NED Chatbot for Admission Related Queries Using Prescriptive Analysis

Wisha Raees[1], Muhammad Ali Ismail[2] Ruqqaiya Aziz[3], Amber Afshan[4]

High Performance Computing Centre (HPCC), Department of Computer & Information Systems Engineering, NED University of Engineering & Technology, Pakistan

Summary
During admissions in NED University of Engineering and Technology thousands of questions or queries are generated related to admission process by the applicants. To solve this problem, an online admission query Chatbot has been developed to respond to student's queries at no time. The operational users can get instant answers of their queries about the admission procedure and different issues related to admission using this chatbot. This online admission query chatbot uses artificial intelligence algorithm and perspective analysis to analyze user's queries and respond accordingly. The Chatbot is trained on 20,000+ queries of the previous three years data. It uses SVM algorithm for classification of queries which is implemented on the R Studio and generating the responses. The GUI is based on JavaScript, J Query and Ajax and my SQL is used for backend.

Key words:
Chatbot, SVM algorithm

1. Introduction

Every year approximately 15,000 students apply for the admissions in NED University of Engineering and Technology and face a lot of queries related to the admission procedure. For answering these thousands of queries, there are only two or three persons who are responsible for handling this hectic process.

This inspired us to make a transformation for NED by implementing a chatbot project for NED so that all this could be done automatically. The elementary notion behind this project is the model of chatbot which is software anticipated to impersonate the human conversations. Chatbot can run on desktop computers and phones, and best of the time it is accessed over the internet. This model of chatbot uses the techniques of data analysis which is by definition a practice to apply statistical procedures and rational methods to evaluate data and helps in prescribing fruitful results. Data analysis is the technique in which data is inspected, cleaned, transformed and modeled into some useful information which derived some effective conclusion and support decision making. There are various specific data analysis methods, few of which comprised of text analytics, data mining, business intelligence and data visualization [1].

Data analytics is of four main types as shown in figure 1. Descriptive analysis is the simplest form of data analysis. It lists and encapsulates the values of each variable in a data set. It also assists you become familiar with a data set and to determine problems with the data. It provides insight into the past by using data aggregation and data mining [1]. Predictive analysis is the most advanced types of analytics and it has the ability to predict what might happen. This analysis is used to predict the future by analyzing the historical data using data mining, statistics, modeling, artificial intelligence, machine learning. It delivers the estimated probability of a certain consequence. Remember it is difficult to predict future with 100% certainty with any statistical algorithm. Prescriptive analysis integrates a feedback loop which combines descriptive and predictive models which influence each other and direct the tendency instead of just detecting them. It is used with the predictive analysis to decide the future performance based on the current and previous data. It not only forecast what will happen and when it will happen but also the reason of happening.

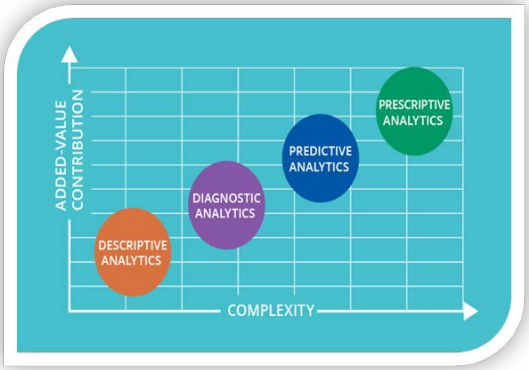

Fig. 1 Types of Data Analysis

This paper introduces a pioneer contribution by presenting a chatbot for responding the queries of students related to the admission procedure that has been developed in the High Performance Computing Centre, NED University of

Manuscript received May 5, 2019
Manuscript revised May 20, 2019

Engineering and Technology. This chatbot applies data analytics techniques and SVM algorithm on data set for providing the run time decision based on existing consequences instantly. The purpose of this paper is to ensure some influence in the domain of data analysis and artificial intelligence by way of they both technologies are useful for problem solving scenario.

Remaining paper comprises of related work in this field which is covered in Section II. Section III explains the flow of the Chatbot. The procedure for collection and categorization of data is described in Section IV. The description of SVM algorithm is given in Section V, which is followed by the Section VI covering the conclusion and future work.

2. Related Work

In 2018, a college chatbot has been developed related to educational system which responds to the queries of college students [2]. Some examples of chatbot technology are virtual assistants like Amazon's Alexa and Google Assistant, and messaging apps, such as WeChat and Facebook messenger [3]. Just like these chatbot this paper also presents a chatbot which is implemented by using SVM Algorithm, REST API and machine learning but the data used for analysis by the chatbot is real data of a university and is still innovative in sense that it helps the university in answering the queries asked by the candidate related to admission without any human intervention. This minimizes the need of human to get over thousands of queries and students to wait for longer time.

3. Flow of Chatbot

The flow of this chatbot system is defined as the user enters the query, the query is going to be saved in excel file and then the excel file is read by R studio on which the query is tested through SVM algorithm and then it returns the category in which the query lies. There are 87 categories on which we have trained the chatbot system and the testing query which is entered by the student must be laid in one of these categories. We have also prepared an answer file in which answers of the categories are written. The category returns by the SVM algorithm is then going to match the answer from the answer file and then write that answer in the output file. In the last, the answer is going to be read by NetBeans and will get displayed on the chat interface of the website. The flow of this chatbot system is presented in figure 2.

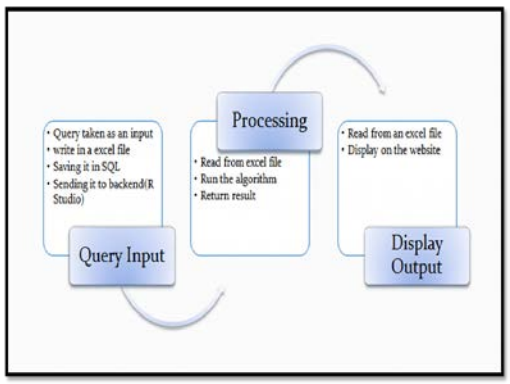

Fig. 2 Flow diagram of Data Processing

4. Collection and Categorization of Data

For any data analytics application, the data follow can be split into many stages. As shown in figure 3. The first step is data collection. Here we have used 3 years of admission related data having 2000+ quarries. The objective of this data collection and categorization was to train the chatbot with all potential and to produce responses according to the classification.

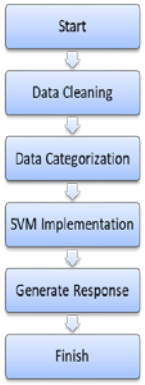

Fig. 3 Flow Chart of Chatbot System

4.1 Collection of Data

For data collection, we have requested the focal person who is accountable for answering the queries of students and asked them about the types of queries which students frequently inquired. We have collected the data of three years from admission office of NED University of Engineering and Technology. The data collected was is in row form and then transferred to suitable form.

4.2 Data Cleaning

Data Cleaning is the most important aspect of data processing. It imparts the data in the consistent form. The

raw data is passed from a series of steps to filter it so as to remove all the unwanted and redundant information. This stage basically removes all the punctuations („„“, “! “, “ :“,“;“,“,“,“,“?“) from the data and replaces them with blank spaces. All the words written in upper-case are converted into lower case. It is done to assure regularity and avoid ambiguity.

4.3 Data Categorization

In this phase, the cleaned data is categorized into the related categories and it is done manually in Microsoft Excel. To determine the trends of the questions the word clouds were extracted determining with few to many possibilities. The following figures 4 & 5 show the extracted word cloud on the 20000+ past three years quarries.

Fig. 4 Word Cloud of Quarries (Dataset)

Fig. 5 Word Cloud of Quarries (Dataset)

Based on the word cloud observations, 87 different categories were made on the dataset. It is also the part of bringing the data into the form so that prescriptive analytical techniques can be applied on it. A related snapshot defanging the categorization on the data set is shown below in the figure 6.

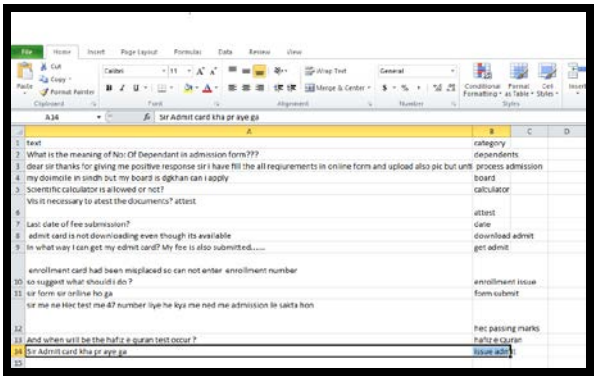

Fig. 6 Categorization of data in MS-Excel

We have prepared the answer file for all the categories so that when SVM predicts the result, it then matches the response of the particular category from the answer file. Below figure 7 shows the answer file in the Microsoft Excel.

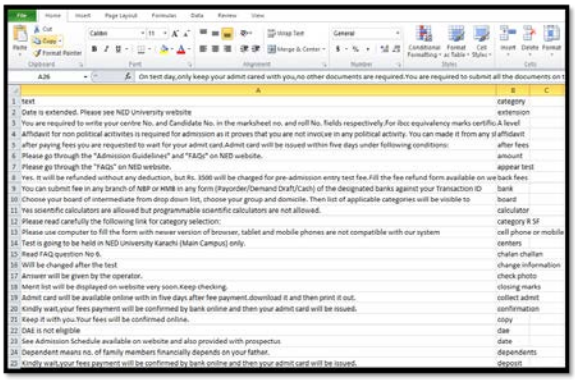

Fig. 7 Answer file in MS-Excel

5. SVM Algorithm

SVM (Support Vector Machine) is a supervised learning model in machine learning which is used for the purpose of data classification. The objective of SVM is to perfectly identify the unseen data. The classification is done by finding hyper-plane that separates the classes very well. SVM is widely used in text categorization, handwriting recognition, image classification and in sciences [4]. The function of SVM is to determine the belonging of the new data with the existing category. SVM is also known as non-binary linear classification.

5.1 How SVM works Well for Text Categorization?

5.1.1 High dimensional input space

When learning text classifiers has to deal with various (more than 10000) features. Since SVMs use over fitting protection which does not necessarily rely on the number of features, they have the potential to deal with these large feature spaces.

5.1.2 Few irrelevant features

One way to evade these high dimensional input spaces is to suppose that most of the features are inappropriate. Unfortunately, there are only very few irrelevant features in text categorization. All features are ranked with respect to their (binary) mutual information. A classifier using only those worst features has a performance much better than random. Since it seems unlikely that all those features are completely redundant, this leads to the estimation that a good classifier should combine many features and that feature selection is likely to compromise performance due to a loss of information.

5.1.3 Document vectors are sparse

For each document, the corresponding document vector holds only few entries which are not zero give both theoretical and empirical evidence for the mistake bound model that additive algorithms, which have a similar inductive bias like SVMs, are well suited for problems with dense concepts and sparse instances. These arguments give evidence that SVMs should perform well for text categorization [5].

5.2 SVM Kernels

The kernel's function is to take the data as input and transform it into the required form. The assortment of Support Vector machine kernel is fairly complicated and data dependent. Different types of kernels use different types of SVM algorithms. Some of them are described below.

5.2.1 Linear kernel

Most of the text classification problems are linearly separable. Linear kernel is used when a number of features are greater than the number of observations. It is much faster to solve the optimization problem for a linear kernel. It is a perverted version of RBF and works well when there is a lot of a feature. It is well suited for text classification and performs more efficiently when the dataset is linearly separable. Training SVM with linear kernel is much faster than any other kernel.

5.2.2 RBF kernel

RBF (Radial Basis Function) Kernel is used to classify data that is not linearly separable and when there is a high variance in dataset. Gamma and C are the parameters of RBF in which gamma defines the distance of the single training example, in which high value means "close" and low value means "far". The parameter C is defined by user and it control the exchange between frequency of error and complexity of decision rule.

5.2.3 Polynomial kernel

Polynomial kernel is suitable in image processing. Polynomial kernel allows the learning of non-linear models by representing the resemblance of training samples in a feature space over polynomials of the original variables.

5.2.4 Sigmoid kernel

Sigmoid kernel is used as an alternative for neural networks. Due to the origin from neural network, sigmoid kernel is quite suitable for support vector machine.

5.2.5 Tuning

For finding the best cost and gamma, SVM tuning is done. The figure 8 & 9 depict the results.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 epsilon cost
       0    8

- best performance: 0.03333333

- Detailed performance results:
   epsilon cost       error dispersion
1       0.0    4 0.04666667 0.06324555
2       0.1    4 0.04666667 0.06324555
3       0.2    4 0.04666667 0.06324555
4       0.3    4 0.04666667 0.06324555
5       0.4    4 0.04666667 0.06324555
6       0.5    4 0.04666667 0.06324555
7       0.6    4 0.04666667 0.06324555
8       0.7    4 0.04666667 0.06324555
9       0.8    4 0.04666667 0.06324555
10      0.9    4 0.04666667 0.06324555
11      1.0    4 0.04666667 0.06324555
12      0.0    8 0.03333333 0.06478835
```

Fig. 8 Tuning to evaluate performance of SVM algorithm

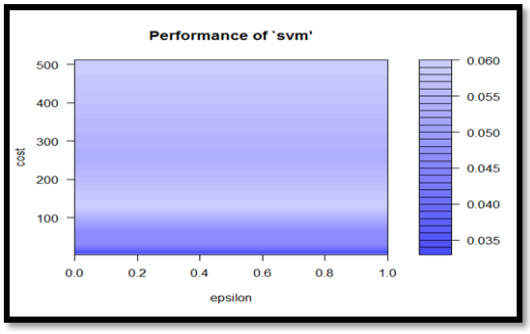

Fig. 9 Graph between epsilon and cost for showing the performance of SVM algorithm

5.3 Best Model

Radial basis kernel has least classification error and is suitable for non-separable data. As our dataset is not linearly separable therefore we decided to choose Radial Basis as SVM kernel.

6. Result

The network was trained for all of the 87 categories. And then it was tested. The accuracy varies from 99% to 90% for different categories. The following figures 10-12 show the accuracy against each category. During the chat if any of the answers is found not related then the user has as option to mark that. Then all the marked quested are answered by the operator directly. Besides, the system is further trained with the correct categorization.

Category	Accuracy
extension	95.8
A level	98.7
affidavit	98.6
after fees	94.2
amount	95.8
appear test	93.4
back fees	96.5
bank	98.5
board	93
calculator	95
category R SF	88.3
cell phone or mobile	90
centers	95
chalan challan	94.7
change information	88.4
check photo	94.6
closing marks	93
collect admit	98.4
confirmation	99
copy	99
dae	95.6
date	98.4
dependents	97.4
deposit	95.4

Fig. 10 Accuracy against each category (a)

A	B	C
document	94.5	
domicile	97.3	
download admit	97.4	
entry test	98.4	
extra	88.5	
field	91.2	
form submit	93.4	
get admit	97.5	
hafiz e Quran	90.4	
hec fee	95.6	
hec passing marks	98.5	
hec test	96.6	
help	99	
how upload admit	99	
hsc year	92.4	
id transaction	95.6	
improvement	91.4	
installment	99	
invalid format	90	
issue admit	99	
last merit	94.5	
link admit	96.4	
marks hec	93.4	
marksheet upload	97.6	
medical test	98	

Fig. 11 Accuracy against each category (b)

A	B	C
merit list	99	
much fee	92.4	
number marksheet	95.6	
ok	100	
open day	99	
pay fees	90	
pay order	99	
photograph	94.5	
picture	96.4	
print admit	97.4	
procedure	97.6	
process admission	95	
process admit	99	
receive admit	99.5	
received fees	96.4	
refund	97.4	
reply	99.6	
result	95	
returned fees	99	
SAT	94.5	
scholarship	96.4	
self finance SF	97.4	
send admit	97.6	
slip	95	
sponsor	99	

Fig. 12 Accuracy against each category (c)

A user interface for this Chatbot is also developed. This UI is integrated with the university admission portal. The candidate can make the related query using the interface. The flowing figure 13 & 14 show some snap shots of the UI.

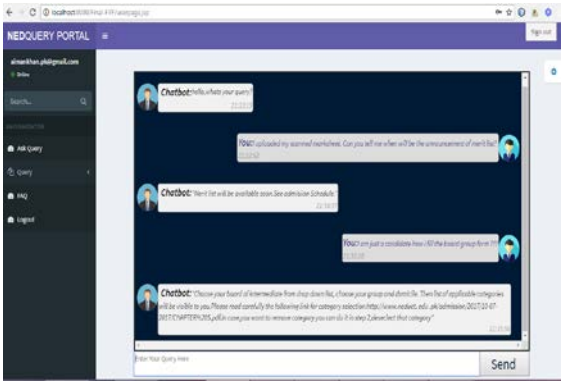

Fig. 13 Chatbot interface

Fig. 14 Chatbot interface

7. Conclusion

This Chatbot System uses prescriptive analysis and is built to minimize the human intercession. The main purpose of this system is to reduce the hectic procedure of delivering answers to the candidates. It enhances the performance of answering as well as reduces the time which leads the students to wait for their queries get answered. The user no longer has to wait for replies as the system will generate the reply on the spot instantly. The system successfully accomplished the task of classifying the queries and generates responses accordingly. The SVM algorithm is used for classification of queries and does a great job in non-linearly separable problems.
The system is being worked out for improved response time and making compatible with roman Urdu.

Acknowledgment

The research is conducted at High Performance Computing Centre (HPCC), NED University of Engineering and Technology .

References

[1] https://www.datasciencegraduateprograms.com/prescriptive-analytics
[2] G. Hiremath, A. Hajare, P. Bhosale, R. Nanaware. ”Chatbot for education system.” International Journal of Advance Research, Ideas and Innovations in Technology,Vol. 4,2018
[3] https://www.investopedia.com/terms/c/chatbot.asp
[4] http://www.statsoft.com/textbook/support-vector-machines
[5] T. Joachim.” Text Categorization with Support Vector Machines: Learning with Many Relevant Features.”
[6] http://scikitlearn.org/stable/aut_examples/svm/plot_rbf_parameters.html