Departamento Administrativo Nacional de Estadística

Diseño
Dirección de Regulación, Planificación, Estandarización y Normalización (DIRPEN)
Dirección General

METODOLOGÍA GENERAL
Estadísticas experimentales: Cálculo de información complementaria de los indicadores 16.b.1 y 16.7.2 a partir de Facebook

DANE INFORMACIÓN PARA TODOS	**METODOLOGÍA GENERAL Estadísticas experimentales: Cálculo de información complementaria de los indicadores 16.b.1 y 16.7.2 a partir de Facebook**	**CÓDIGO:** DSO-EE-MET-01 **VERSIÓN:**01 **FECHA:** 26/05/2022
PROCESO: Capacidades e Innovación	SUBPROCESO: Planificación de Proyectos de Desarrollo de capacidades e innovación definidos	
ELABORÓ: GITs Prospectiva y Análisis de Datos (DIRPEN) y ODS (Dirección General)	REVISÓ:	

TABLA DE CONTENIDO

Lista de Tablas

Lista de ilustraciones

INTRODUCCIÓN

El ritmo acelerado del uso de nuevas fuentes de información diferentes a las convencionales (censos y encuestas) como son: los registros administrativos (RR. AA), las imágenes satelitales, la información proveniente de redes sociales, la información derivada a partir de técnicas del Big Data, del machine learning, métodos bayesianos, redes neuronales; entre otras extiende un horizonte disruptivo para el aprovechamiento de fuentes secundarias de datos en la producción de estadísticas oficiales. Estas fuentes secundarias permiten medir los fenómenos económicos, sociales, culturales y ambientales de una sociedad; y amplía la narrativa al integrar esta información con la proveniente de fuentes tradicionales, como son los censos y encuestas.

A partir de los análisis, metodologías prospectivas y de la inteligencia artificial (IA) los institutos nacionales de estadística (INEs) a nivel mundial logran integrar fuentes de información tradicionales con fuentes secundarias, obtienen mediciones periódicas con un alto nivel de desagregación como herramienta estratégica que soporta las decisiones en política pública y permite comprender los fenómenos de la sociedad. De este modo, se optimiza el aprovechamiento estadístico y se hace un reúso de los datos aplicando métodos y análisis de vanguardia.

El Departamento Administrativo Nacional de Estadística -DANE- ha avanzado en la producción de estadísticas experimentales, entendidas como aquellas derivadas de proyectos en desarrollo que cuentan con aspectos innovadores, ya sea por aprovechamiento de nuevas fuentes de información, la metodología estadística utilizada o una temática nueva no medida anteriormente. Se consideran experimentales porque aún muestran margen de mejora (estandarización, cobertura y metodología) y no han alcanzado todavía la suficiente madurez en cuanto a la fiabilidad, estabilidad o calidad de los datos, como para incluirlos dentro del listado de operaciones estadísticas regulares. En cualquier caso, las estadísticas experimentales son estadísticas oficiales por el decreto 2404 del 2019, y ofrecen nuevas formas de caracterizar cuantitativamente fenómenos en las tres dimensiones del Desarrollo Sostenible: económica, social-demográfica y ambiental del país y además contribuyen a: Mejorar la disponibilidad de estadísticas relevantes y oportunas con los niveles requeridos de desagregación; incluir sectores donde se hayan identificado vacíos de información estadística, e integrar fuentes tradicionales de información, entre otros[1].

Adicionalmente, el DANE, como coordinador del Sistema Estadístico Nacional (SEN) y en el marco del proyecto de Planificación y Armonización Estadística, trabaja por el fortalecimiento y consolidación del

[1] En la página web del DANE es posible acceder a la información relacionada con los proyectos de estadísticas experimentales conforme se vayan desarrollando por la entidad, con el objetivo de ampliar o complementar el proceso de producción de estadística oficial en tres categorías temáticas: Economía, Sociedad y Territorio. Para ver la información sobre estadísticas experimentales para la temática sociedad puede acceder a: https://www.dane.gov.co/index.php/estadisticas-por-tema/estadisticas-experimentales.

SEN mediante los siguientes procesos: la producción de estadísticas estratégicas; la generación, adaptación, adopción y difusión de estándares; la consolidación y armonización de la información estadística y la articulación de instrumentos, actores, iniciativas y productos. Estas acciones tienen como fin mejorar la calidad de la información estadística estratégica, su disponibilidad, oportunidad y accesibilidad para responder a la gran demanda que se tiene de ella. Consciente de la necesidad y obligación de brindar a los usuarios mejores productos, el DANE desarrolló una guía estándar para la presentación de metodologías que contribuye a la visualización y entendimiento del proceso estadístico.

Con este instrumento la entidad elaboró los documentos metodológicos de sus operaciones e investigaciones estadísticas que quedan a disposición de los usuarios especializados y del público en general. Allí se presentan de manera estándar, completa y de fácil lectura las principales características técnicas de los procesos y subprocesos de cada investigación, lo que permite su análisis, control, replicabilidad y evaluación

1. ANTECEDENTES

El proyecto para la generación de estadísticas complementarias para los ODS 16.b.1. y 16.7.2 hace parte de la iniciativa Data For Now, la cual es codirigida por la División de Estadísticas de Naciones Unidas (UNSD), el Banco Mundial, la Alianza Global para los Datos de Desarrollo Sostenible (GPSDD) y la Red de Soluciones de Desarrollo Sostenible (SDSN). Esta iniciativa promueve la transformación de datos (fuentes alternativas y no alternativas) según necesidades o vacíos de información, de productos o de herramientas, a través de la aplicación de técnicas y métodos alternativos para su aplicación.

En el marco de esta iniciativa se generaron diferentes proyectos asociados a los Objetivos de Desarrollo Sostenible (ODS), como la estimación de la pobreza multidimensional a partir de las imágenes satelitales y del aprendizaje de máquinas. Esta medición a partir de las características lumínicas identificadas en los territorios rurales y urbanos obtuvo un patrón de pobreza multidimensional, que complementa la información oficial del Índice de Pobreza Multidimensional (IPM), medición alterna que hace aprovechamiento del alto grado de granularidad de la información del Censo Nacional de Población y Vivienda (CNPV) 2018.

De igual forma, haciendo uso de imágenes satelitales y registros administrativos, se hizo la medición de la distancia entre los centros educativos y los hogares de los estudiantes de educación básica y media y se comparó esa información con las tasas de deserción por departamento, para identificar si dichas distancias estaban incidiendo en el abandono escolar.

Esta iniciativa, entre otras, ha promovido el uso, prueba y fortalecimiento de diferentes procedimientos estadísticos, incluidas técnicas de big data, para poner a disposición de los usuarios, información estadística relevante que permita el análisis de la percepción de la discriminación y de la representatividad política en Colombia dando cuenta de los indicadores de los ODS 16.b.1 y 16.7.2, como respuesta a la falta de información generada a través de fuentes de datos tradicionales.

Si bien la Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos (ACNUDH) planteó algunas líneas de recomendación para el cálculo de estos indicadores dirigidas a la Encuesta de Cultura Política, el DANE diseñó, probó e implementó una metodología para calcular estadísticas complementarias de los indicadores 16.b.1 y 16.7.2 a partir de información obtenida de la red social Facebook, lo cual se adelantó de manera paralela con el proceso de producción estadística de la Encuesta de Cultura Política (ECP).

Para este propósito, se contó con la participación de diversos actores, desde el DANE se consolidó un equipo de trabajo con integrantes de los grupos internos de trabajo (GIT): Prospectiva y Análisis de Datos de la Dirección de Regulación, Planeación, Estandarización y Normalización (DIRPEN), y Indicadores de los Objetivos de Desarrollo Sostenible (ODS) y Seguimiento a la Agenda 2030, de la Dirección General. Adicionalmente, el equipo contó con el apoyo de tres consultores expertos en ciencia de datos y procesamiento de lenguaje natural (PLN), quienes desarrollaron parte de la metodología para la recolección y modelado de los datos, mientras que el equipo DANE se encargó de la implementación de esta estrategia. A su vez, el DANE recibió un curso de procesamiento de lenguaje natural dirigido a

los integrantes del equipo de trabajo de este proyecto y a todo el personal del DANE que tuviera actividades relacionadas con estas metodologías de trabajo.

En particular para el cálculo de estos indicadores asociados a la discriminación, se pueden destacar dos antecedentes, el primero se refiere al trabajo adelantado desde el Barómetro de las Américas, el cual establece cinco formas de discriminación como las más prevalentes en 2016: por condición económica, por color de piel, por situación de discapacidad, por sexo o género y por orientación sexual. El segundo se refiere al Observatorio de la discriminación racial. Ministerio del Interior, el cual incluye datos de personas afrocolombianas (por ejemplo, la comunidad de palenqueros). Sin embargo, no existen datos consolidados a nivel oficial sobre las diferentes formas de discriminación en el país.

2. DISEÑO DE LA OPERACIÓN ESTADÍSTICA

2.1. DISEÑO TEMÁTICO

2.1.1. Formulación de objetivos

a. Objetivo general

Obtener mediciones complementarias de los indicadores del ODS 16, asociados a la percepción de la discriminación y de la representatividad política, a través de Facebook, para los indicadores:

- ODS 10.3.1/ 16.b.1 Proporción de la población que declara haberse sentido personalmente discriminada o acosada en los últimos 12 meses.

- ODS 16.7.2 Proporción de la población que cree que la toma de decisiones es inclusiva y receptiva, por sexo, edad, discapacidad y grupo de población.

b. Objetivos específicos

- Contar con información que permita construir una estimación aproximada del indicador 16.b.1.
- Fortalecer los procesos de innovación mediante el aprovechamiento de fuentes y metodologías no tradicionales, como las redes sociales.
- Consolidar metodologías robustas para la producción de estadísticas experimentales en el DANE.
- Desarrollar ejercicios de investigación reproducible, con resultados que pueden ser compartidos en portal de datos abiertos, GITHUB y página web de ejercicios experimentales.

2.1.2. Alcance

El cálculo de información complementaria de los indicadores 16.b.1 y 16.7.2 a partir de Facebook tiene como producto un conjunto de estadísticas experimentales[2], entendidas de esta manera debido a que son el resultado de un proyecto en desarrollo, el cual tiene aspectos innovadores relacionados con el aprovechamiento de nuevas alternativas de información, como lo son las redes sociales, específicamente Facebook; y con el diseño, prueba e implementación de metodologías innovadoras en el uso de métodos de machine learning y de procesamiento de lenguaje natural para diferentes fases del proceso de producción estadística.

Para el indicador 16.b.1 relacionado con la percepción de discriminación, se tuvo en cuenta la definición de discriminación que hace parte del marco conceptual de la Encuesta de Cultura Política, a decir distinción, exclusión, restricción o preferencia que provenga de autoridades públicas o particulares que tenga por objeto o resultado impedir, menoscabar o anular el reconocimiento o el ejercicio de los derechos y libertades y favorecer la desigualdad basada en prejuicios, estigmatizaciones y estereotipos por motivos como sexo, género, orientación sexual, identidad de género o su expresión, raza, pertenencia étnica, color de piel, origen nacional, familiar o social, lengua, idioma, religión, creencia, cosmovisión, opinión política, ideológica o filosófica, incluida la afiliación a un partido, movimiento político o sindicato, posición económica, edad, estado civil, estado de salud, discapacidad, aspecto físico o cualquier otra condición o situación (Decreto 660 de 2018).

Es importante tener en cuenta que si bien en la Encuesta de Cultura Política, la percepción de discriminación se entiende como el hecho en el cual, alguien se sienta discriminado. No obstante, para el caso del proyecto de Información complementaria al Indicador ODS 16. b.1, la percepción de discriminación incluye el sentirse discriminado y el discriminar. Por eta razón, las mediciones que se presentan se refieren a comentarios que incluyen lenguaje discriminatorio, los cuales pueden ser originados por usuarios de Facebook con la intención de discriminar a otros usuarios, o pueden ser originados por usuarios de Facebook que se han sentido discriminados.

Para el indicador 16.7.2 relacionado con representatividad política, el cual mide los niveles auto declarados de "eficacia política externa", es decir, la medida en que los ciudadanos creen que los políticos y/o las instituciones políticas escucharán y actuarán en función de las opiniones de los ciudadanos de a pie. El metadato definido por Naciones Unidas, Incluye dos dimensiones, traducidas en preguntas relacionadas con la toma de decisiones inclusiva y toma de decisiones receptiva.

Para este indicador, la Encuesta de Cultura Política calcula el porcentaje de personas de 18 años y más, por sexo, según su consideración sobre si el sistema político colombiano permite a las personas como usted opinar sobre lo que hace el gobierno y si el sistema político colombiano permite a las personas como usted tener influencia en la política. Para el caso del proyecto de Información complementaria al Indicador ODS 16.7.2, las mediciones que se presentan se refieren a los usuarios de Facebook con comentarios que incluyen lenguaje relacionado con la toma de decisiones inclusiva, es decir que *tienen algo que decir sobre el gobierno*; y los usuarios de Facebook con comentarios que incluyen lenguaje relacionado con la toma de decisiones receptiva, es decir que *los políticos escuchan lo que tienen que decir.*

[2] Para consultar la definición de estadística experimental, que se implementa en el DANE, consultar: https://www.dane.gov.co/index.php/estadisticas-por-tema/estadisticas-experimentales.

2.1.3. Marco de referencia

El marco de referencia consta de dos partes, la primera sintetiza los conceptos requeridos para el entendimiento de las metodologías implementadas para el modelado de los datos; la segunda integra los conceptos relacionados con las temáticas de discriminación y representatividad política.

2.1.3.1. Conceptos relacionados con el modelado de los datos

Machine Learning

Es una rama de la ciencia de la computación y la Inteligencia Artificial (IA) orientada al entendimiento de diversos sistemas de datos mediante el desarrollo de algoritmos y la programación explícita. Se espera que la máquina aprenda de manera automática y logre identificar patrones, elaborar predicciones y tomar decisiones basadas en la experiencia. El machine learning se divide en aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje reforzado (Attal, 2021). Para la construcción de este marco de referencia nos centramos en la descripción general del aprendizaje supervisado y no-supervisado.

Aprendizaje supervisado

El objetivo de este tipo de análisis es identificar patrones dentro de un conjunto establecido de datos teniendo un input de entrada. Aunque este método tiene la ventaja de requerir menos datos de entrenamiento, lo cual facilita los procesos de análisis, a su vez tiene la desventaja de generar mayores costos asociados a los ejercicios de etiquetado (IBM, s.f.).

Los algoritmos utilizados en aprendizaje supervisado pueden ser de clasificación con detección de fraude, clasificación de imágenes, retención de clientes y diagnóstico; de regresión con previsiones, predicciones, optimización de procesos y nuevas perspectivas.

Aprendizaje semi – supervisado.

Es un enfoque del aprendizaje automático que combina una pequeña cantidad de datos etiquetados con una gran cantidad de datos no etiquetados durante el entrenamiento. Los datos no etiquetados, cuando se utilizan junto con una pequeña cantidad de datos etiquetados, pueden mejorar considerablemente la precisión del aprendizaje (Attal, 2021).

El coste asociado al proceso de etiquetado puede hacer inviable la creación de grandes conjuntos de entrenamiento totalmente etiquetados, mientras que la adquisición de datos no etiquetados es relativamente barata. El aprendizaje semi supervisado es de interés teórico en el aprendizaje automático y como modelo de aprendizaje humano.

Aprendizaje no supervisado

Se utiliza cuando se cuenta con una cantidad masiva de datos sin etiquetar y la máquina se encarga de explorar los datos en busca de posibles patrones. Recibe grandes cantidades de datos y utiliza algoritmos para extraer las características necesarias para etiquetar, ordenar, agrupar y clasificar datos iterativamente y sin intervención humana (Attal, 2021).

Los algoritmos utilizados en aprendizaje no supervisado son de reducción de dimensionalidad con obtención de rasgos, descubrimiento de estructuras, comprensión significativa y visualización de Big Data; de clustering con sistemas de recomendación, marketing orientado y segmentación del cliente. El enfoque de este proyecto fue justamente el de hacer uso de los grandes modelos de lenguajes (LLM, por sus siglas en inglés) para poder clasificar, teniendo poca información de entrada, la información requerida obtenida desde redes sociales.

Modelo de clasificación usando Transformers

El modelo Transformers es un modelo de aprendizaje profundo basado únicamente en mecanismos de atención, prescindiendo por completo de la recurrencia y las circunvoluciones clásicas de los modelos anteriores a él (Vaswani, y otros, 2017). Uno de los modelos más utilizado basado en Transformer es el modelo BERT, sigla de Representaciones de Codificador Bidireccional de Transformers. Este es un modelo utilizado para tareas de PLN, desarrollado en Google por Jacob Devlin y sus colegas en 2018. Fue entrenado con el conjunto de datos de la Wikipedia en inglés (2500 millones de palabras) y BooksCorpus (800 millones de palabras) y desde su lanzamiento ha tenido gran acogida para la realización de tareas de PLN.

Ilustración 1. Variaciones de BERT

Fuente: DANE

BERT es un transformador bidireccional de entrenamiento previo, que usó grandes cantidades de datos textuales sin etiquetar, lo cual le permite identificar patrones de cercanía semántica entre diferentes

términos. BERT utiliza dos estrategias de entrenamiento, (1) modelo de lenguaje enmascarado (MLM), es decir que el 15% de las palabras de cada secuencia se sustituye por un token [MASK]; y (2) predicción de la siguiente frase (NSP), el modelo recibe pares de frases como entrada y aprende a predecir si la segunda frase del par es la siguiente del documento original (Horev, 2018).

En el caso de este proyecto, se exploró el uso de una versión optimizada de BERT, RoBERTa, que cuenta con una metodología de capacitación mejorada, utiliza un 1000 % más de datos y potencia de cómputo. RoBERTa elimina la tarea predicción de próxima oración (NSP) del entrenamiento previo de BERT e introduce un enmascaramiento dinámico para que el token enmascarado cambie durante las épocas de entrenamiento. Usamos RoBERTa también para aplicar una clasificación de emociones, que nos sirvió como información de análisis y contexto, tal y como se explica más abajo.

Además, en el proyecto se uso XLNet, que es un gran transformador bidireccional que utiliza una metodología de entrenamiento mejorada, datos más grandes y más poder computacional para lograr mejores métricas de predicción que BERT en tareas de 20 idiomas.

Modelos de aprendizaje por transferencia (transfer learning): modelo Zero-Shot.

En aprendizaje profundo, es común que algunos modelos pre-entrenados se utilicen como punto de partida en las tareas de procesamiento de imágenes y PLN. Debido a que entrenar un modelo con gigantescos conjuntos de información requiere muchos recursos y tiempo, con aprendizaje por transferencia (transfer learning) y ajustes finos (fine tuning) se toma ventaja de procesos de aprendizaje pasados, como se muestra en la Ilustración 4.

Bajo estas técnicas, se puede aplicar un modelo de reconocimiento para categorías “objetivo” no aprendidas que no cuentan con un etiquetado en el entrenamiento, esto es, que no tienen ningún tipo de marcación por parte de humanos. Utiliza atributos de información auxiliar y transfiere información fuente con muestras marcadas. Para que sea efectivo las imágenes o textos se clasifican como vectores y se realiza una clasificación según las diferentes categorías. Funciona asociando clases observadas y no observadas a través de información auxiliar, que codifica las propiedades distintivas que se observan (Programador Click, s.f.).

Con la clasificación Zero Shot, es posible realizar:

- Análisis de los sentimientos.
- Categorización de noticias.
- Análisis de emociones.

En español hay una versión mejorada del modelo BERT que utiliza el conjunto de datos XNLI, el cual tiene una precisión del 79.9% para vinculación textual, donde la frase A implica/contradice/ninguna de las dos frases B; y clasificación utiliza dos frases para predecir una de las tres etiquetas (MARTIROSYAN, 2021).

Ilustración 2. **Representación de transfer learning y Fine Tuning**

Fuente: DANE

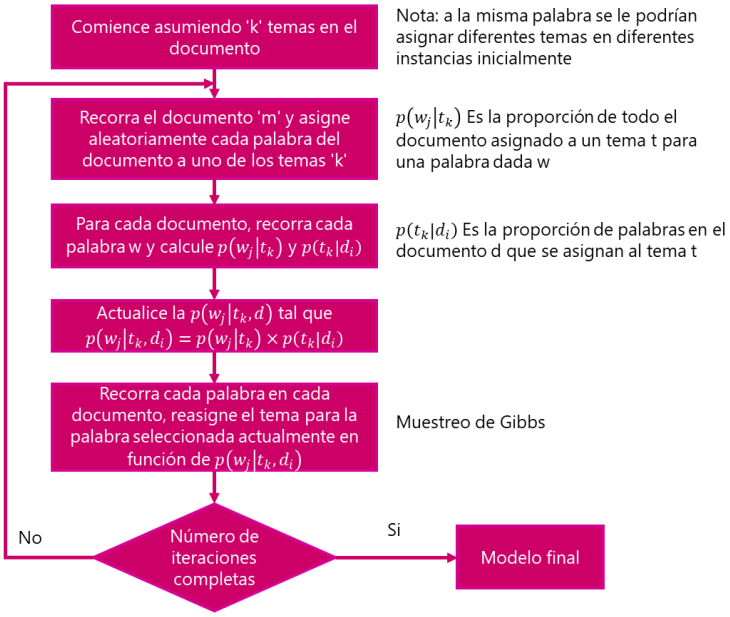
Modelos adicionales

Otros de los modelos que se exploraron dentro de este ejercicio son menos complejos que los mencionados anteriormente, pero permitieron una exploración alternativa de los datos trabajados:

- **LDA (Latent Dirichlet Allocation):** Técnica utilizada para el modelado de temas y asume que los documentos se generan utilizando un proceso generativo estadístico, de modo que cada documento es una mezcla de temas y cada tema es una mezcla de palabras (Ipshita, 2021). Hay tres hiperparámetros en LDA:
    - α → factor de densidad del documento: controla el número de temas esperados en el documento
        - β → factor de densidad de palabras temáticas: controla la distribución de palabras por tema en el documento.
        - K → número de temas seleccionados: define cuántos temas necesitamos extraer.

      Además, cuenta con los siguientes pasos generales

Ilustración 3. Pasos generales del LDA



Fuente: Tomado y traducido de (Ipshita, 2021)

- **Análisis de sentimiento:** Técnica automatizada que sirve para extraer información significativa, la cual se relaciona con actitudes, emociones y opiniones; y se determinan las emociones detrás de una serie de palabras. Es utilizada en procesamiento del lenguaje natural, análisis de texto y ciencia de datos para identificar, extraer y estudiar información subjetiva, es decir, clasifica un texto como positivo, negativo o neutral. El análisis de sentimientos se utiliza para encuestas, análisis de datos en redes sociales, posicionamiento de marca, marketing personalizado y previsión de ventas (QuestionPro, s.f.).

- **Clasificador de emociones:** Este clasificador está basado en XLM-RoBERTa y se entrenó con 138 millones de tweets; está ajustado al idioma español para clasificar ira, asco, miedo, alegría, tristeza, sorpresa, entre otros; y clasificación de sentimientos y predice el sentimiento de un corpus de reseñas, según las estrellas que obtenga entre 1 y 5, es decir 1 y 2 son negativos, 3 es neutro, 4 y 5 son positivos (MARTIROSYAN, 2021).
- **Discurso de Odio:** Es un modelo de BERT multilingüe perfeccionado para la clasificación binaria de lenguaje de odio/no odio, que ha sido adaptado también para el idioma español. Tiene una puntuación de validación de aproximadamente el 74% (MARTIROSYAN, 2021).

2.1.3.2. Conceptos relacionados con la temática de estudio

Indicador 16.b.1.

Según el metadato publicado por Naciones Unidas[3], el indicador 16.b.1. es definido como la *Proporción de la población que declara haber experimentado personalmente discriminación o acoso durante los últimos 12 meses basado en motivos prohibidos por las leyes internacionales de derecho humanos*.

Según dicha metodología, la discriminación se entiende como "cualquier distinción, exclusión, restricción, preferencia u otro trato diferencial que está directa o indirectamente basado en los motivos prohibidos de discriminación y que tienen la intención o efecto de anular o menoscabar el reconocimiento goce o ejercicio, en igualdad de condiciones, de los derechos humanos y libertades fundamentales en el ámbito político, económico, social, cultural o en cualquier otro ámbito de la vida pública".

La ley internacional de derechos humanos provee una lista de los motivos prohibidos de discriminación. La prohibición de "otros estatus" en estas listas indica que estas no son exhaustivas y que otros motivos podrían ser reconocidos por los mecanismos internacionales de derechos humanos. No obstante, en la sección del método de cálculo del metadato, la Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos (OACNUDH) recomienda el uso de los siguiente motivos prohibidos por el derecho internacional de los derechos humanos y la adición de una categoría de "cualquier otro motivo" para incluir los motivos que no se enumeran explícitamente: sexo, edad, discapacidad o estado de salud, origen étnico, color o lenguaje, estado de migración, condición socioeconómica, lugar de residencia, religión, estado civil o condición familiar, orientación sexual o identidad de género, opinión política, otros estatus. El módulo recomienda que esta lista ilustrativa se revise y contextualice a nivel nacional a través de un proceso participativo para reflejar grupos de población específicos y necesidades de recopilación/desglose de datos.

La Encuesta de cultura política[4], publicada por el DANE, la cual ha sido identificada como una de las principales fuentes para medir los indicadores del ODS 16, incluidos los indicadores (16.b.1. 16.5. 17.7.2.) para conocer la percepción de los ciudadanos colombianos ante la discriminación, incluye los mismos motivos de discriminación definidos en el metadato del indicador (Naciones Unidas), no obstante, agrega los motivos por discriminación de identidad y pertenencia cultural y rasgos físicos de su cuerpo. Por su parte el proyecto de información complementaria al Indicador ODS 10.3.1/16. b.1 incluye exactamente los mismos motivos de discriminación que incluye la encuesta de cultura política, cuya descripción se presenta en la Tabla 1.

Tabla 1.Tipos de discriminación

Tipo de discriminación-etiquetas	Descripción

[3] Consultado en: https://unstats.un.org/sdgs/metadata/files/Metadata-16-0b-01.pdf

[4] Para consultar la información publicada de esta encuesta se puede acceder a: https://www.dane.gov.co/index.php/estadisticas-por-tema/gobierno/cultura-politica

Sexo	Corresponde al sexo biológico con el que nació
Edad	Como ser percibido (a) como demasiado viejo (a)
Discapacidad	Como tener dificultades para ver, oír, caminar o moverse, concentrarse o comunicarse
Estado de salud	Tiene una enfermedad u otras afecciones de salud
Origen étnico, color o lenguaje	Como el color de la piel, la vestimenta, la cultura, las tradiciones, el idioma nativo, así como autor reconocerse como indígena o afrodescendiente
Rasgos físicos de su cuerpo	Por ejemplo, tener sobrepeso o bajo peso, tener un tatuaje, una cicatriz o una marca de nacimiento, o que su cuerpo no tenga el aspecto que normalmente se espera de las personas.
Estado de migración	Como nacionalidad, país de nacimiento, refugiados, solicitantes de asilo, estatus migratorio, inmigrantes indocumentados
Condición socioeconómica	Como la situación o el estatus de una persona teniendo en cuenta sus ingresos, trabajo, nivel educativo o propiedad de la tierra, terreno o casa
Lugar de residencia	Como vivir en zonas urbanas o rurales y establecida de manera formal (con los servicios básicos) o informal (invasión)
Religión	Como tener o no una religión o creencias religiosas o considerarse ateo
Estado civil o condición familiar	Ser soltero (a), casado (a), divorciado (a), viudo (a) o la condición de tener o no hijos (as), estar embarazada, ser huérfano (a), adoptado (a) o nacido (a) de padres solteros

Orientación sexual o identidad de género	Como sentirse atraído por una persona del mismo sexo (por ejemplo, ser lesbiana, gay) o de ambos sexos (bisexual), entre otros, o identificarse con un sexo diferente al que nació o no identificarse con ningún sexo
Opinión política	Como expresar opiniones relacionadas con su ideología política, pertenencia a partidos o movimientos políticos o defender los derechos de los demás
Identidad y pertinencia cultural	Como la forma de vestirse, la forma de hablar o de expresarse, o ser campesino (a)

Fuente: ECP, DANE

Indicador 16.7.2.

Según el metadato publicado por Naciones Unidas[5], el indicador 16.7.2. mide los niveles auto declarados de "eficacia política externa", es decir, la medida en que los ciudadanos creen que los políticos y/o las instituciones políticas escucharán y actuarán en función de las opiniones de los ciudadanos de a pie.

Para abordar las dos dimensiones cubiertas por este indicador, el indicador 16.7.2 de los ODS utiliza dos preguntas de encuesta bien establecidas, a saber 1) una pregunta que mide el grado en que las personas sienten que tienen voz en lo que hace el gobierno (se centra en la participación inclusiva en la toma de decisiones) y 2) otra pregunta que mide la medida en que los ciudadanos consideran que el sistema político les permite influir en la política (centrada en la capacidad de respuesta en la toma de decisiones).

Se debe hacer todo lo posible para desglosar los resultados de la encuesta sobre estas dos preguntas por sexo, grupo de edad, nivel de ingresos, nivel educativo, lugar de residencia (región administrativa, por ejemplo, provincia, estado, distrito; urbano/rural), estado de discapacidad, y grupos de población relevantes a nivel nacional.

Conceptos básicos del indicador

Toma de decisiones: Está implícito en el indicador 16.7.2 que la "toma de decisiones" se refiere a la toma de decisiones en la gobernanza pública (y no a toda toma de decisiones).

Toma de decisiones inclusiva: Procesos de toma de decisiones que ofrecen a las personas la oportunidad de "tener algo que decir", esto es, de expresar sus demandas, opiniones y/o preferencias a los responsables de la toma de decisiones.

[5] Consultado en: https://unstats.un.org/sdgs/metadata/files/Metadata-16-07-02.pdf

Toma de decisiones receptiva: Procesos de toma de decisiones en los que los políticos y/o las instituciones políticas escuchan y actúan en función de las demandas, opiniones y/o preferencias de las personas.

2.1.4. Definición de variables e indicadores estadísticos

El ejercicio de descarga de datos fue realizado en dos partes. En la primera parte, se desarrolló una descarga de posts de Facebook de cuentas influenciadoras de dominio público, para los que se incluyeron las siguientes variables:

post_id, text, post_text, shared_text, time, image, video, video_thumbnail, video_id, likes, comments, shares, post_url, link, user_id, username, is_live, factcheck, shared_post_id, shared_time, shared_user_id, shared_username, shared_post_url, available, images, reactions, w3_fb_url, fetched_time.

En la segunda parte, se hizo la descarga de comentarios asociados a los posts descargados anteriormente y para los cuales se consideraron las siguientes variables:

post_time, post_id, text, user, user_comment, las cuales hacen referencia al timestamp en el que se creó el post, el id del post, el texto del post, el nombre de usuario público que realizó el comentario en el post y el comentario del usuario sobre el post, respectivamente.

2.2. DISEÑO ESTADÍSTICO

2.2.1. Universo de estudio

El universo de estudio son los usuarios de Facebook. Para ello, es pertinente caracterizar los usuarios de internet a nivel nacional. Según cifras de la Encuesta de Calidad de Vida, para 2021 el 73% de los colombianos mayores de 5 años usaban internet, cifra que asciende al 79,8% para las cabeceras y un 50,5% para los centros poblados y rural disperso. Para el caso de Facebook, según el estudio “Digital 2021 Global Overview Report” realizado por *We Are Social* y *Hootsuite*, para enero de 2021, Facebook era la segunda red social más utilizada por usuarios de internet colombianos, en edades entre 16 y 64 años con un total de 93.6% de usuarios de internet, superada únicamente por YouTube en el primer puesto con 95.7% y arriba de WhatsAspp con 90.7%[6].

[6] Consultado en: https://www.slideshare.net/DataReportal/digital-2021-colombia-january-2021-v01, pág. 47.

2.2.2. Población objetivo

Buscando minimizar el posible sesgo presente al realizar la descarga de información, y considerando a Facebook como plataforma digital objetivo, se decidió realizar una revisión sobre cuáles son las páginas de Facebook con mayor cantidad de seguidores. Usando información provista por administradores de estadísticas y entidades como Socialbakers.com, se encontró que para el año 2021, figuras como Shakira, James Rodríguez, MALUMA, J Balvin, entre otros encabezaban la lista. Por otra parte, buscando proveer el mayor alcance de público objetivo se optó por identificar los perfiles/páginas/grupos de Facebook con mayor cantidad de seguidores para las categorías deportes, política, economía, orden público, artistas, figuras públicas y alcaldes.

2.2.3. Desagregación temática

ODS 10.3.1/ 16.b.1 Usuarios de Facebook con comentarios que incluyen lenguaje discriminatorio (activo o pasivo).

ODS 16.7.2 Usuarios de Facebook con comentarios que incluyen lenguaje relacionado con la receptividad política.

ODS 16.7.2 Usuarios de Facebook con comentarios que incluyen lenguaje relacionado con la inclusividad política.

2.2.4. Fuente de datos

El proyecto Información complementaria al Indicador ODS 10.3.1/16. b.1, la red social Facebook.

2.2.5. Unidades estadísticas

Unidades de observación

Para el indicador 16.b.1: Usuarios de Facebook que tienen comentarios con lenguaje discriminatorio

Para el indicador 16.7.2: Usuarios de Facebook que tienen comentarios con lenguaje asociado a representatividad política

Unidades de análisis

Para el indicador 16.b.1. Usuarios de Facebook que tienen comentarios con lenguaje discriminatorio y comentarios con lenguaje discriminatorio

Para el indicador 16.7.2. Usuarios de Facebook que tienen comentarios con lenguaje asociado a representatividad política y comentarios con lenguaje relacionado con la representatividad política.

2.2.6. Periodo de referencia

El proyecto Información complementaria al Indicador ODS 10.3.1/16. b.1 tiene como periodo de referencia el comprendido entre los años 2013-2022.

2.2.7. Periodo de recolección/acopio

Para la recolección de los datos se tuvo en cuenta 2 rangos de periodos temporales puesto que la segunda etapa dependía de la primera.

La primera etapa consiste en recolección de posts la cual fue ejecutada entre el 24 de junio de 2021 y el 10 de octubre de 2021, mientras que la segunda etapa de recolección (comentarios asociados a cada post) fue ejecutada entre 23 de Julio de 2021 y el 10 de diciembre de 2021.

2.2.8. Diseño muestral

La muestra de posts y comentarios es una muestra no - probabilística, teniendo en cuenta los criterios de selección establecidos. Sin embargo, es pertinente aclarar la información sobre los comentarios seleccionados para la estimación de los indicadores. En la Tabla 2 se puede observar en primer lugar, tanto la cantidad de comentarios descargados como la cantidad de comentarios seleccionados para el cálculo de la información complementaria de los indicadores de los ODS 16.b.1 y 16.7.2; en segundo lugar, tanto el número de usuarios del total de comentarios descargados con el número de usuarios de los comentarios seleccionados para el cálculo de la información complementaria de los indicadores de los ODS 16.b.1 y 16.7.2

Los usuarios son todos aquellos que han realizado un comentario, categorizado bajo cualquiera de las formas indicadas. Es posible que existan casos de usuarios que hayan realizado comentarios sobre más de una forma de discriminación, por lo que se considera que cada uno de estos hechos debe ser considerado como un único caso, aunque el autor del comentario discriminatorio sea el mismo. De este modo, no se pierde información asociada y se sigue la recomendación de la metodología: "El indicador debe ser un punto de partida para comprender los patrones de discriminación" (ONU, 2021: 4).

Tabla 2. Tamaño de la muestra

Información complementaria al Indicador	Total de comentarios descargados	Comentarios seleccionados	Usuarios del total de comentarios descargados	Usuarios de los comentarios seleccionados
16.b.1	719.902	8.744	503.553	8.177
16.7.2.	187.995	62.000	124.302	50.794
	583.507	275.360	405.693	219.372

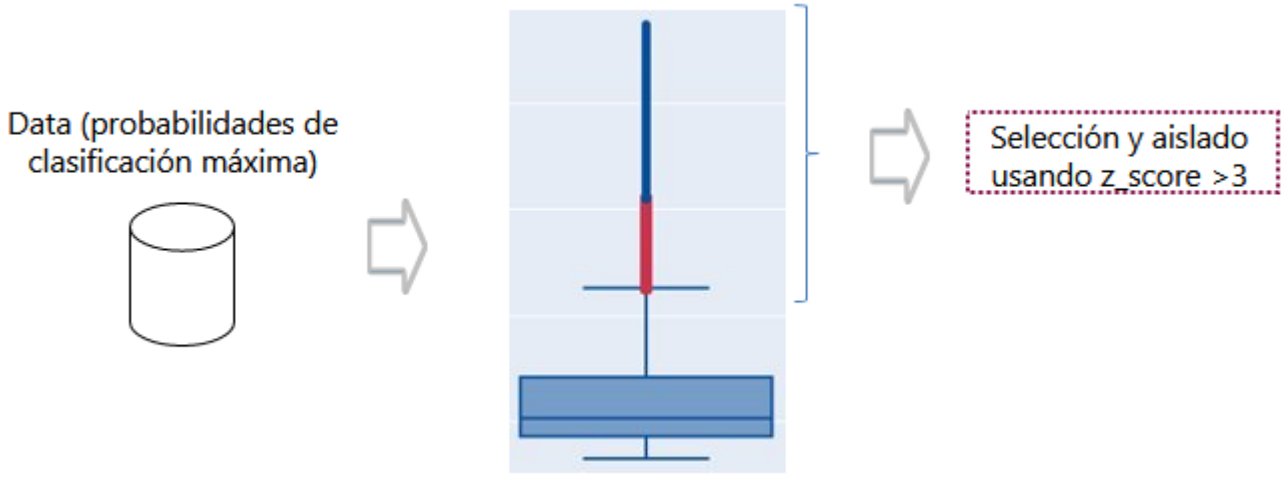
Fuente: DANE

Los comentarios y usuarios de los comentarios seleccionados para el análisis de los datos, en el caso del indicador 16.b.1 son aquellos con probabilidad > 0.5 de corresponder a uno de los motivos de discriminación, según la marcación del modelo.

Los comentarios y usuarios de los comentarios seleccionados para el análisis de los datos, en el caso del indicador 16.7.2 son aquellos con probabilidad > 0.6 de corresponder a una de las formas de representatividad política.

Estos umbrales son identificados tras hacer un análisis de distribución de probabilidad sobre los datos, el cual consta en primer lugar de un análisis de outliers mediante el diagrama de cajas y bigotes junto con un filtrado de los datos por z > 3, como se muestra en la Ilustración 5.

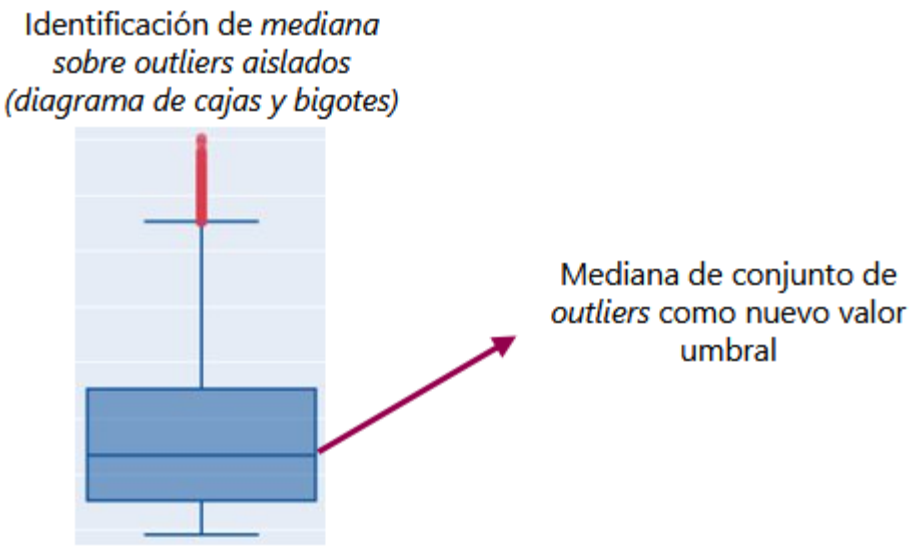
Ilustración 4. Identificación de outliers (diagrama de cajas y bigotes)



Fuente: DANE.

Teniendo en cuenta este resultado, se selecciona el conjunto de outliers y se procede a generar un segundo criterio de filtrado, esta vez apelando a la mediana del conjunto de datos atípicos, buscando así poder identificar aquellos comentarios cuya probabilidad de pertenencia a las formas de discriminación seleccionada sea más alta y, así, tener una mayor seguridad sobre que dicha clasificación es adecuada, como se muestra en la Ilustración 6.

Ilustración 5. Proceso de selección de umbral



Fuente: DANE

2.3. DISEÑO DE RECOLECCIÓN/ACOPIO

2.3.1. Métodos y estrategias de recolección y acopio de datos

Para la recolección de datos, se desarrollaron dos algoritmos de *scraping* en lenguaje de programación Python para la plataforma de Facebook haciendo uso principalmente de las librerías **facebook-scraper** y **bs4** disponibles para este lenguaje de programación.

El primer algoritmo se encarga de descargar los posts de un perfil en específico presentes en una cantidad de páginas especificadas.

Este algoritmo puede resumirse en los siguientes pasos:

1. Se definen/importan las librerías necesarias.
2. Se asigna el perfil a la función **get_posts()** de librería **facebook_scraper**.
3. Se construye el dataframe basado en la data obtenida.
4. Se almacena el dataframe con posts asociados al perfil específico en formato .csv en ubicación local.

El segundo algoritmo se encarga de acceder a cada post de los perfiles previamente obtenidos y descargar los comentarios asociados en una cantidad de páginas especificadas.

Este algoritmo puede resumirse en los siguientes pasos:

1. Se definen/importan las librerías necesarias.
2. Se genera el objeto de controlador/browser usando **selenium**.

3. Se cargan credenciales de acceso a cuenta de *scraping* para adquisición de data.
4. Se procede a iterar sobre cada url de post hasta completar la cantidad de páginas especificadas en modo móvil.
    a. En cada iteración (paginado), se asigna la url a la librería **bs4**, la cual por medio de métodos **find()** y **find_all()**, extrae la información requerida.
    b. La identificación del paginado se realiza mediante la detección del identificador "**View more comments**"
    c. En cada iteración se obtiene la dará del comentario, procediendo a construir y almacenar el dataframe obtenido en formato .csv de forma local.

En adición, como medida para mitigar posibles bloqueos por parte de la plataforma de Facebook, se hace uso de la generación de retardos aleatorios entre pasos, buscando emular comportamiento de navegación natural humana.

2.4. DISEÑO DEL PROCESAMIENTO

2.4.1. Consolidación archivo de datos

Este ejercicio experimental está basado en dos tipos de análisis, el primero es un análisis no supervisado y el segundo es un análisis supervisado.

Para el análisis no supervisado, se usó Spanish language zero-shot classifier, esta es una versión mejorada del modelo en español BERT, que refiere a la parte en español del conjunto de datos XNLI. El modelo tiene una precisión del 79,9 % para las tareas de clasificación y vinculación textual en el conjunto de datos XNLI-es. La vinculación textual asume la siguiente tarea: (la oración A no implica ni contradice la oración B), mientras que la clasificación asume la siguiente tarea (dadas dos oraciones, predice una de las tres etiquetas).

El modelo zero-shot fue utilizado en 2 ejercicios para clasificación de tipos de discriminación y 2 ejercicios de clasificación de tipo de representatividad. En cada uno de los casos, los comentarios son transformados a un embedding y concatenados junto el embedding de las etiquetas con las que se van a comparar. Al tener esto son pasadas por el modelo base (siendo BERT como se mencionó anteriormente) para finalmente obtener una clasificación binaria sobre pertenencia a cada etiqueta dada por una función de salida de tipo *softmax*.

Para discriminación, en el primer ejercicio se consideró un conjunto de 6 etiquetas, las cuales son: "discriminación económica", "discriminación política", "discriminación racial", "discriminación por ser migrante", "discriminación por discapacidad"," discriminación por orientación sexual", "discriminación por ser mujer" y "discriminación no evidenciada". Estas etiquetas el modelo nunca las ha visto (siendo parte de las razones para la selección del modelo zero-shot como clasificador).

Para el segundo ejercicio, se consideró una extensión del primer conjunto, añadiendo nuevas etiquetas para obtener finalmente un conjunto conformado por: "discriminación económica", "discriminación política", "discriminación racial", "discriminación por ser migrante", "discriminación por discapacidad",

“discriminación por orientación sexual”, “discriminación por ser mujer”, “discriminación por sexo”, “discriminación por edad”, “discriminación por estado de salud”, “discriminación por rasgos físicos de su cuerpo”, “discriminación por lugar de residencia”, “discriminación por credo”, “discriminación por estado civil o condición familiar”, “discriminación por identidad y pertinencia cultural” y “discriminación no evidenciada”.

Para representatividad, en el primer ejercicio se consideró un conjunto de 2 etiquetas, las cuales son: “esto es inclusividad política” y “esto es receptividad política”. Por otra parte, para el segundo ejercicio se usaron como etiquetas: tengo algo que decir sobre el gobierno y los políticos escuchan lo que tengo que decir. Al igual que para discriminación, ninguna de esas etiquetas era conocidas con antelación por el modelo.

Con respecto al análisis supervisado se condujo un ejercicio de muestreo aleatoriamente de 1600 comentarios para realizar etiquetado manual, con el cual mediante una estrategia de triple ciego a través de cálculo de coeficiente Kappa, sea posible obtener un indicador de concordancia y confianza para las futuras muestras. De estos 1600 comentarios, 600 fueron seleccionados para conformar la base de discriminación y 1000 para la base de representatividad. Cada comentario en estos ejercicios fue revisado y clasificado por 3 anotadores. En adición, así como los ejercicios no supervisados se realizaron en 2 etapas tanto para discriminación como para representatividad (que incluye inclusividad y receptividad), en este caso fueron etiquetados los mismos comentarios para cada base en 2 momentos diferentes siendo el segundo conducido tras el desarrollo de sesiones de calibración con los equipos de anotación.

2.4.2. Diccionario de datos

Para las bases obtenidas por *scraping*, se generaron dos diccionarios de datos.

El primer diccionario de datos corresponde a la base de posts, como se detalla en la Tabla 3.

Tabla 3. Diccionario de datos para la base de posts

Variable	Tipo	Descripción	Ejemplo
post_id	string	id del post	'2257188721032235'
text	string	Texto del post	-
post_text	string	Texto del post	-
shared_text	-	-	-
time	timestamp	timestamp en el que se creó el post	-
image	string	URL de la imagen encontrada en el post	-
video	-	-	-
video_thumbnail	-	-	-

video_id	-	-	-
likes	numérico	Cantidad de likes del post	1052
comments	numérico	Cantidad de comentarios encontrados	73
shares	numérico	Cantidad de veces que compartieron el post	-
post_url	string	URL del post	-
link	string	URL del post	-
user_id	string	id del autor del post	-
username	string	Nombre del autor del post	-
is_live	booleano	-	-
factcheck	-	-	-
shared_post_id	-	-	-
shared_time	-	-	-
shared_user_id	-	-	-
shared_username	-	-	-
shared_post_url	-	-	-
available	booleano	Si el post está disponible o no	True
images	lista	Lista de URL de imágenes del post	['https://scontent.fhlz2-1.fna.fbcdn.net/v/t1.6435-9/fr/cp0/e15/q65/58745049_2257182057699568_1761478225390731264_n.jpg?_nc_cat=111&ccb=1-3&_nc_sid=8024bb&_nc_ohc=ygH2fPmfQpAAX92ABYY&_nc_ht=scontent.fhlz2-1.fna&tp=14&oh=7a8a7b4904deb55ec696ae255fff97dd&oe=60A36717']
reactions	diccionario	Reacciones y cantidad encontradas en el post (información disponible si la opción “extra_info” está activa.	{'haha': 22, 'like': 2657, 'love': 706, 'sorry': 1, 'wow': 123}
w3_fb_url	string	URL del post	-

fetched_time	-	-	-

Fuente: DANE

El segundo diccionario de datos corresponde a la base de comentarios, como se detalla en la Tabla 4.

Tabla 4. Diccionario de datos para la base de comentarios

Variable	Tipo	Descripción
post_time	timestamp	Tiempo en el que se creó el post
post_id	numérica	id del post
text	string	texto del post
user	string	nombre de usuario público que realizó el comentario en el post
user_comment	string	comentario del usuario sobre el post

Fuente: DANE

2.4.3. Diseño para la generación de los cuadros de salida

Los cuadros de salida son parte esencial para la publicación de resultados del proyecto de mediciones complementarias para los indicadores IDS 16.b.1 y 16.7.2. En la Tabla 5 se listan los principales cuadros de salida, diferenciados para cada una de las variables para las cuales se dispone información:

Tabla 5. Cuadros de salida

Información complementaria al Indicador	Variable	Indicador
16.b.1	Comentarios	Total de comentarios que incluyen lenguaje discriminatorio, con probabilidad>0.5, por motivo de discriminación (Análisis no supervisado-Modelo Zero Shot Classifier-segundo ejercicio)
		Total de comentarios que incluyen lenguaje discriminatorio, con probabilidad>0.5, por año, por motivo de discriminación (Análisis no supervisado-Modelo Zero Shot Classifier-segundo ejercicio)
		Total de comentarios que incluyen lenguaje discriminatorio, con probabilidad>0.5, por motivo de discriminación, para mayo y junio de 2021 (Total de comentarios que incluyen lenguaje discriminatorio, con probabilidad>0.5, por motivo de discriminación, para mayo y junio de 2021)
		Comparación del Porcentaje de comentarios con lenguaje discriminatorio, por motivo de discriminación (Etiquetado manual para modelo supervisado-Análisis no supervisado segundo ejercicio)
	Usuarios	Porcentaje de usuarios con comentarios que incluyen lenguaje discriminatorio, con probabilidad>0.5, por motivo de

		discriminación (Análisis no supervisado-Modelo Zero Shot Classifier-segundo ejercicio)
		Porcentaje de usuarios con comentarios que incluyen lenguaje discriminatorio, con probabilidad>0.5, por motivo de discriminación, por sexo (Análisis no supervisado-Modelo Zero Shot)
		Proporción de usuarios cuyos comentarios se asocian a alguna de las formas de discriminación reconocidas.
		Comparación del porcentaje de usuarios con comentarios con lenguaje discriminatorio, con probabilidad> 0.5, con los resultados de la Encuesta de Cultura Política
16.7.2	Comentarios	Total de comentarios relacionados con representatividad política (Análisis no supervisado-Modelo Zero Shot Classifier-primer ejercicio)
		Total de comentarios para representatividad política, con probabilidad>0.6, por año (Análisis no supervisado-Modelo Zero Shot Classifier-primer ejercicio)
		Porcentaje de comentarios para representatividad política (Análisis no supervisado-Modelo Zero Shot Classifier-segundo ejercicio)
		Porcentaje de comentarios para representatividad política, con probabilidad>0.6 (Análisis supervisado- Etiquetas segundo ejercicio)
	Usuarios	Porcentaje de usuarios con comentarios para representatividad política, con probabilidad>0.6, por sexo (Análisis no supervisado-Modelo Zero Shot Classifier-primer ejercicio)
		Comparación del porcentaje de usuarios con comentarios relacionados con representatividad política, con probabilidad> 0.6, con los resultados de la Encuesta de Cultura Política (Análisis no supervisado, segundo ejercicio)

Fuente: DANE

2.5. DISEÑO DE LA DIFUSIÓN Y COMUNICACIÓN

2.5.1. Diseño de sistemas de salida

Los documentos de publicación de la estadística experimental cálculo de información complementaria de los indicadores 16.b.1 y 16.7.2 a partir de Facebook son el Boletín técnico, la presentación de resultados, los cuadros de salida (anexos) y el tablero de control[7].

[7] El tablero de control es una herramienta que permite visualizar la información producida en el proyecto, basado en la infraestructura de Azure Cloud Services, utilizando la herramienta PowerBi. Para el despliegue e implementación de la infraestructura referenciada se contó con el apoyo de la Consejería de la Presidencia para

Una vez el equipo temático, la coordinación del equipo ODS y la Dirección técnica de Dirpen los han revisado y verificado, estos se envían a un ambiente de pruebas que simula la página web del DANE, con el objetivo de verificar la clara disposición de los resultados y la plena usabilidad del tablero de control.

Estos documentos de publicación se disponen en estadísticas por tema, en la temática Sociedad, en estadísticas experimentales, bajo el título Cálculo de información complementaria de los indicadores 16.b.1 y 16.7.2 a partir de Facebook.

2.5.2. Diseñar los productos de comunicación difusión

El proceso de elaboración de los productos de difusión inició con el diseño de los cuadros de salida por parte del equipo temático, luego estos fueron procesados, contrastados y validados en diferentes escenarios, tanto al interior del DANE como con aquellos actores del sistema estadístico internacional que fuesen relevantes para el estudio, por ejemplo, en Colombia la Consejería de la Presidencia para los DDHH como con ONU Derechos Humanos.

2.5.3. Entrega de productos

La comunicación y promoción de la disponibilidad de los productos generados por esta estadística experimental se realizan mediante la página web del DANE, redes sociales y presentaciones especiales de acuerdo con la solicitud de los usuarios.

2.5.4. Entrega de servicios

El equipo de trabajo de la estadística experimental da soporte a las dudas e inquietudes y solicitudes de los usuarios externos o internos, los cuales envían a la Dirección o dependencias sus solicitudes, mediante correos electrónicos o por medio del Orfeo. La entidad da respuesta puntual dentro del menos tiempo posible sin exceder el legal vigente.

2.6. DISEÑO DE LA EVALUACIÓN DE LAS FASES DEL PROCESO.

La red social Facebook cuenta con la participación de individuos que publican o comentan contenidos, los cuales pueden tener un lenguaje asociado a algún motivo de discriminación o de representatividad en el contexto de la vida política y social de Colombia, no obstante, se ha notado que esta red social es altamente dinámica y presenta múltiples desafíos tanto para la descarga de datos, como para medir la representatividad de los resultados del estudio, con relación a la población colombiana.

---

los derechos humanos.

Con el objetivo de monitorear la calidad de los datos descargados, así como de los modelos desarrollados para calcular los indicadores proxy para discriminación y representatividad política, se cuenta con una estrategia de calidad que abarca desde la fase de descarga de los datos de Facebook y pasa por la verificación de la calidad de los modelos implementados.

En la presente sección, se profundiza sobre los ejercicios adelantados para verificar la calidad del modelado, en particular, a través del cálculo de algunas métricas de desempeño para medir el rendimiento del modelo implementado, la presentación de los resultados en la verificación de la representatividad de los resultados y el despliegue de algunos modelos adicionales de procesamiento de lenguaje natural, los cuales arrojan información contextual útil para el contraste de los proxys calculados.

2.6.1. Métricas de desempeño.

Con el objetivo de medir el desempeño del modelo Zero Shot para los indicadores ODS 16.b.1 y 16.7.2, se establecieron un conjunto de métricas de desempeño: exactitud (accuracy), precisión (precision), recall y f1-score. Debido a que el problema que enfrenta el modelo Zero Shot no consiste en una clasificación binaria, para las métricas precision, recall y f1-score, se establecieron sus respectivos equivalentes: micro, macro y weighted.

Adicionalmente, se estableció como una de las métricas para medir el desempeño, la matriz de confusión, con el fin de visualizar la distribución de la clasificación en términos de Positiva TP, FP, TN y FN. Para hacer el cálculo de estas métricas, se utilizan las bases anotadas como “ground truth”. En primer lugar, se utiliza el filtro *nan*, de tal manera que el cálculo de las métricas se haga sobre aquellas muestras diferentes de *nan*, es decir 541 muestras a evaluar. En segundo lugar, se contrasta el resultado contra su equivalente identificado por el modelo de zero-shot.

A continuación, se precisa el método de cálculo de cada una de las métricas:

Accuracy: La exactitud se calcula como la fracción de predicciones que el modelo realizó correctamente sobre el número total de predicciones (1):

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \quad (1)$$

Donde:

t_n = Verdadero negativos

t_p = Verdaderos positivos

f_n = Falsos negativos

f_p = Falsos positivos.

Estas métricas se pueden encontrar en la matriz de confusión, también conocida como matriz de errores, que no es más que una representación gráfica del diseño de una tabla específico que permite visualizar el rendimiento de un algoritmo.

Precision es la razón entre verdaderos positivos divididos entre el total de verdaderos positivos y falsos positivos (2).

$$precision = \frac{t_p}{t_p + f_p} \quad (2)$$

Micro: Esta calcula la métrica global contando el total de verdaderos positivos, falsos negativos (f_n) y falsos positivos.

Macro: Esta calcula las métricas para cada etiqueta y encuentra su media no ponderada. Cabe notar, que no tiene en cuenta el desbalanceo de las etiquetas.

Weighted: Esta calcula las métricas para cada etiqueta y encuentra su media ponderada por el soporte (el número de instancias verdaderas para cada etiqueta).

Recall: El recall se calcula como la razón de verdaderos positivos divididos por el total de verdaderos positivos y falsos negativos (3).

$$recall = \frac{t_p}{(t_p + f_n)} \quad (3)$$

F1 score: se busca obtener el cociente de dos varianzas, donde su varianza corresponde a la medida de precisión que tiene un test. Esta es expresada como el producto entre el precision y el recall dividido por la suma entre precision y recall (4).

$$F1 = 2 \cdot \frac{precision \cdot recall}{(precision + recall)} \quad (4)$$

Cabe notar que para problemas multiclase y multilabel es la media de la puntuación F1 de cada clase con una ponderación que depende del parámetro medio.

2.6.2. Representatividad de los datos

Un factor importante al considerar la idoneidad del propósito de los datos de Facebook para el análisis de indicadores de los ODS en Colombia es la composición demográfica de la base de usuarios que genera los datos de Facebook. Si difiere significativamente de la población general de Colombia, esto indicará una advertencia importante en su idoneidad del propósito como fuente de información indirecta.

Para caracterizar la composición demográfica de los usuarios de Facebook asociados a los comentarios con lenguaje discriminatorio, se desplegó una estrategia para estimar las variables sexo y edad de los usuarios seleccionados, las cuales deberían ser calculadas con las mismas variables de la población general de Colombia para tener una idea acerca de la representatividad de los datos calculados en el presente estudio.

A continuación se indica el procedimiento para la estimación de la variable sexo. Para el caso de la variable edad, no se cuenta con la información suficiente para adelantar un ejercicio de modelamiento que garantice la representatividad de los datos.

2.6.2.1. Estimación de la variable sexo

El objetivo de estimar la variable sexo es dar insumos para responder a la pregunta sobre la representatividad de los resultados del proyecto, mediante la caracterización de la composición demográfica de los usuarios de Facebook asociados a los comentarios con lenguaje discriminatorio. Asimismo, determinar si los resultados de este proyecto son comparables con los resultados de la pregunta sobre la percepción de discriminación, presentados por la Encuesta de Cultura Política del DANE.

En la recolección de datos para Facebook la incidencia de la variable sexo era muy baja, menor al 10% de los usuarios identificados y disponible para los usuarios influenciadores identificados, los cuales no eran objeto de estudio en este proyecto. En esa medida, se hacía necesario plantear una estrategia que garantizara la obtención de esta variable. Para su estimación, se aplicó una estrategia para predecir esta variable haciendo uso del primer nombre de una persona, en este caso, los usuarios de Facebook asociados a los comentarios con lenguaje discriminatorio o con la representatividad política.

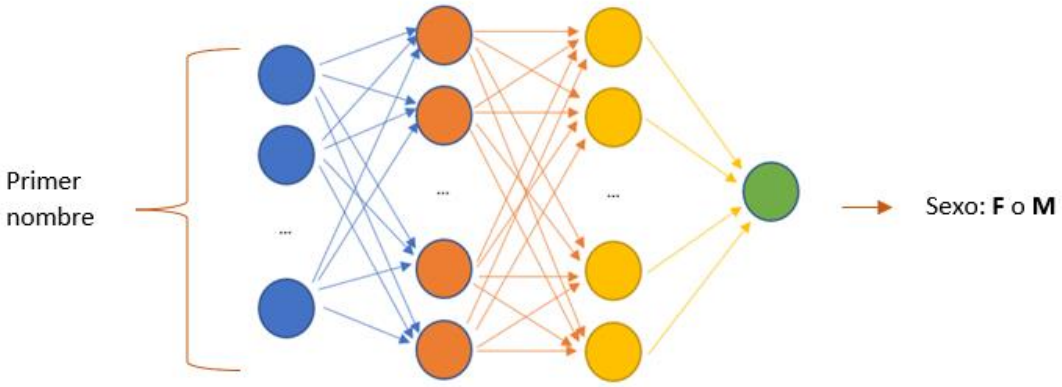
En primer lugar, se hizo una revisión de algunos modelos predictivos pre-entrenados y aunque los desempeños de predicción de estos modelos fueron relativamente altos, padecen de la limitación o sesgo fundamental de ser entrenados con nombres anglosajones o europeos, razón por la cual se procedió a entrenar un modelo propio a partir de los nombres, etiquetados con su respectivo sexo, del Registro Base Estadístico de Población (REBP) del DANE.

El conjunto de datos REBP contiene 63.442.084 registros. Cada registro contiene el primer y segundo nombre de una persona colombiana y una etiqueta de su sexo. El 50.15% de registros están etiquetados como sexo masculino, el 49.85% como sexo femenino y el 0.002% como “nan”.

Con el objetivo de verificar la precisión del pronóstico se adelantaron dos ejercicios experimentales, el primero correspondió a un ejercicio simplificado, en el cual se omitió la columna “segundo nombre”, se eliminaron los registros que no tenían etiqueta de sexo y aquellos registros con más de 15 caracteres o menos de 3 en la columna de primer nombre. El segundo correspondió a un ejercicio que en la fase de preparación de los datos no omite la columna “segundo nombre” y procede con la clasificación entre hombre o mujer.

Como modelo de línea base se entrenó un clasificador “Naive Bayes” y se obtuvo una precisión de 69% sobre el conjunto de prueba. Posteriormente, se ajustaron los hiperparámetros de una red neuronal LSTM (Long Short Term Memory) de una capa oculta, mediante la técnica de búsqueda en grilla en el espacio de los hiperparámetros, como se muestra en la ilustración 7.

Ilustración 6. Esquema del modelo LSTM



Fuente: DANE

El modelo con el mejor desempeño obtuvo una precisión de 88.6% sobre el conjunto de validación y 88.4% sobre el conjunto de prueba. Estos resultados indican un modelo con una generalización aceptable, por lo que se aplicó al conjunto de datos. Al analizar algunos de los errores de predicción del modelo, se encontró que algunos nombres en los que el modelo se equivocaba al predecir el sexo son difíciles, incluso para un humano.

2.6.3. Análisis de información contextual

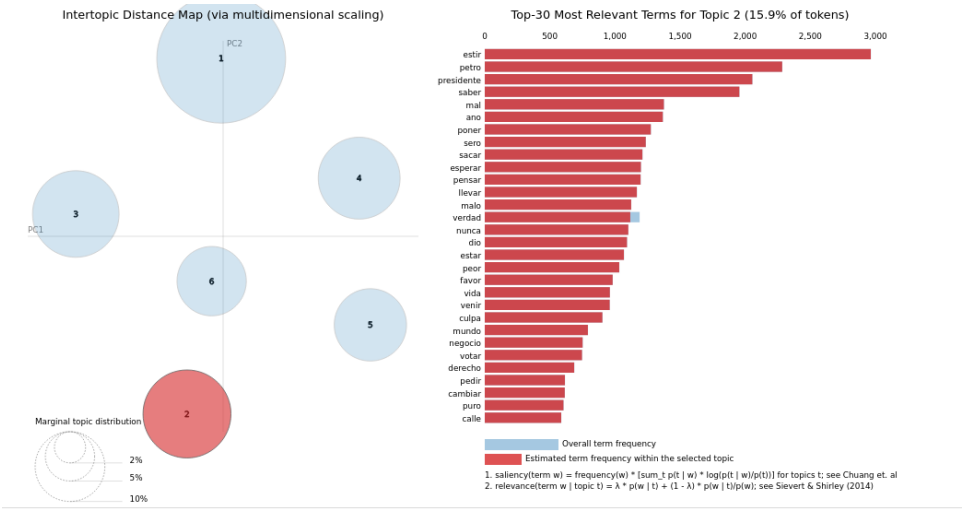
2.6.3.1. Modelamiento de tópicos usando Latent Dirichlet Allocation (LDA)

Dentro de las estrategias planteadas para el análisis contextual de la información se construyeron tres modelos en paralelo: el primero de ellos fue el de modelamiento de tópicos, con el cual se busca dar una contextualización sobre los temas identificados en el conjunto de formas de discriminación, así como en el conjunto de datos en general.

Para ello se utilizó la técnica de modelamiento de tópicos usando un espacio de semántica latente bajo el modelo LDA. Este modelo es un modelo generativo de clasificación, que funciona de manera similar a un modelo probabilístico de semántica latente, solamente que en este caso asumen que la clasificación sigue una distribución a priori de Dirichlet.

El modelo LDA es, básicamente, un clasificador de temas dentro de un conjunto de datos, de tal forma que permite a los investigadores agrupar términos y oraciones en “temas” o “tópicos” que el modelo determina usando las relaciones semánticas definidas matemáticamente en un espacio vectorial. Los resultados de esta aplicación pueden verse en la Ilustración 8 para el caso de la discriminación económica.

Ilustración 8. Temas identificados en el conjunto de comentarios de Facebook asociados a discriminación económica.



Fuente: DANE. GIT ODS

Este modelo se evalúa bajo dos parámetros: coherencia, indica que los tópicos incluyen términos cuya similitud semántica es representativa y perplejidad, que mide el ajuste entre los resultados de aplicar el modelo a un conjunto de entrenamiento en relación con un conjunto de prueba. Basado en el puntaje de coherencia, se calcula, por tópicos, aquellos con mayor valor de coherencia, para definir la cantidad de tópicos.

La Ilustración 8 muestra en detalle los términos encontrados para los primeros 6 grupos de temas, para discriminación económica:

Ilustración 9. Términos identificados en los principales tópicos identificados.

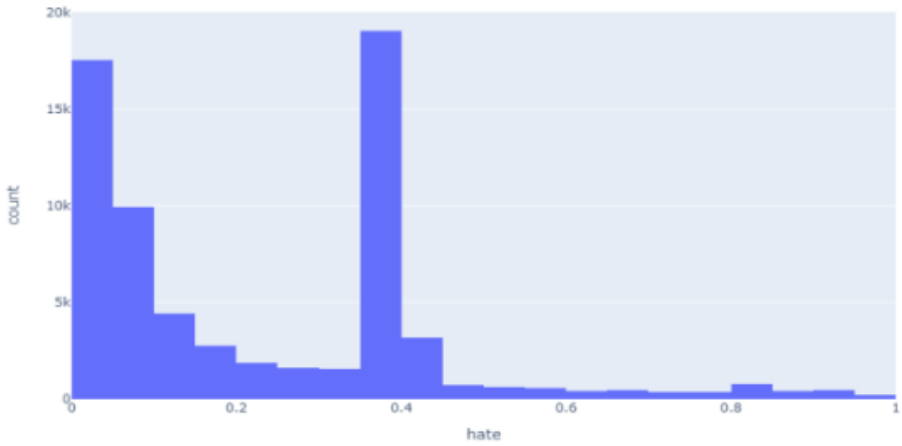
```
[(0,
  '0.022*"subir" + 0.018*"dinero" + 0.017*"gobierno" + 0.017*"robar" + '
  '0.012*"igual" + 0.012*"perder" + 0.011*"dia" + 0.010*"nino" + '
  '0.009*"semana" + 0.009*"generar"'),
 (1,
  '0.043*"dar" + 0.020*"tambien" + 0.019*"quedar" + 0.014*"llegar" + '
  '0.011*"hoy" + 0.011*"casa" + 0.011*"tiempo" + 0.011*"vivir" + 0.009*"aqui" '
  '+ 0.009*"parecer"'),
 (2,
  '0.026*"estir" + 0.020*"petro" + 0.018*"presidente" + 0.017*"saber" + '
  '0.012*"mal" + 0.012*"ano" + 0.011*"poner" + 0.011*"sero" + 0.011*"sacar" + '
  '0.010*"esperar"'),
 (3,
  '0.027*"siempre" + 0.023*"pai" + 0.019*"ir" + 0.018*"paro" + '
  '0.017*"colombiano" + 0.017*"politico" + 0.016*"huevo" + 0.015*"corrupto" + '
  '0.014*"precio" + 0.011*"partido"'),
 (4,
  '0.041*"hacer" + 0.020*"ver" + 0.019*"solo" + 0.018*"decir" + 0.017*"mas" + '
  '0.015*"gente" + 0.014*"pagar" + 0.014*"persona" + 0.013*"bien" + '
  '0.013*"pueblo"'),
 (5,
  '0.041*"querer" + 0.037*"dejar" + 0.014*"jugar" + 0.012*"hijo" + '
  '0.010*"policia" + 0.010*"noticia" + 0.009*"apoyar" + 0.009*"primero" + '
  '0.009*"cambio" + 0.008*"tecnico"')]
```

Fuente: DANE. GIT ODS.

2.6.3.2. Discurso de odio

Para el caso del modelo de discurso de odio, se implementó un modelo clasificador de discurso de odio sobre una muestra de 67.500 comentarios. El modelo evaluó la probabilidad de que el texto contenga incitación al odio. La Ilustración 10 muestra un histograma de la puntuación de odio. Se observa una distribución multimodal, en la que la mayor parte del texto tiene una puntuación baja o media, y unos pocos muestran una presencia alta de incitación al odio.

Ilustración 10. Histograma para el modelo de identificación de discurso de odio



Fuente: DANE

En la Ilustración 11 se pueden ver las nubes de palabras para los diferentes comentarios, en las que se identifica para los niveles bajo y medio el término hacer, mientras que para el nivel alto se identifican los términos *policía, corrupto, colombiano* y *hacer*.

Ilustración 11. Resultados de la clasificación del discurso de odio

Puntaje bajo	Puntaje medio	Puntaje alto

Fuente: DANE.

2.6.3.3. Análisis de sentimiento

Para el caso del análisis de sentimiento, se corrió un modelo que distinguía tres clases de opinión: positiva, negativa y neutra.

En la Ilustración 12 se pueden ver las nubes de palabras para cada clase y la distribución de estas puntuaciones, que se refieren a la probabilidad de pertenencia a la clase de cada comentario. El modelo clasificó la mayoría de los comentarios como neutros y positivos.

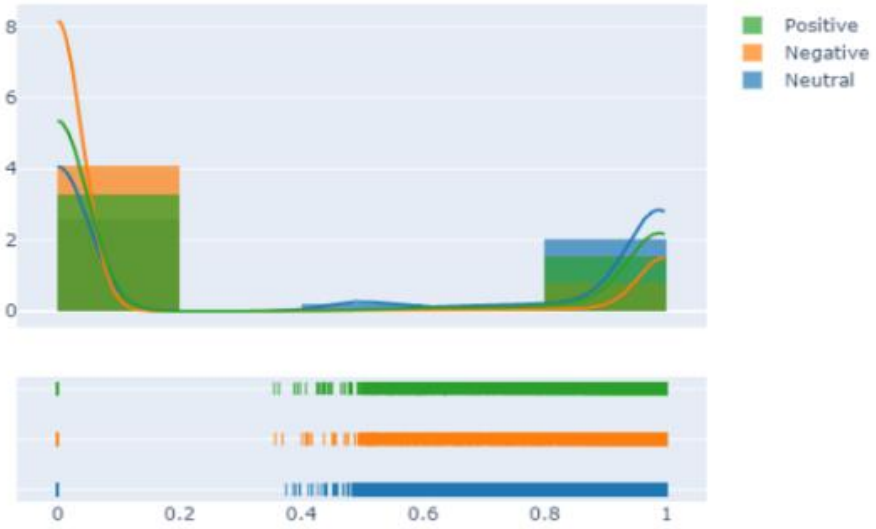
Ilustración 12. Resultados del análisis de sentimiento.

Negativo	Neutro	Positivo

Fuente: DANE. GIT ODS.

En relación con la Ilustración 13, se puede ver que hay una correlación entre la puntuación de la incitación al odio y la puntuación del sentimiento, pero no podemos ver correlación con la opinión negativa. En cambio, observamos una correlación moderada con la opinión neutral. Esto podría explicarse porque es el sentimiento más común. Estos resultados son relevantes porque nos permiten validar que el clasificador de la incitación al odio no describe el mismo patrón que el sentimiento negativo, contribuyendo así a distinguir la información entre el discurso de odio y el sentimiento neutro.

Ilustración 13. Resultados del análisis de sentimiento



Fuente: DANE.

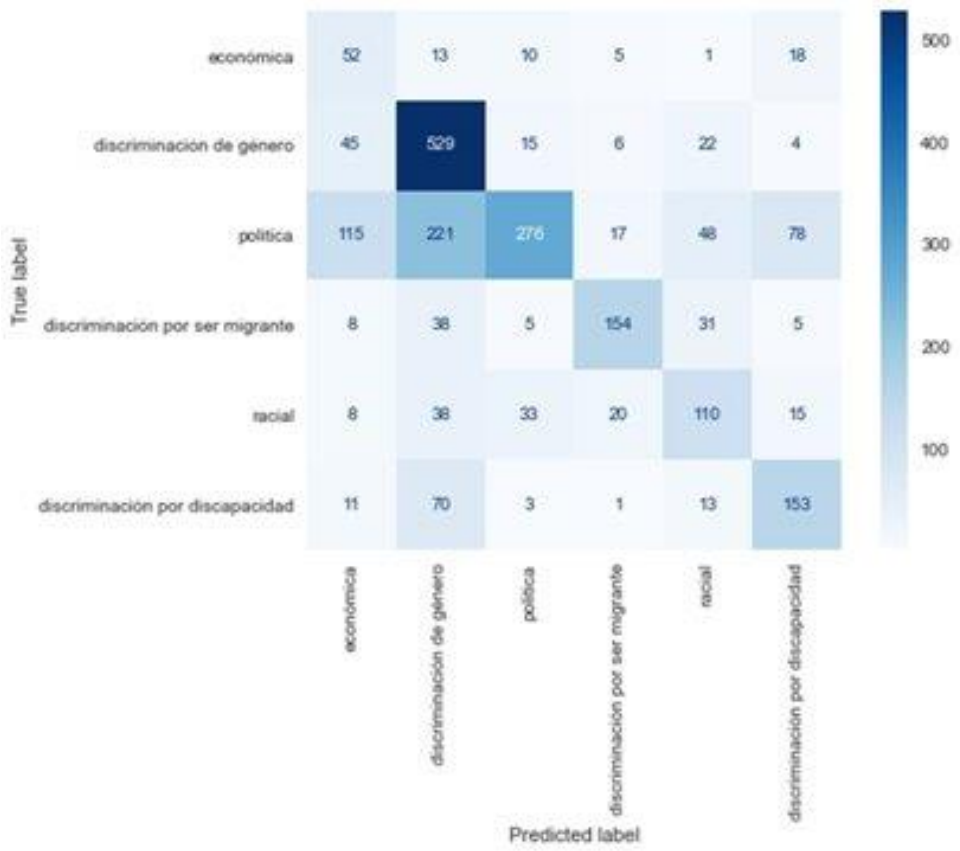
2.6.3.4. Google News

Además de Facebook, los medios de comunicación son una fuente viable de datos lingüísticos. El lenguaje generado en los medios de comunicación es menos volátil que el lenguaje generado en Facebook, por lo tanto, puede ser una fuente de datos de alta calidad que puede calificar y proporcionar contexto para los conocimientos generados desde Facebook.

Por esta razón, se adelantó un ejercicio para aprovechar la información contenida en Google News, el cual contó con el mismo marco de referencia otorgado a través de las etiquetas definidas según los tipos de discriminación. En primer lugar, se hizo una descarga noticias en español para el periodo 2015 a 2021, las cuales contaron con una confirmación por parte de un equipo de etiquetadores quienes verificaron si las noticias descargadas correspondían a la etiqueta.

En segundo lugar, se aplicó el modelo de Clasificación Zero Shot a partir de noticias marcadas con las formas de discriminación, las cuales contaron con un ejercicio adicional de validación de calidad, a través de la construcción de una matriz de confusión para evaluar el desempeño del modelo. En la Ilustración se puede ver esta matriz.

Ilustración 8. Matriz de confusión



Fuente: DANE

Dado que el proceso es muy dependiente de las etiquetas, se hicieron experimentos para mejorar la precisión de las noticias descargadas y comprobar cómo se comportaba el modelo ante estos cambios.

La tendencia de predicción por artículo tiene el siguiente comportamiento:

- La discriminación por ser migrante tiene una precisión del 76%, con un f1 score de 69%.
- La discriminación económica es la clase que representa mayor reto en este experimento porque es difícil distinguirla entre los otros tipos de discriminación.

BIBLIOGRAFÍA

Attal, M. (Diciembre de 2021). *Machine Learning: definición, funcionamiento, usos*. Obtenido de Data Scientest: https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

García, G. (Julio de 2021). Modelo de transformers para la clasificación de texto. Madrid, España: Univesdidad Politécnica de Madrid. Obtenido de https://oa.upm.es/68623/1/TFM_GUILLEM_GARCIA_SUBIES.pdf

Horev, R. (Noviembre de 2018). *BERT Explained: State of the art language model for NLP*. Obtenido de Towards Data Science: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

IBM. (s.f.). *¿Qué es Machine Learning?* Obtenido de IBM: https://www.ibm.com/co-es/analytics/machine-learning

Ipshita. (Julio de 2021). *Topic Modelling using LDA*. Obtenido de Analytics Vidhya: https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13

Khan, S. (Septiembre de 2019). *BERT, RoBERTa, DistilBERT, XLNet — which one to use?* Obtenido de Towards Data Science: https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8

MARTIROSYAN, V. (Diciembre de 2021). ACCELERATING IMPLEMENTATION OF DATA FOR NOW IN COLOMBIA AND SENEGAL. Yerevan, Armenia.

Programador Click. (s.f.). *¿Qué es el aprendizaje zero-shot (aprendizaje zero-shot) artículo uno?* Obtenido de https://programmerclick.com/article/15781581584/

QuestionPro. (s.f.). *Análisis de sentimiento. Qué es y cómo realizarlo*. Obtenido de https://www.questionpro.com/blog/es/herramienta-de-analisis-de-sentimientos/