Students Query Classification System

Venkataramana N[1][0000-0002-4246-3009]Nagesh P[2][0000-0001-9987-3650]
Sivanageswarao G[3][0000-0001-6422-4968] ,Prabha B[4]
[1,2,3,4]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302.
ramana@kluniversity.in

DOI : 10.48047/IJFANS/11/S6/038

Abstract. A University or educational institute generally receives a bulk of complaints posted by students every day. The issues relate to their academics or any issues relates to their education or related to exam sections etc., because of these bulk of complaints received from the students every day makes it difficult for the university to sort out them and classify them and send to their respective department for resolving the issues. In this project, we work on classifying these complaints based on the classes or departments they belong to, using. By using TF-IDF (term frequency-inverse document frequency) it finds terms which are more related to a specific document by converting to vectors. By capturing some keywords in the complaints, adding some weight to the keywords and using different Machine Learning classification's we are classifying the complaint based on these keywords. This classification makes the works easier for the university and saves time which is used to sort them and gives better service for the students. Now they can directly send the complaints to the respective departments with ease**.**

Keywords: Complaints, classification, TF-IDF(term frequency-inverse document frequency), departments, Vectors, Machine Learning.

INTRODUCTION

Nowadays, Educational Institutes are growing day by day more complaints were registered and categorizing them into respective departments making a huge task for the management and it becomes a more consuming time process. The students and Management in educational institutes are facing some problems to their academic-related issues like Fee Issues, Exam Issues, Hostel Issues, and more of like academic Issues. With respective with the Management, there is a lot and bulk of mixed Issues were received and classifying them into different categories of department's is the main objective of this project and we have designed a classification model using the TF-IDF(term frequency-inverse document frequency) to solve this problem effectively. We can solve the Complaints related to the students, facing the issues more accurately by classifying them and triggering them to the respective sub-departments and making the work easier

The Complaints raised by a person related to academic issues are received from a form through a student grievances website, and it is stored in the database table having attributes of Token No, Date, Year, StudentID, EmailId, Grievance Category, Counsellor Name, Cat, Issue Resolver Name, Issue Given Date, No of Days to Resolve, Issue Resolved Status, Final Status and within the amount of time the complaint gets resolved by sending to the sub-department through notifications and the department then checks the issues based on priority and once the issue is resolved then it sends the notification to the issue raised person, so with this, we can solve the bulk of complaints efficiently which were stored in the database and once the issue was solved it will be updated and reflected in departments. So, in this regard, we can classify the issues based on which category the issue belongs to by assigning label numbers to the departments and mapping the issue to the respective department. We can resolve the complaint raised by a person manually, but it takes more amount of time. The issues which were stored in the database within one or two days the exact solution will be given to person whether his/her issue was resolved successfully or not. The classification algorithms combining with TF-IDF(term frequency-inverse document frequency) based on Text Classification which we were using classifies the issue based on the frequency of the works in the given complaint and making vectors by converting the text and divide it into a particular domain and maps the complaint to respective department and role of classification algorithms takes place where we can categorize the data to a particular department based on the labels given pre-definitely for each respective departments and sends a notification to that department admin

The main of this project say that the person should not lose his valuable time and make the work easier by using different machine learning algorithms. So, in this way, we can design, that the person can raise a complaint and get the solution to his/her complaint easier. The web system using technologies is the easiest way to solve the complaints raised by the students in bulk amount. Hence using this developed model, the complaints can be solved easier and faster by classifying to required departments of the Organization

PROPOSED METHODOLOGY

The complaints received are in the form of text, to classify the complaints with the help of classification algorithm, the text needs to be transformed into vectors so that the algorithm will be able to predict the class. To achieve it, we use **TF-IDF** method to convert the text to vectors. TF-IDF means **term frequency-inverse document frequency** which is used to find out which terms are most relevant to a specific topic. It is a statistical metrics used to evaluate how relevant a term/word is to a document in a collection of documents or a corpus

TF-IDF of a word in a document is calculated with the help of two measures TF (term frequency) and IDF (inverse document frequency).

TF (term frequency) is calculated by finding several times a word appeared in a document and the frequency is adjusted with a length of the document or number of words in the document.

$$\text{Term Frequency (TF)} = \frac{\text{Number of times the word appeared in the document}}{\text{Length of document/number of words in document}}$$

IDF (inverse document frequency) of a word or term means how rare or low the word appears in the entire corpus or collection ofdocuments. This can be calculated by dividing the number of documents with respect to the number of documents the term appeared.

$$\text{Inverse Document Frequency (IDF)} = \frac{\text{Total number of documents}}{\text{Total number of documents the word appeared}}$$

If the word or term appears in more no of the document and very common, then it is scaled to '0' else if it is scaled to '1'.

Multiplying the two terms with each other we can obtain the TD-IDF score. Higher the score higher the relevancy of the word with respect to the document.

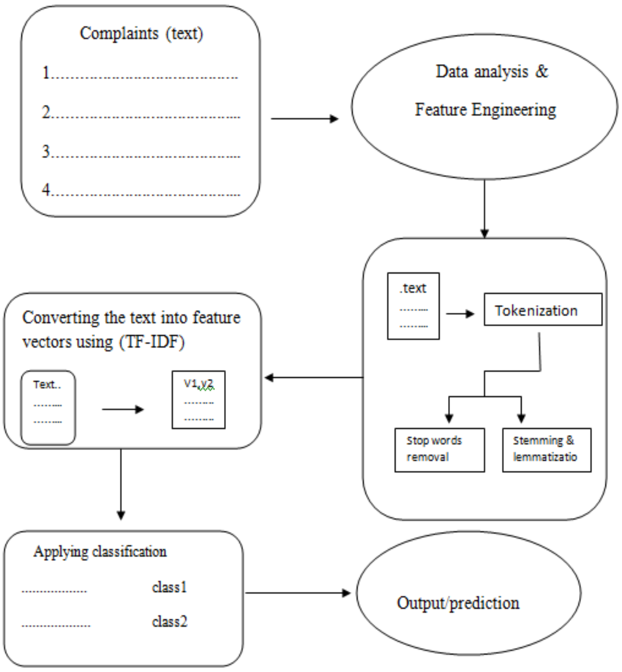
After converting the text, we apply the different classification algorithms like Random Forest Classifier, Linear SVC, Multinomial NB and Logistic Regression

The Dataset "complaints.csv" will be having the attributes as Token No, Date, Year, Student id , Email Id, Grievance Category, Counselor Name, Cat, Issue Resolver Name, Issue Given Date, No of Days to Resolve, Issue Resolved Status, Final Status. By using this "complaints.csv" dataset we will make another copy of Data Frame consisting of Cat (Categories like health issues, examination section, detention, etc.) and Grievance Category consisting of the detailed complaint. Now we will remove the duplicates in the newly created Data Frame let's say "df1" and assign unique Id for each category in other works making a temporary dictionary for future reference. Now we can also know that which section or department is having a greater number of complaints raised by students.

Now, we will be applying the TfidVectorizer which will transform each complaint into a vector, and we will be storing the vectors in an array and we can get the score of Unigrams and Bigrams. After we will map the vectors with most correlated Unigrams and Bigrams for each complaint by removing the stop words. The splitting of data for Training and Testing will take place like 'X' which is having all the Grievance Category and 'y' which is consisting of the target labels we need to predict

By this step, everything will be sorted out with training and testing the data. Now we applydifferent machine learning classification algorithms and predict the output for the given complaints. Now the other part of the project is maintaining the database for sending the notification in bidirectional regarding the complaints. For that when the classification process is completed the predicted output will be taken and based on that prediction, we will trigger the notification for that department employee who will be resolving the complaint. Finally, when the complaint is resolved and once updated on the website the resolved notification will be triggered back to the issue raiser and work will be completed easily without any wasting of time and it will be best when compared to all complaint classifier as it is a one-to-one interaction

BLOCK DIAGRAM



.We can see by the above block Diagram how the process for the classification takes place in the algorithm. The first block specifies the complaints received by the students from a college institution and storing them as Data setfor further processes. The second one specifies the processes of Data analysis such as what type of data is provided and how it should be processed for the next stage and feature engineering.

Then, the third block contains how the provided text is undergoing the nlp methods like tokenization and then removing of stop words and applying the stemming and lemmatization processes. The fourth block specifies the process of converting the text into feature vectors using the TF-IDF. Then the fifth block specifies the classification processes such as Logistic, Linear SVC, etc. Finally, the output will be predicted for the given text which is a complaint by classifying

ALGORITHM

Input:
D: *complaints data (consists of all the complaints)*

Output:
Weight Matrix (which consists of all the weights of terms are called vectors)
Procedure:

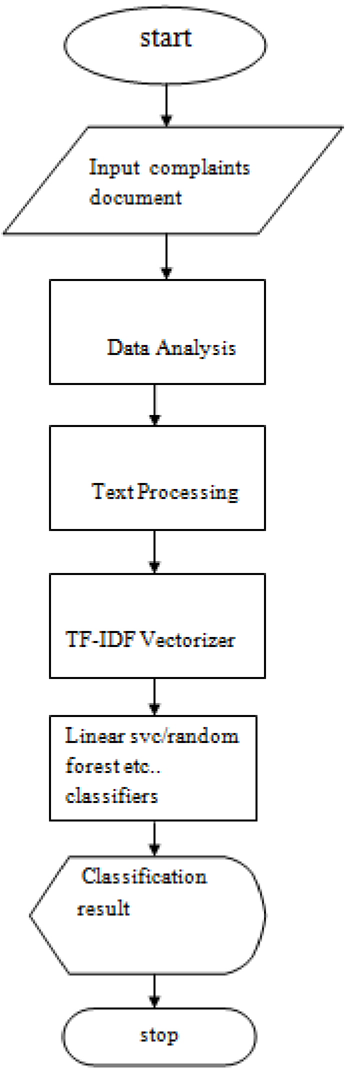
```
1- for each complaint document ($c_i$) do
2-      for each term ($t_j$) in $c_i$ do
3-           TF-IDF score for term $t_j$ in document
               $c_i = TF(c_i, t_j) * IDF(t_j)$
Where, IDF = Inverse Document Frequency
           TF = Term Frequency
TF ($c_i$, $t_j$) = (Term $t_j$ frequency in document $c_i$)
                    ________________________________
                                        (Total words in document $c_i$)

 IDF ($c_i$) = $_{\log 2}$( (Total Documents) / (documents
               with term $t_j$ )  )
4-     End for of term
```

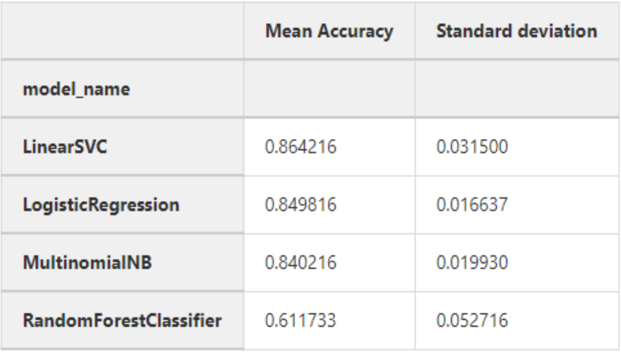
5- *End for of complaint document*

6- *The vectors are stored in an array for training and testing purposes, during classification.*

FLOW CHART

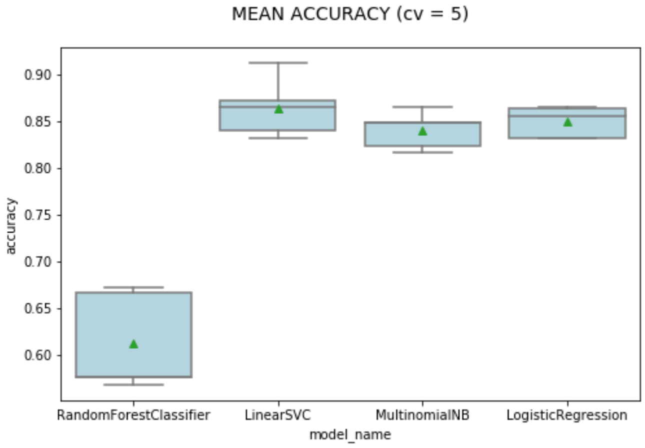


RESULT ANALYSIS



	Mean Accuracy	Standard deviation
model_name		
LinearSVC	0.864216	0.031500
LogisticRegression	0.849816	0.016637
MultinomialNB	0.840216	0.019930
RandomForestClassifier	0.611733	0.052716

By applying the different classification algorithms like Random Forest Classifier, Linear SVC, Multinomial NB and Logistic Regression for getting different predictions for each classification algorithm and produced the mean accuracy of 0.61, 0.86, 0.84, 0.84 and standard deviation of 0.052, 0.031, 0.019, 0.016 respectively for the classification algorithms. After cross-validation is processed for evaluating the accuracies and storing them in a separate Data Frame for further references. By our work, we have chosen Linear SVC as the classification algorithm which is producing more accurate results for the classification process and we are getting an accuracy of 89% for our Dataset.



CONCLUSION

A complete Classification model of complaint classification is created which make every complaint system to work efficiently for maintaining of huge and bulk amounts of complaints and maintaining the one-to-one interaction with the issue raiser and resolver for a University. More productivity of work and consuming more time will vanish and helps in the growth of Institutions. An Effective Complaint Classification System using this machine learning classification's approach combining with TF-IDF (**term frequency-inverse document frequency)** resulting with an accuracy of 89% and Maintaining of Database in an easier way.

References

1. .Ana Catarina Forte and Pavel B. Brazdil. 2016. Determining the Level of Clients' Dissatisfaction from Their Commentaries. In Computational Processing of the Portuguese Language - 12th Int. Conf., PROPOR 2016, volume 9727 of Lecture Notes in Computer Science, pages 74–85. Springer
2. N. S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3):175–185
3. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings
4. .Mrs Sujata Khedkar a , Dr. Subhash Shinde:Deep Learning and Ensemble Approach for Praise or Complaint Classification,sh Shinde, Professor, Computer Engineering Department, LTCE,Koparkhairane, Navi Mumbai, 400050,India,Dr. Subhash Shinde, Professor, Computer Engineering Department, LTCE,Koparkhairane, Navi Mumbai, 400709,IndiaM.A.Fauzi,Automatic complaint classification system using classifier ensembles,January2018
5. Ganesan, Kavita, and Guangyu Zhou. (2016), "Linguistic Understanding of Complaints and Praises in User Reviews." , Proceedings of NAACLHLT.
6. Imam Cholissodin, Maya Kurniawati, Indriati, Issa Arwani Informatics Department, PTIIK, Brawijiaya University, Malang, Indonesia.Classification of Campus E-Complaint Documents using Directed Acyclic Graph Multi-Class SVM Based on Analytic Hierarchy Process 2014
7. Moschitti, A., & Basili, R. (2004). , "Complex Linguistic Features for Text Classification: A Comprehensive Study.", Advances in Information Retrieval, 181–196]Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). "Deep Learning for Hate Speech Detection in Tweets",. Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17
8. Ryan M. Eshleman and Hui Yang. 2014. "Hey #311, Come Clean My Street!": A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints. In 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, pages 477– 484.
9. Ahmad Fauzan and Masayu LeyliaKhodra. 2014. Automatic Multilabel Categorization using Learning to Rank Framework for Complaint Text on Bandung Government. In 2014 Int. Conf. of Advanced Informatics: Concept, Theory and Application (ICAICTA), pages 28–33. InstitutTeknologi Bandung, IEEE.

10. P. Kishore, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, July 2002.
11. Navya Krishna ,Rajarajeswari P et al “ Recognition of Fake Currency Note using Convolutional Neural Networks “ International Journal Of Innovative Technology And Exploring Engineering (IJITEE) , 2019 , Volume-8 Issue-5, pp: 2278-3075
12. S. Rizwana, S.SagarImambi “Enhanced biomedical data modeling using unsupervised probabilistic machine learning technique” International Journal of Recent Technology and Engineering,,2019, vol7.No 6. Pp 579-582
13. Banerjee D, S.SagarImambi. (2019)“ Opinion mining for drug reviews “ International Journal of Innovative Technology and Exploring Engineering. Vol 8 No.7, pp. 1314- pp.13-18