Feature Based Opinion Mining on Student Feedback

Problem Statement

Student feedback have the presence of huge amount of structured data like the grades, enrollment data, progression rates as well as unstructured data like student opinions expressed through surveys, web blogs, twitter, Facebook etc. It becomes highly time and resource consuming to summarize the information manually to reach data led conclusions and decisions. It is crucial to understand the patterns generated by the data like student feedback to effectively improve the performance of the institution and to create plans to enhance institutions' teaching and learning experience. Opinion Mining technique for classifying the students' feedback obtained during evaluation survey that is conducted every semester to know the feedback of students with respect to various features of teaching and learning such as module, teaching, assessments, etc. The extracted and preprocessed datasets can be subjected to various machine learning algorithm such as Support Vector Machine (SVM), Naïve Bayes (NB), K Nearest Neighbor (KNN) and Neural Networks (NN).

Background

Students feedback can help the lecturers understand their students learning behavior and improve teaching. Taking feedback can highlight different issues that the student may have with the lecture. One example of this is when the student does not understand part of the lecture or a specific example. Another example is when the lecturers' teaching pace is too fast or too slow. Feedback is usually collected at the end of the unit, but it is more beneficial taken in real-time.Collecting feedback has numerous benefits for the lecturer andtheir students, such as improvement in teaching and understanding student's learning behavior. Student's feedback improves communication between the lecturer and the students, allowing the lecturer to have an overall summary of the student's opinion. Student feedback can be collected using mobile phones and social media.

Feature based opinion mining deals with the extraction of the different features of the feedback of the student in an educationalorganization. It can be used to extract information from the student feedback about the teaching and learning methods adopted in an educational institute.Student feedback have the features like the grades, enrollment data, progression rates as well as unstructured data like student opinions expressed through surveys, web blogs, twitter, Facebook etc.

Student feedback is collected in form of responses toquestions in a single sentence, it requires sentiment analysis in sentence level. In sentiment classification, machine learning methods have been used to classify each question as positive or negative. Testing of data is done based on training model which is classified using supervised learning algorithm. Evaluation of the total responses for every question and determine the polarity of feedback received in context of the question. The evaluation of response is purely data driven and hence simple while the classification of questions in form of natural language texts involves sentiment analysis. To test the model,collected data from students who posted their views in online discussion forums.

Methodology

Step 1: Data collection

This will involve collection of studentfeedback in the form of structured data like the grades, enrollment data, progression rates as well as unstructured data like student opinions expressed through surveys, web blogs, twitter, Facebook etc.

Step 2: Data Preprocessing

In this phase, the data is prepared for the analysis purpose which contains relevant information. Pre-processing and cleaning of data areone 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.

Step 3: Extraction of Feature Set/Training Data

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 bag of words, -gram, N-gram, POS, TOS tagging etc. The training data can also be prepared by providing them labels and then divide 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.

Step 4: Implementation of Machine Learning Algorithm on Feature Set/Training Data

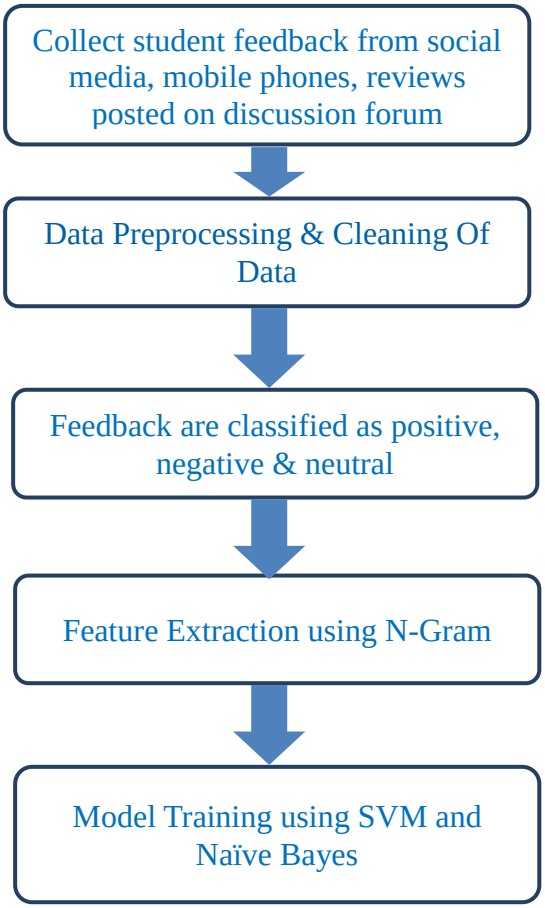
In the educational domain, Naïve Bayes and Support Vector Machines could be the best technique.

Naive classifier is a kind of probabilistic classifier. It is based on Bayes theorem with the assumptions that features between the feature sets are independent of each other. These classifiers are highly scalable. It is a simple method for developing a classifier. The model develops by Naïve Bayes classifier provide a class labels to the features of the feature set/training datasets.

SVM can be used for the analysis of data for both kind of supervised learning algorithms like classification and regression. When a set of training examples are given in which the sets are belongs to one or the other of two categories then SVM algorithm builds a model that will assigns unknown data the one or other categories' this way SVM is a non-probabilistic binary linear classifier. In SVM, representation of feature sets is done in the form of points in space, so that the features of the other categories are separated by clear space that is as wide as possible. Unknown feature is mapped and predicted to belongs to the space on which side of the gap it falls. It is shown as below

Step 5: Testing on Datasets

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



Experimental Design

Dataset

Student feedback can be collected using mobile phones, social media and in form of responses toquestions in a single sentence,from students who posted their views in online discussion forums.

Evaluation Measures

- **Accuracy**:Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

- **Precision:** It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

- **Recall:** It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Software and Hardware Requirements

Python based Computer Vision and Deep Learning libraries will be exploited for the development and experimentation of the project. Tools such as Anaconda Python, juypter notebook and libraries such as Tensorflow, and Keras will be utilized for this process.

References

1. Khan, Khairullah, “Mining opinion componentsfrom unstructured reviews:A review,” Journal of King Saud University – Computer and InformationSciences , Vol. 26, 2014.

2. Jyotsna TalrejaWassan, “Discovering Big Data Modelling for EducationalWorld”, IETC Procedia - Social and Behavioral Sciences, pp:642 – 649, 2015

3. Dirk T. Tempelaar, “In search for the most informative data for feedbackgeneration: Learning analytics in a data-rich context”, :Journal of Computersin 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 InteractiveOnline Learning, Volume 12, Number 1, 2013.

5. Marie Bienkowski, Mingyu Feng, “Enhancing Teaching and LearningThrough Educational Data Mining and Learning Analytics”, Department ofEducation, 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, HodaKorashy, “Sentiment analysisalgorithms and applications: A survey”, Ain Shams Engineering Journal, Vol.5, 1093–1113, 2014.

8. Kumar Ravi, Vadlamani Ravi, “A survey on opinion mining and sentimentanalysis: Tasks, approaches and applications”, Published in Knowledge-BasedSystems Vol. 89 , 14–46, 2015.