Machine learning approach to predicting the acceptance of academic papers

Mikhail Skorikov
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
mikhail.skorikov@northsouth.edu

Sifat Momen
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
sifat.momen@northsouth.edu

Abstract—In this paper, machine learning approaches have been used to predict whether a scientific paper will be accepted in a top-tier AI conferences or not. This shall help authors identify the likelihood of their paper getting accepted in a top-tier AI conference. We have used the PeerRead dataset containing papers collected from major AI conferences that are publicly available. We have achieved an accuracy of 81% using Random Forest classifier. The novelty of the paper lies in accurately predicting whether a scientific paper will be accepted in the top AI conference.

Index Terms—artificial intelligence, machine learning, paper review, classification, acceptance prediction

I. Introduction

There has been a significant rise in the number of conferences being held across the globe every year. For example, EasyChair, a web conference management system, has so far served about 80,000 conferences since 2002 [1]. Out of these conferences, a significant portion of the conferences is related to computer science and artificial intelligence (AI). In recent times, machine learning (ML) has been an active area of research in computer science. With the rise in the popularity of machine learning applications, diverse scientific papers have been added to the knowledge body. The applications of machine learning is now present in disciplines as diverse as medical diagnosis [2], [3], [4], businesses [5], agriculture [6], [7], education [8], [9] as well as arts and music [10].

Our objective is to facilitate new researchers in the field of computer science and artificial intelligence (particularly machine learning) to understand the quality of their manuscript that has been written with the intention of publishing in a reputed conference/journal. In order to achieve this goal, ML-models are created to predict a paper's acceptance in top AI conferences. Although machine learning techniques have been used in different areas pertaining to education, very limited work has been carried out in facilitating early researchers.

The rest of the paper is organised as follows: In section II, we briefly discuss the literature review. Section III outlines the methodology of the research work followed by results in section IV. Finally, in section V, the paper is concluded with remarks on future work.

II. Literature Review

Kang and colleagues [11] collected the first public dataset comprising of scientific peer reviews and made them available for research purposes. They applied Natural Language Processing (NLP) on the dataset and were able to predict whether a paper should be accepted or not with an accuracy of 79%.

Qian and colleagues [12] predicted the number of institutions whose papers will be accepted in the top 8 conferences in computer science. Their research is centered around data mining techniques including feature definition and engineering, finding similar conferences, and dimension reduction methods. They also proposed three ranking models and applied ensemble learning to make the prediction.

Lykourentzo et al. used machine learning techniques to predict the dropouts from e-courses [13]. University dropouts were predicted and a prevention system was experimented on based on students who avail distance learning [14], and forecasting students' grades in e-courses was attempted [15].

Several other attempts have been made to predict student performance. Gray et al. [16] attempted to identify college students at risk of failing their first year of study. Xu et al. [17] predicts student performance in degree programs while addressing the diverse background and course selections of students, the fact that not all courses are equally informative for predictions, and that students' evolving progress is a vital matter for predicting future performance.

Nurhudatiana et al. [18] models research productivity of faculty members of Binus University International, investigating contributing factors to research productivity of junior, intermediate, and senior faculty members. They find that doctoral degrees in junior faculty members are their keys to productivity. The fact that tenured faculty members can only count as productive after two years of consistent publications, it is interesting to note that length of service to the university did not affect research productivity.

Ghosal and colleagues [19] predicted whether a sub-

978-1-7281-9336-6/20/$31.00 ©2020 IEEE

mitted paper is within scope of the journal or not, saving time for both authors and reviewers.

Guo et al. [20] advocated for a decision support system for manuscript submission to academic journals.

III. Methodology

Our research methoodolgy is outlined in figure 1.

A. Data Collection

The data is taken from the PeerRead [11] dataset, which consists of 14.7K machine learning paper drafts and corresponding accept and reject decisions from top-tier venues. The dataset provides the papers in PDF form as well as parsed JSON files that used a tool called Science Parse [21].

TABLE I
PeerRead dataset distribution

Conference	Year	Papers	Accept/Reject
NIPS	2013-2017	2420	2420 / 0
ICLR	2017	427	172 / 255
ACL	2017	137	88 / 49
CoNLL	2016	22	11 / 11
arXiv	2007 - 2017	11778	2891 / 8887
total	2007-2017	14784	5582 / 9202

B. Data Processing

The data-set contained papers from selected years of NIPS, ICLR, CONLL, ACL, and ARXIV conferences. There were about 10,000 reviews altogether. While some papers have reviews and others contain only the acceptance status. We initially selected six attributes of a paper to train the ML models. The six attributes include title length, number of tables, number of figures, number of citations, number of citations within three years of the conference, and the length of the literature review (measured in terms of the number of characters of this section). We did not use the number of equations as an attribute since retrieving such an attribute from the raw papers was difficult to manage reliably without resorting to manual data entry. The average performance was low, with about 65% accuracy and similar F1 scores.

We decided to apply a Bag-of-Words implementation to the abstracts of every paper, taking in 10 most frequent words (see figure 2) as additional features. We also used the conference of submission encoded using the One-Hot Encoding technique as features.

During the above processing actions, we discarded some papers due to improper formatting resulting in unreliable data points.

C. Data Description

In the end, we were left with a total of 23 features (described in Table II) and 14599 instances, of which 9029 papers were rejected, and 5570 were accepted.

D. Classification

The dataset is shuffled and then split into 80% training and 20% testing sets. After that, different machine learning algorithms, including Naive Bayes, Logistic Regression, K- Nearest Neighbours, Decision Tree, Random Forests, Support Vector Machines, and baseline ZeroR, were applied to find performance in predicting the acceptance of the papers.

1) Naive Bayes: It uses the Bayes' theorem to create a simple probabilistic classifier. The Bayes' theorem assumes that each feature is independent of another. Eq. (1) shows the Bayes' theorem in action:

$$P(c|\mathrm{x}) = \frac{P(\mathrm{x}|c)P(c)}{P(\mathrm{x})} \tag{1}$$

Where, c is the class, x is the feature vector, $P(c|x)$ = Posterior probability, $P(x|c)$ = Likelihood, $P(c)$ = Class Prior probability, and $P(x)$ = Predictor Prior probability.

2) K-Nearest Neighbor: It is one of the most intuitive algorithms to understand, and a value of k = 9 with the distance metric as Manhattan distance (1-norm) was used for this dataset.

KNN works by identifying k nearest data points and choosing the majority class from those. There are several distance metrics in use, which is used to calculate the nearest neighbors. Manhattan distance metric (see equation 2) is used here.

$$distance = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

3) Logistic Regression: Logistic regression is a linear regression algorithm transformed using a sigmoid function to make it a classifier capable of predicting probabilities of each class. Linear regression uses a hypothesis function described in Eq. (3). A logistic regression passes the linear regressor into a sigmoid function, making it a probability-calculating classifier, as shown in Eq. (4)

$$h_\theta(X) = \theta_0 + \theta_1 X \tag{3}$$

$$\phi(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

4) Decision Tree: Visualizing and understanding a decision tree is easy, and we found that an entropy-based decision tree with seven nodes deep is optimum for us. Decision tree works with information gain and entropy to split the features such that the data is distributed in the tree as heterogeneously as possible. Eq. (5) shows the formula for entropy, which is calculated for every feature and the highest is selected as the root node:

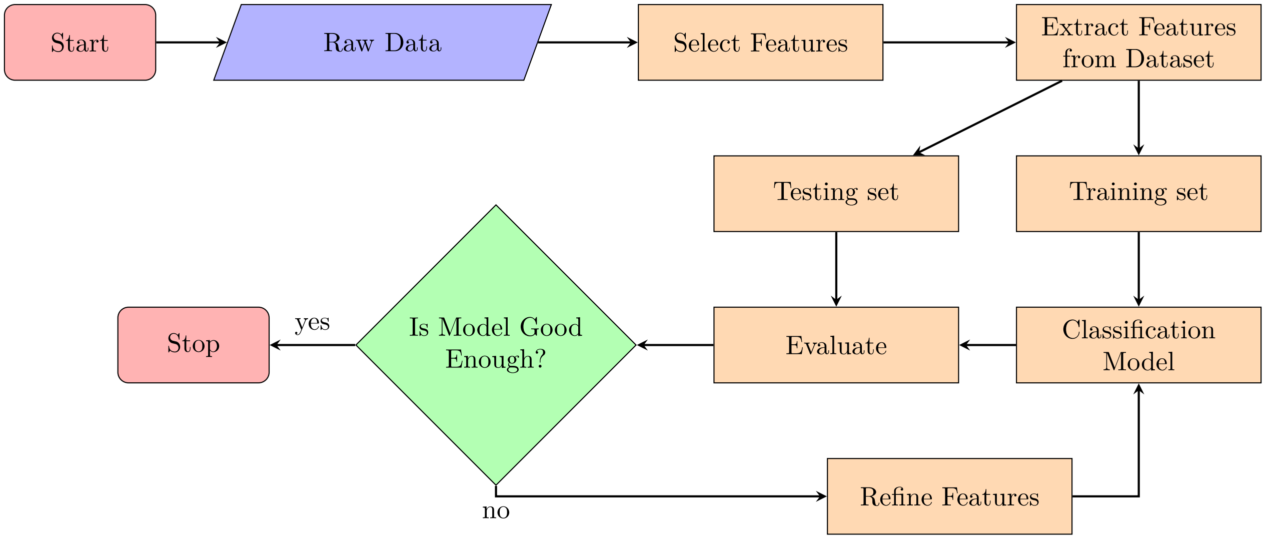

Fig. 1. Flowchart describing the methodology followed during the study

TABLE II
Final feature-list for the dataset

Feature Name	Description	Data Type
titleLength	Number of characters in the title	int
noOfTables	How many tables are present in the paper	int
noOfFigures	How many figures are present in the paper	int
noOfCitations	How many references are cited in the paper	int
noOfRecentCitations	How many of the citations are within 3 years of the paper's year of publishing	int
litRevLength	Number of characters in the Literature Review/Related Works section	int
acl	Whether the conference is ACL or not	1 or 0
arxiv.cs.ai	Whether the conference is in ArXiv's cs.ai or not	1 or 0
arxiv.cs.cl	Whether the conference is in ArXiv's cs.cl or not	1 or 0
arxiv.cs.lg	Whether the conference is in ArXiv's cs.lg or not	1 or 0
conll	Whether the conference is CONLL or not	1 or 0
iclr	Whether the conference is ICLR or not	1 or 0
nips	Whether the conference is NIPS or not	1 or 0
algorithm	Number of times the word comes up in the abstract	int
approach	Number of times the word comes up in the abstract	int
base	Number of times the word comes up in the abstract	int
datum	Number of times the word comes up in the abstract	int
learn	Number of times the word comes up in the abstract	int
method	Number of times the word comes up in the abstract	int
model	Number of times the word comes up in the abstract	int
network	Number of times the word comes up in the abstract	int
problem	Number of times the word comes up in the abstract	int
propose	Number of times the word comes up in the abstract	int

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (5)$$

where, S is the subset of training examples and p_i is the probability of the class.

5) Random Forest: Random forest consists of many decision trees contributing to the prediction that is the modal class. Random forest classifier turns out to be the best classifier, with 10,000 estimators using the entropy method. Since using 10,000 estimators was very resource heavy, we went with 1,000 estimators as the accuracy does not drop very significantly.

6) Support Vector Machine: We used support vector machine with a Radial Basis Function (RBF) kernel as the optimum classifier. Support vector machines work by constructing one or more hyperplanes that can best separate the data.

7) ZeroR: ZeroR does not possess any ability to predict but demonstrate as a baseline for performance. It classifies all data points as the majority class.

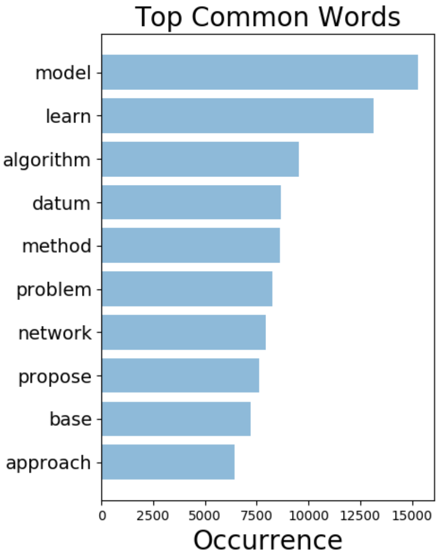

Fig. 2. Most commonly occurring words in abstracts

IV. Results

Different metrics have been used to measure the performance of the models such as accuracy, precision, recall, and F1 scores.

A. Accuracy

One of the most fundamental measures of performance is the accuracy which is defined as the total correct predictions made over the total number of predictions, as seen in Eq. (6):

$$Accuracy = \frac{TP + TN}{FP + TP + FN + TN} \tag{6}$$

where,
FP = False Positive
FN = False Negative
TP = True Positive
TN = True Negative
Often accuracy is measured in percentage (%). Table III demonstrates the accuracy scores for each classifier.

TABLE III
Accuracy Results for Classification Models

Classifier	Training Accuracy	Testing Accuracy
Naive Bayes	78%	79%
K Nearest Neighbor	78%	66%
Logistic Regression	79%	79%
Decision Tree	76%	77%
Random Forest	100%	83%
SVM	63%	63%
Ensemble Voting	83%	80%
ZeroR	61.9%	61.9%

Having applied 10-fold cross-validation on the models showed the performance decreasing slightly. Table IV denotes the accuracy scores and the standard deviations.

TABLE IV
Accuracy Results with 10-fold Cross-Validation

Classifier	Training Accuracy	Testing Accuracy
Naive Bayes	78.1 +/- 0.1%	78.0 +/- 0.7%
K Nearest Neighbor	77.8 +/- 0.2%	70.9 +/- 1.2%
Logistic Regression	78.6 +/- 0.1%	78.6 +/- 0.8%
Decision Tree	79.8 +/- 0.2%	79.1 +/- 0.9%
Random Forest	100 +/- 0.0%	80.5 +/- 0.5%
SVM	63.1 +/- 0.1%	62.9 +/- 0.8%
Ensemble Voting	82.1 +/- 0.1%	79.2 +/- 0.5%

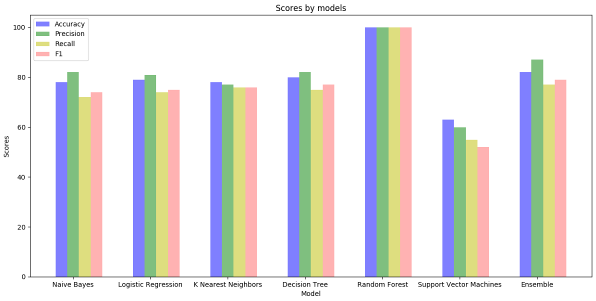

Fig. 3. Training performance of various models

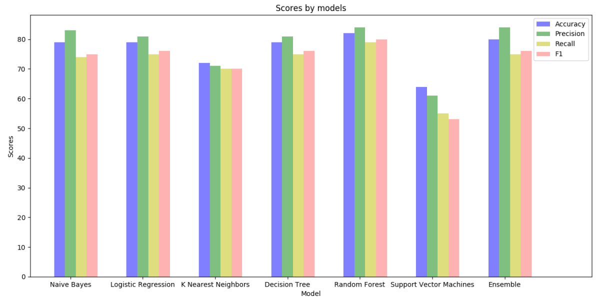

Fig. 4. Testing performance of various models

B. Other Metrics

Accuracy, while important, is not the only metric to judge performance. Tables V and VI as well as Figures 3 and 4 show how the performance of all models vary with recall, precision, and F1 scores during training and testing, respectively.

The formulae for precision, recall, and F1 scores are shown in equations (7), (8), (9), respectively:

$$Precision = \frac{TP}{FP + TP} \tag{7}$$

978-1-7281-9336-6/20/$31.00 ©2020 IEEE

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{9}$$

TABLE V
Training precision, recall, and F1 scores of every model

Classifier	Precision	Recall	F1
Naive Bayes	82%	72%	74%
Logistic Regression	81%	74%	75%
K Nearest Neighbor	77%	76%	76%
Decision Tree	82%	75%	77%
Random Forest	100%	100%	100%
Support Vector Machine	60%	55%	52%
Ensemble Voting	87%	77%	79%

TABLE VI
Testing precision, recall, and F1 scores of every model

Classifier	Precision	Recall	F1
Naive Bayes	83%	74%	75%
Logistic Regression	81%	75%	76%
K Nearest Neighbor	71%	70%	70%
Decision Tree	81%	75%	76%
Random Forest	84%	79%	80%
Support Vector Machine	61%	55%	53%
Ensemble Voting	84%	75%	76%

V. Conclusion and Future work

In this paper, we present machine learning approaches to predict whether an academic paper in the field of AI can will be accepted in top AI conferences or not.

The original model that came with the dataset marked 79% accuracy.Our model on the same dataset shows an improvement in terms of accuracy, precision, recall and F1-score over the previous work.

In future, we plan to work on recommending changes in the manuscript that would improve the acceptance rate in conferences/journals. We also plan to work on identifying the strengths and weaknesses in the paper that affect acceptance rate.

References

[1] (2020) The easychair home page. [Online; accessed February 20, 2020]. [Online]. Available: https://www.easychair.org/

[2] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," Artificial Intelligence in medicine, vol. 23, no. 1, pp. 89–109, 2001.

[3] M. Li and Z.-H. Zhou, "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 37, no. 6, pp. 1088–1098, 2007.

[4] S. Wang and R. M. Summers, "Machine learning and radiology," Medical image analysis, vol. 16, no. 5, pp. 933–951, 2012.

[5] E. Brynjolfsson and A. Mcafee, "The business of artificial intelligence," Harvard Business Review, pp. 1–20, 2017.

[6] D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using spot-5 hrg imagery," Remote sensing of environment, vol. 118, pp. 259–272, 2012.

[7] K. Ahmed, T. R. Shahidi, S. M. I. Alam, and S. Momen, "Rice leaf disease detection using machine learning techniques," in 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE, 2019, pp. 1–5.

[8] J. Y. Chung and S. Lee, "Dropout early warning systems for high school students using machine learning," Children and Youth Services Review, vol. 96, pp. 346–353, 2019.

[9] A. Winkler-Schwartz, V. Bissonnette, N. Mirchi, N. Ponnudurai, R. Yilmaz, N. Ledwos, S. Siyar, H. Azarnoush, B. Karlik, and R. F. Del Maestro, "Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation," Journal of surgical education, vol. 76, no. 6, pp. 1681–1690, 2019.

[10] R. Fiebrink, "Machine learning education for artists, musicians, and other creative practitioners," ACM Transactions on Computing Education (TOCE), vol. 19, no. 4, pp. 1–32, 2019.

[11] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A dataset of peer reviews (peerread): Collection, insights and nlp applications," in Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), New Orleans, USA, June 2018. [Online]. Available: https://arxiv.org/abs/1804.09635

[12] Y. Qian, Y. Dong, Y. Ma, H. Jin, and J. Li, "Feature engineering and ensemble modeling for paper acceptance rank prediction," arXiv preprint arXiv:1611.04369, 2016.

[13] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," Computers & Education, vol. 53, no. 3, pp. 950–965, 2009.

[14] S. B. Kotsiantis, C. Pierrakeas, and P. E. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in International conference on knowledge-based and intelligent information and engineering systems. Springer, 2003, pp. 267–274.

[15] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades," Artificial Intelligence Review, vol. 37, no. 4, pp. 331–344, 2012.

[16] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," in 2014 IEEE International Advance Computing Conference (IACC). IEEE, 2014, pp. 549–554.

[17] J. Xu, K. H. Moon, and M. Van Der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 5, pp. 742–753, 2017.

[18] A. Nurhudatiana and A. Anggraeni, "Decision tree modeling for predicting research productivity of university faculty members," in 2015 International Conference on Data and Software Engineering (ICoDSE). IEEE, 2015, pp. 70–75.

[19] A. E. S. S. T. Ghosal, R. Sonam and P. Bhattacharyya, "Is the paper within scope? are you fishing in the right pond?" in ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2019, pp. 237–240.

[20] G.-M. Guo, "Decision support system for manuscript submissions to academic journals: An example of submitting an enterprise resource planning manuscript," in 2017 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 2. IEEE, 2017, pp. 558–562.

[21] Science parse parses scientific papers (in pdf form) and returns them in structured form. [Online]. Available: https://github.com/allenai/science-parse

978-1-7281-9336-6/20/$31.00 ©2020 IEEE