Capítulo 5
COVID-19: un análisis basado en minería de texto

Victor Daniel Gil vera - Victor.gilve@amigo.edu.co
Catalina Quintero López - Catalina.quinterolo@amigo.edu.co
Universidad Católica Luis Amigó

I. INTRODUCCIÓN

El COVID-19, declarada pandemia mundial por la OMS en mayo de 2020, surgió en China a finales de 2019 [1] y ha generado una emergencia sanitaria a nivel mundial por ser altamente infeccioso. A cifras de noviembre de 2020, en el mundo se han infectado 55,6 millones de personas, se han recuperado 35,8 millones y han muerto 1,34 millones. Esta problemática ha obligado a los diferentes gobiernos del mundo a tomar medidas: cuarentenas obligatorias, medidas de aislamiento social, toques de queda y restricciones sociales para reducir la propagación y evitar el colapso hospitalario, lo que ha impactado directamente en las dinámicas socioculturales, económicas, laborales, etc.

El objetivo de este trabajo fue emplear la técnica de la minería de texto para realizar un análisis de sentimientos para analizar la percepción que tienen las personas sobre las medidas que se han implementado para evitar la propagación del virus a partir de comentarios en la red social Twitter. Se emplearon 3.000 comentarios, se creó un conjunto de datos con texto y usuario, se midieron las puntuaciones de influencia en función de diversos parámetros. El trabajo se divide en las siguientes secciones: en las secciones dos y tres se presenta una contextualización general sobre el COVID-19 y una contextualización sobre la minería de texto y el análisis de sentimientos, respectivamente; en la sección cuatro se presenta la metodología, en la cinco, los resultados y la discusión y, finalmente, las conclusiones.

II. COVID-19

Este virus comenzó en la ciudad de Wuhan, China, y se ha propagado rápidamente en la mayoría de países del mundo [2]. A la fecha ha dejado

más de 1,34 millones de víctimas mortales. La OMS declaró el 30 de enero de 2020 el brote del COVID-19 como una emergencia de salud pública, y en el mes de mayo del mismo año lo declaró pandemia mundial [3].

Los primeros casos estaban relacionados con un mercado de animales vivos y mariscos, o el llamado mercado "húmedo", en Wuhan, provincia de Hubei, que se cerró rápidamente [2]. En promedio, cada individuo con el virus está infectando a otros dos [3]. La transmisibilidad del COVID-19 ha sido similar en todos los países del mundo, los contagios de personas en las principales ciudades del mundo se han vuelto inevitables debido a la exportación sustancial de casos asintomáticos y la ausencia de intervenciones de salud pública a gran escala desde el inicio de la pandemia [2].

La probabilidad de que los pacientes asintomáticos y subclínicos puedan transmitir el virus es alta, especialmente a personas de edad avanzada con comorbilidades [4]. En las enfermedades respiratorias, por lo general, se alojan una gran cantidad de virus o bacterias en la faringe y al tracto respiratorio [4]. Este virus se transmite principalmente de persona a persona por vía aérea, los síntomas más habituales que se presentan en personas contagiadas son: fiebre, tos seca y cansancio, dolores y molestias, congestión nasal, dolor de cabeza, conjuntivitis, dolor de garganta, diarrea, pérdida del gusto o el olfato, erupciones cutáneas o cambios de color en los dedos de las manos o los pies [5], [6], [7].

Estos síntomas suelen ser leves y comienzan gradualmente, algunas de las personas infectadas solo presentan síntomas leves. La mayoría de las personas (cerca del 80 %) se recuperan de la enfermedad sin necesidad de tratamiento hospitalario. Aproximadamente una de cada cinco personas que se infectan acaba presentando un cuadro grave y experimenta dificultades para respirar [8]. Las personas mayores y las que padecen afecciones médicas previas como hipertensión arterial, problemas cardiacos o pulmonares, diabetes o cáncer tienen más probabilidades de presentar cuadros graves. Sin embargo, cualquier persona puede infectarse y caer gravemente enferma [9]. Son grandes los esfuerzos que han realizado los entes gubernamentales para detener su propagación. Día tras día, científicos en todo el mundo aúnan esfuerzos para desarrollar tratamientos y una vacuna efectiva para la inmunización de adultos, jóvenes y niños. La Tabla 1 presenta las vacunas que están desarrollando diferentes compañías farmacéuticas en el mundo.

Tabla 1. Vacunas en desarrolla

	Biontech (Pfizer)	AstraZeneca	Moderna	Janssen	Sanofi	Curevac
País	**EEUU y Alemania**	**Reino Unido**	**EEUU**	**EEUU**	**Francia y Reino Unido**	**Alemania**
Fase actual de ensayos clínicos	III	III	III	III	II	II
Dosis necesarias por persona	2	2	2	1	2	2

En muchos países, debido a la ausencia de una vacuna oficial, se ha recurrido a las cuarentenas y al distanciamiento físico para reducir la tasa de infección y evitar que los sistemas nacionales de salud (clínicas y hospitales) se vean colapsados. Sin embargo, el reto de hacer frente a la pandemia en el ámbito de la salud pública va mucho más allá de estas medidas. Hay problemas crónicos y profundamente arraigados de enfermedad, pobreza y educación, especialmente en América Latina, que, en cierta medida, complican tanto la respuesta inmediata a la crisis como el eventual levantamiento de las restricciones [10], [11], [12].

Existen interrogantes y aspectos inciertos para los cuales no se conoce la respuesta exacta, como por ejemplo la manera en que se replica la enfermedad en diferentes sitios, el tiempo durante el cual los pacientes permanecen infecciosos y el tiempo que deben estar aislados. El virus está mutando en formas más o menos transmisibles [13].

Los adultos mayores con comorbilidades tienen mayor riesgo de desarrollar una enfermedad grave como el síndrome respiratorio agudo severo (SARS-CoV-2). Según las cifras de China, la letalidad ronda cerca al 5 %, pero es concebible que sea mucho mayor [14]. Es difícil para los lugares sin instalaciones de diagnóstico diferenciar entre el coronavirus y la influenza estacional [4]. El alcance de la propagación más allá de China es inminente, situación que se ve reflejada en la cantidad de contagios y muertes que el virus ha generado en países europeos y en EE. UU. [15]. El reservorio animal del virus no ha sido confirmado, pero el

análisis filogenético ha apuntado hacia los murciélagos después de algunas especulaciones no confirmadas de una participación de serpientes y pangolines [16].

En cualquier caso, parece probable que el SARS-CoV-2 se originó en un mercado húmedo [17]. Los mercados húmedos son un riesgo de transmisión de enfermedades zoonóticas a los humanos [18]. China y otros países asiáticos pueden querer revisar su política en estos mercados para intentar minimizar el riesgo, lo que podría significar imponer medidas más estrictas de control de infecciones o incluso prohibirlos por completo [11]. Los consumidores en los países asiáticos donde abundan los mercados húmedos buscan comprar animales vivos y usarlos en su cocina. En ese caso, las consideraciones socioculturales harán que abogar por su eliminación sea una propuesta difícil que deberá adaptarse cuidadosamente a la población asiática. Los mercados húmedos son lugares de intercambio de materiales genéticos de una variedad de animales diferentes, por lo cual se deben vigilar los estándares de higiene y hacer seguimiento de los microbios que circulan dentro de estos [19]. La Tabla 2 presenta el resumen de casos, personas recuperadas y decesos a causa del COVID-19 a noviembre de 2020:

Tabla 2. Cifras oficiales (OMS)

Ubicación	Total de casos	Personas recuperadas	Muertes
Antioquia	200 k	187 k	3785
Colombia	1,23 M	1,14 M	34.929
Mundo	55,6 M	35,8 M	1,34 M

En la revisión del estado del arte, diversas investigaciones afirman que las medidas preventivas adoptadas por los entes gubernamentales pueden ser percibidas como experiencias traumáticas que generan un impacto psicológico amplio y duradero para quienes las sobrellevan, y afectan a diferentes grupos etarios como niños, estudiantes y adultos mayores [20]. Las restricciones establecidas pueden crear una sensación de pérdida de la libertad por la limitación en los desplazamientos, el alejamiento de los seres queridos, la repentina desaceleración de la economía y la incertidumbre constante del posible contagio de la enfermedad, lo que ocasiona la aparición de alteraciones psicológicas [21].

III. MINERÍA DE TEXTO

La minería de texto es el proceso de derivar información de alta calidad de información textual para extraer información útil para la toma eficiente de decisiones [22]. Básicamente, es un proceso de análisis y exploración

de grandes volúmenes de información textual no estructurada, con el fin de identificar conceptos, patrones, temas, palabras claves, tendencias, gustos, etc.

Como rama de la analítica de datos, permite realizar análisis de manera masiva, identificar tendencias, clasificar comentarios y opiniones, clasificar tipos de sentimientos y construir nubes de palabras que permiten obtener una visión general de la información.

En el último quinquenio ha sido empleada y utilizada por científicos y académicos a nivel mundial, debido al desarrollo de plataformas de *Big-Data* y algoritmos de aprendizaje profundo que tienen la capacidad de analizar conjuntos masivos de datos no estructurados.

La minería de texto puede definirse en términos generales como un proceso intensivo de conocimiento en el que un usuario interactúa con una colección de documentos a lo largo del tiempo mediante herramientas de análisis [23]. La Figura 1 presenta el proceso de minería de texto.

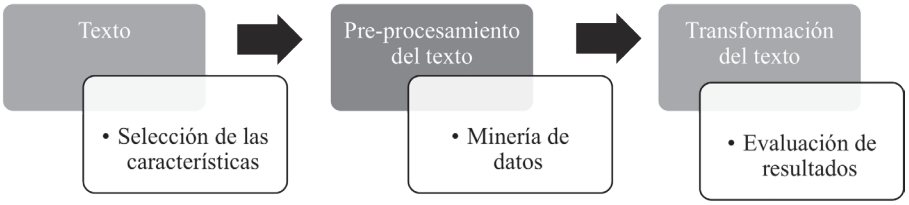

Figura 1. Proceso de minería de texto

3.1 Análisis de sentimientos

Es una rama de la minería de texto, la cual se ha aplicado ampliamente en diferentes contextos: políticos, socioculturales, económicos, entre otros. Diversas investigaciones la han empleado para comprender los pensamientos de las personas. A partir de datos desordenados, las palabras clave extraídas pueden convertirse en conceptos y transformarse en información valiosa [24].

La información disponible en las redes sociales es una fuente de información destacada durante la crisis producida por la pandemia del COVID-19. Los mensajes generados por los usuarios permiten ver un panorama general de lo que piensan las personas, así como conocer estados de ánimo y opiniones. Por la gran cantidad de mensajes emitidos en la redes sociales, es posible realizar un análisis a gran escala de la evolución que ha tenido la pandemia y la percepción de todos los factores asociados a ella [25].

IV. MÉTODOLOGÍA

En este trabajo se empleó el software estadístico R 3.6.2 para analizar 3.000 comentarios de la red social Twitter sobre las medidas adoptadas para evitar la propagación del virus COVID-19. Los paquetes y librerías empleados en el análisis se presentan a continuación:

```
# Install
install.packages("tm")  # for text mining
install.packages("SnowballC") # for text
stemming
install.packages("wordcloud") # word-cloud
generator
install.packages("RColorBrewer") # color
palettes
install.packages("dplyr") # color palettes
install.packages("tidyr") # color palettes
install.packages("tidytext") # color palettes
# Load
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
library("corpus")
library(rtweet")
library("dplyr")
library("tidyr")
library("tidytext")
```

Los *Tweets* fueron seleccionados a través de una API. Se empleó la librería *rtweet*, un *wrapper R* que se comunica con la API de Twitter. Para la extracción de comentarios se creó una *Twitter App*, mecanismo que proporciona Twitter para desarrolladores que quieran acceder a sus contenidos a través de programas, para lo cual proporciona una serie de claves y tokens personalizados. El límite de extracción de comentarios fue de 100 cada 15 minutos (*rate limiting*). La Tabla 3, presenta algunos de los comentarios empleados en el análisis.:

Tabla 3. Comentarios de Twitter

Usuario	Comentario
@NCEAatUSC	COVID-19 has exacerbated feelings of isolation and loneliness in older adults. Practice physically distanced social interactions with older adults to help decrease social isolation and promote connections.
@DanCorcoranTV	Many CT families are worried that social isolation is having a devastating effect on the physical and mental health of their loved ones. Experts say it is a hidden crisis amid the COVID-19 pandemic
@HSScomms	Human–dog relationships during the COVID-19 pandemic: booming dog adoption during social isolation
@cnni	The French health minister said on Tuesday there is a “credible risk” the novel coronavirus outbreak could turn into a pandemic.
@JacariOxford	Local lockdowns and restrictions are leading to increased social isolation for children with English as an additional language and limited opportunities to practise their English..

Una vez seleccionados los *tweets* para hacer el análisis, se almacenaron en un archivo de texto plano .txt, el cual se importó al programa para realizar el proceso de *tokenización*, es decir, el proceso de limpieza en el cual se sustituyen de manera masiva caracteres especiales (“/”, “@” y “|”), espacios en blanco, palabras comunes, etc. Algunas palabras fueron reducidas a su forma raíz y se eliminaron los sufijos. A continuación, se presenta el código empleado para la depuración del texto:

```
# Load the data as a corpus docs <-
Corpus(VectorSource(text))
# Tokenization toSpace <- content_
transformer(function (x , pattern ) gsub(pattern,
“ “, x)) docs <- tm_map(docs, toSpace, “/”) docs
<- tm_map(docs, toSpace, “@”) docs <- tm_
map(docs, toSpace, “\\|”)
# Convert the text to lower case
docs <- tm_map(docs, content_
transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords,
stopwords(“english”))
# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c(“blabla1”,
“blabla2”))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Text stemming
# docs <- tm_map(docs, stemDocument)
```

V. RESULTADOS Y DISCUSIÓN

Una vez depurados los comentarios se procedió a construir la matriz de documentos, en la cual se muestra la frecuencia de aparición de las palabras de los comentarios. El código empleado y la matriz de clasificación para las 10 palabras más frecuentes se presentan a continuación:

```
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```

	word	freq
distancing	distancing	180
quarantine	quarantine	169
social	social	157
public	public	149
people	people	120
pandemic	pandemic	110
infectious	infectious	97
diseases	diseases	83
crowed	crowed	75
understand	understand	69

Posteriormente, se procedió a generar la nube de palabras o *wordcloud*, en la cual aparecen las palabras más frecuentes y cuyo tamaño de fuente varía dependiendo de la frecuencia de aparición de las palabras (Figura 2). El código empleado y el *wordcloud* se presentan a continuación:

```
# wordcloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq,
min.freq = 1,
max.words=200, random.order=FALSE, rot.
per=0.35, colors=brewer.pal(8, “Dark2”))
```

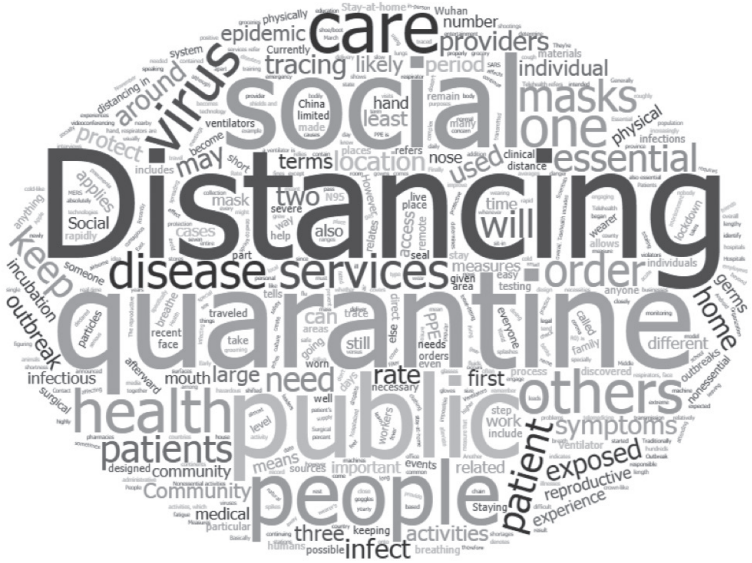

Figura 2. Nube de palabras

El resultado de la nube de palabras muestra claramente que las palabras *"distancing", "quarantine", "social*", "*public"* y "*people"* son las más importantes de todos los *tweets* seleccionados sobre "COVID-19". Se debe aclarar que, los argumentos que componen la función fueron definidos de la siguiente manera:

- *min.freq*: palabras con una frecuencia inferior a "min.freq" no se graficarán (frecuencia 1)
- *max.words*: número máximo de palabras que se graficarán (200 palabras)
- *random.order*: trazar palabras en orden aleatorio. Si es falso, se grafican en frecuencia decreciente (en este trabajo se adoptó el valor FALSO)
- *rot.per*: proporción de palabras con rotación de 90 grados (texto vertical) (0.35)
- *colors:* palabras de color de menor a mayor frecuencia. En nuestro caso adoptamos los valores de la paleta (brewer.pal(8, "Dark2"))

Adicionalmente, se clasificaron los términos más frecuentes, en este caso las que aparecían al menos 5 veces:

```
# Frequency.Terms
findFreqTerms(dtm, lowfreq = 5)
[2] "isolation"  "economy"  "crisis"
"work"
[9] "covid"  "fatal"  "conflict"  "diagnosed"
[25] "poverty"  "health"  "infections"
"masks"  "human"
```

Finalmente, se asociaron términos frecuentes utilizando la función *findAssocs()*. Se descartaron correlaciones inferiores a 0.3. Se identificaron las palabras que estaban asociadas con la palabra "COVID-19" en los *tweets* seleccionados, dando los siguientes resultados:

findAssocs(dtm, terms = "covid", corlimit = 0.3), $covid		
Isolation 0.45	Confinement 0.48	Distancing 0.48
Public 0.40	Health 0.45	Virus 0.43
Keep 0.39	Masks 0.40	Infect 0.39
Economy 0.39	Love 0.39	Social 0.39
Measures 0.39	Rate 0.39	Need 0.39

Finalmente, se clasificaron los Tweets por tipos de sentimientos, para ello empleamos el código:

```
sentiments %>%
  filter(word_count %in% c("very postive",
"positive", "neutral","negative", "very neg-
ative")) %>%
  arrange(word) %>% #sort
  select(-score) %>% #remove this field
```

Lo que nos permitió obtener la siguiente clasificación final:

sentiment	total
very positive	198
positive	443
neutral	224
negative	1683
very negative	452

VI. CONCLUSIONES

Gracias a la minería de texto y al análisis de sentimientos es posible conocer de manera masiva la percepción que tienen las personas sobre un fenómeno determinado. Esta investigación permitió analizar los comentarios en de 3000 personas acerca de las medidas adoptadas para prevenir la propagación del COVID-19, identificar los términos más frecuentes, construir una nube de palabras ("*Wordcloud*") y clasificarlos en cuatro tipos de sentimientos.

La minería de texto, a diferencia de otras técnicas de análisis de información cualitativa, tiene la ventaja de que permite descubrir información nueva, permite analizar grandes bases de datos de información textual mediante la tecnología de *Data mining*, generando resultados que son fáciles de entender, lo que contribuye a la toma eficiente de decisiones, ya que permite detectar la información relevante, tendencias, patrones, entre otros. En efecto, la minería de datos se constituye como una poderosa herramienta para el análisis de información cualitativa.

A pesar de que la mayoría de medidas implementadas para evitar la propagación del virus han afectado considerablemente las dinámicas sociales, culturales, económicas, educativas y demás, estás medidas han ayudado a disminuir los contagios y en efecto las muertes. Hasta que no haya una vacuna oficialmente reconocida por las autoridades mundiales de salud, las personas deben acogerse a estas medidas y buscar apoyo psicológico, económico y demás en caso tal de verse sumamente afectados.

VII. REFERENCIAS

[1] R. Habibi *et al.*, "Do not violate the International Health Regulations during the COVID-19 outbreak," *Lancet*, 2020.

[2] T. Burki, "Outbreak of coronavirus disease 2019," *Lancet Infect. Dis.*, 2020.

[3] D. Chang, H. Xu, A. Rebaza, L. Sharma, and C. S. Dela Cruz, "Protecting health-care workers from subclinical coronavirus infection," *Lancet Respir. Med.*, 2020.

[4] J. Zhang, L. Zhou, Y. Yang, W. Peng, W. Wang, and X. Chen, "Therapeutic and triage strategies for 2019 novel coronavirus disease in fever clinics," *Lancet Respir. Med.*, 2020.

[5] A. Tsatsakis *et al.*, "SARS-CoV-2 pathophysiology and its clinical implications: An integrative overview of the pharmacotherapeutic management of COVID-19," *Food Chem. Toxicol.*, vol. 146, p. 111769, 2020.

[6] P. Zhang *et al.*, "Risk factors associated with the progression of COVID-19 in elderly diabetes patients," *Diabetes Res. Clin. Pract.*, p. 108550, 2020.

[7] K. D. Lee *et al.*, "Providing essential clinical care for non-COVID-19 patients in a Seoul metropolitan acute care hospital amidst ongoing treatment of COVID-19 patients," *J. Hosp. Infect.*, vol. 106, no. 4, pp. 673–677, 2020.

[8] N. M. Vranis, J. M. Bekisz, D. A. Daar, E. S. Chiu, and S. C. Wilson, "Clinical outcomes of COVID-19 positive patients who underwent surgery: A New York City experience," *J. Surg. Res.*, 2020.

[9] X. Wang, L. Lin, Z. Xuan, J. Xu, Y. Wan, and X. Zhou, "Risk communication on behavioral responses during COVID-19 among general population in China: A rapid national study," *J. Infect.*, 2020.

[10] C.-Y. Lin, A. Broström, M. D. Griffiths, and A. H. Pakpour, "Investigating mediated effects of fear of COVID-19 and COVID-19 misunderstanding in the association between problematic social media use, psychological distress, and insomnia," *Internet Interv.*, vol. 21, p. 100345, 2020.

[11] E. A. Severo, J. C. F. De Guimarães, and M. L. Dellarmelin, "Impact of the COVID-19 pandemic on environmental awareness, sustainable consumption and social responsibility: Evidence from generations in Brazil and Portugal," *J. Clean. Prod.*, p. 124947, 2020.

[12] F. J. Elgar, A. Stefaniak, and M. J. A. Wohl, "The trouble with trust: Time-series analysis of social capital, income inequality, and COVID-19 deaths in 84 countries," *Soc. Sci. Med.*, vol. 263, p. 113365, 2020.

[13] W. Liang *et al.*, "Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China," *Lancet Oncol.*, 2020.

[14] Z. Xu, S. Li, S. Tian, H. Li, and L. Kong, "Full spectrum of COVID-19 severity still being depicted," *Lancet*, 2020.

[15] X. Gu, B. Cao, and J. Wang, "Full spectrum of COVID-19 severity still being depicted – Authors' reply," *Lancet*, 2020.

[16] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges," *Int. J. Antimicrob. Agents*, p. 105924, 2020.

[17] D. MacKenzie, "How bad will it get?," *New Sci.*, vol. 245, no. 3269, p. 7, 2020.

[18] T. Solomon, P. Lewthwaite, D. Perera, M. J. Cardosa, P. McMinn, and M. H. Ooi, "Virology, epidemiology, pathogenesis, and control of enterovirus 71," *Lancet Infect. Dis.*, vol. 10, no. 11, pp. 778–790, 2010.

[19] D. MacKenzie, "Wuhan-like virus discovered seven years ago," *New Sci.*, vol. 245, no. 3269, p. 9, 2020.

[20] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[21] S. K. Brooks *et al.*, "The psychological impact of quarantine and how to reduce it: rapid review of the evidence," *Lancet*, 2020.

[22] X. Xie, Y. Fu, H. Jin, Y. Zhao, and W. Cao, "A novel text mining approach for scholar information extraction from web content in Chinese," *Futur. Gener. Comput. Syst.*, 2019.

[23] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[24] W.-L. Chang and J.-Y. Wang, "A 2020 perspective on 'Mine is yours? Using sentiment analysis to explore the degree of risk in sharing economy,'" *Electron. Commer. Res. Appl.*, p. 100934, 2020.

[25] A. Kruspe, M. Häberle, I. Kuhn, and X. X. Zhu, "Cross-language sentiment analysis of European twitter messages duringthe COVID-19 pandemic," *arXiv*, 2020.