INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

Student feedback prediction using machine learning

Nilesh Singh Sengar
nilesh1597@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

Kartik Chourey
kartik1610@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

Shraddha Bajaj
shraddhabajaj1997@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

S. Abinayaa
abi2827@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

ABSTRACT

Advances in natural language processing (NLP) and educational technology, as well as the availability of unprecedented amounts of educationally-relevant text and speech data, have led to an increasing interest in using NLP to address the needs of teachers and students. Educational applications differ in many ways, however, from the types of applications for which NLP systems are typically developed. This paper will organize and give an overview of research in this area, focusing on opportunities as well as challenges. In the proposed methodology, we are collecting student feedback about classroom, exam and lab facilities. After that, we are predicting the dataset using natural language processing. In NLP, we will use 2 approaches: NLTK toolkit and the random forest algorithm approach. After predicting the data, we will be analyzing the positive, negative and neutral nature of the data. This will help in effectively improving the performance of the institution and to create plans to enhance institutions' teaching and learning experience. Student's feedback improves communication between the lecturer and the students, allowing the lecturer to have an overall summary of the student's opinion.

Keywords— *NLP, NLTK, Random forest algorithm*

1. INTRODUCTION

Natural language processing (NLP) has over a 50-year history as a scientific discipline, with applications to education appearing as early as the 1960s. Initial work focused on automatically scoring student texts as well as on developing text-based dialogue tutoring systems, while later work also included spoken language technologies. While research in these traditional application areas continues to progress, recent phenomena such as big data, mobile technologies, social media have resulted in the creation of many new research opportunities and challenges. Commercial applications already include high-stakes assessments of text and speech, writing assistants, and online instructional environments, with companies increasingly reaching out to the research community. NLP can enhance educational technology in several ways. As an example of the first role, NLP is being used to automate the scoring of student texts with respect to linguistic dimensions such as grammatical correctness or organizational structure. As an example of the second role, dialogue technologies are being used to achieve the benefits of human one-on-one tutoring - particularly in STEM domains - in a cost-effective and scalable manner. Examples of the third role include processing text from the web in order to personalize instructional materials to the interests of individual students, automate the generation of test questions for teachers, or (semi-)automate the authoring of an educational technology system. Given the increasing interest in applying natural language processing to education, communities have emerged that now sponsor regular meetings and shared tasks. Recent shared academic tasks have included student response analysis, grammatical error detection (Ng et al. 2014), and prediction of MOOC attrition from discussion forums (Rose and Siemens 2014). There have also been highly visible competitions sponsored by the Hewlett Foundation in the areas of the essay and short answer response scoring. A research problem in the area of NLP for educational applications is usually inspired by a real-world student or teacher need. For example, although word count can very accurately predict many types of essay scores, word count is typically not part of a human's grading rubric and would thus not be useful to mention in student feedback. Finally, since many NLP algorithms are embedded in interactive applications, technical solutions often need to be real-time. In addition to building computer tutors, other uses of dialogue technology for teaching have been explored. Researchers have developed systems that play the role of student peers rather than expert tutors. There has also been interesting in going beyond one-on-one computer-student conversational interaction, by not only enabling human-machine but also improving human-human communication. Dialogue agents have been used to facilitate a student's

dialogue with other human students as in computer-supported collaborated learning or to enable students to observe the training dialogues of other students and/or virtual agents.

2. RELATED WORK

The analysis work "Prediction of Sentiment Analysis on Educational Data based on Deep Learning Approach by Ms Jabeen Sultana" has chosen eight completely different classifiers for comparative analysis of the performance of classifiers and designated the most effective classifier which provides best results among all. The educational dataset is taken from the Kiteboard 360 dataset repository and fed to various classifiers namely SVM, MLP, Decision Tree, K-star, Bayes Net, Simple Logistics, Multi-class Classifier and Random Forest. The data set is trained well on each classifier and a Model is obtained then checked with test data, results are obtained. The obtained results are calculated and evaluated in terms of measures like Accuracy, RMSE, Specificity, Sensitivity and ROC curve area. This paper proposed a student's performance prediction model based on MLP-deep learning method, SVM, Decision tree and other machine learning algorithms. The performance of student's predictive model is evaluated by a set of classifiers, namely; SVM, MLP, Decision Tree, K-star, Bayes Net, Simple Logistics, Multi-class Classifier and Random Forest. Ten-fold cross-validation (CV) is performed. Results indicate that SVM and MLP-Deep learning method conceded good performance in general compared to the other classifiers in terms of classification accuracy, RMSE, sensitivity and specificity and ROC curve area.

2.1 Proposed work

In the proposed methodology, we are collecting student feedback about classroom, exam and lab facilities. After that, we are predicting the dataset using natural language processing. After predicting the data, we will be analyzing the positive, negative and neutral nature of the data. 2 types of approaches are being used in the system.

- Using NLTK toolkit

This approach is used for the prediction of a smaller data set. It gives a generalized result whether the feedback is negative, positive or neutral.

- Using Random Forest Algorithm

This algorithm is being used to predict the data from large datasets and in more precision way. With the use of this algorithm the data is divided into groups; positive and negative, and then within these groups various categories like an exam, library, class, food, staff, bus etc. are analyzed. The analysis is done on the basis of testing and training data. The dataset to be predicted is divided into train and test data (with certain percentage division). After division, the training data is used to predict the accuracy of the algorithm. Then the algorithm is applied onto the test data and the results are predicted. The positive and negative feedback groups are further analyzed and pie charts are plotted with the individual division of the category like how much positive for library feedback, how much negative for exam feedback etc.

Both the approaches use a standard set of 4 phases for prediction which is explained as follows

2.1.1 Dataset collection: In this phase, the data is prepared for the analysis purpose which contains relevant information. Pre-processing and cleaning of data are one of the most important tasks that must be one before dataset can be used for machine learning. The real-world data is noisy, incomplete and inconsistent. So, it is required to be cleaned.

2.1.2 Pre-processing: Raw feedback scraped from twitter or feedback forms are generally result in a noisy dataset. This is due to the casual nature of people's usage of social media and forms. Feedback has certain special characteristics such as emoticons, user mentions, etc. which have to be suitably extracted. Therefore, raw data has to be normalized to create a dataset which can be easily learned by various classifiers. We have applied an extensive number of pre-processing steps to standardize the dataset and reduce its size. We first do some general pre-processing on feedback which is as follows:

- Convert the feedback to lower case.
- Replace 2 or more dots (.) with space.
- Remove punctuations.
- Remove non-English and stop words.
- Replace 2 or more spaces with a single space.
- Remove proper nouns.

2.1.3 Extraction of feature set: In this phase, the cleaned data that is obtained from the data preprocessing phase is used to obtain the feature sets or training data of the student feedback. When we train to a classifier by taking maximum numbers of features, that contains all the irrelevant or redundant features can negatively affect the algorithm performance. So, it is required to carefully select the number and types of features that will be used to train the machine learning algorithms. Various feature selection techniques can be used for selecting features in the feature set/training data. Feature set or training data can be prepared from the cleaned data by using any of the available techniques like a bag of words, -gram, N-gram, POS, TOS tagging etc. The training data can also be prepared by providing those labels and then divides it into two classes like positive class and negative class. The feature sets and training set that has obtained by using any of the above methods will be used for the implementation of machine learning algorithms.

2.1.4 Testing on datasets: Testing of data is done based on a training model which is classified using a supervised learning algorithm. Evaluation of the total responses for every question and determine the polarity of feedback received in the context of the question. The evaluation of response is purely data-driven and hence simple while the classification of questions in the form of natural language texts involves sentiment analysis. To test the model, collected data from students who posted their views in online discussion forums is analyzed.

3. ANALYSIS AND RESULT

- Using NLTK approach for a set of 40-50 feedback data entries the result predicted is

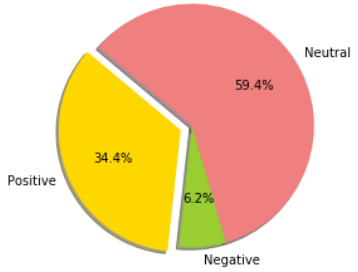

Fig. 1: NTLK approach results

- Using Random forest approach for a set of over 1000 feedback data entries is

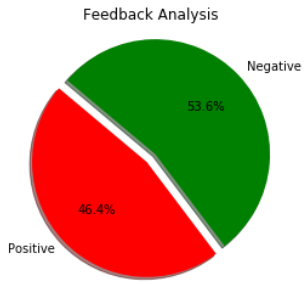

Fig. 2: Random forest approach results

- Negative Feedback Result

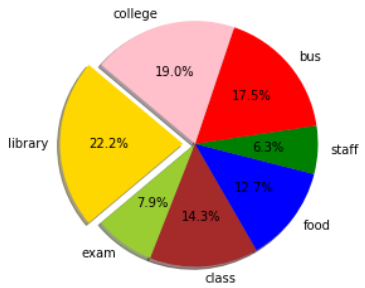

Fig. 3: Negative feedback result

- The frequency of words being used in a negative feedback

Fig. 4: frequency of words used in a negative feedback results

- Positive Feedback Result

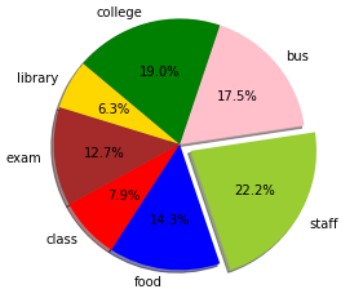

Fig. 5: Positive feedback result

- The frequency of words being used in negative feedback

Fig. 6: frequency of words in negative feedback

4. CONCLUSION AND FUTURE ENHANCEMENT

From the research done and implementation of the algorithm's prediction and analysis of the data can be done. Using the prediction results what improvements are to be required for the betterment of the conditions of the schools' and colleges' and the respective categories can be found out. This good help in improving the schools' and colleges' quality and status. For future improvements and enhancements, more efficient algorithms can be applied. Also, the research can be extended in the field of e-commerce, banking and medical departments also.

5. REFERENCES

[1] Khan, Khairullah, "Mining opinion components from unstructured reviews: A review," Journal of King Saud University – Computer and InformationSciences, Vol. 26, 2014.

[2] Jyotsna Talreja Wassan, "Discovering Big Data Modelling for Educational World", IETC Procedia - Social and Behavioral Sciences, pp:642 – 649, 2015

[3] Dirk T. Tempelaar, "In the search for the most informative data for feedback generation: Learning analytics in a data-rich context", Journal of Computers in Human Behavior, Vol 47, pp: 157– 167, 2015.

[4] Beth Dietz-Uhler and Janet E. Hurn, "Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective", Journal of Interactive Online Learning, Volume 12, Number 1, 2013.

[5] Marie Bienkowski, Mingyu Feng, "Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics", Department of Education, Office of Educational Technology: October 2012.

[6] Pang, B. and Lillian L. s.l, "Opinion mining and sentiment analysis: Foundations and trends in information retrieval", Vol. 2, 2008.

[7] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, Vol.5, 1093–1113, 2014.

[8] Kumar Ravi, Vadlamani Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", Published in Knowledge-Based Systems Vol. 89, 14–46, 2015.

