Un framework para caracterizar Fake News en términos de emociones

A framework for characterizing Fake News in terms of emotions

Luis Rojas Rubio[1] https://orcid.org/0009-0000-6997-4928
Claudio Meneses Villegas[1, 2*] https://orcid.org/0000-0003-1112-4925

Recibido 09 de diciembre de 2024 aceptado 12 de Mayo de 2025
Received: December 09, 2024 Accepted: May 12, 2025

RESUMEN

Las redes sociales se han convertido en uno de los principales canales de información del ser humano debido a la inmediatez e interactividad social que ofrecen, permitiendo en algunos casos publicar lo que cada usuario considere pertinente. Esto ha traído consigo la generación de noticias falsas o Fake News, publicaciones que solo buscan generar incertidumbre, desinformación o sesgar la opinión de los lectores. Se ha evidenciado que el ser humano no es capaz de identificar en su totalidad si un artículo es realmente un hecho o bien una Fake News. Debido a esto es que surgen modelos que buscan caracterizar e identificar artículos basados en minería de datos y machine learning. Este artículo propone un framework de tres capas, cuyo principal objetivo es caracterizar las emociones presentes en Fake News y ser una herramienta que permita asociar el estado emocional y la intención más probable de quien pública una noticia falsa.

Palabras claves: Noticias falsas, emociones, caracterización de noticias falsas.

ABSTRACT

Social networks have become one of the main channels of information for humans due to the immediacy and social interactivity they offer, allowing in some cases to publish what each user considers pertinent. This has led to the generation of false news or Fake News, publications that only seek to generate uncertainty, misinformation or bias the opinion of readers. It has been shown that humans are not able to fully identify whether an article is really a fact or a Fake News. Due to this, models emerge that seek to characterize and identify articles based on data mining and machine learning. This article proposes a three-layer framework, whose main objective is to characterize the emotions present in Fake News and to be a tool that allows associating the emotional state and the most probable intention of the person who publishes a Fake News story.

Keywords: Fake news, emotions, characterization of fake news.

[1] Universidad Católica del Norte. Departamento de Ingeniería de Sistemas y Computación. Antofagasta, Chile. Email: larr014@gmail.com; cmeneses@ucn.cl
[2] Universidad Católica del Norte. Research Core in Artificial Intelligence and Data Science. Antofagasta, Chile. Email: cmeneses@ucn.cl
* Autor de correspondencia: cmeneses@ucn.cl

CC BY

INTRODUCCIÓN

El uso de redes sociales y la presencia de Fake News han generado un impacto negativo en la sociedad [1], generando incertidumbre y desinformación, anexo a esto se ha evidenciado que el ser humano no es capaz de identificar con alta certeza si un artículo es realmente un hecho o bien un Fake News, debido a esto, el problema de la desinformación asociado a la generación y propagación de Fake News se ha tornado relevante a nivel de estados y países que están siendo afectados por este fenómeno de origen humano pero agudizado por tecnologías que facilitan su incremento.

El estado del arte da cuenta de trabajos que buscan caracterizar de diversas formas las Fake News: características textuales [2], tiempos verbales [3], complejidad cognitiva [4], categorización de posturas [5], polaridad emocional [6], [7], entre otros.

La contribución de este artículo se basa en la formulación de un framework para la identificación de emociones presentes en publicaciones de redes sociales. Llevado al ámbito de las Fake News, este framework permite caracterizar no solo las emociones presentes, sino también como estas varían sus características con respecto a publicaciones noFake.

En este sentido, desde el punto de vista del propósito de la investigación, este trabajo busca responder las siguientes preguntas de investigación:

RQ1: *¿Cómo se puede incorporar a un texto que representa una noticia en general, emociones que se asocian a ciertas palabras presentes en el texto?*

RQ2: *¿Cómo se puede cuantificar el grado de emocionalidad de cada tipo de noticias, i.e. emocionalidad presente en Fake News versus las presentes en No Fake News?*

RQ3: *¿Cómo se puede contrastar las variaciones de las emociones expresadas en un conjunto de noticias correspondientes a un dominio versus las expresadas en otro dominio distinto?*

A continuación, el resto del artículo se organiza de la siguiente manera: en primer lugar, se describe la arquitectura general del framework propuesto, así como las capas modulares que lo integran.

Posteriormente, se presentan los resultados obtenidos al aplicar dicho framework a tres conjuntos de noticias pertenecientes a distintos dominios: Política, Celebridades y Covid-19. Luego se discuten los resultados alcanzados en relación con las preguntas de investigación previamente formuladas. Finalmente, se exponen las conclusiones del trabajo desarrollado, destacando sus principales limitaciones e identificando posibles líneas de investigación futura.

FRAMEWORK PROPUESTO

El Framework propuesto (*Figura 1*) busca identificar emociones en un conjunto de publicaciones, para ello se propone que su estructura corresponda a tres capas o módulos: Capa 1: Emocionalización, Capa 2: Cuantificación, y Capa 3: Caracterización.

Cada una de estas capas trabaja en función de las palabras utilizadas, el tipo de publicación y el dominio al que pertenecen.

Emocionalización

La primera capa del framework propuesto requiere de dos conjuntos de datos, NRC-Emotion Lexicon, un conjunto de datos que incorpora palabras, emociones y la intensidad con la cual se relacionan ambas [8]. El segundo conjunto de datos debe estar relacionado a las publicaciones que serán analizadas en función de emociones. Se asume una fase previa de preparación de datos de texto para dicho conjunto, exceptuando la aplicación de Lematización y Stemming, cuya aplicación no es apropiada en este caso.

En el 2019 Anoop, Deepak y Lajish presentan una forma de incorporar emociones a Fake News relacionadas al área de la salud con el propósito de mejorar sus modelos de clasificación [9]. El método implementado busca comparar cada palabra de una publicación con NRC-Emotion Lexicon. La emocionalización puede ser representada con la ecuación (1), donde W es una palabra de una publicación, E la emoción con la cual podría estar relacionada e I corresponde a la intensidad de la relación. Previamente se requiere establecer un umbral (τ) de modo que, si la intensidad de la emoción E respecto a la palabra W resulta ser mayor o igual a τ, entonces se agrega al corpus la palabra W seguida de la emoción E, en caso contrario solo se mantiene la palabra W, sin emoción agregada.

Fuente: Elaboración propia.

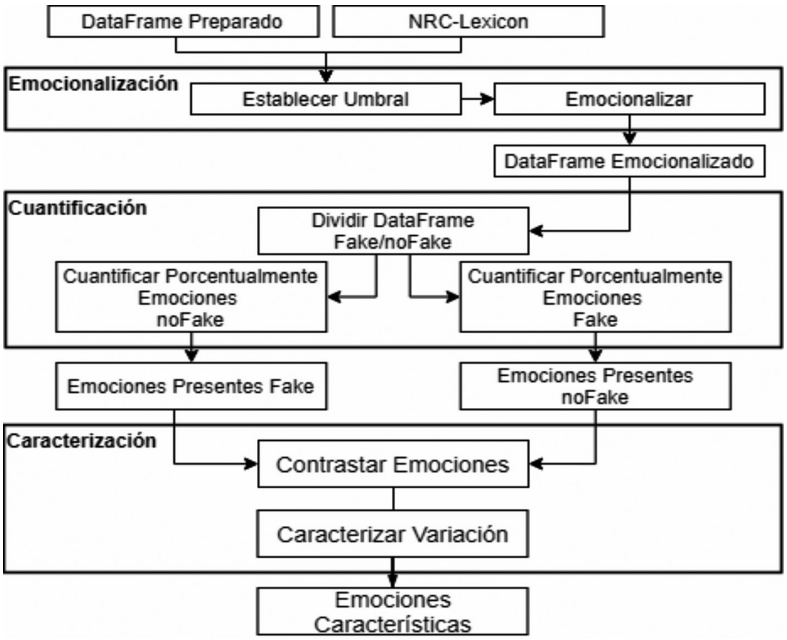


Figura 1. Arquitectura general del framework propuesto.

$$f_{(W,E,I)} = \begin{cases} W+""+E & I \geq \tau \\ W & I < \tau \end{cases} \quad (1)$$

Cuantificación

La capa de cuantificación inicia con el DataFrame emocionalizado generado por la capa de Emocionalización, el cual es dividido en dos nuevos conjuntos en razón al tipo de publicación; Fake y noFake. Posterior a ello, cada DataFrame es sometido a un proceso de cuantificación, la cual se puede expresar de acuerdo a la ecuación (2), donde $E_{(e,i)}$ representa la frecuencia de una emoción (e) en un tipo de publicación (i), y $W_{(i)}$ el número de palabras por tipo de publicación (i). Como salida de esta capa, se calcular una matriz con la frecuencia porcentual de las emociones presentes por tipo de publicación, esto para cada diferente dominio.

$$P_{(e,i)} = \frac{E_{(e,i)} * 100}{W(i)} \quad (2)$$

Caracterización

La capa de caracterización se compone de dos pasos: 1) Contraste de emociones y Caracterización de variaciones. El contraste de emociones puede ser cuantificado por medio del cálculo que se define en la ecuación (3), donde la variación de una emoción ($V_{(e)}$) se define como la diferencia de la cuantificación porcentual de la emoción e para publicaciones noFake ($P_{(e,'noFake')}$) y la cuantificación porcentual de la misma emoción e para publicaciones Fake ($P_{(e,'Fake')}$).

$$V_{(e)} = P_{(e,'noFake')} - P_{(e,'Fake')} \quad (3)$$

Por otra parte, la caracterización de variaciones permite identificar de manera gráfica las variaciones detectadas. Para ello se calcula una matriz de variaciones mediante el cálculo definido en la ecuación (4), donde:

$$C_{(e,\mu)} = \begin{cases} \downarrow & V_{(e)} > \mu \\ \uparrow & V_{(e)} < -\mu \quad |\mu > 0 \\ = & \mu \geq V_{(e)} \geq -\mu \end{cases} \quad (4)$$

- Si $V_{(e)} > \mu$, significa que una emoción e se presenta menos en publicaciones Fake.
- Si $V_{(e)} < -\mu$, se traduce como que una emoción e se presenta más en publicaciones Fake.
- Si $\mu \geq V_{(e)} \geq -\mu$, se interpreta como que una emoción en se presenta en igual medida en publicaciones Fake como en noFake.

APLICACIÓN DEL FRAMEWORK

A modo de validar el uso del framework presentado, se recopilan tres conjuntos de datos: PolitiFact, GossipCop y CoronaFake, donde el primer conjunto de datos se relaciona al dominio de la Política, el segundo a Celebridades y el tercero al Covid-19. Los tres conjuntos abordan tanto publicaciones Fake como noFake, siendo estas recopiladas desde *X* (ex Twitter). Cabe destacar que los tres conjuntos de datos han sido previamente preprocesandos, eliminando URLs, convirtiendo todo a minúsculas, eliminando las StopWords y removiendo los duplicados. La distribución de publicaciones posterior al procesamiento de datos puede ser apreciado en la Tabla 1.

En la Tabla 2 se puede apreciar el resultado de la cuantificación ($P_{(e,\ i)}$), evidenciando cuando frecuente es una emoción respecto a la totalidad de las palabras y tipo de publicación. En negritas se presenta el tipo de publicación de un dominio determinado donde una emoción es más frecuente. Por otra parte, se encuentran subrayados el tipo de publicación del dominio donde la emoción es más frecuente.

En la Tabla 3 se muestra la aplicación de la capa de caracterización, evidenciando el cambio de las emociones presentes entre noticias Fake y noFake de cada uno de los tres dominios explorados en este trabajo, es decir, Política, Celebridades y Covid-19.

Tabla 1. Distribución de publicaciones por Dataset.

Dataset	Fake	No Fake
PolitiFact	130.555	448.397
GossipCop	460.187	445.332
CoronaFake	575	584

Tabla 3. Resultados de la caracterización por dominio.

Emoción	Política	Celebridades	Covid-19
Anger	↑	↑	↑
Anticipation	=	↓	↓
Disgust	↑	↓	↓
Fear	↑	↓	↑
Joy	↓	↓	↑
Sadness	↑	↓	↑
Surprise	↑	↓	↑

DISCUSION DE RESULTADOS

En términos de los resultados expuestos en la aplicación del framework propuesto a tres conjuntos de noticias de tres dominios distintos, y en base a las preguntas de investigación formuladas para direccionar el trabajo experimental de la investigación, se puede evidenciar lo siguiente:

RQ1: ¿Cómo se puede incorporar a un texto que representa una noticia en general, emociones que se asocian a ciertas palabras presentes en el texto?

El framework propuesto aborda la incorporación o agregación de emociones al texto original mediante el uso del NRC-Emotion Lexicon, para buscar que emociones se encuentran fuertemente asociadas a una palabra, para lo cual se define un umbral de la intensidad para que la relación palabra-emoción pueda ser considerada significativa. En ese caso, a la palabra original se le agrega en el

Tabla 2. Resultados de la cuantificación por dominio y tipo de publicación.

	Política		Celebridades		Covid-19	
Emoción	noFake	Fake	noFake	Fake	noFake	Fake
Anger	0,57%	1,11%	0,72%	0,90%	0,33%	0,77%
Anticipation	0,42%	0,42%	1,32%	0,90%	0,45%	0,49%
Disgust	0,32%	0,93%	0,42%	0,38%	1,05%	1,00%
Fear	0,77%	1,84%	1,20%	1,02%	2,56%	3,05%
Joy	0,83%	0,69%	3,29%	2,22%	0,28%	0,37%
Sadness	0,48%	1,33%	1,01%	0,96%	1,63%	1,79%
Surprise	0,20%	0,32%	0,5%	0,31%	0,34%	0,47%

texto original la emoción con la cual está más fuertemente conectada y que supera el umbral definido. Esto se realiza tanto para noticias falsas (fake news) como para noticias no falsas (no fake news). Se experimenta con distintos umbrales de intensidad (e.g., 0.2, 0.4, 0.6, 0.8), encontrándose que empíricamente el umbral de intensidad 0.6 es apropiado para incorporar emociones significativas a las palabras que componen las noticias. Así, los resultados expuestos en las Tablas 2 y Tabla 3, se determinan en base a un umbral de intensidad de 0.6, en la relación entre palabras de cada noticia y las emociones listadas en Tabla 2 y Tabla 3.

RQ2: ¿Cómo se puede cuantificar el grado de emocionalidad de cada tipo de noticias, i.e. emocionalidad presente en Fake News versus las presentes en No Fake News?

Se define el porcentaje en que una determinada emoción *e* está presente en un tipo de noticia *i*, para un dominio determinado, como se formula en la ecuación (2). Cuando esto se aplicó a los tres diferentes dominios, se obtuvieron los resultados que se resumen en la Tabla 2. Aquí se identifica que, para el dominio Política, en las No Fake News es más predominante la emoción Joy que las otras emociones, pero en las Fake News predominan Fear and Sadness. En relación al dominio Celebridades, en las No Fake News son más predominantes las emociones Joy y luego Anticipation, pero en las Fake News se destaca Fear como diferenciadora respecto a las No Fake News en este dominio. En relación al dominio Covid-19, la emoción Disgust se destaca por tener mayor porcentual en No Fake News que su correspondiente valor en Fake News, pero en las Fake News sobresalen las emociones Fear y Sadness, en este dominio.

RQ3: ¿Cómo se puede contrastar las variaciones de las emociones expresadas en un conjunto de noticias correspondientes a un dominio versus las expresadas en otro dominio distinto?

Se define el cálculo de la variación porcentual entre emociones expresadas en Fake News versus No Fake News en un dominio particular, con lo cual se construye una matriz de esta variación porcentual en términos de si una emoción crece (↑), decrece (↓), o permanece en el rango $[-\mu, +\mu]$, definido por el parámetro $\mu>0$. En términos de los 3 dominios, las emociones Fear, Sadness y Surprise experimentan un crecimiento en los dominios Política y Covid-19, a diferencia del dominio Celebridades que experimenta un descenso, de las noticias Fake con respecto a las No Fake. Por lo tanto, en términos de estas 3 emociones, los dominios Política y Covid-19 exhiben similaridad entre ellos, pero son disímiles con respecto al dominio Celebridades (Tabla 3). A su vez, Celebridades es similar en la emoción Disgust con el dominio Covid-19, y con la emoción Joy con el dominio Política.

CONCLUSIÓN

En este artículo se ha propuesto un framework para la identificación de emociones en Fake News de un dominio determinado y se ha presentado su aplicación en tres dominios distintos. Principalmente, este trabajo contribuye al estado del arte con una metodología para detectar emociones en distintos dominios, reconociendo emociones específicas y no polaridades (positiva, negativa, neutra). Este framework podría traer beneficios tales como, mejoras en las tasas de clasificación de Fake News al incorporar las emociones claves como un elemento discriminador, detección temprana de Fake News al identificar la alta predominancia de una emoción asociada usualmente con Fake News en el contexto de un determinado dominio, identificar el estado emocional subyacente y asociado a una cierta probable intención de quien pública una Fake News o en otros ámbitos como la generación de campañas publicitarias, discursos o cualquier medio que requiera abordar emociones específicas.

Como trabajo futuro se espera incorporar al framework herramientas que permitan identificar no solo las emociones, sino también las intenciones de quien publica, buscando establecer relaciones cuantificables entre las ambas. Esto último requerirá definir nuevas métricas para medir la intensidad de la relación emoción versus intención en una fake news.

REFERENCIAS

[1] L. Rodríguez-Fernández, "Desinformación y comunicación organizacional: estudio sobre el impacto de las fake news", *Revista Latina de Comunicación Social*, no. 74, pp. 1714 -1728, doi: 10.4185/RLCS-2019-1406.

[2] B. Horne and S. Adali, "This Just In: Fake news packs a lot in title, uses simpler,

repetitive content in text body, more similar to satire than real news," *in Proceedings of the International AAAI Conference on Web and Social Media(ICWSM)*, vol. 11, no. 1, Montreal, Quebec, Canada, 2017, pp. 759-766, doi: 10.1609/icwsm.v11i1.14976.

[3] V. Perez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," 2017, arXiv: 1708.07104.

[4] N. Nguyen, G. Yan, M. Thai, and S. Eidenbenz, "Containment of misinformation spread in online social networks," in *Proceedings of the 4th Annual ACM Web Science Conference, Association for Computing Machinery*, Evanston, Illinois, USA, 2012, pp. 213-222, doi: 10.1145/2380718.2380746.

[5] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and Resolution of Rumours in Social Media: A Survey," *ACM Computer Surveys*, vol. 51, no. 2, Art. no. 32, 2018, doi: 10.1145/3161603.

[6] F. Qian, Ch. Gong, K. Sharma, and Y. Liu, "Neural user response generator: Fake News detection with collective user intelligence," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence,* (IJCAI), vol. 18, 2018, pp. 3834-3840, doi: 10.24963/ijcai.2018/533.

[7] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatio temporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171-188, 2020, doi: 10.1089/big.2020.0062.

[8] S. Mohammad and P. Turney, "Crowdsourcing a Word-Emotion Association Lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436-465, 2013, doi: 10.1111/j.1467-8640.2012.00460.x

[9] K. Anoop, P. Deepak, and V.L. Lajish, "Emotion cognizance improves health fake news identification," 2020, arXiv: 1906.10365.