Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 3, No.1, April 2021, pp. 0210106-01 ~ 021016-07
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v3i1.8460

Automatic Complaints CategorizSation Using Random Forest and Gradient Boosting

Muchamad Taufiq Anwar[1], Anggy Eka Pratiwi[2], Khadijah Febriana Rukhmanti Udhayana[3]

[1]Politeknik STMI Jakarta, Jl. Letjend Suprapto No.26, Central Jakarta 10510, Special Capital Region of Jakarta, Indonesia

[2]Department of Computer Science and Engineering, Indian Institute of Technology Jodhpur, Jodhpur, India

[3]Department of Informatics, Universitas Bali Internasional, Gg. Jeruk, Tonja, Kec. Denpasar Timur, Denpasar, Bali 80234, Indonesia

taufiq@stmi.ac.id

Abstract. Capturing and responding to complaints from the public is an important effort to develop a good city/country. This project aims to utilize Data Mining to automatize complaints categorization. More than 35,000 complaints in Bangalore city, India, were retrieved from the "I Change My City" website (https://www.ichangemycity.com). The vector space of the complaints was created using Term Frequency–Inverse Document Frequency (TF-IDF) and the multi-class text classifications were done using Random Forest (RF) and Gradient Boosting (GB). Results showed that both RF and GB have similar performance with an accuracy of 73% on the 10-classes multi-class classification task. Result also showed that the model is highly dependent on the word usage in the complaint's description. Future research directions to increase task performance are also suggested.

Keywords: automatic complaints categorization, multi-class classification, data mining, random forest, gradient boosting

1. Introduction

One way of giving a good service to the public is by giving room for the public to post a complaint to the government so that necessary actions can be taken as soon as possible. This is especially important for a developing country such as India to make the cities better. One of the initiatives that already taken is the "I Change My City" website which first debuted in Bangalore City, India, in 2012. This platform is managed by the Swachh Bharat Mission of the Ministry of Urban Development of the Government of India. By using this platform, citizens can post complaints they face in their city and the government will resolve the issue. To be resolved, the issues must be categorized and forwarded to the relevant agencies. Until now, this platform already has nearly 20 million users and 53 million complaints have been submitted, of which 94% percent have been successfully resolved. This research aims to create a

Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 3, No.1, April 2021, pp. 0210106-01 ~ 021016-07
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v3i1.8460

model so that when a new complaint is registered, it will get automatically classified. The classification task can be done by using Data Mining techniques. The automatic categorization of the complaints will help both the user and the government in reducing the overall complaints resolution time.

2. Methods

The research methods are shown in Figure 1. Complaint data are scraped from the "I Change My City" website (https://www.ichangemycity.com). The website itself was started in October 2019. The original scraped data from the website are in JSON format. The JSON file was then transformed into two-dimensional data using Python. The data consists of 44961 data which represent complaints from October 2019 to December 2020. The data are spread across 14 different classes. Each object is having 8 features that include id, category, sub-category, my category (specifically mentioned by the person lodging the problem), description of the complaint, and an associated image to describe the problem. Of the eight features, only two features are used in this research, i.e the complaint description and the complaint category. For this study, we focused only on Bangalore city and its top-10 complaint categories which happened to represent 79% of the total data. The sample data is shown in Table 1.

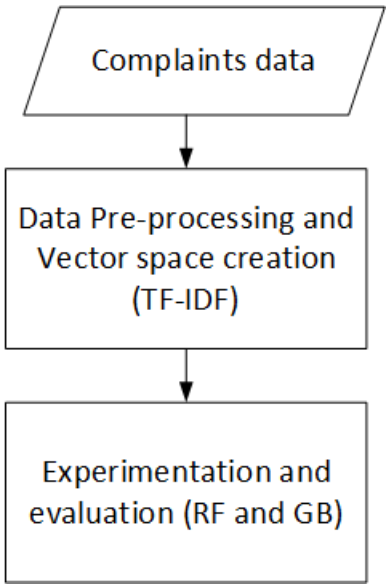

Figure 1. Research methods

Table 1. Sample data

Complaint_description	Complaint_category
Existing road inside GSS layout near Kudlu VGP...	Provide good driveable Roads
Street lights not working in this are 2 to 3 l...	Maintenance/Repair Of Streetlights
47th Cross, 8th Block Jayanagar Road Maintenan...	Fixing/Reparing Potholes

This research aims to categorize data into 10 broad classes. The data is also not perfectly balanced among classes as shown in Table 2. To handle this class imbalance, we utilize ensemble methods such as Random Forest (RF) and Gradient Boosting (GB) since the ensemble method takes care of the class imbalance. The data has majorly two components, a) free-flowing text (the text to be

Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 3, No.1, April 2021, pp. 0210106-01 ~ 021016-07
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v3i1.8460

predicted), and b) complaint categories (target classes). We are using Term Frequency–Inverse Document Frequency (TF-IDF) to give weights for every word to create a vector space for the text. TF-IDF is commonly used in text classification tasks such as in [1]. This research is essentially a multi-class classification problem. The implementation of these classification tasks was done using Python alongside its machine learning libraries i.e the ScikitLearn library.

RF and GB are newer-generation DM techniques that utilize bagging / boosting which improve the performance of the previous "traditional" techniques. These methods are also known as ensemble learning methods since they are using multiple learners (models) for the prediction task. Both of these methods are based on tree structure such as used in C4.5. C4.5 uses a single tree to make a prediction. C4.5 itself is a simple yet versatile classification method and had been used in wildfire modeling [2], rain prediction[3], [4], drug resistance prediction[5], etc. Bagging and boosting methods are a class of classification techniques that increase (boost) its performance by adding bags of models or by iteratively improve the model.

RF is a tree-based classifier that utilizes multiple (a bag of) trees where each tree has its own prediction ability. The creation of each tree is based on a random subset of the entire dataset and also a random subset of the entire attributes. In RF, each tree will give its prediction and the final prediction is taken from the voting from all trees. RF was introduced in 2014 by Breiman [6]. RF had been recently used in various fields including rain prediction[4], drug resistance prediction[5], geospatial pattern alaysis[7], prediction of electricity production [8], battery modeling [9], and landslide susceptibility mapping [10].

GB is another tree-based classifier that improves (boosts) its prediction ability by iteratively building a better tree from the previous iterations. Each iteration will reduce the error of the previous tree. GB was introduced in 2014 and is widely available on multiple platforms including Python, R [11], etc. GB is capable of both classification and regression tasks. GB had been recently used in various fields including rain prediction[4], rainfall prediction[12], transpiration estimation [13], undrained shear strength prediction[14], concrete strength prediction[15], groundwater level prediction[16], and solar irradiation forecasting[17].

3. Results and Discussion

The complaints from Bangalore City consist of 35,375 rows of data. The topmost recurrent complaints are related to road infrastructure, garbage/waste management, stray dogs, and floods as shown in Table 2. The word cloud (top 50 words) for complaints description is shown in Figure 2. This confirms the fact that most complaints are related to road infrastructure and waste management.

Table 2. The top recurring complaints in Bangalore

Rank	Category	Count	Percentage
1	Fixing/Reparing Potholes	10185	29
2	Maintenance/Repair Of Streetlights	6744	19
3	Clearance Of Garbage Dump Or Black Spot	5151	15
4	Repair of Existing Footpaths	3464	10
5	Maintenance Of Dry Waste Collection Centre	2638	7
6	Stray Dog Sterilisation/Animal Birth Control (ABC)	2022	6
7	Garbage Dumping In Vacant Lot/Land	1512	4
8	Provide good driveable Roads	1287	4
9	Desilting of storm water drains	1236	3

Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 3, No.1, April 2021, pp. 0210106-01 ~ 021016-07
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v3i1.8460

10	Collection Of Door-to-door Garbage		1136	3
		TOTAL	**35375**	**100**

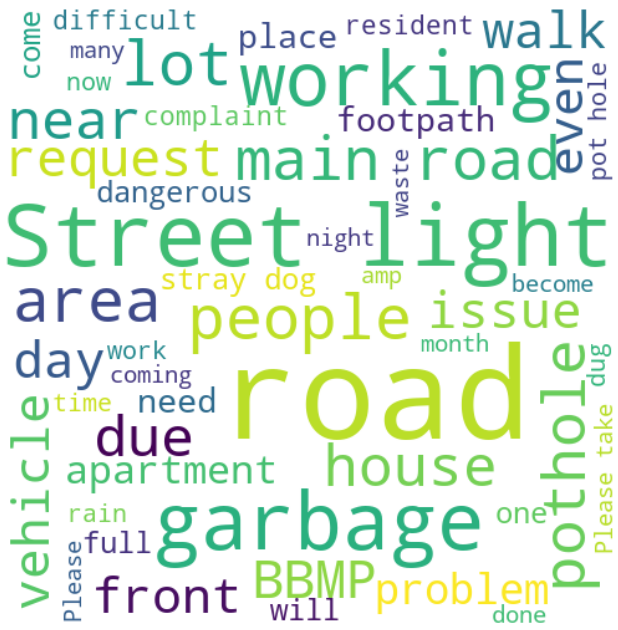

Figure 2. The wordcloud for the complaints

Data were split into 80% training data and 20% testing data. Our result showed that RF (n estimator = 100, using all features) gave better overall accuracy of 0.73 with the best f-1 score in the "Maintenance/Repair Of Streetlights" and "Stray Dog Sterilisation/Animal Birth Control (ABC)" categories. The worst classification performance is on the "Provide good driveable Roads" and "Maintenance Of Dry Waste Collection Centre" categories. This might be caused by similar words being used in several other categories, eg: "waste"/"garbage", and "road"/"street". Whereas the good performing categories have a distinguished use of words such as "streetlights", "dogs", "water", and "potholes". The result of performance metrics of the RF model is shown in Table 3 and the confusion matrix is in Table 4.

Table 3. Performance metrics for RF model

Categories	Precision	Recall	F1-score	Support
Clearance Of Garbage Dump Or Black Spot	0.523224	0.796258	0.631492	962
Collection Of Door-to-door Garbage	0.664634	0.488789	0.563307	223
Desilting of storm water drains	0.804878	0.788845	0.796781	251
Fixing/Reparing Potholes	0.715209	0.901294	0.797541	2087
Garbage Dumping In Vacant Lot/Land	0.764151	0.517572	0.617143	313
Maintenance Of Dry Waste Collection Centre	0.454054	0.159696	0.236287	526
Maintenance/Repair Of Streetlights	0.964784	0.941977	**0.953244**	1396
Provide good driveable Roads	0.588235	0.039216	0.073529	255
Repair of Existing Footpaths	0.615591	0.349618	0.445959	655
Stray Dog Sterilisation/Animal Birth Control (ABC)	0.900474	0.933661	**0.916767**	407

Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 3, No.1, April 2021, pp. 0210106-01 ~ 021016-07
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v3i1.8460

Table 4. Confusion matrix for RF model

Cat	0	1	2	3	4	5	6	7	8	9
0	766	30	7	48	30	63	4	1	5	8
1	99	109	0	2	1	10	1	0	0	1
2	9	0	198	33	0	4	1	0	2	4
3	44	0	17	1881	1	4	26	3	106	5
4	121	3	1	6	162	15	1	0	2	2
5	349	22	7	35	16	84	2	0	4	7
6	16	0	2	42	2	3	1315	0	4	12
7	14	0	3	205	0	0	3	10	20	0
8	31	0	10	370	0	1	8	3	229	3
9	15	0	1	8	0	1	2	0	0	380

The GB method produced a slightly worse accuracy of 0.72 (parameter setting: n estimator = 100, Learning Rate (LR) = 0.1, max features=20, max depth=20). The performance metrics of the GB model are shown in Table 4 and the confusion matrix is in Table 5. The confusion matrix shows that most of the data are misclassified to the “Fixing/Repairing Potholes” category. This might be caused by the words used in that category are also present in the other categories e.g: “street”/”road”. The best performance is, again, on the ”Maintenance/Repair Of Streetlights” and “Stray Dog Sterilisation/Animal Birth Control (ABC)” category which uses a distinctive keyword such as ”streetlight” and “dogs”.

Table 5. Performance metrics for GB model

Category	Precision	Recall	F1-score	Support
Clearance Of Garbage Dump Or Black Spot	0.532641	0.746362	0.621645	962
Collection Of Door-to-door Garbage	0.648000	0.363229	0.465517	223
Desilting of storm water drains	0.848039	0.689243	0.760440	251
Fixing/Reparing Potholes	0.666091	0.923335	0.773896	2087
Garbage Dumping In Vacant Lot/Land	0.725888	0.456869	0.560784	313
Maintenance Of Dry Waste Collection Centre	0.406977	0.13308	0.200573	526
Maintenance/Repair Of Streetlights	0.976709	0.931232	**0.953429**	1396
Provide good driveable Roads	0.394737	0.117647	0.181269	255
Repair of Existing Footpaths	0.617391	0.325191	0.426000	655
Stray Dog Sterilisation/Animal Birth Control (ABC)	0.953125	0.899263	**0.925411**	407

Table 6. Confusion matrix for GB model

Cat	0	1	2	3	4	5	6	7	8	9
0	718	19	6	106	36	66	6	2	3	0
1	115	81	0	8	0	17	2	0	0	0
2	6	2	173	63	0	1	2	0	2	2
3	14	3	11	1927	1	2	13	27	88	1

Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 3, No.1, April 2021, pp. 0210106-01 ~ 021016-07
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v3i1.8460

4	123	1	0	30	143	11	1	0	2	2
5	351	17	7	53	13	70	2	1	6	6
6	4	1	0	77	2	2	1300	1	4	5
7	3	0	1	193	1	0	0	30	27	0
8	10	0	6	403	1	2	3	15	213	2
9	4	1	0	33	0	1	2	0	0	366

Table 7 shows the accuracy result of the RF and GB model on the same parameter setting. From the experiments, it is concluded that RF benefitted more from additional tree depth. Whereas n_estimator and max feature only slightly improve the accuracy. It is also worth noting that GB is slower in training. For future research in multi-class classification tasks, it is recommended to use a large value on the max_depth parameter. Regarding the insight from the word usage and categorization accuracy, it is recommended to create a better categorization / sub-categorization of the complaints by grouping similar categories and splitting distinct categories. For future research, one can explore different methods such as Word2vec, Paragraph2vec, Word embeddings, and other state-of-the-art Natural Language Processing (NLP) techniques to create the vector and use other text classification approaches such as BERT, CNN, etc.

Table 7. Performance comparison of RF and GB model using the same parameter setting

Method	n_estimator	LR	Max_depth	Max features	Accuracy
RF	50	N/A	10	10	0.48
GB	50	0.1	10	10	0.69
RF	100	N/A	10	10	0.49
GB	100	0.1	10	10	0.71
RF	100	N/A	20	10	0.60
GB	100	0.1	20	10	0.72
RF	100	N/A	20	20	0.66
GB	100	0.1	20	20	0.72

4. Conclusion

Capturing and responding to complaints from the public is an important effort to develop a good city/country. This project aims to utilize Data Mining to automatize complaints categorization. More than 35,000 complaints data in Bangalore City, India, were retrieved from https://www.ichangemycity.com. The vector space of the complaints was created using TF-IDF and the multi-class text classifications were done using Random Forest and Extreme Gradient Boosting. Results showed that both RF and GB have a similar performance with an accuracy of 73%, with RF benefitted from larger max_dept whereas GB is slower in training. Result also showed that the model is highly dependent on the word usage in the complaint's description. Future research suggestions to increase the task performance include a) using bagging methods instead of boosting methods, b) creating a better categorization / sub-categorization of the complaints by grouping similar categories and splitting distinct categories. Future research may also experiment using Word2vec, Paragraph2vec, Word embeddings, and other state-of-the-art NLP techniques for the vector space creation and using other text classification approaches like BERT, CNN, etc.

References

[1] E. Zuliarso, M. T. Anwar, K. Hadiono, and I. Chasanah, "Detecting Hoaxes in Indonesian News Using TF/TDM and K Nearest Neighbor," in *IOP Conference Series: Materials Science*

and Engineering, 2020, vol. 835, no. 1, p. 12036.

[2] M. T. Anwar, H. D. Pumomo, S. Y. J. Prasetyo, and K. D. Hartomo, "Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, pp. 409–415.

[3] M. T. Anwar, S. Nugrohadi, V. Tantriyati, and V. A. Windarni, "Rain Prediction Using Rule-Based Machine Learning Approach," *Adv. Sustain. Sci. Eng. Technol.*, vol. 2, no. 1, 2020.

[4] M. T. Anwar, W. Hadikurniawati, E. Winarno, and W. Widiyatmoko, "Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2020, pp. 83–88.

[5] W. Hadikurniawati, M. T. Anwar, D. Marlina, and H. Kusumo, "Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data," in *Journal of Physics: Conference Series*, 2021, vol. 1869, no. 1, p. 12093.

[6] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[7] Z. Xia, K. Stewart, and J. Fan, "Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major us metropolitan area," *Comput. Environ. Urban Syst.*, vol. 87, p. 101599, 2021.

[8] M. Zolfaghari and M. R. Golabi, "Modeling and predicting the electricity production in hydropower using conjunction of wavelet transform, long short-term memory and random forest models," *Renew. Energy*, vol. 170, pp. 1367–1381, 2021.

[9] K. Liu, X. Hu, H. Zhou, L. Tong, D. Widanalage, and J. Macro, "Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification," *IEEE/ASME Trans. Mechatronics*, 2021.

[10] D. Sun, J. Xu, H. Wen, and D. Wang, "Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest," *Eng. Geol.*, vol. 281, p. 105972, 2021.

[11] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost: extreme gradient boosting," *R Packag. version 0.4-2*, pp. 1–4, 2015.

[12] M. T. Anwar, E. Winarno, W. Hadikurniawati, and M. Novita, "Rainfall prediction using Extreme Gradient Boosting," in *Journal of Physics: Conference Series*, 2021, vol. 1869, no. 1, p. 12078.

[13] J. Fan, J. Zheng, L. Wu, and F. Zhang, "Estimation of daily maize transpiration using support vector machines, extreme gradient boosting, artificial and deep neural networks models," *Agric. Water Manag.*, vol. 245, p. 106547, 2021.

[14] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," *Geosci. Front.*, vol. 12, no. 1, pp. 469–477, 2021.

[15] L. Cui, P. Chen, L. Wang, J. Li, and H. Ling, "Application of Extreme Gradient Boosting Based on Grey Relation Analysis for Prediction of Compressive Strength of Concrete," *Adv. Civ. Eng.*, vol. 2021, 2021.

[16] A. I. A. Osman, A. N. Ahmed, M. F. Chow, Y. F. Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Eng. J.*, 2021.

[17] P. Kumari and D. Toshniwal, "Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance," *J. Clean. Prod.*, vol. 279, p. 123285, 2021.