The Importance of Poverty in Sustainability Policies: An Approach to Understanding Online Opinion

Miguel A. Del-Pino
delpinoinjoque@gmail.com / Concordia University

Arezo Bodaghi
arezo.bodaghi@mail.concordia.ca / Concordia University

Pierre Watine
watinepierre@gmail.com / Concordia University

Ketra Schmitt
ketra.schmitt@concordia.ca / Concordia University

Recepción: 1/9/2020 Aceptación: 1/10/2020

Abstract. Twitter data related to poverty and basic income was collected for 24 days in 2019, and then was cleaned and prepared for natural language processing. A 7 % subset of the data was manually labeled for sentiment analysis in order to inform the artificial intelligence (AI), which was trained and verified on this subset. We present the results for both the 7 % verification sample and the entire database. This analysis of public opinion on poverty is situated within the Sustainable Development Goals and the support for poverty reduction policies.

Keywords: social media analytics / poverty / Sustainable Development Goals (SDGs) / sustainable development

La importancia de la pobreza en las políticas de sostenibilidad: un enfoque para comprender la *online opinion*

Resumen. Los datos de Twitter relacionados con la pobreza y los ingresos básicos se recopilaron durante 24 días en el 2019, se limpiaron y se prepararon para el procesamiento del lenguaje natural (*natural language processing*). Un subconjunto del 7 % de los datos se etiquetó manualmente para el análisis de sentimientos con el fin de informar a la inteligencia artificial (IA). La IA fue entrenada y verificada en este subconjunto. Presentamos los resultados tanto de la muestra del 7 % como de toda la base de datos. Este análisis de la opinión pública sobre la pobreza se sitúa dentro de los objetivos de desarrollo sostenible y el apoyo a las políticas de reducción de la pobreza.

Palabras clave: análisis de redes sociales / pobreza / objetivos de desarrollo sostenible (ODS) / desarrollo sostenible

1. INTRODUCTION

Poverty is one of the problems that affect our present as well as our future as humans. Addressing this problem is of utmost importance to governments around the world. Although the proportion of the world's population living below the poverty line has seen a drastic reduction, going from 25.5 % in 2002 to 12.8 % in 2012 (United Nations Statistics Division, 2019), the eradication of poverty remains an extremely complex challenge. Canada is not exempt from this problem: In 2018, about 8.7 % of the population were classified as low income (Government of Canada, 2019). While poverty is a continuing policy challenge, appropriate allocation of resources and effective strategies can help people to exit poverty (United Nations, 2019).

The eradication of poverty is also expressed in the UN's Sustainable Development Goals (SDGs). Goal 1 of the SDGs is to "end poverty in all its forms everywhere" (United Nations, 2019). Canada adopted the 2030 Agenda for Sustainable Development in 2015, which includes the 17 Sustainable Development Goals (Global Affairs Canada - Affaires mondiales Canada, 2017).

While poverty reduction is expressed as a goal for national and international policy, the policy decisions required to allocate resources and opportunities equitably are highly dependent on local attitudes, opinions, and biases regarding poor people. These attitudes are reflected at the voting booth, influencing what policy is made. They are also reflected in conversations, both in person and virtually. Discussions around poverty and universal basic income are flourishing online.

The analysis of online conversations to understand general attitudes presents problems that are related to the fact that the online population on different platforms does not always directly represent the general population. However, widespread access to the Internet makes data collection through online platforms a way to grasp a considerable percentage of results that reflect different attitudes and opinions in the online platforms. In the world, almost half of the population has access to the Internet, whereas in Canada 91 % of the population has access to the Internet (International Telecommunication Union, 2020).

The use of the Internet and specifically Twitter for capturing data is an approach that takes into consideration the posts published from a wide number of individuals (users). The use of Twitter is considerably high in Canada, with 49.7 % of Canada's online population (Slater, 2018).

The increasing popularity and the importance of social media use and communication have led to significant academic interest. This is particularly true as online conversations have begun to drive real-life social movements (Earl, 2016). Public opinion and associated feelings concerning a specific topic can be measured in social media platforms (Reyes-Menendez et al., 2018).

While several qualitative approaches are used to analyze online commentary, one of the most popular methods to the formal analysis of online content is sentiment analysis (Andreotta et al., 2019; Greaves et al., 2013; Neuendorf & Kumar, 2016). Sentiment analysis is part of text mining and utilizes techniques from natural language processing (NLP), information extraction (IE), artificial intelligence (AI), and information retrieval (IR) (Chen et al., 2018; Farzindar & Inkpen, 2015; Patel et al., 2020). These techniques capture opinions efficiently from a text written in syntactically correct and explicit language. However, sentiment analysis techniques demonstrate lower accuracy when processing informal language (Kumar & Garg, 2020).

Both published articles and proposed techniques for sentiment detection in social media have become more common (Kumar & Jaiswal, 2020). Twitter's open application programming interface (API) contributes to its status as the most studied microblog for sentiment analysis, while a supervised classification method called support vector machine (SVM) is the most popular tool for analyzing tweets (Keramatfar & Amirkhani, 2019).

Sentiment analysis has a variety of applications, including healthcare (Tuarob et al., 2014), finance (Oliveira et al., 2017), consumer markets (Arias et al., 2014), and government (Finn & Mustafaraj, 2013). For example, sentiment analysis of health-related tweets can be used for disease surveillance and prevention including HPV (McGregor & Whicker, 2018) and COVID-19 (Chakraborty et al., 2020). These studies developed automatic processing algorithms to score and detect positive and negative tweets in order to understand how topic–specific messages are broadcasted and provide strategies for debunking incorrect information. They found that, while most individuals who write novel tweets share positive or neutral information, negative tweets were more likely to be retweeted. Their finding that negative information is more likely to be amplified suggests the use of fact-checkers to mitigate broadcasting of misleading information.

Several studies (Sanz-Hernández, 2019a, 2019b) evaluated energy poverty press releases to assess the impact of stakeholder engagement on energy policies. They demonstrated that press participation reflects social dynamics but can improve both policy and perception by engaging the public. While these studies did not include social media, this case provides a strong analogue for the value of public engagement in the policy process. Despite extensive academic interest in social media analysis, to our knowledge, no existing study examines public opinion concerning poverty on Twitter.

It is essential to acknowledge that online engagement in general and Twitter in particular are not entirely representative of society. However, the Twitter conversation can drive or reflect national consciousness. We collected and analyzed data about poverty on Twitter and used this data to explore the possibility of employing sentiment analysis to understand online opinion.

2. METHODOLOGY

Twitter data was collected using Nexaintelligence (*Nexalogy,* 2020). Nexalogy's custom algorithms identify the network of sharing, liking, and commenting specific tweets and hashtags, along with geolocation and information on relationships between users.

2.1 Initial Data Capture

Initial data capture was performed to determine the terms that would be used in the search. This data was collected between February 11 and March 7, 2019 (a period of 24 days) and captured 15,923 results. Some of the hashtags we found in the query results were #cdnpoli, #Canada, #poverty, #cdnecon, and #abpoli. Furthermore, some of the top words were *Poverty*, *Canada*, *Canadian*, and *Alberta*. These results helped us to set up our search strategy related to the terms mentioned above.

2.2 Data Collection: Words and Hashtags

Twitter data was collected from March 18 to June 18, 2019. The data collection lasted three months. After this time, it was determined that the number of tweets collected was sufficient and representative. The queries showed 58,247 results. After removing the retweets and replicas, there were 11,386 results left. The query looked for the word *Poverty* or the hashtag #poverty and the post must have contained one or more of the words specified in Table 1: List of Terms and Hashtags.

Table 1
List of Terms and Hashtags

Canada	#abpoli	#Canada
Alberta	#bcpoli	Canadian
British Columbia	#mbpoli	Montreal
Manitoba	#nbpoli	Calgary
New Brunswick	#nlpoli	Edmonton
Northwest Territories	#nspoli	Toronto
Nova Scotia	#onpoli	Ottawa
Nunavut	#skpoli	Prince Edward Island
Ontario	#cdnpoli	Quebec
	#cdnecon	Saskatchewan

Own elaboration

2.3 Data Collection: Geolocation

Data collection was restricted to posts that contained geolocation data. The search was conducted between March 19 and June 18, 2019. Thirty-one thousand eight hundred twenty (31,820) results used the word Poverty or the hashtag #poverty and matched the geolocation criteria. Location was restricted to a 30-km radius from geolocated posts in the cities of Toronto, Montreal, Calgary, Edmonton, and Ottawa. About 52 % of the posts originated within a 30-km radius of Toronto, while about 21 % of the posts came from a 30-km radius of Ottawa. The 30-km radius from Calgary, Edmonton, and Montreal represent approximately 13 %, 9 % and 5 % of total posts, respectively. It should be noted that the search was conducted in English only. After removing the retweets and replicas, 9,507 results remained for analysis.

2.4 Sentiment Analysis

In order to conduct the sentiment analysis, a new AI was coded and trained (Daityari, 2019). Developing a case-specific AI ensured that the algorithm was centered on the distinct subject of this analysis. The first step was to manually evaluate a proportion of the dataset to allow the AI to correctly assess the subject-based sentiment. Manual coding of the tweets involved labeling them as "positive" if they had a positive connotation and "negative" for those with negative connotations. In total, more than 7 % of the database was evaluated manually, which corresponds to more than 1,500 tweets. This manual coding was then used to train the AI. Then, to prepare the training and main samples, the dataset was split into two data frames: one containing tweets with pre-coded sentiments to apply the algorithm and the other one to be used after training the AI. In both data frames, each tweet was then separated word by word and filtered before being fed to the AI. The filter included removing punctuation and neutral words [1], and converting all the words to lower case. To train the AI, two packages were used: the SentimentAnalyzer and NaiveBayesClassifier modules from the NLTK package for Python [2] (Bird et al., 2009). From the database with predefined sentiments, 30 % of the data was fed to train the AI and the rest was used to test its accuracy. From the sample, the AI would create a dictionary and associate sentiments to specific words. If the AI results' accuracy comparing predicted and known sentiment does not reach at least 60 %, another 30 % of the

1 Neutral words include "i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself", "yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itself", "they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that", "these", "those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having", "do", "does", "did", "doing", "a", "an", "the", "and", "but", "if", "or", "because", "as", "until", "while", "of", "at", "by", "for", "with", "about", "against", "between", "into", "through", "during", "before", "after", "above", "below", "to", "from", "up", "down", "in", "out", "on", "off", "over", "under", "again","further", "then", "once", "here", "there", "when", "where", "why", "how", "all", "any", "both", "each", "few", "more", "most", "other", "some", "such", "no", "nor", "not", "only", "own", "same", "so", "than", "too", "very", "s", "t", "can", "will", "just", "don", "should", and "now".

2 The NLTK package is specialized in natural language tools: https://www.nltk.org/.

sample from the database would be selected to train the AI. Once the accuracy threshold has been reached, the other database would be fed to the AI; however, in this case, getting the AI would predict the sentiment associated with the sample tweets. The results are then exported and analyzed. (See Figure 1).

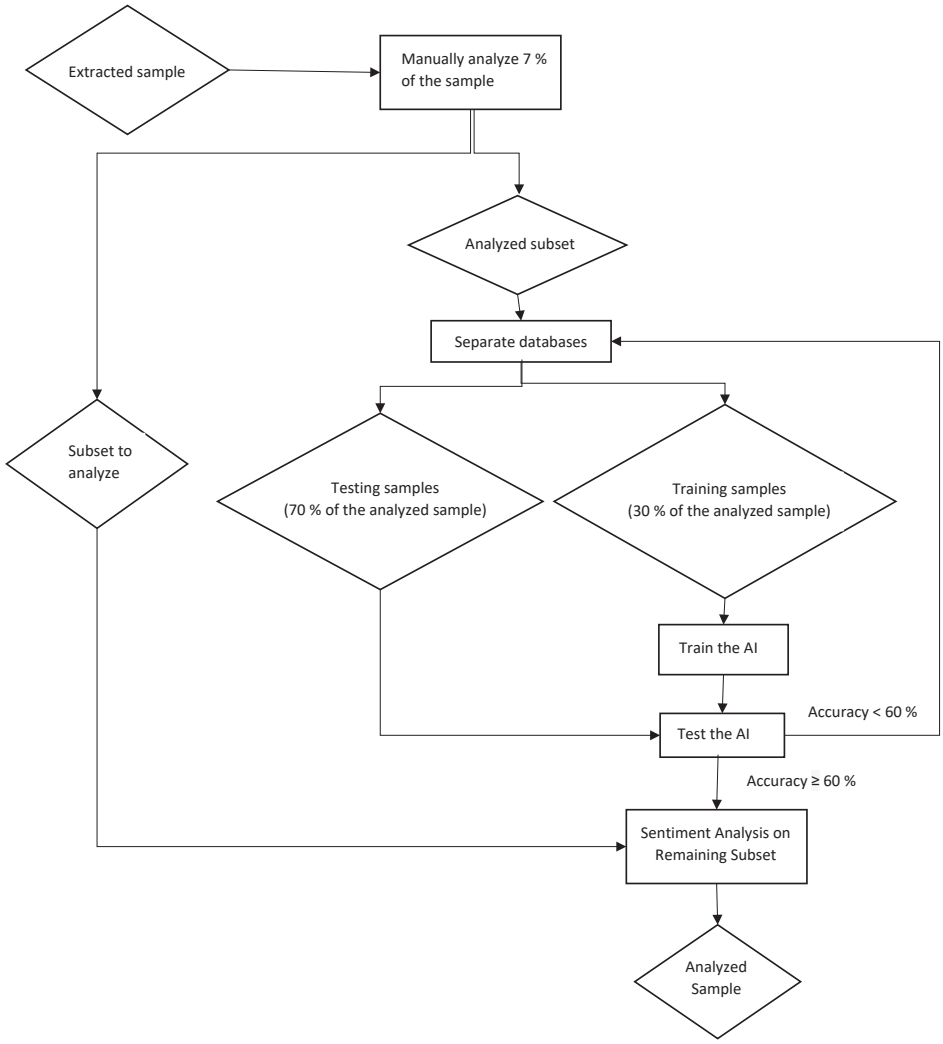

Figure 1. Sentiment Analysis Process
Own elaboration

3. RESULTS AND ANALYSIS

Figure 2 compares the heat map of location-based words and hashtags (on the left-hand side) with geolocated data (on the right-hand side). Location-related words and hashtags like Canada, #abpoli, Alberta, and #bcpoli yielded data from across the Americas as well as from Europe and Africa. In contrast, geolocated tweets about poverty in Canada were generated within the country.

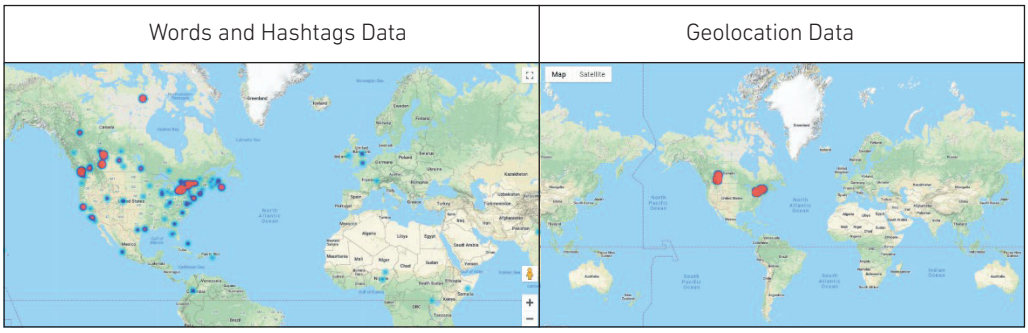

Figure 2. Heat Map
Own elaboration using Google Maps

During this period, the most highly shared link was related to a *New York Times*' opinion article comparing the management of poverty in the United States and Canada (Brooks, 2019). Chief among this discussion was the policy choices behind the level of poverty in both countries. This tweet was viewed as positive through manual annotation. The link was shared 3,148 times. After removing the retweets and replicas as well as the tweets that were evaluated manually, 524 tweets that shared the link evaluated by the AI were left. Our analysis indicated that 389 tweets were classified as positive and 135 tweets were considered negative. So, we can see that almost 75 % of the evaluated posts that contain this link have been listed as positive.

Table 2
Positive and Negative Posts

Class	Tweet
Positive	The world needs more Canadian energy via energy solves poverty
Negative	Approximately 1.3 M CHILDREN in Canada are living in poverty that is 1 in 5

Own elaboration

In addition to the most shared link, Table 2 displays two representative tweets from our dataset. The first row shows a tweet predicted as positive by the AI. This tweet, which focused on Canadian energy, argued that energy availability solves poverty. The second row shows a

representative negative tweet. The language was rated as negative by the AI. We can see that the positive sentence mentions "solves poverty" which has a positive connotation. While the negative sentence indicates that "Approximately 1.3 MILLION CHILDREN in Canada are living in poverty." Poverty has a negative connotation. "Solves poverty" or "Win the war against poverty" are positive sentences which indicate that poverty is detected as a negative word, even though the informative action may be to raise awareness with the intention and (positive) sentiment to change something.

At the end, we evaluated 20,894 results, out of which 44 % were classified as positive, while 56 % were classified as negative (see Table 3).

Table 3
Sentiment Analysis Results

	Number of Results	Percentage
Positive	9,259	44 %
Negative	11,635	56 %
Total	20,894	100 %

Own elaboration

Because poverty is not a simple or neutral concept, the ability of the AI to detect sentiment was limited. During the verification process, our AI was able to detect sentiment for a cleaned dataset at 99.7 % accuracy (Daityari, 2019). In contrast, our AI correctly identified pre-annotated sentiment in our novel dataset at only 60 % accuracy. Indeed, poverty is not a naturally neutral topic. Mentioning poverty generate strong responses, and the concept itself is seen as negative. The tweets highlighted in Table 2 demonstrate the challenges inherent in assigning sentiment values to communications regarding poverty.

4. CONCLUSION

While the proportion of positive tweets regarding poverty is slightly lower than the negative ones, this result does not get us closer to understanding sentiment regarding poverty in Canada. In order for algorithmic approaches to sentiment analysis to be useful in understanding public opinion, this method must be paired with more traditional approaches to understanding opinion. This could be a modified approach to semi-structured interviews, survey instruments or "seed tweets" related to specific anti-poverty policies. The AI could then be trained on this targeted information to analyze actual sentiment related to anti-poverty policy. The results also do not reflect the margin of error associated with the AI. In future works, we will estimate a confidence interval that places the predicted sentiment in a probable range of results.

REFERENCES

Andreotta, M., Nugroho, R., Hurlstone, M. J., Boschetti, F., Farrell, S., Walker, I., & Paris, C. (2019). Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis. *Behavior Research Methods*, *51*(4), 1766-1781. https://doi.org/10.3758/s13428-019-01202-8

Arias, M., Arratia, A., & Xuriguera, R. (2014). Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology*, 5(1), 8:1-8:24. https://doi.org/10.1145/2542182.2542190

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Brooks, D. (2019, April 4). Opinion | Winning the War on Poverty. *The New York Times*. https://www.nytimes.com/2019/04/04/opinion/canada-poverty-record.html

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 Tweets by Deep Learning Classifiers—A study to Show how Popularity is Affecting Accuracy in Social Media. *Applied Soft Computing*, *97*, 106754. https://doi.org/10.1016/j.asoc.2020.106754

Chen, N.-C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans. Interact. Intell. Syst.*, *8*(2), 9:1–9:20. https://doi.org/10.1145/3185515

Daityari, S. (2019, September 26). *How To Perform Sentiment Analysis in Python 3 Using the Natural Language Toolkit (NLTK)*. DigitalOcean. https://www.digitalocean.com/community/tutorials/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk

Earl, J. (2016). *"Slacktivism" that works: "Small changes" matter.* The Conversation. http://theconversation.com/slacktivism-that-works-small-changes-matter-69271

Farzindar, A., & Inkpen, D. (2015). *Natural Language Processing for Social Media*. Morgan & Claypool Publishers. http://gen.lib.rus.ec/book/index.php?md5=10fbd73c15d6d25d8776c08835e45040

Finn, S., & Mustafaraj, E. (2013). Learning to Discover Political Activism in the Twitterverse. *KI - Künstliche Intelligenz*, *27*(1), 17–24. https://doi.org/10.1007/s13218-012-0227-y

Global Affairs Canada-Affaires mondiales Canada. (2017, June 8). *The 2030 Agenda for Sustainable Development*. GAC. https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/priorities-priorites/agenda-programme.aspx?lang=eng

Government of Canada, S. C. (2019, February 26). *Low income statistics by age, sex and economic family type*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1110013501

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *Journal of Medical Internet Research*, *15*(11), e239. https://doi.org/10.2196/jmir.2721

International Telecommunication Union. (2020). *Individuals Using the Internet (Percentage of population)—Canada | Data*. The World Bank. https://data.worldbank.org/indicator/IT.NET.USER.ZS?locations=CA

Keramatfar, A., & Amirkhani, H. (2019). Bibliometrics of Sentiment Analysis Literature. *Journal of Information Science*, *45*(1), 3-15. https://doi.org/10.1177/0165551518761013

Kumar, A., & Garg, G. (2020). Systematic Literature Review on Context-Based Sentiment Analysis in Social Multimedia. *Multimedia Tools and Applications*, *79*(21-22), 15349-15380. https://doi.org/10.1007/s11042-019-7346-5

Kumar, A., & Jaiswal, A. (2020). Systematic Literature Review of Sentiment Analysis on Twitter Uusing Soft Computing Techniques. *Concurrency and Computation: Practice and Experience*, *32*(1), e5107. https://doi.org/10.1002/cpe.5107

McGregor, K. A., & Whicker, M. E. (2018). Natural Language Processing Approaches to Understand HPV Vaccination Sentiment. *Journal of Adolescent Health*, *62*(2), S27-S28. https://doi.org/10.1016/j.jadohealth.2017.11.055

Neuendorf, K. A., & Kumar, A. (2016). Content Analysis. In *The International Encyclopedia of Political Communication* (pp. 1-10). American Cancer Society. https://doi.org/10.1002/9781118541555.wbiepc065

Nexalogy. (2020). *Nexalogy*. https://nexalogy.com/

Oliveira, N., Cortez, P., & Areal, N. (2017). The Impact of Microblogging Data for Stock Market Ppreddiction: Using Twitter to Predict Returns, Volatility, Trading Volume and Survey Sentiment Indices. *Expert Syst. Appl.* https://doi.org/10.1016/j.eswa.2016.12.036

Patel, J., Dubey, R., & Gupta, R. K. (2020). PMI-IR Based Sentiment Analysis Over Social Media Platform for Analysing Client Review. In S. Smys, T. Senjyu, & P. Lafata (Eds.), *Second International Conference on Computer Networks and Communication Technologies* (pp. 204–212). Springer International Publishing. https://doi.org/10.1007/978-3-030-37051-0_23

Reyes-Menendez, A., Saura, J. R., & Alvarez-Alonso, C. (2018). Understanding #World EnvironmentDay User Opinions in Twitter: A Topic-Based Sentiment Analysis Approach. *International Journal of Environmental Research and Public Health*, *15*(11), 2537. https://doi.org/10.3390/ijerph15112537

Sanz-Hernández, A. (2019a). Medios de comunicación y *stakeholders*: Contribución al debate público de la pobreza y justicia energética en España / Media and Stakeholders: Contribution to the Public Debate on Poverty and Energy Justice in Spain. *Revista Española de Investigaciones Sociológicas*, *168*. https://doi.org/10.5477/cis/reis.168.73

Sanz-Hernández, A. (2019b). Social Engagement and Socio-Genesis of Energy Poverty as a Problem in Spain. *Energy Policy*, *124,* 286-296. https://doi.org/10.1016/j.enpol.2018.10.001

Slater, M. (2018). *By the numbers: Twitter Canada at Dx3 2018*. Blog Twitter. https://blog.twitter.com/en_ca/topics/insights/2018/TwitterCanada_at_Dx3.html

Tuarob, S., Tucker, C. S., Salathe, M., & Ram, N. (2014). An Ensemble Heterogeneous Classification Methodology for Discovering Health-Related Knowledge in Social Media Messages. *Journal of Biomedical Informatics*, *49*, 255–268. https://doi.org/10.1016/j.jbi.2014.03.005

United Nations. (2019). *The Sustainable Development Goals Report 2019*. https://www.un-ilibrary.org/content/publication/55eb9109-en

United Nations Statistics Division. (2019, December 20). *SDG Indicators*. Sustainable Development Goal Indicators Website. https://unstats.un.org/sdgs/indicators/database