Diseño de aplicación para visualización de tópicos de noticias sobre COVID-19 en lenguaje español

Lucas Exequiel La Pietra, Esteban Schab, Patricia Cristaldo, Anabella De Battista
Dpto. Ing. en Sistemas de Información. F.R. C. del Uruguay, Univ. Tecnológica Nacional
lelapietra@gmail.com, {schabe,cristaldop,debattistaa}@frcu.utn.edu.ar

Resumen

La evolución de la pandemia de COVID-19 ha tenido gran repercusión en medios periodísticos, en particular en diarios digitales. Las noticias publicadas por los mismos cubren diferentes aspectos relacionados con la evolución de casos y sus diferentes impactos. En este trabajo se presenta la aplicación de técnicas de procesamiento de lenguaje natural para la detección automática de los tópicos principales abordados por medios digitales, y la generación de una aplicación web que permite visualizar los resultados obtenidos en el proceso de topic modeling.

1. Introducción

La rápida propagación del virus *sars-cov-2* a nivel mundial desde los últimos meses de 2019 y durante todo el 2020 ha generado una gran atención pública y los medios de comunicación han presentado gran cantidad de publicaciones en relación al virus y al avance del brote en diferentes países. Los portales de noticias publican artículos periodísticos que brindan aportes sobre este asunto desde diferentes perspectivas.

Cuando se trabaja con volúmenes de textos muy grandes resulta de interés poder clasificar los documentos e identificar cuáles son los tópicos principales presentes en su contenido para poder interpretarlos de mejor manera. Para que esta tarea no resulte muy tediosa y lenta es necesario contar con alguna herramienta que permita automatizar el proceso.

En los últimos años se ha visto un crecimiento exponencial de datos vinculados a diversas disciplinas. Como consecuencia surge la necesidad de contar con nuevas estrategias para extraer conocimientos y relaciones ocultos en estos datos. El Modelado de Tópicos (Topic Modeling en inglés) ha surgido como un método eficaz para descubrir estructuras útiles en colecciones de documentos de textos. Para descubrir estas estructuras en textos han surgido diferentes modelos, uno de ellos es Latent Dirichlet Allocation (LDA) [1] cuyo enfoque asume que es posible dividir a cada documento (o grupo de documentos) en N categorías y, a partir de esa división, coloca cada palabra en base a las observaciones, en una de las categorías.

En este trabajo se presenta la aplicación de Topic Modeling a un corpus de texto en español, generado a partir de la recolección de noticias sobre COVID-19 del mes de mayo de 2020 de los periódicos digitales argentinos Clarín, Infobae y La Nación. El objetivo propuesto fue investigar los patrones subyacentes en los documentos y determinar las temáticas principales de las noticias publicadas por los medios durante esa etapa de la pandemia. Se realizó una comparativa de dos herramientas que implementan el algoritmo LDA y se realizaron las pruebas con la herramienta que obtuvo mejores resultados. Se desarrolló además una aplicación web en la que se presentan los hallazgos del proyecto a través de diferentes visualizaciones.

2. Descripción del problema

El Descubrimiento de Conocimiento en Bases de Datos (KDD por sus siglas en Inglés) se define en [2] como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos". Este es un proceso complejo que incluye la obtención de patrones (que es el objetivo de una etapa del proceso conocida como minería de datos), su evaluación e interpretación. La implementación de un sistema de KDD debe incluir la selección, limpieza y transformación de datos, el análisis de los mismos para la extracción de patrones, la interpretación de esos patrones para convertirlos en conocimiento y finalmente, difundir ese conocimiento para su posterior uso [3]. Una de las fases del proceso de KDD es conocida como Minería de Datos (MD) siendo su objetivo la generación de modelos a partir de la aplicación de métodos de aprendizaje automático y estadísticos.

La Minería de Textos (MT) es un subdominio de la MD cuyo objetivo es extraer información útil de colecciones de documentos a través de la identificación y exploración de patrones interesantes en datos textuales provenientes de fuentes heterogéneas, tales como como libros, páginas web, correos electrónicos, informes o descripciones de productos [4]. La principal característica de la MT es que trabaja con colecciones de documentos de texto no estructurado, generalmente en lenguaje natural, aunque podría ser también código fuente.

Existe una gran variedad de aplicaciones de minería de textos en diferentes dominios, como por ejemplo:

- Minería de opinión: una de las aplicaciones típicas consiste en explorar las opiniones de los usuarios para conocer su percepción sobre productos o servicios brindados por una compañía [5]–[7].
- Revisión sistemática de literatura: la aplicación de estas técnicas permite a investigadores encontrar información en búsquedas bibliográficas de una manera más rápida y eficiente. La aplicación de conceptos provenientes de la cienciometría permite optimizar las búsquedas bibliográficas. Existen varias librerías que brindan soporte para agilizar estas tareas para lenguajes como R o Python [8]–[10].
- Detección de plagios: existen propuestas tendientes a detectar patrones comunes entre documentos con el objetivo de determinar si se está ante una evidencia de plagio entre documentos [11], [12].
- Minería en redes sociales: algunas investigaciones tratan de verificar si es posible aplicar teorías sociales a los datos de las redes sociales y se han realizado pruebas a partir de la aplicación de técnicas de minería de textos que permiten detectar agrupamiento de usuarios, intereses en común, usuarios con mayor grado de influencia, entre otros [13], [14]. Los hallazgos de estas investigaciones tienen una amplia variedad de aplicaciones: Social Media Marketing, Gestión de la cadena de suministros, Minería de Opinión aplicada a nuevos productos o marcas, seguimiento de efectos de la pandemia de COVID-19 [15], [16], entre otros.

El objetivo de este trabajo es aplicar Modelado de Tópicos para descubrir las temáticas subyacentes en noticias de coronavirus publicadas en periódicos digitales nacionales.

3. Materiales y métodos

En el desarrollo de este proyecto se empleó la metodología CRISP-DM que brinda un enfoque detallado de las tareas y actividades a ejecutar en cada etapa. Para la recolección, procesamiento y análisis de la información se desarrolló código en Python. A continuación, se describen las distintas etapas, detallando las actividades realizadas, las herramientas utilizadas y los resultados obtenidos.

3.1. Comprensión del negocio

En esta etapa se concentraron los esfuerzos en comprender los objetivos y requisitos del proyecto desde una perspectiva general, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto.

El objetivo propuesto en este proyecto fue analizar las temáticas principales presentes en las publicaciones de periódicos nacionales sobre COVID-19.

Para cumplir con dicho objetivo se generó un corpus de noticias vinculadas a COVID-19 en lenguaje español obtenidas de los sitios web de tres periódicos nacionales: Clarín, La Nación e Infobae. Las noticias corresponden al mes de mayo de 2020. De cada noticia se guardaron como atributos: fecha, título, cuerpo de noticia, nombre del diario y categoría del diario en la que fue publicada la noticia.

La recolección de las noticias se realizó mediante la técnica de web scraping, para lo cual se desarrollaron scripts en lenguaje Python, que actualmente es ampliamente utilizado en proyectos de ciencia de datos. Para el desarrollo del código requerido para el proyecto se utilizó la IDE Spyder.

Los objetivos propuestos en términos de minería son: lograr la clasificación automática de noticias en base los tópicos principales presentes en sus textos e identificar las temáticas de más relevancia vinculadas a la pandemia sobre las que se publican noticias en periódicos nacionales.

3.2 Comprensión de los datos

Para la recolección de los datos requeridos en el proyecto se desarrolló una aplicación de web scraping que permitió recolectar noticias de los sitios web de los diarios Clarín, Infobae y La Nación.

Identificador del dataset	
Número de variables:	6
Número de observaciones:	1116
Valores perdidos:	14
Filas duplicadas:	0
Tamaño en memoria:	2.1MB
Tipos de variables	
Numéricas:	0
Categóricas:	2
Binarias:	0
Fechas:	1
URL:	1
Texto (único):	2
Tipo no soportado:	0

Tabla 1

Para el desarrollo del código fue necesario estudiar la estructura de las noticias de los tres diarios. Se recuperaron artículos sobre COVID-19 para el período 1 al 31 de mayo de 2020 y se almacenaron en un dataset.

Se recolectaron 1.116 noticias. De cada noticia se recuperaron y almacenaron seis atributos: título, cuerpo, link, fecha, diario y categoría del diario en la que fue publicada.

Una vez que se recolectaron los datos se realizó un análisis exploratorio del dataset generado. A continuación se presentan las características principales de los datos recolectados (ver Tabla 1).

En las Tablas 2, 3, 4, 5, 6 y 7 se puede observar el resumen del análisis exploratorio realizado sobre cada atributo del dataset.

1- Título	
Tipo:	Texto
Descripción:	Título del artículo.
Valores perdidos:	0
Valores distintos:	1116

Tabla 2

2- Link	
Tipo:	URL
Descripción:	URL del artículo.
Valores perdidos:	0
Valores distintos:	1116

Tabla 3

3- Cuerpo	
Tipo:	Texto
Descripción:	Cuerpo del artículo.
Valores perdidos:	14
Valores distintos:	1102

Tabla 4

4- Categoría	
Tipo:	Categórico
Descripción:	Categoría del diario a la que pertenece el artículo.
Valores perdidos:	0
Valores distintos:	49
Valor más repetido:	“El Mundo” (172)
Valor menos repetido:	“Viva” (1)

Tabla 5

5- Diario	
Tipo:	Categórico
Descripción:	Diario al que pertenece el artículo.
Valores perdidos:	0
Valores distintos:	3
Valor más repetido:	“La Nación” (430)
Valor menos repetido:	“Infobae” (313)

Tabla 6

6- Fecha	
Tipo:	Fecha
Descripción:	Fecha en la que se publicó el artículo.
Valores perdidos:	0
Valores distintos:	31
Valor más repetido:	“01/05/2020” (38)
Valor menos repetido:	“30/05/2020” (33)

Tabla 7

3.3 Preparación de los datos

Inicialmente se trabajó con un conjunto de noticias recolectadas entre los días 25 y 30 de abril. Se generó un dataset de prueba que permitió reconocer el formato de los

datos y validar la estructura del dataset para realizar la recolección de las noticias definitivas que son las correspondientes al mes de mayo.

Al diseñar la aplicación para web scraping de noticias, se consideraron los aspectos técnicos necesarios para que la información se almacenara en un formato acorde para evitar la mayor cantidad de operaciones vinculadas a la limpieza y preparación de los datos.

Gracias a haber tomado esos recaudos, la mayor parte de los datos resguardados en el dataset se encontraban en un formato adecuado para realizar el proceso de minería de datos. Solamente fue necesario realizar un proceso de limpieza y preparación de los títulos y la extracción manual de cuerpos de noticia para 14 registros obtenidos del diario Infobae.

3.4 Modelado

Para cumplir con el objetivo de proyecto se decidió implementar un modelo de detección de tópicos, utilizando el algoritmo Latent Dirichlet Allocation (LDA) [1]. Este enfoque considera a cada documento como un conjunto de palabras que pueden combinarse en conjuntos (o tópicos) formados por aquellas palabras que aparecen juntas en el texto con mayor frecuencia. Cada documento del corpus tiene una cierta probabilidad de ser clasificado en cada uno de los tópicos detectados.

Se realizó una comparativa de eficiencia entre el algoritmo LDA base de la librería GenSim y el provisto por la librería Mallet.

Se extendieron las stopwords del idioma español de la librería NLTK con palabras propias del tema a analizar, siendo estas 'coronavirus', 'pandemia', 'covid', 'covid19', 'virus', 'casos', 'personas' y 'cuarentena'.

Para realizar la comparación de los modelos se fijó la semilla de los algoritmos para obtener resultados consistentes y utilizando los mismos parámetros de corpus y diccionario, se crearon 7 modelos por algoritmo, siendo estos modelos de 2, 8, 14, 20, 26, 32 y 38 tópicos. Por último, se realizó el cálculo de coherencia para cada modelo, obteniendo los resultados que se muestran en el siguiente gráfico (ver Figura 1).

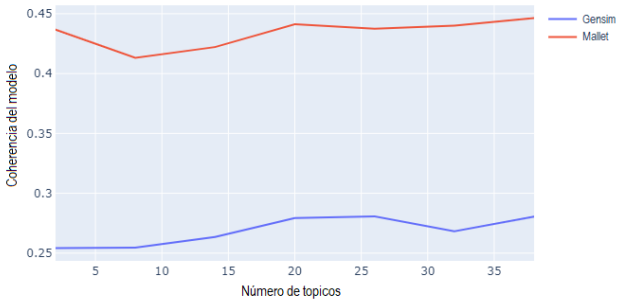

Figura 1

Como se puede apreciar, el algoritmo LDA provisto por Mallet brinda mejores resultados de coherencia para cualquier número de tópicos, por lo que se optó por utilizar esta librería.

Una vez tomada esta decisión, se fijó nuevamente la semilla para el algoritmo, y se crearon 10 modelos variando la cantidad de tópicos, siendo estos de 2, 4, 6, 8, 10, 12, 14, 16, 18 y 20 tópicos. Se realizó el cálculo de coherencia para cada modelo, obteniendo los resultados que se muestran en la Figura 2.

Se optó por utilizar 4 tópicos, los cuales son preprocesados e insertados en la base de datos a utilizar. Los tópicos fueron denominados "Actividades económicas", "Actividades recreativas", "Datos internacionales y avances en la pandemia" y "Estadísticas de la pandemia". Se decidió emplear ese número de tópicos por dos razones principales: por un lado el modelo final con esta cantidad de tópicos mostró uno de los valores de coherencia más alto, solo igualado por 16 tópicos; por otro lado se consideró que esta cantidad resultaba suficiente para categorizar correctamente los temas abarcados en las noticias.

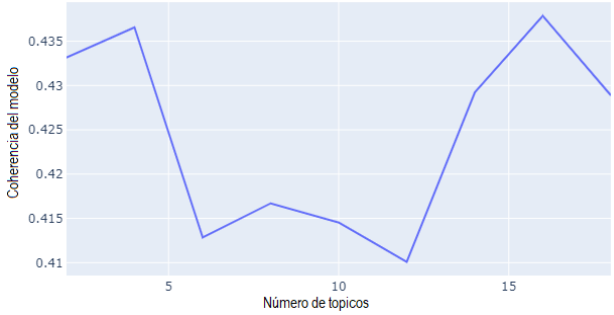

Figura 2

Se insertaron 3 nuevos atributos en el dataframe de noticias: el tópico al que pertenece la noticia, el nombre de dicho tópico y el grado de contribución que posee al mismo (ver Tablas 8, 9 y 10 respectivamente). Se creó un nuevo dataframe que contiene los 4 tópicos, la cantidad de noticias que lo refieren y el porcentaje de actuación sobre el total.

Topico_Dominante	
Tipo:	Integer
Derivado de:	Topic Modeling
Descripción:	Tópico dominante al que pertenece el artículo.

Tabla 8

Porcentaje_Contribucion	
Tipo:	Real
Derivado de:	Topic Modeling
Descripción:	Métrica dada por el algoritmo, que muestra qué tan representativa es el tópico para un artículo, o en otras palabras, que tanto contribuye el artículo a la constitución de un tópico.

Tabla 9

3.5 Evaluación

Actualmente se están desarrollando los métodos para evaluar la bondad del modelo de tópicos obtenidos.

Nombre_Topico	
Tipo:	Categórico
Derivado de:	Topic Modeling
Descripción:	Nombre dado al tópico

Tabla 10

4. Resultados

Una vez que se contó con el modelo para la detección de tópicos probado, se diseñó una aplicación web que, a través de distintas visualizaciones, permite conocer los tópicos subyacentes en el conjunto de noticias analizadas. Para su desarrollo se empleó el lenguaje Python, se empleó la librería Dash para el layout principal y Plotly para las visualizaciones. Se utilizaron componentes de las librerías HTML components, Core components y Bootstrap Components de Dash y se confeccionó una hoja de estilos a medida usando CSS.

En la primera versión de la app se implementaron las funciones para crear un histograma de frecuencia para una noticia, pudiendo elegir el diario, la categoría y la cantidad de palabras a mostrar (ver Figura 4).

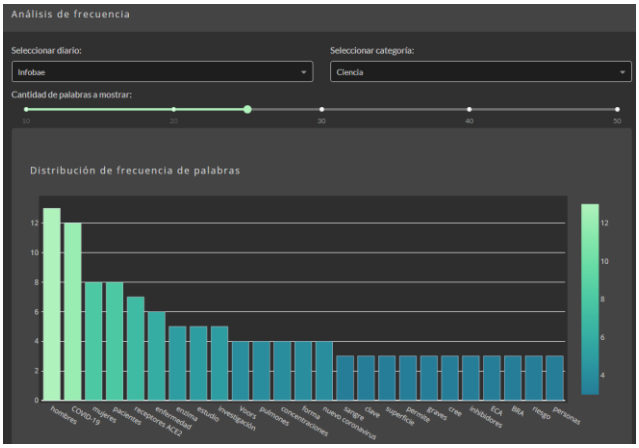

Figura 4

En la segunda versión se implementó la posibilidad de analizar un conjunto de noticias entre dos fechas, utilizando para esto un dataframe con las noticias minadas desde el 25 hasta el 30 de abril.

En la versión 3 se cambiaron los componentes y el layout de la página a componentes pertenecientes a la librería Dash Bootstrap Components, debido a su estética, capacidad de adaptarse a diferentes tamaños de pantalla y facilidad para crear layouts más complejos.

En versiones 3.1, 3.2 y 3.3 se agregaron gráficos que muestran la distribución del conjunto de datos entre los diarios y las categorías de los mismos y un scatterplot que trabajando con los mismos widgets que el histograma de frecuencia, muestra la frecuencia de n-gramas que presentan las noticias (ver Figuras 5 y 6).

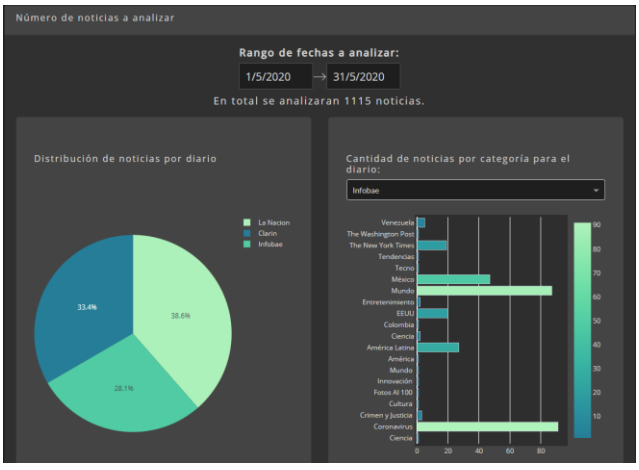

Figura 5

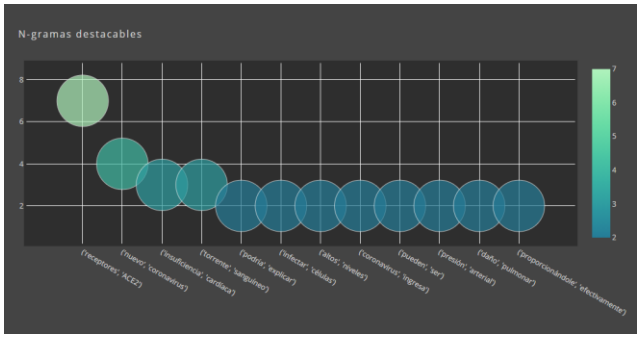

Figura 6

En la versión 3.4 y 3.5 se implementaron los dataframes y gráficos pertenecientes al proceso de Topic Modeling, siendo estos un gráfico circular que muestra la distribución de los tópicos, una tabla que presenta los tópicos encontrados, y un scatterplot parametrizable que muestra la colaboración de las noticias individuales a los tópicos (ver Figura 7).

En la versión 4 se corrigieron detalles estéticos como formatos, letras y tamaños, además de cambiarse la paleta de colores por una referente a los diarios estudiados.

En la versión 4.1 se agregaron los íconos correspondientes y se desactivaron las opciones de depuración. Esta es la versión final desplegada.

A partir del trabajo realizado, se puede concluir que la utilización de técnicas de procesamiento de lenguaje natural para el análisis de noticias resulta adecuada para captar tendencias y ver información qué pasaría desapercibida en un análisis superficial de las mismas. Cabe resaltar que las técnicas utilizadas, a pesar de no ser originalmente pensadas para el español, pudieron adaptarse y se logró consolidar un conjunto de herramientas que permiten analizar y encontrar patrones en grandes volúmenes de datos.

Figura 7

5. Conclusiones

En este trabajo se realizó la aplicación de la técnica de Topic Modeling para detectar automáticamente las temáticas principales abordadas por medios periodísticos digitales en relación a la pandemia de COVID-19. Para la detección de tópicos se empleó la técnica LDA, que se ejecutó mediante la librería Mallet que provee su implementación. Mediante la métrica de coherencia se determinó que con cuatro tópicos sería posible clasificar el conjunto de noticias.

Se desarrolló una aplicación de web scraping para realizar la recolección automática de noticias correspondientes al mes de mayo de 2020. En el diseño de esta aplicación se tomaron los recaudos necesarios para evitar tareas de limpieza y preprocesamiento de los datos. Se recolectaron 1.116 noticias que fueron almacenadas en un dataset en formato csv.

Para la publicación de resultados se desarrolló una aplicación web que se encuentra accesible en la URL https://nlp-noticiascorona.herokuapp.com y que permite visualizar distintos aspectos resultantes del procesamiento del conjunto de noticias como la cantidad de noticias clasificadas recolectadas por cada diario, la cantidad de noticias clasificadas en cada tópico, los bi-gramas y n-gramas de palabras que aparecen con más frecuencia en las noticias.

La ventaja principal de aplicar la técnica de topic modeling a este conjunto de noticias resulta de poder determinar los conceptos generales a partir de los cuales los medios están brindando información, generando una clasificación en cuatro aspectos principales como actividades económicas, recreativas, información vinculada a la pandemia a nivel internacional y estadísticas de la pandemia.

Como trabajo futuro se prevé la extensión de este trabajo a noticias de otras secciones de los diarios, para probar este tipo de técnicas en un conjunto con mayor cantidad de registros. Para lograr ese objetivo se está realizando la recolección de noticias de varias secciones de los tres diarios empleados en este proyecto desde el mes de marzo a la fecha.

Referencias

[1] D. Blei, A. Ng, and M. Jordan. *Latent dirichlet allocation*. J. Mach. Learn. 2003.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. *Knowledge discovery and data mining: towards a unifying framework*. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 82–88. 1996.

[3] J. Hernandez Orallo, M. J. Ramirez Quintana and C. F. Ramírez. *Introducción a la minería de datos*.1° Edición. 2008.

[4] M. Truyens and P. Van Eecke, “Legal aspects of text

mining", *Comput. Law Secur. Rev.*, vol. 30, n.º 2, pp. 153-170, abr. 2014, doi: 10.1016/j.clsr.2014.01.009.

[5] R. A. Laksono, K. R. Sungkono, R. Sarno and C. S. Wahyuni, "Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes," *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, Surabaya, Indonesia, 2019, pp. 49-54, doi: 10.1109/ICTS.2019.8850982.

[6] F. R. Lucini, L. M. Tonetto, F. S. Fogliatto, and M. J. Anzanello, "Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews", *J. Air Transp. Manag.*, vol. 83, p. 101760, mar. 2020, doi: 10.1016/j.jairtraman.2019.101760.

[7] X. Xu and Y. Li, "The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach", *Int. J. Hosp. Manag.*, vol. 55, pp. 57-69, may 2016, doi: 10.1016/j.ijhm.2016.03.003.

[8] M. E. Rose and J. R. Kitchin. "Pybliometrics: Scriptable bibliometrics using a Python interface to Scopus." *SoftwareX* 10 (2019): 100263. https://doi.org/10.1016/j.softx.2019.100263

[9] Bibliometrix R Package [Online]. Available: https://www.bibliometrix.org/ (accedido ago. 04, 2020).

[10] A. Usai, M. Pironti, M. Mital and C. A. Mejri, "Knowledge discovery out of text data: a systematic review via text mining", *J. Knowl. Manag.*, vol. 22, n.º 7, pp. 1471-1488, oct. 2018, doi: 10.1108/JKM-11-2017-0517.

[11] O. Hourrane and E. H. Benlahmar, "Survey of Plagiarism Detection Approaches and Big data Techniques related to Plagiarism Candidate Retrieval", in *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*, Tetouan, Morocco, mar. 2017, pp. 1–6, doi: 10.1145/3090354.3090369.

[12] K. Xylogiannopoulos, P. Karampelas and R. Alhajj, "Text Mining for Plagiarism Detection: Multivariate Pattern Detection for Recognition of Text Similarities", in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, ago. 2018, pp. 938-945, doi: 10.1109/ASONAM.2018.8508265.

[13] A. Akundi, B. Tseng, J. Wu, E. Smith, M. Subbalakshmi and F. Aguirre, "Text Mining to Understand the Influence of Social Media Applications on Smartphone Supply Chain", *Procedia Comput. Sci.*, vol. 140, pp. 87-94, ene. 2018, doi: 10.1016/j.procs.2018.10.296.

[14] J. Tang, Y. Chang and H. Liu, "Mining social media with social theories: a survey", *ACM SIGKDD Explor. Newsl.*, vol. 15, n.º 2, pp. 20–29, jun. 2014, doi: 10.1145/2641190.2641195.

[15] D. Li, H. Chaudhary and Z. Zhang "Modeling Spatiotemporal Pattern of Depressive Symptoms Caused by COVID-19 Using Social Media Data Mining". *Int. J. Environ. Res. Public Health* 2020, 17, 4988. https://doi.org/10.3390/ijerph17144988

[16] Q. Liu *et al.,* "Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach", *J. Med. Internet Res.*, vol. 22, n.º 4, p. e19118, 2020, doi: 10.2196/19118.