International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

PREDICTING STUDENTS RESULT BASED ON TOPIC MODELING -LDA

Suja Babu[1], Aswathy R Kaimal[2]

Suja Babu , Dept. Of computer Science and Engineering, Belivers Chruch Caarmel Engineering CollegeKoonamkara P.O, Ranni-Perunad, Pathanamthitta Dist, Kerala- 689711,India

Prof Aswathi R Kaimal, Dept. Of computer Science and Engineering , Belivers Chruch Caarmel Engineering College,Koonamkara P.O, Ranni-Perunad, Pathanamthitta Dist, Kerala- 689711,India

---

Abstract - *The development of students results prediction in an academic organization and Digital learning environment.* ***Topic modeling*** *can be used for discovering the abstract topics that occur in a collection of students post. Data mining provides many tasks that could be used to study the educational field. Project will be divided into two main parts- one is Topic Modeling-LDA for collection of words and another one is prediction by classification. The machine learning algorithms were tested including Support Vector Machine(SVM), Naïve Bayes, and k-Nearest Neighbor(kNN). As well ,concerned with the interactions between computer compiler and human language techniques were employed to determine the optimum preprocessing configuration to produce relevant results. Found that detecting students result prediction was most successful using a kNN classifier with proper classification preprocessing. This model produced a promosing initial F1-score of 0.92% and an accuracy of 0.89%. And Topic Model-LDA Visualization produce detailed information on each topics which helps to organize, search and understand information.*

Key Words: **Machine Learning, natural language processing, Topic Modeling-LDA, e-learning, sentiment analysis, opinion mining**

1.INTRODUCTION

This document is The Text data mining is the process of deriving high-quality data from text. High-quality information is derived through the patterns and trends through means such as statistical pattern learning. The growth of text-based data sources available from social media tools, online reviews, and other big data communication platforms, research in sentiment analysis, or opinion mining, has become a rapidly increasing in size and changing area of study in recent years. Actually, important research work has away into developing natural language processing tools and related machine learning techniques that try to identify the polarity of audience sentiment around relevant entities and their various attributes. Examples of studies in this area include work with Opinion Fraud Detection in Online Reviews and product review databases.

1.1 Background

e-Learning courses are changing the face of education. There are many machine learning websites on the internet offering courses and certifications online. Websites are most certainly the powerful bridge between the students and the online courses. Website is the first almost every point of contact between them. It has greatly beneficial to students and instructors due to its flexibility, cost reductions, and overall enhanced experience provided by the utilization of advanced technology innovations. In an e-learning platform, instructors should be able to collect data in which students watch the lectures, where they flag sections for questions, and how often they return to the videos for review as our instructor do with our learning platform. Having this data allows instructors to intervene and motivate students who are confused or falling behind.

e-learning data analytics involves the collection, analysis of information gathered from the participants as they undertake an Learning course. Learning analytics is real-time process rather than a retrospective one. It makes one of the important e-learning website features. Historically, teaching online made it difficult to catch when a student was confusing. However, by grip emerging technology and data generated by these tools, it has made it easier for us to help students before it is too late. Identifying where students are confusing allows instructors to spend time with students more effectively and provide personalized instruction. If an individual is observed to be struggling over a particular course module, any one can be offered access to custom tools and resources like links to specific and related websites supplements that throw a greater light on the subject.

The next general background is used as the Topic Modeling. Topic modeling is a type of statistical modeling for discovering the abstract topics that occur in a collection of data. Latent Dirichlet Allocation is an example of topic model and is used to group of text in a document to a particular

topic. It construct a topic per document model and words per topic model, modeled as Dirichlet distributions

1.2 Objectives

The main objective of the project is to use data mining methodologies to study and analyze the school students' result. Data mining provides different tasks that could be used to study the students' result. Here, the classification task is employed to gauge students' performance and deals with the accuracy, confusion matrices and the execution time taken by the various classification data mining algorithms. The growth of data sources available from social media tools, online reviews, and other big data communication platforms, research in sentiment analysis, or opinion mining, has become a rapidly growing area of study in recent years. The evolution of Tutor-Vigilant, a natural language processing (NLP) and machine learning (ML) system are two children of artificial intelligent (AI). NLP alike to those used in sentiment analysis, but applied to the data generated by students in E-learning course management system in order to predict students result. Topic modeling can be used for discovering the abstract "topics" that occur in a collection of documents. A bag-of-words model, or BoW for short, is a way of extracting features from text data for use in modeling, such as with machine learning algorithms. The bag-of-words model is to understand and implement and has seen great success in problems such as language modeling and document classification.

2. EXPERIMENT SETUP

A number of learning methods were including Support Vector Machines (SVM), Naive Bayes, and k-nearest neighbor, and an array of natural language pre-processing techniques were employed determine the optimum classification methods, and to increase the accuracy of the results.

Data mining provides different tasks that could be used to study the student result. Here, work will be divided into two main parts- one is Topic Modeling-LDA and another one is prediction by classification. At first will select our dataset and then perform preprocessing of it. Then apply LDA modeling over the dataset and perform prediction of result. Then will classifiers technique over the dataset and generate some rules which will be analyzed later. At last both result of prediction and accuracy will be visualized by ROC curve.

The followings are the step by step process of our project evaluation.

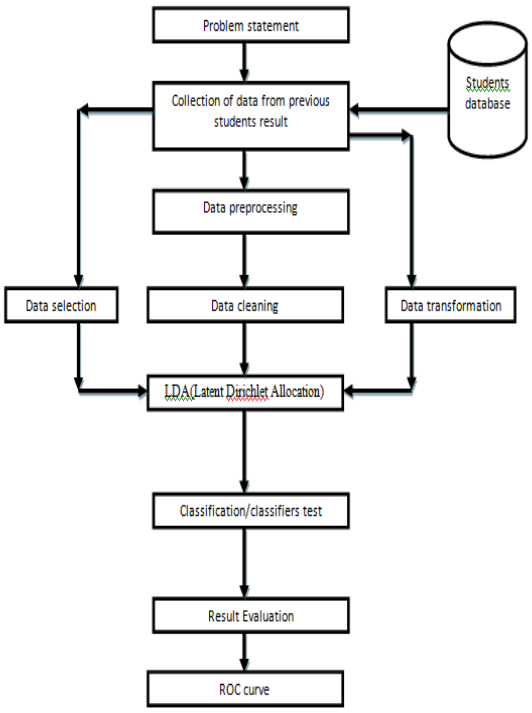
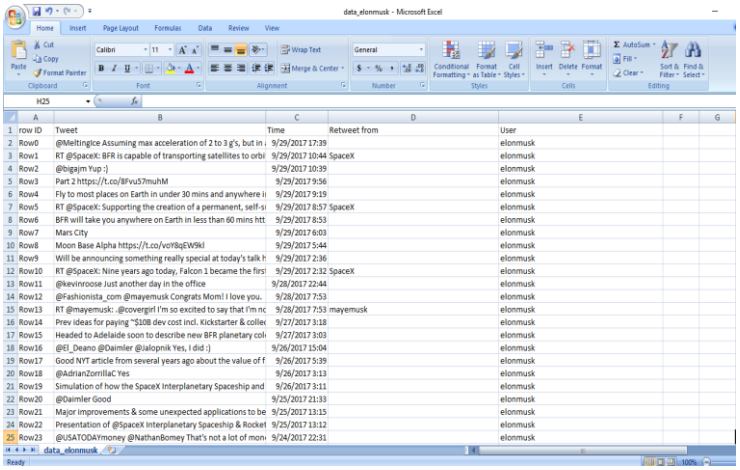


Fig -1: work flow Diagram

2.1 Dataset

Here have collected a dataset which contains the posts of students in learning environment. The data file has to be in 'CSV' format.

The dataset which is in 'CSV' format.



row ID	Tweet	Time	Retweet from	User
Row0	@MeltingIce Assuming max acceleration of 2 to 3 g's, but in	9/29/2017 17:39		elonmusk
Row1	RT @SpaceX: BFR is capable of transporting satellites to orbi	9/29/2017 10:44	SpaceX	elonmusk
Row2	@bigajm Yup :)	9/29/2017 10:39		elonmusk
Row3	Part 2 https://t.co/BFvu57muhM	9/29/2017 9:56		elonmusk
Row4	Fly to most places on Earth in under 30 mins and anywhere i	9/29/2017 9:19		elonmusk
Row5	RT @SpaceX: Supporting the creation of a permanent, self-s	9/29/2017 8:57	SpaceX	elonmusk
Row6	BFR will take you anywhere on Earth in less than 60 mins htt	9/29/2017 8:53		elonmusk
Row7	Mars City	9/29/2017 6:03		elonmusk
Row8	Moon Base Alpha https://t.co/voY8qEW9kI	9/29/2017 5:44		elonmusk
Row9	Will be announcing something really special at today's talk h	9/29/2017 2:36		elonmusk
Row10	RT @SpaceX: Nine years ago today, Falcon 1 became the first	9/29/2017 2:32	SpaceX	elonmusk
Row11	@kevinroose Just another day in the office	9/28/2017 22:44		elonmusk
Row12	@Fashionista_com @mayemusk Congrats Mom! I love you.	9/28/2017 7:53		elonmusk
Row13	RT @mayemusk: .@covergirl I'm so excited to say that I'm nc	9/28/2017 7:53	mayemusk	elonmusk
Row14	Prev ideas for paying ~$10B dev cost incl. Kickstarter & collec	9/27/2017 3:18		elonmusk
Row15	Headed to Adelaide soon to describe new BFR planetary col	9/27/2017 3:03		elonmusk
Row16	@El_Deano @Daimler @jalopnik Yes, I did :)	9/26/2017 15:04		elonmusk
Row17	Good NYT article from several years ago about the value of f	9/26/2017 5:39		elonmusk
Row18	@AdrianZorrillaC Yes	9/26/2017 3:13		elonmusk
Row19	Simulation of how the SpaceX Interplanetary Spaceship and	9/26/2017 3:11		elonmusk
Row20	@Daimler Good	9/25/2017 21:33		elonmusk
Row21	Major improvements & some unexpected applications to be	9/25/2017 13:15		elonmusk
Row22	Presentation of @SpaceX Interplanetary Spaceship & Rocket	9/25/2017 13:12		elonmusk
Row23	@USATODAYmoney @NathanBomey That's not a lot of mon	9/24/2017 22:31		elonmusk

Fig -2: Sample Dataset

2.2 Preprocessing

Data Preprocessing is the first step of evaluation of this project. Here the source data file is selected from local machine. After loading the data in Explorer, refine the data by selecting different options which is known as 'Data Cleaning' and can also select or remove attributes as per our need. The preprocess allows filters to be explain that transform the data in various ways. There are mainly two categories of filters-Supervised and

Unsupervised. Here will choose supervised category filters. All data is labeled and the algorithms learn to predict the output from the input data.

2.3 TopicModeling

Topic Modeling in NLP seeks to find hidden semantic structure in documents. They are probabilistic models that can help you comb through massive amounts of raw text and cluster similar groups of documents together in an supervised way.

2.4 Classification

Classification algorithm is supervised learning algorithms. The main goal of classification is to predict the target class (Pass/ Fail). If the trained model is for predicting of two target classes. It is known as binary classification. Considering the student previous result to predict whether the student will pass or fail. Suppose from train data come to know that your best friend likes the above movies. Now one new movie -test data released. Hopefully, you want to know your best friend like it or not. If you strongly confirmed about the chances of your friend like the movie.

To predict students result by using standard classifiers named are Support Vector Machines (SVM), Naive Bayes, and k-Nearest Neighbors Algorithm (k-NN) for classification. From these classifiers k-NN algorithm is best accuracy for students result prediction. The correctly and incorrectly classified examples show the percentage of test examples that were correctly and incorrectly classified.Here are some others factor in classifier output based on confusion matrix

Using these four parameters then here can calculate Accuracy, Precision, Recall and F1 score.

- Accuracy= TP+TN/TP+FP+FN+TN
- Precision = TP/TP+FP
- Recall = TP/TP+FN

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

3.RESULTS AND DISCUSSION

There have been a number of different machine learning algorithms that have been tested with different data sets, and within different domains, with varying results, so wanted to cast a wide net in determining which would be the best choice to predict students result classifier.

To predict students result by using Topic Modeling and standard classifiers named are Support Vector Machines (SVM), Naive Bayes, and k-Nearest Neighbors Algorithm (k-NN) forclassification. From these classifiers k-NN algorithm is best accuracy for students result prediction

Topic modeling is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents. Latent Dirichlet Allocation(LDA) is an example of topic model and used to group of text in a document to a particular topic. It construct a topic per document model and words per topic model, modeled as Dirichlet distributions. Finally construct LDA-visualization by topic per document model.Here , going to apply LDA to a set of documents and split them into topics.The data set use is a list of students reviews published over a period of 2 years .

Here will perform the following steps:

- Tokenization: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.
- Words that have fewer than 3 characters are removed.
- All stopwords are removed.
- Words are lemmatized
- words are reduced to their root form-stemmed

Preprocess the students posts, saving the results as processed_docs. Bag of Words on the Data set. Create a dictionary from processed_docs containing the number of times a word appears in the training set.Running LDA using Bag of Words: Train our lda model using gensim.models.LdaMulticore and save it to 'lda_model'
Looking at the results in table 1, the kNN algorithm obtained both the highest accuracy and F-score measures, closely followed by SVM and Naïve Bayes.

Table -1: Analysis result

Classifier	Accuracy	Precision	Recall	F score
SVM	.72%	.71%	100%	.83%
nb	.79%	.76%	100%	.86%
kNN	.89%	.86%	100%	.92%

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

4. CONCLUSION

Here, have described how a supervised machine learning system can be used to students result prediction in a digital online learning environment by using Topic Modeling, and initial results have been very satisfactory. By testing a best machine learning algorithm upon which to develop this sort of solution, through to the incorporation of built-in, ongoing supervised learning and, ultimately to the accuracy displayed in confusion matrix.

There have been a number of different machine learning algorithms that have been tested with different data sets, and within different domains, with varying results, so wanted to cast a wide net in determining which would be the best choice to build the students result predict classifier. Here developed three classifiers – an implementation of a Support Vector Machine algorithm called Seqential minimal optimization (SMO), k-nearest neighbor, Naive Bayes. For each classifier computed the Accuracy, Precision, Recall, and

F-score. The results are very promising. Knn algorithm obtained the highest accuracy, precision, recall and F1-Score

Finally, will be working to expand the use of Students result mark prediction through the creation of topic Modeling or integration into popular digital learning environments.

REFERENCES

[1] Steven C Harris, vivekanandhan kumar, "Identifying Student Difficulty in a Digital Learning Environment". IEEE Transactions on Information Forensics and security (volume 13,issue:5), may – 2018.

[2] Singh, V., Saxena, P., Singh, S., & Rajendran, S ,"Opinion Mining and Analysis of Movie Reviews". Indian Journal of Science and Technology, 2017 .R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[3] Singh, V., Saxena, P., Singh, S., & Rajendran, S ,"Opinion Mining and Analysis of Movie Reviews". Indian Journal of Science and Technology, 2017 .

[4] Sahayak, V., Shete, V., & Pathan, A ,"Sentiment Analysis on Twitter Data. International Journal of Innovative Research in Advanced Engineering" (IJIRAE), 2(1), 178-183, (2015).

[5] Bannier, Betsy J ,"Global Trends in Transnational Education", International Journal of INformation and Education Technology vol 6 no. 1, Jan 2016 .

[6] A. Simsek, "Global trends in distance education", presented at the International Conference on Communication, Media, Technology, and Design, Famagusta –North Cyprus, May 2-4, 2013.

[7] R. Bhandari and P. Blumenthal, "International Students and Global Mobility in Higher Education": National Trends and New Directions, New York, NY: Palgrave Macmillian, 2011.