NLP based Model for Classification of Complaints: Autonomous and Intelligent System

Qurat-ul-ain
National University of Sciences and Technology (NUST), Islamabad, Pakistan
qibrahim.ce19ceme@ce.eme.edu.pk

Arslan Shaukat
National University of Sciences and Technology (NUST), Islamabad, Pakistan
arslanshaukat@ceme.nust.edu.pk

Usman Saif
University of Science and Technology Beijing(USTB),Beijing, China
usmansaif531@gmail.com

2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2) | 978-1-6654-9819-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICoDT255437.2022.9787456

***Abstract*—Artificial intelligence nowadays is playing a vital role in our society. It is just minimizing human labor and effort in every field. Industrial sector is feeding their large amount of structured and unstructured data to find out useful information for scientific research. The main alarming thing is how to operate the huge feedback data, which is in the form of complaints i.e., in text format. Here, we have proposed a model which automatically classifies the complaints by analyzing the text with the help of machine learning and NLP (Natural Language Processing) methods. We have initially collected a dataset from a portal containing complaints of citizens. For validation, we have also used another dataset of complaints from the Consumer Complaint Database. After tokenizing, stemming and lemmatization, different feature extraction techniques like count vectorizer and TF-IDF are used to convert all the textual data into numerical data. Then different machine learning algorithms are used to classify the complaints into their categories. In our gathered dataset, 10 different divisions for complaints are used and an accuracy of more than 70% is achieved with all classifiers. Similarly on the Consumer Complaint dataset, 86% accuracy has been achieved. The proposed model is helpful in saving a lot of time, as there is no need to go through each complaint and categorizing manually.**

Keywords—Machine Learning, NLP, Text Classification, KNN, SVM

I. INTRODUCTION

Nowadays, Artificial intelligence (AI) systems are very common and reliable, being increasingly used in different organizations. Different kinds of AI techniques are being employed in different industrial sectors to assist humans and increase work efficiency. One of the AI techniques being used is Natural Language Processing (NLP). NLP handles text that is available in a soft form such as on websites or any available local network where textual data in large amounts can be provided by humans to the system.

To give reliability to citizens, multiple government organizations are providing public services virtually. Those organizations are also answerable to the public. The citizen feedback is developed to make the system more reliable and fix the problem, often these feedbacks (complaints or requests) come to some online portal, in an online form, or via emails. This became a huge and time taking problem to deal with if the support staff handling the feedback forms is limited. The feedback is in massive size as it includes feedback from different citizens belonging to different regions. As the feedback is in the written format like email so NLP techniques are used to classify different complaints based on content.

This research emphasizes the requirements of a government organization, a national administrative authority that focuses on education such as scholarships, accreditation of institutes, and deals with multiple councils. Likewise, it works towards the quality of education in the institutions of Pakistan. This organization also entertains complaints from different citizens regarding education, faculty, scholarships, and much more. It receives around 2-5 thousand complaints on the daily basis. And more than 50 % of these complaints are not even related to the organization. For that, they have to create a specific section to deal with these complaints and assured the quality of all the instituted that are coming under the umbrella of the organization. This becomes a bottleneck issue as it requires more human labor and efforts to analyze the complaints and their categories and which department those complaints pertain to. There are huge chances of human error due to less training of new hiring staff as well. For that doing the research more effectively to get higher chances of accurate results when this system assembles in daily routine, the content of the citizen complaints is taken as it is.

In this paper, we collect and provide a rich data set of 10,000 complaints related to education that pertains to different departments. These complaints are in textual format. After applying NLP techniques and by using Machine-learning algorithms, thus we classify those complaints accurately. To the best of my knowledge, this is the first-time study of this kind regarding the classification of complaints based on provided complainer's textual information in the complaints. This work contributes to discovering and comparing several machine learning algorithms and their performance on the local dataset and on the Consumer Complaints dataset. This work also differs in the comparison approach from the previous works of literature by using multiple and effective approaches to different ML models. The current work offers new insight into how the proposed architectures perform, specifically in classifying the Local Complaints dataset and Consumer Complaints dataset.

Section II presents the brief literature review. Section III is about the proposed methodology. Section IV reports the experimentation and results. Section V presents the conclusion and future work.

II. LITERATURE REVIEW

NLP establishes a center interest in the field of artificial intelligence and software engineering. NLP studies include speculations and strategies that empower viable correspondence among people and computers in regular language. Various researchers have defined NLP as a place of study and the software that explores how computer systems may be used to apprehend and control normal

language text or then again discourse to do helpful things.

Kulkarni et al. [1] Worked on the chat bot's communication in their work. Chatbots are intelligent systems more like virtual assistants. The main goal of the chatbot is that the customer can communicate their inquiries in English and thus inappropriate time the Chabot resolves all queries. For implementation, the system comprises preprocessing, vectorization (BOG), and classification. The disadvantage is that this article did not tell which classifier is more costly in terms of time and space.

Tutika, A. and Nagesh, M.Y.V. [2] Explain the pre-processing of textual information that removes the punctuation marks from text and converts them into a smaller case that gives 100% accuracy. They used NLP libraries to tokenize the text. The radiology report of chest x-rays was used to detect the presence of endotracheal intubation or opacities [3] on the chest by utilizing techniques of natural language processing and machine learning. The NLP algorithms were used, and on preprocessed data, the vectorization and tokenization are being done. The paper gives accuracy and error percentages in the results. The NLP/ML model's performance is quite good on the supervised classification that will take much larger/bigger datasets. The drawback is that with the increase of positive findings, the algorithm of NLP or MP learns itself, but this research is limited by the frequency of positive findings and size of the sample on which the model is trained, although in [4] Count vectorization and Term Frequency-Inverse Document Frequency (TF-IDF) is utilized direct implementation of ML models for more clarity in pre-processing, and then classification is applied. This paper is handy in terms of experimentation and result as it gives precision and also accuracy. This [5] article is handy, it utilizes Neural Network as a classifier and generates the threshold itself. The data is in text format and, for that this article uses the different techniques of NLP, in addition, it uses Word Order Vectorization to improve results. In this paper [8] the author uses two different types of datasets the article level & the sentence level. The dataset is first preprocessed i.e. clean the dataset and check the label that is coming with text is in the correct ratio or not. For feature extraction, they use text vectorization (Count Vectorizer from Skilit-library). The next step in feature extraction is TF-IDF. It shows the importance of the word and the count of the word in that specific document. The next step is a classification for that they use a logistic regression model. The drawback is they only use one classifier.

Razno [6] tries to find the best NLP platform and machine learning process. In this article, the author comes up with different strategies to overcome the issue of poor response through NLP Chatbot. This article is handy as it mentioned how to make an improved and error-free chatbot. They used different pre-processing steps like part of speech (POS). However, he didn't explain any classifier. Whereas in [7], the author uses a neural network model. This paper is handy as it explains how to make a fine-grained dataset, design the Generative Explanation Framework (GEF) which can adopt different models also use minimum risk training method on the proposed framework. The drawback of this paper is that the process is a little bit longer and it didn't give that much good accuracy as compared to other articles.

In this paper [9] author uses analysis of the sentiments from a different platform. That's why they applied various preprocessing steps and check the output. They use logistic regression as a classifier with each different pre-processing step. Here the author uses another approach to use different library TPOT which helps in various steps in the algorithm. The paper is handy as it gives the result of the classifier with each pre-processing as well as gives time estimation of the completion of the task.

In this article [10] different approach is used as they used various vectorization techniques to cover the whole article. KNN, Naïve Bayes & SVM is used as a classifier to obtain accuracy around 91%, 76%, and 81%. This article is handy as they also give the comparison between classifiers

Dien, T.T., Thanh-Hai, N., and Thai-Nghe, N. [11] utilize different machine learning approaches i.e., supervised and unsupervised learning in text classification. The author took multiple datasets from various repositories and perform the different techniques in pre-processing the in-Feature extraction they used the TF-IDF technique and then applied the classifier. They also used MLP containing one hidden layer and 16 neurons in the model. The good point of this paper is that the dataset is taken in a large amount and also have a comparison of the different dataset with each other. In this paper [12] the author presents their finding regarding the classification of complaints through different approaches as well as perform error analysis. The author uses multiple classifiers to check the difference in accuracy rate. Initially, they do Pre-processing and applied TF-IDF for feature extraction then applied classifiers. Here the author uses another technique LSTM as a deep learning algorithm and checks the confusion matrix to perform the error analysis on multiple classes.

Since the information collection process was based only on the identified papers, there were gaps in the gathered data about numerous frameworks. The variety of techniques with a combination of classifiers is different in their dataset. There are different results based on their chosen dataset. Main research gap is that there is minimal work done on the complaints' classification. Thus, it is more challenging to develop a generalized system of complaints classification.

III. PROPOSED METHODOLOGY

NLP is a subset of artificial intelligence, machine learning which gives the chance to humans to communicate with computers when a large amount of data of natural language is given to it. Speech recognition, natural language understanding, and generation are some features of natural language.

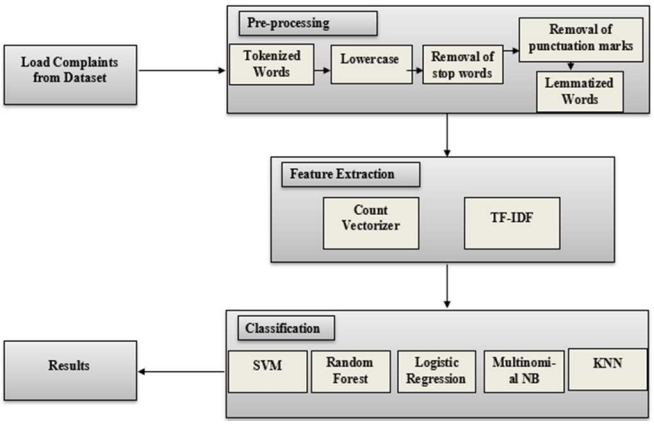

Fig. 1. A framework of the proposed architecture

In supervised machine learning, text classification is the basic problem. In this paper, we will explain how to apply different techniques on multiple classifiers along with two different datasets for correct analysis on them.

The flowchart of our purposed architecture is given in Fig. 1. The architecture consists of preprocessing, feature extraction, and a classification stage. Details of all these components are mentioned next.

A. Data Collection

We use two different data sets in our system. There are three sub-branches of a dataset 1) collect data,2) create data, 3) pre-processing.

Collect data: process to extract the datasets that are publicly available.

Create data: process to create datasets or dataset samples for particular designs which are not available publicly.

Semi Pre-processing: this is the initial process to compile the data in a way that our algorithm can work on it with a good understanding of consistent data.

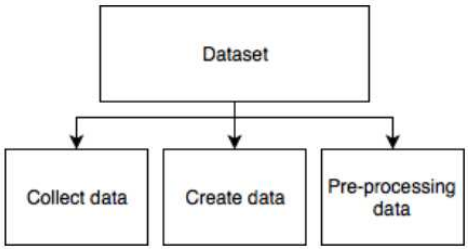

Fig. 2 Problem tree of the data set branch

1) Local complaints dataset: Dataset 1, we have 10 different divisions/classes such as scholarship, information and technology, attestation, coordination, etc. each having a different domain to cater to. We took 10,000 closed complaints from these departments that were saved on the organization's website. Some of these complaints overlap each other. As the complaint relates to two different departments. There is also a duplicate complaint in each department which is removed in initial semi-preprocessing. There are different numbers of complaints (samples). Each department has a different number of complaints i.e., class I contains 340 complaints, on the other hand, class II has 1051 complaints so, the samples size in each department is imbalanced. These departments act as a class/ label in this algorithm whereas the complaints of these departments become the data sample of those classes. Fig 3(a) demonstrates the distribution of data samples in each class.

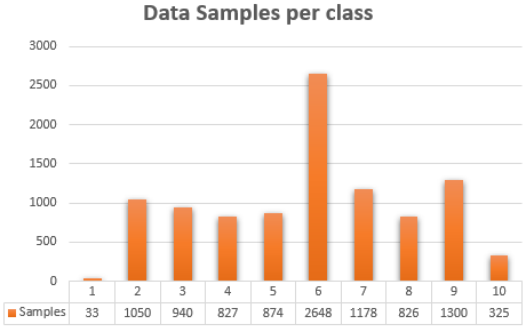

Fig. 3(a) Numbers of data samples per class in the dataset

2) Consumer Complaint Dataset II: Dataset II is an international dataset that is available on Kaggle as well as on Github. The dataset contains 1,62,421 complaints. Initially, we manually removed the duplicate complaints from the data set for cleaning our data. After removing the duplicate complaints, we got 1,24,472 complaints as cleaned data. This dataset sample is related to banking problems i.e credit card, debit card, credit card reporting debit card reporting, and mortgage issues. These all are separate departments and in this algorithm, these departments act as classes or labels containing different data samples. The distribution of data samples is given in fig3(b).

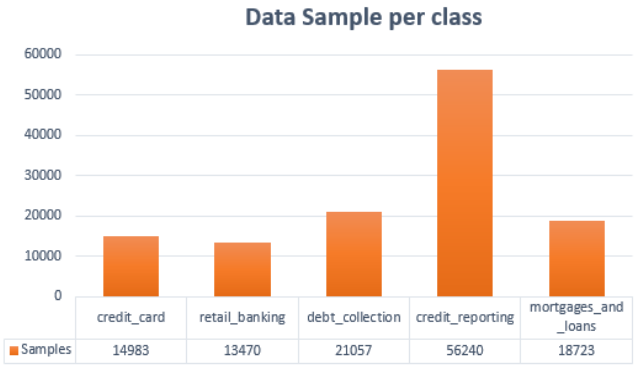

Fig. 3(b) Numbers of data samples per class in dataset II

B. Pre-processing

In the first step, preprocessing is done for reliable feature extraction. In the pre-processing stage, we perform different kinds of operations to clean our data. Initially, the data is cleaned manually. We remove all the roman English or Urdu language complaints from our data manually. The dataset was collected making sure that it has complaints written in English language only.

Following steps for pre-processing are implemented.

1) Tokenization: To split a phrase, sentence, paragraph, or full-text document into separate or single entities (words or terms) we used tokenization. These smaller entities/terms are commonly known as tokens. There are two types of tokenization: Word tokenization and Sentence tokenization.

- *Word Tokenization:* It is used to convert a sentence into several words.
- *Sentence Tokenization:* It is used to convert a paragraph into sentences.

2) Conversion into lower case: After tokenization, we converted our data from the upper case into the lower case. So, in feature extraction, two same words written in upper and lower case are marked as a single entity.

3) Removal of Stop Words: The next step is to remove stop words from our paragraph, those words are usually not needed for further analysis. These were usually common or basic English words.

4) Removal of Puncuatipon Marks: After that, we remove punctuation marks from our text. If we do not remove them, then they are considered as a separate entity and become extra work for the model by increasing computations.

5) Lemmatization: For good results and analysis, we use lemmatization, which uses vocabulary and the morphological

meaning of words. It only gives the base dictionary of a word and removes the inflectional ending; hence it is known as the lemma. For example, the word 'see' can be changed to see or saw depending on the morphological meaning by using lemmatization. It can be seen in the following Table I.

TABLE I. LEMMATIZATION OF WORD

Form	Conversion into Lemma	
	Morphological Information	***Lemma***
Studies	Third-person, singular number, present tense of the verb study	study
Studying	Gerund of the verb study	study

C. Feature Extraction

After pre-processing, our data is still in text format, but we need to convert it to number format. So we need feature extraction methods that convert our pre-processed tokens into numerical format. That is known as vectorization in NLP. Usually, each token is known as a gram. Some models use pairs of tokens, that are known as bigrams. Our technique uses n-gram as there are multiple words in each sentence. Here we have used two different methods to use vectorization such as count vectorizer and tf-idf. Their details are mentioned next. Here we use just the first two methods.

1) Count Vectorization CV: It is used to convert raw text into a numerical vector representation or token count of n-grams. It also gives the high feature representation of text into vector form. This technique is simple and easy as it directly converts them into numeric vectors. In Machine learning such features are used that are easy to utilize and help in classification. Table II demonstrates the model of the count vector.

TABLE II. MODEL OF COUNT VECTOR

Sr. No.	Word Count						
	W1	***W2***	***W3***	***W4***	***W5***	***W6***	**…**
Document 1	1	1	2	0	1	2	…
Document 2	2	0	2	1	2	3	…
Document 3	1	0	1	2	2	0	…
Document 4	2	1	0	3	2	4	…

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop.

2) TF-IDF: Term Frequency-Inverse Document Frequency gives the importance of each word/token in the given document. If the number of the same word appears more than once, then the importance of that word increases but is offset by the frequency of the word in the complaint. Term frequency (TF) shows how many times the same token appears, whereas inverse document frequency (IDF) downscales words that appear a lot across documents [2]. Table III shows the model of the TF-IDF Vectorizer.

$$TF(Word) = \text{No. of times same tokens in a doc./ Total no. of tokens in an in a doc.} \quad (1)$$

$$IDF(Word) = \text{Total no. of doc. in a dataset/ No. of doc. with the same token in it.} \quad (2)$$

TABLE III. MODEL OF TF-IDF VECTORIZER

Sr. No.	Word Count						
	W1	***W2***	***W3***	***W4***	***W5***	***W6***	**…**
Document 1	0.9	0.4	0.1	0.1	0.6	0.8	…
Document 2	0.2	0.2	0.1	0.4	0.4	0.4	…
Document 3	0.1	0.1	0.4	0.9	0.8	0.1	…
Document 4	0.2	0.1	0.9	0.4	0.4	0.1	…

D. Classification

After applying the vectorization models, the whole data is converted into numeric form. So now we can apply machine learning algorithms. We divide our samples into training and testing sets. We use classifiers to check the accuracy of our system, which is calculated as follows.

$$Accuracy = \text{True positives} + \text{True negatives/ Count of total samples} \quad (3)$$

Here, when an actual label is true and the model says it's true then it returns the true positive value (TP), whereas when the actual label is false and models say it is false then it returns the true negative value (TN). The details of the classifiers that we have used are mentioned next.

1) K Nearest Neighbor (KNN): It is a supervised ML model [13]. It is used for the classification of samples with simplicity and effectiveness. It works on the assumption that "similar things are nearer to each other". Euclidean distance is used to calculate the distance of the test sample with its.

2) Logistic Regression: It algorithm uses regression, on samples to classify them in particular classes. [14]. This model works on numerical as well as on categorical data. It utilizes the probability model. It sets the threshold so that classification can be done. If the value is less than 0.5, it places the samples into one class otherwise into another class. By default, logistic regression cannot be used for classification tasks that have more than two class labels, so-called multi-class classification.

By default, logistic regression cannot be used for classification tasks that have more than two class labels, so-called multi-class classification.

Rather, it requires remodeling to support multi-class classification problems. Multinomial Logistic Regression is a Modified version of logistic regression that predicts a multinomial probability (i.e. further than two classes) for each input sample [15].

3) Support Vector Machine (SVM): SVM is used for classification and regression [16]. It makes the hyperplane that differentiates the samples into separate classes.

Usually, SVM is used to classify two classes Multiple classes are also classified in a multi-class model of SVM but these hyperplanes are not simply due to the multi-class model[17].

4) Random Forest: This ML model uses different decision trees [18]. to build separate trees towards an attempt

to form an uncorrelated forest of trees. This gives a prediction that is more accurate than that of any separate tree.

5) Multinomial Naïve Bayes: This is also an ML model that works on the Bayes theorem for text [19]. It assumes individuality between predictors. Simply this model considers a Naive Bayes classifier that assumes the presence of a particular feature in a class that is not related to the presence of any other feature.

IV. Experiments And Results

We are using two different datasets for our generic model. The first dataset is a local dataset which we have collected whereas, the other dataset is taken from Consumer Financial Protection Bureau. We have divided the given datasets into training and testing where we have 80% of samples in the training set and 20% samples in the testing set. We have also used 10 cross-validations, where our full dataset is divided into 10 folds. These 10 folds are used as a testing set in each of the iterations.

A. Results on Local Complaints Dataset

We implement our several machine learning models on the first dataset and obtain different corresponding results. Table IV shows a comparison between different classifiers for a dataset I. Likewise, Table V compares the classifier's accuracy on the datasets I but for cross-validation. In our algorithm, we also use two different feature extraction techniques on those machine learning algorithms.

TABLE IV. Accuracy Of Classifiers

Sr. No.	Classifiers	Accuracies from different techniques	
		Count Vector	***TF-IDF***
1.	Random forest	76%	76%
2.	Multinomial Naive Bayes	76%	74%
3.	Support Vector Machines	78%	80%
4.	Logistic Regression	79%	79%
5.	K-Nearest Neighbor classifier	72%	76%

TABLE V. Cross-Validation Accuracy Of Classifiers

Sr. No.	Classifiers	Accuracies from different techniques	
		Count Vector	***TF-IDF***
1.	Random forest	74%	73%
2.	Support Vector Machines	75%	76%
3.	Logistic Regression	76%	77%
4.	K-Nearest Neighbor classifier	69%	73%

B. Results on Consumer Complaints Dataset

On the second data set, we implement the same algorithm for testing as well as training. Table VI shows a comparison between different classifiers for dataset II whereas, Table VII compares the classifier's cross-validation accuracy on dataset II.

TABLE VI. Accuracy Of Classifiers

Sr. No.	Classifiers	Accuracies from different techniques	
		Count Vector	***TF-IDF***
1.	Random forest	84%	84%
2.	Multinomial Naive Bayes	80%	81%
3.	Support Vector Machines	84%	85%
4.	Logistic Regression	85%	86%
5.	K-Nearest Neighbor classifier	80%	85%

TABLE VII. Cross-Validation Accuracy Of Classifiers

Sr. No.	Classifiers	Accuracies from different techniques	
		Count Vector	***TF-IDF***
1.	Random forest	82%	82%
2.	Support Vector Machines	82%	83%
3.	Logistic Regression	83%	84%
4.	K-Nearest Neighbor classifier	78%	78%

Due to different accuracies obtained from all these ML model Due to different accuracies obtained from all these ML models, we choose Support Vector Machine (SVM) for dataset I. We checked the confusion matrix of accuracy by using both vectorization techniques. Table VIII shows the confusion matrix of Support Vector Machine (SVM) obtained by the TF-IDF technique. we already divide our 10,000 data samples into testing and training parts thus, our testing is on 20% so our confusion matrix is built on 2000 data samples.

TABLE VIII. Confusion matrix of the Support Vector Machine (SVM) using TF-IDF on Dataset I

Class	Corresponding Class									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
1	**5**	1	0	0	0	0	0	0	0	0
2	0	**169**	15	3	8	2	6	1	3	0
3	0	4	**114**	11	3	4	29	9	2	0
4	0	0	6	**123**	18	5	0	1	8	1
5	0	4	0	16	**145**	5	2	4	7	1
6	0	5	2	22	6	**469**	20	3	11	2
7	0	7	6	6	5	36	**175**	1	2	1
8	0	4	16	5	5	3	2	**115**	6	0
9	0	4	5	29	10	11	1	3	**195**	5
10	0	0	0	3	0	4	1	1	3	**56**

Text Multiple issues have been observed while inspecting the misclassified instances. Some of the complaints are compromised on the short text, not providing full information to classify the class. Some of the feed used other languages like Urdu. Some complaints have shown semantic overlap. For example, class III (IT division) is overlapped by class II (HRD division). In all of the above cases, the complaints have been misclassified. The same misconception appears when a short complaint text is written, only including some points. Thus, in the dataset, we identify that some of the complaints had been misclassified by the human operator.

In Consumer Complaint dataset II, the training recall accuracy of our system with the Multinomial Naïve Bayes and Count vectorizer technique is 86.5% whereas, in the article [20] they achieved an accuracy of 86% as training recall. But this accuracy was obtained without removing duplicated samples/complaints. Table IX shows the confusion matrix of Support Vector Machine (SVM) obtained by the TF-IDF technique on Data set-II. The confusion matrix is built on 24895 testing samples.

TABLE IX. CONFUSION MATRIX OF THE SUPPORT VECTOR MACHINE (SVM) USING TF-IDF ON DATASET II

Class	Corresponding Class				
	1	*2*	*3*	*4*	*5*
1	**2349**	289	69	60	250
2	65	**10307**	91	46	11
3	114	823	**3082**	173	44
4	40	344	118	**3164**	77
5	266	176	34	66	**2113**

V. CONCLUSION

In this paper, we have proposed a system to automatically classify the complaints belonging to various departments, written in the English language. In our system, we have used multiple classifiers, but classifiers such as logistic regression and SVM using TF-IDF techniques generate better results. This system can be used in any organization in classifying and distributing the large data of complaints to their designated centers. In the future, we can apply different feature selection methods giving us the best features to use with the classification of complaints. Additionally, we can implement different deep learning models that would give better results. We can also include more than 10 classes in our dataset, increase the data samples, and with a better machine-learning algorithm get more accurate results.

REFERENCES

[1] Kulkarni, C.S., Bhavsar, A.U., Pingale, S.R. and Kumbhar, S.S, "BANK CHAT BOT – An Intelligent Assistant System Using NLP and Machine Learning," International Research Journal of Engineering and Technology, vol. 04 , no. 05 , May -2017.

[2] Tutika, A. and Nagesh, M.Y.V., "Restaurant reviews classification using NLP Techniques," Journal of Information and Computational Science, vol. 9, no. 11, 2019.

[3] Towfighi, S., Agarwal, A., Mak, D.Y. and Verma, A., "Labelling chest x-ray reports using an open-source NLP and ML tool for text data binary classification," medRxiv, November 22, 2019.

[4] Vijayaraghavan, S., Wang, Y., Guo, Z., Voong, J., Xu, W., Nasseri, A., Cai, J., Li, L., Vuong, K. and Wadhwa, E.,"Fake news detection with different models.," arXiv, 2020..

[5] Thompson, J., Hu, J., Mudaranthakam, D.P., Streeter, D., Neums, L., Park, M., Koestler, D.C., Gajewski, B., Jensen, R. and Mayo, M.S.,"Relevant Word Order Vectorization for Improved Natural Language Processing in Electronic Health Records," Scientific Reports, vol. 9, no. 1, pp. 1-9, 2019.

[6] Razno, M., "Machine learning text classification model with NLP approach," Computational Linguistics and Intelligent Systems, vol. 2, pp. 71-73, 2019.

[7] Liu, H., Yin, Q. and Wang, W.Y., "Towards explainable NLP: A generative explanation framework for text classification.," arXiv, 2018.

[8] Oliinyk, V.A., Vysotska, V., Burov, Y., Mykich, K. and Basto-Fernandes, V, "Propaganda detection in text data based on NLP and machine learning," In CEUR Workshop Proceedings, vol. 2631, pp. 132-144, 2020.

[9] Polyakov, E.V., Voskov, L.S., Abramov, P.S. and Polyakov, S.V., "Generalized approach to sentiment analysis of short text messages in natural language processing," Информационно-управляющие системы, no. 1, pp. 2-14, 2020. HGHVGH

[10] Dien, T.T., Loc, B.H. and Thai-Nghe, N., "Article classification using natural language processing and machine learning," in International Conference on Advanced Computing and Applications (ACOMP) , 2019, November.

[11] Dien, T.T., Thanh-Hai, N. and Thai-Nghe, N, "Deep Learning Approach for Automatic Topic Classification in an Online Submission System".

[12] Barbosa, L., Filgueiras, J., Rocha, G., Cardoso, H.L., Reis, L.P., Machado, J.P., Caldeira, A.C. and Oliveira, A.M., "Automatic Identification of Economic Activities in Complaints," in In International Conference on Statistical Language and Speech Processing, 2019, October. HJHK

[13] Brownlee, "Multinomial Logistic Regression With Python", Machine Learning Mastery, 2022. [Online]. Available: https://machinelearningmastery.com/multinomial-logistic-regression-with-python/. [Accessed: 06- Dec- 2021]

[14] T. Pranckevičius and V. Marcinkevičius, "Application of Logistic Regression with part-of-the-speech tagging for multi-class text classification," 2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), 2016, pp. 1-5, doi: 10.1109/AIEEE.2016.7821805.

[15] J. Brownlee, "Multinomial Logistic Regression With Python", Machine Learning Mastery, 2022. [Online]. Available:https://machinelearningmastery.com/multinomial logistic-regression-with-python/. [Accessed: 06- Dec- 2021]

[16] S. Arafin Mahtab, N. Islam and M. Mahfuzur Rahaman, "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine," 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018, pp. 1-4, doi: 10.1109/ICBSLP.2018.8554585.

[17] M. Naveed, Q. Quratulain and A. Shaukat, "Comparison of GLCM based Hand Gesture Recognition Systems using Multiple Classifiers," 2021 International Conference on Robotics and Automation in Industry (ICRAI), 2021, pp. 1-5, doi: 10.1109/ICRAI54018.2021.9651396.

[18] A. Chaudhary, S. Kolhe and R. Kamal, "An improved random forest classifier for multi-class classification", Information Processing in Agriculture, vol. 3, no. 4, pp. 215-222, 2016.

[19] S. Xu, "Bayesian Naïve Bayes classifiers to text classification", Journal of Information Science, vol. 44, no. 1, pp. 48-59, 2016

[20] H. Alpert, Towards Data Science, 05-May-2022.