenece la métrica.	string
metric_name	Nombre de la métrica (ej., Engagement, Confidence).	string
metric_value	Valor calculado de la métrica.	float
metric_factor	Factor asociado a la métrica (ej., Reputation, Expertise).	string
calculation_timestamp	Timestamp en el que se calculó la métrica.	timestamp
calculation_method	Método utilizado para calcular la métrica (ej., nombre y versión del módulo de calidad utilizado).	string
quality_metadata	Información de calidad asociada al cálculo (ej., nivel de completitud de los campos requeridos por el módulo).	JSON
remarks	Observaciones adicionales sobre el cálculo o el resultado.	string

Tabla 5.4: Campos de salida de los módulos de calidad.

5.2.5. Orquestador de Módulos de Calidad

El Orquestador de Módulos de Calidad está diseñado para coordinar y gestionar la ejecución de los distintos módulos de calidad. Este módulo actúa como un intermediario entre el clip enriquecido y los módulos individuales de evaluación de calidad, asegurando que cada evaluación se ejecute de manera eficiente y en el orden necesario.

La función principal del orquestador es garantizar que las métricas generadas por los módulos de calidad, tales como las correspondientes a *Trustworthiness*, *Provenance* o *Verifiability*, sean integrados de manera coherente y se adapten a las necesidades del dominio, establecidas por el usuario experto en calidad de datos. A continuación, se describen sus principales responsabilidades:

- **Coordinación del flujo de datos**: El orquestador recibe el clip enriquecido y lo distribuye a los módulos de calidad relevantes, respetando dependencias y asegurando que cada módulo reciba los datos necesarios para operar correctamente.

- **Ejecución paralela o secuencial**: Según las características de los módulos de calidad, el orquestador decide si la evaluación puede realizarse en paralelo o en forma secuencial. Cuando los módulos son independientes podría adoptarse una ejecución en paralelo. Por el contrario, cuando un módulo depende de los resultados de otros, se adoptará una estrategia secuencial.

- **Gestión de errores y excepciones**: En caso de fallos en algún módulo, el orquestador captura la excepción, registra el error y continúa el flujo con los módulos restantes, minimizando interrupciones en el procesamiento.

- **Consolidación de métricas**: Una vez completadas todas las evaluaciones, el orquestador integra los resultados individuales en la métrica de *credibility*.

- **Escalabilidad y modularidad**: El diseño del orquestador permite la incorporación de nuevos módulos de calidad sin alterar el flujo existente, habilitando la adaptabilidad del sistema frente a nuevos requerimientos o escenarios.

5.3. Sugerencias para la representaciones de las métricas

Para permitir que el usuario final pueda procesar eficientemente las métricas de *credibility*, las mismas se le deben presentar en un formato visual y accesible. De acuerdo al estudio del estado de arte, desarrollado en el capítulo 3, en particular lo estudiado en Kuutila et al. 2024, entendemos fundamental fijar algunos lineamientos sobre este aspecto. Dado que el estudio profundo de la presentación gráfica, así como también su desarrollo, están fuera del alcance de nuestro trabajo, las lineas fijadas deben de considerarse solo como recomendaciones para futuras implementaciones.

Las características de la interfaz gráfica deben de considerar el comportamiento de los usuarios finales en redes sociales. Dado que los usuarios están expuestos a grandes volúmenes de clips, dedicando muy poco tiempo al procesamiento de la información presente en los mismos, se deben de buscar formatos de presentación de alto impacto. A los efectos de cumplir con estos objetivos, se delinean las siguientes recomendaciones.

- Gráficos de resumen: Indicadores visuales como barras, líneas de tiempo y diagramas circulares que destacan las métricas clave de credibilidad (por ejemplo, *trustworthiness*, *reputation*, *verifiability*).

- Panel de control interactivo: Permite al usuario explorar la *credibility* de un

Figura 5.2: Ejemplos de iconos para usuarios finales

clip de manera detallada, visualizando las dimensiones específicas que contribuyen al cálculo global, tal como el grafo de *provenance*.

- Alertas y recomendaciones: Señales visuales simples, como colores o íconos, que indiquen rápidamente si el contenido cumple con criterios de confiabilidad mínima o si existen elementos que deben investigarse más a fondo.

Este enfoque gráfico asegura que el usuario final pueda interpretar rápidamente el valor de las métricas y tomar decisiones informadas con un mínimo esfuerzo cognitivo. Como ejemplo, en la imagen 5.2 presentamos un ejemplo de diseño de íconos que representan distintos valores de métricas de *credibility*, de forma de que sean fácilmente entendibles por usuarios finales.

Capítulo 6

Reconstrucción de Provenance

En este capítulo abordaremos los conceptos y mecanismos necesarios para reconstruir el *provenance* de un clip, a partir de lo cual, es posible obtener las métricas de los factores definidos en el Capítulo 4.

6.1. Conceptos previos

Para abordar la problemática a resolver, resulta fundamental describir cómo un clip suele difundirse en las redes sociales. El agente que disemina el contenido es siempre un usuario, humano o máquina. Al mismo tiempo, la mayoría de las plataformas de redes sociales tienen mecanismos que permiten a los usuarios hacer referencia explícita al usuario desde el cual está obteniendo el clip original. En la práctica, esto es retweet en *X*, compartir en *Facebook* o guardar en *Pinterest*. En algunos casos, como *WhatsApp*, la información de la fuente no está incluida, pero informan que el clip ha sido tomado de otro usuario, colocando una etiqueta que indica que el contenido fue "reenviado".

Es muy interesante observar que plataformas emergentes, como *Tiktok* o *Instagram*, no tienen mecanismos para compartir explícitamente un clip dentro de la propia plataforma. Esto suele deberse a estrategias que alientan la creación de contenido nuevo por parte de sus usuarios. Sin embargo, tienen la funcionalidad de compartir su contenido en otras redes, como *Facebook*, *X* o *WhatsApp*. Este mecanismo permite obtener tráfico de estas redes sociales externas. Esta es una característica que debe considerarse en la reconstrucción del *provenance*, ya que hay redes sociales con alta capacidad de generación de contenido que permiten compartir sus clips de forma nativa con otras plataformas. No considerar este aspecto no permitiría una reconstrucción realista del camino de propagación de la información. En el Cuadro 6.1 se muestran los diferentes mecanismos para compartir que ofrecen las plataformas

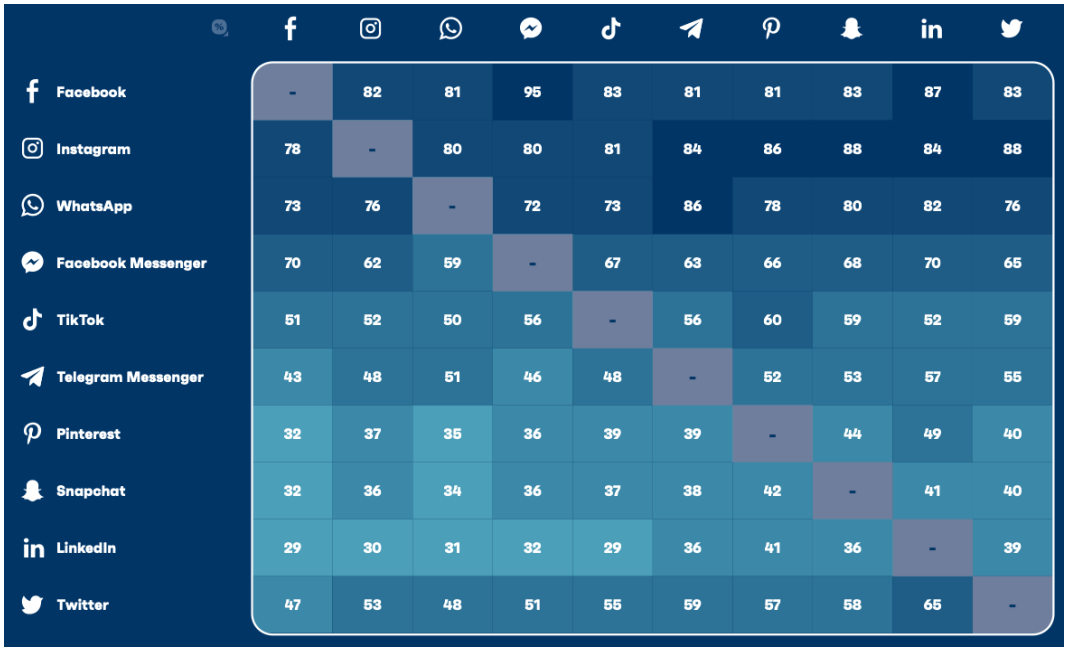

Figura 6.1: Superposiciones en las audiencias de redes sociales. Extraído de GWI, s.f.

de redes sociales.

La acción de compartir externo puede ejecutarse de dos maneras: haciendo uso de las funcionalidades nativas de la plataforma de la red social o manualmente por parte del usuario. En el primer caso, las plataformas suelen utilizar la *URI* (*Uniform Resource Identifier*) del clip como forma de compartirlo en otras plataformas, permitiendo de esta manera generar tráfico desde otras plataformas de redes sociales. Esto permite, más tarde, reconstruir el *provenance* de una manera más sencilla. Cuando la acción es ejecutada manualmente por el usuario, esto es, generando un nuevo clip en la red social destino, se suele no hacer referencia explícita al clip original. Este procedimiento genera que la reconstrucción del *provenance* sea más desafiante. Más adelante, en este mismo capítulo, presentaremos un mecanismo para abordar esta casuística.

En la Figura 6.1 publicada por GWI, s.f., se observa la superposición en las audiencias de redes sociales. En el recuadro se puede leer el porcentaje de usuarios de cada plataforma que también utilizan las siguientes al menos una vez al mes.

Considerando lo comentado anteriormente y los datos estadísticos comentados en las secciones anteriores acerca de la utilización de redes sociales y el enorme crecimiento de ésta, se hace fundamental tener una concepción *multi-red* para el análisis de *provenance*.

Los expertos en redes sociales utilizan el término *cross-sharing* o *cross-posting* para referirse a la actividad de publicar el mismo contenido en múltiples redes sociales. Esta actividad les permite llegar a la mayor cantidad de usuarios, bajo la premisa de que cada red social agrupa usuarios con determinadas características sociodemográficas.

La Figura 6.2 muestra cómo un clip generado en una plataforma de red social *A* puede compartirse en tres plataformas diferentes de manera paralela, mientras que la Figura 6.3 muestra cómo un clip generado en una plataforma de red social *A* puede compartirse en tres plataformas diferentes de manera escalonada.

Describimos a continuación un ejemplo real de *cross-sharing*. *WhatsApp* permite a los usuarios crear y organizar grupos, los cuales están limitados a un máximo de 256 miembros. Por definición, estos grupos son privados, teniendo al menos un administrador que determina quiénes pueden unirse al grupo, enviando invitaciones internamente en *Whatsapp*. Sin embargo, hay otro mecanismo para difundir el grupo, que es a través de un enlace. Este enlace puede ser compartido en otras redes sociales, como *X*, *Facebook* o *Instagram*. De esta manera, el grupo se vuelve público, como se muestra en 6.4. En la otra dirección, los usuarios comparten contenido de otras redes en *WhatsApp*. Según Resende et al. 2019, que estudia contenido compartido en grupos de *WhatsApp* durante la campaña presidencial brasileña, *X*, *Facebook*, *Amino* [1] y *Pinterest* [2] son las plataformas más frecuentes donde las mismas imágenes compartidas se publican en grupos de *WhatsApp*.

Plataforma de Red Social	Compartir Interno	Compartir Externo	URI de Compartir Externo
Facebook	Sí	Sí	Sí
Instagram	No	Sí	Sí
Pinterest	Sí	Sí	Sí
TikTok	No	Sí	Sí
Tumblr	Sí	Sí[3]	Sí
Twitch	Sí	Sí	Sí
X	Sí	Sí	Sí
Whatsapp	Sí	Sí[4]	No
Youtube	No	Sí	Sí

Tabla 6.1: Capacidades de compartir contenido desde las principales plataformas de redes sociales

El *cross-sharing* influye directamente en la capacidad para rastrear, verificar y comprender el *provenance* de un clip. A continuación, presentamos algunos efectos identificados en nuestro trabajo sobre cómo el *cross-sharing* impacta en la construcción del *provenance* de un clip:

Diversificación de las fuentes: El contenido que se comparte entre múltiples plataformas puede originarse de una variedad de fuentes, lo que diversifica y amplía el

[1] www.aminoapps.com
[2] www.pinterest.com

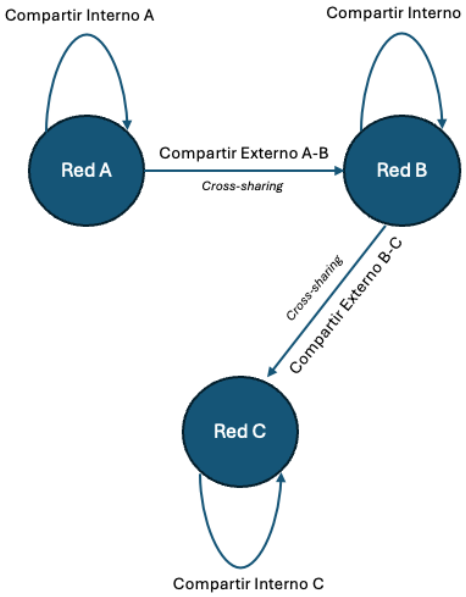

Figura 6.2: Compartir entre plataformas de manera escalonada.

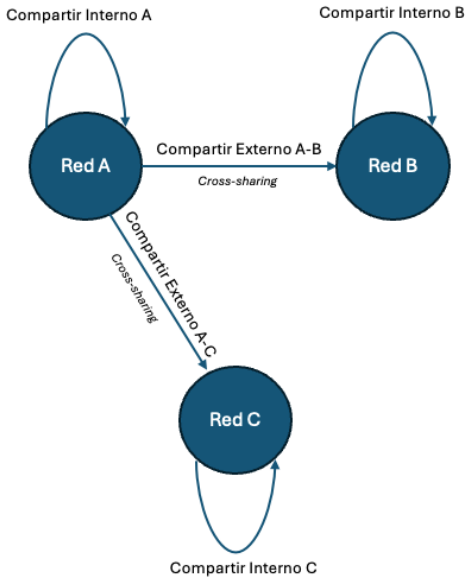

Figura 6.3: Compartir entre plataformas de manera paralela.

alcance del contenido original. Esto puede hacer que sea más difícil rastrear el origen exacto del contenido, especialmente si las plataformas no mantienen un registro claro de la fuente original.

Modificación del contenido: Durante el proceso de *cross-sharing*, el contenido puede ser modificado por los usuarios, lo que altera el mensaje original. Esto puede incluir cambios en el texto, la adición de comentarios o la modificación de imágenes. Estas alteraciones pueden dificultar la verificación de la autenticidad del contenido y la identificación de la fuente original.

Afectación de los metadatos del clip: La reconstrucción del *provenance* se vuelve más compleja debido a que cada plataforma puede tener diferentes métodos para registrar y mostrar el *provenance* del clip. Por ejemplo, algunas plataformas pueden no mostrar explícitamente que un clip fue compartido desde otra plataforma, o pueden perder metadatos cruciales durante el proceso de *cross-sharing*.

Desafíos en la atribución: Identificar al autor original del contenido se vuelve más desafiante cuando el contenido se ha compartido entre múltiples plataformas de redes sociales. Esto puede llevar a confusiones sobre la autoría, especialmente si el contenido se convierte en viral.

Figura 6.4: Enlace entre las plataformas de redes sociales X y WhatsApp

6.1.1. Interacciones Explícitas e Implícitas

Es necesario comprender cómo se manifiesta la influencia a través de interacciones tanto explícitas como implícitas, basándonos para esto en el trabajo previo de Taxidou, 2018. Las interacciones explícitas se refieren a aquellas que se realizan directamente dentro de las plataformas de redes sociales, como los retweets en *X* o los compartir en *Facebook*. Por otro lado, las interacciones implícitas son aquellas que no se registran directamente por las plataformas, por ejemplo cuando los usuarios publican un clip dando crédito al autor original, siguiendo convenciones personales. Otro ejemplo es un contenido de *Facebook* que fue copiado de *X*, donde suele no existir metadatos que permitan determinar a partir de qué clips de *Facebook* fue creado. Estas últimas presentan desafíos significativos en términos de identificación, particularmente cuando no hay indicaciones adicionales del autor original.

Interacciones Explícitas

De acuerdo a Taxidou, 2018 para las interacciones explícitas, se identifican dos casos principales:

- Basadas en enlaces directos, como las respuestas y citas en *X*, donde se proporciona el paso anterior.
- Basadas en la fuente, como los retweets en *X*, donde se proporciona la fuente.

En las plataformas de redes sociales, la funcionalidad de respuesta es una característica muy extendida. La reconstrucción de las respuestas es relativamente simple, ya que el paso anterior está integrado en cualquier respuesta. Esto es, dado un clip es posible recuperar todas las respuestas (clips) que ha recibido. Sin embargo, debido a que las cascadas de respuestas incluyen clips que no son idénticos, a diferencia de los *reposteos*, encontrar mensajes que pertenezcan a la misma cascada de conversación es un desafío.

Por otro lado, la cita se comporta de forma similar al de las respuestas, así como el método para reconstruirlos. Una cita ofrece la posibilidad al usuario de reenviar un clip generado por un tercero incluyendo su propio comentario.

En la Figura 6.5 se representan interacciones explícitas en las que los nodos simbolizan los clips. La información facilitada por las plataformas de redes sociales se destaca en negro. Los datos que necesitan inferirse para trazar completamente los caminos de difusión se muestran en gris. En las interacciones basadas en fuentes, situadas a la izquierda, la conexión con la raíz se denomina “de cualquier paso“, ya que se accede a la raíz a través de uno o más antecesores. Las conexiones directas de un solo paso hacia la raíz deben deducirse. En el caso de las interacciones de

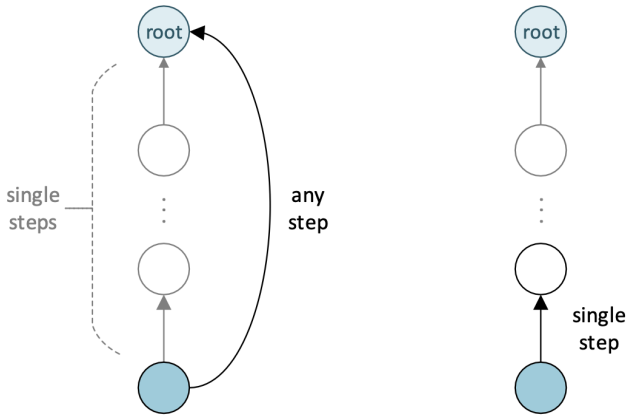

Figura 6.5: Tipos de interacciones explícitas Taxidou, 2018

vinculación directa, a la derecha, una conexión directa de un solo paso conecta con el mensaje anterior, el cual ya está especificado. Sin embargo, es necesario deducir la raíz de esta cadena.

En el análisis de las interacciones explícitas, existen dos dimensiones relevantes a ser consideradas:

- **Número de raíces posibles:** Esto puede ser una única raíz o múltiples raíces. Una única raíz se refiere a un origen único de una cadena de mensajes, mientras que múltiples raíces indican varios orígenes, lo que lleva a caminos de interacción más complejos.
- **Número de aristas de influencia:** Esto se refiere a si hay una sola arista de influencia que emana de un mensaje o varias aristas. Esto influye en cómo se propaga la información y cuán complejas se vuelven las redes de interacción.

Para las interacciones de enlace directo (como respuestas y citas en plataformas como X), el paso anterior en la interacción es conocido, lo que reduce la incertidumbre sobre los caminos de influencia. Sin embargo, la fuente original de la cadena a menudo no se proporciona directamente, lo que requiere un rastreo hacia atrás de los caminos de influencia para identificar la fuente.

En las interacciones basadas en la fuente (como los retweets en X), la fuente original del mensaje se proporciona explícitamente, lo que simplifica la identificación del inicio de la cadena de clips. El desafío aquí radica en inferir el camino completo de influencia que conduce de vuelta a esta fuente, especialmente cuando los pasos intermedios no están documentados explícitamente. De acuerdo a la hipótesis presentada por Taxidou, 2018 los usuarios están expuestos al contenido compartido por sus conexiones sociales. Como resultado, es probable que sean influenciados por

ellas. Esta hipótesis se aplica a las interacciones basadas en fuentes y, en general, a casos donde no se proporciona el influenciador anterior, como en el caso de las interacciones implícitas.

Interacciones Implícitas

En el contexto de las interacciones implícitas, los autores de Taxidou, 2018 afirman que los usuarios están influenciados por fuentes no identificadas, ya sean otros usuarios o fuentes externas, pero no lo manifiestan a través de los mecanismos estándar de las redes sociales, tal como la acción de *citar* en *X*. En nuestro trabajo, una interacción implícita también cubre aquellos casos donde la influencia sucede en diferentes redes sociales (ver Figura 6.3).

Basándonos en el estado del arte Taxidou et al. 2015 Taxidou et al. 2018 Taxidou, 2018, agrupamos las interacciones implícitas en las siguientes categorías.

- **Influencia de usuario con crédito explícito:** Se refiere a las menciones directas de usuarios que reconocen explícitamente la fuente de la información. Por ejemplo, un usuario puede citar o mencionar a otro para darle crédito por la difusión de un clip. Esta influencia no solo incluye aquellos casos donde la mención es efectuada con el identificador de usuario de la misma red social donde reside el clip original, sino que también considera el escenario en el que un usuario menciona un clip publicado en otra red social.

- **Influencia de usuario sin crédito:** Un usuario puede estar influenciado por la información de otros sin reconocer explícitamente la fuente. En estos casos, la influencia se puede inferir si los usuarios tienen conexiones en la propia red social o si el contenido del clip es similar, pero no hay mención directa.

- **Influencia externa:** Ocurre cuando eventos externos afectan a uno o varios usuarios, lo que puede provocar que difundan contenido similar. Estos eventos no se transmiten necesariamente a través de conexiones explícitas en las redes sociales, sino que impactan simultáneamente a varios individuos. Un tipo de influencia externa identificado en nuestro trabajo sucede cuando un clip de una red social influencia a usuarios de otra red, provocando que estos generen clips asociados a la temática original. Un ejemplo de este fenómeno puede observarse cuando eventos relevantes son transmitidos en formato de *streaming* por medio de redes sociales como *YouTube* o *TikTok*. Este fenómeno, además de generar clips en las redes de origen (como comentarios en la propia transmisión), provoca que los usuarios generen clips en otras redes sociales sin hacer ninguna mención a la fuente. El hecho de que los usuarios dispongan

de acceso a múltiples dispositivos en forma simultánea, como *Smart TVs* y teléfonos celulares, potencia aún más este escenario de influencia.

- **Autoinfluencia (promoción):** Los usuarios a menudo vuelven a promocionar su propio contenido para exponerlo nuevamente a su audiencia. Esto puede incluir la eliminación y reescritura de clips en plataformas que no permiten editar el contenido existente.

6.1.2. Definición y cálculo de equivalencia entre clips

Definimos la equivalencia entre clips como una métrica que mide el grado de similitud entre dos clips, considerando tanto su contenido como su contexto. Esta métrica, denotada como $E(c_1, c_2)$, toma un valor entre 0 y 1, donde 0 indica que los clips son completamente diferentes y 1 indica que son idénticos.

En 6.1 se presenta nuestra propuesta de la métrica de equivalencia, donde $S(c_1, c_2)$ representa la cantidad de elementos similares compartidos entre los clips c_1 y c_2, y $M(c_1, c_2)$ es la cantidad máxima de elementos que podrían hacer equivalentes a los dos clips. Estos elementos pueden incluir tanto el contenido visible (texto, imágenes, o videos) como metadatos contextuales (ubicación, tiempo, autoría, entre otros). De esta forma, la métrica permite cuantificar la similitud de manera continua, con valores cercanos a 1 indicando una alta equivalencia.

$$E(c_1, c_2) = \frac{S(c_1, c_2)}{M(c_1, c_2)} \tag{6.1}$$

La complejidad reside en que cada tipo de clip presenta características únicas: los clips de texto pueden ser evaluados en función de su estructura semántica y léxica, mientras que los clips de imagen o video requieren el análisis de características visuales como patrones, colores, objetos y movimientos.

Además, la evaluación de equivalencia debe considerar no solo el contenido explícito del clip, sino también su contexto. Por ejemplo, dos videos pueden compartir escenas similares, pero si han sido editados o reorganizados, su significado o intención puede variar significativamente. Del mismo modo, si el texto que acompaña el video cambia, aunque los videos sean exactamente iguales, la medida de equivalencia de dos clips puede degradarse significativamente.

Por tanto, la equivalencia entre clips no se limita a una mera comparación superficial, sino que implica un análisis multidimensional, que debe abordar aspectos tanto estructurales como contextuales. Este análisis puede incluir técnicas avanzadas de procesamiento de lenguaje natural para clips de texto, y modelos de visión por computadora para clips visuales o audiovisuales, combinados con técnicas de

análisis de metadatos para entender mejor su origen y propósito. A modo de ejemplo, en las Figuras 6.6 y 6.7 se pueden observar dos videos compartidos en fechas diferentes y con texto disímiles. El primero de ellos, publicado en *TikTok* hace referencia a un curso sobre Gestión de Cadáveres donde a los niños se les enseña cómo realizar los últimos ritos para los muertos y rendirles respeto. Durante el curso, los voluntarios toman el lugar de un cadáver y el imán enseña el proceso realizando una demostración sobre los voluntarios. En el segundo, publicado posteriormente en *X*, el texto indica que la "condición de un cadáver ha mejorado a viva". Determinar la equivalencia de dos clips, en estos casos, implica casuísticas complejas.

Figura 6.6: Clip de *X* con video compartido de *TikTok* con alteración en el texto.

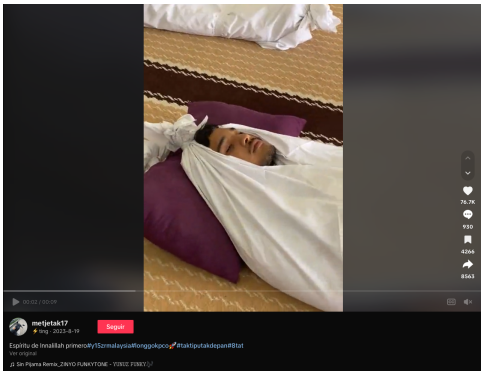

Figura 6.7: Clip original de *TikTok*.

Para no exceder el alcance de nuestro trabajo, hemos decidido restringir la métrica de equivalencia a los clips de texto, considerando que nuestro desarrollo teórico puede aplicarse a métricas de equivalencia más complejas que pueden ser desarrolladas en trabajos futuros.

En De Nies et al. 2016 y Taxidou, 2018 sostienen la hipótesis de que si dos clips son altamente similares, es probable que compartan algún tipo de procedencia. O dicho de otra forma, cuanto mayor sea la similitud entre dos clips, mayor será la probabilidad de que compartan alguna procedencia. En el contexto de nuestro trabajo, mientras mayor sea la equivalencia entre dos clips, es más probable que pertenezcan al mismo *provenance*.

Por lo tanto, para nuestro cálculo de *provenance* es necesario calcular la métrica

de equivalencia entre los clips. De esta forma, es posible agrupar los clips equivalentes por medio de técnicas de clusterización, a los efectos de obtener clusters que contengan clips altamente equivalentes.

6.1.3. Métricas de equivalencia entre clips

Para el caso de clips de texto, utilizaremos la similitud del coseno con ponderación TF-IDF Yun-tao et al. 2005. Consideramos el uso de similitud del coseno como una posible implementación de la métrica 6.1. Esta métrica, permite el cálculo de similitud entre dos documentos de textos, en nuestro caso, clips. Los clips c_1 y c_2 se representan como vectores, generados a partir de un modelo de ponderación TF-IDF aplicado al contenido textual. La fórmula 6.1 se puede instanciar como la expresada en 6.2. Esta expresión define la similitud coseno entre dos clips c_1 y c_2, donde $\cdot$ representa el producto escalar, y $\|c_1\|$ y $\|c_2\|$ son las normas de los vectores.

$$\text{Sim}(c_1, c_2) = \frac{c_1 \cdot c_2}{\|c_1\|\|c_2\|} \tag{6.2}$$

A modo de ejemplo, en la Tabla 6.2 se muestran cuatro ejemplos de *tweets*, haciendo referencia al uso de estatinas. A estos *tweets* no se le agregan otros datos, tales como usuario y links, a los efectos de simplificar el concepto. Luego de aplicar la similitud del coseno, con ponderación TF-IDF, se puede observar en la Tabla 6.3 que los tweets 1 y 3 son los que presentan mayor nivel de similitud, por lo tanto, a pesar de que no existe relación explícita entre ellos, es posible afirmar que ambos pertenecen al mismo *provenance*.

Tweet	**Contenido**
Tweet 1	Increíble que aún haya quien crea que las estatinas curan resfriados #Salud
Tweet 2	¿Sabías que las estatinas pueden mejorar tu memoria? #FakeNews
Tweet 3	Vean esto: Increíble que aún haya quien crea que las estatinas curan resfriados #Salud
Tweet 4	Uso de estatinas como tratamiento para COVID-19 totalmente desmentido #Ciencia

Tabla 6.2: Ejemplos de *tweets* sobre estatinas

	Tweet 1	Tweet 2	Tweet 3	Tweet 4
Tweet 1	1.000000	0.132640	**0.817725**	0.031280
Tweet 2	0.132640	1.000000	0.124747	0.118349
Tweet 3	**0.817725**	0.124747	1.000000	0.029419
Tweet 4	0.031280	0.118349	0.029419	1.000000

Tabla 6.3: Similitud del coseno entre tweets

6.2. Cálculo de Provenance

Para lograr los objetivos marcados, definimos un flujo de procesamiento que aborda los siguientes desafíos: búsqueda de clips equivalentes y reconstrucción del *provenance*. En la figura 6.8 presentamos nuestro flujo de generación de *provenance*. En las siguientes subsecciones presentaremos los desafíos a abordar por cada uno de los componentes.

6.2.1. Búsqueda de Clips Equivalentes

El Módulo de búsqueda de clips equivalentes (*ECSM* por sus siglas en inglés) recibe como entrada un clip normalizado y dispara búsquedas en una o varias plataformas de red social, con el objetivo de detectar clips equivalentes. El criterio de equivalencia es el definido en la expresión 6.2. Una vez obtenidos los clips equivalentes, es posible reconstruir el *provenance*.

La principal función de *ECSM* es implementar estrategias de búsqueda e integración con las API de las redes sociales o los motores de búsqueda. La estrategia de búsqueda podrá variar en función de la plataforma de red social sobre la cual se efectuará la búsqueda y en función de las características del clip de entrada.

El diseño de la estrategia de búsqueda debería cumplir con el objetivo final de la reconstrucción de *provenance*. Este lineamiento de diseño, por tanto, debe buscar implementaciones que recuperen o extraigan clips que sean lo más parecidos al clip de entrada, pero no restringiendo su búsqueda únicamente a patrones de máxima exactitud. Esto se sustenta en la hipótesis de que los clips que forman parte de un mismo *provenance* son similares pero no idénticos, pues el contenido del clip original puede sufrir modificaciones por parte de usuarios intermedios.

La implementación de la estrategia de búsqueda puede ser tan compleja como requiera el nivel de detalle que se espere del *provenance* generado. A modo de ejemplo, pueden buscarse clips de texto casi idénticos al de entrada o podrían buscarse clips que tengan la misma semántica, pero variaciones en el idioma o modificaciones en el parafraseo. En el Cuadro 6.4 proponemos un resumen de las categorías de búsqueda de acuerdo a los requerimientos del *provenance*.

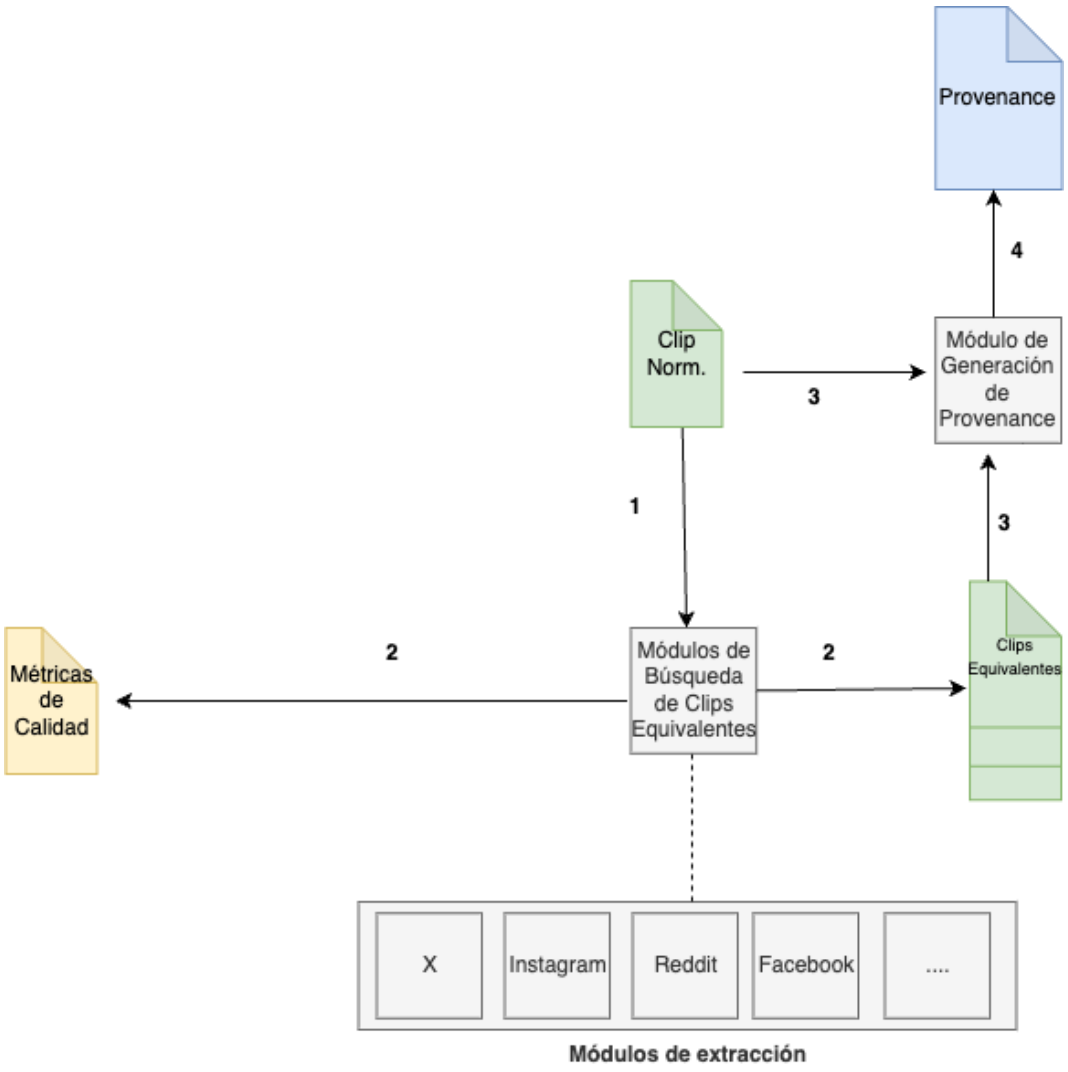

Figura 6.8: Framework para la generación del Provenance

Figura 6.9: Ejemplo Coincidencia Exacta, clip original.

Figura 6.10: Ejemplo Coincidencia Exacta, clip obtenido de la búsqueda.

Figura 6.11: Ejemplo Búsqueda Semántica, clip original

La construcción de parámetros de búsqueda realizada por el *ECSM*, deberá tener en cuenta los parámetros comunes presentados en el Cuadro 6.1. A modo de ejemplo, si el clip original se recibe desde *WhatsApp* y se requiere buscar clips equivalentes candidatos en X, se construye la siguiente consulta de búsqueda:

```
https://api.X.com/1.1/search/tweets.json?
q=url:OriginalClip.Entities["links"].link
```

En esta simple consulta de búsqueda, que utiliza el lenguaje de consulta estándar

Figura 6.12: Ejemplo Búsqueda Semántica, clip obtenido de la búsqueda

de X[1], se devuelven todos los tweets que contienen una *URL* que coincide con la presente en el clip original, en este caso un mensaje de *WhatsApp*.

6.2.2. Módulo de Generación de Provenance

El Módulo de Generación de Provenance (*PGM*, por sus siglas en inglés) toma un conjunto de clips equivalentes con el objetivo de generar una estructura que permita representar el *provenance*. En términos generales, definimos la siguiente información clave: (1) origen del clip original, (2) los clips intermedios, con sus transformaciones y las plataformas de redes sociales en las que han sido procesados. Nuestro enfoque de múltiples redes sociales ofrece más información al usuario final y al mismo tiempo obliga a enriquecer estructuras que comúnmente se utilizan para representar el *provenance*.

Para el modelado del *provenance* hemos tomado como referencia Taxidou et al. 2015, donde se presenta *PROV-SAID*, comentada en la Sección 3. Según Taxidou et al. 2015, los usuarios están representados en el tipo *prov:Agent*. Un mensaje siempre se atribuye a un usuario *prov:Agent* utilizando la relación *prov:wasAttributedTo*.

Para modelar las acciones que un agente puede realizar en un mensaje, se utilizan las acciones de Emisión, Atribución y Derivación.

- **prov-said:EmitClip:** representa la emisión de un clip. El clip generado puede ser del subtipo original, copiado o revisado. Un aspecto relevante de nuestro trabajo es que el clip resultante de la operación puede ser creado en una red social diferente a aquella en la que el usuario ejecutó la operación *EmitClip*.

- **prov-said:Attribution:** representa la acción de asociar un clip con su creador o autor original. En el contexto de nuestro trabajo, la atribución incluye tanto los casos donde el autor es explícitamente reconocido, como aquellos donde la atribución se deriva implícitamente de metadatos contextuales o relaciones previas entre clips.

- **prov-said:Derivation:** captura las relaciones entre clips que se derivan de otros a través de acciones como copiar, revisar o responder. A diferencia de los clips originales, los clips derivados dependen de otros clips para su existencia, lo que significa que pueden rastrearse hasta sus fuentes originales. Nuestro trabajo extiende este concepto al incluir interacciones multi-red, permitiendo modelar derivaciones que cruzan plataformas. Por ejemplo, un clip derivado en *Facebook* puede haberse originado como una respuesta en *X*.

[1]https://developer.X.com/en/docs/X-api/v1/tweets/search/guides/standard-operators

En el listado 6.1 se ilustra cómo los conceptos de emisión, atribución y derivación se reflejan en la creación y transformación de clips entre diferentes redes sociales. La acción de emisión (*prov-said:EmitClip*) se observa en el clip original creado por @Alice en *X* (`alice-status:12345`), identificado como un *OriginalClip*. La atribución (*prov-said:Attribution*) queda representada mediante la relación `prov:wasAttributedTo`, que vincula el clip emitido con su creador, @Alice. Finalmente, la derivación (*prov-said:Derivation*) se evidencia en las transformaciones posteriores del clip: @Bob genera un *RevisedClip* en *Instagram* y @Carol crea un *CopiedClip* en *X*.

Categoría de Estrategia	**Descripción**	**Objetivo para la Reconstrucción de *Provenance***
Coincidencia Exacta	Busca clips que coinciden exactamente con el clip de entrada, sin permitir variaciones.	Identificar puntos de origen o réplicas exactas en la cadena de difusión de un clip. Ver Figuras 6.9 y 6.10.
Búsqueda Aproximada (*Fuzzy Matching*)	Busca clips que son casi idénticos pero que pueden contener pequeños errores tipográficos o variaciones menores en el texto.	Encontrar clips que han sufrido pequeñas modificaciones durante su difusión, pero que aún pertenecen al mismo *provenance*.
Búsqueda Semántica	Busca clips que tienen el mismo significado o intención, aunque el lenguaje o las palabras utilizadas sean diferentes.	Detectar clips cuyos contenidos han sido parafraseados o reformulados, pero que mantienen la misma idea en la cadena de propagación del contenido.
Expansión de Consultas	Genera diferentes variaciones de la consulta original utilizando sinónimos, términos relacionados o conceptos afines, ampliando el alcance de la búsqueda.	Extraer clips que expresan el mismo contenido usando variaciones léxicas o terminológicas, lo cual puede aumentar la granularidad del *provenance*. Ver Figuras 6.11 y 6.12.
Búsqueda Multilingüe	Traduce el clip original a varios idiomas y busca equivalencias en diferentes idiomas, permitiendo detectar publicaciones en idiomas distintos al del clip original.	Permitir la detección de clips en diferentes idiomas, aumentando la granularidad del *provenance*.
Búsqueda por Contexto o *Hashtags*	Busca clips basados en elementos contextuales adicionales, como *hashtags*, *urls*, fechas, ubicación o nombres de usuario mencionados, permitiendo refinar la búsqueda más allá del contenido textual del clip.	Utilizar metadatos para aumentar la búsqueda y aumentar la granularidad del *provenance*.

Tabla 6.4: Estrategias de búsqueda del *ECSM* para la reconstrucción del *provenance* de clips (formato horizontal)

Listing 6.1: Acciones con clips entre diferentes redes sociales

```
prefix alice-status:
   <http://X.com/Alice/status/>
prefix bob-status:
   <http://instagram.com/Bob/>
prefix carol-status:
   <http://X.com/Carol/status/>

// El usuario @Alice tuiteo un mensaje "Hola, mundo!"
prov:entity(alice-status:12345,
    [prov:type='prov-said:OriginalClip',
    prov:label='Hola, mundo!',
    SocialNetwork='www.X.com')],
// El usuario @Bob modifico y re-emitio el
// clip "Hola, mundo!" en Instagram
prov:entity(bob-status:23456,
    [prov:type='prov-said:RevisedClip',
    prov:label='Hola desde mi tambien!']
    \textbf{SocialNetwork='www.instagram.com')},
// El usuario @Carol copio
// el mensaje revisado en X
prov:entity(carol-status:67891,
    [prov:type='prov-said:CopiedClip',
    prov:label='Hola desde mi tambien!
   MT @Alice: Hola, mundo!'],
    SocialNetwork='X.com'))
// alice-status:12345 fue emitido
// por X:Alice
prov:wasAttributedTo(alice-status:12345,
   X:Alice)
```

6.2.3. Cambios introducidos en PROV-SAID

A los efectos que cumplir con los requisitos planteados en las secciones anteriores, es necesario aplicar cambios sobre algunas definiciones de PROV-SAID.

Para trabajar con una terminología acorde con nuestro trabajo, hemos cambiado el término *message* por *clip*. Esto permite utilizar un lenguaje más acorde con las redes sociales. En el Cuadro 6.5 se presentan los cambios que efectuamos a los

nombres de los subtipos de prov:Entity definidos en Taxidou et al. 2015.

Nombre en *PROV-SAID*	**Nuestro para trabajo**
Message	Clip
OriginalMessage	OriginalClip
CopiedMessage	CopiedClip
RevisedMessage	CommentedClip

Tabla 6.5: Mapeo de nombres en las extensiones de PROV-SAID

A continuación se detallan los cambios conceptuales que se introducen a cada entidad.

OriginalClip: Representa clips originales creados en diferentes plataformas. Esto amplía el concepto de *OriginalMessage* en PROV-SAID para permitir múltiples orígenes simultáneos en diferentes contextos.

CopiedClip: Aborda la casuística de compartir contenido entre plataformas, algo no contemplado por PROV-SAID. Este subtipo incluye casos de equivalencia entre clips que son similares pero pueden tener diferencias contextuales, aspecto abordado en nuestra definición de equivalencia de clips.

CommentedClip: Mientras que PROV-SAID considera las revisiones como transformaciones explícitas dentro de una única red social, el concepto de CommentedClip introduce la capacidad de modelar modificaciones tanto explícitas como implícitas. Abarca parafraseos, comentarios y alteraciones en el contenido que pueden cambiar su intención original. Esto es particularmente relevante en el contexto del *cross-sharing*, donde un clip generado en una red social es adaptado y compartido manualmente en otra plataforma.

El tipo *prov:Entity* prevé la inclusión de una lista opcional de atributos que representa información adicional sobre los aspectos fijos de esta entidad. Para nuestro trabajo, es fundamental la presencia de un atributo que especifique el identificador de la plataforma de redes sociales en la que reside el clip. A este atributo lo hemos denominado *SocialNetwork*. Dado que no se ha encontrado en el estado del arte un estándar para referenciar a las plataformas de redes sociales, utilizaremos el *URI* del sitio web de la red social.

Para abordar la evaluación de la credibilidad de los clips, se propone integrar métricas de calidad asociadas a la *trustworthiness* de cada clip. Esta extensión de PROV-SAID es fundamental para permitir el cálculo de las métricas vinculadas al factor *trustworthiness_path* presentado en el capítulo 4. Se profundiza este concepto en la próxima sección.

En la figura 6.13 se describe la creación y atribución de un clip, incorporando los cambios introducidos a PROV-DM a partir de nuestro trabajo. El ejemplo se basa en el presentado en Taxidou et al. 2015.

En el cuadro 6.6 se resumen los cambios aplicados a PROV-SAID.

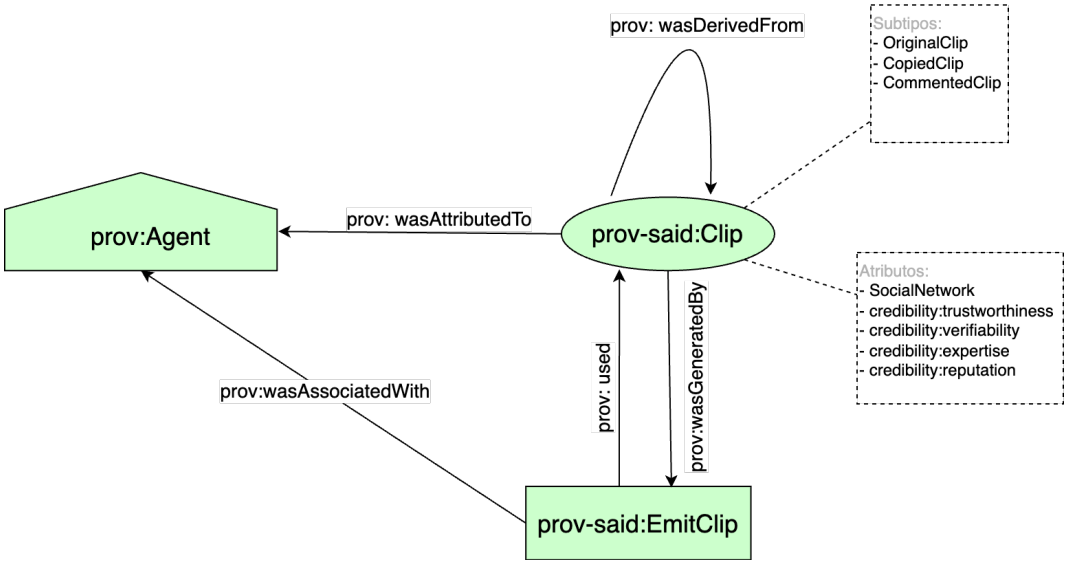

Figura 6.13: Modelo Prov-SAID

Aspecto	**PROV-SAID**	**Nuestro Enfoque**
Entidades clave	Message, OriginalMessage, CopiedMessage, RevisedMessage	Clip, OriginalClip, CopiedClip, CommentedClip
Cobertura de redes sociales	Se centra en una única red social	Adopta una perspectiva multi-red para incluir interacciones entre plataformas
Atributos adicionales	No incluye atributos para identificar las redes sociales	Agrega el atributo *SocialNetwork* para identificar la plataforma de origen
Propagación del contenido	Modela únicamente interacciones dentro de la misma red	Incluye el concepto de *cross-sharing* entre plataformas
Capacidad de modelar métricas de calidad	No considera métricas de calidad	Extiende el modelo para incluir métricas de calidad: *trustworthiness*, *verifiability*, *expertise*, y *reputation*

Tabla 6.6: Comparación entre PROV-SAID y nuestro enfoque

6.2.4. Métricas de calidad

Tal como fue introducido en el capítulo 4, un aporte de nuestro trabajo implica la definición de un factor de calidad derivado del *provenance*: *Trustworthiness Path*. La métrica propuesta es *Trustworthiness Path Stability*. Esta métrica puede

tomar diferentes estrategias, dependiendo del caso de estudio y las características de las redes sociales involucradas. De acuerdo a la definición que se presentó en el capítulo 4, el *Trustworthiness Path Stability* toma valores entre 0 y 1, donde valores cercanos a 1 indican que la confianza se ha mantenido estable. Por tanto, el objetivo central de la métrica es evaluar la consistencia del *trustworthiness* a medida que un clip se propaga, sufre modificaciones y atraviesa diferentes etapas o plataformas en el *provenance*. Esta evaluación permite identificar patrones de confianza en el contenido:

- **Detección de estabilidad:** Si los valores de *trustworthiness* se mantienen cercanos entre versiones consecutivas, la métrica reflejará un nivel alto de estabilidad, lo cual sugiere que el contenido original ha sido preservado en términos de credibilidad.
- **Identificación de degradación:** Fluctuaciones significativas en los valores de *trustworthiness* evidencian inestabilidad en la confianza, lo que puede deberse a modificaciones del contenido, pérdida de calidad de la fuente, o intervenciones de actores con menor *reputation* o *expertise*.
- **Análisis multi-red:** En contextos donde el contenido es compartido entre distintas plataformas por medio de *cross-sharing*, la métrica permite analizar si el *trustworthiness* se ve afectado al migrar de una red social a otra, considerando las características específicas de cada plataforma y sus usuarios.

La métrica debe de ser capaz de detectar degradación del *trustworthiness* y asumir estabilidad en caso de que el *trustworthiness* no varíe significativamente.

Formalmente, la métrica se puede definir de acuerdo a la expresión 6.3. Esta métrica calcula la variación promedio del *trustworthiness* entre nodos consecutivos a lo largo del *provenance*, donde T_i es el valor de *trustworthiness* en el nodo i, n representa el número total de nodos (clips) en el camino del *provenance*, y $|T_{i+1} - T_i|$ corresponde a la diferencia absoluta de *trustworthiness* entre dos clips consecutivos.

El valor resultante se normaliza en el dominio $[0, 1]$, permitiendo la siguiente interpretación: valores cercanos a 1 indican que el *trustworthiness* se ha mantenido estable a través de las versiones. Por otro lado, valores cercanos a 0 reflejan fluctuaciones en el *trustworthiness*, lo que sugiere inestabilidad debido a posibles modificaciones o degradación en la fuente.

$$\text{Trustworthiness\_Path\_Stability} = 1 - \frac{\sum_{i=1}^{n-1} |T_{i+1} - T_i|}{n-1} \tag{6.3}$$

Ejemplo: Consideremos un *provenance* de 4 clips con los siguientes valores de *trustworthiness*:

- $T_1 = 0.9 \quad (OriginalClip)$
- $T_2 = 0.85 \quad (CopiedClip)$
- $T_3 = 0.88 \quad (CommentedClip)$
- $T_4 = 0.86 \quad (CopiedClip)$

Paso 1: Calcular las diferencias absolutas entre nodos consecutivos:

$$|T_2 - T_1| = |0.85 - 0.9| = 0.05$$

$$|T_3 - T_2| = |0.88 - 0.85| = 0.03$$

$$|T_4 - T_3| = |0.86 - 0.88| = 0.02$$

Paso 2: Sumar las diferencias:

$$\text{Suma de diferencias} = 0.05 + 0.03 + 0.02 = 0.10$$

Paso 3: Normalizar para obtener la estabilidad:

$$\text{Trustworthiness\_Path\_Stability} = 1 - \frac{0.10}{4-1} = 1 - \frac{0.10}{3} = 1 - 0.0333 = 0.9667$$

Resultado: La estabilidad del *trustworthiness* es aproximadamente 0.97, lo que indica una alta estabilidad a lo largo del *provenance*.

Ciertos escenarios de aplicación podrían requerir asignar un peso diferenciado a aquellos pasos del *provenance* donde el clip ha sido compartido en otra red social. El experto del dominio puede considerar que estas transiciones influyen significativamente en la credibilidad del contenido.

Adicionalmente, aunque la fórmula presentada para el cálculo del *Trustworthiness_Path_Stability* se basa en la variación promedio del *trustworthiness* entre nodos consecutivos del *provenance*, es relevante considerar métodos alternativos. Por ejemplo, la desviación estándar puede ser utilizada para medir la dispersión de los valores de *trustworthiness* a lo largo de todo el *provenance*, proporcionando una perspectiva global.

En el listado 6.2 se muestra un ejemplo de un nodo del *provenance*, al cual se incorporan los valores de las métricas.

Listing 6.2: Ejemplo de entidad con métricas de trustworthiness

```
prov:entity(alice-status:12345,
```

```
[prov:type='prov-said:OriginalMessage',
prov:label='Hola, mundo!',
SocialNetwork='www.instagram.com'),
credibility:trustworthiness=0.85,
credibility:verifiability=0.92,
credibility:expertise=0.88,
credibility:reputation=0.87])
```

Capítulo 7

Prueba de Concepto

En este capítulo se detalla la implementación de una instancia del modelo de calidad descrito en el Capítulo 4 y del flujo de procesamiento presentado en el Capítulo 5. A partir de estas implementaciones, se diseñó una prueba de concepto en colaboración con un experto de dominio. El objetivo fue evaluar la aplicabilidad del modelo de calidad utilizando un conjunto acotado de clips provenientes de diversas redes sociales.

7.1. Implementación Flujo de Procesamiento

La implementación de la instancia del Flujo de Procesamiento, presentado en el Capítulo 5, se realiza en *Google Cloud Platform* (GCP de aquí en adelante). Esta decisión se basa en las capacidades de GCP para integrar servicios en modalidad *serverless*, esto es, sin la necesidad de gestionar infraestructura. GCP proporciona herramientas como *Pub/Sub* para la transmisión de datos en tiempo real, *Cloud Storage* para el almacenamiento persistente y *Cloud Functions* para la implementación de los módulos de features y de calidad. Además, *Cloud Composer*, basado en *Apache Airflow*, facilita la orquestación de las tareas del flujo de procesamiento (tanto la orquestación de features como la de módulos de calidad), facilitando la modularidad y simplificando el control en la ejecución de cada etapa. En la Figura 7.1 se presenta el diagrama de arquitectura, utilizando los íconos habituales para los componentes de GCP.

En las siguientes secciones se describe, en detalle, la implementación de cada una de las etapas del flujo de procesamiento.

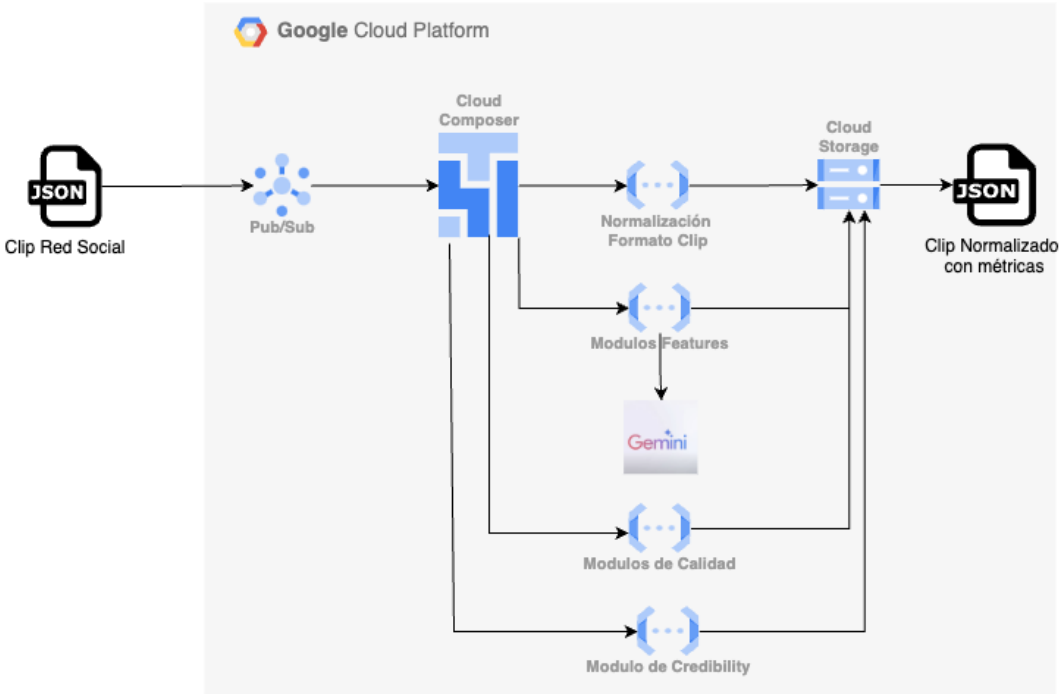

Figura 7.1: Arquitectura del flujo del procesamiento

7.1.1. Ingesta de Clips

La ingesta de clips es la etapa inicial del flujo de procesamiento. Su objetivo es capturar, almacenar y preparar los clips originales provenientes de diversas plataformas de redes sociales. El proceso comienza con la publicación de un clip en *Pub/Sub*. Cada mensaje enviado a un *topic* representa un clip en formato JSON, que incluye tanto el contenido de la publicación como sus metadatos: fecha de creación, identificador del autor, enlaces externos, y otros atributos relevantes para el análisis posterior. La implementación propuesta configura un topic por cada tipo de clip.

Una vez publicados en *Pub/Sub*, los clips son almacenados en *Cloud Storage*, que funciona como repositorio. Esta etapa asegura la disponibilidad y persistencia de los clips, permitiendo que las siguientes fases del flujo puedan procesarlos, así como también recuperarlos en caso de que falle alguna de las etapas.

7.1.2. Normalización

La normalización de los clips tiene como propósito ajustar la estructura de los datos con el objetivo de lograr coherencia y uniformidad de los campos, independientemente de la plataforma de origen. Esta tarea es realizada por una *Cloud Function* que se activa automáticamente al detectar un nuevo archivo en el *bucket* de *Cloud*

Storage. La función procesa cada clip, mapeando los campos específicos de cada red social con el formato de clip normalizado, presentado en el capítulo 5. La estructura del archivo es presentada en el Apéndice D.

El resultado de este procesamiento es almacenado nuevamente en *Cloud Storage* en una ubicación específica que servirá como entrada para las siguientes etapas del flujo. Este proceso de normalización permite estandarizar los clips, facilitando la generación de features en la etapa posterior.

7.1.3. Orquestación de Módulos de Features

Una vez que los clips normalizados están disponibles, se ejecuta el Módulo de Orquestación de Features, implementado mediante *Cloud Composer*. Este componente organiza y coordina el flujo de trabajo necesario para la generación de características asociadas a los clips. La orquestación se define a través de un DAG (*Directed Acyclic Graph*), que permite estructurar las tareas en pasos independientes y ejecutar en paralelo aquellas que no presentan dependencias.

El módulo de generación de features incluye tareas específicas como la extracción de características de usuario, como por ejemplo, la experiencia profesional extraída de la *minibio*. Esta separación modular facilita la incorporación de nuevas tareas o la modificación de las existentes sin afectar el flujo general.

7.1.4. Orquestación de Módulos de Calidad

Una vez generadas las características de los clips en la etapa anterior, se inicia la Orquestación de los Módulos de Calidad. Esta orquestación, al igual que en la generación de *features*, es implementada mediante *Cloud Composer*.

El flujo de trabajo organiza la ejecución de los distintos Módulos de Calidad, asegurando que las tareas se realicen de manera secuencial o en paralelo, dependiendo de las dependencias entre ellas. Cada módulo evalúa una dimensión específica asociada a la calidad del clip: *provenance*, *expertise*, *reputation*, *verifiability* y *trustworthiness*. La orquestación centralizada facilita la integración de los resultados parciales generados por cada módulo, permitiendo que puedan ser analizados y utilizados en etapas posteriores. En particular, el módulo encargado del cálculo de la métrica de *trustworthiness* utiliza como parámetros las métricas de *expertise*, *reputation* y *verifiability*. Estos valores son obtenidos directamente del clip normalizado.

7.1.5. Módulos de Calidad

Los Módulos de Calidad, implementados por medio de *cloud functions* tienen como objetivo calcular las métricas asociadas a los factores y dimensiones de calidad. Cada módulo se especializa en una métrica específica, aplicando los métodos de cálculo definidos en función del factor evaluado. Si bien un módulo puede implementar múltiples métodos para calcular una misma métrica, en nuestra experimentación se desarrolla una única métrica por factor.

Cada módulo recibe como entrada los clips normalizados, generados en la etapa anterior, y aplica las transformaciones necesarias para calcular la métrica correspondiente. Una vez obtenidos los resultados parciales de cada módulo son incorporados al clip normalizado. Esto permite que el procesamiento posterior tenga acceso a las métricas calculadas.

7.1.6. Módulo de Credibilidad

El Módulo de Credibilidad, implementado como una *Cloud Function*, integra los resultados generados por los Módulos de Calidad con el objetivo de calcular una métrica única de credibilidad para cada clip. La credibilidad se determina a partir de la combinación ponderada de las dimensiones de calidad calculadas previamente, tal como se presentó en el Capítulo 4. En nuestra implementación se utilizan ponderadores fijos para cada métrica. Desarrollos posteriores, pueden utilizar *Vertex* en caso de que se requiera resolver la inferencia de ponderadores por medio de técnicas de *Machine Learning*.

7.1.7. Implementación del módulo de reconstrucción de provenance

El módulo de *provenance* fue implementado mediante una *Cloud Function* que recibe como entrada el clip normalizado. Se desarrollaron dos técnicas de búsqueda de clips equivalentes: una utiliza la API de la red social, y la otra emplea el servicio de búsqueda de Google.

Para identificar las interacciones explícitas dentro de la red *X*, se utilizó la API oficial en su versión gratuita[1] para obtener los identificadores de los usuarios que compartieron el clip. En el caso de interacciones explícitas en otras redes sociales, así como de interacciones implícitas, se utilizó el servicio de búsqueda *Google Custom Search API*. Este servicio se invoca con el texto del mensaje incluido en el clip,

[1] https://developer.x.com/en/docs/x-api/getting-started/about-x-api

eliminando caracteres especiales y emoticones. La búsqueda se restringe a los sitios web de las siguientes redes sociales: *Facebook*, *Instagram*, *X*, *Tiktok* y *Reddit*.

El JSON retornado por la API se procesa para determinar qué resúmenes o *snippets* son equivalentes al texto del mensaje del clip original, utilizando `TfidfVectorizer` de la librería *scikit-learn*[1], con una métrica de similitud de coseno. Los resultados con un valor superior a 0.95 se consideran equivalentes.

Los resultados de búsqueda considerados equivalentes se descargan con la librería *Requests*[2] y se procesan mediante *BeautifulSoup*[3], extrayendo los campos necesarios para generar un clip normalizado.

Finalmente, una vez obtenidos los clips normalizados, se genera un archivo JSON siguiendo el esquema *PROV-SAID* ampliado, presentado en el Apéndice E.

7.2. Implementación de los módulos para el caso de estudio

En esta sección se describe la implementación de los módulos específicos para el caso de estudio de estatinas. El revelamiento de las necesidades funcionales se hizo en conjunto con el usuario experto de dominio, de forma de poder determinar cuáles son las necesidades específicas para el cálculo de las métricas. Utilizando como guía el modelo de calidad presentado en el Capítulo 4, se determinaron los factores y métricas más relevantes, para luego determinar los features necesarios.

Se deciden implementar los factores y métricas descritos en el Cuadro 7.1.

Dimensión de Calidad	Factor	Métrica
Trustworthiness	**Reputation**	Engagement
	Verifiability	Access to Source Data
	Expertise	Certification
Provenance	**Trustworthiness Path**	Trustworthiness Path Stability

Tabla 7.1: Factores y métricas para la experimentación

[1]https://scikit-learn.org/stable/
[2]https://pypi.org/project/requests/
[3]https://pypi.org/project/beautifulsoup4/

Listing 7.1: Ejemplo de salida JSON para el cálculo del feature

```
{
    "module_name": "MedicalProfessionModule",
    "feature_name": "MedicalProfession",
    "feature_value": "Nefrologo",
    "confidence_level": 0.95,
    "feature_details": {
        "description": "Profesion basada en perfil de
           usuario"
    },
    "calculation_details": {
        "calculated_at": "2024-12-22T10:00:00",
        "calculation_method": "Gemini Language Model (
           gemini-1.5-flash)",
        "processing_time_ms": 1500
    }
}
```

7.2.1. Implementación de los módulos de features

En esta sección se detallan los módulos diseñados para la generación de *features* relacionados con perfiles médicos. Cada módulo tiene un propósito específico, como evaluar la profesión médica, la educación, la verificabilidad y la influencia médica. Estas *features* son necesarios para calcular métricas de calidad.

Estructura General de los Módulos

Los módulos implementados comparten una estructura común que incluye componentes como la carga de perfiles en formato JSON, la interacción con el modelo generativo *Gemini* de *Google*, y el procesamiento de las respuestas generadas. Este proceso culmina con la validación de los resultados y su almacenamiento en el clip normalizado.

El flujo de trabajo consiste en tomar como entrada un perfil médico en formato JSON, extraído desde el clip normalizado, procesarlo mediante un modelo generativo con *prompts* específicos y generar como salida el clip normalizado actualizado que incluye las *features* generadas.

El formato de salida es consistente para todos los módulos, asegurando una estructura estándar que facilita su integración y análisis. Un ejemplo de la estructura del JSON se muestra en el Listado 7.1.

Módulos Implementados

En los siguientes párrafos se describen los módulos de *features* implementados. La salida generada por cada módulo es almacenada en el campo `features.generated_features` del clip normalizado. La estructura del salida del módulo se presenta en el Cuadro 7.2.

Campo	Descripción
`module_name`	Nombre del módulo que generó el feature.
`feature_name`	Nombre del feature generado.
`feature_value`	Valor generado del feature.
`confidence_level`	Nivel de confianza asociado al *feature* (valor en $[0, 1]$).
`feature_details.description`	Descripción adicional sobre cómo se generó el *feature*.
`calculation_details.calculated_at`	Marca temporal indicando cuándo fue generado el *feature*.
`calculation_details.calculation_method`	Método o modelo utilizado para generar el *feature*.
`calculation_details.processing_time_ms`	Tiempo de procesamiento en milisegundos para generar el feature.

Tabla 7.2: Estructura de la salida de un módulo de feature

Profesión Médica Este módulo tiene como objetivo determinar la profesión médica asociada con un perfil y asignar un índice de confianza basado en la información disponible. El *prompt* utilizado guía al modelo para analizar si la profesión médica es válida, considerando la relevancia del contenido relacionado con medicina.

La respuesta esperada del modelo sigue un formato estructurado con dos líneas: la primera identifica la profesión, y la segunda proporciona un índice numérico entre 0 y 1 que representa la calidad de la profesión detectada. Un ejemplo de salida es:

```
Profesion: Cardiología
profession_quality: 0.9
```

Educación Médica Este módulo evalúa el nivel educativo del usuario relacionado con la medicina. El *prompt* identifica cualquier referencia a títulos médicos o educativos relevantes. Un ejemplo de salida es el siguiente:

```
Education: Médico Cirujano
education_quality: 1.0
```

Verificabilidad Médica Este módulo evalúa la verificabilidad de las afirmaciones médicas de un usuario. El *prompt* instruye al modelo a buscar referencias a instituciones reconocidas, como universidades o hospitales, y asignar un índice de confianza basado en la calidad de esta evidencia. Un ejemplo de salida es:

```
Verifiability: Alta (Universidad Nacional de Medicina)
verifiability_quality: 0.85
```

Influencia Médica El objetivo de este módulo es calcular un índice de influencia médica basado en las recomendaciones y comentarios textuales asociados al perfil. El modelo analiza el contenido y genera una respuesta estructurada con una breve descripción y un índice de confianza. Un ejemplo de salida es:

```
Influence: Excelente reputación, basada en comentarios positivos.
influence_quality: 0.95
```

7.2.2. Implementación de los módulos de calidad

En los próximos párrafos se describen los módulos de calidad implementados, de acuerdo a las métricas presentadas en el Cuadro 7.1. La salida generada por cada módulo es almacenada en el campo `quality_metrics` del clip normalizado, cuya estructura se presenta en el Cuadro 7.3.

Campo	Descripción
`clip_id`	Identificador único del clip al que pertenece la métrica.
`metric_name`	Nombre de la métrica evaluada.
`metric_value`	Valor calculado de la métrica, en el rango $[0, 1]$.
`metric_factor`	Factores o fórmula utilizada para calcular la métrica.
`calculation_timestamp`	Marca temporal indicando cuándo se realizó el cálculo de la métrica.
`calculation_method`	Método o fórmula exacta empleada para calcular la métrica.
`quality_metadata`	Metadatos y ponderadores vinculados con el cálculo.
`remarks`	Observaciones adicionales sobre la métrica. Este campo puede ser nulo si no se requiere anotación extra.

Tabla 7.3: Estructura de la salida de un módulo de calidad

Confidence

La métrica *Confidence* fue diseñada para evaluar la confianza asociada a un creador de contenido en redes sociales dentro del dominio de la medicina, específicamente enfocándose en expertos en colesterol y enfermedades cardiovasculares. Esta métrica combina variables que reflejan la influencia y credibilidad del usuario en este dominio.

La métrica se define como una combinación ponderada de dos aspectos principales:

- **Influence**: Nivel de influencia del usuario, representado por un valor entre 0 y 1. Este valor es generado previamente por el módulo Influencia Médica, que considera las recomendaciones o comentarios de otros usuarios, siempre y cuando estén asociadas a su conocimiento y experiencia en medicina y colesterol.
- **Audience Size**: Tamaño de la audiencia del usuario, que incluye seguidores, amigos o conexiones, dependiendo de la plataforma. Este indicador refleja el alcance potencial del contenido relacionado con la medicina.

La fórmula definida para calcular *Confidence* es:

$$\text{Confidence} = \alpha \cdot \text{Influence} + (1 - \alpha) \cdot \frac{\log(\text{Audience Size} + 1)}{\log(\text{Max Audience Size} + 1)}$$

Donde:

- α es un parámetro de ponderación que asigna mayor peso a la influencia ($\alpha = 0.7$ en nuestra implementación).
- $\log(\text{Audience Size} + 1)$ se utiliza para normalizar el tamaño de la audiencia.
- $\log(\text{Max Audience Size} + 1)$ asegura que la normalización se encuentra en un rango de $[0, 1]$, utilizando un tamaño máximo de audiencia ($\text{Max Audience Size} = 100,000$) como referencia.

En caso de que no se encuentren valores para *Influence* o *Audience Size* en el clip normalizado, se asume un valor de 0. Esto asegura que el cálculo de la métrica nunca falle, incluso en situaciones donde los datos son incompletos. En estos casos, el resultado de *Confidence* será 0.

El uso del logaritmo natural permite ajustar el impacto del tamaño de la audiencia, evitando que valores extremadamente altos de *Audience Size* dominen el cálculo. Según el experto de dominio, no es habitual en este contexto encontrar referentes con gran tamaño de audiencia. La normalización garantiza que el valor resultante de la métrica se encuentre en el rango $[0, 1]$.

Certification

La métrica *Certification* fue diseñada para evaluar el nivel de certificación profesional y académica de un creador de contenido en el dominio de la medicina, con un enfoque específico en expertos en colesterol y enfermedades cardiovasculares. Esta

métrica combina dos aspectos: la calidad de la educación formal del usuario y la relevancia de su profesión médica declarada.

Certification se define como una combinación ponderada de los siguientes factores:

- **Education Quality**: Un valor entre 0 y 1 que indica la calidad y relevancia de la formación académica del usuario en el campo de la medicina. Este valor es generado por el módulo Educación Médica.
- **Profession Quality**: Un valor entre 0 y 1 que refleja la relevancia de la profesión declarada por el usuario en el contexto médico, generado por el módulo Profesión Médica.

La fórmula utilizada para calcular *Certification* es:

$$\text{Certification} = \alpha \cdot \text{Education Quality} + (1 - \alpha) \cdot \text{Profession Quality}$$

Donde:

- α es un parámetro de ponderación que asigna mayor peso a la calidad educativa formal. En nuestra implementación, se establece como $\alpha = 0.6$.
- *Education Quality* y *Profession Quality* son valores generados por los módulos correspondientes, con un rango entre 0 y 1.

La selección de $\alpha = 0.6$ se basa en tres consideraciones. Primero, la educación formal es clave para establecer credibilidad profesional en el ámbito médico. Los títulos y especializaciones cumplen estándares reconocidos y son potencialmente verificables. Segundo, la experiencia profesional complementa la formación, pero su relevancia varía según el área del usuario. Por ejemplo, un médico general con poca experiencia en colesterol puede ser menos relevante que un cardiólogo. Finalmente, $\alpha = 0.6$ asigna un 60 % de peso a la educación formal y un 40 % a la profesión, priorizando la certificación académica sin descartar la experiencia.

En caso de que alguno de los atributos (*Education Quality* o *Profession Quality*) no esté presente en el clip normalizado, se asume un valor de 0.

Si *Education Quality* o *Profession Quality* no están disponibles, su valor se define como 0. El cálculo asegura que *Certification* siempre esté en el rango $[0, 1]$.

Access to Source Data

La métrica *Access to Source Data* se diseñó para evaluar la capacidad de verificación de las credenciales y afirmaciones hechas por un creador de contenido en

redes sociales dentro del dominio de la medicina. Esta métrica mide específicamente la calidad y disponibilidad de los datos fuente que respaldan estas afirmaciones.

El cálculo de esta métrica utiliza directamente el valor (*medicalVerifiability*) generado por el módulo Verificabilidad Médica.

En caso de que el módulo *MedicalVerifiabilityModule* no haya generado un valor para el feature *MedicalVerifiability*, se asigna un valor por defecto de 0.

7.3. Ejecución del experimento

En los siguientes párrafos se presenta el detalle del diseño y ejecución del experimento.

7.3.1. Metodología de clasificación por expertos de dominio

Para la ejecución del experimento, se partió de un conjunto de 97 clips, extraídos de diferentes redes sociales, con *keywords* vinculadas a estatinas y colesterol. Sobre esta muestra, se realizó una preselección de once clips [1] para que fueran clasificados en términos de credibilidad por el experto del dominio. Estos clips fueron seleccionados con dos criterios, variedad en términos de redes sociales y un juicio no experto para determinar nivel de credibilidad, de forma de tener un balance entre clips creíbles y no creíbles.

El experto de dominio recibió estos once clips provenientes de tres redes sociales: *Instagram*, *Facebook*, *X* y *TikTok*. Si bien se trata de una cantidad acotada de clips, esta selección permitió un análisis detallado de cada caso.

Se le indicó al experto del dominio que clasificara cada clip en una de las siguientes categorías:"creíble", "medianamente creíble", "no creíble" y "no clasificable". Además, se le solicitó asignar a cada clip una puntuación en una escala de 0 a 10, donde 10 representa "muy creíble" y 0 "nada creíble".

En colaboración con el experto de dominio, se definieron los umbrales para clasificar los niveles de credibilidad, como se detalla en el Cuadro 7.4. Es importante destacar que la categoría de credibilidad baja abarca la mitad del rango total (0.00 a 0.50). Esto refleja un enfoque conservador, propio del dominio médico. Este criterio permite enfatizar la importancia de diferenciar claramente entre contenidos de baja credibilidad y aquellos con mayor potencial de confianza.

[1] https://github.com/dsgarcia/clips-credibility

Nivel de Credibilidad	Inferior	Superior
Alto	0.76	1.00
Medio	0.51	0.75
Bajo	0.00	0.50

Tabla 7.4: Rangos de Niveles de Credibilidad

# Clip	Nivel de Confianza	Puntuación
1	Medio	0.6
2	Medio	0.7
3	Medio	0.6
4	Medio	0.7
5	Medio	0.6
6	Bajo	0.0
7	Bajo	0.0
8	Bajo	0.0
9	Bajo	0.1
10	Bajo	0.0
11	Bajo	0.0

Tabla 7.5: Niveles de confianza y métricas asignadas por usuario experto

7.3.2. Justificación métricas y ponderadores

A continuación se justifican los valores de los ponderadores seleccionados para el caso de estudio.

Métricas y ponderadores para el cálculo de *Trustworthiness*

De acuerdo a los definido por el experto de dominio, la selección de métricas para el cálculo del *Trustworthiness* fueron las siguientes: *Access to Source Data*, *Certification* y *Confidence*.

- ***Access to Source Data*:** evalúa la capacidad del contenido de ser verificable a través de datos fuente, como instituciones médicas y publicaciones científicas. Se asignó un peso de 0.4.
- ***Certification*:** mide el nivel de acreditación o validación profesional del creador del contenido, como títulos médicos, afiliaciones a instituciones reconocidas o certificaciones relevantes. Se le otorgó un peso del 0.4.
- ***Confidence*:** refleja la influencia y el alcance del usuario dentro de la red social, considerando elementos como el tamaño de la audiencia y las interacciones

positivas de los seguidores. Se le asignó un peso de 0.2. Se definió menor a los otros factores, para evitar que un alto número de seguidores domine el cálculo. Se entendió que esto podría asignar mayores índices de confianza en cuentas de usuarios que, a pesar de tener muchos seguidores y buenos comentarios, podrían pertenecer a disciplinas pseudocientíficas.

La Fórmula 7.1 resume los pesos asignados, por el experto de dominio, a las métricas de cada factor.

$$\text{Trustworthiness} = 0.4 \cdot \text{Access to Source Data} + 0.4 \cdot \text{Certification} + 0.2 \cdot \text{Confidence} \tag{7.1}$$

Ponderadores para el cálculo de *Credibility*

La métrica *Credibility* combina dos dimensiones: *Trustworthiness Path Stability* y *Trustworthiness*, con ponderadores asignados de 0.6 y 0.4, respectivamente, tal como se observa en la ecuación 7.2

Estos pesos, también asignados en conjunto con el experto de dominio, buscan darle relevancia a la estabilidad del contenido a lo largo de su *provenance*. Según el criterio del experto del dominio, que el clip sea compartido por otras cuentas confiables aumenta la credibilidad del contenido original.

$$\text{Credibility} = 0.6 \cdot \text{Trustworthiness Path Stability} + 0.4 \cdot \text{Trustworthiness} \tag{7.2}$$

En caso de que un clip no tenga *provenance* disponible, se le asigna una ponderación de 0 al Trustworthiness Path Stability y de 1 al Trustworthiness.

7.3.3. Análisis de resultados

El Cuadro 7.6 muestra los resultados finales de las métricas de *Trustworthiness*. Desde esta perspectiva, sin aún considerar la dimensión *Provenance*, se puede observar que los primeros cinco clips están dentro de la franja de credibilidad media, esto es, clips creíbles. Tres son provenientes de *Facebook* y dos de *X*.

En los Cuadros 7.7, 7.8, 7.9 y 7.10 se presentan los resultados obtenidos por los módulos de generación de features Profesión Médica, Educación Médica, Verificabilidad Médica e Influencia Médica.

Los resultados muestran que los clips con referencias explícitas a instituciones médicas reconocidas, como universidades u hospitales, obtuvieron puntuaciones su-

periores en *AAccess to Source Data*, impactando fuertemente en el factor *Verifiability*. Esto puede observarse en los clips 2, 3, 4 y 5 del Cuadro 7.9. Sin embargo, los clips que carecen de estas referencias presentaron valores significativamente bajos, reflejando la importancia de la verificabilidad de las afirmaciones, tal como se observa en los clips 6 a 11.

El valor de *Certification* también demostró ser relevante para identificar fuentes confiables. Los clips publicados por médicos especializados, con títulos claros y experiencia documentada, obtuvieron altos valores en esta métrica. Esto puede observarse en los clips 2, 3, 4 y 5. En estos casos, en los Cuadros 7.8 y 7.7 se observa la claridad y calidad de la evidencia presentada en los perfiles de esos clips. Esta métrica impacta positivamente en el factor *Expertise*.

Por otro lado, los perfiles con títulos vagos o autoproclamados, como "naturópata", siendo el caso del clip 7, fueron calificados con valores bajos, evidenciando la sensibilidad del modelo ante perfiles pseudocientíficos.

La métrica *Confidence* reflejó la influencia de los usuarios en la red social, pero con un peso menor en el cálculo total de *Trustworthiness* para evitar sesgos derivados del tamaño de la audiencia. Esto fue clave para evitar que perfiles con gran cantidad de seguidores influyeran de manera desproporcionada. Por ejemplo, el clip 1, aunque proveniente de un perfil popular, fue evaluado principalmente por la calidad de sus referencias, tal como se puede observar en el Cuadro 7.10, obteniendo un valor alto en la métrica *Confidence*.

Al analizar las métricas de *credibility* presentadas en el Cuadro 7.11, se observa el impacto del factor *Trustworthiness Path Stability*. Los clips 1 y 5, que fueron compartidos por otros usuarios y mantuvieron una procedencia estable durante su propagación, obtuvieron los puntajes más altos de *credibility*, con valores de 0.80 y 0.78, respectivamente. Este resultado respalda la hipótesis de que los clips cuya estructura de *provenance* permanece inalterada y cuyos nodos contribuyen positivamente en términos de *Trustworthiness* aumentan su credibilidad. En otros términos, los clips compartidos por usuarios con un alto nivel de *Trustworthiness* son considerados más confiables que aquellos que no fueron compartidos o lo fueron por fuentes con menor respaldo.

Para los clips 2 a 4 y 6 a 11 no hay *provenance* disponible, esto eso, no hay evidencia de que hayan sido compartidos. En estos casos, el ponderador asignado al *Trustworthiness Path Stability* fue de 0 y 1 para el *Trustworthiness*.

En términos de eficiencia, el tiempo promedio de ejecución del flujo fue de 5220 milisegundos. Si bien es necesario realizar pruebas de carga con un volumen mayor de clips, los resultados preliminares son alentadores, ya que se utilizó una muestra representativa. Esto sugiere un buen desempeño inicial, aunque es imprescindible

validar la escalabilidad en condiciones más exigentes.

Los resultados sugieren que el modelo tiene potencial para distinguir los contenidos con respaldo médico verificable de aquellos con escasa o nula información confiable, aunque se requieren más datos para una evaluación concluyente.

Se observó una alta concordancia entre las categorías de credibilidad generadas por el modelo y las clasificaciones realizadas por el experto de dominio, lo que respalda la efectividad del modelo de calidad. En el Cuadro 7.12 se presentan las comparaciones entre las puntuaciones del experto y las métricas obtenidas por el modelo.

La discrepancia se evidenció únicamente en los clips 1 y 5, donde el modelo asignó un nivel de credibilidad "Alto", mientras que el experto los clasificó como "Medio". Esta diferencia se debe al aporte de la métrica *Trustworthiness_Path_Stability*. Es relevante señalar que el experto de dominio no tuvo acceso a todos los perfiles que compartieron los clips, especialmente cuando estos fueron difundidos en otras plataformas de redes sociales, por lo que, tendió a asignar un puntaje de credibilidad menor, coincidente con la métrica de *Trustworthiness*, esto es, sin la afectación positiva del *Trustworthiness_Path_Stability*

En conclusión, el modelo de calidad demostró ser efectivo al evaluar la credibilidad de los clips analizados, destacando la relevancia de las dimensiones de *Trustworthiness* y *Provenance*. No obstante, se recomienda ampliar la muestra y validar el enfoque en entornos más diversos para garantizar su robustez y escalabilidad.

# Clip	Red	Access to Data	Certification	Confidence	Trustworthiness
1	Facebook	0.20	0.92	0.87	0.6227
2	Facebook	0.60	1.00	0.22	0.6844
3	Facebook	0.60	0.96	0.28	0.6806
4	X	0.50	1.00	0.27	0.6371
5	X	0.60	1.00	0.27	0.6945
6	Facebook	0.00	0.08	0.30	0.0929
7	Facebook	0.10	0.20	0.28	0.1754
8	X	0.10	0.04	0.27	0.1101
9	X	0.10	0.00	0.23	0.0870
10	X	0.10	0.00	0.20	0.0808
11	TikTok	0.20	0.56	0.30	0.3649

Tabla 7.6: Métricas de *Trustworthiness*

# Clip	Feature Value	Confidence Level	Processing Time (ms)
1	Cardiólogo	0.8	1395
2	Cardióloga	1.0	1288
3	Cardiólogo	0.9	1328
4	Médico Internista	1.0	1307
5	Nefrólogo	1.0	1314
6	Médico Integrativo	0.2	1457
7	Naturópata	0.2	1320
8	Ninguno	0.1	1301
9	Ninguno	0.0	1268
10	Ninguno	0.0	1431
11	Doctor	0.2	1270

Tabla 7.7: Valores del feature Profesión Médica

# Clip	Feature Value	Confidence Level	Proc. Time (ms)
1	Cardiólogo	1.0	1286
2	Cardióloga, Maestría en Diabetes	1.0	1316
3	Medicina y Profesor en Medicina	1.0	1430
4	Especialista en Medicina Interna	1.0	1393
5	Nefrólogo	1.0	1314
6	Ninguno	0.0	1296
7	Naturópata	0.2	1345
8	Ninguno	0.0	1257
9	Ninguno	0.0	1269
10	Ninguno	0.0	1418
11	Doctor	0.8	1281

Tabla 7.8: Valores del feature Educación Médica

# Clip	Feature Value	Conf. Level	Proc. Time (ms)
1	Bajo (Cardiólogo Universitario, Director de comunicaciones SIAC y LAHRS; falta información sobre universidades o instituciones específicas)	0.2	1534
2	Medio (Hospital Provincial Dr René Favaloro 'In Memoriam', Facultad de Ciencias Biomédicas - Universidad Austral, U.N.R.)	0.6	1520
3	Media (USAL - Universidad del Salvador, Centro Médico Sala Salud, Hospital Argerich)	0.6	1460
4	Bajo (Docente IRM-USFQ)	0.5	1420
5	Medio (Hospital Civil de Guadalajara, Universidad de Guadalajara)	0.6	1426
6	Baja (Ninguna universidad, hospital u otra organización relevante mencionada)	0.0	1936
7	Baja (Ninguna universidad, hospital u otra organización relevante mencionada)	0.1	1389
8	Bajo (Universidad de Barcelona y Hospital Clínic Barcelona mencionados en el mensaje, pero sin verificación de la afiliación del usuario a ninguna institución médica.)	0.1	1535
9	Baja (ninguna evidencia de educación médica o afiliación a instituciones médicas)	0.1	1429
10	Bajo (Ninguna universidad, hospital u otra organización reconocida es mencionada en el perfil. Solo se menciona el Dr. Ludwig Johnson sin ninguna evidencia de verificación independiente.)	0.2	1669
11	Media (Hospital Universitario Nacional, Facultad de Medicina, Universidad Nacional de Colombia)	0.7	1600

Tabla 7.9: Valores del feature Verificabilidad Médica

# Clip	Valor	Conf.	Proc. Time (ms)
1	Excelente reputación, basado en múltiples testimonios positivos	0.95	37256
2	Índice de reputación médica no disponible	0.00	1466
3	Índice de reputación no disponible debido a la falta de información	0.00	1456
4	Índice de reputación médica no disponible	0.00	3462
5	Índice de reputación médica no disponible	0.00	1329
6	Índice de reputación médica no disponible	0.00	1420
7	Índice de reputación médica no disponible	0.00	1362
8	Índice de reputación no disponible	0.00	831
9	Índice de reputación médica no disponible	0.00	1385
10	Índice de reputación no disponible	0.00	1828
11	Índice de reputación médica no disponible	0.00	1983

Tabla 7.10: Valores del feature Influencia Médica

# Clip	Trustworthiness	Trustworthiness_Path_Stability	Credibility
1	0.62	0.92	0.80
2	0.68	0.00	0.68
3	0.68	0.00	0.68
4	0.64	0.00	0.64
5	0.69	0.83	0.78
6	0.09	0.00	0.09
7	0.18	0.00	0.18
8	0.11	0.00	0.11
9	0.09	0.00	0.09
10	0.08	0.00	0.08
11	0.36	0.00	0.36

Tabla 7.11: Valores de las métricas de *Credibility*

# Clip	Nivel Cred. Exp.	Puntuación Exp.	Nivel Cred. Mod	Credibility	Desviación
1	Medio	0.6	Alto	0.80	0.20
2	Medio	0.7	Medio	0.68	0.02
3	Medio	0.6	Medio	0.68	0.08
4	Medio	0.7	Medio	0.64	0.06
5	Medio	0.6	Alto	0.78	0.18
6	Bajo	0.0	Bajo	0.09	0.09
7	Bajo	0.0	Bajo	0.18	0.18
8	Bajo	0.0	Bajo	0.11	0.11
9	Bajo	0.1	Bajo	0.09	0.01
10	Bajo	0.0	Bajo	0.08	0.08
11	Bajo	0.0	Bajo	0.36	0.36

Tabla 7.12: Puntuación experto de dominio y métrica de credibilidad (resaltando diferencias)

Capítulo 8

Conclusiones

Para concluir nuestro trabajo, se resumen los principales aportes y las posibles líneas de trabajo futuro.

8.1. Aportes de nuestro trabajo

Los principales aportes de nuestro trabajo se resumen en los siguientes puntos.

- Se realiza un estudio del estado del arte en torno a la evaluación de la confianza en redes sociales. Este análisis permite identificar los diferentes abordajes de la temática en la literatura, así como también, relevar los vacíos existentes en los enfoques previos. Esta revisión proporciona una base para justificar nuestra propuesta de un marco teórico, que conecta la *Credibility* con conceptos como *Trustworthiness*, *Reputation*, *Expertise*, *Verifiability* y *Provenance*.

- Se presenta un modelo de calidad diseñado para evaluar la *Credibility* en clips en redes sociales, estructurado en torno a dos dimensiones principales: *Trustworthiness* y *Provenance*. La dimensión *Trustworthiness* se define como una dimensión de calidad que integra tres factores: *Reputation*, *Verifiability* y *Expertise*. El factor *Reputation* incluye métricas como *Engagement* y *Confidence*, que evalúan la interacción generada por el clip y la cantidad y calidad de los seguidores de su creador. *Verifiability* mide la capacidad de verificar la información contenida en un clip mediante métricas como *Access to Source Data*, *Modification Transparency* y *Access to Verification Metadata*. Finalmente, *Expertise* se enfoca en evaluar

el conocimiento y la experiencia del creador del clip a través de métricas como *Certification*, *Production*, *Influence* y *Experience*. Por otra parte, la dimensión *Provenance* permite rastrear el origen y la evolución del contenido a lo largo de su ciclo de vida. Se introduce un nuevo factor de calidad que denominamos *Trusthworthiness Path*. Esto ofrece un perspectiva de la evolución de la confianza a lo largo del tiempo, desde que un clip es generado en la red social, hasta que un usuario final consume la información.

- Se introduce un enfoque para reconstruir el *provenance* de un clip, con el objetivo de calcular las métricas de calidad definidas en el modelo presentado previamente. Mientras que la literatura existente se enfoca en la reconstrucción del *provenance* dentro de una única plataforma, este trabajo incorpora el impacto del *cross-sharing*, entendido como la práctica de compartir un clip entre diferentes plataformas. Para abordar este desafío, se define el concepto de equivalencia entre clips, que permite desarrollar métricas para medir la similitud entre clips de distintas redes sociales y determinar si forman parte del mismo flujo de información.

- se presenta un flujo de procesamiento modular y agnóstico de las redes sociales, diseñado para medir la *credibility* de clips, independientemente de la plataforma de origen. Se incluye el diseño de un formato genérico para representar clips, lo que permite unificar y normalizar los datos provenientes de distintas plataformas y asegurar su análisis uniforme. Además, se definen los usuarios del flujo, especificando sus necesidades y responsabilidades. El diseño también facilita la incorporación de nuevos métodos para el cálculo de métricas de calidad, manteniendo la flexibilidad conceptual frente a futuros requerimientos.

Por último, se desarrolla una implementación del flujo de procesamiento para evaluar la efectividad de las métricas. Esta implementación, montada sobre GCP (Google Cloud Platform), permite calcular las métricas de *credibility* al analizar un conjunto de clips previamente clasificados por expertos del dominio. De esta forma, se valida la utilidad práctica del modelo de calidad propuesto.

8.2. Trabajo Futuro

Los resultados obtenidos identifican áreas de investigación y de desarrollo adicionales:

- **Equivalencia de clips en múltiples formatos:** Extender los métodos de equivalencia de clips para incluir datos multimedia, como imágenes y videos, integrando la evaluación de texto y contenido audiovisual. Esto contribuye al *provenance* al permitir rastrear y relacionar publicaciones que combinan distintos formatos, incluso cuando han sido compartidas o transformadas entre plataformas.

- **Optimización de las métricas de credibilidad:** Explorar técnicas de aprendizaje automático para mejorar la inferencia de ponderadores en el cálculo de métricas. Esto incluye el uso modelos tradicionales y redes neuronales, para identificar patrones complejos en los datos que influyen en la *credibility*. Al mismo tiempo, esto permite retroalimentar al modelo de calidad con las etiquetas aplicadas por los usuarios experto del dominio

- **Escalabilidad y desempeño:** Mejorar el flujo de procesamiento para manejar grandes volúmenes de datos en tiempo real.

- **Visualización de grafos de *provenance*:** Desarrollar herramientas que representen de manera visual y navegable los grafos de *provenance*, facilitando a los usuarios finales y expertos del dominio el análisis de flujos de información.

- **Formatos de presentación de métricas de calidad:** Dado que los usuarios están expuestos a grandes volúmenes de clips y dedican poco tiempo al procesamiento de la información, es relevante desarrollar formatos de presentación de alto impacto para las métricas de calidad, facilitando su interpretación y uso. Estos formatos deben ser intuitivos y accesibles, permitiendo una rápida comprensión de la *credibility* de un clip, incluso en contextos de alta carga informativa. Un enfoque prometedor podría incluir visualizaciones interactivas, como gráficos simplificados o indicadores de nivel de confianza, adaptados al perfil del usuario y al dispositivo utilizado (escritorio o móvil). Además, sería beneficioso explorar la personalización de las representaciones, ajustándolas a las

preferencias culturales y lingüísticas del usuario consumidor, como se discute en el capítulo 3. Este desarrollo también debería considerar la integración de mecanismos que garanticen la privacidad de los datos, especialmente en contextos donde las métricas se calculen en tiempo real durante la interacción del usuario con el contenido.

- **Procesamiento en tiempo real respetando la privacidad:** Aprovechar las capacidades de los grandes modelos de lenguaje (LLM por sus siglas en inglés) para capturar y procesar datos de pantalla, permitiendo evaluar un clip en el momento en que el usuario lo visualiza en su dispositivo, ya sea escritorio o móvil, garantizando el respeto por la privacidad. El modelo podrá generar el clip para ingresarlo al flujo de procesamiento. Esto evita la necesidad de integración con la *API* de las diferentes plataformas de redes sociales, lo que reduce la dependencia de políticas externas y posibles restricciones. Adicionalmente, este enfoque abre la posibilidad de personalizar las métricas de calidad en función del contexto del usuario, adaptando las evaluaciones al idioma, las preferencias culturales y el propósito de la interacción con el contenido. Para garantizar la privacidad, es importante implementar técnicas de procesamiento local en el dispositivo del usuario, evitando la transferencia de datos sensibles a servidores externos.

- **Incorporación de factores culturales y lingüísticos:** Considerar los factores culturales y el lenguaje del usuario consumidor para ajustar y mejorar las métricas de *credibility*, asegurando su relevancia y aplicabilidad en contextos diversos. Esto implica adaptar las métricas a las particularidades culturales que pueden influir en la percepción de la *credibility*, como los valores sociales, las normas locales y las diferencias en la interpretación de ciertos contenidos. Además, se deben incluir estrategias para manejar variaciones lingüísticas, como dialectos, modismos y estructuras gramaticales propias del usuario consumidor, lo que permitirá una evaluación más precisa. La integración de estos factores también permite personalizar las visualizaciones y recomendaciones generadas, promoviendo una mayor confianza y usabilidad de las herramientas desarrolladas.

- **Incorporación de metadatos de calidad en el *provenance*:** Inte-

grar metadatos de calidad en el *provenance* generado, permitiendo una evaluación más precisa y completa de la *credibility* de los datos a lo largo de su ciclo de vida. Estas métricas consideraran aspectos como su completitud, coherencia y granularidad. Por ejemplo, una métrica de completitud podría medir qué proporción de las transformaciones e interacciones relevantes están documentadas en el *provenance*. Esto permitiría identificar posibles agujeros en la trazabilidad y mejorar la capacidad de los usuarios para confiar en los datos analizados.

- **Integración de perfiles de usuario en múltiples redes sociales:** Consolidar los perfiles de un mismo usuario en distintas plataformas para enriquecer las métricas de *credibility*, aprovechando *features* más diversos y completos. Una estrategia efectiva es utilizar los enlaces que los usuarios suelen incluir en sus minibios para identificar y relacionar sus perfiles en diferentes redes sociales. Esto permitiría aumentar la cantidad de datos disponibles para evaluar factores como *Reputation* y *Expertise*. Al mismo tiempo, esta entrategia permite reconstruir el *provenance* entre múltiples redes sociales de una manera más efectiva.

- **Encuestas para analizar los criterios de credibilidad en usuarios no expertos:** Proponer encuestas mediante plataformas como *Amazon Mechanical Turk* o *Prolific*. En estas, se pedirá a usuarios no expertos que cataloguen clips como confiables o no confiables. El objetivo es comparar los resultados de un modelo de calidad validado por expertos con las evaluaciones de los usuarios no expertos. Esto permitirá medir los criterios de credibilidad utilizados por este grupo. Este experimento resulta de interés para evaluar el sesgo de la población respecto al uso de estatinas y el colesterol.

Referencias bibliográficas

14:00-17:00. (s.f.). ISO/IEC 25012:2008. Consultado el 31 de diciembre de 2022, desde https://www.iso.org/standard/35736.html

89th EAS Congress, Helsinki 2021 – EAS. (s.f.). Consultado el 11 de febrero de 2023, desde https://eas-society.org/academy/89th-eas/

AlMansour, A. A., y Communication, T. S. o. D. I. a. W. (2014). A MODEL FOR RECALIBRATING CREDIBILITY IN DIFFERENT CONTEXTS AND LANGUAGES - A TWITTER CASE STUDY. *International Journal of Digital Information and Wireless Communications*, *4*(1), 53. Consultado el 24 de febrero de 2023, desde https://www.academia.edu/5873994/A_MODEL_FOR_RECALIBRATING_CREDIBILITY_IN_DIFFERENT_CONTEXTS_AND_LANGUAGES_A_TWITTER_CASE_STUDY

Al-Rakhami, M. S., y Al-Amri, A. M. (2020). Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter [Conference Name: IEEE Access]. *IEEE Access*, *8*, 155961-155970. https://doi.org/10.1109/ACCESS.2020.3019600

Alrubaian, M., Al-Qurishi, M., Alamri, A., Al-Rakhami, M., Hassan, M. M., y Fortino, G. (2019). Credibility in Online Social Networks: A Survey. *IEEE Access*, *7*, 2828-2855. https://doi.org/10.1109/ACCESS.2018.2886314

Amazon Mechanical Turk. (s.f.). Consultado el 19 de junio de 2023, desde https://www.mturk.com/

Amintoosi, H., y Kanhere, S. S. (2014). A Reputation Framework for Social Participatory Sensing Systems. *Mobile Networks and Applications*, *19*(1), 88-100. https://doi.org/10.1007/s11036-013-0455-x

Arolfo, F., Rodriguez, K. C., y Vaisman, A. (2020). Analyzing the Quality of Twitter Data Streams. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-020-10072-x

Atske, S. (2021, septiembre). News Consumption Across Social Media in 2021. Consultado el 29 de diciembre de 2022, desde https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/

Barbier, G., Feng, Z., Gundecha, P., y Liu, H. (2013). Provenance Data in Social Media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *4*(1), 1-84. https://doi.org/10.2200/S00496ED1V01Y201304DMK007

Batini, C., Palmonari, M., y Viscusi, G. (2012). The Many Faces of Information and their Impact on Information Quality. Consultado el 2 de enero de 2023, desde https://www.semanticscholar.org/paper/The-Many-Faces-of-Information-and-their-Impact-on-Batini-Palmonari/0342ad26413fd79bbb1eca95e09dc7bb8640a676

Batini, C., Rula, A., Scannapieco, M., y Viscusi, G. (2015). From Data Quality to Big Data Quality. *J. Database Manage.*, *26*(1), 60-82. https://doi.org/10.4018/JDM.2015010103

Batini, C., y Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*. Springer International Publishing. Consultado el 22 de mayo de 2019, desde https://www.springer.com/us/book/9783319241043

Bizer, C. (2007). Quality-Driven Information Filtering in the Context of Web-Based Information Systems. Consultado el 20 de mayo de 2019, desde https://refubium.fu-berlin.de/handle/fub188/10062

Bourne, L. E., Kole, J. A., y Healy, A. F. (2014). Expertise: defined, described, explained. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00186

Brown, S. (2021, junio). The case for new social media business models. Consultado el 29 de diciembre de 2022, desde https://mitsloan.mit.edu/ideas-made-to-matter/case-new-social-media-business-models

Buneman, P., Khanna, S., y Tan, W. C. (2001). Why and Where: A Characterization of Data Provenance. *ICDT*. https://doi.org/10.1007/3-540-44503-X_20

Byrd, K., Mansurov, A., y Baysal, O. (2016). Mining Twitter data for influenza detection and surveillance. *Proceedings of the International Workshop on Software Engineering in Healthcare Systems*, 43-49. https://doi.org/10.1145/2897683.2897693

Canini, K. R., Suh, B., y Pirolli, P. L. (2011). Finding Credible Information Sources in Social Networks Based on Content and Social Structure. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 1-8. https://doi.org/10.1109/PASSAT/SocialCom.2011.91

Cassa, C. A., Chunara, R., Mandl, K., y Brownstein, J. S. (2013). Twitter as a Sentinel in Emergency Situations: Lessons from the Boston Marathon Explosions. *PLoS Currents*, *5*. https://doi.org/10.1371/currents.dis.ad70cd1c8bc585e9470046cde334ee4b

Castillo, C., Mendoza, M., y Poblete, B. (2011). Information credibility on Twitter, 675-684. https://doi.org/10.1145/1963405.1963500

Choi, W., y Stvilia, B. (2015). Web credibility assessment: Conceptualization, operationalization, variability, and models. *Journal of the Association for Information Science and Technology*, *66*. https://doi.org/10.1002/asi.23543

Chuai, Y., Tian, H., Pröllochs, N., y Lenzini, G. (2023, julio). The Roll-Out of Community Notes Did Not Reduce Engagement With Misinformation on Twitter. Consultado el 26 de abril de 2024, desde https://arxiv.org/abs/2307.07960v1

Community Notes. (s.f.). Consultado el 26 de abril de 2024, desde https://communitynotes.x.com/guide/es/about/introduction

De Nies, T., Mannens, E., y Van de Walle, R. (2016). Reconstructing Human-Generated Provenance Through Similarity-Based Clustering. En M. Mattoso y B. Glavic (Eds.), *Provenance and Annotation of Data and Processes* (pp. 191-194). Springer International Publishing. https://doi.org/10.1007/978-3-319-40593-3_19

Dinh, L., y Parulian, N. (2020). COVID-19 pandemic and information diffusion analysis on Twitter [_eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.252]. *Proceedings of the Association for Information Science and Technology*, *57*(1), e252. https://doi.org/10.1002/pra2.252

Dong, H., Halem, M., y Zhou, S. (2013). Social Media Data Analytics Applied to Hurricane Sandy. *2013 International Conference on Social Computing*, 963-966. https://doi.org/10.1109/SocialCom.2013.152

Farmacovigilancia - OPS/OMS — Organización Panamericana de la Salud. (s.f.). Consultado el 21 de junio de 2023, desde https://www.paho.org/es/temas/farmacovigilancia

Flemming, A. (2010). Quality Characteristics of Linked Data Publishing Datasources. http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources

Fogg, B. J., y Tseng, H. (1999). The Elements of Computer Credibility [event-place: Pittsburgh, Pennsylvania, USA]. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 80-87. https://doi.org/10.1145/302979.303001

Gabel, S., Reichert, L., y Reuter, C. (2022). Discussing conflict in social media: The use of Twitter in the Jammu and Kashmir conflict [Publisher: SAGE Publications]. *Media, War & Conflict*, *15*(4), 504-529. https://doi.org/10.1177/1750635220970997

Gahr, M., Eller, J., Connemann, B. J., y Schönfeldt-Lecuona, C. (2017). Underreporting of adverse drug reactions: Results from a survey among physicians. *European Psychiatry*, *41*, S369. https://doi.org/10.1016/j.eurpsy.2017.02.377

Gaziano, C., y McGrath, K. (1986). Measuring the Concept of Credibility [Publisher: SAGE Publications]. *Journalism Quarterly*, *63*(3), 451-462. https://doi.org/10.1177/107769908606300301

Gil, Y., y Artz, D. (2007). Towards Content Trust of Web Resources. https://doi.org/10.2139/ssrn.3199370

Gomide, J., Veloso, A., Meira, W., Jr., Almeida, V., Benevenuto, F., Ferraz, F., y Teixeira, M. (2011). Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter. *Proceedings of the 3rd International Web Science Conference*, 3:1-3:8. https://doi.org/10.1145/2527031.2527049

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., y Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election [Publisher: American Association for the Advancement of Science]. *Science*, *363*(6425), 374-378. https://doi.org/10.1126/science.aau2706

Gupta, A., y Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, 2-8. https://doi.org/10.1145/2185354.2185356

Gupta, A., Lamba, H., Kumaraguru, P., y Joshi, A. (2013). Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. *Proceedings of the 22nd International Conference on World Wide Web*, 729-736. https://doi.org/10.1145/2487788.2488033

GWI. (s.f.). On-demand Consumer Research — GWI. Consultado el 30 de abril de 2024, desde https://www.gwi.com

Hargittai, E., Fullerton, L., Menchen-Trevino, E., y Thomas, K. Y. (2010). Trust Online: Young Adults' Evaluation of Web Content [Number: 0]. *International Journal of Communication*, *4*(0), 27. Consultado el 3 de marzo de 2023, desde https://ijoc.org/index.php/ijoc/article/view/636

Hovland, C., Janis, I., y Kelley, H. (1953). *Communication and persuasion*. Yale University Press.

Huracán Sandy [Page Version ID: 148397782]. (2023, enero). Consultado el 19 de junio de 2023, desde https://es.wikipedia.org/w/index.php?title=Hurac%C3%A1n_Sandy&oldid=148397782

Jenkins, H., Clinton, K., Purushotma, R., Robison, A. J., y Weigel, M. (s.f.). Confronting the Challenges of Participatory Culture: Media Education for the 21st Century.

Kolodny, L. (2023, septiembre). Elon Musk says Twitter, now X, is moving to monthly subscription fees and has 550 million users. Consultado el 23 de abril de 2024, desde https://www.cnbc.com/2023/09/18/musk-says-twitter-now-x-is-moving-to-monthly-subscriptions.html

Kušen, E., y Strembeck, M. (2018). Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. *Online Social Networks and Media*, *5*, 37-50. https://doi.org/10.1016/j.osnem.2017.12.002

Kuutila, M., Kiili, C., Kupiainen, R., Huusko, E., Li, J., Hosio, S., Mäntylä, M., Coiro, J., y Kiili, K. (2024). Revealing complexities when adult readers engage in the credibility evaluation of social media posts [arXiv:2303.09656 [cs]]. *Computers in Human Behavior*, *151*, 108017. https://doi.org/10.1016/j.chb.2023.108017

Kwon, S., Cha, M., Jung, K., Chen, W., y Wang, Y. (2013). Prominent Features of Rumor Propagation in Online Social Media [ISSN: 2374-8486]. *2013 IEEE 13th International Conference on Data Mining*, 1103-1108. https://doi.org/10.1109/ICDM.2013.61

Leu, D., McVerry, J., O'Byrne, W., Byrne, O., Kiili, C., Zawilinski, L., Everett-Cacopardo, H., Kennedy, C., y Forzani, E. (2011). The New Literacies of Online Reading Comprehension: Expanding the Literacy and Learning Curriculum. *Journal of Adolescent & Adult Literacy International Reading Association*, *55*, 5-14. https://doi.org/10.1598/JAAL.55.1.1

Leu, D., Forzani, E., Burlingame, C., Kulikowich, J., Sedransk, N., Coiro, J., y Kennedy, C. (2013, marzo). The New Literacies of Online Research and Comprehension: Assessing and Preparing Students for the 21st Century With Common Core State Standards. En S. Neuman (Ed.), *Quality Reading Instruction in the Age of Common Core Standards* (pp. 219-236). International Reading Association. https://doi.org/10.1598/0496.16

McCroskey, J. C., y Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement [Place: United Kingdom Publisher: Taylor & Francis]. *Communication Monographs*, *66*, 90-103. https://doi.org/10.1080/03637759909376464

MedHelp - Health community, health information, medical questions, and medical apps. (s.f.). Consultado el 21 de junio de 2023, desde https://www.medhelp.org/

Middleton, S. E., Middleton, L., y Modafferi, S. (2014). Real-Time Crisis Mapping of Natural Disasters Using Social Media [Conference Name: IEEE Intelligent Systems]. *IEEE Intelligent Systems*, *29*(2), 9-17. https://doi.org/10.1109/MIS.2013.126

Muerte y Discapacidad por Desinformación sobre las estatinas – Fundación Hipercolesterolemia Familiar. (s.f.). Consultado el 28 de enero de 2023, desde https://www.colesterolfamiliar.org/muerte-y-discapacidad-por-desinformacion-sobre-las-estatinas/

Oremus, W., Harwell, D., y Armus, T. (2023). A tweet about a Pentagon explosion was fake. It still went viral. *Washington Post*. Consultado el 23 de mayo de 2023, desde https://www.washingtonpost.com/technology/2023/05/22/pentagon-explosion-ai-image-hoax/

Política de verificación antigua en X. (s.f.). Consultado el 17 de diciembre de 2023, desde https://help.x.com/es/managing-your-account/legacy-verification-policy

Powering AI with human insight - Toloka AI. (2022, diciembre). Consultado el 17 de marzo de 2024, desde https://toloka.ai/

Prolific — Quickly find research participants you can trust. (s.f.). Consultado el 17 de marzo de 2024, desde https://www.prolific.com/

Quality-Driven Query Answering for Integrated Information Systems. (2002). En F. Naumann, G. Goos, J. Hartmanis y J. Van Leeuwen (Eds.). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45921-9

Ramos, D. M. V. (s.f.). Internet y las redes sociales: los nuevos médicos consultantes. *Revista Uruguaya de Cardiología*, *vol.34 no.1*. Consultado el 15 de diciembre de 2024, desde http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S1688-04202019000100056

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., y Benevenuto, F. (2019). (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. *The World Wide Web Conference*, 818-828. https://doi.org/10.1145/3308558.3313688

Ross, A. S., y Rivers, D. J. (2018). Discursive Deflection: Accusation of "Fake News" and the Spread of Mis- and Disinformation in the Tweets of President Trump [Publisher: SAGE Publications Ltd]. *Social Media + Society*, *4*(2), 2056305118776010. https://doi.org/10.1177/2056305118776010

Ruby, D. (2022, noviembre). Global Social Media User Statistics (2023) — Demographics & Trends. Consultado el 29 de diciembre de 2022, desde https://www.demandsage.com/social-media-users/

Sakaki, T., Toriumi, F., y Matsuo, Y. (2011). Tweet Trend Analysis in an Emergency Situation. *Proceedings of the Special Workshop on Internet and Disasters*, 3:1-3:8. https://doi.org/10.1145/2079360.2079363

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., y Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, *54*(Supplement C), 202-212. https://doi.org/10.1016/j.jbi.2015.02.004

Schomberg, J. P., Haimson, O. L., Hayes, G. R., y Anton-Culver, H. (2016). Supplementing Public Health Inspection via Social Media. [Publisher: Public Library of Science (PLoS)]. *PLoS ONE*, *11*(3), e0152117. https://doi.org/10.1371/journal.pone.0152117

Sen, M., y Morozov, E. (s.f.). Analysing the Twitter social graph: Whom can we trust?

Serra, F., Peralta, V., Marotta, A., y Marcel, P. (2022). Modeling Context for Data Quality Management. En J. Ralyté, S. Chakravarthy, M.

Mohania, M. A. Jeusfeld y K. Karlapalem (Eds.), *Conceptual Modeling* (pp. 325-335). Springer International Publishing. https://doi.org/10.1007/978-3-031-17995-2_23

Shah, A., Ravana, S. D., Hamid, S., y Ismail, M. A. (2015). Web credibility assessment: Affecting factors and assessment techniques. *Information Research*, *20*.

Simmhan, Y., Plale, B., y Gannon, D. (2005). A Survey of Data Provenance in e-Science. *SIGMOD Record*, *34*, 31-36. https://doi.org/10.1145/1084805.1084812

Social media as a news source worldwide 2024. (s.f.). Consultado el 9 de diciembre de 2024, desde https://www.statista.com/statistics/718019/social-media-news-source/

Swire-Thompson, B., y Lazer, D. (2020). Public Health and Online Misinformation: Challenges and Recommendations [_eprint: https://doi.org/10.1146/annurev-publhealth-040119-094127]. *Annual Review of Public Health*, *41*(1), 433-451. https://doi.org/10.1146/annurev-publhealth-040119-094127

Taxidou, I. (2018). *Information diffusion and provenance in social media.* [PhD Thesis]. University of Freiburg, Freiburg im Breisgau, Germany.

Taxidou, I., De Nies, T., Verborgh, R., Fischer, P. M., Mannens, E., y Van de Walle, R. (2015). Modeling Information Diffusion in Social Media as Provenance with W3C PROV. *Proceedings of the 24th International Conference on World Wide Web*, 819-824. https://doi.org/10.1145/2740908.2742475

Taxidou, I., y Fischer, P. M. (2014). Online analysis of information diffusion in twitter. *Proceedings of the 23rd International Conference on World Wide Web*, 1313-1318. https://doi.org/10.1145/2567948.2580050

Taxidou, I., Fischer, P. M., y Zablocki, M. (2018). A comparative study of social interactions and their combination on social media. *Proceedings of the VLDB Endowment*, *11*(9).

Wang, S., Su, F., Ye, L., y Jing, Y. (2022). Disinformation: A Bibliometric Review [Number: 24 Publisher: Multidisciplinary Digital Publishing Institute]. *International Journal of Environmental Research and Public Health*, *19*(24), 16849. https://doi.org/10.3390/ijerph192416849

Whitehead, J. L. (1968). Factors of source credibility. *Quarterly Journal of Speech*, *54*(1), 59-63. https://doi.org/10.1080/00335636809382870

Yang, C. C., Yang, H., y Jiang, L. (2014). Postmarketing Drug Safety Surveillance Using Publicly Available Health-Consumer-Contributed Content in Social Media. *ACM Trans. Manage. Inf. Syst.*, *5*(1), 2:1-2:21. https://doi.org/10.1145/2576233

Yelp. (s.f.). Consultado el 22 de junio de 2023, desde https://www.yelp.com/

Yun-tao, Z., Ling, G., y Yong-cheng, W. (2005). An improved TF-IDF approach for text classification. *Journal of Zhejiang University-SCIENCE A*, *6*(1), 49-55. https://doi.org/10.1631/BF02842477

APÉNDICES

Apéndice A

Artículos de dominios específicos

A pesar de los avances y resultados obtenidos en los trabajos analizados, se observa que ninguno de ellos aborda explícitamente la evaluación de la calidad de los datos utilizados. Asumen que la calidad con la que provienen de la plataforma de red social, es adecuada para los análisis posteriores.

A.1. Situaciones de emergencia

En Cassa et al. 2013 se analiza el uso de las redes sociales, específicamente X, como una fuente de información en tiempo real durante situaciones de crisis. Se enfoca en el caso de los ataques con bombas en el maratón de Boston en 2013, utilizando mensajes de X geolocalizados para analizar la respuesta inicial en las redes sociales. Los resultados muestran que los mensajes relacionados con la emergencia comenzaron a aparecer solo tres minutos después de las explosiones y proporcionaron información oportuna sobre el incidente. Sin embargo, se señala la necesidad de tener precaución al usar datos de redes sociales, ya que también pueden generar informes falsos. Por lo tanto, se sugiere el desarrollo de estrategias de clasificación y filtrado para mejorar la precisión de la información obtenida de las redes sociales durante situaciones de crisis. En general, se destaca el potencial de utilizar las redes sociales como una herramienta complementaria en la respuesta a emergencias, sin embargo no abordan los aspecto asociados a la calidad de los datos.

Sakaki et al. 2011 se centra en analizar cómo el terremoto que ocurrió en

Japón en marzo de 2011 afectó el uso de X. Los investigadores utilizaron registros de tráfico de la red para inferir qué hicieron las personas en la web durante y después del terremoto. También analizaron registros de actividad en X para ver cómo se utilizó la plataforma durante y después del desastre. Encontraron que en las áreas más afectadas, como Iwate y Miyagi, la frecuencia de tweets disminuyó debido a los daños causados por el terremoto y las dificultades para acceder a Internet. Sin embargo, en áreas menos afectadas, como Tokio y Osaka, el uso de X no se vio afectado significativamente. A diferencia de las áreas cercanas al epicentro del terremoto, donde hubo un descenso notable en la frecuencia de tweets, estas regiones mostraron una continuidad en el uso de la plataforma de redes sociales. En estas áreas, el uso dominante de X seguía siendo a través de computadoras personales (PCs), y la frecuencia de tweets se mantuvo estable antes y después del terremoto. Esto puede estar relacionado con el hecho de que estas áreas no experimentaron el mismo nivel de daño y de interrupción en la infraestructura de red, lo que permitió a los usuarios seguir accediendo a Internet y publicar en X con normalidad. Estos hallazgos sugieren que la capacidad de las personas para usar X y comunicarse a través de esta plataforma puede variar según la ubicación geográfica y el impacto de eventos catastróficos. Mientras que en áreas altamente afectadas, como Miyagi e Iwate, el uso de X se vio afectado debido a la devastación y la falta de acceso a Internet, en lugares menos dañados como Tokio y Osaka, los usuarios pudieron mantener una continuidad en su actividad en X. Estos resultados resaltan la importancia de comprender cómo las redes sociales, como X, pueden desempeñar un papel crucial en la comunicación y el intercambio de información durante eventos catastróficos. En este trabajo, tampoco se hace referencia explícita a la calidad de los datos procesados.

En Middleton et al. 2014 se presenta una plataforma de mapeo de crisis en redes sociales para desastres naturales. Utilizan información de ubicación de diversas fuentes para identificar áreas en riesgo y luego analizan los tweets en tiempo real para generar mapas de crisis. Evalúan la precisión de la identificación de ubicaciones y comparan los mapas generados con evaluaciones oficiales posteriores al evento. En particular, la plataforma presentada utiliza técnicas de geoparsing, esto es, la extracción información geográfica a partir de datos semi-estructurados. En particular, desde los tweets, se extraen lugares, calles y regiones. Los datos geográficos utilizados provienen de diferentes fuentes, como gazetteer (diccionarios geográficos), mapas de calles e información geográfica

voluntaria (VGI por sus siglas en inglés). Estos datos se utilizan para mejorar la precisión del geoparsing y generar mapas de crisis en tiempo real. La plataforma también utiliza análisis estadísticos para calcular una línea de base de las menciones de ubicaciones en los tweets y establecer un umbral para determinar qué ubicaciones se muestran en el mapa de crisis. Además, los autores efectuaron pruebas de precisión y comparaciones con fuentes de información verificadas, como imágenes de satélite y aéreas, para evaluar la calidad de los mapas generados. Es interesante mencionar que este trabajo presenta una arquitectura de uso general para cualquier situación de desastre natural, siendo de las primeras que hemos detectado en nuestro estudio del estado del arte. Cabe destacar que no aborda profundamente la credibilidad de la información procesada. Los autores asumen que el volumen de tweets, reportando un desastre sobre una misma área geográfica, es condición suficiente para asegurar confianza sobre la credibilidad de los datos.

En Gupta et al. 2013 se analiza el uso de *X* durante el huracán Sandy en 2012 para difundir imágenes falsas del desastre. Los autores se basaron en trabajos anteriores sobre la credibilidad en redes sociales, en particular Canini et al. 2011. Se identificaron 10.350 tweets únicos que contenían imágenes falsas, de un total de 1.782.526 tweets, que se compartieron en la red social durante el huracán. En la Figura A.1 se observa un ejemplo de fotografía falsa que circuló en *X*. Se realizó un análisis de caracterización para comprender los patrones temporales, de reputación social y de influencia en la propagación de estas imágenes falsas. Se detectó que el 86 % de los tweets que difundían las imágenes falsas eran retweets, lo que significa que muy pocos usuarios publicaron tweets originales con estas imágenes. Además, se encontró que solo 30 usuarios (0.3 % de los usuarios) fueron responsables del 90 % de los retweets de las imágenes falsas. El análisis también reveló que solo hubo un 11 % de superposición entre la red de retweets y la red de seguidores, lo que indica que la red de seguidores de un usuario en X tiene poco impacto en la propagación de estas imágenes falsas. Para la clasificación automática, los autores utilizaron dos técnicas de Aprendizaje Automático sobre dos conjuntos de características, las correspondientes al usuario de *X* y las correspondientes al contenido del tweet. Es importante destacar que este trabajo no ejecuta un trabajo de clasificación sobre la imagen compartida, sino sobre las características mencionadas, lo que implícitamente permite inferir la credibilidad del tweet, un aspecto central para nuestro trabajo. Por este motivo, nos resulta de interés

Figura A.1: Ejemplo de fotografía falsa extraída de Gupta et al. 2013

detallar los grupos de atributos seleccionados por los autores, los cuales están listados en las tablas A.1 y A.2.

También en Dong et al. 2013, los autores analizan el uso de datos de X durante el huracán Sandy. En este caso, desarrollan un sistema automatizado para recolectar, analizar y visualizar datos de X, integrando herramientas como limpieza de datos y preprocesamiento con algoritmos de aprendizaje automático. Si bien el artículo no aborda directamente aspectos asociados a la calidad de los datos, entendemos que sí lo hacen de forma implícita, mencionando la importancia de procesar los tweets para eliminar ruido (mensajes que no hacen referencia al huracán) y duplicados. De esta forma, mejoran las utilidad de los

Tabla A.1: Características del usuario

Características del usuario
Número de Amigos
Número de Seguidores
Relación Seguidores-Amigos
Número de listas
El usuario tiene una URL
El usuario es verificado

Tabla A.2: Características del tweet

Descripción de la variable
Longitud del Tweet
Número de Palabras
¿Contiene signo de interrogación?
¿Contiene signo de exclamación?
Número de signos de interrogación
Número de signos de exclamación
¿Contiene emoticón feliz?
¿Contiene emoticón triste?
¿Contiene pronombre de primera persona?
¿Contiene pronombre de segunda persona?
¿Contiene pronombre de tercera persona?
Número de caracteres en mayúscula
Número de palabras con sentimiento negativo
Número de palabras con sentimiento positivo
Número de menciones
Número de hashtags
Número de URLs
Recuento de retweets

datos para su posterior análisis.

A.2. Farmacovigilancia

De acuerdo a la OPS (Organización Panamericana de la Salud) "Farmacovigilancia - OPS/OMS — Organización Panamericana de la Salud", s.f., la farmacovigilancia es un campo de la medicina que se centra en el monitoreo y evaluación de la seguridad de los medicamentos una vez que han sido aprobados y están en uso generalizado en la población. El objetivo principal de la farmacovigilancia es detectar y prevenir los efectos adversos de los medicamentos y mejorar la seguridad del paciente. Esta disciplina implica la recolección sistemática y el análisis de los informes de eventos adversos y reacciones adver-

sas a los medicamentos (ADR: Adverse Drug Reaction), que pueden provenir de los pacientes, los profesionales de la salud y otros actores. Estos informes se utilizan para identificar nuevos y potenciales riesgos de seguridad y para evaluar la relación beneficio-riesgo de los medicamentos.

Uno de los desafíos que presenta la farmacovigilancia son las bajas tasas de reporte de efectos adversos a los medicamentos. Según Gahr et al. 2017 de 316 médicos elegibles, solo 176 completaron un formulario reportando un ADR (tasa de respuesta = 55.7 %). La mayoría de los médicos (77.8 %) afirmaron que informan raramente (33.5 %), muy raramente (33.5 %) o nunca (10.8 %) los ADR que han observado a la autoridad competente. La mayoría (69.9 %) no había informado ningún ADR.

El uso de plataformas de redes sociales, como *X* y *Facebook*, ha abierto nuevas oportunidades para la farmacovigilancia. Estas plataformas ofrecen una gran cantidad de datos generados por los usuarios en tiempo real, lo que permite un monitoreo más rápido de los eventos adversos asociados con el uso de medicamentos. La información proporcionada por los usuarios en las redes sociales puede ayudar a identificar efectos secundarios inesperados, nuevas interacciones medicamentosas y patrones de uso de medicamentos. Una de las mayores dificultades detrás de esta técnica, radica en verificar la credibilidad de las publicaciones.

En Yang et al. 2014 se discute el uso de datos de redes sociales para la farmacovigilancia. Los autores argumentan que los datos de redes sociales pueden ser una herramienta valiosa para la farmacovigilancia porque son más completos y oportunos que los datos tradicionales de vigilancia pasiva. El trabajo propone un enfoque de minería de reglas de asociación, para identificar la asociación entre un medicamento y una ADR a partir de las publicaciones. Evalúan la efectividad de su enfoque al compararlo con las alertas publicadas por la Administración de Alimentos y Medicamentos (FDA, por sus siglas en inglés). Los resultados muestran que su enfoque es capaz de identificar posibles ADR con un alto grado de precisión. Un aspecto interesante de este trabajo es que no utiliza redes sociales de uso general. Los experimentos los basa en publicaciones en la red social *MedHelp* "MedHelp - Health community, health information, medical questions, and medical apps", s.f. Los autores afirman que la sección de medicamentos de *MedHelp* es uno de los componentes más importantes y populares de la plataforma. Contiene decenas de miles de medicamentos presentados en su base de datos. Además de los datos técnicos de las

drogas, contiene una sección de Publicaciones en donde los usuarios pueden iniciar un hilo de discusión sobre un determinado medicamento, en el cual todos los usuarios pueden comentar. Se pueden encontrar miles de hilos de discusión bajo cada uno de ellos. En el proceso de extracción y análisis de comentarios presentado en este trabajo, no se hace referencia a ninguna evaluación de la credibility de los datos.

En Sarker et al. 2015 se revisan estudios publicados sobre el uso de redes sociales en farmacovigilancia. Identifican dos enfoques principales. El primero utiliza léxicos para detectar ADR a partir de diccionarios médicos, el cual sufre limitaciones debido al lenguaje informal que usan los usuarios en redes sociales. Esto provoca dificultades para coincidir términos y genera falsos positivos. El segundo enfoque se basa en aprendizaje supervisado. Estos métodos dependen de datos etiquetados manualmente por expertos, sin embargo, la falta de *datasets* anotados limita su uso y dificulta la comparación entre estudios. Los autores proponen un *framework* conceptual para la detección de ADR en redes sociales. Este marco organiza el proceso en cuatro etapas: recolección de datos, filtrado, extracción de menciones de ADR y análisis estadístico. Tal como se muestra en la Figura A.2, el *framework* proporciona un flujo claro para procesar los datos, sin embargo, no considera explícitamente la credibilidad de la información extraída.

Figura A.2: *Framework* propuesto en el estudio de Sarker et al. 2015

A.3. Inspección de Salud Pública

Según Schomberg et al. 2016, en Estados Unidos se reportan, anualmente, 38.4 millones de enfermedades transmitidas por alimentos, y se estima que ocurren un total de 48 millones de enfermedades transmitidas por alimentos

(reportadas y no reportadas), los autores afirman que el sistema tradicional de inspección de establecimientos no es suficiente para reducir estos guarismos. En el trabajo se discute el uso de datos de redes sociales, específicamente reseñas de restaurantes en Yelp "Yelp", s.f., para predecir violaciones del código de salud en restaurantes. Los investigadores construyeron un modelo predictivo utilizando reseñas de Yelp de restaurantes en San Francisco. El modelo incluyó palabras clave relacionadas con regulaciones del código de salud y síntomas de enfermedades transmitidas por alimentos, así como otros factores como el número de reseñas, el número de estrellas y el rango de precios del restaurante. A pesar de que los autores reportan que los modelos predictivos arrojan buenos resultados, no incorporan ningún aspecto asociado a la credibilidad de los datos extraídos de *Yelp*.

En Byrd et al. 2016, los autores exploran el uso de datos de X como una herramienta para la vigilancia de enfermedades, centrándose en la identificación de brotes de influenza. El trabajo demuestra cómo los datos de redes sociales pueden ser procesados, clasificados y visualizados en tiempo real para analizar la propagación de enfermedades. Para ello, los autores recopilan tweets a través de palabras clave relacionadas con síntomas de la gripe y emplean técnicas de procesamiento de lenguaje natural y análisis de sentimientos para distinguir los mensajes relevantes. El estudio presenta un sistema capaz de extraer patrones temporales y geográficos que permiten monitorear la evolución de los brotes en diferentes regiones. Si bien el trabajo destaca el valor de los datos de X en la vigilancia epidemiológica, no aborda explícitamente la calidad de los mismos. Los autores se centran en la utilidad de la información para identificar tendencias y desarrollar sistemas de alerta temprana, pero no consideran dimensiones formales de calidad.

En Gomide et al. 2011 se presenta un sistema de vigilancia epidemiológica en tiempo real, también basado en datos de X, enfocado en monitorear la propagación del dengue. El trabajo utiliza técnicas de análisis espacio-temporal y procesamiento de lenguaje natural para identificar tweets relevantes, para posteriormente, construir modelos predictivos sobre la incidencia de la enfermedad en diferentes regiones de Brasil. Aunque el artículo no aborda explícitamente la temática de calidad de datos, menciona la necesidad de realizar filtrados para excluir contenido irrelevante, como experiencias personales. Esta práctica se presenta como una estrategia para mejorar la precisión del análisis, aunque no se formaliza en términos de dimensiones o métricas de calidad.

Apéndice B

Ejemplo para inferir nuevos ponderadores usando un árbol de regresión

Este modelo se propone para evaluar la *credibility* de los clips basándose solo en la dimensión *Trustworthiness*. Por tanto, se deben de manejar métricas para *Reputation*, *Expertise* y *Verifiability*. La ponderación de estas métricas se ajusta mediante el entrenamiento de un árbol de regresión [1].

Preparación de Datos: Inicialmente, se definen los ponderadores para cada métrica de calidad de datos y se proporcionan los ponderadores establecidos por un usuario experto. Notar que estos ponderadores son específicos de un caso de uso concreto. Los ponderadores iniciales son $P_e = [0.3, 0.4, 0.3]$ para *Reputation*, *Expertise* y *Verifiability*, respectivamente. Al mismo tiempo, se cuenta con los valores de *credibility* determinados por el usuario experto, los cuales pueden diferir de los calculados a partir de los ponderadores. El usuario experto puede haber advertido que la *credibility* de un determinado clip no coincidía con la calculada, por lo cual apelando a su juicio establece un nuevo valor de *credibility*.

[1] https://github.com/dsgarcia/inference

Los vectores de características se definen como sigue:

$$\begin{aligned} \text{Reputation} &: [0.8, 0.6, 0.5], \\ \text{Expertise} &: [0.7, 0.9, 0.4], \\ \text{Verifiability} &: [0.5, 0.3, 0.6], \\ \text{CredibilityExpert} &: [0.75, 0.85, 0.65] \end{aligned}$$

Entrenamiento del Modelo: Se utiliza un `DecisionTreeRegressor` de `sklearn` con una profundidad máxima de 3. El modelo se entrena con los datos proporcionados, resultando en la asignación de nuevos ponderadores a cada una de las métricas basadas en su importancia calculada. En la sección C del Apéndice presentamos el código fuente utilizado.

Resultados y Análisis: Los nuevos ponderadores calculados por el modelo resaltan la importancia relativa de cada métrica en la evaluación de la credibilidad. Se re-calculan los vectores de características según los nuevos ponderadores.

Árbol de decisión: Para ayudar a comprender cómo el modelo evalúa la importancia de las distintas métricas en la determinación de la credibilidad de los clips, se proporciona en la Figura B.1 la visualización del árbol de decisión generado.

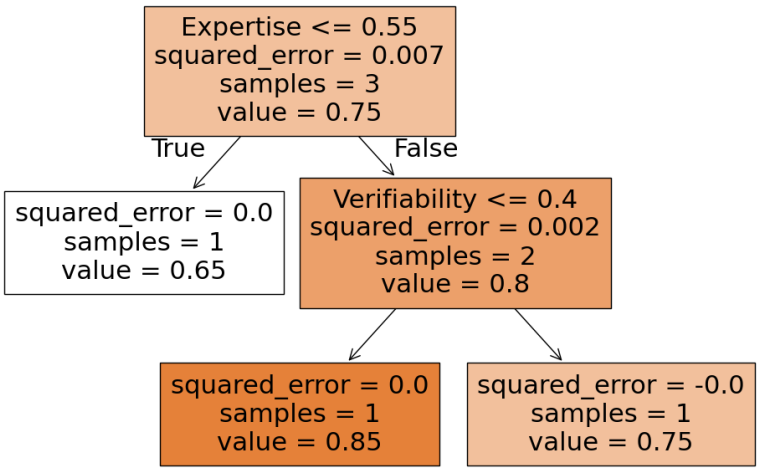

Figura B.1: Árbol de decisión para inferencia de ponderadores

Conclusión: Los resultados obtenidos del modelo de regresión de árbol de decisión muestran diferencias entre la *credibility* calculada con los ponderadores

originales y la *credibility* re-calculada utilizando los ponderadores derivados del modelo. Específicamente, los nuevos ponderadores predichos por el modelo, [0, 0.75, 0.25] ha priorizado la métrica de *Expertise*, mientras que la importancia de *Verifiability* ha sido reducida y la de *Reputation* ha obtenido un peso de 0. Esta redistribución de ponderadores sugiere que, según el modelo, *Expertise* es la dimensión más crítica para evaluar la *credibility* en este contexto.

La comparación entre la *credibility* experta y la *credibility* calculada con los nuevos ponderadores muestra discrepancias que subrayan la complejidad de calibrar modelos de evaluación de *credibility*. La aplicación de técnicas de aprendizaje automático ofrecen una herramienta de adaptabilidad y recalibración periódica de los modelos de evaluación de *credibility*. Esto permite reflejar con la mayor precisión posible el aporte del análisis efectuado por los usuarios expertos en un determinado caso de uso o dominio.

Apéndice C

Código utilizado para el ejemplo de Machine Learning

El código que se muestra a continuación no solo entrena un modelo para evaluar la credibilidad de los datos, sino que también compara la credibilidad calculada con los nuevos pesos del modelo con la credibilidad calculada a partir de los pesos originales y con la credibilidad proporcionada por un experto:

```
import numpy as np
from sklearn.tree import DecisionTreeRegressor, plot_tree
import matplotlib.pyplot as plt

# Inicializacion de variables y pesos originales
weights = np.array([0.3, 0.4, 0.3])  # Ponderadores originales
reputation_vector = np.array([0.8, 0.6, 0.5])
expertise_vector = np.array([0.7, 0.9, 0.4])
verifiability_vector = np.array([0.5, 0.3, 0.6])

# Credibilidad establecida por el usuario experto
credibility_expert = np.array([0.75, 0.85, 0.65])

# Combinacion de vectores en una matriz de caracteristicas
X = np.vstack((reputation_vector, expertise_vector,
    ↪ verifiability_vector)).T

# Entrenamiento del modelo de arbol de regresion
tree_reg = DecisionTreeRegressor(max_depth=3)
tree_reg.fit(X, credibility_expert)
```

```
# Importancias de las caracteristicas proporcionadas por el
    ↪ modelo
new_weights = tree_reg.feature_importances_
print("Nuevos␣pesos␣predichos␣por␣el␣modelo:", new_weights)

# Calculo de credibilidad con pesos originales y nuevos
credibility_original = np.dot(X, weights)
credibility_new = np.dot(X, new_weights)

# Comparacion de la credibilidad
print("Credibilidad␣con␣pesos␣originales:",
    ↪ credibility_original)
print("Credibilidad␣con␣nuevos␣pesos:", credibility_new)
print("Diferencia␣entre␣credibility_expert␣y␣credibility_new:",
    ↪  credibility_expert - credibility_new)

# Visualizacion del arbol de decision
plt.figure(figsize=(12, 8))
plot_tree(tree_reg, feature_names=["Reputation", "Expertise", "
    ↪ Verifiability"], filled=True)
plt.show()
```

Apéndice D

Esquema clip normalizado

```
{
  "clip": {
    "identification": {
      "clip_id": "string",
      "platform": "string",
      "clip_url": "string"
    },
    "content": {
      "message": "string",
      "message_type": "string"
    },
    "dates": {
      "creation_date": "datetime",
      "update_time": "datetime or null"
    },
    "geolocation": {
      "has_geo_data": "boolean",
      "coordinates": {
        "longitude": "float or null",
        "latitude": "float or null"
      }
    },
    "interaction": {
      "hashtags": ["string"],
      "links": ["string"],
      "mentions": ["string"],
```

```
    "reactions_breakdown": {
      "like": "int or null",
      "love": "int or null",
      "angry": "int or null",
      "other": "int or null"
    }
  },
  "multimedia": {
    "media_id": "string or null",
    "duration": "string or null",
    "audio_track": "string or null"
  },
  "comments": [
    {
      "comment_id": "string",
      "user_id": "string",
      "text": "string",
      "comment_date": "datetime"
    }
  ],
  "performance_metrics": {
    "view_count": "int or null",
    "impression_count": "int or null"
  },
  "metadata": {
    "extraction_metadata": {
      "extracted_at": "datetime",
      "processing_time_ms": "int",
      "source_details": "string"
    }
  },
  "features": {
    "generated_features": [
      {
        "module_name": "string",
        "feature_name": "string",
        "feature_value": "any",
        "confidence_level": "float",
```

```
          "feature_details": {
            "description": "string"
          },
          "calculation_details": {
            "calculated_at": "datetime",
            "calculation_method": "string",
            "processing_time_ms": "int"
          }
        }
      ]
    }
  },
  "user": {
    "identification": {
      "user_id": "string",
      "name": "string"
    },
    "personal_information": {
      "full_name": "string or null",
      "is_private": "boolean",
      "location": "string or null",
      "audience_size": "int or null"
    },
    "contact": {
      "business_contact_method": "string or null",
      "public_email": "string or null",
      "websites": ["string"]
    },
    "biography": "string or null",
    "experience_and_education": {
      "experience": ["string"],
      "education": ["string"]
    }
  },
  "quality_metrics": [
    {
      "clip_id": "string",
      "metric_name": "string",
```

```
      "metric_value": "float",
      "metric_factor": "string",
      "calculation_timestamp": "datetime",
      "calculation_method": "string",
      "quality_metadata": {
        "key": "any"
      },
      "remarks": "string or null"
    }
  ],
  "features": {
    "generated_features": [
      {
        "module_name": "string",
        "feature_name": "string",
        "feature_value": "any",
        "confidence_level": "float",
        "feature_details": {
          "description": "string"
        },
        "calculation_details": {
          "calculated_at": "datetime",
          "calculation_method": "string",
          "processing_time_ms": "int"
        }
      }
    ]
  }
}
```

Apéndice E

Esquema PROV-SAID con nuestras extensiones

Listing E.1: Esquema JSON para PROV-SAID con las extensiones incluidas en nuestro trabajo

```
{
  "id": "http://provenance.ecs.soton.ac.uk/prov-json/
    ↪ schema#",
  "$schema": "http://json-schema.org/draft-04/schema#",
  "description": "Schema for a PROV-JSON document adapted
    ↪  for PROV-SAID",
  "type": "object",
  "additionalProperties": false,
  "properties": {
    "prefix": {
      "type": "object",
      "patternProperties": {
        "^[a-zA-Z0-9_\\-]+$": { "type": "string", "format
          ↪ ": "uri" }
      }
    },
    "entity": {
      "type": "object",
      "additionalProperties": { "$ref": "#/definitions/
        ↪ entity" }
    },
    "activity": {
```

```
      "type": "object",
      "additionalProperties": { "$ref": "#/definitions/
        ↪ activity" }
    },
    "agent": {
      "type": "object",
      "additionalProperties": { "$ref": "#/definitions/
        ↪ agent" }
    },
    "wasGeneratedBy": {
      "type": "object",
      "additionalProperties": { "$ref": "#/definitions/
        ↪ generation" }
    },
    "used": {
      "type": "object",
      "additionalProperties": { "$ref": "#/definitions/
        ↪ usage" }
    },
    "wasDerivedFrom": {
      "type": "object",
      "additionalProperties": { "$ref": "#/definitions/
        ↪ derivation" }
    },
    "wasAttributedTo": {
      "type": "object",
      "additionalProperties": { "$ref": "#/definitions/
        ↪ attribution" }
    },
    "wasAssociatedWith": {
      "type": "object",
      "additionalProperties": { "$ref": "#/definitions/
        ↪ association" }
    }
  },
  "definitions": {
    "attributeValues": {
      "anyOf": [
```

```
        { "type": "string" },
        { "type": "number" },
        { "type": "boolean" },
        { "type": "array", "items": { "type": "string" }
          ↪ }
      ]
    },
    "entity": {
      "type": "object",
      "title": "Entity",
      "properties": {
        "prov:type": {
          "type": "string",
          "enum": ["prov-said:OriginalClip", "prov-said:
            ↪ CopiedClip", "prov-said:CommentedClip"]
        },
        "prov:label": { "type": "string" },
        "prov-said:SocialNetwork": { "type": "string", "
          ↪ format": "uri" },
        "credibility:trustworthiness": { "type": "number"
          ↪ , "minimum": 0, "maximum": 1 },
        "credibility:verifiability": { "type": "number",
          ↪ "minimum": 0, "maximum": 1 },
        "credibility:expertise": { "type": "number", "
          ↪ minimum": 0, "maximum": 1 },
        "credibility:reputation": { "type": "number", "
          ↪ minimum": 0, "maximum": 1 }
      },
      "required": ["prov:type", "prov:label", "prov-said:
        ↪ SocialNetwork"],
      "additionalProperties": { "$ref": "#/definitions/
        ↪ attributeValues" }
    },
    "agent": {
      "type": "object",
      "title": "Agent",
      "properties": {
```

```
        "prov:type": { "type": "string", "enum": ["prov:
          ↪ Agent"] },
        "prov:label": { "type": "string" }
      },
      "required": ["prov:type", "prov:label"],
      "additionalProperties": { "$ref": "#/definitions/
        ↪ attributeValues" }
    },
    "activity": {
      "type": "object",
      "title": "Activity",
      "properties": {
        "prov:type": { "type": "string", "enum": ["prov:
          ↪ Activity"] },
        "prov:startTime": { "type": "string", "format": "
          ↪ date-time" },
        "prov:endTime": { "type": "string", "format": "
          ↪ date-time" }
      },
      "required": ["prov:type"],
      "additionalProperties": { "$ref": "#/definitions/
        ↪ attributeValues" }
    }
  }
}
```