Recibido: 29 de diciembre de 2020 Revisado: 5 de febrero de 2021 Aceptado: 8 de febrero de 2021

[redacted]idad de la información IET

DOI: 10.1049 / ise2.12021

The Institution of Engineering and Technology WILEY

O PA PER DE BÚSQUEDA DE IGI NA L RE

Un modelo híbrido para la detección de noticias falsas: aprovechar el contenido de las noticias y los comentarios de los usuarios en las noticias falsas

Marwan Albahar

Departamento de Ciencias de la Computación, Universidad Umm Al Qura, La Meca, Arabia Saudita

Correspondencia

Marwan Albahar, Departamento de Ciencias de la Computación, Universidad Umm Al Qura, La Meca, Arabia Saudita. Correo electrónico:mabahar@uqu.edu.sa

Abstracto

Hoy en día, las plataformas de redes sociales como Twitter se han convertido en un medio popular para que las personas difundan y consuman noticias debido a su fácil acceso y la rápida proliferación de noticias. Sin embargo, la credibilidad de las noticias publicadas en estas plataformas se ha convertido en un problema importante. En otras palabras, las noticias escritas que contienen información inexacta con el objetivo de engañar a los lectores se han difundido rápidamente en estas plataformas. En la literatura, esta noticia se llama noticias falsas. Detectar tales noticias en las plataformas de redes sociales se ha convertido en una tarea desafiante. Uno de los principales desafíos es identificar información útil que se explota como una forma de detectar noticias falsas. Se incorpora un modelo híbrido que comprende una red neuronal recurrente (RNN) y una máquina de vectores de soporte (SVM) para detectar noticias reales y falsas. Se utilizó un RNN con unidades recurrentes con compuertas bidireccionales para codificar datos textuales, incluido el contenido de noticias y comentarios, en vectores de características numéricas. Las características codificadas se enviaron a un SVM con kernel de función de base radial para clasificar la entrada dada de noticias reales y falsas. Los experimentos con el conjunto de datos del mundo real arrojan resultados alentadores y demuestran que el marco propuesto supera a los métodos más avanzados.

1 | INTRODUCCIÓN

Los rápidos avances en los teléfonos móviles y el uso generalizado de Internet han reformulado las interacciones sociales. Debido a sus características particulares, fácil acceso a la información, bajo costo de generar noticias y rápida proliferación de noticias, las redes sociales se han convertido en una plataforma atractiva para que las personas difundan y consuman noticias. Además, se ha demostrado que las redes sociales capturan información relacionada con un evento en curso como COVID-19, así como los intereses y opiniones de las personas. Así, más personas prefieren seguir el desarrollo de una historia o un evento en las redes sociales que en los medios tradicionales como la televisión o los formatos tradicionales de noticias. Sin embargo, las noticias en las plataformas de redes sociales de alguna manera carecen de credibilidad y confiabilidad en comparación con las noticias en los medios tradicionales. En otras palabras,

Por el contrario, las fuentes de la mayoría de las noticias publicadas en las plataformas de redes sociales pueden ser difíciles de verificar, lo que facilita la manipulación del contenido para lograr varios objetivos [1]. En consecuencia, dicha información, difundida rápida y ampliamente, puede ser utilizado para propagar artículos de noticias inexactos [2]. Con la ausencia de una verificación rigurosa, los usuarios bien intencionados pueden contribuir involuntariamente a la rápida difusión de noticias falsas. La rápida difusión de noticias falsas puede afectar negativamente a la sociedad y a las personas. Su daño puede extenderse a empresas y gobiernos. Por ejemplo, las noticias falsas sobre una organización podrían ser difundidas por usuarios malintencionados o el spam podría causar un daño significativo a la imagen de la organización en la sociedad. De ahí que la detección de fake news se haya planteado como un área de investigación importante. El surgimiento de las redes sociales implica que las noticias reales y falsas se presentan de manera similar y, a veces, es difícil de distinguir. Primero, el contenido de las noticias falsas se combina con datos reales y falsos para atraer lectores. En segundo lugar, se difunde información amplia y variada en las redes sociales; por ejemplo, muchos usuarios anónimos comunican información ruidosa.3,4] utilizando numerosas funciones de noticias de las redes sociales, como información de texto, funciones de usuario y comentarios de los usuarios. Sin embargo, el aprendizaje del contexto no ha sido diseñado para noticias falsas. Específicamente, no es posible obtener predicciones precisas de noticias falsas basándose únicamente en el contenido textual [5,6]

porque el contenido de las redes sociales suele ser breve y fragmentado. En segundo lugar, otro estudio sobre detección de noticias falsas se centró en analizar las primeras características del difusor de noticias, ignorando así las opiniones de los comentarios de los usuarios posteriores [7,8]. Es difícil distinguir las noticias falsas de una noticia masiva. Esta pregunta lleva a uno a considerar diferentes enfoques para determinar la validez de una noticia. En los últimos años, el proceso de identificación de noticias falsas ha seguido creciendo lentamente en un camino de investigación en evolución. Definimos dos características clave relacionadas con la identificación de noticias falsas, a saber, las respuestas de los usuarios y el contenido de la información.

El contenido de la información distribuida es principalmente lo que debe clasificarse como verdadero o falso. Al analizar las características textuales de los artículos de diferentes páginas de noticias falsas, Horne y Adali [9] identificó características únicas de noticias falsas que difieren del contenido de noticias verdaderas y luego las comparó con artículos de sitios web periodísticos creíbles. Sus resultados indican que las noticias falsas son más largas, con más frases en mayúsculas y menos palabras vacías. Perez - Rosas y col. [10] encontró que los artículos de noticias falsas tenían más términos sociales, palabras verbales y palabras temporales, lo que implica que el texto tendía tanto al presente como al futuro y no a volverse más fáctico y objetivo. Newman y col. [11] encontró que las historias engañosas tenían menor sofisticación semántica, menos frases, palabras más pesimistas sobre los sentimientos y más palabras sobre el movimiento. Silverman [12] señala que el 13% de 1600 artículos de noticias tenían titulares y contenido incoherentes, utilizando titulares declarativos junto con cuerpos de artículos que eran escépticos sobre la afirmación.

Respuestas de usuario en las redes sociales cuentan con datos auxiliares bastante útiles en el análisis de noticias falsas. Se considera que tienen señales de identificación más fuertes que el contenido de la información, principalmente porque las reacciones de los usuarios y los patrones de difusión son más difíciles de explotar que el contenido de la información y contienen datos de veracidad obvia [13]. Este conocimiento secundario en forma de interacciones del usuario (me gusta, comentarios, respuestas o acciones) incluye información rica recopilada en la estructura de propagación (árbol) que muestra la dirección del flujo de información, los detalles de la marca de tiempo de las interacciones, la información textual de las interacciones del usuario y el usuario. información de perfil de los usuarios que participan en interacciones. Zubiaga y col. [14] declaró que a través de su enfoque, los comentarios de los usuarios se pueden diferenciar. Se incluyen cuatro formas en la categorización más utilizada: ayuda, rechazo, pregunta y comentario (que puede ser neutral o no relacionado). También señalaron que, según las fases de difusión, la esencia de las respuestas de los usuarios difiere. En el caso de los rumores, cuando se considera la vida útil completa de un rumor, la mayoría de los usuarios aceptan los rumores reales y un porcentaje más alto rechaza los rumores falsos. Qian y col. [15] encontró que las noticias falsas parecen obtener más respuestas negativas y cuestionamientos que las noticias reales. Se descubrió que los usuarios parecen respaldar los rumores independientemente de su credibilidad al analizar solo las primeras reacciones a los rumores. Esto es evidente porque los usuarios tienen problemas para evaluar la integridad en las primeras etapas. Sin embargo, en la práctica, la mayoría de estos métodos requieren un análisis de las declaraciones de integridad por parte de profesionales capacitados, lo que dificulta la automatización y la falta.

generalizabilidad sobre diferentes temas y dominios. A continuación, enumeramos el tipo de métodos existentes y sus limitaciones en la Tabla1 . Para hacer frente a las limitaciones mencionadas anteriormente de los métodos existentes, abordamos el problema de la detección de noticias falsas en las redes sociales mediante el desarrollo de un modelo híbrido. El modelo incorpora tanto el contenido de las noticias como la información potencial en los comentarios y consta de dos fases. El proceso de incrustación se llevó a cabo utilizando unidades recurrentes con compuerta bidireccional (GRU) y una máquina de vectores de soporte (SVM) con un kernel gaussiano para la clasificación.

2 | OBRAS RELACIONADAS

Si bien el problema de la identificación de noticias falsas es relativamente reciente, ha atraído una atención considerable. Se han propuesto numerosos métodos para detectar noticias falsas dentro de varios conjuntos de datos. Esta sección proporciona un resumen de la literatura actual y relevante sobre la identificación de noticias falsas. Para identificar las noticias falsas, actualmente existen tres enfoques: basado en la propagación, análisis de fuentes e investigación basada en el contenido. El enfoque basado en la propagación postula que los patrones de distribución de las noticias falsas son diferentes de los de las noticias confiables. Según el mapa de propagación, estos patrones de distribución incluirán noticias como falsas o verdaderas [17]. El segundo enfoque evalúa el contexto y los patrones de la noticia, lo que permite la detección temprana de noticias falsas [18]. El enfoque de la investigación basada en el contenido se basa en la detección de características lingüísticas tanto léxicas como sintácticas. Supone que los artículos falsos se formulan utilizando un lenguaje sintáctico y engañoso [10,19,20,21,22]. En [23] para la detección de la postura. El título y el cuerpo de cada artículo están codificados con vectores de pensamiento de omisión. Las características hechas a mano incluyen gramos, caracteres y puntos TF-IDF ponderados entre cada artículo y el cuerpo del título. Los autores en [24] propuso el uso de la atención neuronal con redes neuronales recurrentes bidireccionales (RNN) para codificar un artículo de noticias completo, las dos primeras oraciones del artículo y el título del artículo. Estas representaciones luego se combinan con características hechas a mano, como se usa en [23]. Rubin y col. [25] dividió el tema de la identificación de noticias falsas en tres categorías. Estas categorías son engaños a gran escala, noticias falsas humorísticas y fabricación severa. En [26], Conroy et al. propuso un enfoque híbrido para detectar noticias falsas. Su enfoque híbrido incorpora enfoques de análisis de redes y señales lingüísticas. Para la verificación de la noticia, se utilizó un modelo de espacio vectorial en [27]. Los autores utilizaron el conocimiento en línea para detectar información engañosa. Dadgar y col. [28] dividió las noticias con TF - IDF y SVM en categorías separadas. En [20], los autores hicieron referencia a señales satíricas para detectar noticias falsas o engañosas. El modelo basado en SVM evaluó un conjunto de datos que contenía 360 artículos de noticias. En [29], Jin y col. estableció puntos de vista opuestos en las redes sociales para verificar las noticias y validar su concepto en un conjunto de datos del mundo real. Ha habido una discusión detallada de los métodos de validación, algoritmos de minería de datos y conjuntos de datos para la identificación de noticias falsas [30]. Ahmed y col. clasificación empleada

TABLA 1 Resumen de métodos y limitaciones para la detección de noticias falsas [dieciséis]

Método	Limitaciones
Basado en contenido	Se requieren profesionales capacitados para analizar las declaraciones de integridad, lo que dificulta la automatización.
	La falta de generalización entre idiomas, temas y dominios no explota y extrae por completo la información sintáctica del contenido. A menudo es un desafío
	determinar la veracidad solo mediante el análisis del texto y, por lo tanto, se necesita información adicional o verificación de hechos. Los parámetros del
Basado en propagación	modelo no son específicos del usuario y se asumen las mismas constantes de velocidad y probabilidades para todos los usuarios.
Basado en fuente	El monitoreo de la red requiere que la estrategia se actualice en función de las topologías de red cambiantes. Por lo tanto, la supervisión de la red en
	Las redes a gran escala pueden resultar costosas debido a la complejidad de la red.

enfoques y análisis de n-gramas para identificar noticias falsas y spam de opinión [31]. En [32], Gilda usó árbol de decisión acotado, SVM, bosque aleatorio, aumento de gradiente y algoritmos de gradiente estocástico. Los mejores resultados del autor se obtuvieron con el método de descenso de gradiente estocástico. Ruchansky y col. [33] utilizó un algoritmo híbrido llamado capturar, puntuar e integrar (CSI) para la detección de noticias falsas, donde se combinaron tres características para una predicción más precisa. Shu y col. [34] propuso un modelo de detección de noticias falsas que considera la asociación de interacciones del usuario relacionadas, el sesgo del editor y la postura de las noticias.

Además, se utilizaron conjuntos de datos de detección de noticias falsas del mundo real para verificar la eficiencia del modelo. Long y col. [35] utilizó un algoritmo híbrido novedoso centrado en redes de memoria a corto plazo a largo plazo (LSTM) basadas en la atención para problemas de detección de noticias falsas. El método se comparó con otros conjuntos de datos de detección de noticias falsas. Figueira y Oliveira [36] revisó el estado actual de las noticias falsas, sugiriendo un enfoque opuesto de las noticias falsas para aumentar la conciencia empresarial para detectar automáticamente las noticias falsas. Un nuevo algoritmo automatizado [10] fue presentado por Perez-Rosas et al. Los autores mostraron un modelo de clasificación basado en la combinación de detalles léxicos, sintácticos y semánticos. Buntain y Golbeck [37] introdujo un método automatizado para detectar noticias falsas. Utilizaron este enfoque en tres conjuntos de datos de acceso público. Bessi [38] utilizó las redes sociales en línea para examinar las propiedades estadísticas de las noticias falsas, la especulación y las afirmaciones no probadas. Zu y col. [39] han presentado un enfoque competitivo para reducir el efecto de la información incorrecta que se centra en la correlación entre la información original incorrecta y la información actualizada. Shu y col. [40] consideró la confianza del usuario para crear un nuevo algoritmo para la detección de noticias falsas. Se han desarrollado varios otros modelos para la detección de noticias falsas [32,33,41,42,43]. En [42], el autor adoptó una perspectiva diferente y abordó el tema como un problema de procesamiento del lenguaje natural (PNL). El autor utilizó el aprendizaje profundo basado en PNL para la detección de noticias falsas. Además, el autor propuso un nuevo diseño que incorporó mecanismos 'parecidos a la atención' con una red de convolución.

Además, el autor documentó los resultados de comparar diferentes redes neuronales (como recurrente, LSTM, GRU y convolucional con atención aumentada). En consecuencia, la arquitectura RNN con GRU superó a LSTM. El autor también evaluó el conjunto de datos en otros clasificadores, como soporte vectorial, disminución del gradiente estocástico, etc. En [44], los autores utilizaron un detector automatizado con métodos de aprendizaje profundo a través de una red jerárquica de tres niveles de atención (3HAN). 3HAN se ha utilizado para construir un vector de noticias con tres niveles de atención correspondientes a la entrada de las oraciones, palabras y titulares de un artículo de noticias siguiendo una forma jerárquica de abajo hacia arriba. Una característica distintiva de un artículo de noticias falso es su título, por lo que relativamente pocas frases y palabras en un artículo son más importantes que el resto. Debido a sus tres niveles de atención, 3HAN otorga una importancia diferencial a las partes de un artículo. Los autores observaron la eficacia de 3HAN con una precisión del 96,77% a través de experimentos en un gran conjunto de datos del mundo real.

3 | CONJUNTO DE DATOS

Usamos un conjunto de datos públicos, FakeNewsNet [13], que se recopila especialmente para la detección de noticias falsas. FakeNewsNet contiene noticias etiquetadas de dos sitios web,politifact.com y gossipcop. com. El contenido incluye información lingüística y visual, todos los tweets y retweets de cada noticia y la información de usuario de Twitter correspondiente. Además, el conjunto de datos contiene declaraciones anotadas e incorpora información espacio-temporal e información sobre el contexto social. Además, los autores introdujeron una línea de actualización continua de datos para noticias falsas actuales. Las estadísticas detalladas del repositorio de FakeNewsNet se muestran en la Figura1 [13].

3.1 | Preprocesamiento

Para hacer que el conjunto de datos sea _t para el modelo, las oraciones se tokenizaron y se eliminaron los signos de puntuación y las palabras vacías. Las palabras vacías son palabras menos importantes y no definen ningún contexto. Por ejemplo, considere la siguiente línea:

El FBI también investigó a grupos liberales que tenían progresista en sus nombres. . . el FBI básicamente estaba mirando a todo el mundo.

Después de tokenizar y eliminar la puntuación y las palabras vacías, obtenemos lo siguiente:

	Category	Features	PolitiFact		GossipCop	
			Fake	Real	Fake	Real
News Content	*Linguistic*	# News articles	432	624	5,323	16,817
		# News articles with text	420	528	4,947	16,694
	Visual	# News articles with images	336	447	1,650	16,767
Social Context	*User*	# Users posting tweets	95,553	249,887	265,155	80,137
		# Users involved in likes	113,473	401,363	348,852	145,078
		# Users involved in retweets	106,195	346,459	239,483	118,894
		# Users involved in replies	40,585	18,6675	106,325	50,799
	Post	# Tweets posting news	164,892	399,237	519,581	876,967
	Response	# Tweets with replies	11,975	41,852	39,717	11,912
		# Tweets with likes	31692	93,839	96,906	41,889
		# Tweets with retweets	23,489	67,035	56,552	24,955
	Network	# Followers	405,509,460	1,012,218,640	630,231,413	293,001,487
		# Followees	449,463,557	1,071,492,603	619,207,586	308,428,225
		Average # followers	1299.98	982.67	1020.99	933.64
		Average # followees	1440.89	1040.21	1003.14	982.80
Spatiotemporal Information	*Spatial*	# User profiles with locations	217,379	719,331	429,547	220,264
		# Tweets with locations	3,337	12,692	12,286	2,451
	Temporal	# Timestamps for news pieces	296	167	3,558	9,119
		# Timestamps for response	171,301	669,641	381,600	200,531

FIGURA 1 Estadísticas del repositorio de FakeNewsNet

< FBI>, <investigado>, <liberal>, <grupo>, <progresista>, <nombres>, <FBI>, <básicamente>, <buscando >, <todo el mundo>

4 | CONTENIDO DE NOTICIAS Y COMENTARIOS INCORPORADOS

En las siguientes secciones, discutimos el proceso de incrustar las palabras, oraciones y comentarios por separado. Las oraciones y los comentarios incrustados se concatenan antes de enviarlos a la SVM.

4.1 | Incrustación de palabras

Para codificar una palabra, el codificador de palabras basado en RNN se utiliza para aprender la representación de oraciones. En un RNN, cuando la secuencia se vuelve más larga, la memoria antigua se desvanece. Por lo tanto, para determinar las dependencias de RNN a largo plazo y garantizar la memoria persistente, se adoptaron GRU en el RNN. Además, se incorporó el GRU bidireccional [traducción automática neuronal mediante el aprendizaje conjunto de alinear y traducir] para capturar la información contextual de las anotaciones. Hay dos GRU en un GRU bidireccional; uno, Gf, es adelante, y el otro, GB, está al revés. Hacia adelante GRU Gf lee el Ila oración de la palabra X_I $_1$ para $X_{I,}$ mientras que GRU al revés $GRAMO_B$ lee el Ila frase de la palabra $X_{I,}$ para $X_{I.}$

$$O_F = GRU\ x_{I\ t};\ t \in F1;\ldots;\ metro_I gramo$$

$$O_B = GRU\ x_{I\ t};\ t \in F metro_I;\ldots;\ 1gramo$$

La anotación de la palabra $X_{I\,t}$ se obtuvo concatenando estados de salida hacia adelante y hacia atrás $O_{I\ t} = <O_F; O_B>$. La notación $O_{I\,t}$ incluye toda la información del conjunto oración basada en cada palabra $X_{I,t}$

4.2 | Incrustación de oraciones

El mismo enfoque utilizado para la codificación de palabras se utiliza para codificar oraciones. El RNN con unidades GRU se utiliza para codificar cada frase de noticias. El vector de palabra anotada O_k se usa para aprender la representación de la oración S_k mediante el uso las unidades GRU bidireccionales. Se puede mostrar matemáticamente de la siguiente manera:

$$S_F^k = GRU\ O_t^{\ k};\ k \in F1;\ 2\ldots;\ norte gramo$$

$$S_B^k = GRU\ O_{k\ t};\ k \in F NORTE;\ldots;\ 2;\ 1gramo$$

Tanto las anotaciones hacia adelante como hacia atrás fueron concatenadas. Nated para obtener la anotación de la oración. $S_k = <S_{k\ F_0}, S_B>$ los notación S_k aprende la información contextual de las oraciones que se encuentran en la localidad de la oración k.

4.3 | Incorporación de comentarios de usuario

Los comentarios de los usuarios son importantes en este escenario, ya que las personas expresan sus opiniones sobre diferentes tipos de noticias publicando comentarios, reacciones y compartiendo opiniones escépticas. Por lo tanto, los comentarios contienen potencialmente información semántica útil para discriminar entre noticias reales y falsas. Para codificar dichos comentarios se utilizó el RNN con GRU bidireccional y es el mismo método aplicado para las noticias en el apartado anterior.

5 | ANÁLISIS TEORICO

Supongamos que nos dan un artículo de noticias. $norte$ eso contiene S frases F_{Si} gramo$_{i=1}$. Cada oración$_{SI}$ consiste en Ki palabras F_{wij} gramo$_{j=1}$. Dejar C ser el conjunto de T comentarios sobre noticias $norte$ y cada comentario consta de Qj palabras. Tratamos el

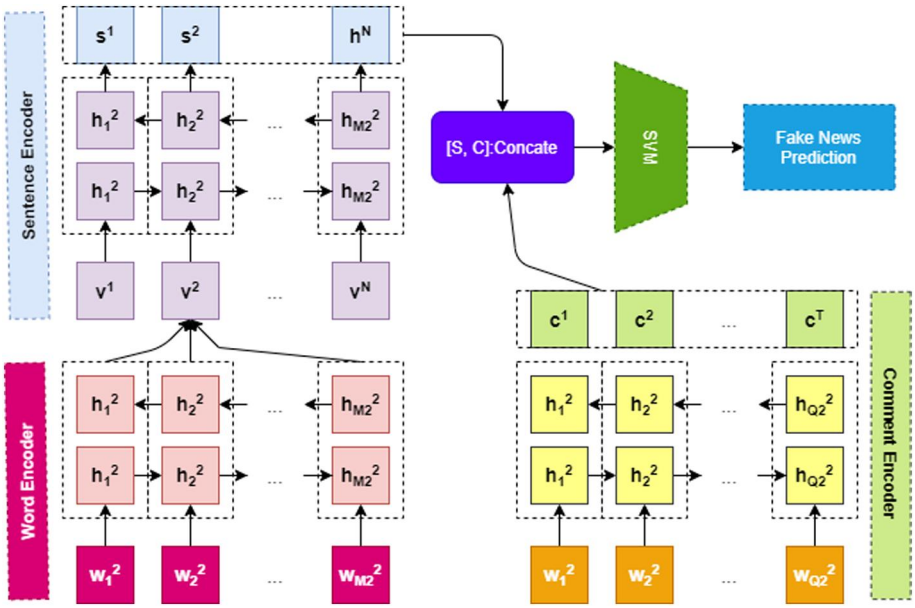

FIGURA 2 Modelo propuesto basado en red neuronal recurrente con unidades recurrentes cerradas bidireccionales y máquina de vectores de soporte. Los comentarios son incrustado utilizando el mismo método utilizado para la incrustación de palabras

Problema de detección de noticias falsas como clasificación binaria. Por lo tanto, cada noticia puede ser cierta.*y* = 1 o falso *y* = 0. Además, aplicamos el proceso de incorporación que se explicó en la Sección 4. Después de la incrustación, el contenido de las noticias y los comentarios se concatenaron para cada artículo de noticias correspondiente y se proporcionaron como entrada a la SVM kernelised. La salida de los GRU se limitó a 512 unidades y, después de la concatenación de los comentarios, el tamaño se amplió a 1024 unidades. Por lo tanto, el tamaño del vector de características está limitado a 1024 unidades. El SVM con un kernel gaussiano toma la entrada y la transforma en un espacio de alta dimensión donde la entrada se vuelve más separable. Las noticias se pueden clasificar fácilmente en las clases correspondientes con alta precisión. Los comentarios agregan más información semántica que tiene el potencial de ayudar en la detección de noticias falsas. Esto se debe a que las personas expresan sus emociones u opiniones hacia las noticias falsas a través de publicaciones en las redes sociales en forma de opiniones dudosas, haciendo preguntas relacionadas, expresando sus propias explicaciones y reaccionando sensacionalmente. Por lo tanto, estos rasgos agregan una dimensión adicional a la tarea de detección. Como se muestra en la figura2, los comentarios codificados se combinan con las oraciones del contenido de las noticias para crear una longitud vectorial de 1024 unidades, lo que mejora aún más la detección de noticias falsas. El método de incrustación es el mismo que para la incrustación de oraciones, pero lógicamente, el proceso acumula la información complementaria que está potencialmente presente en los comentarios. El tamaño del vector de características se establece en 1024 unidades, ya que este vector se convierte en la entrada de la SVM para su posterior procesamiento. El rendimiento de la SVM en este tamaño es más preciso que para 512 o 128 unidades. Además, la característica de entrada se transforma nuevamente a una dimensión superior; por lo tanto, aumentar el tamaño a más de 1024 unidades aumenta la sobrecarga de procesamiento y, por lo tanto, tiene un efecto insignificante en la precisión.

6 | MODELO

En este estudio, SVM y RNN con GRU bidireccionales están incorporados para hacer un modelo híbrido para detectar nuestro espectro que no da un dulce. Tanto SVM como RNN son modelos secuenciales, y la combinación de ambos superó los algoritmos existentes. En el primer paso, los tweets etiquetados se alimentan al modelo RNN con unidades GR bidireccionales y la inserción de palabras y frases. En la segunda fase, las palabras codificadas, las oraciones y las características del documento extraído se alimentan a las funciones intermedias y los resultados del modelo.

6.1 | Arquitectura

En nuestro enfoque, RNN y SVM se introducen para detectar noticias reales y falsas. Se utilizó RNN con GRU bidireccionales para incrustar el contenido y arroja resultados en codificación one-hot como una matriz 2D. Esta salida se entregó como entrada al SVM para entrenarlo a detectar noticias reales y falsas en las entradas. La SVM está incorporada con un kernel gaussiano, como se muestra en la Ecuación (3), para mejorar el poder de detección de la SVM:

$$KDX; X\text{-}Þ \text{¼} \exp - \frac{X - X^{2}}{2\sigma 2}$$

Para determinar el tamaño del vector de características y el número máximo de palabras, elegimos las oraciones más largas posibles del contenido. Aquellas oraciones que tienen una longitud igual o menor que el número máximo de palabras se rellenaron para hacer

TABLA 2 Comparación con el estado del métodos de arte

Conjuntos de datos	La medida	Nuestro modelo	HAN	TCNN – URG	HPA – BLSTM	CSI
PolitiFact	Precisión	**0,912**	0,837	0,712	0,846	0,827
	Precisión	**0,910**	0,824	0,711	0,894	0,847
	Recordar	**0,961**	0,896	0,941	0,868	0,897
	Puntuación F1	**0,932**	0,860	0,810	0,881	0,871
GossipCop	Precisión	**0,802**	0,742	0,736	0,753	0,772
	Precisión	0,730	0,655	0,715	0,684	**0,732**
	Recordar	**0,790**	0,689	0.521	0,662	0,638
	Puntuación F1	**0,762**	0,672	0,603	0,673	0,682

Abreviaturas: CSI, capturar, puntuar e integrar; HAN, red de atención jerárquica; HPA-BLSTM, memoria a corto plazo hipotalámica-pituitaria-adrenocortical-bidireccional; TCNN – URG, transferencia de red neuronal convolucional–generador de respuesta del usuario.

Nota: Los mejores resultados están resaltados en negrita.

ellos oraciones de 30 palabras. Todas estas oraciones de tamaño único se asignaron a capas GRU para producir oraciones incrustadas en forma de codificación 2D one-hot. Finalmente, esta salida se entregó como entrada a la SVM para detectar clases de noticias verdaderas y falsas. Una ilustración de nuestro modelo propuesto se muestra en la Figura2.

6.2 | Sintonia FINA

Elegimos las métricas de rendimiento de precisión, precisión, recuperación y puntuación F1 para evaluar y comparar el rendimiento del modelo. En el proceso de formación, seleccionamos el 75% de las muestras al azar y evaluamos el desempeño del 25% restante de la muestra. Este proceso se repitió cinco veces y las medidas promedio se describen en la Tabla2. Los parámetros de SVM se establecieron en 0,001 y 1,0, respectivamente. Se demostró experimentalmente que estos valores eran los mejores valores de parámetro.

7 | RESULTADOS Y COMPARACIÓN

Para comparar aún más nuestro modelo propuesto, elegimos los modelos representativos existentes descritos en la discusión que sigue. Para comparar con estos modelos, se calculan diversas medidas como la exactitud, la puntuación f, la precisión y la sensibilidad, como se muestra en la Tabla2. Sin embargo, a menudo es difícil medir el desempeño del modelo usando exactitud, recuperación y precisión, por lo que necesitamos examinar la curva AUC-ROC, que permite una tasa de falsos positivos porque traza la tasa de verdaderos positivos contra una tasa de falsos positivos. Cifras3 y 4 mostrar los puntajes AUC micro y macro promedio y por clase logrados con el modelo propuesto y mostrar puntajes AUC consistentes, lo que indica predicciones estables del modelo propuesto. En los siguientes métodos, hipotalámico-pituitario-adrenocortical-bidireccional largo
memoria a corto plazo (HPA – BLSTM) [45] y CSI [33] incorporan el contenido de las noticias y consideran los comentarios de los usuarios, mientras que otros métodos funcionan basándose únicamente en el contenido de las noticias. Los resultados en cuanto a la precisión y la puntuación F1 de nuestro modelo propuesto se compararon con los modelos de vanguardia existentes proporcionados

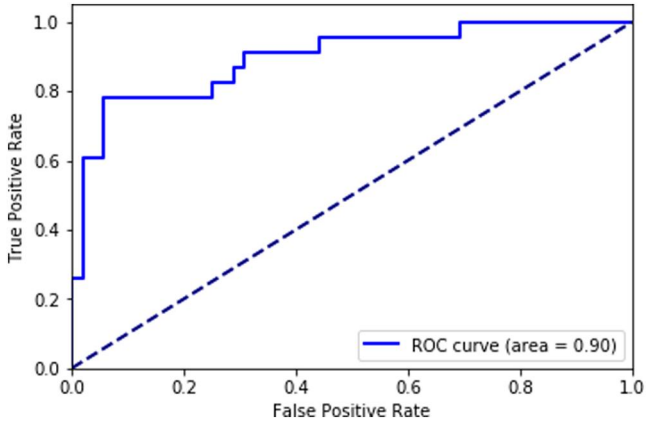

FIGURA 3 Curva AUC-ROC para el conjunto de datos PolitiFact

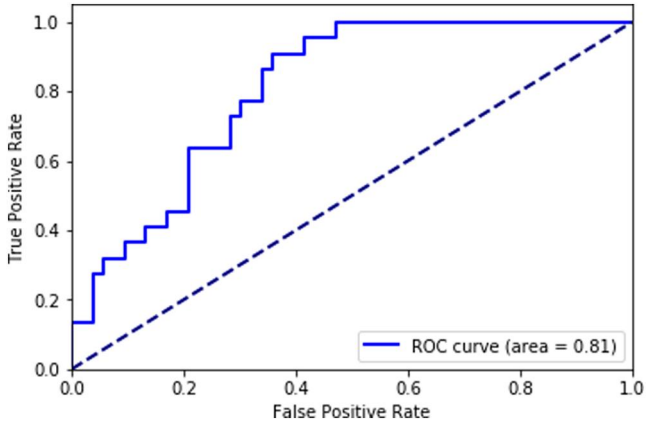

FIGURA 4 Curva AUC-ROC para el conjunto de datos de GossipCop

en cifras 5 y 6. En mesa2, varias medidas de desempeño indican que nuestro trabajo supera a los modelos existentes.

HAN [46]: Para detectar noticias falsas, la atención jerárquica network (HAN) aplica el marco de las redes neuronales de atención jerárquica en el contenido de las noticias. En esto

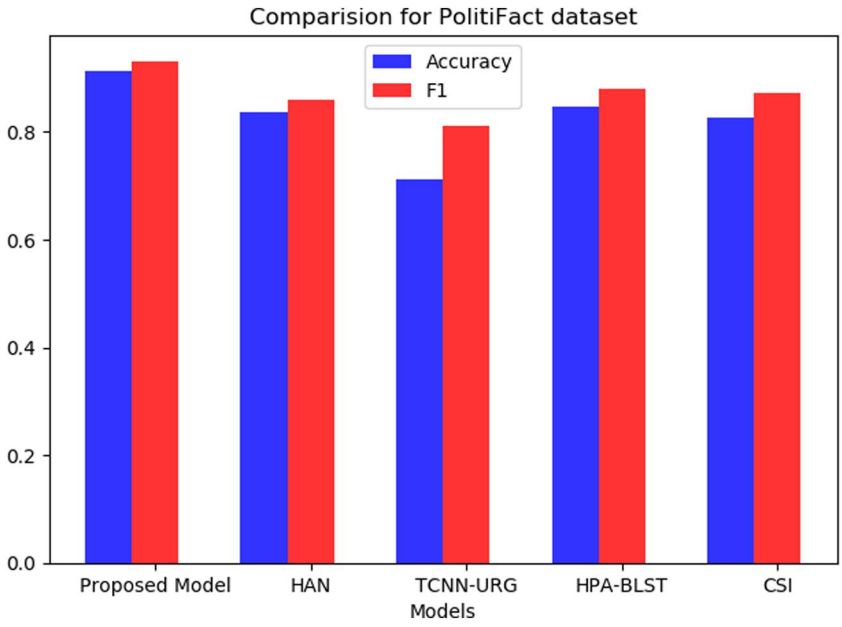

FIGURA 5 Comparación de nuestro modelo con modelos existentes en términos de precisión y puntuación F1 para el conjunto de datos PolitiFact

marco, la atención a nivel de palabra se usa en cada oración, mientras que la atención a nivel de oración se usa en cada documento.

TCNN – URG [15]: La transferencia neuronal convolucional El marco generador de respuesta de red-usuario (TCNN-URG) contiene dos componentes principales: una red neuronal convolucional y un autocodificador variacional condicional. La red neuronal convolucional es responsable de aprender las representaciones del contenido de las noticias, mientras que el codificador automático variacional se utiliza para registrar los atributos de los comentarios de los usuarios.

HPA – BLSTM [45]: HPA – BLSTM utiliza HAN marco basado en una red neuronal. Aprende la incorporación de noticias de la participación de los usuarios a nivel de palabra, nivel posterior y nivel secundario en las redes sociales.

CSI [33]**:** CSI es un aprendizaje profundo híbrido basado en LSTM marco que modela representaciones de noticias basadas en la incrustación de Doc2Vec tomando como entradas el contenido de las noticias y los comentarios de los usuarios.

8 | DISCUSIÓN Y LIMITACIONES

Para considerar solo métodos basados en contenido de noticias, las señales semánticas y sintácticas pueden capturarse de manera eficiente a través del marco HAN basado en redes neuronales de atención jerárquica. Nuestro trabajo se basa en contenidos de noticias, pero además incorpora datos de comentarios. De los resultados en la tabla2, se puede observar que los comentarios contienen información complementaria; esto nos ayudó a mejorar el rendimiento de detección del método. El rendimiento de nuestro método propuesto es superior al de los marcos HPA-BLSTM y CSI. HPA – BLSTM y CSI son métodos basados en comentarios de usuarios y el rendimiento de dichos métodos es superior al de los métodos basados en contenido de noticias. Considerando la incorporación de tanto los comentarios como el contenido de las noticias revelaron que los comentarios de los usuarios tienen más poder discriminativo que el contenido de las noticias o los comentarios por sí solos. Sin embargo, debido a que nuestro modelo propuesto explota los comentarios de los usuarios para mejorar la tasa de detección, esto lo hace altamente vulnerable a los ataques a través de comentarios adversarios (como UNITRIGGER [47], HOTFLIP [48], TextBugger [49] y MALCOM [50]). En particular, generar comentarios de alta calidad y coherencia en los artículos mediante el uso de un conjunto menos diverso de palabras muy relevantes puede inducir a error a nuestro modelo.

Además de estos ataques, GPT - 2 [51] es un método moderno que puede producir textos falsos. GPT - 2 es un modelo de lenguaje masivo basado en transformadores con 1.5 mil millones de parámetros con el objetivo básico de predecir la siguiente palabra en cualquier texto a partir de palabras anteriores [51]. Los términos 'texto real' y 'texto falso' se refieren solo a si el texto fue generado por una máquina o un humano. Sin embargo, GPT - 2 no presenta un argumento claro sobre la veracidad de un contenido. Hay casos en los que un modelo de lenguaje produce una expresión válida y también hay casos en los que un humano escribe una declaración falsa. Argumentamos que el texto generado por máquina no es lo mismo que el texto falso. Identificar si es probable que una parte del contenido sea generado por una máquina puede ayudar a medir la credibilidad. Por ejemplo, la incorporación de patrones lingüísticos puede detectar eficazmente comentarios escritos por humanos para identificar un comentario malicioso generado por una máquina. Si bien esto puede mejorar el rendimiento del modelo para detectar noticias falsas, también eleva el listón de posibles ataques.

En una dirección diferente, algunos estudios [52,53,54] han sugerido que la detección de noticias falsas no debería simplificarse como enfoques supervisados que se centran en etiquetas binarias individuales. En cambio, debe modelarse y estudiarse como un espectro continuo de valores. Particularmente, debemos considerar la complejidad de abrazar la manipulación y el engaño para detectar una coordinación sospechosa. En consecuencia, la solución potencial

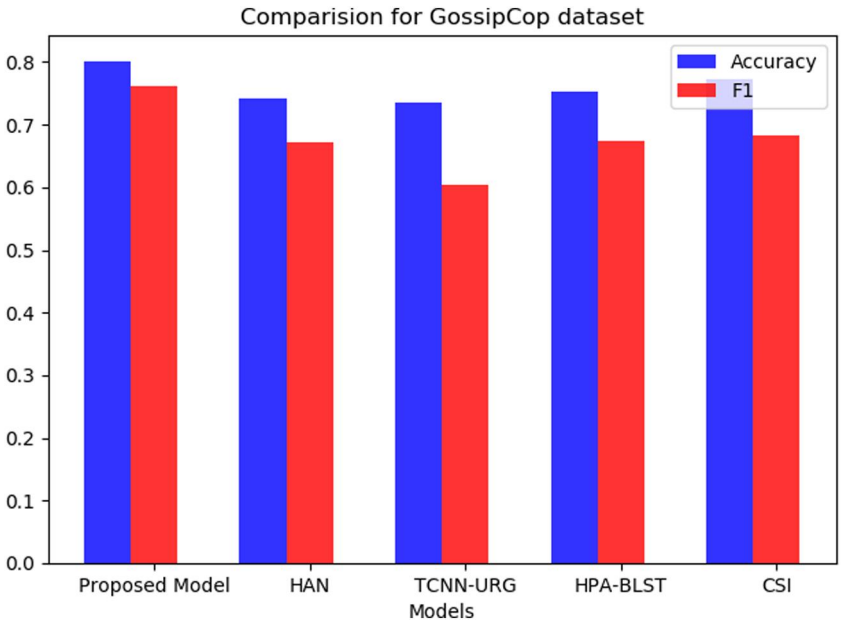

FIGURA 6 Comparación de nuestro modelo con modelos existentes en términos de precisión y puntuación F1 para el conjunto de datos de GossipCop

no debe incluir etiquetas lineales demasiado simplistas, sino que debe producir medidas de cooperación polifacéticas cuestionables.

Finalmente, una limitación de SVM es que su rendimiento depende del tamaño del vector de características. A medida que el vector de características aumenta, el rendimiento aumenta, pero el rendimiento disminuye para los vectores de características pequeños. Por lo tanto, el tamaño mínimo del vector de características (salida de las GRU) se limitó a 512 unidades.

9 | CONCLUSIÓN

Hemos propuesto un modelo híbrido basado en un RNN y un SVM para detectar rumores en el contenido de noticias en conjuntos de datos de FakeNews compuestos por dos subpartes, los conjuntos de datos PolitiFact y GossipCop. Se utilizó RNN para codificar contenido de noticias y comentarios para presentar la representación. Las características se proporcionaron como entrada a una SVM con un kernel gaussiano para detectar rumores (noticias reales o falsas) en los datos de entrada. Los resultados en cuanto a la precisión y la puntuación F1 de nuestro modelo propuesto se compararon con los modelos de vanguardia existentes. Varias medidas de desempeño muestran que nuestro trabajo ha superado a los modelos existentes.

Los investigadores deben prestar más atención a comprender las estructuras de noticias falsas para comprender sus patrones y su difusión en el universo digital. La presentación online de fake news se ha ido adaptando al crecimiento digital y sigue adquiriendo nuevos formatos cada vez más significativos. Las investigaciones futuras deberían discutir el efecto de las noticias falsas y la desinformación de manera amplia para las últimas formas. El fenómeno de las fake news con un abanico más amplio de fuentes en diversos escenarios también será fundamental en futuros estudios.

CONFLICTO DE INTERESES

El autor declara que no tengo ningún conflicto de intereses.

ORCID

Marwan Albahar https://orcid.org/0000-0003-3586-3423

REFERENCIAS

1. Allcott, H., Gentzkow, M .: Redes sociales y noticias falsas en las elecciones de 2016. J. Econ. Perspect. 31, 211–36 (2017)
2. Zhang, Y., et al .: propagación de rumores e información autorizada modelo considerando super propagación en redes sociales complejas. Phys. Stat. Mech. Apl. 506, 395–411 (2018)
3. Girgis, S., Amer, E., Gadallah, M .: Algoritmos de aprendizaje profundo para detección de noticias falsas en texto en línea. 13th IEEE International Conference on Computer Engineering and Systems (ICCES) (2018)
4. Liu, Y., fang Brook Wu, Y .: Detección temprana de noticias falsas en las redes sociales a través de la clasificación de rutas de propagación con redes recurrentes y convolucionales. AAAI. 354–361 (2018)
5. Popat, K .: Evaluación de la credibilidad de las afirmaciones en la web. En: Actas de la 26a Conferencia Internacional sobre World Wide Web Companion, págs. 735–739. Comité Directivo de Conferencias Internacionales de la World Wide Web (2017)
6. Gupta, A., et al .: TweetCred: Evaluación de la credibilidad en tiempo real del contenido de Twitter, págs. 228–243. Springer International Publishing (2014)
7. Yang, F., et al .: Detección automática de rumor en sina weibo. En: Actas del taller ACM SIGKDD sobre semántica de datos de minería, vol. 12. MDS Association for Computing Machinery, Nueva York (2012)
8. Castillo, C., Mendoza, M., Poblete, B .: Credibilidad de la información en twitter. En: Actas de la 20ª Conferencia Internacional sobre la World Wide Web, vol. 11, págs. 675–684.WWW Association for Computing Machinery, Nueva York (2011)
9. Horne, BD, Adali, S .: Esto acaba de llegar: las noticias falsas tienen mucho título, usan contenido más simple y repetitivo en el cuerpo del texto, más similar a la sátira que a las noticias reales, (2017) arXiv, eprint: 1703.09398
10. P_erez - Rosas, V., et al .: Detección automática de fake news. En: Coling '18, págs. 3391–3401 (2018)
11. Newman, ML et al .: Palabras mentirosas: predicción del engaño a partir de estilos lingüísticos. Pers. Soc. Psychol. Toro. 29 (5), 665–675 (2003)
12. Silverman, C .: Mentiras, malditas mentiras y contenido viral: cómo los sitios web de noticias difunden (y desacreditan) rumores en línea, afirmaciones no verificadas y desinformación. Tow Center de Periodismo Digital. Universidad de Columbia (2015)
13. Shu, K., et al .: Un repositorio de datos con contenido de noticias, contexto social e información espacio-temporal para estudiar noticias falsas en las redes sociales. Big Data. 8 (3), 171–188 (2020)

14. Zubiaga, A., et al .: Detección y resolución de rumores en redes sociales. Computación ACM. Surv. 51, 1–36 (2018)
15. Qian, F. et al .: Generador de respuesta neuronal del usuario: detección de noticias falsas con inteligencia colectiva del usuario. En t. Conf. Conjunta Artif. Intell. 18, 3834–3840 (2018)
16. Zhou, X., Zafarani, R .: Una encuesta de noticias falsas: teorías fundamentales, métodos de detección y oportunidades. Computación ACM. Surv. 53 (2020)
17. Vosoughi, S., Roy, D., Aral, S .: La difusión de noticias verdaderas y falsas en línea. Ciencias. 359 (6380), 1146-1151 (2018)
18. Baly, R., et al .: Predecir la facticidad de los informes y el sesgo de las fuentes de los medios de comunicación. CoRR. abs / 1810, 01765 (2018)
19. Afroz, S., Brennan, M., Greenstadt, R .: Detectando engaños, fraudes y engaños en el estilo de escritura en línea. En: Simposio de IEEE sobre seguridad y privacidad de 2012, págs. 461–475 (2012)
20. Rubin, V., et al .: ¿Noticias falsas o verdad? utilizando señales satíricas para detectar noticias potencialmente engañosas. En: Actas del segundo taller sobre enfoques computacionales para la detección de engaños, págs. 7–17, Asociación de Lingüística Computacional, San Diego, CA (2016)
21. Rashkin, H., et al .: Verdad de diferentes matices: análisis del lenguaje en noticias falsas y verificación de hechos políticos. En: Actas de la Conferencia de 2017 sobre métodos empíricos en el procesamiento del lenguaje natural, págs. 2931–2937, Asociación de Lingüística Computacional, Copenhague (2017)
22. Potthast, M., et al .: Una investigación estilométrica sobre noticias falsas y hiperpartidistas. Actas de la 56ª Reunión Anual de la Asociación de Lingüística Computacional, Documentos largos, vol. 1 (2018)
23. Bhatt, G., et al .: Combinación de características neuronales, estadísticas y externas para la identificación de posturas de noticias falsas. Compañero de la Conferencia Web 2018 en la Conferencia Web 2018. WWW '18 (2018)
24. Borges, L., Martins, B., Calado, P .: Combinando características de similitud y aprendizaje de representación profunda para la detección de posturas en el contexto de verificación de noticias falsas. J. Data Infor. Qual. 11 (2019)
25. Rubin, VL, Chen, Y., Conroy, NK: Detección de engaños para noticias: tres tipos de falsificaciones. Proc. Assoc. En para. Sci. Technol. 52 (1), 1-4 (Ene. 2015)
26. Conroy, NK, Rubin, VL, Chen, Y .: Detección automática de engaños: métodos para encontrar noticias falsas. Proc. Assoc. En para. Sci. Technol. 52 (1), 1 a 4 (2015)
27. Chen, Y., Rubin, VL, Conroy, N .: Hacia la verificación de noticias: métodos de detección de engaños para el discurso de las noticias. En: Actas de la 48ª Conferencia Europea sobre los principios del descubrimiento del conocimiento y la minería de datos (PKDD), 48ª ed. (2015)
28. Dadgar, SMH, Araghi, MS, Farahani, MM: Un enfoque novedoso de minería de texto basado en TF-IDF y una máquina de vectores de apoyo para la clasificación de noticias. En: 2016 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, págs. 112-116 (2016)
29. Jin, Z., et al .: Verificación de noticias mediante la explotación de puntos de vista sociales conflictivos en microblogs. En: Actas de la trigésima Conferencia AAAI sobre Inteligencia Artificial, AAAI, vol. 16, págs. 2972–2978.AAAI Press (2016)
30. Shu, K., et al .: Detección de noticias falsas en las redes sociales. SIGKDD Expl. Newsl. 19 (1), 22 a 36 (2017)
31. Ahmed, H., Traore, I., Saad, S .: Detectar mensajes no deseados de opinión y noticias falsas mediante la clasificación de texto. Asegurar. Intimidad. 1 (1), e9 (2017)
32. Gilda, S .: Aviso de infracción de los principios de publicación de IEEE: evaluación de algoritmos de aprendizaje automático para la detección de noticias falsas. 2017 IEEE 15th Student Conference sobre investigación y desarrollo. (PUNTUACIÓN) (2017)
33. Ruchansky, N., et al .: Un modelo profundo híbrido para la detección de noticias falsas. Actas de la ACM de 2017 sobre la Conferencia sobre Gestión de la Información y el Conocimiento (2017)
34. Shu, K., Wang, S., Liu, H .: Explotación de la triple relación para la detección de noticias falsas. ArXiv. vol. abs / 1712, 07709 (2017)
35. Long, Y., et al .: Detección de noticias falsas a través de perfiles de oradores con múltiples perspectivas. En: Actas de la Octava Conferencia Internacional Conjunta sobre Procesamiento del Lenguaje Natural, págs. 252-256, Federación Asiática de Procesamiento del Lenguaje Natural (2017)
36. Figueira, Á., Oliveira, L .: El estado actual de las fake news: retos y oportunidades. Procedia Comput. Sci. 121, 817–825 (2017)
37. Buntain, C., Golbeck, J .: Identificación automática de noticias falsas en hilos populares de Twitter. En: 2017 IEEE International Conference on Smart Cloud, págs. 208–215 (2017)
38. Bessi, A .: Sobre las propiedades estadísticas de la desinformación viral en las redes sociales en línea. Phys. Stat. Mech. Apl. 469, 459–470 (2017)
39. Azam, N., Yao, J .: Comparación de la frecuencia de términos y métricas de selección de características basadas en la frecuencia de documentos en la categorización de texto. Expert Syst. Apl. 39 (5), 4760–4768 (2012)
40. Shu, K., Wang, S., Liu, H .: Comprensión de los perfiles de usuario en las redes sociales para la detección de noticias falsas. En: Conferencia IEEE de 2018 sobre procesamiento y recuperación de información multimedia, (MIPR), Miami, págs. 430–435 (2018)
41. Zhang, J., et al .: Detección de noticias falsas con modelo de red de difusión profunda. ArXiv. vol. abs / 1805, 08751 (2018)
42. Bajaj, S .: ¡El Papa tiene un nuevo bebé! detección de noticias falsas mediante aprendizaje profundo (2017)
43. Trovati, M., Hill, R., Bessis, N .: Un método de detección de mensajes no genuino basado en conjuntos de datos no estructurados. En: 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, (3PGCIC), Cracovia, págs. 597–600 (2015)
44. Singhania, S., et al .: Una red neuronal profunda para la detección de noticias falsas. En Procesamiento de información neuronal, págs. 572–581, Springer International Publishing (2017)
45. Guo, H., et al .: Detección de rumores con red jerárquica de atención social. Actas de la 27a Conferencia Internacional ACM sobre Gestión de la Información y el Conocimiento (2018)
46. Yang, Z., et al .: Redes de atención jerárquica para la clasificación de documentos. En: Actas de la Conferencia de 2016 del Capítulo Norteamericano de la Asociación de Lingüística Computacional. Tecnologías del lenguaje humano (2016)
47. Cer, D., et al .: Universal Sentence Encoder (2018)
48. Ebrahimi, J., et al .: HotFlip: Ejemplos de confrontación de caja blanca para clasificación de texto. Actas de la 56a Reunión Anual de la Asociación de Lingüística Computacional (2018)
49. Li, J., et al .: TextBugger: generar texto de confrontación contra aplicaciones del mundo real. Proceedings 2019 Simposio sobre seguridad de redes y sistemas distribuidos (2019)
50. Le, T., et al .: Generación de comentarios maliciosos para atacar modelos neuronales de detección de noticias falsas, págs. 2020 (2009) ArXiv, abs / .01048
51. Radford, A, et al .: Los modelos lingüísticos son aprendices multitarea sin supervisión (2019)
52. Starbird, K .: La propagación de la desinformación: bots, trolls y todos nosotros. Naturaleza. 571, 449–449 (2019)
53. Cresci, S .: Una década de detección de bots sociales. Comun. ACM. 63, 72–83 (2020)
54. Nizzoli, L., et al .: Comportamiento coordinado en las redes sociales en las elecciones generales del Reino Unido de 2019 (2020) arXiv e-prints, p. arXiv: 2008.08370

Cómo citar este artículo: Albahar M. Un modelo híbrido para la detección de noticias falsas: aprovechar el contenido de las noticias y los comentarios de los usuarios en las noticias falsas. *IET Inf. Asegurar* . 2021; 15: 169-177.https://doi.org/10.1049/ise2.12021