Civil Complaints Management System by using Machine Learning Techniques

Sayali Bhosale[1], Sonali Patankar[2], Kshitija Kadam[3], Rujuta Dhere[4], Prof. Manisha Desai[5]

[1,2,3,4](Student, Dept. of Computer Engineering, RMD Sinhgad School of Engineering, Warje, India.)
[5](Asst. Prof., Dept. of Computer Engineering, RMD Sinhgad School of Engineering, India)

Abstract - *In today's fast paced world, citizens' needs are getting increased and becoming more and more complex. But nobody has a time to wait for a long time to register their complaint manually as traditional method do. As metropolitan citizens necessitate a common platform where they can register and keep track of their complaint, we established the model that addresses this problem. The holistic approach of our model is to be effective in bring down the complaints through urging them on the basis of severity of complaint. The proposed system is elicited from the concepts of NLP, sentiment analysis and machine learning models. It integrates the different abilities that analyses the citizens' sentiment to prioritize the complaints and categorize them into respective departments. So that service executives can directly use it to identify citizen's major needs and provide them appropriate facilities.*

Key Words*: *Classification, NLP, Machine Learning, Prioritization, Sentiment Analysis

1.INTRODUCTION

Nowadays enormous number of complaints get registered at government offices which didn't even get noticed due to poor management and lack of appropriate system which will mainly focus on sorting and settling of complaints.

Large dataset gets generated by registered complaints which made it difficult to sort and to resolve them on time. Sentiment analysis helps in finding the urgency and severity of problem. As a result, priority will be assigned to most important complaints based on the polarity score of statement.

In previous years, different algorithms were used for sentiment analysis such as Lexicon Based Sentiment Analysis, Naïve bayes, Support Vector Machine (SVM) etc. Also, for clustering purpose k-means and BA (Bees Algorithm) algorithms were commonly used. As per the study, we noted that using Machine Learning techniques and data mining tools can be useful for our model. As nearest neighbor classifier generally lag the performances of SVMs and Random Forest Classifiers, we decided to use Random Forest Classifier for classification of departments.

The classification model needs to be evaluated to determine how well the model performs for the desired categorization. The accuracy value of the model will represent how the overall dataset classified correctly or not.

Proposed system provides an integrated and accessible portal for public to register the complaints and to keep track of the same. Resulting to which public participation can get increased and this can make the government active towards their duties.

2. LITERATURE SURVEY

[1] Xingming Chen, Yanghui Rao, Haoran Xie, Fu Lee Wang, Yingchao Zhao, Jian Yin published "Sentiment Classification Using Negative and Intensive Sentiment Supplement Information" in Data Science and Engineering (2019) 4:109–118

To tackle the semantic composition issue of existing deep neural networks for sentiment classification is the research's main aims. They modelled the distinct effects of negative and intensive words through a LSTM network using Negative Expression Modelling and Intensive Expression Modelling. To validate the effectiveness, models were incorporated into three deep neural networks i.e. CNN, LSTM and CharSCNN. Python libraries such as TensorFlow and keras were used along with Stochastic Gradient Descent with Adadelta for training. For all datasets the proposed NIM-CNN model performed better than these baselines. Intensive words played a more significant role in fine-grained sentiment classification tasks. The proposed model has addressed the sentiment reversing effect of negative words and the sentiment shifting effect of intensive words.

[2]Gutha Jaya Krishna,Vadlamani Ravi,Bheemidi Vikram Reddy,Mohammad Zaheeruddin,Harshal Jaiswal,P. Sai Ravi Teja,Rohit Gavval, published "Sentiment Classification of Indian Banks' Customer Complaints" in 2019 IEEE Region 10 Conference (TENCON 2019)

The proposed model in this paper uses raw textual data for pre-processing, with the use of techniques like document term matrix (DTM) driven by TF-IDF (Term Frequency - Inverse Document Frequency) , embedding model likeWord2Vec and psycho-linguistic method like Linguistic Inquiry and Word Count (LIWC). For categorization "moderate" and "extreme' are the labels given to the raw textual complaints by the three human annotators. Researchers performed three feature extraction methods on

the data and concluded that the LIWC in combination with Random Forest and Naïve Bayes techniques, gives the statistically significant performance with a pairwise t-test on results of AUC obtained from multiple classifiers.

[3] Ms. Manushree A. M, Mr.Adarsh M. J, Dr. PushpaRaviKumar published "A Comparative Method for Different Aspect Based Products Features in Online Reviews of Different Languages" in 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology, May 19-20, 2017, India.

This work put forward the idea of overcoming the most frequently occuring spell mistakes. Also gives the solution to deal with the multiple languages at a time and comparison between SentiWordNet and TextBlob shows the accuracy difference in the models while computing the data. Result was stated in terms of polarity score which can be assigned to each sentence in the review by using TextBlob. This work represses the most commonly committed spell mistakes by the reviewer in context of needed features, also a solution to deal with reviews written in other languages.

[4] "Evaluating Citizen Satisfaction and Prioritizing Their Needs Based on Citizens' Complaint Data" by Mostafa Ghodousi, Ali Asghar Alesheikh, Bahram Saeidian, Biswajeet Pradhan, Chang-Wook Lee published in Sustainability 2019, 11, 4595;doi:10.3390/su11174595 (www.mdpi.com/journal/sustainability)

To prioritize the metropolitan needs and to estimate citizens' satisfaction is the aim of the research done on a data-mining of a '137 system' (citizens' complaint system) of the region in Iran. To achieve this, the K-means clustering and Bees Algorithms(BA) were used. Each of these two algorithms was executed using different methods. In the first method, prioritization and estimation of satisfaction has been done separately, whereas in the second method, prioritization and estimation of satisfaction were done simultaneously. To compare the clustering results in these two methods, an index was presented quantitatively. The results of these methods showed the superiority of the second method. The index of the second method for the first needs in K-means was 0.299 more than the first method and it was the same in two methods in BA. Also, the results of the BA clustering were much better at it because of the S (silhouette) and CH (Calinski-Harabasz) indexes.

3. PROPOSED ARCHITECTURE

In the system architecture there are two main parts given below:

Natural Language Processing (NLP):

It is basically the interaction between human languages and computers. Through NLP, system learns to accurately manage and apply overall linguistic meaning to text excerpts like phrases or sentences. We applied Natural Language Processing (NLP) techniques to generate structured data automatically out of texts from raw dataset. Ambiguities such as ambiguous segmentation, Noisy short text, ambiguous instance will be handled in the system with help of some function in NLP and knowledgebase WordNet.

Sentiment Analysis:

Sentiment analysis is a text classification tool that can be used to analyse a piece of text and determine whether the underlying sentiment is positive, negative or neutral. It used in the system to determine the polarity of the complaint text entered by citizen. Sentiment score range is between -1.0 (highly negative) to +1.0 (highly positive). Based on the sentiment analysis text score, intensity is calculated.

3.1 Flow of the system:

1) User will enter the Complaint and Address in portal.
2) Once click on Submit complaint will be added to the database.
3) System will send the same complaint to the Machine Learning model/NLP engine.
4) Machine learning model will extract the needed information from the complaint.
5) This information will be used for classifying the complaint to its respective department and the polarity of it, which will help to decide its priority.
6) Citizen can see the Complaint ID and Department of the complaint once he submits the complaint.
7) The complaints will be arranged according to their priority. The complaint with highest priority will be displayed first.
8) Once the complaint is resolved Admin can change the status of the complaint as Close.

Figure 1 shows the flow of the proposed system.

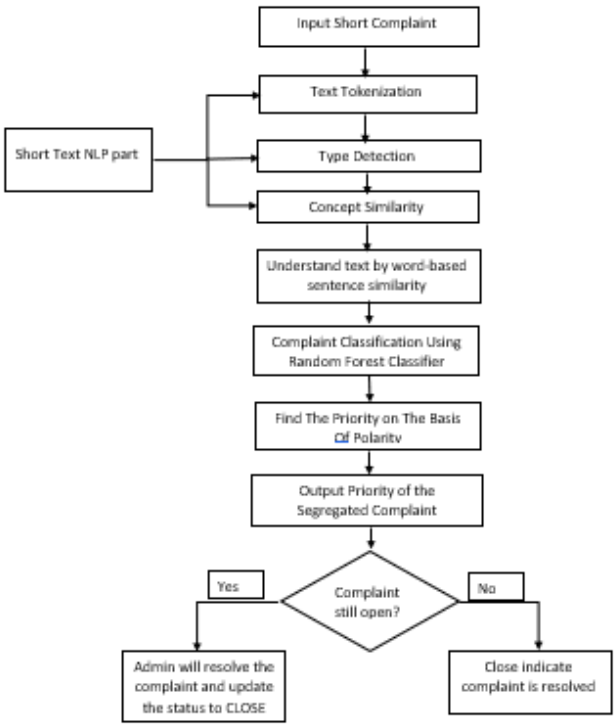

Fig -1: Flow of the System

4. ALGORITHMIC STRATEGY

1.2 Pre-Processing:

The pre-processing will help to transform the unstructured data into structured data.

a) Tokenization: Tokenization is an essential strategy for most NLP tasks. It parts a sentence or archive into tokens which are words or expressions. CountVectorizer is used in this study. It will tokenize the complaint along with performing very basic pre-processing like removing the punctuation marks, converting all the words to lowercase.
b) Tfidf Transformer: Term frequency inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in collection.
c) Stemming, is a process of taking the stem. For example, the word "using", "uses", and "used" will have the same stem as "use".

4.2 Classification

The classification process is done to classify the report data into some specified categories or classes. The purpose of this process is to build a predictive model that able to classify documents automatically to some known classes effectively and efficiently. In this study, Random Forest Classifier was used to establish a classification model. The classification process will divide the data into five classes that are: Water Supply, Electricity, PWD, Sewage and other. Random forests are methods that learn similarity in the presence of many irrelevant attributes. A random forest samples bags of attributes at nodes of a decision tree and selects the best one among them for splitting. It uses the selected attribute to recursively partition the data in the nodes, until each node contains only the instances belonging to a particular class [6]. Once the data was created it was shuffled. 75% of the data was used to train the model and remaining 25% was dataset, which was unknown to the model was used for testing purpose.

4.3 Prioritization

The polarity score varies from -1 to +1 indicating the intensity of the sentence polarity which helps to analyse the positivity and negativity of the selected complaints. For that we have used the TextBlob library.

5. RESULT AND EVALUATION

Various experiments that are done for Complaint Management system involves the And these are built with different Classification Algorithm. Following are the Classification Algorithms:

- SVM
- KNN
- Random Forest,
- Decision Tree

We have studied them and decided on Random Forest as it gives the more accuracy comparing to others. In our case it gives 75%. The accuracy value of the model will represent how the overall document classified correctly. The higher accuracy value means the model is better and more accurate in doing classification.

Also, the additional function we added is Prioritization. The complaints will be Prioritized based on the polarity of the Complaint. For which we have used TextBlob library in which there is inbuild function Polarity. So not only the complaints will be classified but prioritize as well.

Citizen can register the complaint and Complaint ID will be generated once Citizen Submit the complaint. In Fig 2 we can see that through Complaint ID citizen can check the Complaint Status if it is resolved or not. We can also see the Date and Time at which complaint has been registered.

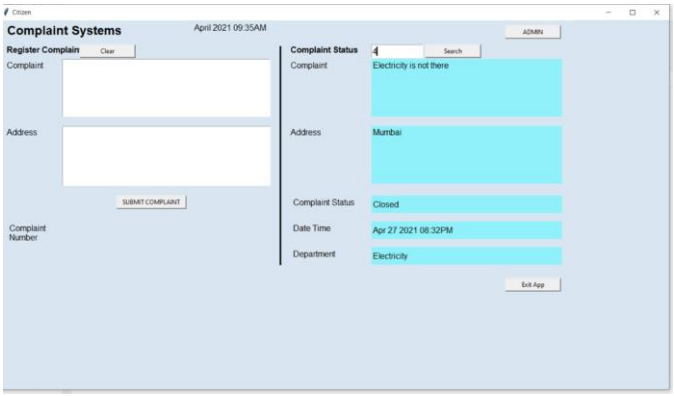

Fig -2: Citizen Screen

In the given Fig 3 we can see that Admin can select the Department and he will be able to see the complaints regarding to it. The complaint with highest priority will be displayed first. Once the complaint is resolved Admin can close the complaint.

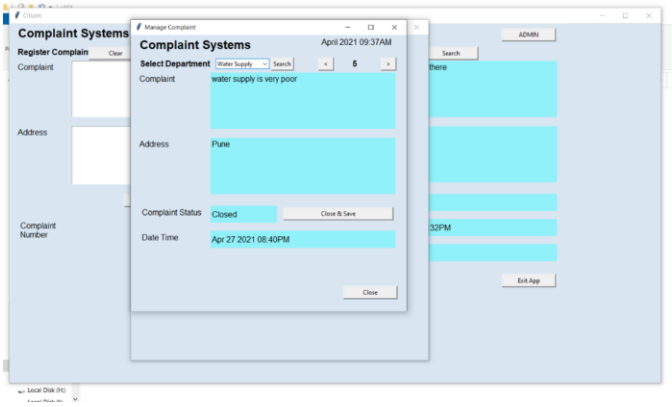

Fig -3: Admin's Screen

6. CONCLUSIONS

In this work we proposed an effective model of Complaint Management System. In which Complaint Classification and Prioritization is done. The complaints are getting classified using Random Forest Classifier algorithm into various department with accuracy 75%. The complaints will be displayed to admin in prioritized manner which is the main goal of the system. The challenging task is analysis of complex sentences and misspelled words in complaint. Another disadvantage is that the system may not consider spelling mistakes in language other than English therefore this challenge can be taken for future work

REFERENCES

[1] Xingming Chen, Yanghui Rao, Haoran Xie, Fu Lee Wang, Yingchao Zhao, Jian Yin, "Sentiment Classification Using Negative and Intensive Sentiment Supplement Information" in Data Science and Engineering (2019) 4:109–118 (Springer).

[2] Gutha Jaya Krishna,Vadlamani Ravi,Bheemidi Vikram Reddy,Mohammad Zaheeruddin,Harshal Jaiswal,P. Sai Ravi Teja,Rohit Gavval, "Sentiment Classification of Indian Banks' Customer Complaints" in 2019 IEEE Region 10 Conference (TENCON 2019)

[3] "A Comparative Method for Different Aspect Based Products Features in Online Reviews of Different Languages" , Ms. Manushree A. M, Mr.Adarsh M. J, Dr. PushpaRaviKumar in 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India.

[4] "Evaluating Citizen Satisfaction and Prioritizing Their Needs Based on Citizens' Complaint Data" by Mostafa Ghodousi, Ali Asghar Alesheikh, Bahram Saeidian, Biswajeet Pradhan, Chang-Wook Lee in Sustainability 2019, 11, 4595;doi:10.3390/su11174595 (www.mdpi.com/journal/sustainability)

[5] "Priority Based Sentiment Analysis for Quick Response to Citizen Complaints" Published by Kamalakshi V. Deshmukh, Prof. Sankirti S. Shiravale in 2018 3rd International Conference for Convergence in Technology (I2CT)

[6] "Nearest Neighbor Classifiers versus Random Forests and Support Vector Machines" Published by Saket Sathe, Charu C. Aggarwal in 2019 IEEE International Conference on Data Mining (ICDM)