The 10th IEEE International Conference on Awareness Science and Technology (iCAST 2019)

Sifting for Deeper Insights from Public Opinion: Towards Crowdsourcing and Big Data for Project Improvement

Jean Marie Tshimula,[1,2,3] Mary Muthoni Njuguna,[1,4] Thierry Roger Bayala,[1,5] Mbuyi Mukendi Didier,[2,6]
Achraf Essemlali,[1,3] Hugues Kanda,[2,7] Numfor Solange Ayuni[1,4]
[1]African Bridge and Research Society [2]Groupe de Recherche de Prospection et Valorisation des Données
[3]University of Sherbrooke, Canada [4]Tohoku University, Japan [5]Miyagi University, Japan
[6]University of Kinshasa, DR Congo [7]University of La Rochelle, France
e-mail: kabj2801@usherbrooke.ca

***Abstract*—Over the years, there seems to be a unidirectional top-down approach to decision-making in providing social services to the masses. This has often led to poor uninformed decisions being made with outcomes which do not necessarily match needs. Similarly from the grassroots level, it has been challenging to give opinions that reach the governing authorities (decision-making organs). The government consequently sets targets geared towards addressing societal concerns, but which do not often achieve desired results where such government endeavors are not in harmony with societal needs. With public opinions being heard and given consideration, societal needs can be better known and priorities set to address these concerns. This paper therefore presents a priority-based voting model for governments to collect public opinion data that bring suggestions to boost their endeavors in the right direction using crowdsourcing and big data analytics.**

***Index Terms*—Business Model, Public Opinion, Public Participation, Big Data**

I. Introduction

Since time immemorial, governments always decide projects and initiatives that directly impact citizens. This is in fulfilling their mandated role of formulating and enforcing public policy. However, more often than not, these policies are not reflective of the societal needs which are encountered by the populace on a daily basis. How can a government gather the population's outlooks? Collecting public opinion across the country using time-worn methods for example going door-to-door using questionnaires would be tedious and time-consuming. In order to conduct such an exercise extensive financial and human resources would be required for creation of teams in charge of data collection, data encoding, and data analysis and even security and transportation for staff. Hence, this might effectively have some influence on governments' decision-making process for improvement of social services to masses.

The advent of the digital age has fundamentally changed the speed of acquisition and sharing of data by means of web-based platforms. Crowdsourcing is a pervasive technology that enables citizens to participate in activities that benefit themselves and their communities [1], [2]. This then affords a new opportunity to create an avenue for public opinion to be acquired by the governments for use in the decision-making process. The advantage of crowdsourcing over the paper-based survey questionnaire is that the government can verify whether respondents are legitimate citizens thanks to features like *"national identity number,"* and allows for setting up multiple-choice questions or written answer questions with the possibility of correcting misspellings since the data encoding from the survey may include many typos due to the illegible writings of respondents and many duplicates due to the imprudence of data encoders. Based on this assumption, the probability that the analysis would be biased is high.

Fundamentally, a country is composed of cities each of which faces its own realities and issues. When the problems of a specific city become recurrent, it is more likely to be obvious that local people will enroll them to the list of priorities and questions that concern the city most. If there exists a bridge between the government and citizens, citizens may play a major role by providing and reporting in real-time, useful information to the government. Furthermore, the government should be able to collect public opinion feedback data from all over the country in order to improve existing projects and assess the effectiveness of ongoing projects in solving the problems they were intended to solve. This constitutes a big data challenge to efficiently and correctly convert this data into a set of priorities in order to uncover effective hypotheses and unimaginable solutions that will boost the governments endeavors in the right direction by establishing a strong road-map (by city or region, or for the whole country).

The rest of the paper is organized as follows. We describe the literature review of previous work in Section 2. In section 3, we introduce our methodology framework and describe various methods in detail. In section 4 we present data and report our results. In section 5, we conclude and discuss some future research directions.

II. Related Work

Prior research has tackled the priority-based model issue. For instance, Tshimula and Togashi (2018) in [3] built a machine learning-based workflow for the project viability to classify the most fruitful sectors and to highlight the most relevant topics that investors may focus on for investing in Africa using the project portfolio of the African Development Bank (AfDB). This research found seven promising sectors which have a significant impact on the business for successful growth in the African continent. In addition, all of them fit the High Five priority agenda defined by the AfDB. The AfDB projects have been initiated and implemented in cooperation

with the regional member countries (RMCs). At this level, the RMCs participate in the elaboration of continental projects, thus, it is also the same thing they are expected to do with their respective citizens for the local projects.

Ribeiro et al. (2011) in [4] introduced crowdMOS, a crowdsourcing for the mean opinion score (MOS) to measure in both efficient and practical ways the scores attained by internet users that attended MOS-like listening study. Furthermore, crowdMOS also foresees a wide range of open-source features to effectuate *"subjective opinion experiments in a customizable and user-friendly way, completely shielding the researcher from the details surrounding crowdsourcing or Mechanical Turk, and from the bookkeeping required for user studies."* Wu et al. (2015) in [5] addressed the algorithmic optimizations towards the diversity of opinion of crowdsourcing marketplaces in order to select the right workers that will form a sage crowd.

Kittur et al. (2011) in [6] proposed *crowdForge*, a framework to accomplish complex and independent tasks from a set of many small contributions in crowdsourcing markets using micro-task markets. The aforementioned studies [3], [4], [5], [6] strongly inspire the present research to introduce a priority-based voting model referring to citizens as a wise crowd who can interact with governing authorities to voice their viewpoints (or suggestions). Our approach is built in the way to help governments deal with bulky data stemming from citizens. Thereafter, they can easily map societal problems by geographical position in order to more accurately prepare future expenditure commitments. We use big data analytics to turn public opinion into actionable insights for improving government projects, prioritizing projects and providing a better platform for public policy and management, to name a few.

III. Public Participation Framework

We introduce a new workflow that addresses the problem of public opinions from a different angle. Our workflow is composed of the following:

Sectors. The economy of a nation is composed of several sectors with a proportion of the population affiliated to activities of each sector, classifying them into primary, secondary and tertiary sectors [7]. However, our paper focuses on a breakdown of these sectors and will therefore comprise education, health, infrastructure, water, environment, and others. Citizens do pay taxes and in turn should be provided a fair society where good health, water, environment, roads, housing, and others, are assured [11]. In most literature, however, an approved development strategy constituent is the conviction of key critical sectors [8], which is the focus of this paper.

Geographic Information. Citizens are dispersed within a country based on their origin or destination. These locations can be cities or regions, making up geographic information or geographical location [8]. Different geographical locations have sectors with the most concerns, requiring immediate government intervention while others follow suite.

Application. With the rapid increase in the use of technology, an application software, easily accessible by a bulk of the population, is designed to collect data on public opinions. Through this app, public opinions or votes will be channeled directly to the appropriate unit, without interference by governing authorities.

Big Data System. Big data is viewed as enormous sets of data which are huge, complex and varied structures, difficult to store, analyzed and visualized, for advance results [10]. The big data system is, therefore, the appropriate system to collect and analyze the data involving public opinions on various geographical locations and prioritized sectors for projects implementation by the government. Coordinated by a data control unit, the big data system will collect all the data from the information provided by the citizens through the app. This data control unit should be independent to avoid data tampering or some influences from the government. The big data system stands as a data repository on public opinions and all data collected will be recorded and analyzed.

Insight. Results obtained from Big Data analysis will provide insights to government projects implementation based on the prioritized sectors as voted by the citizens. This will ensure efficient projects to match the most wanted needs of the public.

Government. The government here is regarded as an open government which synchronizes the vision of the economy and voice of the citizens, through its governing authorities. Such an open government is deemed to be efficient, democratic, legitimate and not corrupt [9]. The government is in charge of formulating, initiating and funding projects to allow the growth and development of the economy.

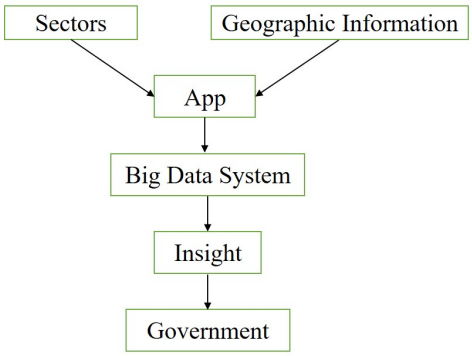

Fig. 1: Public participation workflow for transforming stances.

Our proposed workflow involves three players; Citizens, Data Control Unit, Government. Through the designed application, citizens will provide information on their respective geographical locations and the sectors placed in order of priority so as to guide government and local authorities projects implementation, meeting the most economic/societal needs. This will allow the masses to contribute to decision-making, thereby guiding economic activities and reforms. Through crowd-sourcing, the data collection unit will receive inputs from the citizens across the country and will perform analyses to get results that match the needs of the society/economy.

The results will provide insights into the government on preference projects to address societal concerns. Through the same platform the citizens can give feedback on ongoing or finished projects.

IV. Experimental Setup

A. Data Description

We ran the experiments on the questionnaire dataset of 40.000 respondents. We focused our study on the Democratic Republic of the Congo, where the dataset was collected. The questionnaire implies three questions, the dataset itself includes six variables and each row represents the respondent's answer. While the first question of the questionnaire is about prioritizing three sectors, we split the answer into three variables where each choice consisted of a variable according to their order. We used questions below to collect the dataset:

- With your local region in mind, what are the three sectors of priority for the government to focus on? For this question, we expect the three answers to be comprised in the following list: *education, health, transportation, energy, construction* and *agriculture*. Therefore, we built four variables from this question, namely region and three choices depending upon sectors of priority. We principally use the region of the respondent as the first variable, and we then take independently the three choices based on the order that they are selected by the respondent. Even though the three choices consist of the priority of the respondents, we ask them to specify which of these sectors is the first, second and third choice. For randomly selected sectors which miss the indication of the priority, we simply label their order by relying on how they are provided. It is worth recalling that the three choices must be different from each other. We assume that the order of choices made by citizens is of great importance since this somehow appoints what should be treated with more importance than another. Based on the order of choices, we further build the target feature of predictive models.
- Do you think that public participation is an effective tool for prioritizing government developmental agenda for different regions? The expected answer is *yes* or *no*.
- Based on your sectors of priority for development, how much time do you think is necessary for successful project completion? Here, the expected answer is *6, 12* or *24 months*.

It is worth to mention that the number of sectors in our questionnaire is not exhaustive, we only consider six sectors for the experiments. We fairly considered the respondents' opinions, regardless of their educational qualification. We assume that the opinion of every single citizen is important. For instance, some villages may be formed of relatively uneducated people, it would be unfair not taking into consideration the opinions of locals simply because they are less educated. They mostly live in these areas and better know their issues of concern. We restricted our study[1] to the aforementioned variables. However, in the future, we would like to add other features to better know the respondents such as income, family size, gender, marital status, affinity score between individuals within the community to measure their togetherness [12], access to financial services and financial inclusion [13], etc. We also aim to build a web- and mobile-based application to enable rapid data collection at scale. For the security aspect, the app can use the national id cards of respondents in order *(i)* to identify them as nationals, *(ii)* to avoid duplicate participation of a single respondent in the survey and *(iii)* to filter out respondents whom the age does not meet the survey requirements. We would like to use blockchain to ensure the immutability of data by keeping it as genuine as it was collected, i.e. by resisting to the modification of the data.

B. Data Distribution

The visualization is an important step in the machine learning pipeline for building models. It helps discover hidden patterns within data, reveals outliers and depicts the data distribution by giving the demography of each feature. From this step, we can easily rule out models that do not perform well with our features and retain the promising models. Figure 2 shows the distribution of sectors upon the priority of the three choice slots for every single region. While figure 3 represents the distribution of the time that respondents expecting their sectors of priority are going to be realized. We, however, grouped the expectation by region, we report that results are diversified based on choices made, although agriculture, education, and health appear to be slightly prevalent when clustering the data. We found that most people would like to have their desires to be implemented within 12 months.

The data distribution indicates that data from multiple groups have the same variance, in other words, there is homogeneity of variances. There is no missing data and there is therefore no need to apply the process of data standardization. Furthermore, the analysis of the data distribution reveals which features are more to consider, which task to predict and what models to use.

C. Feature Sets

To build the models, we took into account six features as indicated in Section IV-A. More specifically, we reduced the number of these descriptive features to four features (*priority, region, expectation* and *choice*), where the last feature was obtained by stacking the values of the three choices. Additionally, we created a new feature for indicating the order of priority made by the respondent in the choice of the three sectors of interest for the development of his/her region. We call this feature 'choice importance', then use it as the target feature. The target feature is considered as being a class, and the set of these classes includes (1, 2 and 3). Recall that the class 1, 2 and 3 respectively denote the priority of the first, second and third choice. As a result, we used the following

[1]The dataset will be released at *https://www.github.com/afbrs/pubop*

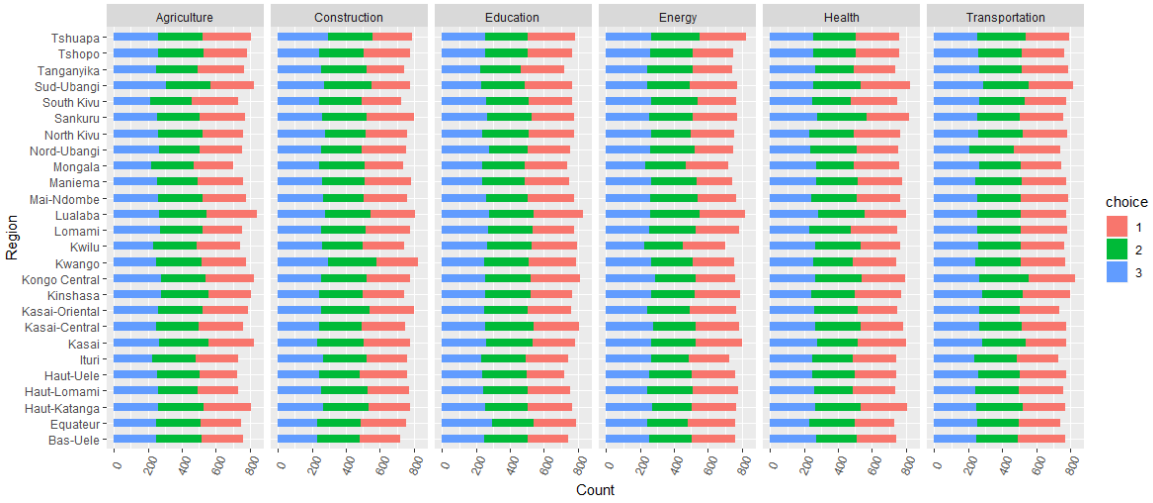

Fig. 2: Distribution of the order of choices upon sectors by region.

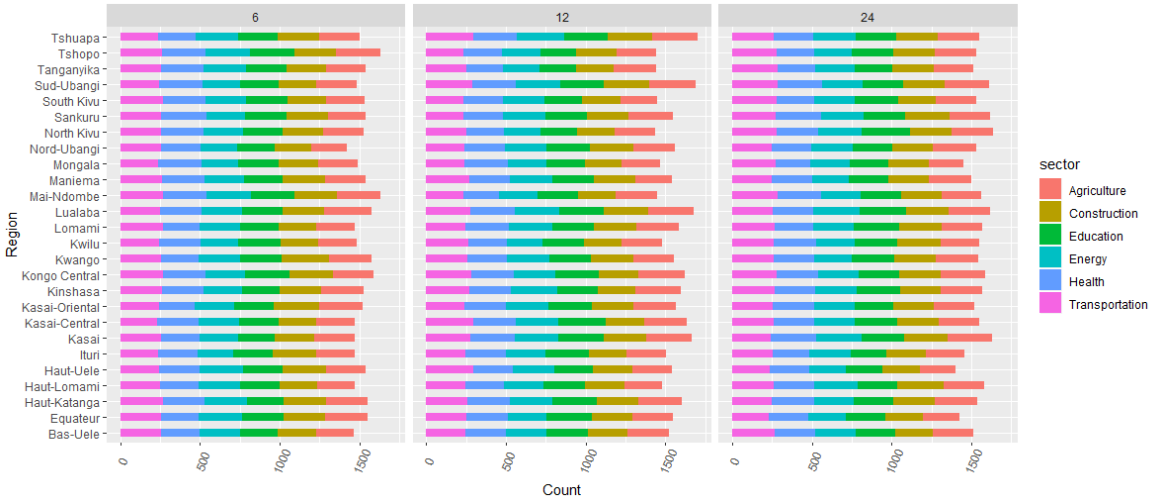

Fig. 3: Distribution of the envisaging time for the implementation of the chosen sectors by region.

five features: *priority, expectation, region, choice*, and *choice importance*.

D. Model Evaluation

We predicted the target feature *choice importance* for each region based on *(1)* sectors of interest (choice) of the respondents for the development of their regions, *(2)* the time (*expectation*) the respondents envisaged that their desires would be implemented by the government, and *(3)* the priority that they grant to their choices. We constructed the model using the feature *choice importance* as the response variable and other features as explanatory variables. We used seven different classifiers, namely Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Decision Tree (DT), Extra Trees (ET), Logistic Regression (LR), Perceptron (PERC) and Multi-Layer Perceptron (MLP). To evaluate the performance of the prediction of the importance of choices, we performed five-fold cross-validation for training and test data. To measure the reliability of our model, we computed the following metrics: Accuracy, Recall, Precision, F-1 score and Root Mean Squared Error (RMSE).

TABLE I: Prediction results for the seven classifiers

Classifier	Accuracy	Recall	Precision	F-1 score	RMSE
ET	0.997	0.997	0.996	0.997	0.003
SVM	0.574	0.432	0.440	0.436	0.426
LR	0.637	0.554	0.607	0.579	0.510
PERC	0.976	0.992	0.965	0.978	0.024
MLP	0.751	0.680	0.789	0.731	0.252
LDA	0.998	0.998	0.998	0.998	0.002
DT	1.000	1.000	1.000	1.000	0.000

E. Discussion

Table I shows that Decision Tree achieves the best performance and outperforms other classifiers in terms of all utilized metrics. SVM poorly performed the task and obtained smaller values for all the metrics compared to other classifiers,

although its accuracy is above-average, this does not guarantee promising results since its RMSE is too big. We group the results obtained in Table 1 into two different tiers based on their performance. The first-tier includes ET, PERC, LDA, and DT, whereas the second tier comprises LR, SVM, and MLP. We report that first-tier classifiers achieved the best performance and competitive results which even border up 100% in terms of accuracy and F-1 score, although DT is the only classifier to obtain a null RMSE. Owing to this, DT is the classifier for this task. LDA is greater than ET in terms of Accuracy but ET has achieved the second best RMSE which can be likely rounded to zero. We note that the performance of second-tier classifiers is not worse to some extent, except the fact that their RMSEs are quite bigger. However, when it comes to orient the decision-making process, the first-tier classifiers show a high ability to predict sectors in which the government should prioritize for the development of regions.

V. Conclusion and Future Work

We propose a priority-based voting model for helping governments collect public opinion data at scale and transform them into insights capable of orienting the governmental endeavors in the right direction. Experiments on real-data provide promising results and demonstrate the effectiveness of the used model. While ensuring transparency and equity, the proposed solution can also help improve development indices of some regions in the country and propel the development of sectors which are still lagging behind. Future research avenues are related to adding more features including income, social status (or position) and socioeconomic status of respondents, etc. and to reinforcing the security layer of the workflow by integrating the blockchain technology.

Acknowledgment

We would like to thank Jemimah Kibira from Tohoku University for fruitful discussions, suggestions for the survey questionnaires and for helpful feedback on earlier drafts of this manuscript. We would also like to thank Japan International Cooperation Agency (JICA) for the African Business Education (ABE) Initiative for youth through which the authors came to know each other and collaborate on this paper. Jean Marie Tshimula, Mary Muthoni Njuguna, Thierry Roger Bayala and Numfor Solange Ayuni are ABE alumni.

References

[1] D.C. Brabham, "Using Crowdsourcing In Government," IBM Center for the Business of Government, 2013, [online]: https://bit.ly/2JFTkVM.

[2] S. Hosio, J. Goncalves, V. Kostakos, J. Riekki, "Crowdsourcing Public Opinion using Urban Pervasive Technologies: Lessons from Real-Life Experiments in Oulu," Policy & Internet, 7(2), 203–22, 2015.

[3] J.M. Tshimula, A. Togashi, "Machine Learning-Based Framework for the Analysis of Project Viability," In IEEE ICCCS, 80–84, 2018.

[4] F. Ribeiro, D. Florencio, C. Zhang, M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," ICASSP'11, 2416–2419.

[5] T. Wu, L. Chen, P. Hui, C.J. Zhang, W. Li, "Hear the whole story: Towards the diversity of opinion in crowdsourcing markets," In VLDB Endowment, 8(5), 485–496, 2015.

[6] A. Kittur, B. Smus, R. Kraut, "Crowdforge: crowdsourcing complex work," In UIST'11, 43-52, 2011.

[7] M. Rosenberg (2007), "Sectors of the Economy," Retrieved Jul 15, 2014.

[8] G.J. Hewings, "The empirical identification of key sectors in an economy: a regional perspective," Developing Economies, 20(2), 173–195.

[9] A.J. Meijer, D. Curtin, and M. Hillebrandt, "Open government: connecting vision and voice," IRAS, 78(1), 10–29, 2012.

[10] S. Sagiroglu, and D. Sinanc, "Big data: A review," In CTS, 42–47, 2013.

[11] V. Morabito, "Big data and analytics for government innovation," In Big data and analytics (pp. 23–45). Springer, Cham, 2015.

[12] J.M. Tshimula, B. Chikhaoui, and S. Wang, "HAR-search: A Method to Discover Hidden Affinity Relationships in Online Communities," In IEEE/ACM ASONAM 2019.

[13] J.M. Tshimula, A. Togashi, N.S. Ayuni, "Towards Financial Inclusion-based Monetization Model for Startups Drive," In TEMSCON, 2018