Hoja de ruta para producir frecuentemente información estadística representativa

mediante el uso conjunto de información de redes sociales y encuestas

Roadmap for Frequently Producing Representative Statistical Information

through the Joint Use of Social Networking and Survey Data

Víctor Alfredo Bustos y de la Tijera, Silvia Laura Fraustro Velhagen, Noemí López Delgado y Ricardo Antonio Olvera Navarro*

Desarrollamos una propuesta metodológica para que las oficinas nacionales de estadística produzcan información representativa sobre múltiples temas, con mayor frecuencia, utilizando en conjunto datos de encuestas de hogares y publicaciones en redes sociales. La propuesta se basa en dar un nuevo rol a los datos como insumo para el entrenamiento de algoritmos de aprendizaje automático (ML, por sus siglas en inglés). Comenzamos clasificando a los encuestados según sus datos registrados en el cuestionario. Las publicaciones en las redes sociales de estos, si las hubiera, heredan sus etiquetas de clase. Utilizándolas como entrada, se entrenan algoritmos de ML. Para el seguimiento, las recientes en el momento de las nuevas recopilaciones de la encuesta se etiquetan y se entrenan de nuevo los algoritmos. En cualquier caso, cuando se considera apropiado el resultado de entrenar un algoritmo, se utiliza para etiquetar automáticamente grandes volúmenes de publicaciones actuales y futuras de usuarios no incluidos en la encuesta. El seguimiento futuro se lleva a cabo

We developed a methodological proposal for National Statistical Offices (NSOs) to produce representative information on multiple topics, with greater frequency, using household survey data and social media posts together. The proposal is based on giving a new role to the data as input for training machine learning (ML) algorithms. We begin by classifying respondents according to their data recorded in the questionnaire. Their social media posts, if any, inherit their class tags. Using them as input, ML algorithms are trained. For follow-up, recent ones at the time of new survey collections are tagged and algorithms are trained again. In either case, when the result of training an algorithm is deemed appropriate, it is used to automatically tag large volumes of current and future posts from users not included in the survey. Future tracking is carried out through tweets posted between survey rounds. The above procedure also has application in selection bias mitigation. In this case, a minimal set of sociodemographic (SD) variables collected through surveys can be used to develop a da-

* Instituto Nacional de Estadística y Geografía (INEGI), alfredo.bustos@inegi.org.mx, silvia.fraustro@inegi.org.mx, nohemi.delgado@inegi.org.mx y ricardo.olvera@inegi.org.mx, respectivamente.

a través de tuits publicados entre rondas de encuestas. El procedimiento anterior también tiene aplicación en la mitigación del sesgo de selección. En este caso, se puede usar un conjunto mínimo de variables sociodemográficas (SD) recopiladas a través de encuestas para desarrollar una base de datos de autores etiquetada según SD. Se hará referencia a esta durante los estudios temáticos para mitigar la falta de representatividad de la población de usuarios. Para que todo lo anterior funcione, las respuestas a las encuestas y las publicaciones en redes de los usuarios-informantes deben ser vinculadas. Proponemos una forma de conseguirlo. Un futuro levantamiento de la Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares del Instituto Nacional de Estadística y Geografía se empleará para estudiar la viabilidad de la propuesta, puesto que ya investiga el uso de las redes sociales y recopila información sociodemográfica.

Palabras clave: redes sociales; representatividad; sesgo de selección; etiquetado.

Recibido: 22 de septiembre de 2021.
Aceptado: 5 de noviembre de 2021.

tabase of authors labelled according to SD. This will be referenced during the thematic studies to mitigate the lack of representativeness of the user population. For all of the above to work, survey responses and user-informant network postings must be linked. We propose a way to achieve this. A future survey of the National Survey on Availability and Use of Information Technologies in Households (ENDUTIH in Spanish) of the National Institute of Statistics and Geography (INEGI) will be used to study the feasibility of the proposal, since it already investigates the use of social networks and collects sociodemographic information.

Key words: official statistics; machine learning; social networks; tagging; representativeness; selection bias.

Young Girl Using Smart Phone Social Media Concept/ bombuscreative/ iStock

Introducción

La sola posibilidad de complementar las fuentes tradicionales de información mediante la incorporación de los así llamados grandes datos o *Big Data* ha dado lugar a una amplia bibliografía en años recientes. Numerosos organismos nacionales e internacionales han llevado a cabo diversas acciones para estudiar su uso en sus actividades cotidianas. Por ejemplo, la CBS holandesa está entre las primeras oficinas nacionales de estadística (ONE) en iniciar el estudio de estas fuentes alternativas, así como de sus implicaciones para la estadística oficial (ver Struijs *et al*., 2014 y Struijs y Daas, 2014). A su vez, la Organización de Naciones Unidas (ONU) creó en el 2014 el Grupo Global de Trabajo (GWG, por sus siglas en inglés) para *Big Data* en la estadística oficial[1,2] (ver UNSD, 2015; Jansen, 2020; Smith, 2018). De acuerdo con Snyder (2015), sus términos de referencia[3] asignan al GWG, entre otras, las tareas de "… aportar una visión estratégica, dirección y coordinación de un programa global de Big Data para la estadística oficial, incluyendo los indicadores para la *Agenda 2030 para el desarrollo sostenible*. También promueve el uso práctico de fuentes *Big Data* de grandes datos, la promoción del desarrollo de capacidades, el entrenamiento y el intercambio de experiencias…". Durante su primera reunión en octubre 2014 en Beijing, el GWG estableció ocho equipos de trabajo:

1. Datos de redes sociales.
2. *Big Data* y los Objetivos de Desarrollo Sostenible (ODS).
3. Datos de telefonía móvil.
4. Temas transversales.
5. Mejorar acceso a fuentes *Big Data*.
6. Promoción y comunicación.
7. Capacitación, habilidades y fortalecimiento de capacidades.
8. Imágenes de satélite y datos geoespaciales.

Estos han sesionado y alcanzado diferentes avances. En Jansen (2020)[4] se hace un gran resumen de dichos avances para casi todos los grupos creados por el GWG. Cabe resaltar que no están señalados aquellos para los equipos sobre integración de datos ni el que se refiere al uso de información de redes sociales, al que declara en receso.[5]

Un ámbito de aplicación inmediato para los trabajos del GWG está dado por la *Agenda 2030*, la cual adoptó un marco global de monitoreo amplio abarcando 231 indicadores dentro de 17 ODS (ver Van Halderen *et al*., 2021). Como señalan Lokanathan *et al*. (2017), "… aprovechar fuentes de datos nuevas y existentes (tanto del sector público como del privado) con el fin de monitorear el progreso hacia los ODS, así como para lograrlo, no está exento de desafíos…"; enfatizan que las diferencias en lo que denominan *datificación*[6] entre las economías desarrolladas y las que están en vías de serlo impedirán que estas hagan un uso óptimo de la información disponible. Los mismos autores indican que "… es importante recordar que a pesar del gran acervo de literatura y aplicaciones que ya existen, el estado del arte en aplicaciones enfocadas en el desarrollo innovador de estas nuevas fuentes de datos aún se encuentra en sus etapas embrionarias…"; adicionalmente, exponen que será necesario poner particular atención para "… abordar los dilemas éticos y de privacidad…" que surgirán de estas fuentes de datos.

Para América Latina, en Data-Pop Alliance (2016) se indica que "… destacan los riesgos y las oportunidades que *Big Data* presenta a las oficinas nacionales de estadística en Latinoamérica en el contexto de los ODS…". Entre los primeros, incluye barreras institucionales para la administración del cambio y la innovación, restricciones para el acceso y la completez de los datos, retos técnicos y

1 https://unstats.un.org/bigdata/

2 United Nations Global Working Group (GWG) on Big Data for Official Statistics (https://unstats.un.org/bigdata/), y sus seis International Conferences on Big Data for official statistics (ej. https://unstats.un.org/unsd/bigdata/conferences/2020/).

3 https://unstats.un.org/bigdata/documents/TOR%20-%20GWG%20-%202015.pdf

4 International Symposium on the Use of Big Data for Official Statistics, 16-18 October 2019, Hangzhou, China http://www.stats.gov.cn/english/InternationalTraining/2019/202009/t20200930_1792523.html

5 El reporte más reciente del Social Media WG para 2017 se puede encontrar en https://unstats.un.org/unsd/bigdata/conferences/2017/gwg/GWG%20Task%20Team%20on%20Social%20Media%20Data%20-%202017%20report.pdf

6 Tendencia tecnológica que convierte aspectos de nuestra vida en datos que posteriormente se transforman en información.

metodológicos, brechas en capacidades humanas, así como riesgos éticos y políticos. Por otro lado, sobresalen cinco tendencias que considera propicias en la región: la experiencia latinoamericana en el movimiento de datos abiertos;[7] la aparición de asociaciones públicas y privadas sobre el tema de *Big Data*;[8] la presencia de comités, instituciones y grupos de trabajo fuertes, y que abarcan a toda la región; el desarrollo de mejores prácticas adaptables; y la existencia de una red interdisciplinaria de innovación que involucra tanto a las ONE como a otros actores. Se concluye desarrollando una hoja de ruta regional para el aprovechamiento de *Big Data* en la estadística oficial y en el seguimiento a los ODS.

Aun considerando legislar para que las oficinas nacionales de estadística accedan a las fuentes *Big Data*, como se sugiere en Destatis (2019), es claro que no se puede garantizar que las empresas de redes sociales sigan funcionando. Hay experiencias de estas que han visto reducido el número de sus usuarios o hasta han desaparecido, por ejemplo, *Sixdegrees*, *Friendster* y *My Space*. La producción de estadística oficial habrá de establecer con adecuada anticipación las estrategias a instrumentar para garantizar la continuidad y comparabilidad de los resultados.

7 Iniciativa Latinoamericana por los Datos Abiertos (ILDA), https://idatosabiertos.org/acerca-de-nosotros/

8 Ver también DNP (2017) y Dutra (2018).

Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos

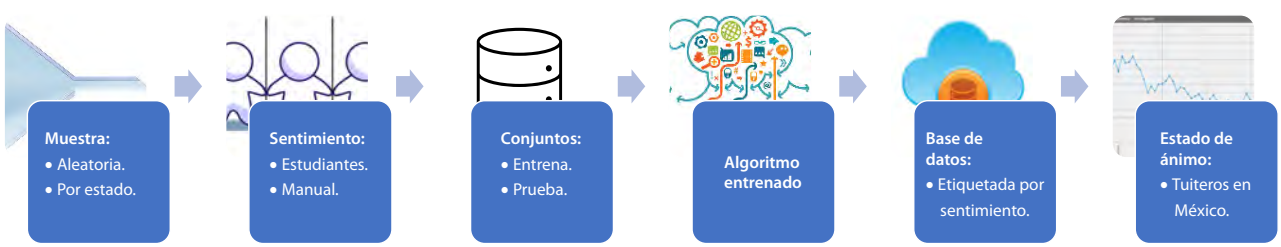
Este proyecto[9] se encuentra entre los usos más destacados que el Instituto Nacional de Estadística y Geografía (INEGI) ha dado al aprendizaje automatizado, así como a textos provenientes de *Twitter* (INEGI, 2017). En sus inicios, la única certeza era que los usuarios de esta red que publicaban textos georreferenciados se representaban a sí mismos; es decir, que la inferencia que resultara del análisis de sus textos aplicaría a ese segmento de habitantes, pero no necesariamente al resto de la población mexicana. De ahí el nombre que se dio al proyecto.

El procedimiento seguido en aquel momento para llegar a la publicación se resume en la figura 1. Entre los mensajes capturados se obtuvo una muestra aleatoria. Cada mensaje en esta fue etiquetado manualmente por uno o más estudiantes según su interpretación del sentimiento (positivo, neutro o negativo) del autor al publicarlo. Se obtuvo, así, un conjunto de tuits etiquetados que fueron separados para formar conjuntos de entrenamiento y prueba. Con este insumo se entrenaron y evaluaron diversos algoritmos para elegir la combinación más adecuada. A esta parte del proceso se la denominó *Pío análisis*. A partir de ella fue posible etiquetar de manera automática millo-

9 https://www.inegi.org.mx/app/animotuitero/#/app/multiline

Figura 1

Proceso del estado de ánimo de los tuiteros en México



Fuente: elaboración propia.

nes de mensajes almacenados en nuestra base de datos, así como los que cotidianamente son capturados con el fin de dar seguimiento diario al estado de ánimo de los tuiteros.

Limitaciones por resolver

Sesgo de selección

Al igual que ocurre en el caso de otros estudios observacionales, es necesario preguntarnos si la muestra disponible de usuarios de una red social es representativa de la población para la cual se quiere inferir. Es claro que, mientras más se aleje la primera de la segunda en términos de variables relevantes a la materia de estudio, se corre el riesgo de que los resultados que se produzcan apliquen solamente para la muestra; es decir, de que se vean afectados por un sesgo en relación con la población objetivo, ante la sub o sobrerrepresentación de una o más subpoblaciones importantes.

La incertidumbre que dio lugar al nombre del proyecto *Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos* motivó la ampliación de la Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH) del INEGI, cuyo fin era el de indagar el uso de redes sociales en estos. De esta manera, Bustos *et al*. (2021) pudieron identificar discrepancias sociodemográficas entre los usuarios de redes sociales y la población mexicana en general, con lo cual, además de que se justifica la elección del nombre del proyecto, quedan establecidas sus limitaciones y alcances. Destacan que, para el caso mexicano, las poblaciones de usuarios de redes sociales son, en promedio, más jóvenes, con un estatus socioeconómico más alto y un mayor nivel educativo que el resto de la población. Cabe destacar que, aun cuando subrepresentadas en términos relativos, hay en estas subpoblaciones usuarios de casi todos los grupos de edad, así como de todos los niveles de escolaridad y socioeconómicos.

Sobre el denominado sesgo de selección en redes sociales, es poco lo que se ha escrito (ver Golub, 2010 o Stark, 2015) y menos todavía acerca de sus implicaciones en la estadística oficial. Iacus *et al*. (2020) elaboran una propuesta metodológica para corregirlo a través de modelos comúnmente usados para estimación en dominios pequeños. En este caso, se ajustan los modelos mencionados a los agregados al nivel de dominio, afectados por el sesgo de selección, usando como variables explicativas agregados similares provenientes de la estadística oficial; en otras palabras, se busca corregir agregados y no a la base de datos misma. Por su parte, Kim y Tam (2020) recurren al mismo tipo de modelos, pero su enfoque, basado en lo que ha venido denominándose integración de datos, permite el tratamiento de errores de medición en las variables tanto para *Big Data* como para la muestra probabilística. Para la corrección del sesgo de selección en la estimación de proporciones, Tam y Kim (2019) introducen dos enfoques, para el caso en el que los registros de la encuesta pueden ser vinculados con los de la fuente *Big Data* y para lo contrario, recurriendo a modelos logísticos.

A causa de lo anterior, por el momento no ha sido posible instrumentar una estrategia que nos permita acercarnos al estado de ánimo de los mexicanos, según *Twitter*. Para ello, requeriríamos etiquetar adicionalmente los tuits capturados de acuerdo con características sociodemográficas (SD) de sus autores, como se muestra en la figura 2. Sin embargo, ya que la ENDUTIH tiene como unidad de análisis a las personas, en términos de las cuales hemos caracterizado el sesgo de selección y su posible corrección, es deseable dejar de trabajar con mensajes individuales. Si las características SD de los tuiteros fueran conocidas, en el momento de calcular totales para la construcción de índices se podrían reponderar sus etiquetas temáticas según la relación guardada entre los tamaños relativos de la clase a la que pertenecen en la población abierta, por un lado, y en la de tuiteros que publicaron durante el periodo en cuestión, por el otro. Para la clase identificada con las etiquetas $i, j,\ldots, k$, se representa su tamaño relativo en relación con la población de referencia (R) por $\%P^{R}_{ij\ldots k}$ y el de la de la muestra de redes sociales (SN) para el periodo por $\%P^{SN}_{ij\ldots k}$. Entonces, el factor de corrección para

representantes de esa subpoblación se presenta en la expresión (1):

$$\omega_{ij\ldots k} = \frac{\%P^{R}_{ij\ldots k}}{\%P^{SN}_{ij\ldots k}} \qquad (1)$$

Un valor menor a 1 del cociente (1) indicará que la subpoblación está sobrerrepresentada en esa red social. En caso contrario, se estará frente a una subrepresentación. A partir de ese momento, ese usuario será ponderado de acuerdo con el anterior cociente al ser incluido en agregaciones espaciales y/o temporales, tomando en cuenta sus publicaciones recientes. Sin embargo, tal información no se encuentra a nuestra disposición en general.

Si la información requerida estuviera disponible para cada tuitero *t* que publicó al menos un tuit en la región y en el periodo considerados, se estaría en condiciones de recalcular el Índice de Positividad del estado de ánimo de los tuiteros como se muestra en (2). En dicha expresión, los índices $I_{+}(t)$ e $I_{-}(t)$ toman valor 1 según si el ánimo del tuitero en el periodo es considerado positivo o negativo, respectivamente:

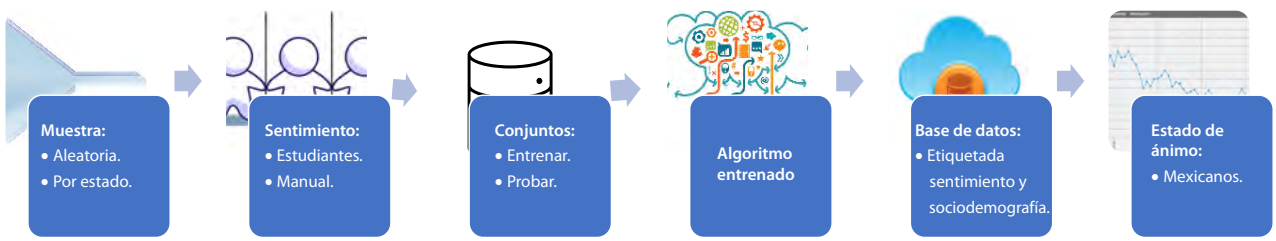
$$\textit{Índice de Positividad} = \frac{\sum_t \omega_{ij\ldots k}(t) I_{+}(t)}{\sum_t \omega_{ij\ldots k}(t) I_{-}(t)} \qquad (2)$$

Consulta directa a los usuarios

Otro asunto no resuelto se refiere la incertidumbre introducida cuando *expertos* etiquetan los mensajes. Ejemplo de ella es el hecho de que un mismo mensaje no reciba la misma etiqueta al ser evaluado por distintos individuos. En tanto se estudia la manera de incorporar dicha incertidumbre en el entrenamiento de algoritmos, parece deseable explorar alternativas. Por ejemplo, podría haberse considerado preguntar de manera directa al autor de la publicación sobre su estado de ánimo al momento de escribirla, como se muestra en la figura 3. Es claro que ello no elimina totalmente la incertidumbre, ya que el propio proceso de recordación carece de certezas.

Figura 2

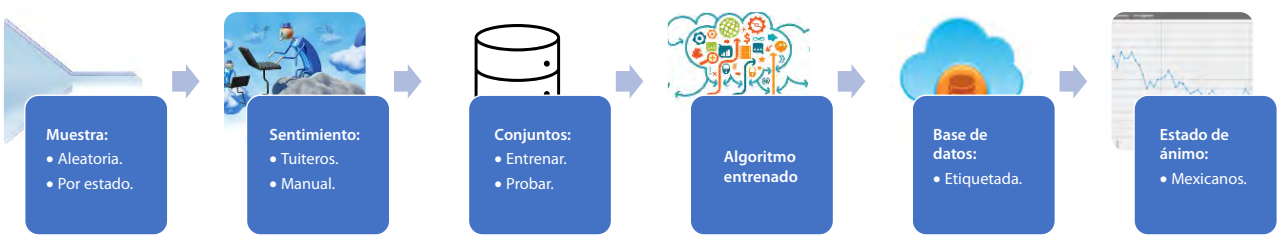
Proceso del estado de ánimo de los mexicanos reduciendo sesgo de selección



Fuente: elaboración propia.

Figura 3

Proceso del estado de ánimo de los mexicanos consultando autores



Fuente: elaboración propia.

Sin embargo, existen otras formas de acercarse a los usuarios de *Twitter* que, a nuestro juicio, reducen algunos de los mencionados riesgos. Las comentaremos más adelante como parte de nuestra propuesta.

Predicción de la información requerida

El hecho de que nuestra base de datos de tuits georreferenciados no cuente con información sobre la edad, el sexo, el nivel académico ni el estrato socioeconómico impide llevar a cabo la indicada ponderación de tuiteros. En *Twitter*, los usuarios pueden decidir mantener confidencial su perfil. En consecuencia, no podemos acudir a dicha información.

La segunda opción obvia es la de predecir las características de interés de los tuiteros cuyos tuits públicos almacenamos. Ejemplificaremos cómo proceder haciendo uso de la ENDUTIH, aunque cualquier otra encuesta que recopile la información necesaria podría servir. A lo largo del llenado del cuestionario de la Encuesta se han registrado datos sociodemográficos de cada integrante del hogar. Adicionalmente, los informantes que declaran ser usuarios recientes de alguna red han sido identificados.[10] Se requiere vincular aquella información con los mensajes públicos de cada usuario. Para ello, sugerimos que al concluir la entrevista después de darle una breve explicación del objetivo que perseguimos, así como de señalarle los artículos de la *Ley del Sistema Nacional de Información Estadística y Geográfica* que garantizan el uso confidencial de sus datos (ver anexo), se solicite al tuitero que nos proporcione su *username* o que publique un mensaje que contenga un código personalizado y conocido solo por ambas partes para evitar errores al registrar su nombre de usuario. El informante está en total libertad de atender nuestra solicitud. Si lo hiciera, nos aportará su nombre de usuario, lo que nos permitirá acudir a sus publicaciones. Estas podrán ahora ser etiquetadas con la información socioeconómica recogida en el cuestionario de la Encuesta. De esta manera, habremos creado un conjunto de entrenamiento, y uno menor de prueba, que serán insumo para la prueba de diversos algoritmos de aprendizaje automatizado. Los resultados de estos se evaluarán para determinar una combinación adecuada. En su primera mitad, la figura 4 ilustra el anterior procedimiento. En ella, el resultado intermedio en la forma de un algoritmo entrenado se destaca con un color de fondo diferente; ello nos permitirá señalar dónde y cuándo se usará posteriormente dicho algoritmo. A partir de este momento, dejaremos de hacer uso individual de los datos de los informantes de la ENDUTIH.

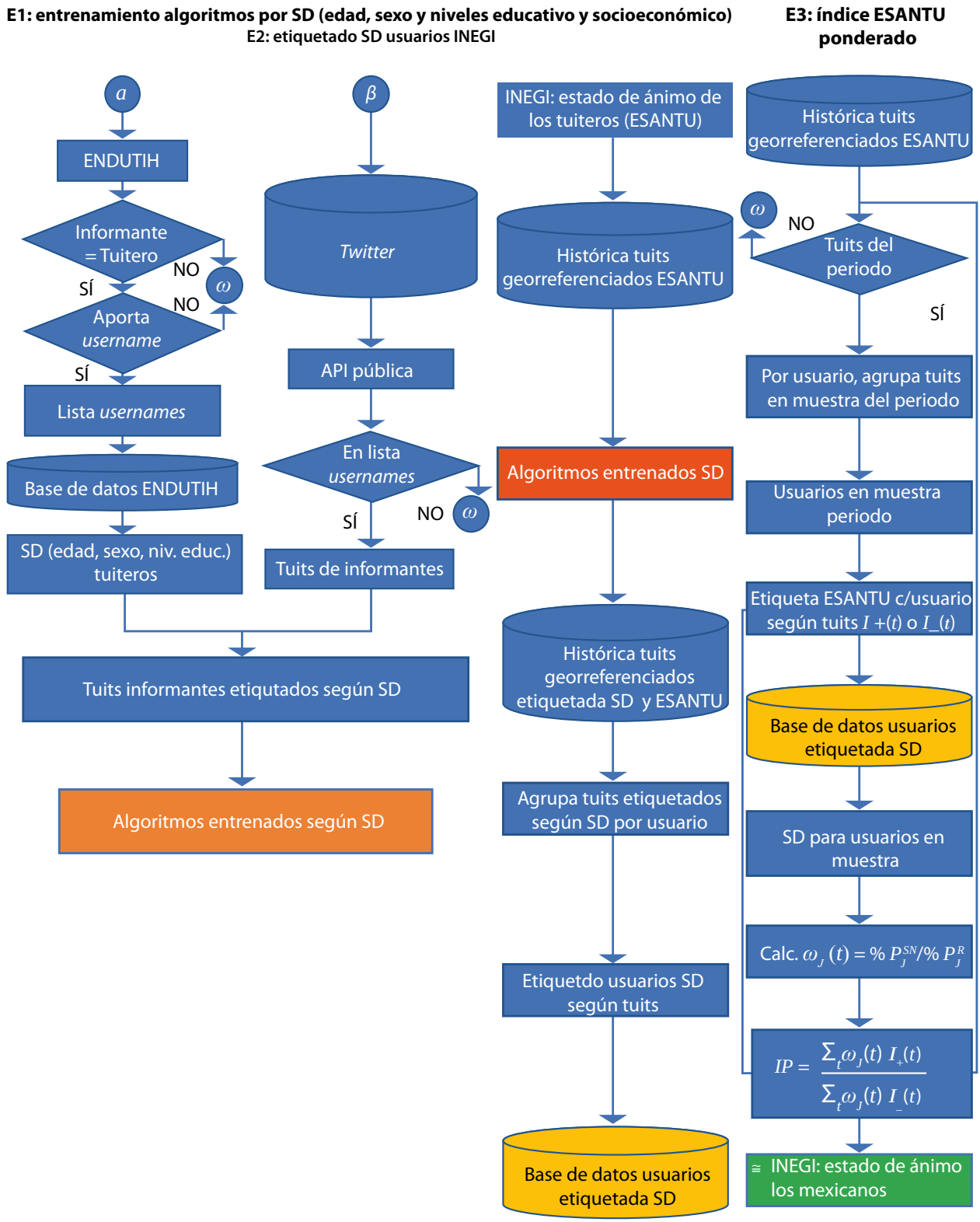
Si la calidad de los resultados lo permite, se estará en condiciones de proceder a (E2 en figura 4) etiquetar sociodemográficamente (SD) a los usuarios autores de tuits georreferenciados en la base de datos en poder del INEGI. Inicialmente, será necesario aplicar el algoritmo a los mensajes contenidos en la base de datos histórica para predecir la información SD requerida. Acto seguido, todos los mensajes etiquetados de un mismo usuario se reunirán para decidir las etiquetas que, a su vez, le correspondan de acuerdo con esa información. Ello se debe a que no hay garantía de que todos sus mensajes hayan recibido las mismas etiquetas, por lo cual habrá que determinar las que se le asignarán. A partir de este momento se tendrá un nuevo resultado intermedio en la forma de una base de datos de usuarios etiquetados según sus características sociodemográficas. Recurriremos a esta para asignar una ponderación a cada usuario incluido en las muestras por periodo y ubicación, que se usará en el cálculo de estadísticas o indicadores.

De acuerdo con el etiquetado del tuitero, será posible revisar los resultados de proyectos, como el *Estado de ánimo de los tuiteros*. Sin pérdida de generalidad, supongamos que se desea producir resultados en un día particular para una entidad federativa. Primero se determinará el estado de ánimo más probable de cada tuitero $(I_+(t) \circ I_-(t))$ con base en el etiquetado automático de sus publicaciones, de ese día y lugar, por el algoritmo en uso actualmente. Acto seguido, será determinado el tamaño relativo de cada una de las subpoblaciones $J = A,B,\ldots,Z$ representadas

10 En caso de recurrir a otra encuesta, sería necesario lograr dicha identificación.

Figura 4

Reducción del sesgo de selección en el estudio del estado de ánimo en México



Fuente: elaboración propia.

en la muestra de tuits publicados el mismo día y en el mismo lugar. A partir de la expresión (1), se determinará la ponderación que permitirá mitigar el sesgo de selección en los resultados. La relación de positividad se calculará como el cociente de la suma de ponderaciones de los tuiteros positivos entre la de las correspondientes a los negativos, según muestra la expresión (2).

Se procederá de manera similar para actualizar resultados mediante la captura diaria de tuits. Cuando se obtengan publicaciones de un usuario que no se encuentre en la base de datos, este será etiquetado en alguna subpoblación sociodemográfica y añadido a la información histórica.

Para el caso en el que el informante rehúse atender nuestra petición, estaremos ante la posibilidad de un nuevo sesgo de selección por no respuesta, semejante al estudiado en la teoría de muestreo. En esta ocasión, no obstante, la propia Encuesta aporta información suficiente que nos permite comparar entre lo que debería haberse recibido y lo que finalmente se recibió. En tanto se cuente con respuestas favorables de representantes de cada una de las clases definidas, será posible reponderar lo recibido si esto fuera necesario.[11] Cabe destacar, sin embargo, que en realidad no tenemos claro si, con el fin de entrenar un algoritmo, es o no útil reponderar los casos disponibles ni conocemos resultados en la literatura disponible que nos indiquen cómo hacerlo en la fase de entrenamiento, en su caso. Contra lo que ocurre al calcular indicadores con datos de la muestra, podemos equiparar la fase de entrenamiento de algunos algoritmos de ML (ej., redes neuronales) con el ajuste de modelos no lineales a esta. Para este último caso, no es un requisito que la muestra sea representativa, por lo que no es usual reponderar las unidades muestrales. En todo caso, se requiere que estén presentes dos o más casos para todas y cada una de las subpoblaciones.

11 Durante el proceso de revisión de este trabajo se recibieron indicaciones de que este tema ha sido estudiado en la literatura de ciencia de datos. Sin embargo, no fueron puestas a nuestra disposición las correspondientes referencias.

Seguimiento continuo a temas sociales

Además de permitirnos el etiquetado de la base de datos de usuarios de *Twitter*, el procedimiento descrito para el *Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos* nos aporta un camino a seguir para lograr la producción de información estadística con base en el uso combinado de encuestas y publicaciones en redes sociales. De esta manera, es posible sugerir que también, en el caso de otras encuestas, se solicite el permiso de acceso a las publicaciones de usuarios de *Twitter*, residentes en las viviendas seleccionadas, algunas de cuyas características en su carácter de informante han sido recogidas durante la entrevista. La identificación de aquellas con etiquetas para los textos publicados en fechas cercanas a la de la entrevista permitirá entrenar uno o más algoritmos cubriendo múltiples temas. El etiquetado automático mediante dicho algoritmo dará lugar a etiquetas adicionales para cada tuit. Bajo estas nuevas condiciones y la reponderación obtenida, será posible hacer un seguimiento estadístico frecuente a los temas estudiados por cada encuesta.

De esta manera, por ejemplo, es de esperar que cada levantamiento de la Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE) aportará etiquetas acerca del delito y sus causas, o su repercusión sobre sus víctimas. Con base en sus publicaciones, dichas etiquetas serán automáticamente asignadas a los autores que publiquen en fechas cercanas y posteriores al levantamiento, lo que permitirá dar seguimiento estadístico frecuente a esos temas. Del mismo modo, la Encuesta Nacional de Ocupación y Empleo (ENOE) aportará elementos para seguir la evolución de la ocupación, además de la de los recogidos en la hoja de datos sociodemográficos (defunciones, nacimientos, así como migración y sus causas). De manera adicional, el uso de paneles rotatorios abre la posibilidad de estudiar la evolución de la forma en que tuitea una persona cuyo estatus ocupacional cambia durante el tiempo que permanece en muestra.

En términos de la prevención del delito, la Encuesta de Cohesión Social para la Prevención de la Violencia y la Delincuencia (ECOPRED) aportaría el etiquetado acerca de los factores de riesgo y exposición a situaciones de violencia y delincuencia que enfrentan los jóvenes de 12 a 29 años de edad. Por su parte, la Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) lo haría sobre aquellas situaciones de violencia emocional, económica, patrimonial, física y sexual ejercida en contra de las mujeres de 15 años y más, ocurrida en distintos ámbitos (escolar, laboral, comunitario, familiar y de la pareja). Con ello se abre, además, la posibilidad de diseñar encuestas sobre temas emergentes, como los contemplados por algunos de los indicadores para los ODS de la ONU, o que no hemos podido estudiar con anterioridad, como la salud mental en adolescentes, para vincularlos con publicaciones en redes sociales y darles también seguimiento frecuente.

Figura 5

Proceso del seguimiento diversos tópicos de los mexicanos usando encuestas

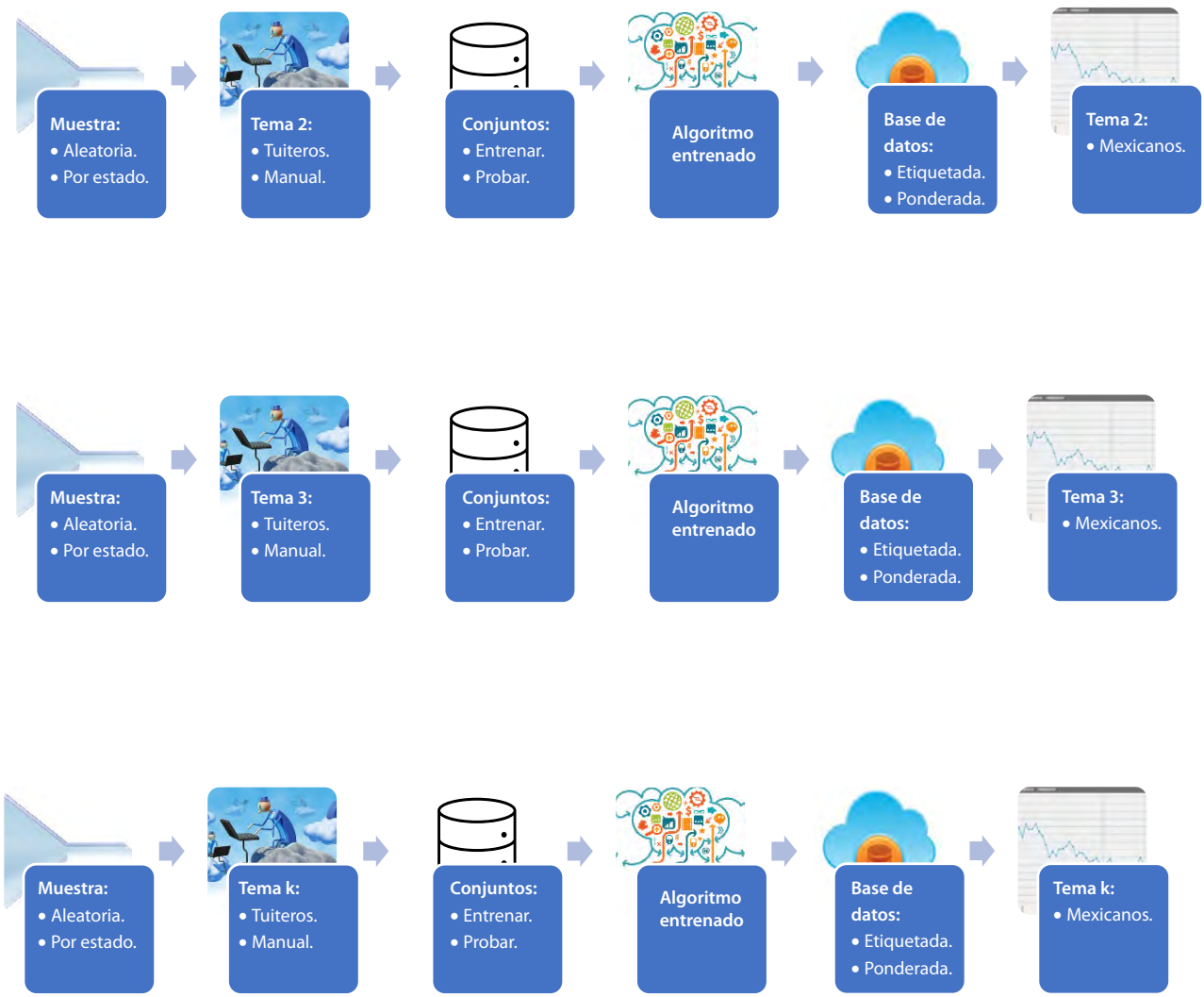
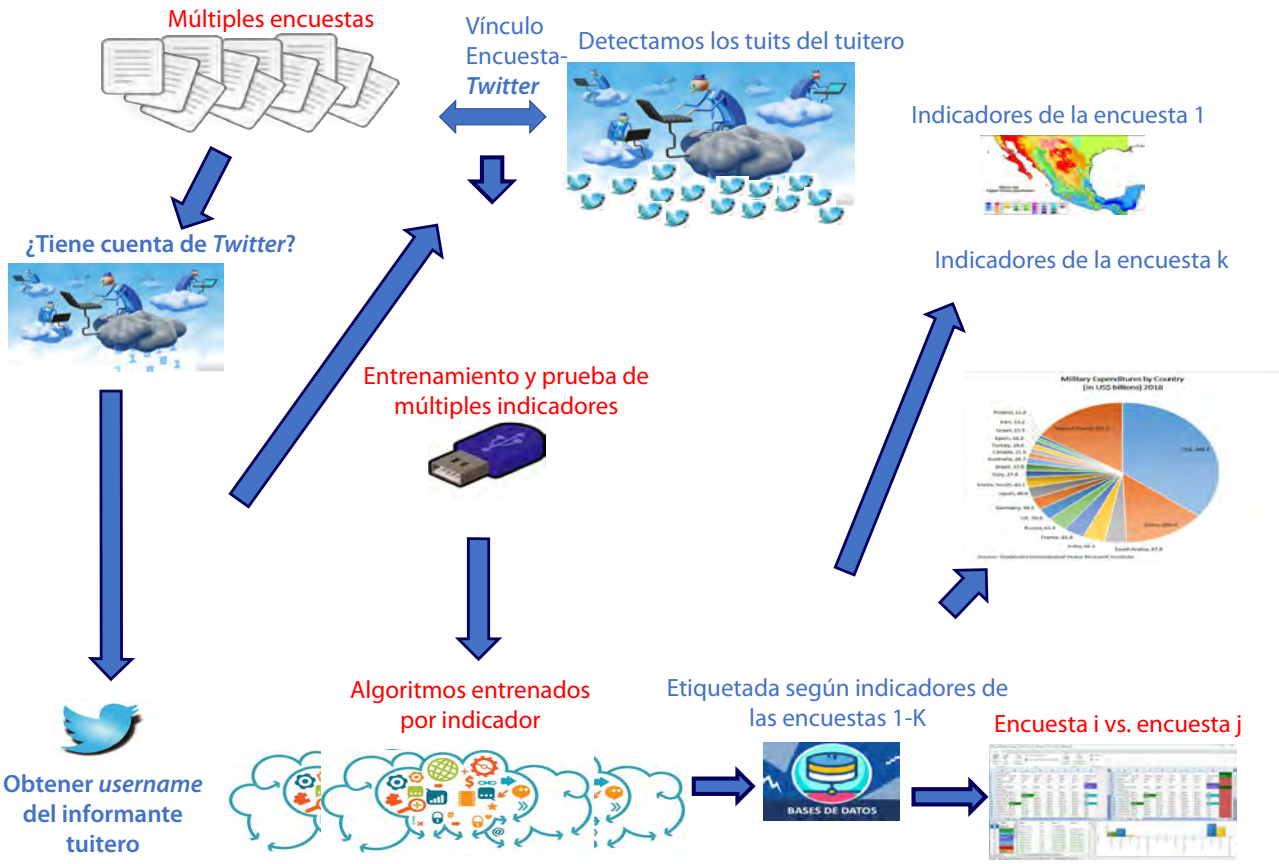


Figura 6

Proceso del seguimiento de diversos tópicos y de sus interrelaciones usando *Twitter*



Fuente: elaboración propia.

En consecuencia, cada usuario será clasificado desde muy diversos puntos de vista. De esta manera, se abre la posibilidad de relacionar las temáticas de distintas encuestas en formas hasta ahora impensables. Por ejemplo, es posible pensar en dar seguimiento a la salud mental de aquellas personas que han perdido el empleo o que han sufrido algún otro tipo de pérdida; otro podría ser el cambio en la confianza de los consumidores o víctimas recientes de algún delito. En fin, se extiende una gama interesante de posibilidades que no estaban a nuestro alcance ni en el caso del levantamiento de un censo de población.

Más aún, diversas encuestas permiten identificar cambios en el entorno inmediato del usuario/informante, como: el nacimiento de un nuevo integrante del hogar, el retorno de un migrante o el matrimonio de un pariente, que pueden influenciar la forma en la que el tuitero se expresa en la red. Nos gustaría pensar que el uso de toda o parte de esta información como etiquetas asociadas a sus mensajes permitirá entrenar con razonable precisión nuevos algoritmos. Esto nos llevaría a relacionar sus publicaciones con eventos no asociados directamente con el usuario/informante.

Comentarios finales

La experiencia acumulada en el uso de información de redes sociales ha sido invaluable para

su explotación en la producción de estadística oficial. El grupo de técnicos y profesionales que laboran en el INEGI y que se han capacitado en el uso de las técnicas relevantes alcanza ya un número importante. Asimismo, los planes para el fortalecimiento de la infraestructura muestran avances valiosos. Es preciso reconocer que a lo largo de los diferentes trabajos reportados en este documento se ha cumplido con el propósito didáctico que alentó los primeros esfuerzos. Esta propuesta es una más de las derivadas de dicha experiencia.

En ella hemos delineado un camino para el seguimiento continuo de temas sociales y demográficos de interés para las oficinas nacionales de estadística. Esta ruta se apoya en los datos recolectados dentro de los programas tradicionales de producción de información estadística oficial de cualquier ONE a través de encuestas por muestreo. De esta manera, se reduce el sesgo de selección en los indicadores con los que se dará seguimiento prácticamente continuo a diversos temas; asimismo, se reduce la incertidumbre detrás de esos indicadores, pues se evita depender de la opinión de expertos para el etiquetado de los conjuntos de entrenamiento y prueba.

En general, tanto si se trata del etiquetado en la base de datos de usuarios como si lo es del seguimiento de algún tema objeto de estudio de alguna encuesta, la hoja de ruta propuesta para optimizar la combinación de datos de encuestas con los de las redes sociales debe abarcar los siguientes pasos, como mínimo:

a) Aplicar el cuestionario de la encuesta.
b) Al finalizar el llenado, indagar si algún informante es usuario de redes sociales.
c) Solicitarle la publicación de un mensaje.
d) En su caso, vincular sus respuestas en el cuestionario con sus publicaciones.
e) Incorporarlo al conjunto de entrenamiento.
f) Entrenar un algoritmo.
g) Predecir las etiquetas para la base de datos, o para las nuevas publicaciones, a través del mismo algoritmo.
h) Elaborar agregados usando las reponderaciones correspondientes.

Parece útil sugerir una postura proactiva que informe a los usuarios de redes sociales acerca de los alcances del ejercicio que la oficina nacional de estadística se propone realizar. Se sugiere el uso de las propias redes sociales para difundir los alcances y las limitaciones en el uso de la información, así como los análisis y resultados que se derivarán de su uso. Es de esperarse que los mensajes transmitidos por esta vía faciliten la labor de los entrevistadores. Por supuesto, y para aportar una descripción más amplia de las acciones a realizar, las publicaciones de la ONE harán referencia a alguna sección en su página oficial. En ella, el informante encontrará, además, la base legal que garantice el uso con fines estadísticos de su información, y ninguno otro.

La adaptación de nuestra propuesta a la explotación de registros administrativos con fines similares no nos resulta inmediata. Para este caso, habrá que desarrollar formas de vincular información contenida en un registro administrativo con las publicaciones del ciudadano. Sin embargo, será necesario ser cautos para evitar hacer públicos datos cuyo uso inadecuado por terceros pueda implicar un daño a nuestro informante. Por ejemplo, sugerimos evitar el uso de la Clave Única del Registro de Población (CURP) utilizada en México para fines oficiales, pero cuya publicación en un tuit permitiría vincular la información de diversos registros públicos con la publicada en la red social.

Fuentes

Brakel, J.; E. van den Söhler, P. Daas y B. Buelens. "Social media as a data source for official statistics; the Dutch Consumer Confidence Index", en: *Survey Methodology*. Vol. 43, No. 2, December. Statistics Canada, 2017, pp. 183-210. Catalogue No. 12-001-X (DE) https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2017002/article/54871-eng.pdf?st=LYTvhP20, consultado el 21 de mayo de 2021.

Bustos y de la Tijera, V. A., A. A., Coronado Iruegas, S. L. Fraustro Velhagen, G. Leyva Parra, N. López Delgado, R. A. Olvera Navarro, A. M. Romo Anaya y

V. Silva Cuevas. "Caracterización del sesgo de selección en redes sociales en México a través de algunas características sociodemográficas de sus usuarios", en: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. En prensa 2022.

Data-Pop Alliance. *Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America*. Data-Pop Alliance, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute, November 2016.

Destatis. *Access to Big Data for statistical purposes* (Note by the Federal Statistical Office of Germany, Economic Commission for Europe). Paris, Conference of European Statisticians, 67th plenary session, 26-28 June 2019 (DE) https://undocs.org/ECE/CES/2019/20, consultado el 21 de mayo de 2021.

DNP. *Definición de la estrategia de Big Data para el Estado colombiano y para el desarrollo de la industria de Big Data en Colombia: estado del arte y análisis comparativo de estrategias nacionales de Big Data, noviembre de 2017* (DE) https://datapopalliance.org/wp-content/uploads/2018/09/Documento1_VersionFinal_DNP.pdf, consultado el 11 de noviembre de 2021.

Dutra. *Las organizaciones deben implementar una estrategia centrada en los datos*. 2018 (DE) https://www.telefonica.com/es/web/public-policy/blog/articulo/-/blogs/las-organizacioneSDeben-implementar-una-estrategia-centrada-en-loSDatos, consultado el 21 de mayo de 2021.

Golub, B., and M. O. Jackson. "Using selection bias to explain the observed structure of Internet diffusions", en: *Proc Natl Acad Sci USA*. Vol. 107, No. 24, June 15, 2010, pp. 10833-10836.

Iacus, S., G. Porro, S. Salini y E. Siletti. "Controlling for Selection Bias in Social Media Indicators through Official Statistics: a Proposal", en: *Journal of Official Statistics*. Vol. 36, No. 2, 2020, pp. 315-338 (DE) http://dx.doi.org/10.2478/JOS-2020-0017, consultado el 21 de mayo 21 de 2021.

INEGI. *Estado de ánimo de los tuiteros en los Estados Unidos Mexicanos. Documento metodológico V 2.0.* 2017 (DE) https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825099718, consultado el 29 de octubre de 2021.

Jansen, R. *UN Global Working Group (GWG) on Big Data and its Task Teams*. Hangzhou, China, International Symposium on the Use of Big Data for Official Statistics, National Bureau of Statistics of China, Oct. 16-18, 2020 (DE) http://www.stats.gov.cn/english/pdf/202010/P020201012399997943871.pdf, consultado el 21 de mayo de 2021.

Kim, J. K., and S. M. Tam. "Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference", en: *International Statistical Review*. John Wiley & Sons Ltd on behalf of International Statistical Institute, 2020, doi:10.1111/insr.12434.

Lokanathan, S., T. Perera-Gomez y S. Zuhyle. *Mapping Big Data Solutions for the Sustainable Development Goals* [Draft]. LIRNEasia, 2017 (DE) https://lirneasia.net/2017/03/mapping-big-data-solutions-sustainable-development-goals/, consultado el 21 de mayo de 2021.

Smith, H., *Big Data for Official Statistics*. Workshop on Big Data for Economic Statistics: Challenges and Opportunities, 11 September 2018, Rio de Janeiro, Brazil, https://unstats.un.org/Unsd/nationalaccount/workshops/2018/rio/UNSD.PDF, consultado el 11 de noviembre de 2021

Snyder, N. *UN Global Working Group on Big Data*. UNECE Workshop on Statistical Data Collection, Washington, D. C., 29 April-1 May 2015.

Struijs, P., B. Braaksma, and P. Daas. *Official statistics and Big Data, Big Data & Society*. April–June 2014, pp. 1-6, DOI: 10.1177/2053951714538417 (DE) https://journals.sagepub.com/doi/abs/10.1177/2053951714538417, consultado el 21 de mayo de 2021.

Struijs, P. y P. Daas. "Quality approaches to big data in official statistics, European Conference on Quality", en: *Official Statistics*. 2014 (DE) http://www.pietdaas.nl/beta/pubs/pubs/Q2014_session_33_paper.pdf, consultado el 21 de mayo de 2021.

Stark, T. H. "Understanding the selection bias: Social network processes and the effect of prejudice on the avoidance of outgroup friends", en: *Social Psychology Quarterly*. 78(2), 2015, pp. 127-150 (DE) https://doi.org/10.1177/0190272514565252, consultado el 21 de mayo de 2021.

Tam, S. M., and J. K. Kim. "Big Data ethics and selection-bias: An official statistician's perspective", en: *Statistical Journal of the IAOS*. 34, 2018, pp. 577-588, DOI 10.3233/SJI-170395.

Van Halderen, G., I. Bernal, T. Sejersen, R. Jansen, N. Ploug y M. Truszczynski. *Big Data for the SDGs*, *Country examples in compiling SDG indicators using non-traditional data sources*. Working Paper Series. ESCAP Statistics Division, SD/WP/12/January 2021 (DE) https://www.unescap.org/sites/default/d8files/knowledge-products/SD_Working_Paper_no12_Jan2021_Big_data_for_SDG_indicators.pdf, consultado el 21 de mayo de 2021.

UNSD. *Report of the Global Working Group on Big Data for Official Statistics*. New York, Statistical Commission Forty-sixth session, 3-6 March, 2015 (DE) https://documentSDds-ny.un.org/doc/UNDOC/GEN/N14/692/71/PDF/N1469271.pdf?OpenElement, consultado el 21 de mayo de 2021.

Anexo

Módulo Experimental Aprendizaje Automatizado basado en Encuestas en Hogares (MAAEH) 2022

Texto que será leído por el entrevistador al informante seleccionado en la ENDUTIH 2022 después de finalizar el llenado del cuestionario de la encuesta:

"Con el fin de informar a los mexicanos y las mexicanas, el INEGI explora nuevas formas de producir

estadísticas. En este nuevo módulo experimental queremos entrenar algoritmos de inteligencia artificial con su información para **establecer si la forma en la que escriben los usuarios mexicanos de *Twitter* depende de su edad, sexo, y escolaridad**. Si gracias a este proyecto se establece que sí hay relación, más adelante, usando solo los tuits que publican en el país **millones de tuiteros que NO están en la muestra de la ENDUTIH**, podremos hacer un seguimiento frecuente de la evolución de estas y otras características en la población de México.

Solicitamos su apoyo para avanzar en este proyecto de la siguiente forma: le invitamos a que nos comparta ahora su nombre de usuario en *Twitter* o, si lo prefiere, a que envíe dentro de los siguientes siete días un tuit público, o uno directo, a @INEGI_INFORMA, que incluya un número personalizado que le daré, y que solo usted y el INEGI conocerán. Si acepta nuestra invitación, esta información será la vía: a) para leer todos sus tuits de dominio público hasta esta fecha, ____ de ____ de 2022; b) para vincularlos con sus respuestas a la ENDUTIH; y c) para establecer si hay o no una relación entre los primeros y estas. Los textos de sus tuits públicos serán tratados conforme a las disposiciones del **artículo 37, párrafo primero, de la *Ley del Sistema Nacional de Información Estadística y Geográfica*** en vigor, al igual que la información que nos brindó en la ENDUTIH 2022. Los resultados de esta investigación no serán divulgados por ningún medio, ya que serán utilizados estrictamente para análisis interno del Instituto.

El número que le solicitamos que incluya en su mensaje es el XXXXXXXXXX".