Análisis de categorías de noticias digitales para determinar tópicos cubiertos entre 2011-2015, mediante varios algoritmos de clustering en un dataset de tweets con temática de salud

Freddy L. Abad L., Edisson S. Reinozo T.

Ingeniería de Sistemas, Facultad de Ingenieria
Universidad de Cuenca
Cuenca, Ecuador

***Abstract*— En la actualidad, las redes sociales son una parte vital en la práctica del periodismo digital. Dado esto los medios de comunicación eligen la red social Twitter, para generar y compartir sus artículos. Esto genera una gran cantidad de datos que cubren diferentes situaciones del mundo real, como por ejemplo la salud. El uso de redes sociales como práctica periodística, plantean la necesidad de evaluar los tweets por tópicos principales para así evaluar a cada medio y la evolución de categorías tratadas en esta red social. Inicialmente estos tweets deben ser procesados por distintas técnicas de tratamiento de datos para establecer métricas de similitud y asentar las palabras más representativas. Posteriormente, el uso de distintos algoritmos de clusterización establecerá los mejores rendimientos respecto a los datasets procesados y la categorización de medios y noticias. Así este informe permite responder a varias preguntas metódicamente demostrable con datos reales.**

Keywords*— *Text Analysis, Text Mining, K Means, Spectral, Hierarchical, Twitter, Health

I. Introducción

En la coyuntura actual, el uso de técnicas de minería de texto ha tenido un auge para el entendimiento de grandes cantidades de texto generado por redes sociales [1]. Este auge se debe, en gran manera, a la madurez de varias técnicas de procesamiento de texto y datos, así como el desarrollo y popularización de varias técnicas de clusterización a través de algoritmos no supervisados [2]. Esta popularización a su vez se debe al acceso libre de las librerías que facilitan su ejecución [3], en el caso de este informe, el uso de estas técnicas se realizó mediante los distintos paquetes de Scikit Learn, NLTK como API de Python. Todo lo mencionado con anterioridad converge en la propuesta de este informe, en el cual se busca categorizar automáticamente por tweet y por medio de comunicación, los datos en formato de texto obtenidos de la red social de Twitter entre los años 2011-2015.

Este informe permite evidenciar la interacción generada por los medios de comunicación digitales y la evolución de temas en el tiempo. Además, define las categorías principales y su evolución en el tiempo, para concluir en el impacto y establecer la relación, de por ejemplo una enfermedad, en la cantidad de tweets que se generan por los medios.

Este informe inicialmente aborda los temas de preprocesamiento mediante el tratamiento y unificación de los archivos fuentes. Posteriormente se procesan los datos, mediante el tokenizado de frases, eliminación de stopwords, stemming y lematización para finalmente obtener los TF-IDF, y obtener las palabras más representativas de cada documento-tweet. Asimismo, se entrena los datasets procesado mediante los algoritmos de clusterización de K Means, Spectral y Hierarchical. En estos algoritmos se configuran y testean sus hyperparametros para obtener mejores rendimientos al momento de clusterizar tweets por medio de comunicación y por tiempo. La entrada del proyecto son tweets sin tratar, mientras la salida consiste en un dataset procesado y las categorías respectivas según el tiempo y el medio de comunicación.

A. Objetivo General

Categorizar los tópicos que cubren los medios de comunicación en el área salud en la red social Twitter y determinar si existe una variación de estos en el tiempo, mediante la aplicación de algoritmos no supervisados de clusterización.

II. Trabajos relacionados

El análisis de flujos de datos de redes sociales es el resultado de varias fases y tópicos de conocimiento, tales como el procesamiento y análisis de similitud textual, la identificación de eventos y subeventos en tiempo real, el uso de BOW (bag of words) y técnicas de enriquecimiento de texto y el uso de algoritmos de agrupamiento supervisado o no supervisado.

En referencia al análisis de similitud de texto, Pohl, et. al. exploró el problema de identificación de subeventos en tiempo real con datos de social media [4]. Además, propone la indexación y agrupación en línea de flujos de datos para generar informes situacionales, esto en un contexto de situaciones masivas como epidemias, huracanes, etc. Asimismo, en situaciones eventuales masivas, Sighn et. al. analiza el rendimiento de algoritmos no supervisados para el agrupamiento por similitud textual, específicamente, Simple K Means y Spectral K Means [5].

El uso de estos algoritmos, se ve contrastado por Xiangfeng et. al., que evita usar la técnica de BOW, ya que dificulta la comprensión semántica del texto. En su lugar, utilizan un método de agrupación basado en incrustaciones de palabras, esto en datos de tipo texto de tópicos de salud pública [6]. Refiere que la incrustación de palabras es una tendencia fuertemente usada en el procesamiento del lenguaje natural en la actualidad. Este método "permite aprender los vectores óptimos de palabras circundantes y los vectores pueden representar la información semántica de las palabras". Así un tweet puede representarse como unos pocos vectores y dividirse en grupos de palabras similares, y según su medida de similitud puede clasificarse como relacionado o no relacionado a una temática, como el ébola o la influenza. Este método presenta buenos rendimientos con precisiones pertenecientes al quintil más alto. Al igual que K Means, la incrustación de palabras es no supervisada, presentando la ventaja de no requerir etiquetado previo de los datos por humanos, de capacitación y puede extenderse fácilmente a otros problemas de clasificación u otras enfermedades.

Bajo la premisa de este informe, se considera importante la investigación de Alsayat et. al. [7]. Esto debido a su objetivo

de estudiar el comportamiento social humano mediante el análisis una gran cantidad de flujos de datos provenientes de redes sociales. Los autores de esta, apuntan la necesidad de distinguir entre usuarios regulares, medios digitales y opinión por líderes o personalidades públicas. Este análisis se debe a la distinción de uso del lenguaje, los medios de comunicación y los líderes de opinión regularmente, hacen uso de un mejor lenguaje, esto optimiza los procesos de stemming y lematización en la fase de procesamiento de datos. Partiendo de esta hipótesis, buscan mejorar la "granularidad de las comunidades de usuarios y su comportamiento mediante un marco para la detección de comunidades usando el algoritmo de agrupación de K-Means junto con el algoritmo genético y el método de distancia de reagrupación optimizada (OCD) para agrupar datos". Esta propuesta, resulta interesante al proponerse superar el problema de "K-Means para elegir los mejores centroides iniciales utilizando el algoritmo genético, así como maximizar la distancia entre los grupos mediante el agrupamiento por pares utilizando OCD para obtener grupos precisos". Esta propuesta, fue validada por los autores y ofrece mejores resultados de agrupación y proporciona un caso de uso novedoso de agrupación de comunidades de usuarios en función de sus actividades, esto optimizado y escalable para la agrupación en tiempo real de datos de redes sociales.

Finalmente, se contrasta el uso de K Means y Spectral K Means, con otros algoritmos de clustering, como el Hierarchical Clustering. La investigación de Hussain Shah et. al. [8], propone una metodología para la agrupación de nombres de marcas, basada en datos de la red social Twitter. Para tal propósito, proponen un algoritmo denominado BNACA, o Algoritmo de agrupación de aglomerados de nombres de marca, una extensión del algoritmo de agrupación jerárquica estándar. El algoritmo utiliza como enlace único la medida de similitud. Esta investigación presenta resultados relevantes y competitivos con el uso de algoritmos similares, como K Means o Spectral. Mostrando claramente en formato de dendograma las máximas similitudes entre marcas a través de los tweets.

III. METODOLOGÍA

Previo la ejecución de la metodología, se detalla las características del dataset a analizar a continuación.

A. *Descripción del dataset*

El dataset proviene del UCI ML Repository [9], con tematica "Health News in Twitter Data Set". Estos datos fueron recopilados en 2015, en un rango de años de 2011 a 2015, mediante la API de Twitter. Estos tweets refieren a noticias de salud física, mental y emocional de más de 15 agencias de noticias de salud como BBC, CNN entre otras. El número de instancias se define en la tabla 1.

TABLE I. CONTEO DE INSTANCIAS

Medio Digital	***Conteo***
Everyday Health	3239
CBC Health	3741
CNN Health	4061
Fox News Health	2000
GDN HealthCare	2997
Good Health	7864
Kaiser Health News	3509
LA Times Health	4171
MSN Times Health	3199
NBC Health	4215
NPR Health	4837
NY Times Health	6245
Reuters Health	4719
US News Health	1400
WSI Health	3200
TOTAL	**64117**

El dataset inicialmente fue procesado debido a inconsistencias que se tenían en los archivos fuentes. Una de las inconsistencias radica en la existencia del carácter separador de columnas "|". Los archivos tenían originalmente 3 columnas, sin embargo, por estos caracteres existentes en la 3era columna, muchas de las líneas se tomaban como 4 o 5 columnas, entorpeciendo la carga de datos. Esta inconsistencia, se replicaba en miles de líneas de al menos 7 archivos fuentes, para solucionarlo se verificó y arreglo manualmente. La segunda inconsistencia radica en tweets con comillas sin cerrar, al utilizar el stack provisto por Pandas [10] estas comillas, por defecto se tomaban como separadores, inclusive cuando se definen los caracteres separadores "|". La solución se redujo a la declaración explícita de omitir estas comillas. Para finalizar el preprocesamiento de los datos, se añadió una columna con la fuente de datos, es decir, el medio que compartió el tweet.

El resultado fue un dataset de 64117 tweets (ver Tabla I), este dataset se ordenó por timestamp de publicación y se contabilizo globalmente, es decir sin hacer discriminación de medios de comunicación, obteniéndose la Figura 1, la cual indica días de picos máximos y mínimos de publicacion global de tweets.

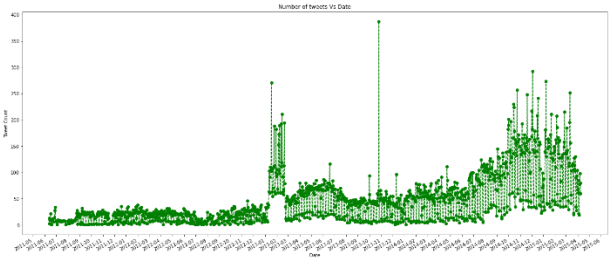

Fig. 1. Serie de tiempo del número de tweets por día en el rango de 2011-2015. (*Autoria propia*)

Así mismo, se realizó la Figura 2 de conteo de tweets por medio de comunicación por día. Por razones de espacio se muestran solo 4 medios, las gráficas con todos los medios se encuentran en: http://bit.ly/37r33ec. Los medios de la figura 2, muestran las divergencias en la regularidad y consistencia de publicación, por ejemplo, la figura 2.a el medio LA Times, presenta más regularidad de publicación, a diferencia del medio 2.b, NBC que muestra consistencia de publicación en varios rangos de tiempo, es decir, manejan un estándar de

publicaciones por día, sin embargo, existe un rango de tiempo en el cual no realizó publicaciones. Al igual, el medio 2.c y 2.d, Good Health y Fox News que muestra irregularidad de publicación, teniendo días picos de publicación y otros donde no se publica nada.

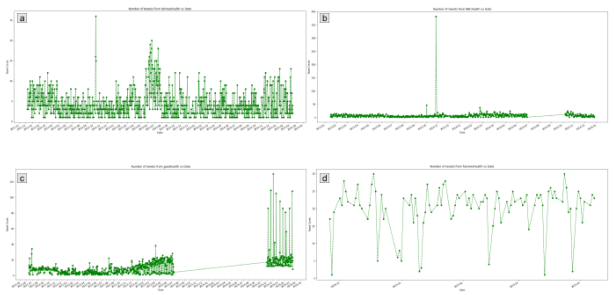

Fig. 2. Gráficas en el tiempo de numero de tweets de 4 medios de comunicación del total de medios en análisis. (*Autoria propia*)

B. Procesamiento de datos

El proceso para la ejecución de este proyecto se dividió en 3 fases:

1) Tokenización

La Tokenización es el proceso de convertir un texto dividiéndolo en las unidades que lo conforman, entendiendo por unidad el elemento más sencillo con significado propio para el análisis en cuestión, en este caso, las palabras. Este proceso se realizó mediante funciones primitivas de Python, además del proceso TfidfVectorizer [11] se realiza en adición.

2) Eliminación de stopwords

Las stopwords son palabras que aportan significado irrelevante al texto, incrementando la complejidad temporal al momento de realizar stemming y lematización, además de ralentizar los algoritmos TF-IDF para hallar la relevancia de palabras en cada documento. Estas palabras pueden ser "The", "and", "about", entre otros. Este proceso se realizó con el paquete incluido en TfidfVectorizer [11], y el paquete NLTK [12]. Además, previamente a este paso, se realizó la eliminación de palabras con poca relevancia, es decir, todas aquellas que no se consideren palabras, excluyendo "#" y "@", tales como RT abreviación de retweet.

3) Stemming y lematización del texto

Estos procesos permiten hallar la palabra raíz morfológica de cada palabra de un tweet, esto con el fin de disminuir tópicos de relevancia según las métricas de similitud establecidas por TF-IDF [11]. Este proceso se realizó mediante el paquete NLTK [12]. La relevancia de este paso radica, además, en el establecimiento de 3 datasets, uno solo ejecutando stemming, uno solo ejecutando lematización y uno ejecutando los dos procesos, con la finalidad de establecer el dataset con mejor rendimiento en este caso de estudio.

4) TF-IDF

El proceso TF-IDF calcula la relevancia, para mejorar las posiciones de contenido en las palabras clave de mayor relevancia [11]. Este proceso es radical para la exploración del dataset. Este proceso se realizó usando el paquete TfidfVectorizer [11] en 3 iteraciones, cada uno con distinto dataset, establecido en la etapa previa de stemming y lematización. Este proceso se complementa con el uso de Wordcloud en la etapa de exploración de datos, mediante el paquete WordCloud [13].

C. Clustering

El proceso de clustering se realizó con 3 algoritmos distintos: K Means, Spectral, Hierarchical. Esto con la finalidad de comparar su rendimiento relacionado a datasets de texto provenientes de redes sociales. La principal diferencia de estos algoritmos radica en:

- KMeans aplica análisis de conglomerados evaluando principalmente por distancia.
- Spectral analiza conglomerados evaluando principalmente por conectividad.
- Hierarchical aplica análisis de agrupación divisivos o aglomerativos para construir una jerarquía de agrupaciones [14][15].

Este proceso de clustering se realizó en diversas subfases, esto para evaluar el rendimiento con los distintos dataset procesados ya sea solo con stemming (d_stem), solo con lematización (d_lem) o con stemming y lematización (d_stem_lem). La Tabla II, muestra los k optimos establecios por:

- "Elbow curve" en el caso de KMeans con métrica inertia.
- "Dendograma" en el caso de Hierarchical con la métrica "Ward Linkage" que mide distancia entre clústeres a través de la suma de diferencia de cuadrados en los distintos clústeres [14].
- "Heurística de Eigengap" en el caso de Spectral Clustering con la métrica diferencia entre eigenvalor consecutivos [16].

Es importante mencionar, que Hierarchical Clustering presenta problemas de convergencia en datasets grandes. Al requerir el cálculo y almacenamiento de una matriz de distancia n × n, el costo computacional temporal y espacial es elevado [15]. En este proyecto, se evidenció al agotamiento de recursos de maquinas virtuales en Google Colab, como en maquinas locales. Así, para solucionar en parte, el dataset se redujo su dimensionalidad al 12% mediante SVD, así como la reducción de instancias al 10% randomicamente.

Spectral Clustering no tuvo un buen desenvolvimiento en este caso de estudio. Esto se debe a los pasos que necesita para establecer las matrices de similitud de distancias, para el establecimiento de Eigenvalores [15]. Estos pasos intermedios, complican la selección de k-optimos y posteriormente la evaluación de los modelos, por los costes temporales y espaciales (Google Colab agotó sus recursos con 100GB de memoria y 12 GB de RAM, situación similar en la máquina local, no permitiendo evaluar a este algoritmo).

TABLE II. K ÓPTIMOS ESTABLECIDOS PARA LOS DISTINTOS DATASETS

	d_stem	***d_lem***	***d_stem_lem***
K Means Elbow Curve Métrica "Inertia"	7	7	8
Spectral Clustering Heurística de Eigengap Metrica diferencia entre eigenvalue consecutivos	-	-	-
Hierarchical Clustering Dendograma Métrica "Ward Linkage"	6	9	5

La segunda fase consistió en la elección del mejor algoritmo de aprendizaje para cada dataset, en la Tabla III se visualiza los resultados obtenidos. La elección de los modelos optimos para cada algoritmo y por dataset, se realizó mediante el análisis a través del Wordcloud. Este análisis se realizó estableciendo algún sentido grupal a las palabras más relevantes de cada clúster. Estos resultados permiten establecer que el mejor algoritmo para cada dataset es: d_lem con K Means, d_stem_lem con Hierarchical.

TABLE III. SELECCIÓN DE ALGORITMO ÓPTIMO PARA LOS DISTINTOS DATASETS

	d_stem	***d_lem***	***d_stem_lem***
K Means Métrica "Inertia"	No Óptimo	Óptimo	No Óptimo
Spectral Clustering Metrica diferencia entre eigenvalue consecutivos	-	-	-
Hierarchical Clustering Métrica "Ward Linkage"	No Óptimo	No Óptimo	Óptimo

Finalmente, se establece el mejor modelo, es decir, aquel que tenga mejor rendimiento global algoritmo-dataset, K Means con el dataset de Lematización.

IV. RESULTADOS Y DISCUSIÓN

Seleccionado el modelo óptimo se procedió a establecer las etiquetas correspondientes a cada clúster, esto mediante la exploración de palabras más relevantes en cada clúster. En la Figura 3, se visualiza una muestra de 4 clústeres de los 7 finales (la figura completa se puede visualizar en http://bit.ly/3s0OKoq), los cuales mediante WordCloud [13] se puede establecer tópicos principales, el primero etiquetado como ébola, el segundo como seguro de salud, el tercero como fitness y el ultimo como tercera edad. El resultado de etiquetado se detalla en la Tabla IV.

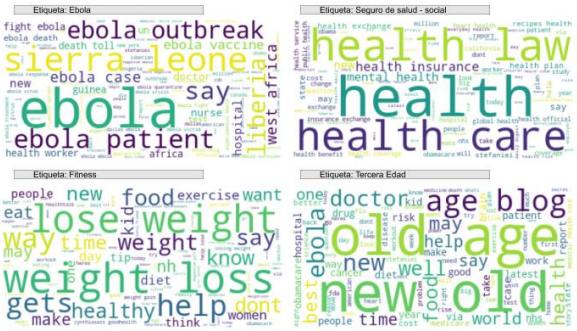

Fig. 3. Wordcloud para una muestra de 4 de los 7 clusters finales. (*Autoria propia*)

TABLE IV. SELECCIÓN DE ALGORITMO ÓPTIMO PARA LOS DISTINTOS DATASETS

Clúster	**Etiqueta**
Cluster 0	Bienestar
Cluster 1	Seguros de Salud-Social
Cluster 2	Medicamentos
Cluster 3	Fitness
Cluster 4	Investigaciones médicas
Cluster 5	Tercera edad
Cluster 6	Ebola

Asimismo, se realizó un análisis mediante el modelo entrenado, para establecer los tópicos principales tratados por los medios. Los tópicos se detallan en la Tabla V, se establecen 3 tópicos principales, colocados descendentemente por grado de relevancia. A forma de verificación, se puede notar que el medio Kaiser Health News está formado por Bienestar, Seguros de Salud, Tercera Edad. Este medio de comunicación pública "periodismo sobre temas de atención médica concerniente especialmente a personas con bajos ingresos, vulnerables al costo de la atención médica, como las personas *sin seguro*, las personas con enfermedades crónicas o Beneficiarios de Medicaid-Medicare" (programa de cobertura de seguridad social estatal de EEUU para todas las *personas mayores de 65 años*) [17] [18].

TABLE V. TOPICOS PRINCIPALES POR MEDIO DE COMUNICACIÓN

Medio Digital	***Tópicos Principales***
Everyday Health	Bienestar, Fitness, Tercera Edad
CBC Health	Bienestar, Ébola, Seguros de Salud
CNN Health	Bienestar, Investigaciones Médicas, Fitness
Fox News Health	Bienestar, Investigaciones médicas, Ébola
GDN HealthCare	Bienestar, Seguros de salud, Fitness
Good Health	Bienestar, Fitness, Investigaciones medicas
Kaiser Health News	Bienestar, Seguros de Salud, Tercera Edad
LA Times Health	Bienestar, Investigaciones Médicas, Fitness
MSN Times Health	Bienestar, Investigaciones Médicas, Medicamentos
NBC Health	Bienestar, Investigaciones Médicas, Medicamentos
NPR Health	Bienestar, Tercera Edad, Seguros de Salud
NY Times Health	Bienestar, Ébola, Seguros de Salud
Reuters Health	Bienestar, Ébola, Seguros de Salud
US News Health	Bienestar, Fitness, Seguro de Salud, Tercera Edad
WSI Health	Bienestar, Seguros de Salud, Ébola

Finalmente, la figura 4 establece la variación de tópicos de los tweets globales respecto al tiempo. Esta figura evidencia, por ejemplo, el incremento del tópico ébola en el año 2014, debido a los brotes en Guinea [19]. Complementando este análisis, la Tabla VI establece los principales tópicos por año. A forma de verificación de tópicos, se puede notar nuevamente que en el año 2014, aparece el tópico ébola, debido al brote mencionado de ébola en África [19]. Adicionalmente se tiene la figura 5, que muestra la variación de tópicos de los tweets de 4 medios de comunicación respecto al tiempo. La figura de todos los medios se puede encontrar en: http://bit.ly/3s5G9kq.

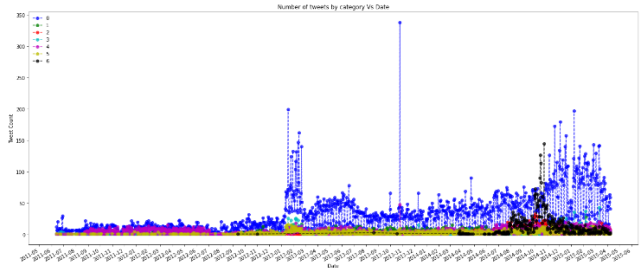

Fig. 4. Gráfica de evolución de las categorías de los tweets globales en el tiempo.

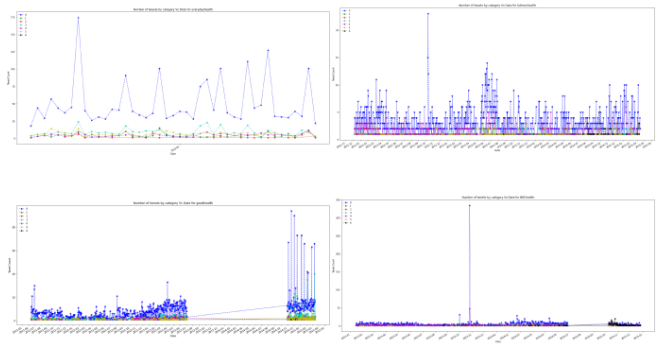

Fig. 5. Grafica de evolución de las categorías de los tweets por medio de comunicación en el tiempo

TABLE VI. TÓPICOS PRINCIPALES POR AÑO

Año	Etiqueta
2011	Bienestar, Investigaciones médicas, Fitness
2012	Bienestar, Investigaciones médicas, Fitness
2013	Bienestar, Seguros de Salud, Investigaciones médicas
2014	Bienestar, Ebola, Investigaciones médicas
2015	Bienestar, Investigaciones médicas, Fitness

V. CONCLUSIÓN

Este informe detalló metódicamente el proceso para categorizar los tópicos que cubren los medios de comunicación en el área salud en la red social Twitter, además de determinar la variación de estos en el tiempo, mediante la aplicación de algoritmos no supervisados de clusterización y su posterior análisis.

En el caso de clustering de datos se notó que, entre los métodos de procesamiento para la obtención de las raíces, el que mejor rendimiento obtuvo fue lematización, esto en conjunto con el modelo que mejor rendimiento obtuvo el cual es K Means.

La etapa de exploración de datos evidenció que los medios de comunicación deben establecer mayor continuidad en la publicacion de tweets. Además de ser complementados en un rango de tiempo en el cual la periodicidad de publicación de tweets englobe a todos los medios, ya que se notó que hay medios que tienen un rango menor al global.

El proceso de entrenamiento de los algoritmos de clustering tomó tiempos considerables por lo cual, se recomienda tener un equipo con buenas prestaciones (CPU, GPU, RAM).

VI. CONTRIBUCIONES

El presente trabajo contribuye exponiendo el rendimiento de los algoritmos no supervisados de clusterización K Means, Spectral, Jerárquico en datos de tipo texto. Además, expone una posible metodología para el análisis de categorías tratadas para la identificación de tópicos principales en cada año y un posible patrón de comportamiento en el uso de los medios de comunicación en redes sociales. Es decir, identificar meses-años donde se hablan más de un tema, por ejemplo, el ébola, evidenciaría posibles brotes de la enfermedad o posicionamiento recurrente en la opinión pública. Este análisis tendría múltiples dimensiones de aplicación, por ejemplo, permitiría a un gobierno sustentar la necesidad de atender los brotes en la población, cada cierto tiempo, o a su vez, mejorar los procesos de publicación de noticias en un medio digital, al proveer variedad de temáticas.

El desarrollo de este proyecto se dividió, de tal manera que **Abad F.** propuesta dataset, metodología del proyecto, preprocesamiento de datos, unificación de datos, búsqueda bibliografía, clustering KMeans. **Reinozo E**. procesamiento de dataset, métodos de evaluación de k optimos, clustering Hierarchical. **Trabajo en grupo**: Análisis exploratorio de dataset, Clustering Spectral, filtración y análisis de resultados, establecimiento de gráficas relevantes para resultados y conclusiones, informe y presentación.

REFERENCIAS BIBLIOGRAFICAS

[1] C. C. Aggarwal and H. Wang, “Text Mining in Social Networks,” in *Social Network Data Analytics*, Springer US, 2011, pp. 353–378.

[2] C. H. Lee, “Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams,” *Expert Syst. Appl.*, vol. 39, no. 18, pp. 13338–13356, Dec. 2012, doi: 10.1016/j.eswa.2012.05.068.

[3] H. Sheikha, “Text mining Twitter social media for Covid-19 Comparing latent semantic analysis and latent Dirichlet allocation,” 2020. Accessed: Feb. 18, 2021. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:hig:diva-32567.

[4] D. Pohl, A. Bouchachia, and H. Hellwagner, “Online indexing and clustering of social media data for emergency management,” *Neurocomputing*, vol. 172, pp. 168–179, Jan. 2016, doi: 10.1016/j.neucom.2015.01.084.

[5] K. Singh, H. K. Shakya, and B. Biswas, “Clustering of people in social network based on textual similarity,” *Perspect. Sci.*, vol. 8, pp. 570–573, Sep. 2016, doi: 10.1016/j.pisc.2016.06.023.

[6] X. Dai, M. Bikdash, and B. Meyer, “From social media to public health surveillance: Word embedding based clustering method for twitter classification,” May 2017, doi: 10.1109/SECON.2017.7925400.

[7] A. Alsayat and H. El-Sayed, “Social media analysis using optimized K-Means clustering,” in *2016 IEEE/ACIS 14th International Conference on Software Engineering Research, Management and Applications, SERA 2016*, Jul. 2016, pp. 61–66, doi: 10.1109/SERA.2016.7516129.

[8] T. Hussain Shah, N. Naveed, and Z. Rauf, “A Methodology for Brand Name Hierarchical Clustering Based on Social Media Data,” Dec. 2018. doi: 10.36785/JAES.V8I1.238.

[9] Amir Karami, “UCI Machine Learning Repository: Health News in Twitter Data Set.”

http://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter (accessed Feb. 18, 2021).

[10] PyData Editors, “pandas - Python Data Analysis Library.” https://pandas.pydata.org/ (accessed Feb. 18, 2021).

[11] Juan González Villa, “TF IDF: herramientas para mejorar la relevancia de tus contenidos - USEO.” https://useo.es/tf-idf-relevancia/ (accessed Feb. 18, 2021).

[12] P. Editors, “nltk · PyPI.” https://pypi.org/project/nltk/ (accessed Feb. 18, 2021).

[13] P. Editors, “wordcloud · PyPI.” https://pypi.org/project/wordcloud/ (accessed Feb. 18, 2021).

[14] Maklin Cory, “Hierarchical Agglomerative Clustering Algorithm Example In Python | by Cory Maklin | Towards Data Science.” https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019 (accessed Feb. 18, 2021).

[15] Abhishek Gupta, “Difference between K means and Hierarchical Clustering - GeeksforGeeks.” https://www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/ (accessed Feb. 18, 2021).

[16] L. Zelnik-Manor and P. Perona, “Self-Tuning Spectral Clustering.” Accessed: Feb. 18, 2021. [Online]. Available: http://www.vision.caltech.edu/lihi/Demos/SelfTuningClustering.html.

[17] Wikipedia Editors, “Kaiser Family Foundation - Wikipedia.” https://en.wikipedia.org/wiki/Kaiser_Family_Foundation (accessed Feb. 18, 2021).

[18] W. Editors, “Medicare - Wikipedia, la enciclopedia libre.” https://es.wikipedia.org/wiki/Medicare (accessed Feb. 18, 2021).

[19] D. Gatherer, “The 2014 Ebola virus disease outbreak in West Africa,” *Journal of General Virology*, vol. 95, no. PART 8. Society for General Microbiology, pp. 1619–1624, 2014, doi: 10.1099/vir.0.067199-0.