United Nations Economic Commission for Europe

Conference of European Statisticians (CES)

7th Expert Meeting on Statistics for SDGs

12 - 13 April 2023, Geneva

Working paper WP.S5_1

Session 5: Expanding the SDGs monitoring with non-traditional data sources

Use of Social Networks for Measuring Perception of Discrimination

Prepared by Andrés ARÉVALO, Karen CHAVEZ, Andrés PÉREZ, Julieth SOLANO, Grace TORRES
Departamento Administrativo Nacional de Colombia (DANE), Colombia
Vahan MARTIROSYAN
United Nations Development Programme, Armenia

Introduction

1. DANE has been working with custodian agencies in strategies to fill data gaps. 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* and, SDG *16.7.2 Proportion of population who believe decision-making is inclusive and responsive, by sex, age, disability, and population group* (DNP, 2018).

2. DANE has deployed a strategy to measure some aspects of these indicators, using traditional sources like the Victimization Survey (2000), the Coexistence and Citizen Security Survey (ECSC, by its acronym in Spanish), carried out every two years, and the Political Culture Survey (ECP, by its acronym in Spanish), the latter 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).

3. Given this particular context, citizen-generated data, like social networks, and citizen science 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 the Internet penetration in the country. According to the National Quality of Life Survey - ECV of 2021 internet usage 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 Colombia (around 78% of the total population of the country), and the percentage of people from 14 to 64 using Facebook as its main social platform was 91,4%, and the potential Facebook audience was in this lapse of 36 million people (Datareportal, 2021). However, citizen-generated data as social media presents some of the problems that citizen science also presents: if citizen science involves the processes where citizens become data providers and users, then citizen-generated data requires scientific standards, ethical considerations, and data management, among others (Heigl et al, 2019, p. 3). Likewise, the result's poor quality and statistical relevance are among the main concerns in these two research fields for scientists (Pateman and West, 2017, p. 3).

4. Despite social networks like Facebook to be one of the inputs to understand marginalized communities (NETWORK, 2022) a few research has been done to address the problem of discrimination as a natural language processing problem in social media.

5. 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 & Antebi-Griszcaa (2020), where 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, contrary to the results of Lucero 2020, where LGBTQ population feels this social network as a safe place to interact with some other members of this community. Mancini and Imperato (2020) also used Facebook as their data source, studying the behaviour of different online groups in that network in order to understand the process by which online intergroup contact makes individuals more sensitive to detect discrimination (2020, p. 8). Brooks, Shmargad and Williams (2018) researched the discrimination that comes from the very algorithms and how the bias, the lack of data and audits inform a clear picture of how data systems and algorithms could, in fact, make discriminatory decisions against people.

6. 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).

7. Therefore, the question that motivates this study is whether Facebook is a useful and feasible source to generate official statistics broadly speaking and on discrimination, in this research. To address this question, a deep learning methodology is proposed to obtain complementary measurements for both SDG indicators 16.b.1 and 16.7.2, which take full advantage of Facebook information and can be used to contrast and complement information for Colombia's Political Culture Survey.

2. Method

8. 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. Most data quality assessment methodologies also consider the reputation or believability of the source of data in question.

9. The Language Modelling based on Transformers models concentrates on the quality of language classification models employed to extract information from the text in Facebook posts and comments. The pre-trained large language models were used for text classification, in a technique called zero-shot text classification. To obtain this data for statistical uses is a significant data quality bottleneck, given the lack of large, labelled 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 modelling that can at once be used to generate large, labelled datasets more affordably and quickly. Preliminary results show the challenges to reach the accurate representativity for the collected data and the possible solutions on it.

10. 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.

11. Since a pre-trained Zero-Shot 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 SGD indicators under study: perception of discrimination and representativeness, respectively. In addition, each one of these exercises has two sub exercises in which different target labels were proposed.

12. Taking the 771.502 records available, a total of 503.553 users have been identified but only 8.177 (corresponds just to 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. This is due to the performance of the very model, with a low metrics associated to discrimination.

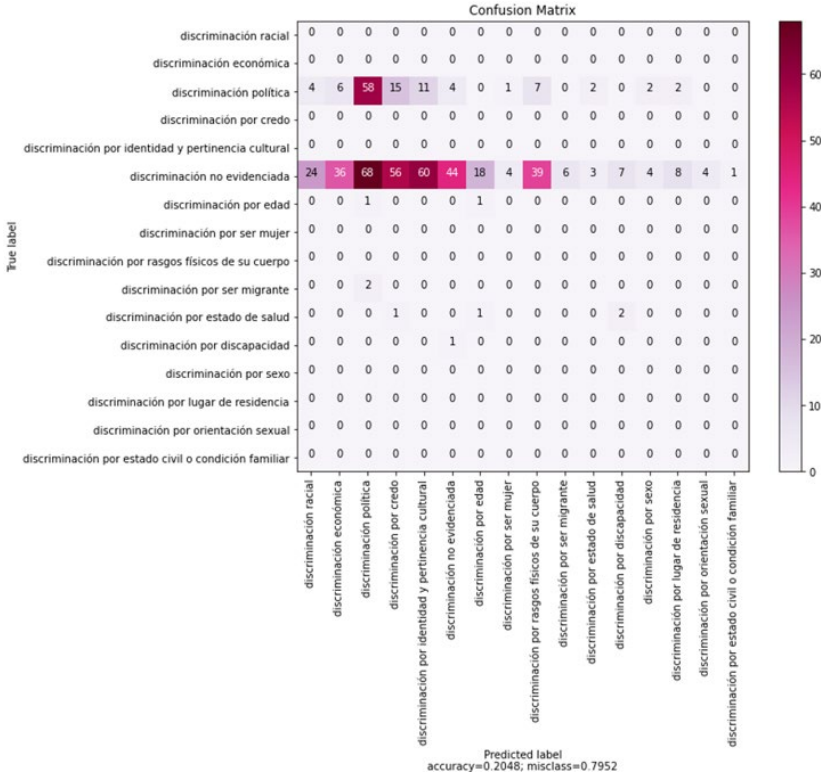
13. For the case of the indicator SDG 16.7.2, a total of 405.693 users were identified and 219.372 (54% of the users) were included once we applied the outlier's analysis.

3. Model performance

14. The 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.

Figure 1

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



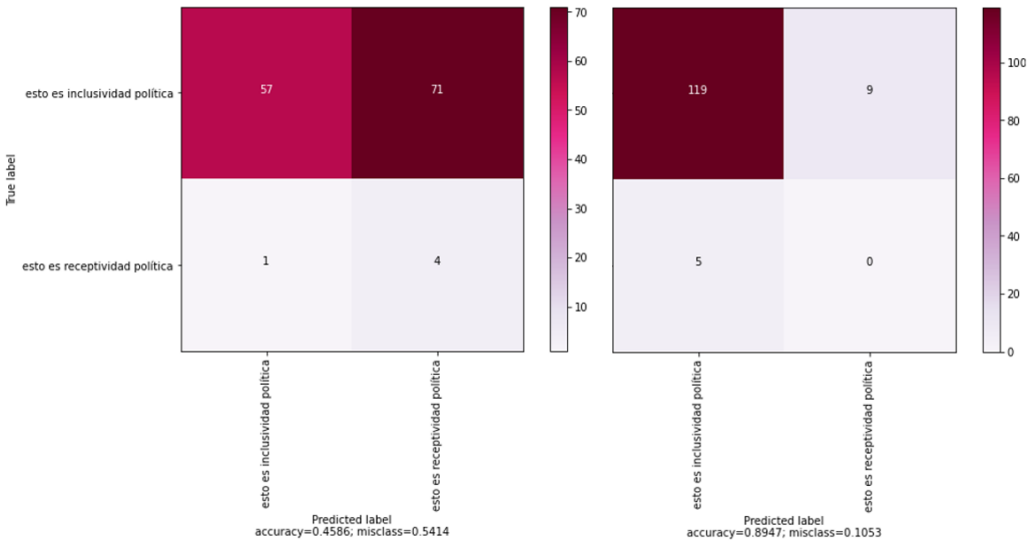
15. From this, it is confirmed that the imbalance of existing classes confounds the base model, putting on evidence the need to deepen the exercise with a larger number of balanced samples, in addition to a re-training of the same seeking to specialize the model in the domain of information under study.

Table 1
Metric results for discrimination performance

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*

16. On the other hand, for the representativeness case, Figure 2 shows the confusion matrix obtained for the two representativeness exercises. These exercises had as 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” for the first and second representativeness exercise respectively.

Figure 2
Representativeness confusion matrix with the second set of labels. True label corresponds to ground truth values. Predicted label (corresponds to predicted values). Left) A - exercise. Right) B - exercise.



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

Table 2
Metric results for representativeness performance for both exercises

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***

4. Indicators production

17. 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 be considered as a single 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 followed: "The indicator should be a starting point for understanding patterns of discrimination" (UN, 2021:4). As shown in Table 3 shows the percentages of users whose comments were labelled by the model as discriminatory.

Table 3
Discrimination types (in percentage) disaggregated by users

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%
Not 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%

18. For SDG indicator 16.7.2 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: 44,6% 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.

5. Conclusions and future work

19. For the discrimination case, low classification accuracy is observed. At first instance, when analyzing the confusion matrix, a low variability is found for the actual labels. In the second instance, the model identifies non-evidenced discrimination as political, cultural identity and belonging, creed, non-evidenced, physical features, economic, racial and age. Considering both the discrimination case and the representativeness case, a fine-tuning process is necessary to obtain better domain adaptability.

20. Based on the proposed method results, it is possible to estimate a proxy indicator of SDG 16.b.1 due to the closeness between the obtained value for official information and the alternative one. However, a difference in the types of discrimination more prevalent between the ECP and the exercise is found. The main difference between the ECP and the exercise 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.

21. From the above, it could be concluded that the conceptual differences in the capture of the phenomena between the official information and the results of this research (in the ECP the cases of discrimination of which people have been victims are identified, while in this analysis comments related to discrimination are generally identified considering having been a victim or having discriminatory comments), suggests that more studies are needed to ensure the data representativity, particularly in relation with the approximation to identify victims of discrimination. 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”.

22. 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. In this order of ideas, it is important to establish that the indicators generated as part of this research could be not considered as an official estimation for discrimination or the SDG indicator. Similarly, 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.

23. In addition, a few methodological challenges were identified as follows: the design of a robust methodology for stable data ingestion, for which strategies such as simulating Facebook users; 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 favourable behaviour, the less likely it is to be closed; finally, increasing the number of profiles, posts and comments extracted, which improves the quality of the metrics added for demographic indicators such as age and gender.