CITIZEN SCIENCE:
THEORY AND PRACTICE

Civil Society Data for Sustainable Development Goal 16 Monitoring: A Case Study of the Use of Social Networks for Measuring Perception of Discrimination

COLLECTION:
CONTRIBUTIONS OF
CITIZEN SCIENCE TO
THE UN SDGS

CASE STUDIES

VICTOR AREVALO CABRA
KAREN CHÁVEZ QUINTERO
ANDRÉS D. PÉREZ
GRACE TORRES PINEDA
JULIETH SOLANO VILLA
VAHAN MARTIROSYAN

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

In recent years, the use of nontraditional data sources in statistical production has been increasing, given the additional need for more timely and disaggregated data. In the scope of nontraditional sources, citizen science represents an innovative approach to filling data gaps and including citizens as part of the recent innovation processes of national statistical offices (NSOs) in both the production of statistics and its role as data stewards of the national statistical systems (NSSs). The National Statistical Office of Colombia (DANE, acronym in Spanish) has structured a project within the framework of the Data4Now initiative for Sustainable Development Goal (SDG) indicators from SDG 16, using social networks as an alternative source of data generated by citizens, and natural language processing (NLP) to extract actionable intelligence from this data. In this paper, we describe the proposed strategy followed by DANE to estimate SDG indicators 16.b.1 and 16.7.2 as a pilot exercise in the framework of experimental statistics. Preliminary results illustrate potential use cases of unconventional information streams to analyze social phenomena through virtual environments such as online social media, and raise compelling challenges regarding representativity and quality assurance based on the statistical standards used by NSOs.

CORRESPONDING AUTHOR:

Victor Arevalo Cabra

Departamento Administrativo Nacional de Estadística (DANE), CO

vaarevaloc@dane.gov.co

KEYWORDS:

Facebook; discrimination; experimental statistics; citizen-generated data; Sustainable Development Goals; natural language processing; machine learning

TO CITE THIS ARTICLE:

Arevalo Cabra, V, Quintero, KC, Pérez, AD, Pineda, GT, Villa, JS and Martirosyan, V. 2023. Civil Society Data for Sustainable Development Goal 16 Monitoring: A Case Study of the Use of Social Networks for Measuring Perception of Discrimination. *Citizen Science: Theory and Practice,* 8(1): 38, pp. 1–16. DOI: https://doi.org/10.5334/cstp.590

INTRODUCTION

The 2030 Agenda for Sustainable Development was adopted by 193 countries in the 70th Session of the United Nations General Assembly. With the adoption of this Agenda, the United Nations member states and multilateral organizations recognized the need for data to monitor the progress towards the 169 targets and 17 goals. This global framework for the Sustainable Development Goals (SDGs) is composed of 231 unique indicators (adding sub-indicators and disaggregation, the total number of indicators goes beyond one thousand) (United Nations 2022). In this context, policy needs to be agreed upon as the most relevant global Agenda until 2030, increasing the demand for quantitative measures on a scale not previously experienced, causing additional challenges for national statistical systems (NSSs) worldwide, and opening opportunities for innovations in sources and methods to satisfy additional data needs not met with traditional sources and methods.

Considering the additional data needs for SDGs adopted at a global level and defined as part of the national SDGs policy, defined in the Public Policy Document CONPES 3918 of 2018, the National Statistical Office of Colombia (DANE, acronym in Spanish) has been working with custodian agencies on strategies to fill data gaps (DNP 2018). Two of these identified gaps are SDG 16 indicators: SDG 16.b.1, Proportion of the population that declares having personally felt victim of discrimination or harassment in the previous 12 months on grounds of discrimination prohibited by international humanitarian law (United Nations 2018); and SDG 16.7.2, Proportion of population who believe decision-making is inclusive and responsive, by sex, age, disability, and population group (United Nations 2023).

DANE has deployed a strategy to measure some aspects of these indicators, using traditional sources like the Victimization Survey; the Coexistence and Citizen Security Survey, carried out every two years; and the Political Culture Survey (ECP, by its acronym in Spanish), the latter of which has become a barometer to measure the perception of the impact of public policies on the consolidation of democracy in the country (DANE 2021). However, the periodicity of this information affects a wider description of the discrimination phenomenon in the country, as no annual data is available from DANE, and no other national institution reports this information for the forms of discrimination as defined in the metadata. This also occurs in other parts of the world. Defined as "any distinction, exclusion, restriction or preference or other differential treatment that is directly or indirectly based on prohibited grounds of discrimination, and which has the intention or effect of nullifying or impairing the recognition, enjoyment, or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life" (United Nations 2018), only 31 countries have reported information on discrimination over the period 2014–2019. In that period, one in five persons reported having personally experienced discrimination on at least one ground of discrimination prohibited by international human rights law (OHCHR 2020)—on the one hand, a high number, considering the number of countries that report. On the other hand, this number of countries represents a challenge in terms of the global report of the phenomenon.

According to Nicolas Fasel, chief statistician at UN Human Rights, "States need to tackle discrimination more comprehensively and address its overlapping and cumulative forms as well as its consequences on everyday life. The collection of disaggregated data, using a human rights approach is a first step that can go a long way to tackling this" (OHCHR 2020).

Given this context, citizen-generated data like social networks represent a potential statistical data source for the measurement of this phenomenon. In the Colombian case, the adoption of social networks such as Facebook is high, as much as the internet penetration in the country. According to the National Quality of Life Survey, in 2021, internet usage in Colombia corresponds to 79.9%, and the frequency of internet use for 76% of people aged 5 and over corresponds to every day of the week. In that year, there were 39 million users of social media in the country (around 78% of the total population), and the percentage of people from 14 to 64 using Facebook as its main social platform was 91.4%. In 2021, the potential Facebook audience was 36 million people (Datareportal 2022).

This use of citizen-generated data from social media is understood as a wider set of processes and techniques, which includes the well-known concept of citizen science (Haklay et al. 2021, p. 30), which is defined herein as "people, who are not professional scientists, taking part in research, i.e., co-producing scientific knowledge. This involves collaborations between the public and researchers/institutes but also engages governments and funding agencies" (Haklay et al. 2021, p. 18). Therefore, citizen-generated data from social media shares some of the same issues as citizen science, such as the needs for scientific standards, for ethical considerations, and for data management, among others (Heigl et al. 2019, p. 3). These concerns are as old as citizen science is. According to Droege (in Bonney et al. 2009, p. 978), public participation in scientific research, at least for fields such as bird watching, dates back to the end of the 19th century and the beginning of the 20th century, and it includes participation in different stages of the scientific work such as collecting, processing, and analysis, as well as the assessment of the results.

It is worth noting that some ethical considerations arise from embracing this definition: Citizen-generated data, as defined above, including data collected from informed and uninformed citizens, poses questions on data privacy, on the right of the public to be informed about the use of data, and on the active role citizens play in the new conception of scientific endeavor. The poor quality of results and statistical relevance are among the other concerns in this field (Pateman and West 2017, p. 3).

It should also be emphasized that social media is a space where people can express themselves freely. It is usually thought of and used as a place to make people feel free to express themselves, air their grievances (Miller et al. 2020), engage in self-identification in a broader public sphere (Lee 2019), and find community through online contact (Mancini and Imperato 2020) in order to create safe spaces from discrimination, especially for population groups that historically have been victims of bigotry, like the LGBTIQ+ community (Marciano and Antebi-Griszca 2020). But these groups also feel discrimination attitudes at distinct levels in social media, as posed in Lucero (2017), even online intergroup contact makes individuals more sensitive to detect discrimination (Marciano and Antebi-Griszca 2020). In this context, discussions on Facebook regarding discrimination might reflect the perception of people in the offline world, as is suggested by Marciano and Antebi-Griszca (2020), including possible associated mental health issues.

Because of this, social networks like Facebook can be one of the tools to understand marginalized communities (NETWORK 2022). One of the unique factors of internet communication is anonymity (Amichai-Hamburger and Furnham 2007, pp. 1041–1042), which created a protective environment for people to express themselves. The authors write, "the protective cloak of anonymity allows people to share aspects of the self online with far fewer costs and risks" (Amichai-Hamburger and Furnham 2007, p. 1038). In a quantitative study, Mancini and Imperato (2020, p. 9) found similar results, for online intergroup contact on Facebook makes people more attentive to detecting sexual discrimination.

However, little research has been done to address the problem of discrimination as a natural language processing problem in social media. Some studies are qualitative and focused on several population groups: For example, social media has been used as a new space to humiliate the Dalit community in India, with hate speech used against them, and legal repercussions following (Sajlan, 2022); and there has been discrimination against the Muslim community on Facebook (Awan 2016). Another study is that of Ben-David and Matamoros (2016), who have studied political violence and its different characteristics in Spain, including discrimination. This focus is based on Latour's network-actor theory, in which humans and nonhumans have agency which implies that different technological devices such as the like and share buttons play a prominent role in the identification of the different aspects of political discrimination in Spain.

From the quantitative point of view, the literature related to hate speech is vast, but the differences between this subject and forms of discrimination are not clearly defined. It is worth mentioning the paper by Marciano and Antebi-Griszca (2020), in which different types of discrimination (e.g., political or sexual identity) are identified as prevalent in several contexts like Facebook interactions, online dating, and the offline world. This is contrary to the results of Lucero (2017), who reports that the LGBTQ population feels this social network is a safe place to interact with some other members of the community. Mancini and Imperato (2020) also used Facebook as their data source, studying the behavior of different online groups in that network to understand the process by which online intergroup contact makes individuals more sensitive to discrimination (p. 8). Brooks, Shmargad, and Williams (2018) researched the discrimination directly from the algorithms, studying how the bias, the lack of data, and the audits inform a clear picture of how data systems and algorithms could, in fact, make discriminatory decisions against people.

As can be seen so far, no study addressed the use of Facebook as a statistical data source for official information about discrimination, especially as a data source to estimate SDGs indicators, first and foremost following the people's lived experience in the definition of metrics associated with SDGs (Pateman and West 2017).

Therefore, the question that motivates our study is whether Facebook is a useful and feasible source from which to generate official statistics, both broadly speaking and specifically on discrimination. To address this question, we propose a deep learning methodology to obtain complementary measurements for both 16.b.1 and 16.7.2 SDG indicators from Facebook data, which can be used to contrast and complement information for Colombia's Political Culture Survey.

The remainder of this paper is organized as follows: The Methods section explains in detail the proposed method and strategies. The Dataset section presents the dataset and preprocessing strategy. Experimental Evaluation shows the experimental evaluation of the method for both SGD indicators 16.b.1 and 16.7.2, the experimental setup, results, and discussion. Finally, Conclusions and Future Work reports the main conclusions and future work.

METHODS

The concepts, tasks, and models associated with language modelling are vast and they include different discrete and probabilistic models such as n-grams models, vector semantics, neural language models, and deep learning approaches to language processing. Key to any of these methods is the notion that data quality is dynamic and changes as the data undergo transformations. The same data can be both an output from a data source, as well as a source of data. Therefore, the methodology consists of two principal components. On one hand, data collection concentrates on measures taken to assess and increase the quality of the data collection process. It concerns the resilience and reliability of data gathering procedures and the fitness–for–purpose of the source of data for the relevant analysis. On the other hand, data quality assessment consider the reputation or believability of the source of data in question, as most data quality assessment methodologies do. This includes aspects of privacy and data access for scientific purposes.

Our methodology concentrates on the quality of language classification models employed to extract information from the text in Facebook posts and comments. This is a significant data quality bottleneck when working with social media data, given the lack of large, labeled datasets and the high level of entropy in language data available in social media. This methodology seeks to address these constraints by providing a framework for more accurate and flexible language modeling that can at once be used to generate large, labeled datasets more affordably and quickly. It is worth noting that some of the activities of our methodology follow the guidelines of the CRISP – DM process (Chapman et al, 2000) and its newer updates (IBM, 2021).

DATA COLLECTION

The primary data collection method is data scraping. This technique is based on automated browsing that allows for simulation of a user's behavior and collection of the data visualized on the screen (Mancosu and Vegetti 2020, p. 6). The tool proposed as part of this methodology is a Facebook automation bot that only collects posts and comments from public Facebook pages and profiles. The bot is coded in the Python programming language and uses web browser automation software to browse Facebook. The profile pages selected for data collection were chosen based on their relevance to the political environment in Colombia and curated manually to represent diverse viewpoints and backgrounds.

It is important to note that no data from private profiles were collected, following the ethical and privacy considerations associated with the use of citizen-generated data for academic and statistical production. This aspect was also analyzed considering the Colombian legal framework, specially the 1993 Statistical Act and the Decree 2404 of 2019 that define the concept of alternative data sources for statistical production, which includes social media, and establish its conditions of use (DANE 2019).

STRATEGY

Annotated Spanish-language datasets for several types of discrimination were not available during the time of the experiment. Furthermore, the cost of building a large custom dataset to train neural networks for discrimination detection was prohibitive in the context of the exercise. Therefore, pre-trained large language models were used for text classification, in a technique called zero-shot text classification. The pre-trained large language model was a version of the popular BERT neural network, which was already trained on massive quantities of Spanish-language text. The model used has an accuracy of 79.9% for textual entailment (determining whether two statements are contradictions, entailments, or neutral to one another) and topic classification. These scores were obtained using the popular XNLI-es dataset. A small subset of the data was then sub-sampled and annotated manually, to measure the performance of the zero-shot approach.

In order to minimize the potential impact of inaccurate predictions using the pre-trained model, outlier analysis and benchmarking are carried out using the confidence scores for each prediction made by the model. An adequate confidence threshold is determined to ensure model confidence for discrimination classification, as explained in detail in the section "Outlier analysis."

For the labelled comments approach, a random sample from the original dataset was extracted and manually annotated by DANE's experts. The labelling took place in three iterations, each by a different annotator, in order to ensure an appropriate agreement between annotators reviewed by Cohen's Kappa Score as explained in detail in the section "Labeling."

ZERO-SHOT MODEL

Zero-shot learning (ZSL) is a machine learning paradigm whereby a pre-trained model is used to predict labels that it did not explicitly see during the training process. The zero-shot classification model (Pushp 2017) extends inference to new categories without prior explicit semantic information. This methodology used a version of the BERT neural network that was finetuned on the Spanish portion of the XNLI dataset.

OUTLIER ANALYSIS

The purpose of this analysis is to obtain a confidence coefficient for the filtering and selection of those classifications in which the model is more certain to belong to a specific category. To obtain this coefficient, those values above three standard deviations over the mean of the classification probability coefficients are identified. Once these values (outliers) have been detected and isolated, the median of this set is calculated. This coefficient will be selected as the candidate threshold of confidence threshold when ensuring that a sample classified by the model corresponds to that category.

LABELING

For both 16.b.1 and 16.7.2 SDG indicators, the same tagging strategy was defined. This strategy consists of randomly extracting *n* samples from the comments dataset and tagging the same comments by three different annotators until an acceptable inter-annotator agreement is obtained. This inter-rater agreement level is identified through the calculation of Cohen's kappa coefficient, which is a statistic used to measure inter-rater reliability (and intra-rater reliability) for qualitative (categorical) items.

The Label Box platform was used to generate collaborative annotations. Three independent projects per annotation set containing the same samples were created. This was to ensure independence between annotators to minimize tagging bias.

DATASET

SCRAPING

The scraped dataset contains 771,502 records of public Facebook comments, obtained from different users, mainly from Colombia. The dataset presents wide variability, seeking to mitigate the latent bias given by the very nature of the data source. To achieve this objective, 66 profiles of public figures were considered, and categorized as follows: artists, economy, government, mayors, news, politics, public opinion, public order, sports, and others. From these profiles, posts were collected between the periods of June and October 2021. Once these posts were obtained, comments on these posts made between July and December 2021 were also collected. The selection of these collection periods has no reason beyond the periods of data collection during the research.

In addition, from this main set, a sample of 1,000 random comments was filtered to obtain 541 testing samples used as "ground truth" for evaluation purposes. Finally, within the anonymized version of the dataset, variables such as the date on which the post was made, the text of the post, and the user's comment were included.

PREPROCESSING

A standard preprocessing strategy was used, consisting of special character removal, Unicode symbol removal, and lowercasing. This preprocessing was performed both on the text coming from the post and the text corresponding to the users' comments.

EXPERIMENTAL EVALUATION

EXPERIMENTAL SETUP

Since a pre-trained zero-zhot model was used to generate predictions, the default values with a raw representation of the information were used to set up a classification baseline. Two exercises were carried out corresponding to the SGDs indicators under study: perception of discrimination and representativeness, respectively. In addition, each one of these exercises has two sub exercises in which the proposed target labels were as follows:

- Discrimination:
  - A: "discriminación económica," "discriminación política," "discriminación racial," "discriminación por ser migrante," "'discriminación por discapacidad," "discriminación por orientación sexual," "discriminación por ser mujer," and "discriminación no evidenciada."
  - B: "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," "discriminación no evidenciada."
- Representativeness:
  - A: "esto es inclusividad política" and "esto es receptividad política."
  - B: "tengo algo que decir sobre el gobierno" and "los políticos escuchan lo que tengo que decir."

The predictions were generated using a Nvidia GeForce RTX 3080 card with an approximate execution time of 16.03 hours for each proposed exercise. Taking the 771,502 records available, a total of 503,553 users have been identified but only 8,177 (corresponds to just 1% of the total records and 2% of the identified) users have been selected for the analysis, based on the outlier's analysis mentioned above (see the section "Outlier analysis"). This is due to the performance of the model, with a low metrics associated with discrimination as is shown below (see the section "Model performance").

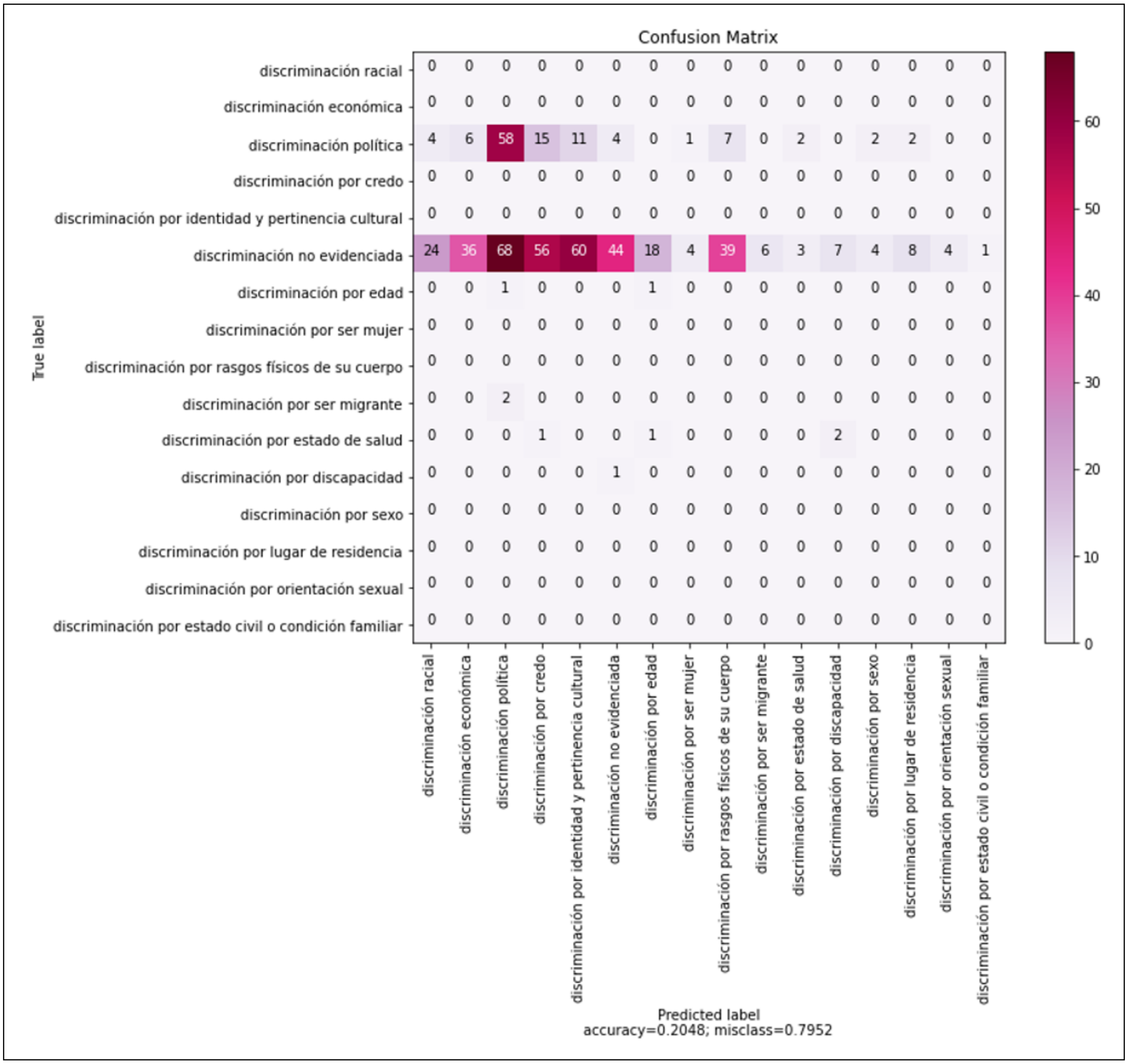
For the case of the indicator SDG 16.7.2, a total of 405,693 users were identified and 219,372 (54% of the users) were included once we applied the outlier's analysis.

For the second discrimination exercise as well as for the two representativeness exercises, the following metrics were calculated: accuracy, precision, recall, and F1-score. In the case of the precision, recall, and F1-score metrics, their respective equivalents were also calculated: micro, macro, and weighted (since the problem does not consist of binary classification). In addition, the confusion matrix is obtained to visualize the classification distribution in terms of true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

MODEL PERFORMANCE

The full evaluation set reported in the section "Scraping" was used for each of the proposed exercises. Predictions for the perception of discrimination and political representativeness were generated and compared against the annotations made by the experts. Figure 1 shows the confusion matrix obtained for the 16-label discrimination exercise.

As can be seen, the largest amount of information is originally classified as non-evidenced discrimination,



	discriminación racial	discriminación económica	discriminación política	discriminación por credo	discriminación por identidad y pertinencia cultural	discriminación no evidenciada	discriminación por edad	discriminación por ser mujer	discriminación por rasgos físicos de su cuerpo	discriminación por ser migrante	discriminación por estado de salud	discriminación por discapacidad	discriminación por sexo	discriminación por lugar de residencia	discriminación por orientación sexual	discriminación por estado civil o condición familiar
discriminación racial	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación económica	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación política	4	6	58	15	11	4	0	1	7	0	2	0	2	2	0	0
discriminación por credo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación por identidad y pertinencia cultural	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación no evidenciada	24	36	68	56	60	44	18	4	39	6	3	7	4	8	4	1
discriminación por edad	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
discriminación por ser mujer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación por rasgos físicos de su cuerpo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación por ser migrante	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación por estado de salud	0	0	0	1	0	0	1	0	0	0	0	2	0	0	0	0
discriminación por discapacidad	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
discriminación por sexo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación por lugar de residencia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación por orientación sexual	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discriminación por estado civil o condición familiar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1 Discrimination confusion matrix with 16 labels. True labels correspond to ground truth values. Predicted label (corresponds to predicted values).

where the model is confused by identifying this type of discrimination as political discrimination, discrimination based on identity and cultural relevance, and discrimination based on creed, mainly at 15.8%, 15.71% and 14.66% respectively. In addition, the classification results obtained for this exercise are summarized in Table 1.

From this, it is confirmed that the imbalance of existing classes confounds the base model, putting on evidence the need to deepen the exercise with a larger number of balanced samples, in addition to a re-training of the same, seeking to specialize the model in the domain of information under study.

For the external political efficacy case, Figure 2 shows the confusion matrix obtained for the two exercises. These exercises had as a differential the use of two (2) distinct sets of labels for the zero-shot model, corresponding to: "*esto es inclusividad política—esto es receptividad política*" and "*tengo algo que decir sobre el gobierno—los políticos escuchan lo que tengo que decir*." The first one associated with political inclusiveness and the second with political responsiveness.

The classification results obtained for these exercises are summarized in Table 2.

From this, it could be concluded that the second set of labels on the base model allows for a better classification of political representativeness. However, as in the case of discrimination, when considering the other metrics, it is necessary to specialize the model in this information domain with a considerable sample of examples, to improve the results obtained so far.

PRODUCTION OF INDICATORS

Users have been defined as all who have made a comment, categorized under any of the types of discrimination. There may be cases of users who made comments on more than one form of discrimination, so each of these facts should

METRIC	TYPE			
	–	MACRO	MICRO	WEIGHTED
Accuracy	0.2047	–	–	–
Precision	–	0.0873	0.2047	0.7822
Recall	–	0.6958	0.2047	0.2047
F1 Score	–	0.0485	0.2047	0.2625

Table 1 Metric results for discrimination performance.

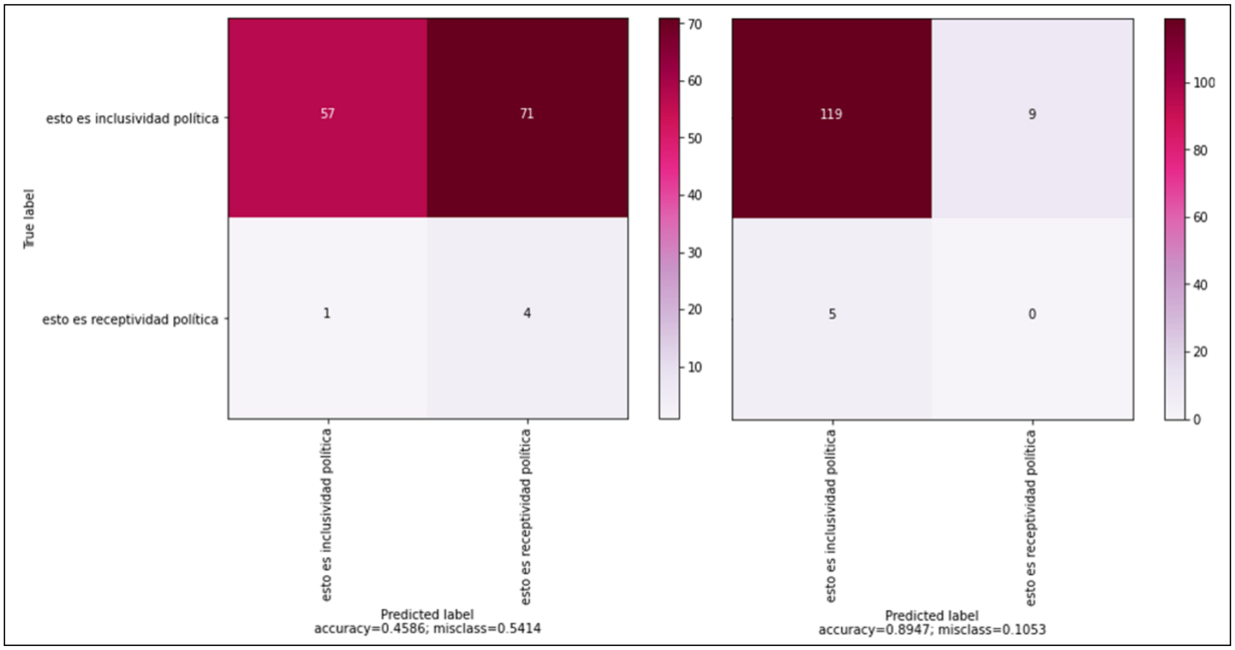


Figure 2 External political efficacy confusion matrix with the second set of labels. True label corresponds to ground truth values. Predicted label (corresponds to predicted values). Exercise A (*left*), Exercise B, (*right*).

be considered as a particular case, even if the author of the comment is the same. In this way, no associated information was lost, and the recommendation of the methodology is as follows: "The indicator should be a starting point for understanding patterns of discrimination" (United Nations 2018, p. 4). Table 3 shows the percentages of users whose comments were labelled by the model as discriminatory.

As it is presented in Figure 3, the highest proportion of users whose comments include discriminatory language correspond to political discrimination type (55.4%), which can be explained due to the political context during that period, which included demonstrations, mobility restrictions in the main cities of the country, and the highest peak of deaths and infection rate by Covid-19. The next types of discrimination in importance were cultural identity (11.50%) and religion (7.9%). The categories with the lowest participation were ethnicity (0.1%) and sex (0.2%).

To get a proxy value of these results comparable with the results of DANE's ECP, the proportion of users whose comments were associated with some of the recognized types of discrimination were calculated; this proxy indicator arises from the quotient between users whose comments were associated with some of the types of discrimination over the total number of users. (See Equation 1). In this

EXERCISE	METRIC	TYPE			
		–	MACRO	MICRO	WEIGHTED
A	Accuracy	0.4586	–	–	–
B		**0.8947**	–	–	–
A	Precision	–	**0.518**	0.4586	**0.9478**
B		–	0.4798	**0.8947**	0.9235
A	Recall	–	**0.6226**	0.4586	0.4586
B		–	0.4648	**0.8947**	**0.8947**

Table 2 Metric results for political external efficacy performance for both exercises.

TYPE OF DISCRIMINATION	ABSOLUTE VALUES	PERCENTAGE
Religion	650	7.95%
Disability	111	1.36%
Economic	500	6.11%
Age	306	3.74%
Civil status	15	0.18%
Cultural identity	940	11.50%
Migrant condition	30	0.37%
Women	14	0.17%
No_identified	432	5.28%
Political opinion	4.533	55.44%
Physical aspects	477	5.83%
Ethnicity	12	0.15%
Place of residence	50	0.61%
Health condition	87	1.06%
Sex	20	0.24%
Total users	8.177	100.0%

Table 3 Discrimination types (in percentage) disaggregated by users.

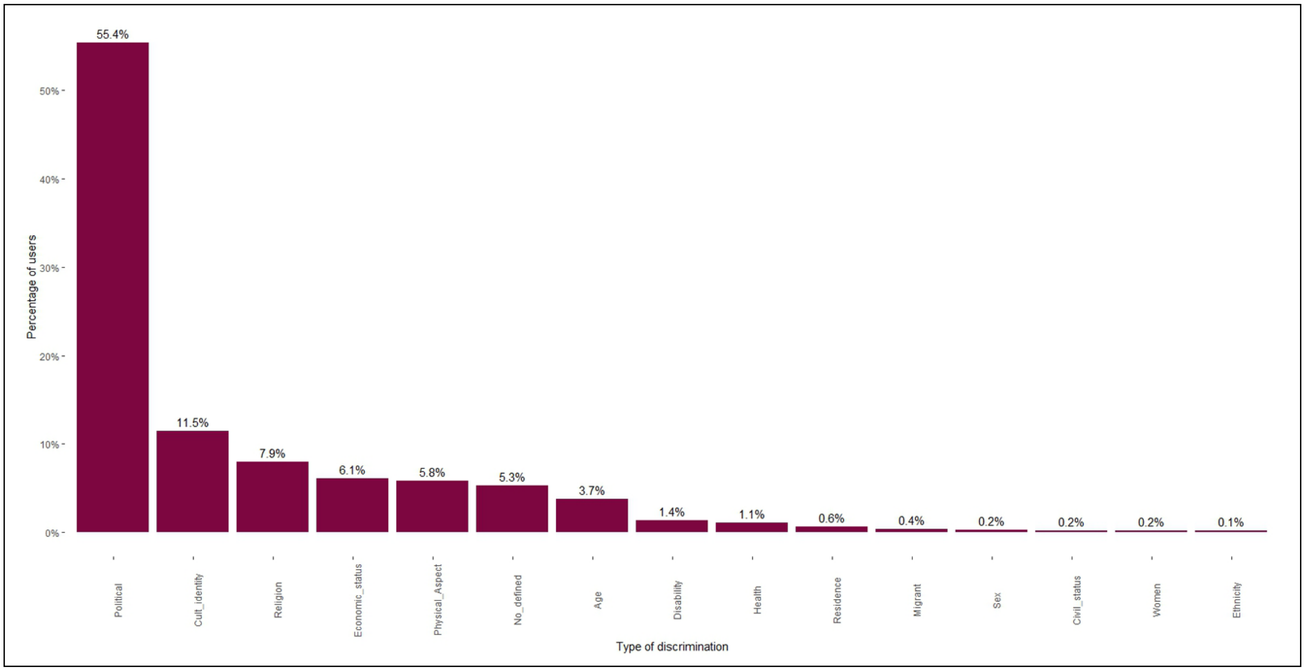


Figure 3 Discrimination comments by user and type.

expression, $Users_{prob0.5}$ corresponds to the total number of users with discrimination-related comments whose probability was higher than 0.5, $Users_{total}$ represents the total users with discrimination-related comments, and $Users_{discri}$ corresponds to the proportion of users whose comments were associated with some of the recognized types of discrimination.

$$Users_{discri} = \frac{Users_{prob0.5}}{Users_{total}} * 100. \qquad (1)$$

In this case, the value is 1.9%, which means that 1.9% of the users have made some comment with a high probability of having discriminatory content.

This excludes users in the non-evidenced category since they are comments that the model cannot assign to any of the forms of discrimination, although it cannot be affirmed that there is no discriminatory content in them. The comparative results are presented in Figures 4 and 5.

The comparison with the proportion of users with comments related to discrimination confirms that there are differences in the prevalence of the types of discrimination in the two sources observed, thus for users of social networks such as Facebook, the type of discrimination that has more comments was politics, whereas types of discrimination associated with age and economic discrimination were observed in the ECP.

For SDG indicator 16.7.2, as shown in Figure 6, inclusive decision-making has a significant prevalence. 79.5% of the users made comments that the model has been associated with the latter. The difference between men and women is short: 44.7% of comments made by men were labelled as inclusive, compared with 34.8% of comments made by women labelled as inclusive. A similar proportion is observed in responsive decision-making.

As it is presented in Figure 7, the official statistics from Colombia show a similar tendency: The ECP show a higher percentage of inclusive decision-making for the people who respond to the ECP, and show a very short difference between both sexes (21.0% of men consider their decision-making process inclusive, whereas 20.1% of women consider theirs inclusive). The same proportions are presented in responsive decision-making.

CONCLUSIONS AND FUTURE WORK

For the discrimination case, low classification accuracy is observed. When at first analyzing the confusion matrix, a low variability is found for the actual labels. In the second instance, the model identifies types of discrimination as political, cultural identity and belonging, creed, physical features, economic, racial, and age, as well as non-evidenced discrimination. From these results, it could be inferred that the model has a certain bias for political and creedal discrimination.

Regarding indicator 16.7.2, although the results for Exercise A show a low performance, better results are obtained for Exercise B, which presents similar results with the ECP. This suggests that for external political efficacy with a zero-shot base model, the best way in which the reference labels are presented to the model corresponds to

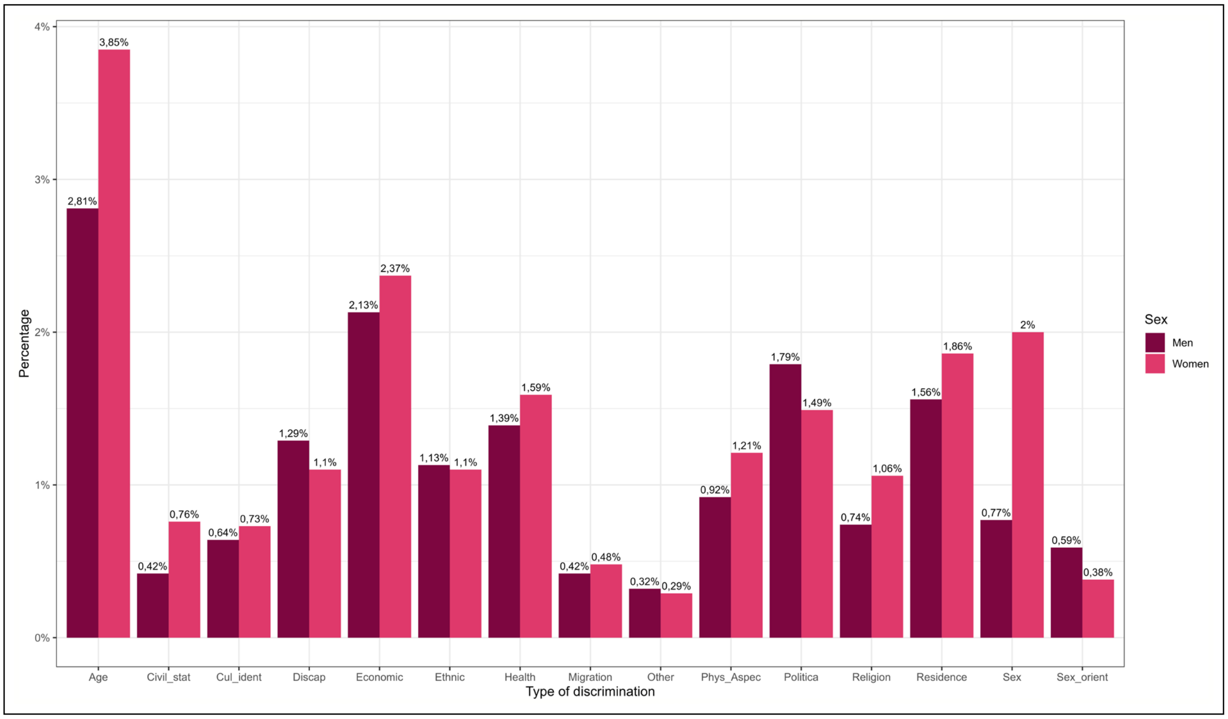


Figure 4 Proportion of people who felt discriminated against, by sex. Colombia's Political Culture Survey.

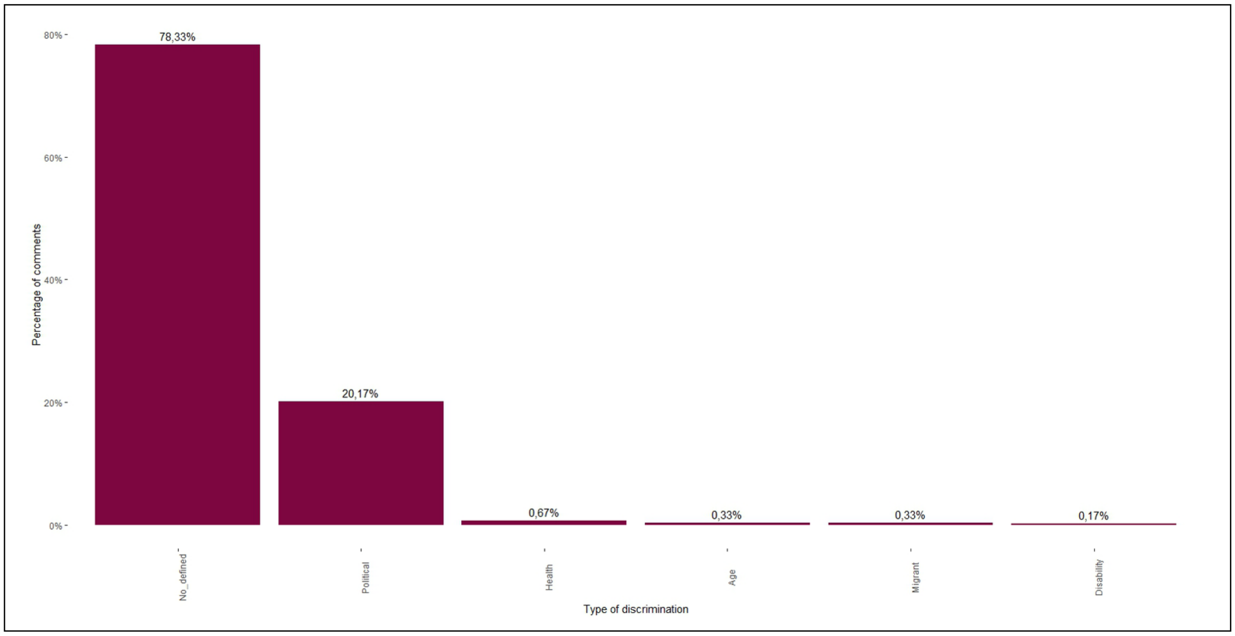


Figure 5 Supervised analysis. Proportion of types of discrimination.

Exercise B. Considering both the discrimination case and the external political efficacy case, it is necessary to perform a fine-tuning process to obtain better domain adaptability.

Based on the proposed method results, it is possible to estimate a proxy indicator of SDG 16.b.1 because of the closeness between the obtained value for the $\text{Users}_{\text{discri}}$ and the affirmative response percentage by the different types of discrimination, as analyzed in ECP. However, a key difference in the most prevalent types of discrimination between the ECP and the exercises was found. The main difference between the ECP and the exercises was in the type of discrimination by age, although the types of discrimination related to the economic situation and political opinions are amongst the more prevalent types in both measurements. Therefore, the estimation should be made with caution, noticing that bias and the differences identified in the types of discrimination could affect the results.

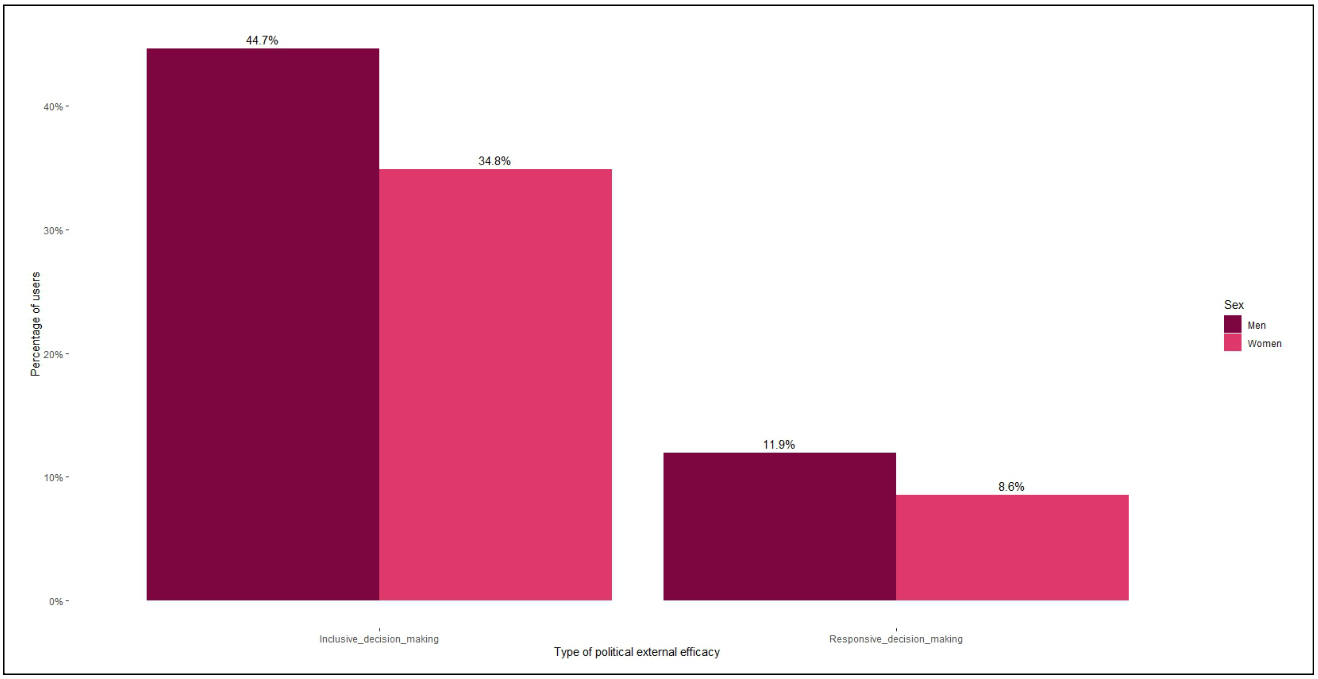


Figure 6 Proportion of users by sex and types of political external efficacy.

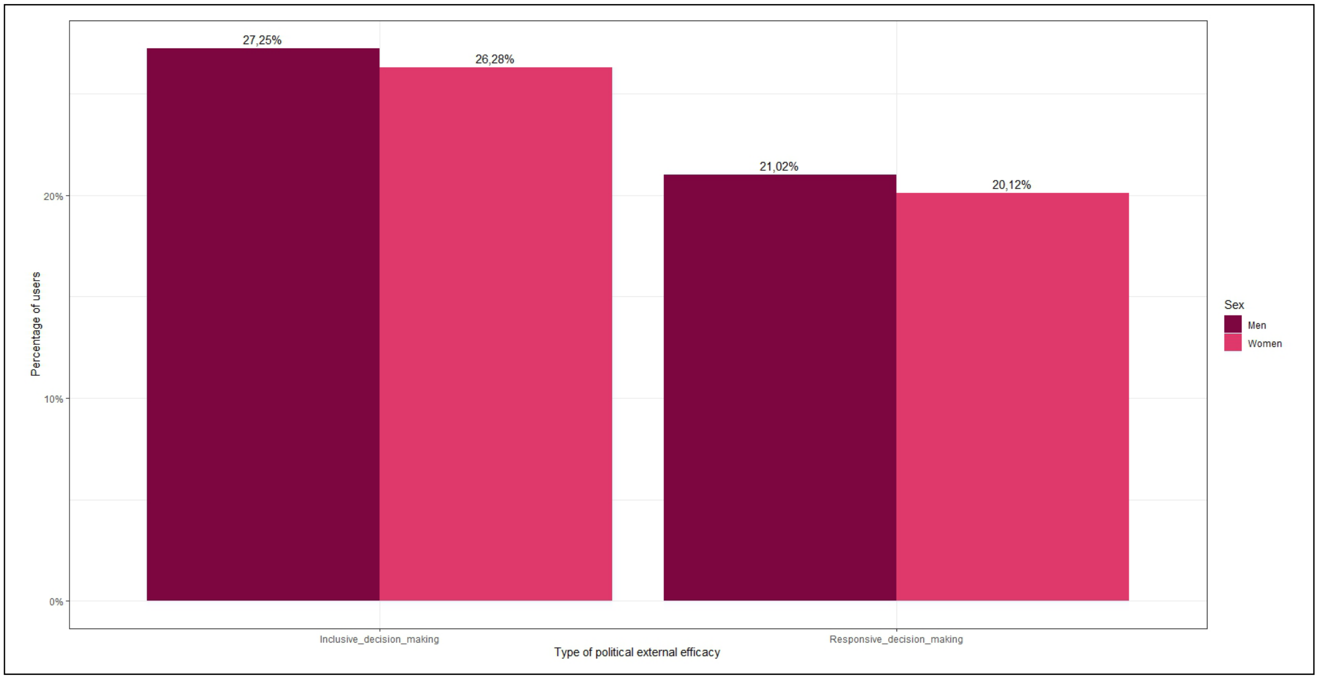


Figure 7 Colombia's Political Culture Survey. Political external efficacy by sex.

From the above, the conceptual differences in the capture of the phenomena between the official information and the results of this research play a prominent role in further researches on the subject.

In the ECP, the types of discrimination of which people have been victims are based on guidelines posed on the United Nations indicator metadata, whereas in this analysis, comments related to discrimination are only generally identified based on the semantic features found by the applied language models, and more research based on these models is required to identify victims of discrimination. Therefore, ECP and a language processing model approach should be understood as two different points of view to describe different aspects for the discrimination phenomenon. This conclusion suggests that more studies are needed to ensure data representativity, particularly concerning the approximation to identify victims of discrimination and the comparison of the traditional and nontraditional data for producing statistical information.

Hence, these results should be considered as contextual or complementary information, presenting the specific dynamic of social media in which people reveal their situation related to discrimination. If so, it is important to establish that the indicators generated as part of this research could not be considered as an official estimation for discrimination or for the SDGs indicator. Similar to the 16.b.1 SDG indicator, the 16.7.2 SDG indicator results obtained with the proposed method are comparable, and provide context to the ECP.

These findings are consistent with the literature on citizen science challenges, as shown in Pateman and West (2017): "Citizen science could make contributions in three types of process linked to the SDGs: defining national and subnational targets and metrics, monitoring progress and implementing action."

In addition, a few methodological challenges were identified. First, the design of a robust methodology for stable data ingestion, which includes use of bots or crawlers in Facebook. Second, the development of dummy accounts is also an important factor to increase the stability of Facebook data capture. This considers that the longer an account is active and shows favorable behavior, the less likely it is to be closed. Finally, increasing the number of profiles, posts, and comments extracted to add demographic indicators such as age and gender that guarantee data representativity and improve the quality of the metrics.

These challenges posit a key question regarding citizen science and the use of citizen-generated data: How can we assure the quality of data? Based on these results, it is not enough just to count on a well-designed and implemented methodology (Pateman and West 2017, p. 3) to resolve these issues. A comprehensive analysis of data should be included the methodology, there should be an audit process, and the follow-up should be based on statistical standards (e.g., the Generic Statistical Business Production Model). As stated in Fritz et al. (2019), "The quality of data from citizen science can be evaluated using the same measures as any other official data… This includes measures such as positional and thematic accuracy, temporal currency of the data, completeness and representativeness over space and time, and whether the data are fit-for-purpose." This must be accompanied by capacity-building in the institutions and a fluent dialogue with the Civil Society Organizations that work on the subject.

Therefore, it can be concluded that Facebook data is not a feasible official data source given the various challenges presented and the estimated numbers for both indicators. Given that no representativity can be assured, no further comparison with the current data can be made. This reduces the scope of this data source to be a contextual data source but not an official one in NSOs.

Aspects associated with citizen participation, as it is understood in the main definitions of citizen science (Haklay et al. 2021, pp. 15–18), must also be considered. In the case of the labelers in the supervised analysis, they received training in technological devices as well as in the definition of the project's main concepts, for at least two different periods. This improved the responsiveness indicator results, but the improvement does not occur with discrimination indicator 16.b.1, although the labelers received the same training for both indicators. This kind of lesson argues in favor of being more open to including different types of collaboration involving citizen participation, but a strictness when a methodological approach is required. According to Pateman and West (2017, p. 3) "when a study is well designed and implemented, the quality of citizen-collected data is, in fact, comparable to that collected by professional scientists."

Our research shows that the data collection process from social networks also raises ethical concerns in two aspects: the use of citizen-generated data from social media as a relevant data source in scientific research, and the use of "black box" models and the bias they have (Franzen et al. 2021, p. 190). The issues connected to opacity and bias in machine learning models have brought to light the need for more transparency in the designing of algorithms and the data used for training to prevent or mitigate adverse effects. According to Franzen et al. (2021), black box is a system "in which we can observe the inputs and outputs but not the internal process. Machine learning algorithms like neural networks and deep learning are so intrinsically complex that it is virtually unworkable to get to the bottom of their operations and internal decision-making processes." One of the reasons for using models such as zero–shot and BERT models in this project was these are well known, and their technical details are widely worked by different researchers. In addition, its use is transparent in the sense Franzen is referring to: "The idea behind explainable AI radiates from the implementation of algorithms that are understandable to a human expert who can discern the internal mechanisms and understand what is happening" (Franzen et al. 2021, p. 191). In this very way, the scripts, notebooks, and a manual were written and disseminated to explain how this exercise was made.

However, some semantic and linguistic considerations on the BERT model could not be studied in terms of the accuracy of these semantic relations between the comments and the types of discrimination, and how we can mitigate the bias behind it. This is an open question, and more studies are required. In this project, supervised analysis, in which labelers reviewed comments, was the main strategy employed to reduce the possible bias of these models. Consideration of this subject remains an open question, too, and it could be a promising research line because of its impact on the production of official statistical information.

However, using data from scraping raises serious concerns about the data privacy of social media users, hence the discussion of this topic in the development of the project. It was also considered that the use of social media data and administrative records present similarities. For instance, these data sources are created with a specific purpose and not necessarily for statistical production. Decree 2404 of 2019 states, for the case of administrative records: "The data protection and information security conditions of the microdata custodian shall be prioritized. The parties involved in the exchange shall guarantee that the information shall not be used for purposes other than statistical and shall maintain confidentiality" (DANE 2019, p. 13). The same criteria could be applied for social media data since both share various features such as unstructured formats, automatized data collection, high velocity and volume, and, in some cases variability. It also presents differences: Users give their data because it is mandatory (like taxes or health registers), but social media data are shared voluntarily in that network. In both cases, data privacy and confidentiality are required to guarantee that the statistical information for the public is trustworthy.

Given this confluence, DANE considered social media as a potential alternative data source of statistical information, and fosters its usage in a mandatory fashion (DANE 2019, p. 14) or in a voluntary one, as suggested in the National Code of Good Practices (DANE 2022, p. 23).

Based on the latter, the project also establishes the use of social media for another goal—to create technical capacities to use deep learning models for improving statistical processes. In that sense, data scraping for social media should be understood, too, as a technical device to collect data, as it is stated in some paradigmatic legal cases. One example is *Sandvig v. Sessions* in the United States District Court for the District of Columbia (Mancosu and Vegetti 2020, p. 6), in which scraping was considered in that sense. According to Mancosu and Vegetti (2020, p. 6), "[s]craping is merely a technological advance that makes information collection easier; it is not meaningfully different from using a tape recorder instead of taking written notes, or using the panorama function on a smartphone instead of taking a series of photos from different positions." This consideration was also addressed in the project, based on Facebook Terms of Service and in the current legal Colombian framework.

It is worth noting that new projects related to citizen-generated data are being developed in DANE, where citizens do have an active role in the data collection; hence, a comparison could be made to evaluate the best approach to work with citizen-generated data and with citizen science in general. A more active, participatory approach could be more fruitful, based on the experiences of other countries (Haklay et al. 2021).

Therefore, the deep learning approach using transformer models like the zero-shot model, represents a starting point to study different SDG indicators associated with perception or information retrieval from both citizen-generated data and citizen science perspectives.

Further research could be conducted in three ways. First, produce a model retraining both for discrimination and for representative domain specialization to address and possibly enhance the obtained results based on the carried analysis. This entails the adequacy of the estimation formula and the proposed method in this research. Second, launch a new approach, broadening the data source, as alternative sources such as Twitter might be more feasible to tap into. Finally, produce a model complexity analysis in order to evaluate model alternatives for discrimination classification.

In terms of public policy, the feasibility of these kinds of alternative sources should be explored in other SDG indicators. Going forward, researchers should try to assess the data quality and strength of the methodological design by taking into consideration the role of citizens in conceiving statistical production for the 2030 Agenda. To this end, they should develop guidelines for using social media data and citizen-generated data for statistical production in general, fostering the participation of civil society. This is the necessary next step to broaden the scope of the citizen science in Colombia.

ACKNOWLEDGEMENTS

We gratefully recognize the sponsorship of Data for Now Initiative (with partners, the UN Statistics Division's Department of Economic and Social Affairs, the Global Partnership for Sustainable Development Data, the World Bank, and the Sustainable Development Solutions Network) and DANE for supporting the research project "*Proxy measurement of indicator 10.3.1 / 16.b.1 and 16.7.2, using techniques based on data mining and advanced machine learning techniques through social networks*" in the framework of DANE's experimental statistics. We also would like to thank the Presidential Council for Human Rights and International Affairs, and the Ministry of Technologies and Communications. We would especially like to recognize the DANE specialists who were responsible for the labeling tasks. In addition, we want to extend our thanks to Andrés Jaque and Andrés Rosso who gave us their support as consultants, as well as Adriana Leguizamón and Diego Bravo for their participation in carrying out exploratory exercises for the project development. Finally, we would like to encourage and highlight the project researchers: Grace Andrea Torres, Victor Andrés Arevalo, Vahan Martirosyan, and Andrés D. Pérez for their significant contribution.

FUNDING INFORMATION

The publication of this paper was funded by the Departamento Administrativo Nacional de Estadística (DANE) and the journal *Citizen Science Theory and Practice*. The authors would like to express their gratitude to both institutions for providing the resources that made this paper possible.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

This paper was conceived after preliminary results showed the importance and impact of this idea. The initial proposal to focus on citizen science was made by Ms. Karen Chavez. The writing tasks were performed by Mr. Víctor Andrés Arévalo, Mr. Andrés D. Pérez, Mr. Vahan Martirosyan, and Ms. Grace Torres. The data collection and processing were done by Mr. Pérez and Mr. Martirosyan, as well as the method section. Quality metrics were calculated by Mr Pérez. The data analysis was done by Mr. Arévalo, as well as the data visualization, with major revisions by Ms. Chavez and Ms. Julieth Solano. The national context of use of alternative sources was written by Ms. Torres, and the discussion of citizen science was written by Mr. Arévalo. Final editing, grammar and style corrections were made by Mr. Arévalo.

AUTHOR AFFILIATIONS

Victor Arevalo Cabra orcid.org/0000-0001-9068-1265
Departamento Administrativo Nacional de Estadística (DANE), CO
Karen Chávez Quintero orcid.org/0000-0001-5502-9319
Departamento Administrativo Nacional de Estadística (DANE), CO
Andrés D. Pérez orcid.org/0000-0001-5410-4938
Departamento Administrativo Nacional de Estadística (DANE), CO
Grace Torres Pineda orcid.org/0000-0002-1693-4866
Departamento Administrativo Nacional de Estadística (DANE), CO
Julieth Solano Villa orcid.org/0000-0002-8929-8565
Departamento Administrativo Nacional de Estadística (DANE), CO
Vahan Martirosyan orcid.org/0009-0006-7213-8588
United Nations Development Program, AM

REFERENCES

Amichai-Hamburger, Y and **Furnham, A.** 2007. The Positive Net. *Computers in Human Behavior*, 23: 1033–1045. DOI: https://doi.org/10.1016/j.chb.2005.08.008

Awan, I. 2016. Islamophobia on Social Media: A Qualitative Analysis of the Facebook's Walls of Hate. *International Journal of Cyber Criminology*, 10(1). January–June 2016. DOI: https://doi.org/10.5281/zenodo.58517

Ben-David, A and **Matamoros Fernández, A.** 2016. Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. *International Journal Of Communication,* 10: 27. Available at: https://ijoc.org/index.php/ijoc/article/view/3697/1585. (Last accessed 10 February 2023).

Bonney, R, Cooper, CB, Dickinson, J, Kelling, S, Phillips, T, Rosenberg, KV and **Shirk, J.** 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, 59(11): 977–984. Available at: http://www.bioone.org/doi/full/10.1525/bio.2009.59.11.9. (Last accessed 09 June 2023).

Brooks, CF, Shmargad, Y and **Williams, BA.** 2018. How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy*, 8: 78–115. Available at: https://www.jstor.org/stable/10.5325/jinfopoli.8.2018.0078. (Last accessed 09 June 2023).

Chapman, P, Clinton, J, Kerber, R, Khabaza, T, Reinartz, TP, Shearer, C and **Wirth, R.** 2000. CRISP-DM 1.0: Step-by-step data mining guide. Available at: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf. (Last accessed 21 May 2023).

DANE (Departamento Administrativo Nacional de Estadística de Colombia). 2019. Decree 2404 of 2019. Available at: https://www.dane.gov.co/files/acerca/Normatividad/decretos/DECRETO-2404-DE-2019.pdf. (Last accessed 05 May 2023).

DANE (Departamento Administrativo Nacional de Estadística de Colombia). 2021. Political Culture Survey website. Available at: https://www.dane.gov.co/index.php/estadisticas-por-tema/gobierno/cultura-politica. (Last accessed 05 August 2022).

DANE (Departamento Administrativo Nacional de Estadística de Colombia). 2022. National Code of Good Practice. Available at: https://www.dane.gov.co/files/sen/bp/Codigo_nal_buenas_practicas-2022.pdf. (Last accessed 05 May 2023).

Datareportal. 2022. *Digital 2021: Colombia.* 11 February 2021. Available at: https://datareportal.com/reports/digital-2021-colombia. (Last accessed 10 February 2023).

Departamento Nacional de Planeación de Colombia (DNP). 2018. CONPES 3918 of 2018. Strategy for the Implementation of the Sustainable Development Goals

(SDGs) in Colombia. Available at: https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/3918.pdf. (Last accessed 05 May 2023).

Franzen, M, Kloetzer, L, Ponti, M, Trojan, J and **Vicens, J.** 2021. Machine Learning in Citizen Science: Promises and Implications. In: Vohland, K, et al. (eds.), *The Science of Citizen Science*. Springer. DOI: https://doi.org/10.1007/978-3-030-58278-4_2

Fritz, S, See, L, Carlson, T, Haklay, M, Oliver, J, Dilek, F, Mondardini, R, Brocklehurst, M, Shanley, L, Schade, S, When, U, Abrate, T, Anstee, J, Arnold, S, Billot, M, Campbell, J, Espey, J, Gold, M, Hager, G, He, S, Hepburn, L, Hsu, A, Long, D, Masó, J, McCallum, I, Muniafu, M, Moorthy, I, Obersteiner, M, Parker, A, Weisspflug, M and **West, S.** 2019. Citizen science and the United Nations Sustainable Development Goals. *Nature Sustainability*, 2(10): 922–930. DOI: https://doi.org/10.1038/s41893-019-0390-3

Haklay, M, Dörler, D, Heigl, F, Manzoni, M, Hecker, S and **Vohland, K.** 2021. What Is Citizen Science? The Challenges of Definition. In: Vohland, K, et al. (eds.), *The Science of Citizen Science*. Springer. DOI: https://doi.org/10.1007/978-3-030-58278-4_2

Heigl, F, Kieslingerb, B, Paulc, K, Uhlikd, J and **Dörlera, D.** 2019. Toward an international definition of citizen science. *PNAS Journal*, 116(17). DOI: https://doi.org/10.1073/pnas.1903393116

IBM. 2021. IBM SPSS Modeler CRISP-DM Guide. 2021. Available at: https://www.ibm.com/docs/it/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf. (Last accessed 21 May 2023).

Lucero, L. 2017. Safe spaces in online places: social media and LGBTQ youth. *Multicultural Education Review*, 9(2): 117–128. DOI: https://doi.org/10.1080/2005615X.2017.1313482

Lee, D. 2019. Muslim Women on the Internet: Social Media as Sites of Identity Formation. *Journal of South Asian and Middle Eastern Studies*, 42(3): 20–34. Spring 2019. DOI: https://doi.org/10.1353/jsa.2019.0018

Mancini, T and **Imperato, C.** 2020. Can Social Networks Make Us More Sensitive to Social Discrimination? E-Contact, Identity Processes and Perception of Online Sexual Discrimination in a Sample of Facebook Users. *Social Science Journal,* 9(4): 47. DOI: https://doi.org/10.3390/socsci9040047

Mancosu, M and **Vegetti, F.** 2020. What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media + Society*, 1–11. DOI: https://doi.org/10.1177/2056305120940703

Marciano, A and **Antebi-Gruszka, N.** 2020. Offline and online discrimination and mental distress among lesbian, gay, and bisexual individuals: The moderating effect of LGBTQ Facebook use. *Media Psychology*. DOI: https://doi.org/10.1080/15213269.2020.1850295

Miller, GH, Marquez-Velarde, G, Williams, AA and **Keith, VM.** 2020. Discrimination and Black Media Use: Sites of Oppresion and Expression. *Sociology of Race and Ethnicity*, 7(2): 1–17. DOI: https://doi.org/10.1177/2332649220948179

NETWORK, T. 2022. Producing and supporting citizen-generated data. Available at: https://secureservercdn.net/198.71.233.45/bj7.5fd.myftpupload.com/wp-content/uploads/2021/07/Producing-and-supporting-citizen-generated-data.pdf. (Last accessed 07 September 2022).

OHCHR. 2020. New global data on human rights showcased in Sustainable Development Goals Report, 14 July 2020. Available at: https://www.ohchr.org/en/stories/2020/07/new-global-data-human-rights-showcased-sustainable-development-goals-report. (Last accessed 06 June 2023).

Pateman, R and **West, S.** 2017. How could citizen science support the Sustainable Development Goals? *Stockholm Environment Institute*. http://www.jstor.com/stable/resrep17213. (Last accessed 10 February 2023).

Pushp, PK and **Srivastava, MM.** 2017. Train once, test anywhere: Zero-shot learning for text classification. *ArXiv*. DOI: https://doi.org/10.48550/arXiv.1712.05972

Sajlan, D. 2022. Hate Speech against Dalits on Social Media: Would a Penny Sparrow be Prosecuted in India for Online Hate Speech? *CASTE: A Global Journal on Social Exclusion,* 2(1): 77–96. DOI: https://doi.org/10.26812/caste.v2i1.260

United Nations. 2018. SDGs 16.b.1. Indicator Metadata. Available at: https://unstats.un.org/sdgs/metadata/files/Metadata-16-0b-01.pdf. (Last accessed 07 September 2022).

United Nations. 2022. IAEG-SDGs: Inter-agency and Expert Group on SDG Indicators. Available at: https://unstats.un.org/sdgs/iaeg-sdgs/. (Last accessed 07 September 2022).

United Nations. 2023. SDGs 16.7.2 Indicator Metadata. Available at: https://unstats.un.org/sdgs/metadata/files/Metadata-16-07-02.pdf. (Last accessed 05 May 2023).

TO CITE THIS ARTICLE:

Arevalo Cabra, V, Quintero, KC, Pérez, AD, Pineda, GT, Villa, JS and Martirosyan, V. 2023. Civil Society Data for Sustainable Development Goal 16 Monitoring: A Case Study of the Use of Social Networks for Measuring Perception of Discrimination. *Citizen Science: Theory and Practice,* 8(1): 38, pp. 1–16. DOI: https://doi.org/10.5334/cstp.590

Submitted: 02 October 2022 **Accepted:** 10 May 2023 **Published:** 27 June 2023

Citizen Science: Theory and Practice is a peer-reviewed open access journal published by Ubiquity Press.