Este artículo ha sido aceptado para su publicación en un número futuro de esta revista, pero no ha sido completamente editado. El contenido puede cambiar antes de la publicación final. Información de la cita: DOI 10.1109 / TKDE.2021.3103833, IEEE Transacciones sobre el conocimiento y la ingeniería de datos

Codificador-decodificador controlado por categoría para falsos Detección de noticias

Lianwei Wu, Yuan Rao, Cong Zhang, Yongqiang Zhao y Ambreen Nazir

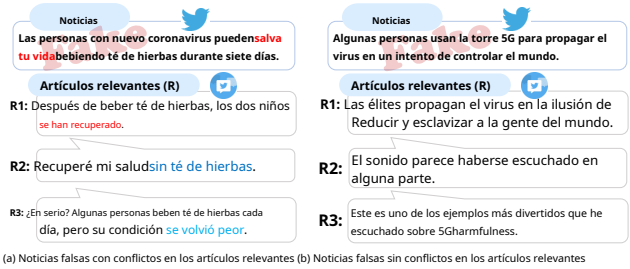
Abstracto—Los enfoques basados en datos existentes típicamente capturan representaciones indicativas de credibilidad de artículos relevantes para la detección de noticias falsas, como opiniones escépticas y contradictorias. Sin embargo, estos métodos todavía tienen varios inconvenientes: 1) Debido a la dificultad de recopilar noticias falsas, la capacidad de los conjuntos de datos existentes es relativamente pequeña; y 2) existe una cantidad considerable de noticias no verificadas que carecen de voces conflictivas en los artículos relevantes, lo que dificulta que los métodos existentes identifiquen su credibilidad. Especialmente, las diferencias entre noticias verdaderas y falsas no se limitan a si hay características de conflicto en sus artículos relevantes, sino que también incluyen diferencias ocultas más extensas a nivel lingüístico, como las perspectivas de la expresión emocional (como la emoción extrema en las noticias falsas). , estilo de escritura (como el título impactante en clickbait), etc., los métodos existentes son difíciles de capturar completamente estas diferencias. Para capturar diferencias más generales y de amplio alcance entre noticias verdaderas y falsas, en este artículo, directamente de las diferentes categorías de noticias en sí, proponemos un modelo de codificador-decodificador controlado por categorías (CED) para generar ejemplos con características diferenciadas por categorías y ampliar la capacidad del conjunto de datos para lograr un efecto de mejora de datos, mejorando así la detección de noticias falsas. Específicamente, para hacer que los ejemplos generados enriquezcan más las características de las noticias, desarrollamos un codificador guiado por noticias para guiar artículos relevantes para generar representaciones de contexto semántico de noticias. Para impulsar los ejemplos generados para que contengan más características diferenciadas por categorías, Diseñamos un decodificador controlado por categoría que se basa en una unidad de patrón compartido para capturar, respectivamente, características compartidas dentro de la categoría dentro de noticias verdaderas o falsas, y emplea una unidad de restricción para forzar a los dos tipos de características compartidas a ser más diferentes para resaltar características diferenciadas entre categorías. . Los resultados experimentales en tres conjuntos de datos demuestran la superioridad de CED.

Términos del Índice—Descodificador de codificador, detección de noticias falsas, análisis de redes sociales, procesamiento de lenguaje natural

F

1 Introducción

As el megáfono en la nueva era, las redes sociales con su igualdad y ocultación únicas permiten que todos se conviertan en participantes y comentaristas de noticias, para transmitir información libremente. Sin embargo, la información no solo contiene información veraz y confiable, sino también muchas noticias falsas. Las investigaciones ilustran que, durante las elecciones presidenciales de EE. UU. (2016), la tasa de tweets de los usuarios que tuitean enlaces a sitios web contiene noticias clasificadas como falsas más de cuatro veces más grandes que las de los medios tradicionales, a veces incluso agregando al 25% de los tweets en Twitter difunden noticias falsas [1]. Más que eso, el 1% de los usuarios están expuestos al 80% de las noticias falsas [2]. La difusión de noticias falsas ha tenido un impacto negativo sin precedentes en la vida personal, la estabilidad social y el patrón político. Por lo tanto,



(a) Noticias falsas con conflictos en los artículos relevantes (b) Noticias falsas sin conflictos en los artículos relevantes

Fig. 1. La descripción intuitiva de las características del conflicto.

Actualmente, la mayoría de los estudios existentes se centran principalmente en capturar características de texto [3], [4], [5] y características de metadatos [6], [7] confiando en estructuras y semántica de publicaciones y comentarios y propuestos modelos estructurados en árbol basados en neuronas recursivas redes para aprender representaciones para la detección de rumores. Shu*et al.* [9] desarrollaron redes de atención conjunta para explotar tanto publicaciones como en redes neuronales profundas para la detección de noticias falsas, que han comentarios para capturar de forma conjunta$k$ oraciones dignas de verificación para lograr un gran éxito. Especialmente, debido a la mayoría de las publicaciones en la detección de noticias falsas sociales. Más recientemente, varios métodos de concentración mediática se basan en textos breves y falta de semántica suficiente, en descubrir rasgos conflictivos entre noticias y relevantes, sus comentarios relevantes o artículos relevantes son artículos ampliamente explotados, incluyendo diferenciales [10], dudosos [11], desaprobadores y se confirmó que son potentes indicadores de credibilidad, donde [12], y funciones refutatorias [13] para la detección de noticias falsas. Como los artículos relevantes representan una serie de artículos o comentarios que ejemplos concretos, Popat*et al.* [10] intentó desacreditar la discusión falsa de una noticia específica. En detalle, Ma*et al.* [8] consideró afirmaciones mediante la construcción de un modelo de interacción basado en la atención para aprender palabras diferenciales y contradictorias de artículos relevantes. Wu*et al.* [12] propuso redes de fusión de interacción adaptativa que cumplen la fusión de interacción cruzada entre afirmaciones y artículos relevantes para capturar sus conflictos semánticos y características en desacuerdo para detección. También desarrollaron redes de atención interactivas jerárquicas conscientes de la evidencia [14] para capturar la semántica de credibilidad que discute las partes cuestionables de las afirmaciones de los artículos relevantes como fragmentos semánticos de conflicto para la verificación de las afirmaciones.

Las características conflictivas podrían ilustrarse intuitivamente con un caso típico. Como se muestra en la Figura 1 (a), 'empeoró' en el

- *L. Wu, Y. Rao, Y. Zhao y A. Nazir estaban con Xi'an Key Lab. de Social Procesamiento de datos de inteligencia y complejidad, la Escuela de Ingeniería de Software, Universidad de Xi'an Jiaotong; Laboratorio clave conjunto de Shaanxi para artefactos Inteligencia (Sublaboratorio de la Universidad Xi'an Jiaotong), Xi'an, Shaanxi 710054, China*
  Correo electrónico: stayhungry@stu.xjtu.edu.cn , raoyuan@mail.xjtu.edu.cn , {yongqiang1210, ambreen.nazir} @ stu.xjtu.edu.cn
- *C. Zhang trabajaba con Xi'an Huanyu Satellite Control and Data Application Co. Ltd. Xi'an 710065, China. Correo electrónico: congzhang0825@gmail.com*

el artículo relevante 3 (R3) tiene una semántica de conflicto con "salva tu vida" en las noticias y "se ha recuperado" en el artículo relevante 1 (R1). Los modelos que capturan características conflictivas tan efectivas entre noticias y artículos relacionados o entre artículos relevantes podrían lograr un rendimiento notable. Sin embargo, estos modelos todavía tienen varias limitaciones generales.**Primero**, debido a la dificultad de recopilar noticias falsas, la capacidad de los conjuntos de datos públicos existentes es relativamente pequeña (como el conjunto de datos PHEME solo contiene 2246 elementos, que se describen en detalle en la Sección 4.1), y los modelos actuales que capturan características de conflicto no pueden expandir la capacidad de conjuntos de datos.

Segundo, cuando se aprovechan artículos extra relevantes para detectar noticias falsas, los modelos existentes carecen de filtrado de características, que es fácil de introducir características de ruido no relacionadas con las noticias, lo que interfiere con el rendimiento de los modelos. **Tercera**, hay una cantidad considerable de noticias no verificadas que carecen de voces conflictivas en los artículos relevantes, lo que dificulta que los métodos existentes identifiquen su credibilidad. Particularmente, las diferencias significativas entre noticias verdaderas y noticias falsas no solo se limitan a si hay características de conflicto en sus artículos relevantes, sino que también incluyen diferencias ocultas más extensas a nivel lingüístico, como las perspectivas de expresión emocional (como emociones extremas en noticias falsas) [15], [16], estilos de escritura (como títulos impactantes en clickbait) [17], [18], [19], [20], etc., mientras que los modelos existentes basados en conflictos son incapaces de explorar las características de diferencia entre noticias falsas desde perspectivas tan amplias. Por lo tanto, cómo se explica la arquitectura de CED y se explica el diseño de cada diseño un modelo eficaz que toca directamente el módulo más amplio en detalle. Los resultados experimentales y la discusión son El alcance y las diferencias universales entre noticias verdaderas y falsas es la clave para mejorar el rendimiento de la detección de noticias falsas.

Para abordar los problemas anteriores, en lugar de capturar detalladamente las características de los conflictos, nos esforzamos por explorar la amplia gama de diferencias entre las categorías de noticias que son verdaderas y falsas directamente. Con base en esta idea, en este artículo proponemos**C**categoría revisado **mi**ncoder-**D**modelo ecoder (en adelante, CED) para generar noticias falsas es un tema de investigación de larga data. Comió ejemplos con características diferenciales entre categorías y los lectores pueden consultar [15] para una encuesta reciente. En lugar de ampliar la capacidad del conjunto de datos para lograr una mejora de los datos, los métodos anteriores de detección de noticias falsas mediante el control manual, mejorando así la detección de noticias falsas. Los ejemplos son características de estructuración, nos centramos en la detección automática de noticias falsas generadas mediante la recopilación de características diferenciadas entre categorías con la ayuda del modelo de generación de texto. Por tanto, nuestra de los dos tipos de funcionalidades compartidas intracategoría, donde el trabajo se relaciona con dos grupos de tareas: fake news automáticas detección y generación de lenguaje natural.

Las características compartidas dentro de la categoría denotan las características comunes dentro de cada categoría de noticias (es decir, noticias verdaderas o falsas), y las características diferenciadas entre categorías se refieren a las características que distinguen entre noticias verdaderas y falsas. Específicamente, con el fin de capturar valiosas funciones relacionadas con las noticias de artículos relacionados, CED diseña un módulo codificador guiado por noticias, que se basa en contenido y metadatos. El methtrue basado en contenido y fake news para orientar sus artículos relevantes en la generación de ods se extraen principalmente de los siguientes aspectos: gramática verdadera y ejemplos fake con características de noticia, respectivamente. [21], semántica [22], emociones [23], estilos [3], [5] y posturas Explorar diferencias ocultas más extensas entre verdadero y [24], [25]. Por ejemplo, Chen*et al.* [26] atención combinada de noticias falsas, el modelo CED construye un mecanismo decodificador controlado por categorías con el módulo Recurrent Neural Networks (RNN) para enfocar, que tiene como objetivo consolidar la diferenciación entre características de texto con diferentes atenciones para capturar las noticias verdaderas y falsas, que diseña dos unidades, es decir, representación de patrón oculto. Wu*et al.* [25] diseñó una unidad de cribado múltiple y una unidad de restricción. La primera unidad refuerza el método de aprendizaje de tareas con una capa de intercambio seleccionada para filtrar y capturar las características compartidas dentro de la categoría dentro de cada categoría. Aunque este tipo de enfoque es eficaz, sus artículos relevantes), respectivamente, y la unidad de restricción obliga a que la captura de la diversidad de características sea limitada. Ambos tipos de características compartidas deben ser más diferentes, para resaltar los métodos basados en metadatos se centran en extraer características que surgen características diferenciadas entre categorías. Finalmente, combinamos las fuentes de redondeo [27], las publicaciones [28], [29], los comentarios [9], los usuarios [30], los ejemplos generados verdaderos y falsos con los datos originales de las noticias [31] y las redes de propagación [8]. para la detección de noticias falsas. para formar un conjunto de datos mejorado para mejorar la capacidad de detección. Concretamente, el rápido desarrollo de las redes sociales ha brindado a Experimentos en tres conjuntos de datos públicos, es decir, Snopes, PolitiFact y cada usuario anónimo la oportunidad de convertirse en un editor de PHEME, lo que demuestra la efectividad de CED. Los principales aportes de la información, que ha incrementado mucho la medición del papel, son los siguientes: complejidad de la credibilidad de las fuentes de información, lo que la hace

- Se explora una nueva perspectiva sobre la detección de noticias falsas, que considera la selección de características compartidas dentro de una categoría (noticias verdaderas o falsas) y la captura de características de individualización entre categorías (noticias verdaderas y falsas) para generar ejemplos con características diferenciales para fortalecer las noticias falsas. detección.
- Se desarrolla el módulo codificador guiado por noticias, que se basa en la coincidencia semántica y la fusión entre noticias y artículos relevantes para hacer que los ejemplos generados posean una semántica de noticias más valiosa (Sección 4.4.1 y 4.5.1).
- La unidad de patrón compartido con un mecanismo de filtrado de puerta es capaz de capturar características compartidas dentro de la categoría dentro de noticias verdaderas o falsas que contienen características indicativas de credibilidad de múltiples perspectivas (Sección 4.4.2 y 4.5.3), y la unidad de restricción propuesta es capaz de Restringir la independencia y la diferencia de los dos tipos de características compartidas para subrayar las características diferenciadas entre categorías (Sección 4.4.3).
- Nuestro enfoque en tres conjuntos de datos del mundo real logra mejoras superiores sobre las líneas de base de última generación (Sección 4.3). Además, la combinación de los ejemplos generados por nuestro modelo con los datos originales puede mejorar significativamente el rendimiento de los métodos de referencia (Sección 4.4.6).

Las partes restantes de este documento están organizadas de la siguiente manera. Revisamos algunos trabajos relacionados en la Sección 2. La Sección 3 presenta descrito en la Sección 4, y finalmente, la Sección 5 concluye nuestra trabajar y describe las direcciones para el trabajo futuro.

2 Rexaltado Work

2.1 Detección automática de noticias falsas

Los métodos para la detección automática de noticias falsas se dividen

Es difícil obtener un mejor rendimiento para la credibilidad de las características de la categoría de noticias basadas en fuentes para la detección de noticias falsas. Nuestros métodos basados. El marco general de métodos basados en el usuario y en la red se muestra en la Figura 2, que consiste en la necesidad de un aliado guiado por noticias para construir perfiles de usuario y redes de propagación, módulo codificador y módulo decodificador controlado por categoría. Los respectivamente, que requieren mucha mano de obra porque son tan complejos, se basan en noticias verdaderas y falsas para guiarlos, respectivamente, como los métodos para construir características manualmente. Debido a los artículos relevantes para producir una representación de contexto rica en características de credibilidad verdaderamente ricas (como características cuestionables, de argumentación y noticias falsas, y estas últimas tienen como objetivo controlar el decodificador o apoyar las voces) en comentarios (o artículos) relacionados con noticias, generar ejemplos con características diferenciadas por categorías (es decir, los métodos basados en comentarios se concentran en capturar características diferenciadas conflictivas entre noticias verdaderas y falsas). Particularmente, y la semántica cuestionable de los comentarios o artículos relevantes, el módulo decodificador controlado por categorías diseña patrones compartidos para detectar noticias falsas. Como ejemplos concretos, Ma *et al.* [8] unidad construida para capturar respectivamente las características compartidas dentro de las redes neuronales recursivas estructuradas en árbol (RvNN) para capturar la categoría (es decir, noticias verdaderas con sus artículos relevantes o noticias falsas, la representación oculta de ambas estructuras de propagación con sus artículos relevantes) y desarrolla la unidad de restricción a la fuerza (a partir de comentarios) y los contenidos del texto. También propusieron que los dos tipos de características compartidas fueran más diferentes, para obtener una nueva red de atención jerárquica de extremo a extremo [13] que se centra en las características diferenciadas entre categorías.

sobre aprender a representar una semántica coherente, así como su relación semántica con la afirmación de los artículos relevantes para la detección de noticias falsas. Este tipo de métodos adquiere efectivamente características de credibilidad indicativas, y se ha convertido en uno de los enfoques.Como se muestra en la Figura 3, el módulo codificador guiado por noticias es una característica de detección de noticias falsas actualmente, pero tiene problemas para combatir la capa de codificación de secuencia, una capa de coincidencia guiada, algunas noticias que carecen voces cuestionables en su capa relevante y de fusión. Primero, la secuencia de codificación de la capa de comentarios o artículos. En base a esto, en lugar de leer intencionalmente el contenido del artículo y las noticias relevantes, y aprende a encontrar voces dudosas a partir de noticias específicas, desde la perspectiva de sus representaciones contextuales por separado. Luego, guiados de las categorías de noticias, hacemos un gran esfuerzo para explorar la capa coincidente que reúne la información de noticias para cada palabra en artículos relevantes que reflejan las partes importantes del contexto de esta palabra. Finalmente, la capa de fusión fusiona la semántica de las noticias agregadas en cada palabra del artículo relevante para producir la representación final del contexto guiada por las noticias. Especialmente, el verdadero codificador guiado por noticias y el codificador guiado por noticias falsas generan representaciones de la misma manera. El seguimiento diferencias entre noticias falsas y verdaderas de sus artículos relevantes para mejorar la detección de noticias falsas.

2.2 Generación de lenguaje natural

El propósito principal de la tarea de generar lenguaje natural eración es generar automáticamente una pieza de proceso detallado de alta calidad que se aplica tanto a noticias verdaderas como falsas. texto en lenguaje natural con la ayuda del conocimiento existente, que se ha aplicado ampliamente en muchos campos, como la unidad recurrente neutralizada (Bi-GRU) para codificar artículos relevantes y traducción automática ral [32], resumen abstractivo [33], secuencias de noticias. Para cualquier secuencia X, cada palabra $w_I \in X$ escritura creativa [34], e incluso detección de noticias falsas [35]. El primero está incrustado en un D-Vector de incrustación dimensional Los métodos [36] para la tarea incluyen plantillas, estructuras $v_I \in \mathrm{R}_D$. Particularmente, adoptamos X_a, X_C para indicar el [37] basado, el codificador-decodificador basado, etc., donde codificador-decodificador incrustaciones del artículo relevante y la noticia, respectivamente.

Los métodos [38], [39] han logrado un rendimiento prometedor en esta tarea. Específicamente, CopyRNN [40] considera en primer lugar la generación proceso como una tarea de aprendizaje secuencia a secuencia y aplica un marco codificador-decodificador ampliamente disponible [41] con atención [42] y mecanismos de copia [43]. Basándose en CopyRNN, se han propuesto recientemente varias extensiones [44]. Además, muchos estudios introducen modelos de generación de codificador-decodificador [35], [45] para la tarea de detección de noticias falsas. Como ejemplo concreto, Ma *et al.* [35] propuso un modelo de codificador-decodificador de estilo GAN para generar voces inciertas o conflictivas para presionar al discriminador para aprender representaciones indicativas de rumores más fuertes para la detección de noticias falsas. Sin embargo, estos métodos solo se enfocan en una perspectiva de características de diferencia entre noticias verdaderas y falsas, es decir, características de conflicto, ignoran las diferencias basadas en perspectivas más amplias, como las perspectivas de expresión de emociones y estilos de escritura, etc. Las características de diferencia general y de amplio alcance entre noticias falsas y verdaderas son un tema crítico para la detección de noticias falsas. En base a esto, en este trabajo, desarrollamos un modelo generativo controlado por categorías para recopilar características compartidas dentro de la categoría y las características diferenciales entre las categorías para la detección de noticias falsas.

3 Categoria-revisadomincoder-Decoder Metroodel

En esta sección, proponemos un codificador controlado por categoría modelo de decodificador (CED) que genera ejemplos con la diferencia- agregan la semántica filtrada relevante de las noticias para

3.1 Módulo codificador guiado por noticias

Capa de codificación de secuencia Empleamos un bidireccional

Luego, cada palabra se asigna hacia adelante y hacia atrás. estados ocultos (denotados como $\overrightarrow{\mathbf{h}}_I$ y $\overleftarrow{\mathbf{h}}_I$) con lo siguiente operaciones definidas:

$$\overrightarrow{\mathbf{h}}_I = \mathrm{GRU}\,(v_I, \mathbf{h}_{yo1}) \quad (1)$$

$$\overleftarrow{\mathbf{h}}_I = \mathrm{GRU}\,(v_I, \mathbf{h}_{I+1}) \quad (2)$$

La concatenación de $\overrightarrow{\mathbf{h}}_I$ y $\overleftarrow{\mathbf{h}}_I$, $[\overrightarrow{\mathbf{h}}_I, \overleftarrow{\mathbf{h}}_I]$, sirve como w_I's estado oculto en el codificador, denotado como $\mathbf{h}_I \in \mathrm{R}_D$. $\mathbf{h}_{a\ I}$ y $\mathbf{h}_{C\ j}$ atender como los vectores contextuales para el I-ésima palabra del artículo correspondiente y la j-ª palabra de la noticia, respectivamente.

Capa a juego guiada En la capa de coincidencia, considerando que puede haber información en el artículo relevante que revela la verdad de las noticias falsas, lo que conduce a una semántica conflictiva entre el artículo relevante y la noticia. Para mantener las diferentes relaciones entre los diferentes tipos de noticias y comentarios, diseñamos un mecanismo cerrado que apunta a diferentes tipos de noticias, es decir, filtrando la semántica del conflicto entre noticias falsas y artículos relevantes, y filtrando la semántica ruidosa entre noticias verdaderas y artículos relevantes. Además, para reflejar la importancia de la información en el artículo relevante basada en el hecho de que la información altamente relacionada con las noticias en el artículo relevante debe contener más información central, utilizamos un mecanismo de atención guiado por noticias para

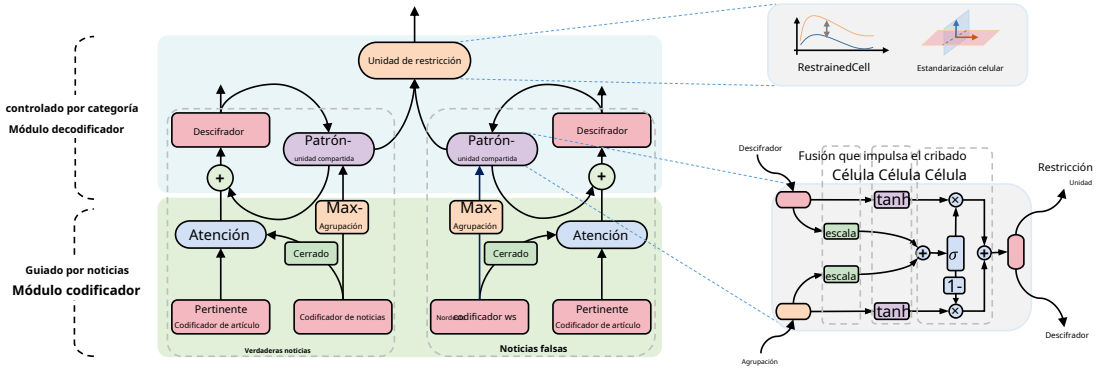


Fig. 2. Arquitectura general de CED. El modelo consta de dos módulos: módulo codificador guiado por noticias (los detalles se muestran en la Figura 3) y módulo decodificador controlado por categoría que incluye una unidad de patrón compartido y una unidad de restricción, excepto el decodificador.

cada palabra del artículo relevante. Formalmente, el proceso de agregación del I-th palabra (en el paso I) en el artículo correspondiente $\text{att}_I = \text{attn}(\mathbf{h}_{a_I}, [\mathbf{h}_{c_1}, \mathbf{h}_{c_2}, \ldots, \mathbf{h}_{c_{|c|}}]; \mathbf{W}_{att})$ se representa como:

$$\text{att}_I = \sum_{j=1}^{|c|} \alpha_{yo,j} \mathbf{h}_{gramo_j} \quad (3)$$

$$\mathbf{h}_{gramo_j} = \sigma(\mathbf{W}_{gramo}[\mathbf{h}_{a_I}; \mathbf{h}_{c_j}] + \mathbf{B}_{gramo}) \quad (4)$$

$$\alpha_{yo,j} = \text{Exp}(z_{yo,j}) / \sum_{k=1}^{|c|} \text{Exp}(z_{yo,k}) \quad (5)$$

$$z_{yo,j} = (\mathbf{h}_{a_I})^T \mathbf{W}_{att} \mathbf{h}_{c_j} \quad (6)$$

dónde $\mathbf{W}_{gramo}$, $\mathbf{W}_{att}$, y $\mathbf{B}_{gramo}$ son parámetros entrenables. $\sigma(\cdot)$ es función de activación sigmoidea, y $\alpha_{yo,j}$ es el normalizado puntaje de atención entre $\mathbf{h}_{a_I}$ y $\mathbf{h}_{c_j}$.

Así, a través de estas dos estructuras, el emparejamiento guiado La capa no solo puede capturar las partes altamente relacionadas con este noticias del artículo relevante, sino que también obtienen la semántica conflictiva entre el artículo relacionado y la noticia.

Capa de fusión de fusión Finalmente, el vector original $\mathbf{h}_{a_I}$ de el artículo relevante y el vector de información agregada att_I actuar como las entradas a la capa de fusión de fusión de información:

$$\overrightarrow{\mathbf{metro}}_I = \text{GRU}([\mathbf{h}_{a_I}; \text{att}_I], \overrightarrow{\mathbf{metro}}_{yo-1}) \quad (7)$$

$$\overleftarrow{\mathbf{metro}}_I = \text{GRU}([\mathbf{h}_{a_I}; \text{att}_I], \overleftarrow{\mathbf{metro}}_{I+1}) \quad (8)$$

$$\mathbf{metro}_I = \lambda \mathbf{h}_{a_I} + (1-\lambda)[\overrightarrow{\mathbf{metro}}_I, \overleftarrow{\mathbf{metro}}_I] \quad (9)$$

dónde $[\mathbf{h}_{a_I}, \text{att}_I] \in \mathbb{R}^{2D}$, $\overrightarrow{\mathbf{metro}} \in \mathbb{R}^{D/2}$, $[\overrightarrow{\mathbf{metro}}, \overleftarrow{\mathbf{metro}}] \in \mathbb{R}^{D}$, y $\mathbf{metro} \in \mathbb{R}^{D}$. $\mathbf{h}_{a_I}$ en la ecuación. (9) es una conexión residual, y $\lambda \in (0, 1)$ es el hiperparámetro. Finalmente, obtenemos la representación contextual guiada por noticias del contexto (es decir, $\mathbf{metro} = [\mathbf{metro}_1, \mathbf{metro}_2, \ldots, \mathbf{metro}_{|a|}]$), dónde $|a|$ denota la longitud de la secuencia del artículo relevante, que se considera un repositorio para decodificar posteriormente. Particularmente, $\mathbf{metro}_{cierto}$ y $\mathbf{metro}_{falso}$ denotar una verdadera representación contextual guiada por noticias y noticias falsas-representación contextual guiada, respectivamente.

3.2 Módulo decodificador controlado por categoría

Para hacer que el decodificador genere ejemplos verdaderos y falsos con más características diferenciadas por categorías, primero diseñamos una unidad de patrones compartidos para capturar las características compartidas dentro de la categoría (noticias verdaderas con sus artículos relevantes o noticias falsas con sus artículos relevantes). Simultáneamente, exploramos la restricción

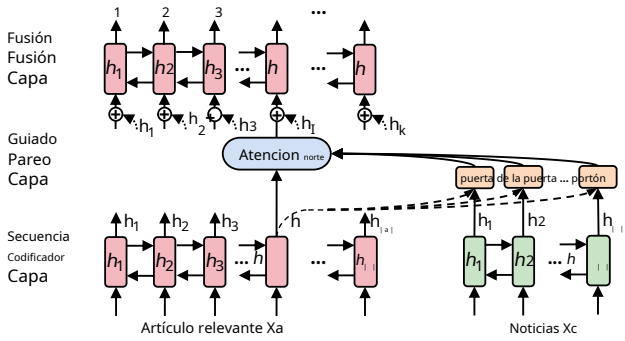


Fig. 3. La instantánea del módulo codificador guiado por noticias en el paso I.

unidad para restringir los dos tipos de características compartidas para hacerlas más diferentes, para obtener características diferenciadas entre categorías. A tiempo, el decodificador controlado generará más ejemplos con características diferenciadas entre noticias verdaderas y falsas.

3.2.1 *Unidad de patrón compartido*

Para adquirir las características compartidas dentro de las noticias verdaderas o falsas, respectivamente, diseñamos una unidad de patrón compartido que consta de celda de detección, celda de filtrado y celda de fusión, que elige características del ejemplo generado y la noticia. En particular, la unidad de patrón compartido para noticias verdaderas y la de noticias falsas.

son lo mismo. Tome el ejemplo generado en el paso de tiempo t como un ejemplo, los detalles de la unidad de patrón compartido se describen como:

Celda de cribado Adoptamos la transformación afín para filtrar la semántica de ews y la semántica generada para descubrir vectores eigen-invariantes.

$$S_{mi}(mi_{t-1}) = \mathbf{W}_{se} mi_{t-1} + \mathbf{B}_{se} \quad (10)$$

$$S_c(\mathbf{mi}_{piscina}) = \mathbf{W}_{Carolina\ del\ Sur} \mathbf{mi}_{piscina_c} + \mathbf{B}_{Carolina\ del\ Sur} \quad (11)$$

dónde $\mathbf{mi}_{piscina}$ denota el vector de agrupación máxima de noticias a t en-capa de codificación. $mi_{t-1} \in R_D$ son las incrustaciones del $(t-1)$ -la palabra predicha en el decodificador (en la subsección 3.2.2). $\mathbf{W}_{se} \in \mathbb{R}^{D \times D}$, $\mathbf{B}_{se} \in \mathbb{R}^{D}$, $\mathbf{W}_{Carolina\ del\ Sur} \in \mathbb{R}^{2D \times D}$, y $\mathbf{B}_{se} \in R_D$ son parámetros entrenables.

Impulsar la celda Para mejorar la capacidad no lineal de características semánticas, aprovechamos la función de activación - tanh para mapear la semántica de noticias y la semántica generada.

$$t_{mi}(mi_{t-1}) = \tanh(\mathbf{W}_{mi} mi_{t-1} + \mathbf{B}_{mi}) \quad (12)$$

$$t_c(\mathbf{mi}_{piscina}) = \tanh(\mathbf{W}_c \mathbf{mi}_{piscina} + \mathbf{B}_c) \quad (13)$$

dónde $\mathbf{W}_{mi} \in \mathrm{R}^{D \times D}$, $\mathbf{B}_{mi} \in \mathrm{R}^{D}$, $\mathbf{W}_{C} \in \mathrm{R}^{2D \times D}$, y $\mathbf{B}_{C} \in \mathrm{R}^{D}$ son parámetros entrenables.

Celda de fusión Para capturar las funciones compartidas dentro de una categoría (es decir, noticias verdaderas o falsas) de la semántica de las noticias y la semántica generada, construimos una celda de fusión para integrar la semántica eigen-invariante y la semántica no lineal tics.

$$\alpha = \sigma(\mathbf{W}_{ce}[S(mi_{t-1}); S(\mathbf{mi}_{piscina})] + \mathbf{B}_{ce}) \quad (14)$$

$$\mathbf{F}_t = \alpha_t(mi_{t-1}) + (1 - \alpha_t)(\mathbf{mi}_{piscina}) \quad (15)$$

dónde $\mathbf{W}_{ce} \in \mathrm{R}^{2D \times D}$, $\mathbf{B}_{ce} \in \mathrm{R}^{D}$ son parámetros entrenables. Especialmente, $\mathbf{F}_{cierto_t}$ (o $\mathbf{F}_{falso_t}$) son las características compartidas dentro de las noticias verdaderas (o noticias falsas) con sus artículos relevantes.

3.2.2 *Descifrador*

Después de aprender la representación contextual guiada por noticias, Utilice un decodificador GRU basado en la atención para generar una secuencia de palabras como ejemplo. Para hacer ejemplos verdaderos y falsos generados por el decodificador ricos en características de categoría, confiamos en las características intracategoría aprendidas de la unidad de patrón compartido para controlar la generación del decodificador. Concretamente, cuando generando el t-ésima palabra en el ejemplo, el decodificador emite un conjunto de todos los parámetros que se pueden aprender. Además, usamos búsqueda de haz con el ancho de haz 1 como algoritmo de decodificación durante la prueba.

vector de estado oculto $\mathbf{h}_t \in \mathrm{R}^{D}$ y pone una atención global sobre **metro**. La atención desea obtener representaciones indicativas de **metro** y los integra en un vector de contexto $\tilde{\mathbf{h}}_t$ definido como:

$$\mathbf{h}_t = \mathrm{GRU}([mi_{t-1}; \mathbf{F}_t; \tilde{\mathbf{h}}_{t-1}], \mathbf{h}_{t-1}) \quad (16)$$

$$\mathbf{C}_t = \mathrm{attn}(\mathbf{h}_t, \mathbf{metro}, \mathbf{W}_1) \quad (17)$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_2[\mathbf{C}_t; \mathbf{h}_t]) \quad (18)$$

dónde $t = 1, 2, \ldots, L_y$, y L_y es la longitud del conjunto de datos generado, es decir, A = {{GRAMO} ∪ {PAG}}. Más detalladamente, ejemplo. $mi_{t-1} \in R^{D}$ son las incrustaciones del (t−1) -th A = {{GRAMO$_F$} ∪ {PAG$_F$} ∪ {GRAMO$_T$} ∪ {PAG$_T$}}, dónde GRAMO$_F$ y GRAMO$_T$ palabra pronosticada en la que mi_0 ¿Las incrustaciones del inicio resienten ejemplos falsos y verdaderos generados, respectivamente, y PAG$_F$ simbólico. $\mathbf{F}_t \in \mathrm{R}^{D}$ son las salidas de la unidad de patrón compartido en el tiempo y PAG$_T$ representan todas las noticias falsas y verdaderas originales, respectivamente. paso t. En el lado de las verdaderas noticias, $\mathbf{F} = \mathbf{F}_{cierto_t}$ y en la falsificación lado de noticias, $\mathbf{F} = \mathbf{F}_{falso_t}$. $\mathbf{C}_t \in \mathbb{R}$ es el vector atencional, y $\tilde{\mathbf{h}}_t \in \mathrm{R}^{D}$ es el vector atencional en el paso de tiempo t.

La distribución de probabilidad predicha sobre la predefinida vocabulario V para el paso actual se calcula mediante:

$$Pr(y_t / y_{<t}, X_a, X_C, \mathbf{F}) = \mathrm{softmax}(\mathbf{W}\tilde{\mathbf{h}}_t + \mathbf{B}) \quad (19)$$

que refleja la probabilidad de que una palabra sea la t-th palabra en el ejemplo generado. Aquí, $y_{<t}$ se refiere a ($y_1, y_2, \ldots, y_{t-1}$). $\mathbf{W} \in \mathrm{R}^{V \times D}$ y $\mathbf{B} \in \mathrm{R}^{V}$ son pesas entrenables. En particular, <llamada de socorro> y <EOS> se aplican en la secuencia generada para habilitar el decodificador para comprender el principio y el final de una oración.

3.2.3 *Unidad de restricción*

Para hacer que los ejemplos verdaderos y falsos generados por decodificador contengan más características diferenciadas por categorías, es decir, explorando las características más diferenciales en noticias verdaderas y falsas (características intercategorías), diseñamos una unidad de restricción para restringir los dos tipos de características compartidas capturadas de las verdaderas o falso Noticias. En detalle, la unidad de restricción consta de dos celdas, es decir, celda restringida y celda de estandarización.

Celda restringida La celda restringida fomenta los dos tipos de funciones compartidas para que sean lo más diferentes posible.

$$L_{simi} = \sum_{I}^{D} \left(\frac{\mathbf{F}_{cierto_{ti}}}{\mathbf{F}_{falso_{ti}}} \right) \quad (20)$$

$$L_{diff} = 1 / L_{simi} \quad (21)$$

dónde D denota la longitud vectorial de ambos $\mathbf{F}_{cierto_t}$ y $\mathbf{F}_{falso_t}$ *mi*.

Célula de estandarización La estandarización de la celda alivia la correlación entre los dos tipos de características compartidas y asegura su independencia. Concretamente:

$$L_{rc} = \| \mathbf{F}_{cierto_t} - \mathbf{F}_{falso_t} \|_F^2 \quad (22)$$

dónde $\| \cdot \|_F^2$ es la norma de Frobenius al cuadrado [46].

Toda pérdida de la unidad de restricción podría integrarse como:

$$L_r = \alpha_1 L_{rc} + (1 - \alpha_1) L_{diff} \quad (23)$$

dónde α_1 es el hiperparámetro.

3.3 Capacitación

Durante la etapa de entrenamiento, aplicamos descenso de gradiente estocástico para minimizar la función de pérdida de nuestro modelo:

$$L = \sum_{norte=1}^{norte} \sum_{t=1}^{L_y} \text{Iniciar sesión}((Pr(y_t / y_{<t}, X_a, X_C, \mathbf{F}; \Theta)) + L_r) \quad (24)$$

dónde *norte* es el número de instancias de entrenamiento, y Θsignifica el todos los parámetros que se pueden aprender. Además, usamos búsqueda de haz con el ancho de haz 1 como algoritmo de decodificación durante la prueba.

3.4 Detección de noticias falsas

Después de la generación y optimización, ejemplos generados obtener características diferenciadas disponibles entre categorías de noticias egorías. Para mejorar la capacidad de detección, combinamos ejemplos borrados con datos originales para formar un nuevo aug-

Especialmente, la cantidad de ejemplos generados es la misma que la cantidad de noticias originales, es decir, generamos un nuevo ejemplo para cada noticia original.

A continuación, utilizamos el conjunto de datos aumentado A para verificar las noticias mediante redes de auto-atención de múltiples cabezas [47] que no solo capturan las dependencias globales de toda la secuencia, sino que también aprenden características de la estructura de la secuencia. Lo expresamos brevemente:

$$\mathbf{mi} = \text{uno mismo-atención}([\mathrm{PAG_I}; \mathrm{GRAMO_I}], \theta_{att}) \quad (25)$$

$$pag = \mathrm{softmax}(\mathbf{W}_{\mathrm{pag}}\mathbf{mi} + \mathbf{B}_{\mathrm{pag}}) \quad (26)$$

dónde [PAG$_I$;GRAMO$_I$] representa la concatenación de una noticia PAG$_I$ y un ejemplo generado GRAMO$_I$ en un artículo en UNA. **mi** es el vector de credibilidad aprendido por las redes de autoatención. θ_{att}, $\mathbf{W}_{\mathrm{pag}}$, y $\mathbf{B}_{\mathrm{pag}}$ son todos parámetros entrenables.

Finalmente, entrenamos las redes para minimizar la entropía cruzada. error para una sola instancia con etiqueta de verdad fundamental y:

$$L_{tarea} = -\sum_{y \text{Iniciar sesión} pag} \quad (27)$$

4 miXPERIMENTOS

En esta sección, para demostrar ampliamente la efectividad de CED, primero describimos tres conjuntos de datos de referencia públicos (es decir, Snopes, PolitiFact y PHEME) y entornos experimentales. Luego evaluamos sistemáticamente el desempeño de CED en estos conjuntos de datos en torno a las siguientes perspectivas: el desempeño del modelo, la ablación del modelo, la evaluación la utilidad de los ejemplos generados en diferentes líneas de base

métodos, el efecto de la cantidad de ejemplos generados y el proceso de formación. A continuación, visualizamos las características. aprendido por los módulos de CED para una comprensión más transparente para los usuarios finales. Finalmente, las limitaciones de CED se analizan de acuerdo con los resultados experimentales anteriores.

4.1 Conjuntos de datos

Evaluamos CED y demostramos su generalidad mediante el desempeño experimentos en tres conjuntos de datos públicos competitivos: Snopes, PolitiFact y PHEME. Sus detalles se muestran a continuación.

Snopes y PolitiFact son proporcionados por Popat *et al.* [10], que contiene 4.341 y 3.568 noticias, junto con 29.242 y 29.556 artículos relevantes recopilados de varias clasificaciones web mediante la aplicación de perturbaciones a la palabra incrustación. fuentes, respectivamente. Para las etiquetas de los dos conjuntos de datos: cada uno suena en redes neuronales recurrentes. las noticias en Snopes se etiquetan como verdaderas y falsas, mientras que cada noticia en PolitiFact se divide originalmente en uno de los siguientes seis modelos de decodificador para producir categorías de veracidad de voces inciertas o en conflicto: verdadero, mayormente verdadero, medio verdadero, mayormente falso, para presionar al discriminador para que aprenda un rumor más fuerte falso, y pantalones en llamas. Para capturar las diferentes características de noticias verdaderas y falsas, fusionamos verdadero, en su mayoría verdadero, y medio verdadero como verdadero, y las otras categorías se tratan como falsas.

PHEME [48] se aprovecha para la clasificación binaria de verdadero y falso con respecto a un tweet (noticias) a través de su relevante tweets (comentarios). El conjunto de datos incluye el aprendizaje de conversaciones de Twitter para representar evidencia coherente, así como su hilos asociados con nueve eventos de interés periodístico que contienen relación semántica con el reclamo de verificación del reclamo. Charlie Hebdo, disparos en Ottawa, etc. Un hilo de conversación En Snopes y PolitiFact, empleamos el 10% de los conjuntos de datos para consta de un tweet y una serie de comentarios. Filtramos el ajuste de los hiperparámetros y realizamos reclamos de validación cruzada de 10 veces con menos de 10 tweets y equilibramos el número en el resto de los conjuntos de datos. Recurrimos tomicro- / macro-promediado instancias de las dos categorías, es decir, 1.123 noticias verdaderas F1, puntuación F1 específica de la clase como métricas de evaluación. Además, nosotros y 1.123 noticias falsas (2.246 hilos de conversación). implementar nuestro modelo con Tensorflow[1]. Debido a que DeClarE no es

Dividimos los conjuntos de datos en entrenamiento, validación y pruebas de código abierto, los replicamos en función de sus estructuras de red con subconjuntos con la proporción de 70%, 10% y 20%, respectivamente. Theano[2]. Otras líneas de base se implementan a través de los códigos fuente que publicaron. Como se muestra en la Tabla 1, observamos que:

- En comparación con SVM, CNN y LSTM, DeClarE obtiene el mejor rendimiento, presentando al menos un 3,5% y un 3,2% de aumento en micF1 en Snopes y PolitiFact, respectivamente, lo que indica que DeClarE no solo aprende la semántica profunda del contenido de las noticias, sino que también captura palabras destacadas. de artículos relevantes para mejorar la representación de las características de credibilidad.
- El rendimiento de ADV-VIR es obviamente mejor que el de DeClarE, lo que expresa que la mejora de datos basada en restricciones podría elevar la capacidad de detección de los modelos. GAN-ED obtiene un rendimiento más notable que ADV-VIR, lo que explica que confiar en GAN para guiar el modelo de codificador-decodificador para generar características en conflicto para aumentar las representaciones indicativas podría mejorar de manera efectiva el rendimiento. dEFEND logra un desempeño más eminente que GAN-ED, lo que refleja la utilidad de la interacción entre noticias y comentarios. Además, HAN es superior a dEFEND, lo que ilustra que es eficaz para la detección considerar evidencia coherente y características comunes de artículos relevantes.
- CED supera consistentemente a los otros métodos de referencia, presentando 82.8% y 80.5% inmicF1 en Snopes y PolitiFact, respectivamente, y sus ventajas podrían expresarse en las dos perspectivas siguientes: 1) Nuestro modelo es capaz de fortalecer las representaciones de características indicativas de credibilidad mediante la utilización de el contenido de las noticias para guiar el modelo codificador-decodificador para capturar palabras destacadas de artículos relevantes; y 2) CED

1. https://www.tensorflow.org
2. http://deeplearning.net/software/theano

4.2 Configuración

Ajustamos estrictamente todos los hiperparámetros en el conjunto de datos de validación y obtenemos el mejor rendimiento a través de una búsqueda de cuadrícula pequeña. Los detalles de los hiperparámetros ajustados se muestran a continuación:
1) El modelo basado en BERT previamente entrenado [49] se utiliza para inicializar la secuencia de incrustaciones de noticias y artículos relevantes; 2) La dimensión D está configurado en 768; 3) Bi-GRU es la capa única y el tamaño oculto de GRU se asigna como 200; 4) Dado que la longitud de cada noticia y artículo relevante es diferente, aquí realizamos un relleno de ceros respectivamente estableciendo una longitud máxima;
5) El parámetro α_1 finalmente se entrena como 0,65, 0,60 y 0,80 en Snopes, PolitiFact y PHEME, respectivamente; 6) En auto-Las redes de atención, cabezas de atención y bloques se establecen en 6 y 4, respectivamente; 7) El abandono de la atención de varios cabezales se establece en 0,6; 8) Los regularizadores L2 con capas completamente conectadas, así como la deserción, se utilizan para el entrenamiento; 9) Todos los modelos se entrenan con el optimizador Adam [50] con una tasa de aprendizaje de 0,002 y un tamaño de mini lotes de 128 para minimizar la pérdida de entropía cruzada categórica; y 10) Reducimos la tasa de aprendizaje a la mitad cuando la perplejidad de la evaluación deja de disminuir. La detención anticipada se aplica cuando la perplejidad de la validación deja de caer para tres evaluaciones de generación continua.

4.3 Comparación de rendimiento

4.3.1 Resultados de Snopes y PolitiFact

Comparamos CED y las siguientes líneas de base de vanguardia en Snopes y PolitiFact:

SVM: Una SVM lineal adopta una colección de características lingüísticas creadas a mano en torno al contenido de noticias para la detección de noticias falsas [51].

CNN: Un modelo de CNN captura la semántica de las noticias a través de diferentes tamaños de ventana convolucional que sirven como *norte*-gramas para la detección de noticias falsas [52]. Aquí, solo consideramos el contenido de noticias sin características de metadatos.

LSTM: LSTM modela secuencias de palabras de noticias para aprender y representar la semántica para la verificación de hechos [53].

Declarar: Popat *et al.* [10] proponen un modelo de evaluación consciente de la evidencia para agregar señales de artículos relevantes, el lenguaje de los artículos y la confiabilidad de sus fuentes.

ADV-VIR: Un método basado en la argumentación de datos [54] extiende el entrenamiento de adversarios y adversarios virtuales al texto incrustación.

GAN-ED: Mamá *et al.* [35] desarrollar un codificador de estilo GAN Presentaciones indicativas para la detección de rumores.

defender: Las redes de co-atención de oraciones-comentarios [9] explotan tanto los contenidos de las noticias como los comentarios para capturar conjuntamente oraciones explicables top-k dignas de verificación para la detección.

HAN: Una red de atención jerárquica [13] se centra en

TABLA 1
Comparación de rendimiento en Snopes y PolitiFact

Métodos	Snopes				PolitiFact			
	micF1	macF1	Puntuación F1 Verdadero	Falso	micF1	macF1	Puntuación F1 Verdadero	Falso
SVM	0,704	0,649	0,511	0,786	0,658	0,623	0,516	0,710
CNN	0,721	0,636	0,460	0,812	0,654	0,610	0,505	0,745
LSTM	0,689	0,642	0,517	0,771	0,673	0,629	0,527	0,753
DeClarE	0,756	0,690	0,550	0,831	0,705	0,686	0,542	0,775
ADV-VIR	0,783	0,710	0,556	0,845	0,733	0,714	0,571	0,796
GAN-ED	0,792	0,746	0,567	0,851	0,771	0,743	0,602	0,809
dEFEND	0,801	0,750	0,581	0,860	0,778	0,748	0,611	0,813
HAN	0,803	0,753	0,646	0,864	0,783	0,751	0,622	0,821
Nuestro	**0,828**	**0,778**	**0,667**	**0,887**	**0,805**	**0,774**	**0,656**	**0,842**

TABLA 2
Comparación de desempeño de CED con las líneas de base en PHEME

Precisión de los métodos		Precisión		Recordar		Puntuación F1	
		Falso	Verdadero	Falso	Verdadero	Falso	Verdadero
Rango DT	0.562	0,588	0,549	0,421	0,704	0,491	0,617
DTC	0.581	0,582	0,579	0,473	0,788	0,478	0,584
SVM-TS	0,651	0,663	0,642	0,617	0,786	0,639	0,663
INCLINARSE	0,704	0,724	0,687	0,675	0,734	0,699	0,710
CNN	0,665	0,671	0,661	0,652	0,679	0,661	0,669
GRU	0,742	0,737	0,754	0,753	0,730	0,745	0,739
CVM	0,767	0,760	0,782	0,787	0,752	0,768	0,764
SHG	0,774	0,765	0,782	0,787	0,752	0,776	0,767
GAN-ED	0,781	0,773	0,791	0,796	0,766	0,784	0,778
defender	0,790	0,782	0,804	0,806	0,772	0,794	0,788
Nuestro	**0,803**	**0,795**	**0,814**	**0,819**	**0,788**	**0,807**	**0,801**

aprovecha las características basadas en categorías para guiar al codificador-decodificador a generar representaciones indicativas de categoría, destacando las diferencias entre noticias verdaderas y falsas.

4.3.2 Resultados en PHEME

Para evaluar más a fondo la efectividad de CED, llevamos a cabo experimentos en el conjunto de datos PHEME. En particular, debido a que nuestro modelo en este documento no tiene en cuenta las relaciones de asociación entre artículos relevantes, en lugar de utilizar el conjunto de datos FEVER que posee asociaciones entre artículos relacionados, explotamos tres conjuntos de datos, es decir, Snopes, PolitiFact y PHEME, que contienen Comentarios / artículos relevantes relativamente independientes, para evaluar nuestro CED. Comparamos nuestro CED con las siguientes líneas de base en PHEME:

Rango DT: El método de clasificación basado en el árbol de decisiones [55] identifica los rumores de tendencia a través de la búsqueda de grupos completos de publicaciones cuyo tema es una afirmación fáctica en disputa.

DTC: Un modelo de clasificador de árbol de decisión [56] utiliza una serie de características hechas a mano que incluyen características basadas en mensajes, basadas en usuarios, basadas en temas y basadas en propagación de tweets para evaluar la credibilidad de la información.

SVM-TS: Un modelo de clasificación de SVM lineal [57] captura las características temporales de la información del contexto social basándose en la serie temporal del ciclo de vida del rumor para su detección.

CNN: Un modelo basado en CNN [58] obtiene representaciones de rumores al enmarcar los tweets relevantes como secuencias de longitud fija.

GRU: Un modelo basado en RNN con GRU [59] aprende representaciones de tweets relevantes a lo largo del tiempo para la detección de noticias falsas.

CVM: El método de puntos de vista en conflicto [60] utiliza modelos temáticos para descubrir semánticas en conflicto y construye una red de propagación de credibilidad de tweets vinculados con relaciones de apoyo u oposición para generar resultados de evaluación.

SHG: StylizedHeadlineGeneration [45] genera titulares legibles y realistas para ampliar los datos de entrenamiento originales para mejorar la capacidad de clasificación de la detección de clickbait.

GAN-ED y **defender:** Se han introducido en la subsección 4.3.1.

Hacemos uso del 10% de los tweets en PHEME para ajustar los hiperparámetros, y el resto de las afirmaciones se realizan para una validación cruzada de 5 veces. Además, utilizamos exactitud, precisión, recuperación y puntuación F1 como métricas de evaluación.

Como se muestra en la Tabla 2, obtenemos las siguientes observaciones:

- En todas las líneas de base, las primeras tres líneas de base (es decir, DT-Rank, DTC y SVM-TS) basadas en características hechas a mano tienen un desempeño claramente peor que las otras líneas de base, lo que refleja una degradación de la precisión de entre 1,4% y 22,8%, lo que explica que la Los métodos extraídos automáticamente pueden descubrir más características indicativas de credibilidad ocultas que los métodos hechos a mano.
- En los métodos extraídos automáticamente, GRU logra un rendimiento más superior que BOWandCNNon PHEME, lo que ilustra que GRU es capaz de capturar características complejas ocultas indicativas de credibilidad más allá de patrones explícitos y superficiales. CVM supera a GRU en PHEME, lo que demuestra que excavar las características diferenciadas entre noticias y comentarios conduce a mejorar el rendimiento de las noticias falsas. GAN-ED y SHG aumentan respectivamente un 0,7% y un 1,4% el rendimiento en precisión en PHEME en comparación con CVM, lo que indica que los métodos extraídos automáticamente que dependen del modelo generativo producen más características de credibilidad para la detección de noticias falsas.
- Nuestro modelo supera constantemente a otras líneas de base en PHEME, presentando al menos un 15,2% de aumento que los métodos hechos a mano y al menos un 1,3% de aumento en la precisión que los métodos extraídos automáticamente, lo que confirma la efectividad de CED basándose en el modelo generativo de codificador-decodificador.
- Según la Tabla 1 y 2, el desempeño de las noticias falsas es más notable que el de las noticias verdaderas, mientras que el desempeño de las noticias falsas en PHEME no tiene un fenómeno similar. La razón puede ser que los artículos relevantes de noticias falsas en Snopes y PolitiFact contienen más características de credibilidad de las noticias falsas, mientras que los comentarios en PHEME incluyen más ruido irrelevante, como bromas irrelevantes y AD, que pueden interferir con la adquisición de características de credibilidad de los modelos.
- En comparación con la Sección 4.3.1 y 4.3.2, nuestro modelo emplea diferentes modelos de referencia para diferentes conjuntos de datos. La razón es que algunos modelos de referencia solo son apropiados para un conjunto de datos en particular, es decir, algunos modelos de referencia funcionan bien en ciertos conjuntos de datos y de manera temporal en otros conjuntos de datos.

4.3.3 Análisis de calidad de la generación de modelos en tres conjuntos de datos

Para analizar la calidad generada de CED, adoptamos tres

INCLINARSE: Una línea de base ingenua para la detección de noticias falsas se basa en métricas de evaluación, que incluyen Perplexity, BLEU-1 y BLEUon Bag-Of-Words para obtener la representación del texto de las noticias y 3, para evaluar el rendimiento de generación de nuestro modelo como clasificador de detección de construcciones con SVM lineal. así como los modelos de generación competitiva, es decir,**ADV-VIR**,

TABLA 3
Análisis de calidad de CED frente a varias líneas de base generadas

Métodos	Snopes			PolitiFact			PHEME	
	Perp.	B-1	B-3	Perp.	B-1	B-3	B-1	B-3
ADV-VIR	35,24	19,27	2,34	38,23	18,78	1,97	38,50	13,04 1,64
SHG	30,49	21,43	2,89	33,84	20,39	2,42	35,98	14,90 2,02
GAN-ED	25,72	24,56	3,26	27,61	22,16	2,76	32,20	16,85 2,45
Nuestro	**22,45**	**26,12**	**3,68**	**25,26**	**23,44**	**3,12**	**30,42**	**18,57 2,78**

TABLA 4
Resultados de la prueba de ablación de nuestro CED en los tres conjuntos de datos

Métodos	Snopes		PolitiFact		PHEME
	micF1	macF1	micF1	macF1	Precisión
- Noticias Enc.	0,804	0,760	0,783	0,758	0,786
- Cerrado	0,812	0,769	0,792	0,761	0,793
- Patrón	0,786	0,741	0,776	0,737	0,779
- restricción	0,795	0,753	0,783	0,755	0,787
- BERT Emb.	0,811	0,765	0,793	0,766	0,794
- SelfAtt.	0,813	0,769	0,798	0,769	0,798
Sección de la economía	0,828	0,778	0,805	0,774	0,803

SHG, y **GAN-ED**. La perplejidad es la medida estándar para evaluar modelos de lenguaje, y BLEU [61] mide las proporciones de las co-ocurrencias de*norte*-gramas entre el ejemplo generado y el artículo relevante original.

La Tabla 3 muestra el desempeño de todos los métodos comparados en los tres conjuntos de datos. Pudimos aprender que: Primero, entre los tres métodos generados, SHG y GAN-ED funcionan mejor que ADV-VIR. Los dos modelos tienen arquitecturas de red similares, es decir, dos arquitecturas codificador-decodificador, lo que confirma la eficacia de la arquitectura. Simultáneamente, también lo adoptamos a nuestro modelo como marco básico. En segundo lugar, en comparación con SHG, GAN-ED utiliza una estrategia de estilo GAN para capturar características conflictivas de artículos relevantes y generar secuencias de texto de mayor calidad, lo que indica que hay abundantes conflictos o semánticas controvertidas en los comentarios relacionados con las noticias (o artículos relevantes). En tercer lugar, nuestro modelo supera a todas las líneas de base con un gran margen, que refleja al menos un 0,42%,
0.36% y 0.33% de aumento en BLEU-3 en los tres conjuntos de datos, respectivamente. Una de las principales razones es que las líneas de base son la falta de restricciones en el discriminador o la capa de decodificación, mientras que nuestro modelo filtra la información de ruido y mejora la semántica valiosa con la ayuda de la unidad de patrón compartido en el decodificador.

4.4 Discusión

4.4.1 Análisis de ablación

Para evaluar la efectividad de diferentes componentes de CED, ablate CED en los siguientes modelos simplificados: 1) **-Noticias Enc.** 3) **-fusión** indica que la celda de fusión se reemplaza por concatenación indica que CED elimina los componentes del codificador de noticias de verdadero y operativo. Adicionalmente,**-Patrón** significa que la unidad compartida de patrones son módulos de noticias falsos, respectivamente; 2)**-Cerrado**significa que CED reemplaza eliminado de CED, que se ha introducido en la Sección 4.4.1. De mecanismo de puerta con concatenación; 3)**-Patrón** significa que CED elimina las unidades compartidas de patrones de los módulos de noticias verdaderas y falsas y adopta la operación de concatenación como forma de conexión; 4) **- restricción** representa que CED elimina la unidad de restricción para la detección de noticias falsas, 5) **-BERT Emb.** es que CED reemplaza las incrustaciones de BERT a las incrustaciones de word2vec como incrustaciones de entrada, y 6) **-SelfAtt.** es que CED reemplaza las redes de auto atención a BiLSTM para la detección de noticias falsas.

TABLA 5
Resultados de la ablación de la unidad de patrón compartido

Métodos	Snopes		PolitiFact		PHEME
	micF1	macF1	micF1	macF1	Precisión
- impulsar	0,809	0,766	0,797	0,763	0,797
- poner en pantalla	0,804	0,762	0,792	0,756	0,794
- fusión	0,798	0,758	0,785	0,745	0,787
- Patrón	0,786	0,741	0,776	0,737	0,779
Sección de la economía	0,828	0,778	0,805	0,774	0,803

Como se muestra en la Tabla 4, tenemos las siguientes observaciones:

- **Efectividad del codificador de noticias.** CED aumenta al menos un 1,7% el rendimiento que -News Enc., Lo que indica que la eficacia de las noticias guía el modelo codificador-decodificador para generar ejemplos para la detección de noticias falsas.
- **Efectividad de unidad cerrada.** Análisis de los resultados de -Gated y CED, CED obtiene un rendimiento superior apoyándose en la unidad cerrada, lo que ilustra la eficacia de CED utilizando el mecanismo de puerta para filtrar las noticias.
- **Efectividad de la unidad de patrón compartido.** Cuando se compara con -Pattern, CED mejora significativamente el rendimiento con la ayuda de la unidad de patrón compartido, lo que explica la efectividad de la unidad de patrón compartido que captura las características compartidas dentro de la misma categoría.
- **Efectividad de la unidad de restricción.** Al introducir la unidad de restricción, CED aumenta el rendimiento en comparación con - Restricción, que confirma que la efectividad del CED diseña la unidad de restricción para adquirir características diferenciadas entre categorías entre noticias verdaderas y falsas.
- **Efectividad de las incrustaciones de BERT.** Las incrustaciones de entrada de CED se reemplazan con incrustaciones de word2vec mediante incrustaciones de BERT, y el modelo logra un rendimiento más débil, lo que demuestra la eficacia del modelo utilizando incrustaciones de BERT como incrustaciones de entrada.
- **Efectividad del marco de CED.** Al reemplazar las redes de auto atención con BiLSTM para la detección de noticias falsas, CED refleja una reducción del rendimiento del 1,5% (micF1), 0,7% (micF1) y 0,5% (precisión) en los tres conjuntos de datos, lo que ilustra la eficacia de las redes de auto atención. Además, en comparación con las Tablas 1 y 2, CED con la ayuda de BiLSTM es superior a los últimos modelos de referencia (como HAN y dEFEND), lo que transmite la eficacia de nuestro marco CED.

4.4.2 Evaluación de la unidad de patrón compartido

La Tabla 5 proporciona el desempeño de cada parte de la unidad de patrón compartido por los siguientes modelos simplificados: 1) **-impulsando** significa que la unidad de patrón compartido elimina la celda de impulso; 2)**-poner en pantalla** implica que la unidad de patrón compartido elimina la celda de cribado; y Tabla 5, obtenemos las siguientes observaciones:

- La ablación de cualquier parte de la unidad de patrón compartido podría reducir el rendimiento de la unidad de patrón compartido, debilitando el rendimiento de 0,8% a 3,0% en micF1 en Snopes y PolitiFact y de 0,6% a 1,6% en precisión en PHEME, lo que demuestra la eficacia de cada parte de la unidad de patrón compartido.

- En comparación con el refuerzo, el cribado cumple con una reducción del 0,5% en micF1 en Snopes y PolitiFact, respectivamente, y una reducción del 0,3% en la precisión en PHEME, que presenta que para la unidad de patrón compartido, la captura de características invariantes eigen es más eficaz que el fortalecimiento de características específicas. Funciones para mejorar el rendimiento de la detección de noticias falsas.
- -La fusión refleja el peor desempeño en comparación con los modelos que solo eliminan una parte, lo que explica que la integración orgánica de la semántica de noticias y la semántica generada es capaz de capturar características compartidas de manera más efectiva para mejorar el desempeño de la detección.

4.4.3 Evaluación de la unidad de restricción

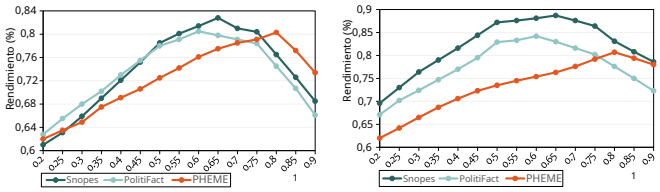
Para evaluar la contribución de cada celda de la unidad de restricción a las características diferenciadas entre categorías, probamos el rendimiento de CED con el parámetro α_1. Específicamente, nosotros tomamos α_1 de 0,2 a 0,9 en unidades de 0,05 en los tres conjuntos de datos y pruebe el rendimiento general (micF1 en Snopes y PolitiFact, y precisión en PHEME) del modelo y el desempeño específico en noticias falsas (es decir, el término False F1-score). Como se muestra en la Figura 4, observamos que:

- En general, cuando α_1 es inferior a 0,5, el rendimiento del modelo mejora con el aumento de α_1, pero el rendimiento del modelo en Snopes y PolitiFact es significativamente más rápido que eso en PHEME. Al mismo tiempo, cuando el modelo obtiene el rendimiento óptimo en tres conjuntos de datos, α_1 en Snopes es más similar al de PolitiFact, pero es bastante diferente al de PHEME, los valores específicos de α_1 en los tres conjuntos de datos son 0,65, 0,60 y 0,80, respectivamente. Estos reflejan, desde cierto aspecto, la similitud de la estructura de datos entre los conjuntos de datos de Snopes y PolitiFact, así como las diferencias con los de PHEME.
- Cuando el modelo alcanza el rendimiento más excelente en los tres conjuntos de datos, sus parámetros α_1 son todos superiores a 0,5, que indica que la celda de estandarización en la unidad de restricción contribuye más a la captura de la diferenciación entre categorías más características que la celda restringida. Además, cuando α_1 supera 0,75, el rendimiento del modelo cae drásticamente en Snopes y PolitiFact, que expresa que debilita el rendimiento, y cuando α_1 es menor que 0.3 y mayor que el papel de la celda restringida no es propicio para el desempeño de 0.7, el desempeño del CED se vuelve pobre, lo que muestra el del modelo. Tomados en conjunto, estos muestran que la efectividad de las dos células de nuestro modelo en la fusión de la codificación original se complementan entre sí, y su combinación orgánica podría generar características y características de información agregada para generar el máximo potencial. ejemplos. Además, en general, con el cambio de α_1,
- En el término de False F1-score, cuando α_1 es inferior a 0,5, el rendimiento del modelo no cambia significativamente, el rendimiento del modelo mejora rápidamente y cuando flota alrededor del 2%, lo que también muestra que la integración de α_1 es mayor que 0.5, su rendimiento aumenta lentamente en las funciones de secuencia y tiene un efecto limitado en el codificador guiado por noticias.

Snopes y PolitiFact. Sin embargo cuando α_1 va de 0,2 a 0,8, en segundo lugar, medimos el impacto de la deserción de múltiples cabezales exhibe una tendencia estable de ascenso en PHEME. Esto puede ser redes de auto-atención sobre el rendimiento del modelo, cambiamos porque las noticias falsas en Snopes y PolitiFact tienen más obvio el valor de abandono de 0 a 1 para obtener el valor experimental. características indicativas de credibilidad basadas en categorías, en comparación con las de PHEME.

4.4.4 La limitación del método que captura las características del conflicto

Seleccionamos al azar 200 elementos de tres conjuntos de datos, es decir, Snopes, PolitiFact y PHEME, y hacemos una comparación manual y encontramos que el 32%, 36% y 26% de los elementos de los tres conjuntos de datos no tienen conflictos en los artículos relevantes. Para confirmar la limitación de la detección de noticias falsas utilizando características en conflicto, realizamos experimentos para comparar el modelo CED y un método típico que se basa en las características de conflicto entre noticias y comentarios para detectar noticias falsas, AIFN [12]. Se muestran los resultados experimentales



(a) Cambios en el desempeño general de (b) Cambios en el desempeño de CED en CED con α_1. noticias falsas con α_1.

Fig. 4. Cambios de rendimiento de CED con α_1 de la unidad de restricción en Snopes, PolitiFact y PHEME.

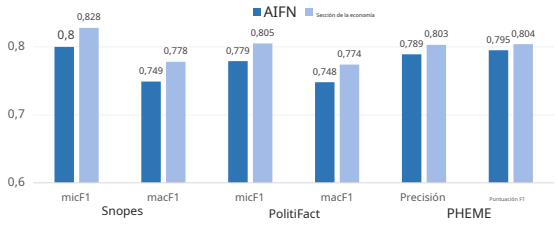
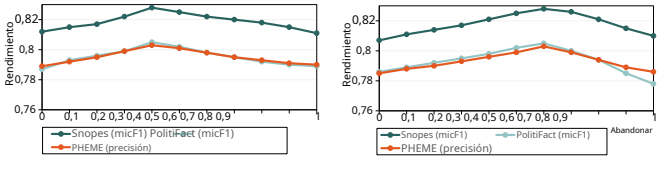


Fig. 5. Comparación de rendimiento entre nuestro CED y AIFN.

En la Figura 5, encontramos que nuestro modelo es obviamente mejor que el modelo AIFN, mostrando al menos 2.8% (micF1), 2.6% (micF1) y Rendimiento del 1,4% (precisión) en los tres conjuntos de datos, respectivamente, lo que confirma las ventajas de nuestro CED para capturar diferencias de categoría y las limitaciones de capturar la semántica de conflictos para la detección de noticias falsas.

4.4.5 El impacto de varios parámetros en el rendimiento del modelo

En esta sección, examinamos el efecto de λ en noticias guiadas codificador y abandono en el rendimiento del modelo. Primero nosotros evalúe la influencia de λ en el módulo codificador guiado por noticias medir el cambio de rendimiento del modelo de 0 a 1. Los resultados experimentales se muestran en la Figura 6 (a). Nosotros observa que cuando λ es 0.4, el modelo logra el mejor resultados del modelo, como se muestra en la Figura 6 (b). Observamos que, el valor de abandono de 0 a 1 para obtener el valor experimental.



(a) El impacto de λ en el modelo de desempeño (b) El impacto de la deserción de las redes de atención de desempeño personal en el modelo

Fig. 6. El impacto de λ y abandono de las redes de auto atención sobre el rendimiento del modelo

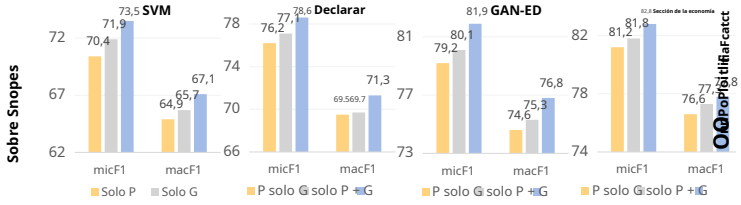


Fig. 7. Comparación de rendimiento de los modelos típicos basados en noticias originales y los ejemplos generados en Snopes.

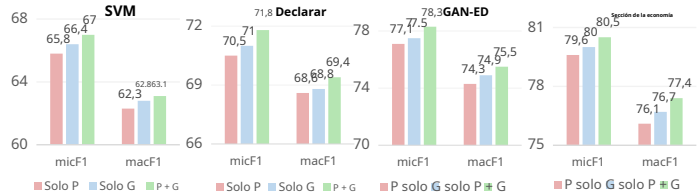


Fig. 8. Comparación de rendimiento de los modelos típicos basados en originales noticias y los ejemplos generados en PolitiFact.

Sobre PolitiFact

Al principio, el rendimiento del modelo mejora con el aumento de la deserción, lo que confirma que la utilización de la deserción para evitar el sobreajuste del modelo podría mejorar efectivamente el rendimiento de CED. En particular, cuando la deserción es 0,6, el rendimiento del modelo es óptimo. Después de eso, el rendimiento del modelo disminuye con el aumento de la deserción. Por lo tanto, elegimos un abandono de 0.6 para obtener el desempeño más excelente de CED.

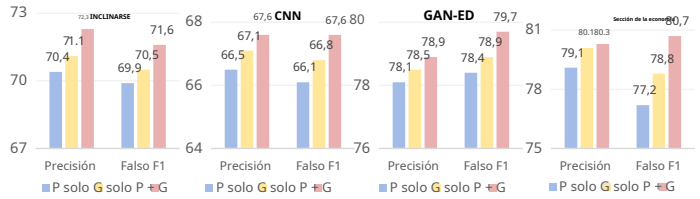


Fig. 9. Comparación de rendimiento de los modelos típicos basados en noticias originales y los ejemplos generados en PHEME.

En PHEME

4.4.6 Evaluación de la utilidad de los ejemplos generados

Para evaluar la utilidad de los ejemplos generados con ricas características diferenciadas entre categorías, realizamos experimentos Mentos en múltiples modelos en tres tipos de datos, es decir, solo los ejemplos generados de CED. Para comprender mejor el impacto de los ex- generados (solo G), y la integración de los datos originales y las amples en el modelo, realizamos dos grupos de experimentos con ejemplos generados (G + P). Los resultados experimentales son para analizar los cambios en el desempeño del CED con se muestra en las Figuras 7, 8 y 9. Observamos que:

- Los modelos logran un rendimiento más satisfactorio en los tres Los conjuntos de datos que usan solo los ejemplos de generación que solo el uso examinan el rendimiento general de CED usando la métrica del conjunto de datos original. Específicamente, los cuatro modelos se muestran en micF1 en Snopes y PolitiFact, y la métrica de precisión aumenta al menos un 0,4% en micF1. Esto ilustra que los ejemplos en PHEME, y el otro grupo (Figura 10 (b)) investiga generados por CED incluyen más intercategoría diferenciada el desempeño de la categoría específica (noticias falsas) utilizando características que el conjunto de datos original. la métrica de puntuación F1 falsa en los tres conjuntos de datos. Desde el
- En general, todos los modelos con ejemplos generados logran figura, obtenemos las siguientes observaciones: diferentes grados de mejoras que estos sin generar ejemplos ed. SVMobtains la mejora más obvia en Snopes y PHEME, y DeClarEwins aumenta un 1,3% inmicF1 en PolitiFact. Todos ellos elaboran la utilidad de los ejemplos generados y también transmiten que las mejoras del modelo provienen principalmente de los ejemplos generados.
- El rendimiento de la mayoría de los modelos es más destacado en Snopes que en PolitiFact. Por ejemplo, SVM (P + G) y DeClarE (P + G) logran un aumento de 3.1% y 2.4% en micF1 en Snopes que SVM (solo P) y DeClarE (solo P), respectivamente, mientras reflejan un aumento de 1.2% y 1.3% en PolitiFact, respectivamente. La razón principal se puede atribuir a las diferencias de clasificación. En detalle, PolitiFact originalmente divide las noticias en seis categorías, y las dividimos en dos categorías, es decir, noticias verdaderas y falsas, lo que da como resultado que algunas características semi-verdaderas y semifalsas se mezclen con noticias verdaderas o falsas para afectar la captura de características compartidas dentro de la misma categoría.
- Los modelos logran un aumento de rendimiento relativamente bajo en PHEME. Específicamente, los modelos obtienen una mejora de más del 2.5% (en micF1) en los dos primeros conjuntos de datos, mientras que solo reflejan una mejora de rendimiento promedio de aproximadamente 1.2% (en precisión) en PHEME. La razón puede ser que utilizamos comentarios de noticias en PHEME como insumos de CED, y su calidad de texto es más pobre que la del

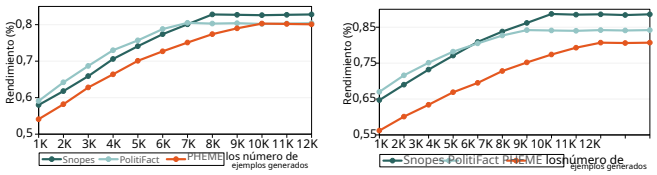
4.4.7 Efecto de la cantidad de ejemplos generados en CED

De la subsección 4.4.6, hemos aprendido que el generado ejemplos juegan un papel decisivo en la mejora del rendimiento del conjunto de datos original (solo P), solo los ejemplos generados de CED. Para comprender mejor el impacto de los ex- generados (solo G), y la integración de los datos originales y las amples en el modelo, realizamos dos grupos de experimentos con ejemplos generados (G + P). Los resultados experimentales son para analizar los cambios en el desempeño del CED con el número de ejemplos generados capturados de los tres conjuntos de datos, como se muestra en la Figura 10. Un grupo (Figura 10 (a)) examinan el rendimiento general de CED usando la métrica del conjunto de datos original. Específicamente, los cuatro modelos se muestran en micF1 en Snopes y PolitiFact, y la métrica de precisión aumenta al menos un 0,4% en micF1. Esto ilustra que los ejemplos en PHEME, y el otro grupo (Figura 10 (b)) investiga el desempeño de la categoría específica (noticias falsas) utilizando la métrica de puntuación F1 falsa en los tres conjuntos de datos. Desde el figura, obtenemos las siguientes observaciones:

- En la etapa inicial, es decir, el número de ejemplos generados es inferior a 6K, el rendimiento de CED continúa mejorando con el aumento del número de ejemplos generados, y el cambio de rendimiento general es generalmente consistente con el de una sola categoría (falso news), lo que muestra que el aumento de ejemplos generados proporciona continuamente al modelo más características diferenciadas entre categorías para distinguir la credibilidad de las noticias.
- A medida que aumenta el número de ejemplos generados, el modelo finalmente logra un rendimiento estable y óptimo. Cuando los ejemplos generados siguen aumentando, es difícil que el modelo continúe mejorando su rendimiento, lo que indica que CED ha obtenido las características indicativas de máxima credibilidad de los ejemplos, proporcionando una guía constructiva para el entrenamiento y la optimización del modelo.
- El número de ejemplos generados necesarios para que el modelo alcance un rendimiento estable es diferente, es decir, 8K, 7K y 10K en Snopes, PolitiFact y PHEME, respectivamente. Esto explica por un lado que el modelo de formación en PHEME es más laborioso que en Snopes y PolitiFact.

4.4.8 Análisis del proceso de formación

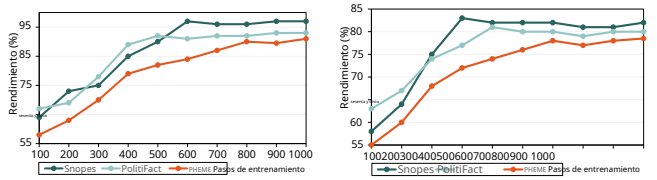
Además de la influencia de los aspectos anteriores en CED, también realizamos un estudio sobre el proceso de formación de los artículos relevantes de CED en Snopes y PolitiFact. en los tres conjuntos de datos. Durante el entrenamiento, un lote de datos es



(a) El desempeño en general cambios de CED con el número de ejemplos generados.

(b) Los cambios del desempeño específico formance de CED sobre noticias falsas con el número de ampollas.

Fig. 10. Cambios en el rendimiento de CED con el número de ejemplos generados en Snopes, PolitiFact y PHEME.



(a) Cambios de rendimiento de CED en el conjunto de entrenamiento en tres conjuntos de datos.

(b) Cambios de rendimiento de CED en conjunto de validación establecida en tres conjuntos de datos.

Fig. 11. Cambios de rendimiento de CED en el conjunto de entrenamiento y conjunto de validación en tres conjuntos de datos durante el proceso de entrenamiento.

siempre que se proporcionen, los pesos de nuestro modelo se actualizan en la dirección que minimiza la pérdida, que se define como un paso de entrenamiento. Analizamos el proceso de entrenamiento del modelo mediante la observación del rendimiento (en micF1 en Snopes y PolitiFact, en precisión en PHEME) de CED en el conjunto de entrenamiento y el conjunto de validación bajo pasos de entrenamiento incrementales. los Los resultados experimentales se muestran en la Figura 11. Claramente podríamos realizar experimentos para visualizar cómo se obtienen las siguientes observaciones controladas por categorías: el decodificador enriquece la semántica de los ejemplos generados para

- La precisión de nuestro modelo aumenta gradualmente y tiende a poseer más características diferenciales de categoría. La muestra se mantiene estable a medida que aumenta el número de pasos de entrenamiento. Específicamente, noticia verdadera y una noticia falsa de Snopes, nuestro modelo obtiene el mejor rendimiento en precisión y presenta algunos contenidos generados en las Figuras 13 y 14, 600, 500 y 800 pasos de entrenamiento en el conjunto de entrenamiento, y en aproximadamente 500, 400 y 700 pasos de entrenamiento en el conjunto de validación en los tres conjuntos de datos, respectivamente.
- CED es difícil de converger en la fase inicial de formación. En detalle, en el paso de entrenamiento de 300 en el conjunto de validación, nuestro modelo solo asegura un 75,69% (en micF1), 74,34% (en micF1) y 68,21% (en precisión) de rendimiento en Snopes, PolitiFact y PHEME, respectivamente. Especulamos que la unidad de patrón compartido de CED no ha aprendido suficientes características comunes dentro de la categoría, lo que conduce a características diferenciadas entre categorías no obvias.
- Nuestro modelo obtiene la convergencia más rápida en Snopes y PolitiFact (en aproximadamente 600 y 500 pasos de entrenamiento en el conjunto de entrenamiento, respectivamente), mientras logra una convergencia relativamente más lenta en PHEME (en aproximadamente 800 pasos de entrenamiento en el conjunto de entrenamiento). La razón podría ser que en Snopes y PolitiFact, los artículos de alta calidad con texto extenso, como el Hemos observado en la última subsección que los insumos generados del modelo, podrían capturar más características de credibilidad, ejemplos verdaderos y falsos son capaces de generar diferente credibilidad en el proceso de generación, comparado con la muy semántica. Explorar elocuentemente las diferencias en la generación. Los comentarios escasos con texto breve se utilizan como entradas del modelo en PHEME.

4.5 Estudio de caso

Para obtener una comprensión más transparente de las características capturados por nuestro modelo, visualizamos las salidas de noticias mapeadas en los valores correspondientes en $mi_{piscina}$ de noticias y el codificador guiado y el de decodificador controlado por categoría, respectivamente. Los detalles son los siguientes:

4.5.1 La visualización del codificador guiado por noticias

Para expresar intuitivamente las características que el CED ha aprendido de las noticias y artículos relevantes, llevamos a cabo experimentos para visualizar los resultados del codificador guiado por noticias. Específicamente, primero empleamos la operación de agrupación máxima para agrupar las salidas de la fusión de fusión capa *metro*$_{1}$ y luego mapearlos en los elementos correspondientes en la capa de entrada, respectivamente. Finalmente los patrones interesantes se obtienen y visualizan en la Figura 12. Observamos que:

- El codificador sin guía de noticias solo obtiene la semántica clave de los artículos relevantes (como 'sin oleada de crímenes' y 'sin amenaza al acecho'), mientras que el codificador que usa la guía de noticias no solo adquiere la semántica clave de los artículos relevantes, sino que también gana la semántica clave de las noticias, como "prohibir los coches negros" y "frenar el calentamiento global" (artículo relevante 1).
- El codificador que utiliza la guía de noticias puede centrarse en las partes más relevantes de las noticias a través de los artículos relevantes, como 'climas más cálidos' y 'muchas preguntas' en el artículo 2 relevante y 'frenar el calentamiento global' en la semántica de las noticias.
- Nuestro modelo también es capaz de capturar algunas de las características semánticas centrales ricas y diversas de artículos relevantes, lo que ayuda a revelar las partes incorrectas de la afirmación, por ejemplo, 'agencias de California' y 'regulaciones extralimitadas', mientras que el codificador sin guía de noticias solo asegura Palabras no clave que no están fuertemente relacionadas con la semántica de las noticias, por ejemplo, "competencia lanzada" y "agencias de California" (artículo 3 relevante).

4.5.2 La visualización del decodificador controlado por categorías

Para expresar intuitivamente las características que el CED ha aprendido de las noticias y artículos relevantes, llevamos a cabo experimentos para visualizar los resultados del decodificador controlado por categorías. Específicamente, primero empleamos la operación de agrupación máxima para agrupar las salidas de la fusión de fusión capa *metro*$_{1}$ y luego mapearlos en los elementos correspondientes en la capa de entrada, respectivamente. Finalmente los patrones interesantes se obtienen y visualizan en la Figura 13. Observamos que:

- Los ejemplos generados presentan una gramaticalidad débil, pero son relevantes para las noticias de entrada, como el contenido generado 'la cita errónea de clinton de giulian' y el contenido de noticias 'candidato dicho' en la Figura 13.
- El decodificador con control de categoría indica una semántica de credibilidad rica, por ejemplo, 'absurdo' y 'poco convincente' (en palabras rojas), mientras que el decodificador sin control de categoría solo se enfoca en la semántica clave, por ejemplo, 'musulmán' y 'no permitido' (en palabras grises)) en la Figura 14.
- Los ejemplos generados producen diferentes semánticas indicativas de credibilidad que apuntan a diferentes categorías de información, como 'será una broma' (palabras rojas) en noticias falsas y palabras azules 'sin rumor solo hecho' (palabras azules) en noticias verdaderas.

4.5.3 La visualización de la unidad de patrón compartido

Hemos observado en la última subsección que los insumos generados del modelo, podrían capturar más características de credibilidad, ejemplos verdaderos y falsos son capaces de generar diferente credibilidad en el proceso de generación, comparado con la muy semántica. Explorar elocuentemente las diferencias en la generación. de ejemplos verdaderos y falsos, visualizamos las características capturadas en unidades de patrón compartido que corresponden a noticias verdaderas y noticias falsas, respectivamente. Específicamente, según los tres conjuntos de datos, primero buscamos estos elementos con los valores más grandes de F_t en la celda de fusión de unidades de patrón compartido, entonces estos elementos son

Noticias: Californiaplanningbanlos coches negros orden bordillocalentamiento global	***Noticias:*** Californiaplanningbanlos coches negros orden bordillocalentamiento global
Artículo 1 pertinente: Describió que todavía tenemos que escuchar, pero incluso si eso resulta ser ***Artículo 1 pertinente:*** Descrito que todavía tenemos que escuchar, pero incluso si ese resulta ser el caso, claramente no hay onda delictiva, no hay peligro siempre presente para los automovilistas en todas partes, no, el caso, claramente, no hay onda del crimen, no hay peligro siempre presente para los automovilistas en todas partes, no hay amenaza al acecho en los lotes de estacionamiento de la centralización, la leyenda urbana de California está planeando la amenaza. theurbantheurbanlegendcalifornia está planeando prohibir los coches negros para frenar el calentamiento global de la respuesta de Snopes a prohibir los coches negros a fin de frenar el calentamiento global de la respuesta de Snopes.	
RelevantArticle2: ¿Son los coches blancos más fríos que los coches oscuros? ***RelevantArticle2:*** ¿Son los coches blancos más fríos que los coches oscuros? los climas tienen muchas preguntas acerca de sus autos, deberías vender tu auto azul oscuro porque los climas tienen muchas preguntas sobre sus autos, deberías vender tu auto azul oscuro porque será como anoven en Phoenix será como anoven en Phoenix	
Artículo 3 pertinente: en cualquier disputa hay una competencia campal entre california ***Artículo 3 pertinente:*** En cualquier disputa, existe una competencia campal entre las agencias de California para las cuales es la más absurda en su implementación de las agencias de regulación de alcance superior. regulaciones	
(a) Codificador sin guía de noticias	(b) Codificador que utiliza la guía de noticias

Fig. 12. La visualización del codificador con / sin guía de noticias sobre una noticia no verificada con tres artículos relevantes.

Noticias [falso]: El principal candidato demócrata dijo que la fuerza destructiva del libre mercado sin restricciones es la América moderna
Ejemplo generado: Los detalles sobre ***Ejemplo generado:*** El libre mercado es uno de estos problemas con respecto a Giulian, uno de nuestros mayores activos y Hillary dijo una vez que cita errónea de Clinton, estas palabras son el libre mercado sin restricciones es el más absurdo y poco convincente, será una broma ... fuerza destructiva en la América moderna ...
(a) El decodificador con control de categoría (b) El decodificador sin control de categoría

Fig. 13. La visualización del decodificador con / sin control de categoría sobre una noticia falsa.

Noticias [Verdadero]: dice roy moore dijo que practican la fe musulmana no miembro del congreso
Ejemplo generado: No hay rumor solo ***Ejemplo generado:*** Alabama dijo que un hecho que Roymoore dice que los musulmanes no deberían ser musulmanes practicantes no deberían poder servir en el congreso ... en el congreso por razón de su fe. actualmente enfrenta acusaciones de mala conducta sexual ...
(a) El decodificador con control de categoría (b) El decodificador sin control de categoría

Fig. 14. La visualización del decodificador con / sin control de categoría sobre una noticia verdadera.

palabras generadas en el decodificador y luego encontrar las palabras específicas correspondientes en la secuencia de noticias y la secuencia de decodificación. Los resultados visualizados se muestran en la Figura 15. Tenemos las siguientes observaciones:

- En la unidad de patrón compartido para noticias verdaderas, muchas palabras poseen una semántica cercana, como "de acuerdo" y "seguro". Igualmente, en la unidad de patrones compartidos para noticias falsas, varias palabras tienen significados similares:

ings, como 'lío'. Estos ilustran que nuestro modelo podría, aunque el CED supera a los métodos de vanguardia anteriores,

Conozca las características de credibilidad compartida dentro de la categoría dentro de los ejemplos verdaderos y falsos generados.

- Ambas unidades de patrones compartidos capturan palabras que son en su mayoría palabras de uso común, como 'obtener' y 'hecho' para noticias verdaderas, y 'obviamente' y 'pero' para noticias falsas, que rara vez involucran palabras de dominio o palabras específicas que reflejan una noticias específicas o un evento. Presenta que la unidad de patrón compartido podría obtener efectivamente las características eigen-invariantes.
- Ambas unidades de patrones compartidos adquieren palabras cuestionables y escépticas, como 'probablemente', 'correcto' y 'conjetura' para noticias verdaderas y 'sospechoso', 'increíble' e 'incorrecto' para noticias falsas, lo que indica que nuestro modelo es capaz de aprender características indicativas de credibilidad, y también confirma nuestro consenso de que tanto las noticias verdaderas como las falsas pueden ser cuestionadas.
- Las palabras más suaves y emocionales (como 'delicioso', 'alegre' y 'triste') se capturan en la unidad de patrón compartido para noticias verdaderas, mientras que la unidad de patrón compartido para noticias falsas aprende más palabras negativas o palabras extremas, como 'extremadamente ',' rabia 'y' maníaco ', lo que demuestra que la unidad de patrón compartido mejora efectivamente las características diferenciadas por categorías a nivel de emoción.

Múltiples perspectivas	Unidad de patrón compartido para True News	Unidad de patrón compartido para noticias falsas
Perspectiva de credibilidad	Sin duda; De acuerdo; Claro; Seguro; Sea correcto; Creer; Sí; Hecho	Pero; Obviamente; Seriamente; Realmente lío; Incorrecto; No poder; Absolutamente
Perspectiva del conflicto	Probablemente; Correcto; Adivinar	Sospechar; Increíble; Incorrecto; Confundir; Probablemente
Perspectiva de la emoción	Encantador; Agradecer; Alegre; Triste; Aburrido	Extremadamente; Rabia; Maníaco Desesperación
Perspectiva de estilo	Deberían; Considerar; Presumiblemente; Reportado	Impactante; Inesperadamente; Completamente; Absolutamente

Fig. 15. La visualización de unidades de patrones compartidos para noticias verdaderas y la de noticias falsas en tres conjuntos de datos.

- Algunas palabras con estilos diferentes son capturadas por ambas unidades de patrón compartido. Específicamente, la unidad de noticias falsas podría capturar algunas palabras relacionadas con estilos impactantes, como 'impactante', 'inesperado' y 'absolutamente', mientras que la unidad de noticias verdaderas se enfocaría más en palabras objetivas relacionadas con el estilo, como 'debería', 'considerar' y 'informado'.
- En general, existen diferencias entre las palabras capturadas en base a noticias verdaderas y falsas desde múltiples perspectivas, que no solo aprenden las características del conflicto a menudo capturadas por los modelos existentes, sino que también obtienen las diferencias desde las perspectivas de emoción extrema y estilos de escritura.

4.6 Análisis de errores

ods en tres conjuntos de datos (de la subsección 4.3), también descubrimos los siguientes defectos basados en los experimentos anteriores:

- De la subsección 4.5.3, se encuentra que CED puede capturar características indicativas de escepticismo compartido tanto de noticias verdaderas como falsas, pero no puede aprender con precisión las partes falsas del contenido de noticias no verificado, lo que indica que nuestro modelo es difícil de aplicar a la interpretabilidad de noticias falsas.
- En la subsección 4.4.8 se encuentra que nuestro modelo no solo logra un rendimiento más débil, sino que también requiere más tiempo de entrenamiento en PHEME en comparación con Snopes y PolitiFact. La razón es que PHEME está estructurado como 'noticias-comentarios', mientras que Snopes y PolitiFact están estructurados como 'artículos relevantes para las noticias', en los que los comentarios son inferiores a los artículos relevantes en términos de extensión, calidad del texto y relevancia con las noticias. Esto explica que nuestro modelo tiene un rendimiento relativamente bajo en términos de rendimiento y eficiencia computacional al tratar con comentarios de baja calidad en comparación con el tratamiento de artículos relevantes de alta calidad.
- Además, nuestro modelo está diseñado para la clasificación binaria de noticias falsas y verdaderas, y nuestro modelo no se puede aplicar directamente a la detección detallada de noticias falsas (como

como los seis tipos de noticias de PolitiFact). Para la detección detallada de noticias falsas, proponemos extender nuestro CED de dos tipos de ejemplos generados a múltiples tipos de ejemplos, y luego fortalecer las diferencias entre estos ejemplos por unidad de restricción.

5 Conclusión

En este artículo, propusimos un nuevo modelo de codificador-codificador (CED) controlado por categorías para generar ejemplos con características diferenciadas entre categorías (es decir, noticias verdaderas y falsas) para noticias falsas. detección de noticias, que aprendió características diferenciales entre categorías tures de características compartidas dentro de la categoría. Nuestro modelo diseñó una unidad de patrón compartido cerrado para fortalecer el aprendizaje de la representación de las características compartidas dentro de la categoría dentro de noticias verdaderas o falsas y desarrolló una unidad de restricción para forzar a los dos tipos de características compartidas dentro de la categoría a ser más diferentes para obtener inter-características diferenciadas por categoría. Verificamos el desempeño de CED en tres conjuntos de datos de referencia del mundo real, es decir, Snopes, PolitiFact y PHEME. A partir de estos experimentos, hemos observado que el CED superó los métodos de extracción profunda hechos a mano, superficiales y automáticos existentes. Observamos que los ejemplos generados pudieron mejorar eficazmente el rendimiento de diferentes tipos de métodos. También obtuvimos una comprensión transparente para los usuarios finales sobre las características capturadas por nuestro modelo a través de la visualización de los diferentes módulos de CED.

Este trabajo ha abordado la detección de noticias falsas en las redes sociales mediante el aprendizaje de la semántica del contenido de noticias y artículos relevantes para capturar las características compartidas dentro de la categoría para explorar características diferenciadas entre categorías. Si bien las redes sociales son ricas en información de metadatos, estudiaremos las siguientes direcciones en el futuro. Primero, mejoraremos nuestro marco CED fusionando múltiples características de metadatos, por ejemplo, perfil de usuario, para fortalecer la captura de las características dentro de una categoría. En segundo lugar, desarrollaremos variantes (semi) supervisadas para la unidad de patrón compartido, para aprender características compartidas detalladas y orientadas a categorías para diferentes tipos de información, en lugar de limitarnos a la clasificación binaria entre noticias verdaderas y falsas. Tercera,

Aagradecimientos

El trabajo de investigación está respaldado por el 'Programa nacional de investigación y desarrollo clave en China' (2019YFB2102300); 'las Universidades (Disciplinas) de Nivel Mundial y los Fondos de Orientación para el Desarrollo de Características para las Universidades Centrales de China' (PY3A022); Proyectos del Fondo del Ministerio de Educación (No. 18JZD022 y 2017B00030); Ciencia y tecnología de Shenzhen Proyecto gy (JCYJ20180306170836595); Gastos operativos de investigación científica básica de las universidades centrales (n. ° ZDYF2017006); Xi'an Navinfo Corp. y Centro de Ingeniería del Proyecto de Análisis de Datos Espacio-temporales de Inteligencia de Xi'an (C2020103); Proyecto de ciencia y tecnología del distrito de Beilin de Xi'an (GX1803).

Referencias

[1] A. Bovet y HA Makse, "Influencia de las noticias falsas en Twitter durante las elecciones presidenciales estadounidenses de 2016", *Nat. Comun.*, vol. 10, no. 1, pág. 7, 2019.

[2] LFBS-TN Grinberg, K. Joseph y D. Lazer, "Noticias falsas en Twitter durante las elecciones presidenciales estadounidenses de 2016", *Ciencias*, vol. 363, no. 6425, págs. 374–378, 2019.

[3] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff y B. Stein, "Una investigación estilométrica sobre noticias falsas y hiperpartidistas" *preimpresión de arXiv arXiv: 1702.05638*, 2017.

[4] D. Khattar, JS Goud, M. Gupta y V. Varma, "Mvae: Multimodal codificador automático variacional para la detección de noticias falsas ", en *Proc. La conferencia web*. ACM, 2019, págs. 2915–2921.

[5] T. Gröndahl y N. Asokan, "Análisis de texto en entornos adversarios: ¿El engaño deja un rastro estilístico? " *Computación ACM. Surv.*, vol. 52, no. 3, pág. 45, 2019.

[6] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu y H. Liu, "Unsupervised detección de noticias falsas en las redes sociales: un enfoque generativo ", en *Proc. AAAI*, 2019.

[7] F. Yang, SK Pentyala, S. Mohseni, M. Du, H. Yuan, R. Linder, ED Ragan, S. Ji y XB Hu, "Xfake: detector explicable de noticias falsas con visualizaciones", en *Proc. La conferencia web*. ACM, 2019, págs. 3600–3604.

[8] J. Ma, W. Gao y K.-F. Wong, "Detección de rumores en Twitter con redes neuronales recursivas estructuradas en árbol ", en *Proc. ACL*, 2018, págs. 1980–1989.

[9] K. Shu, L. Cui, S. Wang, D. Lee y H. Liu, "defiende: detección explicable de noticias falsas", en *Proc. SIGKDD*, 2019, págs. 395–405.

[10] K. Popat, S. Mukherjee, A. Yates y G. Weikum, "Declaran: Desacreditar las noticias falsas y las afirmaciones falsas mediante el aprendizaje profundo consciente de la evidencia ", en *Proc. EMNLP*, 2018, págs. 22–32.

[11] Y. Nie, H. Chen y M. Bansal, "Combinando extracción y verificación de hechos con redes de coincidencia semántica neuronal", en *Proc. AAAI*, vol. 33, 2019, págs. 6859–6866.

[12] L. Wu e Y. Rao, "Redes de fusión de interacción adaptativa para la detección de noticias falsas", en *Proc. ECAI*, 2020.

[13] J. Ma, W. Gao, S. Joty y K.-F. Wong, "Integración de pruebas a nivel de oraciones para la verificación de afirmaciones con redes de atención jerárquica", en*Proc. ACL*, 2019, págs. 2561-2571.

[14] L. Wu, Y. Rao, X. Yang, W. Wang y A. Nazir, "Redes de atención interactiva jerárquica consciente de la evidencia para la verificación de afirmaciones explicables", en *Proc. IJCAI*, 2020.

[15] K. Shu, A. Sliva, S. Wang, J. Tang y H. Liu, "Detección de noticias falsas en las redes sociales: una perspectiva de la minería de datos", *Proc. SIGKDD*, vol. 19, no. 1, págs. 22–36, 2017.

[16] B. Ghanem, P. Rosso y F. Rangel, "Un análisis emocional de información falsa en las redes sociales y artículos de noticias", *TOIT*, vol. 20, no. 2, págs. 1–18, 2020.

[17] Y. Chen, NJ Conroy y VL Rubin, "Contenido en línea engañoso: reconocimiento de clickbait como" noticias falsas "," en *Actas del ACM de 2015 sobre el taller sobre detección de engaños multimodal*, 2015, págs. 15-19.

[18] VL Rubin, N. Conroy, Y. Chen y S. Cornwell, "¿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*, 2016, págs. 7–17.

[19] W. Zhong, D. Tang, Z. Xu, R. Wang, N. Duan, M. Zhou, J. Wang y J. Yin, "Detección neural de deepfake con estructura fáctica del texto", en *Proc. EMNLP*, 2020, págs. 2461–2470.

[20] JP Baptista y A. Gradim, "Comprensión del consumo de noticias falsas: una revisión", *Ciencias Sociales*, vol. 9, no. 10, pág. 185, 2020.

[21] S. De Sarkar, F. Yang y A. Mukherjee, "Asistir a sentencias para detectar noticias satíricas falsas", en *Proc. COLING*, 2018, págs. 3371–3380.

[22] NT Tam, M. Weidlich, B. Zheng, H. Yin, NQV Hung y B. Stantic, "Desde la detección de anomalías hasta la detección de rumores utilizando flujos de datos de plataformas sociales" *Proc. Dotación VLDB*, vol. 12, no. 9, págs. 1016–1029, 2019.

[23] A. Giachanou, P. Rosso y F. Crestani, "Aprovechando las señales emocionales para la detección de credibilidad", en *Proc. SIGIR*. ACM, 2019, págs. 877–880.

[24] J. Ma, W. Gao y K.-F. Wong, "Detecta el rumor y la postura de forma conjunta mediante el aprendizaje neuronal multitarea", en*Proc. La conferencia web*, 2018, págs. 585–593.

[25] L. Wu, Y. Rao, H. Jin, A. Nazir y L. Sun, "Absorción diferente del mismo intercambio: aprendizaje multitarea tamizado para la detección de noticias falsas", *preimpresión de arXiv arXiv: 1909.01720*, 2019.

[26] T. Chen, X. Li, H. Yin y J. Zhang, "Llamar la atención sobre los rumores: redes neuronales recurrentes basadas en la atención profunda para la detección temprana de rumores", en *Proc. PAKDD*. Springer, 2018, págs. 40–52.

[27] N. Ruchansky, S. Seo e Y. Liu, "Csi: Un modelo híbrido profundo para la detección de noticias falsas", en *Proc. CIKM*. ACM, 2017, págs. 797–806.

[28] Q. Zhang, A. Lipani, S. Liang y E. Yilmaz, "Detección de información errónea asistida por respuesta a través del aprendizaje profundo bayesiano", en *Proc. La conferencia web*. ACM, 2019, págs. 2333–2343.

[29] K. Zhou, C. Shu, B. Li y JH Lau, "Detección temprana de rumores", en *Proc. NAACL*, 2019, págs. 1614-1623.

[30] K. Shu, X. Zhou, S. Wang, R. Zafarani y H. Liu, "The Role of User Profile for Fake News Detection", *arXiv: 1904.13355*, 2019.

[31] Q. Li, Q. Zhang y L. Si, "Detección de rumores mediante la explotación de la información de credibilidad del usuario, la atención y el aprendizaje multitarea", en *Proc. ACL*, 2019, págs. 1173-1179.

[32] K. Cho, B. Van Merriënboer, D. Bahdanau e Y. Bengio, "Sobre las propiedades de la traducción automática neuronal: enfoques de codificador-decodificador", *preimpresión de arXiv arXiv: 1409.1259*, 2014.

[33] R. Paulus, C. Xiong y R. Socher, "Un modelo profundamente reforzado para el resumen abstracto", *preimpresión de arXiv arXiv: 1705.04304*, 2017.

[34] A. Fan, M. Lewis e Y. Dauphin, "Hierarchical neural story generation", en *Proc. ACL*, 2018, págs. 889–898.

[35] J. Ma, W. Gao y K.-F. Wong, "Detecta rumores en Twitter mediante la promoción de campañas de información con aprendizaje generativo contradictorio", en *Proc. La conferencia web*. ACM, 2019, págs. 3049–3055.

[36] A. Gatt y E. Krahmer, "Encuesta sobre el estado del arte en la generación de lenguaje natural: tareas básicas, aplicaciones y evaluación", *J. Artif. Intell. Res.*, vol. 61, págs. 65-170, 2018.

[37] K. Sohn, H. Lee y X. Yan, "Aprendizaje de la representación estructurada de resultados mediante modelos generativos condicionales profundos", en *Proc. NeurIPS*, 2015, págs. 3483–3491.

[38] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk y Y. Bengio, "Aprendizaje de representaciones de frases mediante el codificador-decodificador rnn para traducción automática estadística", en *Proc. EMNLP*, 2014, págs. 1724-1734.

[39] J. Juraska, P. Karagiannis, KK Bowden y MA Walker, "Un modelo de conjunto profundo con alineación de ranuras para la generación de lenguaje natural secuencia a secuencia", en *Proc. NAACL*, 2018, págs. 152-162.

[40] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky y Y. Chi, "Deep keyphrase generation", en *Proc. ACL*, 2017, págs. 582–592.

[41] I. Sutskever, O. Vinyals y QV Le, "Secuencia a secuenciar el aprendizaje con redes neuronales", en *Proc. NeurIPS*, 2014, págs. 3104–3112.

[42] D. Bahdanau, K. Cho e Y. Bengio, "Traducción automática neuronal mediante el aprendizaje conjunto de alinear y traducir", *preimpresión de arXiv arXiv: 1409.0473*, 2014.

[43] J. Gu, Z. Lu, H. Li y VO Li, "Incorporando Mecanismos de Copia en el Aprendizaje de Secuencia a Secuencia", en *Proc. ACL*, 2016, págs. 1631-1640.

[44] W. Chen, Y. Gao, J. Zhang, I. King y MR Lyu, "Codificación guiada por título para la generación de frases clave", en *Proc. AAAI*, vol. 33, 2019, págs. 6268–6275.

[45] K. Shu, S. Wang, T. Le, D. Lee y H. Liu, "Generación de titulares profundos para la detección de clickbait", en *Proc. ICDM*. IEEE, 2018, págs. 467–476.

[46] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan y D. Erhan, "Redes de separación de dominios", en *Proc. NeurIPS*, 2016, págs. 343–351.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, AN Gomez, Ł. Kaiser e I. Polosukhin, "Atención es todo lo que necesitas", en *Proc. NeurIPS*, 2017, págs. 5998–6008.

[48] E. Kochkina, M. Liakata y A. Zubiaga, "Todo en uno: aprendizaje multitarea para la verificación de rumores", en *Proc. COLING*, 2018, págs. 3402–3413.

[49] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, "Bert: Preentrenamiento de transformadores bidireccionales profundos para la comprensión del lenguaje", en *Proc. NAACL*, 2019, págs. 4171–4186.

[50] DP Kingma y J. Ba, "Adam: Un método para la optimización estocástica", *preimpresión arXiv arXiv: 1412.6980*, 2014.

[51] J. Thorne y A. Vlachos, "Verificación automatizada de hechos: formulaciones de tareas, métodos y direcciones futuras", en *Proc. COLING*, 2018, págs. 3346–3359.

[52] WY Wang, "mentiroso, mentiroso jadeando: un nuevo conjunto de datos de referencia para la detección de noticias falsas", en *Proc. ACL*, 2017, págs. 422–426.

[53] H. Rashkin, E. Choi, JY Jang, S. Volkova y Y. Choi, "Truth of varying shades: Analyzing language in fake news and policy fact-check", en *Proc. EMNLP*, 2017, págs. 2931–2937.

[54] T. Miyato, AM Dai e I. Goodfellow, "Métodos de entrenamiento de adversarios para la clasificación de texto semi-supervisada", en *Proc. ICLR*, 2017.

[55] Z. Zhao, P. Resnick y Q. Mei, "Mentes inquisitivas: detección temprana de rumores en las redes sociales a partir de publicaciones de investigación", en *Proc. La conferencia web*, 2015, págs. 1395–1405.

[56] C. Castillo, M. Mendoza y B. Poblete, "Information credibility on twitter", en *Proc. La conferencia web*, 2011, págs. 675–684.

[57] J. Ma, W. Gao, Z. Wei, Y. Lu y K.-F. Wong, "Detecta rumores usando series de tiempo de información de contexto social en sitios web de microblogueo", en *Proc. CIKM*, 2015, págs. 1751-1754.

[58] F. Yu, Q. Liu, S. Wu, L. Wang y T. Tan, "Un enfoque convolucional para la identificación de información errónea", en *Proc. IJCAI*, 2017, págs. 3901–3907.

[59] J. Ma, W. Gao, P. Mitra, S. Kwon, BJ Jansen, K.-F. Wong y M. Cha, "Detectando rumores de microblogs con redes neuronales recurrentes", en *Proc. IJCAI*, 2016, págs. 3818–3824.

[60] Z. Jin, J. Cao, Y. Zhang y J. Luo, "Verificación de noticias mediante la explotación de puntos de vista sociales conflictivos en microblogs", en *Proc. AAAI*, 2016.

[61] K. Papineni, S. Roukos, T. Ward y W.-J. Zhu, "Bleu: un método para la evaluación automática de la traducción automática", en *Proc. ACL*, 2002, págs. 311–318.

Lianwei Wu actualmente está trabajando para obtener el doctorado en la Escuela de Ingeniería de Software de la Universidad Xi'an Jiaotong, China. Ha publicado más de 10 artículos de investigación en importantes conferencias y revistas internacionales, entre las que se incluyen: AAAI, IJCAI, ACL, EMNLP, ECAI, IEEE Transactions on Affective Computing y Information Sciences, etc. Sus intereses de investigación incluyen la evaluación de la credibilidad de la información, el análisis de redes sociales y procesamiento natural del lenguaje.

Yuan Rao recibió el doctorado de la Universidad de Xi'an Jiaotong, China, en 2005. Fue investigador postdoctoral en la Universidad de Tsinghua de 2005 a 2007. Fue profesor visitante en el laboratorio LTI. de la Universidad Carnegie Mellon de 2015 a 2016. Actualmente es profesor asociado y subdirector del departamento de análisis empresarial y tecnología, en la Escuela de Ingeniería de Software de la Universidad Xi'an Jiaotong. También es Director del Laboratorio de Inteligencia Social y Procesamiento de Datos de Complejidad (SICDP), Universidad de Xi'an Jiaotong. Sus principales intereses de investigación incluyen el procesamiento del lenguaje natural, el análisis de redes sociales y el aprendizaje automático.

Cong Zhang recibió una maestría de INSEEC, Francia, en 2015. Actualmente es asistente de recursos humanos en Xian Huanyu Satellite Control and Data Application Co. Ltd. Sus intereses de investigación incluyen el análisis de redes sociales y la minería de datos.

Yongqiang Zhao recibió la licenciatura en ingeniería mecánica de la Universidad de Ciencia y Tecnología de Taiyuan, Taiyuan, China, en 2017. Actualmente está trabajando para obtener la maestría en la Escuela de Ingeniería de Software de la Universidad Xi'an Jiaotong. Sus intereses de investigación incluyen visión por computadora, detección de objetos y subtítulos de imágenes.

Ambreen Nazir recibió la licenciatura y la maestría en ingeniería de software de la Universidad de Ciencia y Tecnología, Taxila, Pakistán en 2012 y 2014, respectivamente. Trabajó como profesora en Comsats Institute of Information and Technology, Pakistán durante 2 años. Actualmente está trabajando para obtener el doctorado en la Universidad de Xi'an Jiaotong, enfocándose en el análisis de sentimientos a nivel de aspecto y sus aplicaciones, y cómo avanzar hacia una forma de análisis de sentimientos más contextual y semántica.