Universidad Católica
San Pablo

FACULTAD DE INGENIERÍA Y COMPUTACIÓN

DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN

Escuela Profesional de Ciencia de la Computación

Aplicación de Nubes de Palabras (Word Clouds) para la Identificación de Noticias Falsas en Plataformas Digitales

Tesis

Presentada por el bachiller

Renzo Estefanero Chavez

Para Optar por el Título profesional de:

Licenciado en Ciencia de la Computación

Asesor: Mg. Gina Lucia Muñoz Salas

** Arequipa, Mes Año **

A todas esas personas que me apoyaron en el desarrollo de este papel, gracias

Abreviaturas

NLP Procesamiento de lenguaje natural

TWC Nubes de Palabras Tradicionales

SWC Nubes de Palabras Semanticas

IA Inteligencia Artificial

CBOW *continuous bag of words*

GloVe Global Vectors for Word Representation

Agradecimientos

Aquí deberás colocar a quien y porque agradeces. Ejemplo:

En primer lugar deseo agradecer a Dios por haberme guiado a lo largo de estos cinco años de estudio.

Agradezco a mis padres por el apoyo brindado para forjarme como un profesional.

Agradezco a la universidad, mi *alma matter*, por haberme cobijado y brindado la formación que ahora me permitirá ayudar a construir una mejor sociedad.

Agradezco de forma muy especial a mi asesor Prof. Dr./Mag. nombre 1 por haberme guiado en esta tesis. ...

Deseo agradecer de forma especial a mis profesores: nombre 1, nombre 2, nombre 3 porque fueron ejemplos que deseo seguir en mi vida profesional.

Deseo agradecer al personal administrativo de la universidad: nombre 1, nombre 2, nombre 3. Muchas gracias por la atención brindada y porque siempre estuvieron dispuestas a ayudarnos.

Resumen

Aquí deberás colocar hasta 300 palabras como máximo, describiendo el problema que intentas resolver, la justificación, aportes o soluciones que planteas, qué tanto los haz alcanzado en calidad de resultados obtenidos.

Abstract

Here you should enter up to 300 words maximum, describing the problem you are trying to solve, the justification, contributions or solutions you are proposing, how much you have achieved them in terms of obtained outcomes.

Índice general

Índice de figuras

Capítulo 1

Introducción

1.1. Motivación y Contexto

En la era digital actual, las noticias falsas se propagan rápidamente, aprovechando las plataformas sociales y los medios digitales. Este fenómeno se caracteriza por la difusión de información errónea o manipulada, con el objetivo de engañar a la audiencia o generar confusión. Una de las características más peligrosas de las fake news es su capacidad para replicarse rápidamente, superando en muchos casos la velocidad de las noticias verdaderas.

Diversos estudios han abordado este fenómeno, y uno de los más significativos es el de Vosoughi et al. (2018), quienes demostraron que las noticias falsas se propagan aproximadamente un 70 % más rápido que las noticias verdaderas en plataformas como Twitter [20]. Este tipo de desinformación, a menudo vinculada a temas políticos, conspiraciones o alertas sensacionalistas, tiene un impacto desmesurado en la percepción pública, contribuyendo a la desinformación colectiva.

Como se observa en la Figura 1.1, el gráfico ilustra claramente esta diferencia en la velocidad de propagación entre noticias falsas y verdaderas. Mientras que las noticias falsas logran una difusión más acelerada, las noticias verdaderas siguen un patrón de propagación significativamente más lento, lo que subraya el riesgo que representan las fake news en términos de rapidez e impacto.

Este gráfico refuerza la necesidad urgente de herramientas que permitan detectar tempranamente este tipo de contenidos, facilitando así la intervención en etapas tempranas del proceso de propagación. Dado que las noticias falsas se diseminan mucho más rápido que las verdaderas, se requieren metodologías ágiles y accesibles que permitan a los investigadores, periodistas y usuarios comunes realizar un análisis preliminar y explorar patrones de desinformación de manera eficiente.

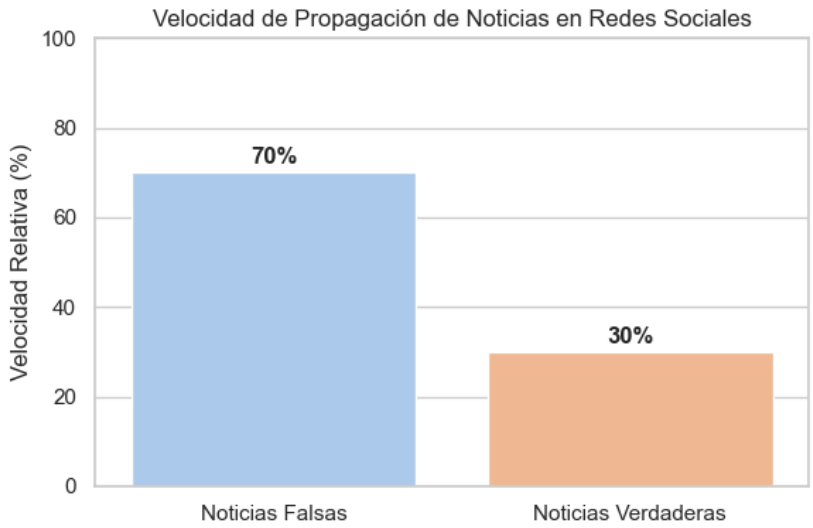

Figura 1.1: Velocidad relativa de propagación de noticias falsas y verdaderas en redes sociales, basado en [20].

1.2. Planteamiento del Problema

La detección de noticias falsas sigue siendo un desafío importante debido a su habilidad para imitar la estructura y estilo de noticias verídicas. Aunque se han desarrollado modelos de aprendizaje automático para abordar este problema, su aplicación generalizada es limitada debido a su complejidad y costo.

Actualmente, no existen suficientes herramientas simples, visuales y accesibles que permitan un análisis preliminar del contenido textual con el objetivo de identificar señales de desinformación. En particular, se carece de metodologías que integren la visualización y el análisis semántico de manera combinada, lo cual limitaría la capacidad de instituciones, investigadores o usuarios comunes de explorar ágilmente las diferencias entre noticias verdaderas y falsas.

1.3. Nubes de Palabras Tradicionales y Semánticas

Las nubes de palabras tradicionales Nubes de Palabras Tradicionales (TWC) representan gráficamente las palabras más frecuentes en un texto, asignando mayor tamaño tipográfico a aquellas de mayor frecuencia. Sin embargo, no consideran el contexto ni las relaciones entre palabras.

En cambio, las nubes de palabras semánticas Nubes de Palabras Semanticas (SWC) utilizan *word embeddings* (como Word2Vec o GloVe) para capturar la similitud semántica entre términos. Esto permite no solo visualizar frecuencia, sino también la proximidad semántica entre palabras, facilitando la detección de agrupaciones temáticas relevantes, como términos emocionales o narrativas específicas.

1.4. Objetivos

1.4.1. Objetivo General

Explorar el uso de TWC y SWC como herramienta visual y exploratoria para identificar patrones textuales y semánticos que puedan estar asociados a noticias falsas.

1.4.2. Objetivos Específicos

- Revisar los fundamentos teóricos y técnicos relacionados con las nubes de palabras y su aplicación en el análisis de texto.
- Recolectar y preparar un conjunto de datos compuesto por noticias verdaderas y falsas.
- Generar nubes de palabras tradicionales y semánticas para ambas categorías.
- Analizar visualmente las diferencias linguïsticas y semánticas entre noticias verdaderas y falsas.
- Evaluar la utilidad de estas visualizaciones como herramienta preliminar de apoyo en la detección de noticias falsas.

1.4.3. Objetivos de diseño de la visualización

- Garantizar la claridad visual para la identificación de patrones lexicológicos y semánticos.
- Favorecer el agrupamiento explícito de términos con alta proximidad semántica.
- Facilitar la interpretación para usuarios sin formación técnica especializada.
- Mantener consistencia visual entre las diferentes visualizaciones para facilitar comparaciones.

1.5. Organización de la Tesis

El Capítulo 2 presenta el marco teórico relacionado con las nubes de palabras, el procesamiento de lenguaje natural y la problemática de las noticias falsas.

El Capítulo 3 describe la metodología propuesta para generar y analizar nubes de palabras tradicionales y semánticas.

Posteriormente, se presentan los resultados de la aplicación metodológica, seguidos de las conclusiones y propuestas de trabajos futuros.

1.6. Cronograma

Día	Fecha	Actividad	Detalles
Sábado	12/04/2025	Redacción de la Introducción	Objetivos, contexto, motivación y planteamiento del problema.
Sábado	19/04/2025	Marco Teórico / Estado del Arte	Revisión y clasificación de papers, resumen de propuestas encontradas.
Sábado	26/04/2025	Redacción de la Propuesta	Modelos de representación de datos, justificación de WordClouds y métodos de construcción.
Sábado	03/05/2025	Refinamiento de la Propuesta	Mejoras y corrección de errores.
Sábado	10/05/2025	Plan de Tesis (Versión Parcial)	Revisión y ajustes de secciones anteriores.
Sábado	24/05/2025	Implementación de la Propuesta	Desarrollo inicial de WordClouds semánticas.
Sábado	31/05/2025	Evaluación de la Implementación	Pruebas con datasets seleccionados.
Sábado	07/06/2025	Ajustes en Implementación y Documentación	Explicación detallada del proceso, métricas utilizadas.
Sábado	14/06/2025	Experimentación	Recopilación de datos mediante encuestas.
Sábado	21/06/2025	Experimentación	Análisis de resultados de encuestas.
Sábado	28/06/2025	Redacción de Resultados	Integración final del documento de tesis.

Capítulo 2

Word Embeddings y Análisis Semántico de Noticias Falsas

2.1. Marco Teórico

El análisis de noticias falsas constituye un desafío contemporáneo con importantes implicancias sociales, al implicar el estudio del impacto de los textos en la formación de la opinión pública. La detección automática de este tipo de contenido se ha consolidado como un campo emergente dentro de la Inteligencia Artificial (IA), especialmente a través del uso de técnicas de Procesamiento de lenguaje natural (NLP). Entre los enfoques más utilizados destacan las representaciones vectoriales de palabras, conocidas como **embeddings**, que permiten transformar palabras en vectores en espacios de alta dimensión, capturando relaciones semánticas y contextuales de manera eficaz.

2.1.1. Procesamiento de Lenguaje Natural (NLP)

El **Procesamiento de Lenguaje Natural** (NLP) es una rama interdisciplinaria de la inteligencia artificial que estudia la interacción entre los sistemas computacionales y el lenguaje humano. Su objetivo principal es dotar a las máquinas de la capacidad para comprender, interpretar, generar y manipular texto o discurso de manera similar a como lo hace una persona. Entre sus aplicaciones más relevantes se encuentran la traducción automática, la clasificación de texto, el análisis de sentimientos y la detección de noticias falsas [4, 17, 22].

Las técnicas desarrolladas en este campo se apoyan en modelos estadísticos y computacionales capaces de procesar grandes volúmenes de datos textuales para identificar patrones lingüísticos y semánticos. Uno de los enfoques fundamentales es el uso de **embeddings de palabras**, representaciones vectoriales densas que posicionan las palabras en un espacio continuo de alta dimensión. En estas representaciones, términos con significados similares tienden a agruparse, lo que permite capturar relaciones semánticas y sintácticas de manera efectiva [12, 15]. Esta propiedad resulta particularmente útil en tareas como la detección de desinformación, donde es crucial identificar matices del lenguaje y estructuras

discursivas sutiles que podrían indicar contenido engañoso [16].

2.1.2. Word Embeddings

Las representaciones vectoriales de palabras, también conocidas como **word embeddings**, son una técnica fundamental en el campo del NLP, ya que permiten representar palabras como vectores densos en un espacio de alta dimensión. Estas representaciones preservan relaciones semánticas y sintácticas, permitiendo que palabras con significados similares estén cercanas entre sí en dicho espacio. Por ejemplo, términos como "desastre", "caos" y "emergencia" tienden a agruparse debido a su carga semántica y emocional similar.

Uno de los modelos más influyentes es **Word2Vec**, propuesto por Mikolov et al. (2013) [11, 12], que emplea redes neuronales poco profundas para aprender las representaciones a partir del contexto en que aparecen las palabras. Este modelo se basa en dos arquitecturas principales: *skip-gram*, que predice el contexto de una palabra dada, y *continuous bag of words* (CBOW), que predice la palabra central a partir de su contexto cercano. El entrenamiento de Word2Vec produce vectores donde relaciones semánticas pueden ser capturadas de manera aritmética, como en el ejemplo clásico:

$$\text{vector}(\text{“rey”}) - \text{vector}(\text{“hombre”}) + \text{vector}(\text{“mujer”}) \approx \text{vector}(\text{“reina”})$$

Otro enfoque relevante es **GloVe**, desarrollado por Pennington et al. (2014) [15], el cual combina ventajas de los métodos basados en conteo global y de predicción local. A diferencia de Word2Vec, GloVe utiliza una matriz de coocurrencia de palabras construida a partir de todo el corpus para generar sus vectores, optimizando una función de costo que modela las relaciones proporcionales de ocurrencia entre pares de palabras. Esto le permite capturar mejor relaciones globales en el texto, como similitudes analógicas y semánticas.

Finalmente, **FastText**, propuesto por Bojanowski et al. (2017) [3], extiende el modelo skip-gram de Word2Vec al considerar no solo las palabras, sino también los subcomponentes morfológicos de cada palabra (n-gramas de caracteres). Esto lo hace especialmente eficaz en lenguas con alta flexión y para representar palabras poco frecuentes o fuera del vocabulario (OOV, por sus siglas en inglés). Al componer la representación de una palabra a partir de sus subpalabras, FastText ofrece mayor robustez y generalización en tareas del lenguaje.

2.1.3. Generación de Nubes de Palabras

Las **nubes de palabras** (*word clouds*) son representaciones visuales que destacan los términos más frecuentes de un texto mediante variaciones en el tamaño tipográfico. En estas visualizaciones, cuanto mayor es la frecuencia de una palabra en el corpus, más grande y prominente se muestra. Esta técnica ofrece una manera rápida e intuitiva de explorar el contenido temático de un conjunto de datos textual, especialmente útil en etapas tempranas del análisis exploratorio.

Tradicionalmente, las nubes de palabras se construyen a partir de conteos de frecuencia sin considerar el contexto semántico de los términos. A pesar de esta limitación, su simplicidad y capacidad visual las han convertido en herramientas populares en disciplinas como el periodismo computacional, la sociolingüística, y el análisis de medios. En el ámbito de la detección de noticias falsas, su uso permite observar patrones léxicos comunes que podrían ser indicativos de desinformación. Por ejemplo, trabajos como el de Ahmed et al. (2018) han empleado nubes de palabras para visualizar términos recurrentes en noticias clasificadas como falsas frente a noticias verificadas, revelando el uso frecuente de palabras sensacionalistas o emocionalmente cargadas [2].

Para superar las limitaciones de las nubes basadas exclusivamente en frecuencia, se han desarrollado variantes más sofisticadas como las Nubes de Palabras Semanticas SWC. Estas emplean técnicas de representación vectorial, como los **word embeddings**. De este modo, en lugar de mostrar simplemente las palabras más frecuentes, estas visualizaciones permiten identificar clústeres semánticos y explorar relaciones conceptuales dentro del texto [7]. Este enfoque es particularmente útil cuando se trabaja con textos donde el contenido latente no puede inferirse únicamente a partir de la frecuencia superficial.

Además de su aplicación en la detección de noticias falsas, las nubes de palabras se han utilizado en dominios como la minería de opiniones [23], la exploración de discursos políticos [6], el análisis de emociones en redes sociales [9], y la educación digital, donde permiten sintetizar grandes volúmenes de contenido generado por usuarios o estudiantes. Estas aplicaciones demuestran su versatilidad como herramienta de visualización accesible, que, cuando se combina con métodos computacionales avanzados, puede enriquecer significativamente el análisis textual.

2.1.4. Word Embeddings y su Aplicación en la Detección de Noticias Falsas

La utilidad de los **word embeddings** en la detección automática de noticias falsas radica en su capacidad para representar de forma semántica el contenido textual, lo que permite identificar patrones lingüísticos comunes en textos engañosos. A diferencia de enfoques tradicionales basados únicamente en conteo de palabras, estas representaciones vectoriales permiten capturar matices sutiles del lenguaje, como el uso reiterado de términos alarmistas o emocionalmente cargados (por ejemplo, “urgente”, “crisis” o “desastre”), que suelen estar presentes en contenidos desinformativos [16].

Diversos estudios han demostrado que incorporar embeddings mejora el rendimiento de los modelos de clasificación en esta tarea. Por ejemplo, Wang (2017) presentó el conjunto de datos **LIAR**, compuesto por más de 12,000 declaraciones etiquetadas con veracidad, donde se demostró que el uso de representaciones semánticas superaba a métodos tradicionales basados en características superficiales [21]. De manera complementaria, Shu et al. (2018) introdujeron **FakeNewsNet**, un repositorio más amplio que combina el contenido textual con metadatos sociales y temporales, lo cual ha permitido entrenar modelos más robustos y contextualmente informados [17].

Además del uso directo en clasificación, los **embeddings** también pueden visuali-

zarse para entender mejor las relaciones semánticas entre palabras presentes en noticias verdaderas y falsas. Una técnica ampliamente utilizada para este propósito es **t-SNE** (t-distributed Stochastic Neighbor Embedding), que permite reducir la dimensionalidad de los vectores de palabras y proyectarlos en un plano 2D. En la Figura 2.1, se muestra una visualización generada con un subconjunto de palabras comunes en noticias falsas, donde puede observarse una agrupación semántica coherente, revelando el uso repetitivo de ciertos marcos discursivos.

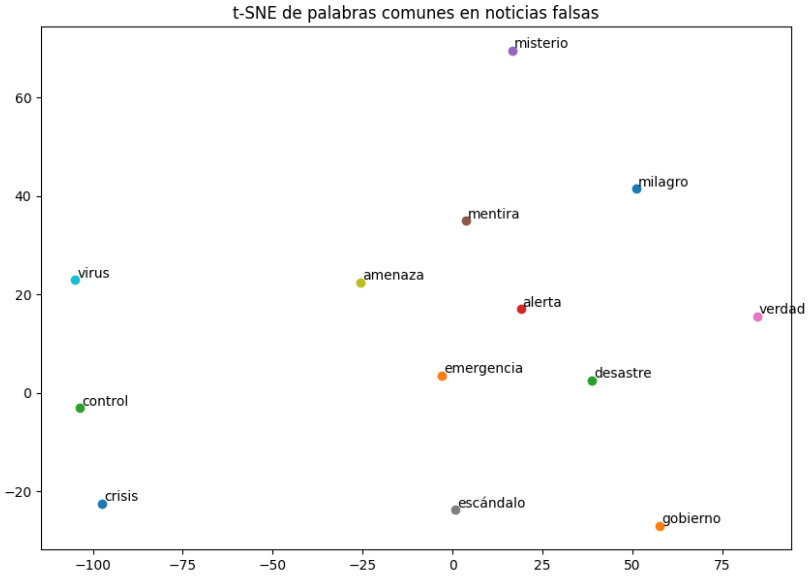

Figura 2.1: Proyección t-SNE de embeddings de palabras comunes en noticias falsas.

Este tipo de análisis no solo permite comprender el léxico utilizado en contenidos desinformativos, sino también explorar cómo las palabras se agrupan en torno a temas específicos como salud, política o catástrofes, facilitando el desarrollo de sistemas explicables de detección automática.

2.2. Estado del Arte

En la última década, el uso de **word embeddings** y técnicas de NLP ha ganado protagonismo en el desarrollo de sistemas automáticos para la detección de noticias falsas. Modelos como **Word2Vec**, **GloVe** y **FastText** han demostrado una notable capacidad para capturar relaciones semánticas y sintácticas en grandes volúmenes de texto, facilitando así la identificación de patrones lingüísticos característicos del discurso desinformativo. Estos enfoques han sido implementados con éxito en diversos estudios. Por ejemplo, Wang (2017) presentó el corpus **LIAR**, que contiene más de 12,000 declaraciones etiquetadas según su veracidad, y demostró cómo los embeddings mejoran el rendimiento de los clasificadores en comparación con modelos de texto plano [21]. De forma complementaria, Shu et al. (2018) desarrollaron **FakeNewsNet**, un repositorio más rico que integra contenido textual, información social y temporal, lo que permitió explorar la detección desde una perspectiva

más contextual [17].

Varios autores han propuesto técnicas basadas en embeddings como una forma de enriquecer los sistemas de detección con características semánticas. Rashkin et al. (2017) analizaron el lenguaje de noticias falsas, sátiras y hechos reales utilizando representaciones distribuidas para identificar matices emocionales y de veracidad [16]. Del mismo modo, Karimi y Tang (2019) combinaron embeddings con redes neuronales convolucionales (CNN) y modelos de atención para capturar dependencias de largo alcance en la estructura lingüística de textos falsos [8].

En cuanto a la visualización y análisis interpretativo, el uso de **nubes de palabras** ha evolucionado hacia enfoques más semánticamente informados. Feldman (2013) fue uno de los primeros en señalar las limitaciones de las nubes de palabras basadas solo en frecuencia y propuso su integración con análisis de sentimientos y técnicas de NLP para mejorar su capacidad interpretativa [4]. Más recientemente, Heylen et al. (2015) introdujeron el concepto de **nubes semánticas**, las cuales utilizan vectores de palabras para representar relaciones conceptuales, permitiendo visualizar agrupamientos semánticos más significativos [7]. Estas visualizaciones han sido adoptadas en el análisis de discursos políticos, minería de opiniones y estudios sobre desinformación, permitiendo observar patrones emocionales o temáticos de forma intuitiva.

Finalmente, se han reportado experimentos que combinan técnicas de reducción de dimensionalidad, como **t-SNE** o **UMAP**, con embeddings entrenados sobre corpora de noticias para representar gráficamente los clústeres léxicos predominantes en contenidos falsos. Esto ha facilitado la identificación visual de términos con fuerte carga emocional o ideológica. Por ejemplo, Shu et al. (2020) presentan un enfoque visual para analizar la propagación de fake news usando embeddings proyectados en espacios bidimensionales para revelar dinámicas semánticas en redes sociales [18].

Estos avances reflejan una evolución constante en las metodologías empleadas, donde la integración de representaciones semánticas, técnicas de visualización y datos contextuales permite abordar la detección de noticias falsas desde una perspectiva más explicativa y robusta.

2.3. Comparación entre Nubes de Palabras Tradicionales y Semánticas

Aunque las **nubes de palabras** tradicionales permiten observar la frecuencia de los términos más comunes en un corpus, su principal limitación es que no capturan relaciones semánticas entre palabras. Como resultado, términos que aparecen juntos o que son conceptualmente similares no necesariamente se agrupan, dificultando el análisis de patrones discursivos complejos.

Para superar estas limitaciones, se proponen las **nubes semánticas**, construidas a partir de proyecciones de embeddings mediante técnicas como **t-SNE**. Estas visualizaciones agrupan palabras según su proximidad semántica, más allá de su frecuencia, lo que resulta útil para identificar clústeres temáticos o emocionales.

La Figura 2.2 ilustra esta diferencia: mientras que la nube tradicional (izquierda) representa las palabras más frecuentes en el corpus de noticias falsas, la nube semántica (derecha), generada con **t-SNE**, revela agrupaciones de palabras con connotaciones emocionales o alarmistas, como "crisis", "emergencia" y "urgente", que forman un clúster semántico coherente. Esta comparación permite observar cómo cada tipo de visualización resalta diferentes aspectos del lenguaje.

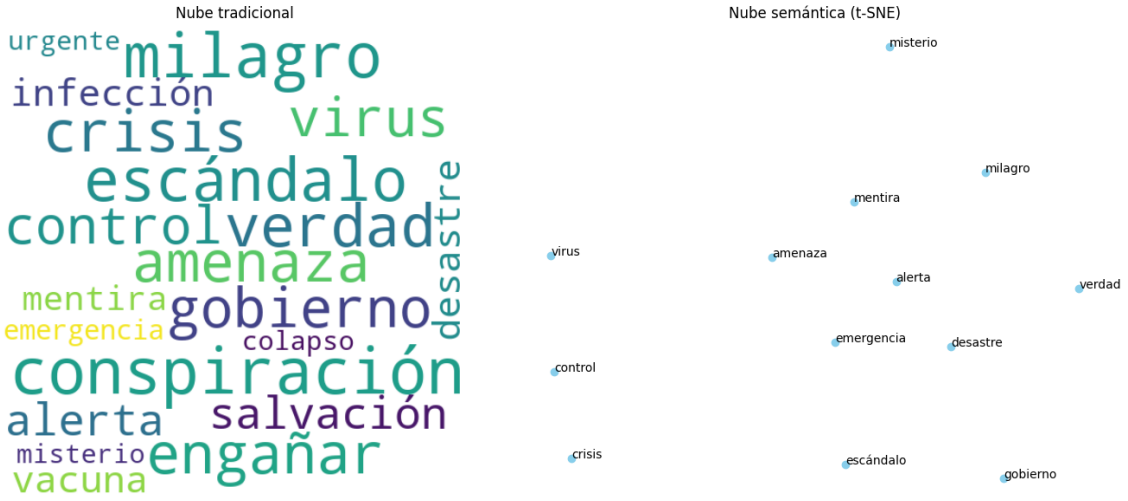

Figura 2.2: Comparación entre nube tradicional (izquierda) y nube semántica con t-SNE (derecha) a partir de noticias falsas.

2.4. Análisis Comparativo entre Noticias Falsas y Verdaderas

El análisis comparativo de las **nubes de palabras** y los **embeddings semánticos** nos permite observar cómo se agrupan las palabras en noticias falsas y verdaderas. En este tipo de análisis, se puede identificar que las noticias falsas tienden a utilizar un vocabulario emocionalmente cargado, mientras que las noticias verdaderas emplean un lenguaje más neutro y verificable. Esta diferencia en el uso del lenguaje es fundamental para diferenciar las noticias falsas de las verdaderas, utilizando herramientas basadas en **NLP** y **embeddings**.

Este enfoque analítico, apoyado en técnicas de visualización como las **nubes semánticas** y los **embeddings** de palabras, ofrece un marco robusto para la detección de patrones semánticos que caracterizan a las noticias falsas, proporcionando una metodología eficiente para su identificación automática.

2.5. Conclusiones

El uso de técnicas de **NLP**, como los **embeddings** de palabras, ha demostrado ser una herramienta poderosa en la detección de noticias falsas. Las representaciones semánticas

obtenidas a través de modelos como **Word2Vec** y **GloVe** permiten identificar patrones lingüísticos comunes en las noticias falsas, lo que facilita su clasificación automática. Las **nubes semánticas**, basadas en estos **embeddings**, ofrecen una representación visual clara de cómo las palabras se agrupan según su significado, lo que proporciona una comprensión más profunda de las características semánticas y emocionales del contenido textual. Estas herramientas, junto con las técnicas de reducción de dimensionalidad como **t-SNE**, son esenciales para la detección eficiente y precisa de desinformación en el ámbito digital.

Capítulo 3

Propuesta Metodológica para el Análisis Exploratorio de Noticias Falsas mediante Nubes de Palabras Semánticas

3.1. Introducción

Este capítulo presenta una metodología estructurada para realizar un análisis exploratorio de noticias falsas mediante el uso de **nubes de palabras semánticas**. El objetivo principal es diseñar un sistema dinámico e interactivo que permita identificar patrones lingüísticos y semánticos característicos de la desinformación, facilitando su detección incluso para usuarios no especializados. A través de la comparación entre noticias verdaderas y falsas, se emplearán técnicas avanzadas de procesamiento de lenguaje natural (NLP) para extraer y representar visualmente relaciones clave entre términos.

3.2. Estructura General del Pipeline

La propuesta metodológica se organiza en cinco fases principales, cada una de las cuales aborda un componente esencial del análisis. Estas fases se estructuran de manera secuencial y modular, permitiendo una implementación flexible y fácilmente escalable:

1. **Recolección y preparación del corpus de noticias**: incluye la obtención de noticias etiquetadas como verdaderas o falsas desde fuentes públicas y su almacenamiento estructurado.

2. **Preprocesamiento y normalización textual**: implica limpiar y transformar los textos (eliminación de stopwords, lematización, tokenización, etc.) para su posterior análisis.

3. **Generación de nubes de palabras semánticas mediante embeddings**: se generan vectores con modelos como Word2Vec o GloVe y se proyectan con t-SNE para producir

agrupamientos visuales significativos.

4. **Implementación de un análisis visual e interactivo**: se presentan visualizaciones que permiten explorar dinámicamente los clústeres semánticos.

5. **Comparación y análisis de patrones entre noticias falsas y verdaderas**: se identifican diferencias y similitudes en los patrones de uso del lenguaje mediante las visualizaciones obtenidas.

La Figura 3.1 muestra de manera esquemática la arquitectura general del pipeline propuesto.

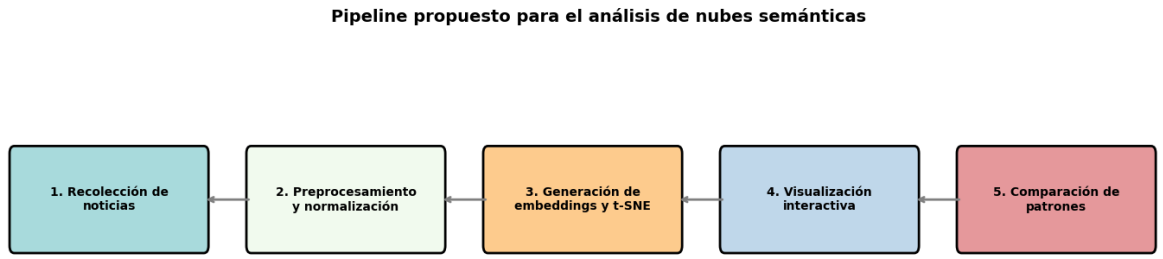

Figura 3.1: Arquitectura del pipeline para la generación y análisis de nubes semánticas.

3.3. Fase 1: Recolección y Preparación del Corpus

El primer paso del pipeline consiste en la recolección y estructuración de un corpus que contenga tanto noticias falsas como verdaderas. Este conjunto de datos será la base para el análisis semántico posterior, ya que permitirá generar las representaciones vectoriales que alimentarán las visualizaciones.

3.3.1. Fuentes y Criterios de Selección

El corpus será construido a partir de las siguientes fuentes reconocidas:

- **LIAR** [21]: un conjunto de datos etiquetado que incluye más de 12,000 declaraciones categorizadas según su veracidad, útil para tareas de clasificación y análisis semántico.
- **FakeNewsNet** [17]: contiene noticias completas junto con metadatos sociales y temporales, lo que permite un análisis más contextualizado del contenido y su propagación.

Los siguientes criterios se aplicarán para la selección y filtrado del corpus:

Equilibrio de clases Se garantizará una distribución balanceada entre noticias falsas y verdaderas para evitar sesgos en el análisis comparativo.

Diversidad temática Se incluirán noticias de distintos ámbitos (política, salud, tecnología, etc.) con el fin de cubrir una variedad representativa de estilos discursivos.

Extensión mínima Se establecerá un umbral mínimo de 100 palabras por noticia, asegurando así la presencia suficiente de contenido semántico para el análisis vectorial.

Idioma Se priorizarán noticias en español, ya sean originales o traducciones controladas, a fin de mantener la coherencia lingüística en el preprocesamiento y representación semántica.

Objetivo de esta fase: Construir un corpus representativo, equilibrado y temáticamente diverso que sirva como base confiable para la comparación entre noticias falsas y verdaderas a través de herramientas semánticas.

3.4. Fase 2: Preprocesamiento del Texto

El preprocesamiento del texto es una etapa fundamental que garantiza que los datos estén en condiciones óptimas para ser utilizados en análisis semántico. Esta fase tiene como objetivo limpiar, normalizar y estructurar los textos para mejorar la calidad de las representaciones vectoriales y facilitar su posterior visualización.

3.4.1. Técnicas de Preprocesamiento

Las principales operaciones aplicadas durante esta fase son:

- **Conversión a minúsculas**: Unifica la representación textual y evita duplicidades entre palabras que difieren solo en capitalización.
- **Eliminación de puntuación, números y caracteres especiales**: Se eliminan elementos no léxicos que no aportan valor semántico al análisis.
- **Tokenización**: Se divide el texto en unidades léxicas (tokens), que generalmente corresponden a palabras. Este paso es clave para el análisis palabra por palabra.
- **Lematización**: Cada palabra se transforma en su forma canónica (lema), lo que permite agrupar variantes morfológicas bajo una sola representación semántica. Por ejemplo, “confirmado” y “confirmar” se reducen a un mismo lema.
- **Eliminación de *stopwords***: Se eliminan palabras funcionales que no aportan información relevante, como artículos, preposiciones o conjunciones (ej. “el”, “la”, “de”, “y”).

3.4.2. Implementación del Preprocesamiento

Para la implementación de estas tareas se utilizarán herramientas especializadas:

- **SpaCy**: Se empleará para la tokenización y lematización eficiente en idioma español. Esta herramienta permite un análisis morfosintáctico preciso y de alto rendimiento.
- **NLTK**: Complementará el preprocesamiento con la eliminación de *stopwords* y otras operaciones básicas. También se usará para crear listas personalizadas de palabras irrelevantes, según el contexto del corpus.

Objetivo de esta fase: Obtener un corpus textual limpio, coherente y normalizado que optimice la generación de embeddings y permita una visualización semántica precisa y significativa en las fases siguientes.

3.5. Fase 3: Generación de Nubes de Palabras Semánticas

En esta fase se generarán **nubes de palabras semánticas** a partir del análisis de relaciones entre palabras representadas mediante **embeddings**. A diferencia de las nubes tradicionales, estas representaciones no se basan únicamente en la frecuencia, sino en la similitud semántica entre términos, lo que permite identificar clústeres de palabras con significados o contextos similares.

3.5.1. Técnicas de Embeddings

Para capturar las relaciones semánticas presentes en el corpus, se utilizarán las siguientes técnicas de representación vectorial:

- **Word2Vec**: Se entrenará un modelo *Word2Vec* sobre el corpus recolectado. Esta técnica convierte las palabras en vectores de alta dimensión mediante el análisis de sus contextos en el texto, agrupando términos como “crisis” y “desastre” debido a su cercanía semántica.
- **Global Vectors for Word Representation (GloVe)**: También se evaluará el uso de embeddings preentrenados de GloVe, que se basan en estadísticas de co-ocurrencia de palabras dentro del corpus. Este enfoque es especialmente eficaz para capturar relaciones globales de contexto.

3.5.2. Reducción de Dimensionalidad

Los vectores generados tendrán típicamente entre 100 y 300 dimensiones, por lo que se aplicarán técnicas de reducción de dimensionalidad para permitir su visualización:

- **t-SNE** (t-Distributed Stochastic Neighbor Embedding): Se utilizará para proyectar los vectores en un espacio bidimensional, preservando las relaciones locales entre palabras.
- **UMAP** (Uniform Manifold Approximation and Projection): Se considerará como alternativa a t-SNE, ya que ofrece mayor velocidad de procesamiento y mejor conservación de la estructura global del espacio semántico.

3.5.3. Implementación de la Nube Semántica

Las nubes de palabras semánticas se construirán a partir de los vectores proyectados en dos dimensiones. En estas visualizaciones:

- Palabras cercanas en el plano corresponderán a términos semánticamente relacionados.
- Se utilizarán colores, tamaños y agrupaciones visuales para destacar clústeres temáticos o emocionales.
- Se compararán las nubes generadas a partir de noticias falsas y verdaderas, permitiendo analizar cómo varía el lenguaje entre ambas.

Para enriquecer la interacción, se explorará el uso de librerías como **D3.js** y **Plotly**, que permiten generar visualizaciones dinámicas. Esto permitirá a los usuarios hacer clic sobre las palabras para obtener información adicional, como ejemplos de uso en el corpus o su frecuencia relativa.

Objetivo de esta fase: Generar visualizaciones semánticas interactivas que representen de manera clara y estructurada las relaciones entre palabras, facilitando la identificación de patrones discursivos y emocionales distintivos entre noticias falsas y verdaderas.

3.6. Fase 4: Implementación de un Análisis Visual e Interactivo

En esta fase se desarrollará un entorno de visualización interactivo que permitirá a los usuarios explorar las **nubes semánticas** de forma dinámica. El objetivo es facilitar la comprensión de las relaciones lingüísticas presentes en el corpus y permitir la navegación entre términos y sus contextos reales.

3.6.1. Interactividad y Exploración

La interactividad es clave para que usuarios no expertos puedan interpretar fácilmente los patrones semánticos del lenguaje utilizado en las noticias. Para ello, se utilizarán bibliotecas como **D3.js** o **Plotly**, que permiten:

- Realizar **clic** sobre palabras para acceder a ejemplos de noticias donde aparecen.
- **Resaltar clústeres temáticos o emocionales** con colores y herramientas de selección.
- Implementar **filtros dinámicos** por fecha, categoría temática o nivel de veracidad (falsa/verdadera).

Estas funciones mejoran la accesibilidad del sistema y ofrecen un medio intuitivo para explorar patrones de desinformación.

3.6.2. Actualización de la Nube Semántica: Enfoques Comparativos

La incorporación continua de nuevas noticias al corpus plantea el desafío de mantener actualizadas las visualizaciones. Se consideran dos enfoques:

- **Nube Incremental**: Las nuevas palabras se integran al modelo existente sin reentrenar o reestructurar la nube completa. Este enfoque es más eficiente en términos computacionales, pero puede producir una representación desbalanceada si los nuevos términos alteran significativamente el espacio semántico.
- **Nube Recalculada**: Cada vez que se incorpora nueva información, se vuelve a generar la nube completa, incluyendo el reentrenamiento (o revectorización) y reproyección. Si bien este enfoque es más costoso, garantiza una representación más precisa y coherente de los patrones lingüísticos emergentes.

Recomendación: Para contextos donde la precisión y actualización semántica son prioritarias (por ejemplo, monitoreo de desinformación en tiempo real), se sugiere optar por la recalculación completa. Sin embargo, para aplicaciones con recursos limitados, el enfoque incremental puede resultar adecuado si se implementan mecanismos de corrección periódica.

Objetivo de esta fase: Implementar un sistema visual e interactivo que permita a los usuarios navegar por el espacio semántico generado, accediendo de manera intuitiva a los contextos reales de uso de las palabras y facilitando así el análisis comparativo entre tipos de noticias.

3.7. Fase 5: Comparación y Análisis de Patrones entre Noticias Falsas y Verdaderas

Esta fase se enfoca en el análisis comparativo entre las nubes semánticas generadas para noticias falsas y verdaderas, con el fin de identificar diferencias significativas en el uso del lenguaje. La visualización de estas diferencias permite observar patrones discursivos que podrían indicar la presencia de desinformación.

3.7.1. Indicadores Clave de Comparación

Para estructurar el análisis, se proponen los siguientes indicadores:

- **Presencia de términos emocionales o alarmistas**: Evaluar la frecuencia y centralidad de palabras con alta carga emocional en noticias falsas, como "urgente", "alerta", "desastre" o "escándalo".
- **Agrupaciones temáticas dominantes**: Identificar si los clústeres semánticos en noticias falsas se concentran en temas como conspiraciones, salud pública o política polarizada.
- **Diversidad léxica**: Comparar la variedad de términos utilizados en ambos tipos de noticias para determinar si las falsas presentan un vocabulario más repetitivo o limitado.
- **Densidad de clústeres semánticos**: Analizar si las agrupaciones de palabras en las noticias falsas tienden a ser más densas y homogéneas, lo que podría indicar una narrativa centrada en emociones específicas o ideas recurrentes.

Objetivo de esta fase: Obtener una representación visual y cuantitativa de las diferencias lingüísticas entre noticias falsas y verdaderas. Este análisis permite identificar patrones discursivos propios de la desinformación, lo cual puede servir como base para desarrollar futuros modelos automáticos de detección más explicables y transparentes.

3.8. Ventajas y Limitaciones

3.8.1. Ventajas

- **Accesibilidad**: Las visualizaciones interactivas permiten a usuarios no especializados explorar el contenido de manera intuitiva, facilitando la comprensión de patrones lingüísticos complejos.
- **Exploración dinámica del corpus**: La posibilidad de interactuar con las palabras dentro de la nube semántica, accediendo a las noticias relacionadas, enriquece el análisis al conectar directamente las representaciones visuales con el contenido original.
- **Implementación eficiente y modular**: El uso de herramientas como **D3.js** o **Plotly** permite desarrollar visualizaciones interactivas de forma ágil, adaptable y reutilizable en diferentes contextos o conjuntos de datos.

3.8.2. Limitaciones

- **Enfoque exploratorio**: La metodología propuesta no reemplaza modelos predictivos más sofisticados, ya que se orienta principalmente al análisis visual e interpretativo.

- **Dependencia del corpus**: La representatividad y diversidad del corpus influye directamente en la calidad de las visualizaciones y de los patrones detectados.
- **Subjetividad en la interpretación**: Dado que se trata de análisis visuales, los resultados pueden estar influenciados por la interpretación del analista, especialmente en la identificación de clústeres semánticos o patrones emocionales.

3.9. Resumen

La metodología propuesta integra técnicas de procesamiento de lenguaje natural, generación de embeddings y visualización interactiva para llevar a cabo un análisis exploratorio de noticias falsas. Mediante el uso de **nubes de palabras semánticas**, es posible representar de forma clara las relaciones entre términos y explorar cómo varía el uso del lenguaje entre noticias falsas y verdaderas.

El carácter interactivo del sistema permite una exploración profunda del corpus, y la posibilidad de recalcular la nube semántica garantiza una representación precisa y actualizada ante nuevos datos. Asimismo, la comparación de patrones semánticos entre tipos de noticias proporciona *insights* significativos sobre las estructuras discursivas asociadas a la desinformación, sentando las bases para desarrollos futuros en modelos de detección más robustos y explicables.

Bibliografía

[1] Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211

[2] Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1). https://doi.org/10.1002/spy2.100

[3] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

[4] Feldman, R. (2013). Techniques and Applications for Sentiment Analysis. *Communications of the ACM*, 56(4). https://doi.org/10.1145/2436256.2436274

[5] Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social Clicks: What and Who Gets Read on Twitter?. In *Proceedings of the 2016 ACM SIGMETRICS International Conference* (pp. 179–192). https://doi.org/10.1145/2896377.2901462

[6] Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297.

[7] Heylen, K., Tummers, J., Peeters, D., & Geeraerts, D. (2015). Visualising lexical variation with word clouds: The case of *weinig* and *een beetje* in Dutch. *Digital Scholarship in the Humanities*, 30(1).

[8] Karimi, H., & Tang, J. (2019). Learning Hierarchical Discourse-level Structure for Fake News Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3432–3442.

[9] Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50, 723–762.

[10] Lazer, D. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The Science of Fake News. *Science*, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781.*

[12] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, 26.

[13] Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579–2605.

[14] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

[15] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

[16] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937.

[17] Shu, K., Mahudeswaran, D., & Liu, H. (2018). FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information. *arXiv preprint arXiv:1809.01286*.

[18] Shu, K., Wang, S., & Liu, H. (2020). Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 3476–3482.

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems*, 5998–6008.

[20] Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread of True and False News Online. *Science*, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[21] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 422–426.

[22] Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys (CSUR)*, 53(5), 1–40.

[23] Shah, F., Asghar, M. Z., & Ahmad, N. (2018). Opinion mining and sentiment analysis of social media content: A systematic review. *International Journal of Computer Applications*, 179(5), 1–15. https://doi.org/10.5120/ijca2018917209