iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.36180

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI June 2021- Available at www.ijraset.com

Machine Learning Based Essay Grading System

Ojasvi Daga[1], Abhishek Abhyankar[2], Kshitij Kamat[3], Prof. Vishal Mamde[4]
[1, 2, 3, 4]*School of Electronics and Communication Engineering MIT World Peace University, Pune*

Abstract: ***Machine Learning and automation has progressed immensely over the years and has tend to make human lives simpler with reducing human effort and time on tasks by enabling a machine to perform them. One such task is to grade essays. Essay writing is an integral part for anyone willing to learn a language or skill or to simply exhibit one's thoughts and ideas on a topic. This leads us to the reason why essay grading is important. When a work is scored against some parameters, a scope of improvement is possible. Hence, when essays are graded and feedbacks are provided, it guides the writer to analyse the work and to have a better understanding of the topic in general. Although, manual grading of essays could create discrepancy because of being graded by different individuals having different perceptions of the same content. It also consumes a lot of human time and effort. Therefore, automatic grading of essays could prove to be the saviour. In this project, we build a machine learning model which grades essays based on various features extracted using Natural Language Processing. We also test the model's performance using several regression models like Linear, Lasso, and Ridge, and methods like Artificial Neural Network to find the best fit giving the maximum correlation with human grades.***

LIST OF CONTENTS

LIST OF FIGURES

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI June 2021- Available at www.ijraset.com

I. CHAPTER I

A. *Introduction*

The impact that computers have on our writings have been in use for 40 years now. Even the most basics of computers like processing of words, is of great help to authors in updating their writing material. Although, the computers have the capability to function as a more effective cognitive tool. Revision and feedback are important parts of writing material. Essays are crucial testing tools for assessing academic achievement, integration of ideas and ability to recall. However, students in general require input from their teachers to master the art of writing. Manual grading of essays takes up a significant amount of the instructors' valuable time, and hence making it an expensive process. Automated grading, if proven effective, will not only reduce the time for the assessment but will thus significantly reduce the costs.

B. *Review of Related Literature*

1) *Project Essay Grader (PEG):* Project Essay Grade (PEG) is one of the earliest implemented (1996) automated essay grading system [1]. This scoring system was developed to make essay scoring more practical and effective and it relies on style analysis of surface linguistic features of a text block. PEG does not use NLP approach but instead uses proxy measures to predict the intrinsic quality and computer approximations. As no Natural Language Processing (NLP) is used so it does not take lexical content in account. Proxies refers to the structure of essay which consist of average word length, essay length, number of semicolons or commas, counts of preposition, parts of speech and so on. One of the best things about PEG is that PEG's predicted scores are in close agreement to those of human raters. It achieved a correlation score of 0.87 with human raters [1]. Second is, this system is computationally tractable which means it can track the errors made by the users. However, PEG purely relies on a statistical approach based on the assumption that the quality of essays is reflected by the measurable proxies, thus has been criticized for not including semantic features of the essays and only focusing on the structure. Since PEG uses structure of essay to score it was easy to cheat by increasing the length of the essay.
2) *Intelligent Essay Assessors (IEA):* IEA uses Latent Semantic Analysis (LSA) which is a computational model of human knowledge representation [2]. It is also a method for extracting semantic similarity of words and passages from text. It is based on statistical analysis of large amount of text (typically thousands to millions of words) and focuses on conceptual content, that is the knowledge conveyed in an essay. LSA approach was able to perform at the same reliability level as the trained ETS graders with approximately 0.87 correlation. The biggest advantage of IEA is that it can flag those essays that are off topic, so that it becomes easy for graders to grade.
3) *Electronic Essay Rater (E-Rater):* ETS developed e-rater has been used for scoring essays since Feb'1999 and has been continuously upgrading with newer versions [3]. It generates advisory flag when it encounters anomalous essay & writing samples such as exceeding length, repetition material & off topic response. It uses 10 features to evaluate the essay, that are - grammar, mechanics, style, usage, organization, development, word choice, word length, positive features, differential word use. E-rater has up to 97% agreement with human raters. It follows Multiple Linear Regression methodology, but can also use Support Vector Machines, Random Forest, and K-Nearest Neighbours. SVM performs better but MLR gives the basis of score for each feature, hence preferably used.
4) *Linear Regression:* Linear regression is an important tool for statistical analysis. Its broad spectrum of uses includes relationship description, estimation, and prognostication [4]. Linear regression is used to study the linear relationship between a dependent variable Y and one or more independent variables X.
5) *Lasso Regression:* LASSO (Least Absolute Shrinkage and Selection Operator) regression is a type of linear regression that uses shrinkage [5]. It a shrinkage and variable selection method for regression models. LASSO regression aims to identify the variables and corresponding regression coefficients that lead to a model that minimizes the prediction error. In practical sense, it constrains the complexity of the model.
6) *Support Vector Regression (SVM):* Support Vector Regression uses the same principle as Support Vector Machines, but for regression problems. In this method, the points that are within the decision line boundary are considered. This model acknowledges the presence of non-linearity in the data [6].
7) *Ridge Regression:* Ridge Regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. Ridge regression solves the problem of overfitting which might occur in a linear regression model [7].
8) *Long Sort-Term Memory (LSTM):* Long Short-Term Memory is a type of Artificial Recurrent Neural Network which is used in the field of deep learning. It can process data sequentially and it keeps hidden state through time. Since the traditional RNN

tends to forget its previous extracted features (vanishing mode), another long-term state is introduced to append the data into the memory which was previously forgotten [8].

II. CHAPTER II

A. *Objectives of the Project*

The aim of this project is to train and develop a Machine Learning model to grade essays. This will be a supervised ML model that will assess essays and automatically grade them. The basic idea is to search for features which can model the attributes such as vocabulary, grammar, structure, relevance, content etc, and design a system to assess and grade the essays.

The model will have three parts –

1) Learning the input essays (Machine Learning).
2) Extracting its features using Natural Language Processing.
3) Scoring the essays using various Regression Models.

B. *Scope of the Project*

Automatic essay grading is a very useful machine learning application. It has been studied several times, using various techniques like latent semantic analysis etc. The current approach tries to model the language features like fluency, grammatical correctness, domain information content of the essays, and tries to fit the best polynomial in the feature space using various regression methodologies.

III. CHAPTER III

A. *Block Diagram*

We start with collecting the dataset of graded essays and splitting it into 70-30 to begin training our Machine learning model. The training set will extract features such as spellings, word length, vocabulary, nouns, adjectives, adverbs, length of the essay etc.

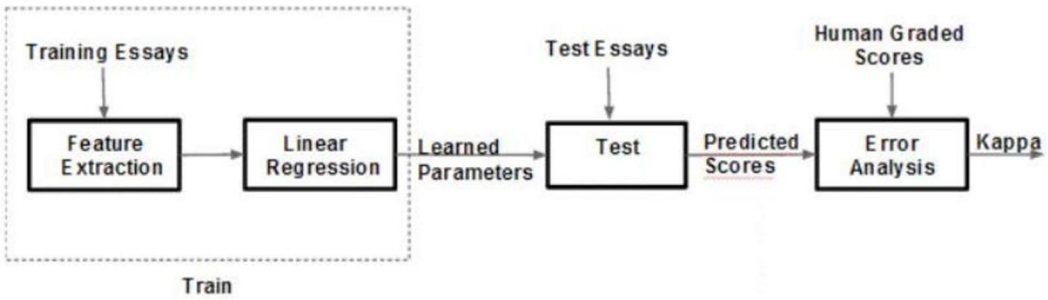


Fig 3.1 Implementation Methodology

We assign weights to the features based on the scores of the essays trained. We then test the model on the remaining 30% of the essays by comparing their features to that of the trained essays and then grade them. Finally, to verify the accuracy of the model, we can verify the predicted scores with the actual grades from the dataset.

IV. CHAPTER IV

A. *System Design and Methodologies*

1) *Feature Extraction*

a) Bag of Words (BOW)
b) Number of characters in an essay
c) Number of words in an essay
d) Number of sentences in an essay
e) Average word length of an essay
f) Number of lemmas in an essay
g) Number of spelling errors in an essay

h) Number of nouns in an essay
i) Number of adjectives in an essay
j) Number of verbs in an essay
k) Number of adverbs in an essay
l) Relevance of content with the topic

2) Parameters Used to Judge Model Credibility

a) Mean Squared Error: In statistics, the mean squared error of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

b) Variance: The variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean. Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean.

c) Cohen Kappa Score: It measures the agreement between two raters. It is a quantitative measure of reliability for two raters that are rating the same thing, corrected how often that the raters may agree by chance.

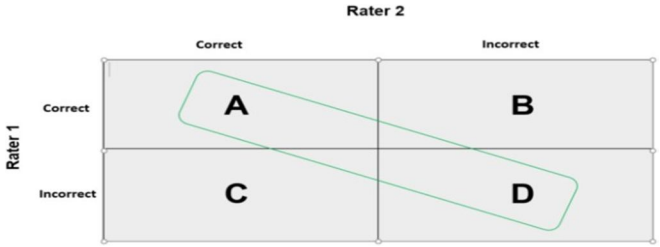


Fig 4.1 Cohen Kappa Score

- Po = Number in agreement / Total
- Pcorrect = (A+B/Total) * (A+C/Total)
- Pincorrect = (C+D/Total) * (B+D/Total)
- Pe = Pcorrect + Pincorrect

Thus, Cohen Kappa coefficient was calculated using - K = Po – Pe / 1 – Pe

B. Implementation, Testing and Debugging

1) Exploratory Data Analysis on the Data

The following are the plots of the relationships between the individual features extracted and the domain scores.

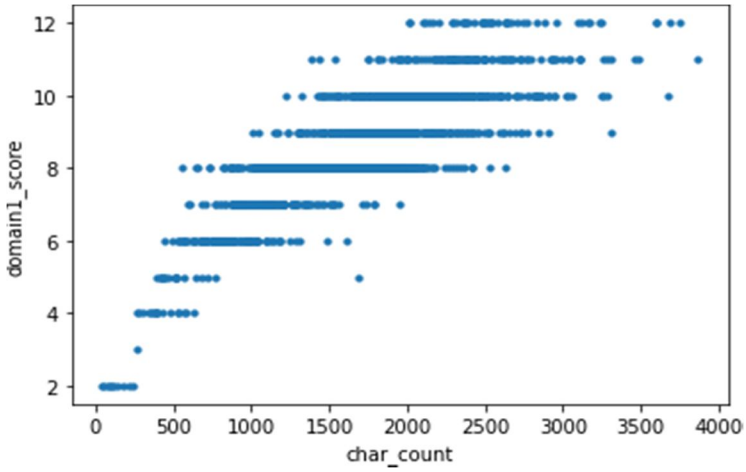


Fig 4.2 Character Count Scatter Plot

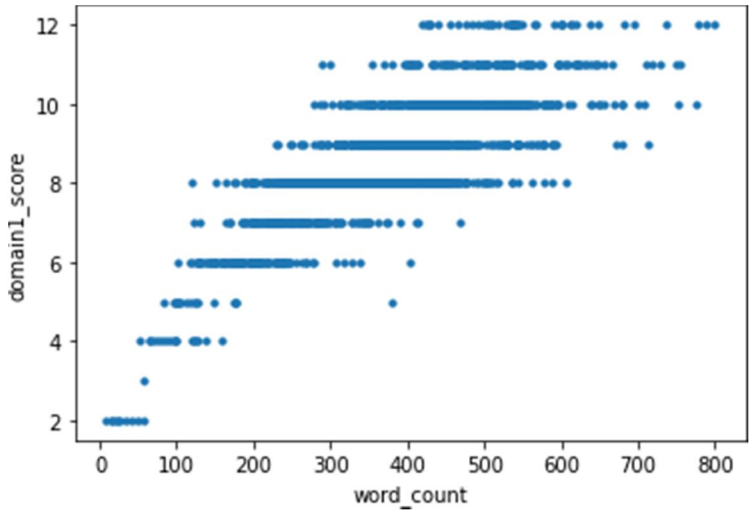


Fig 4.3 Word Count Scatter Plot

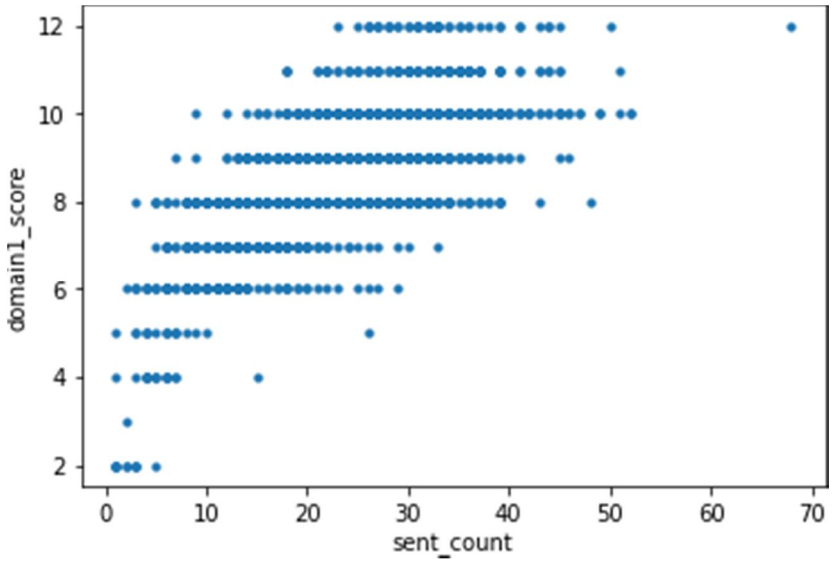


Fig 4.4 Sentence Count Scatter Plot

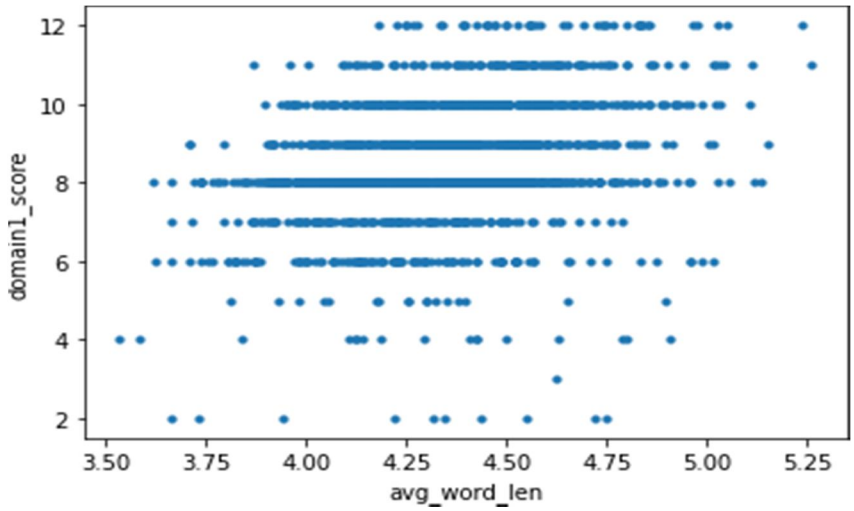


Fig 4.5 Average Word Length Scatter Plot

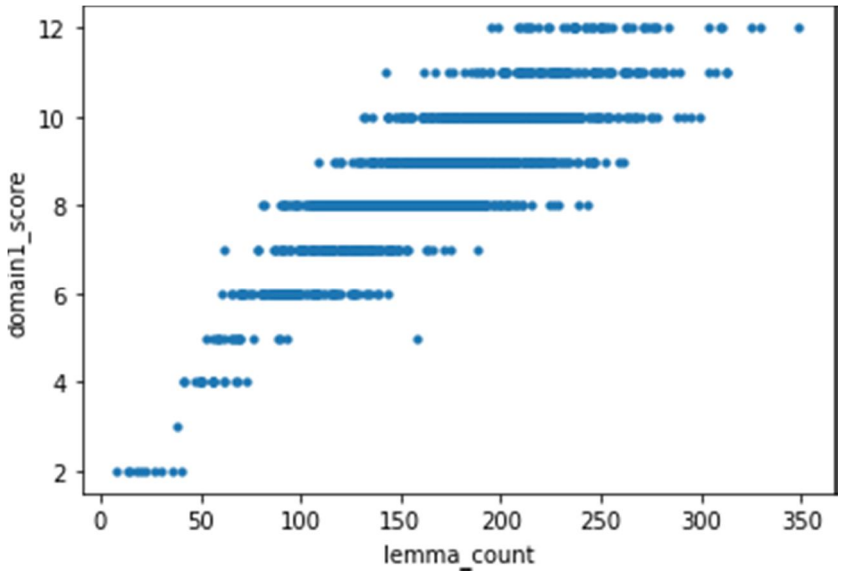


Fig 4.6 Lemma Count Scatter Plot

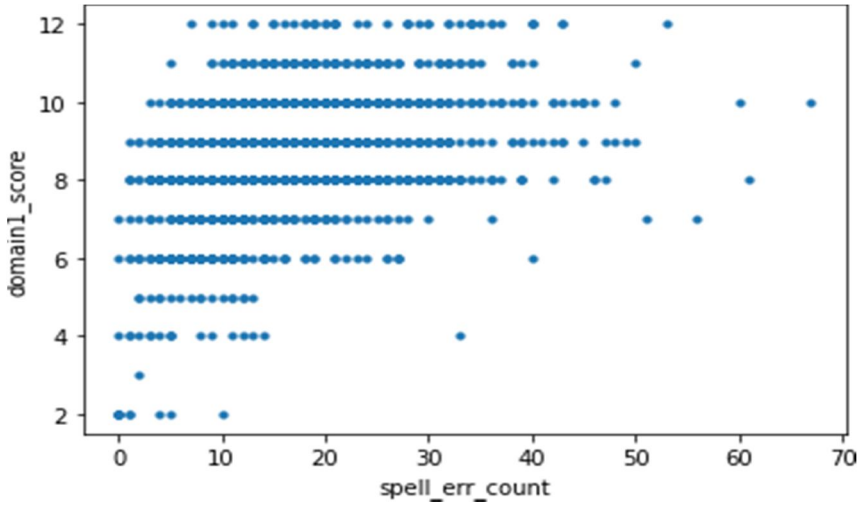


Fig 4.7 Spelling Error Scatter Plot

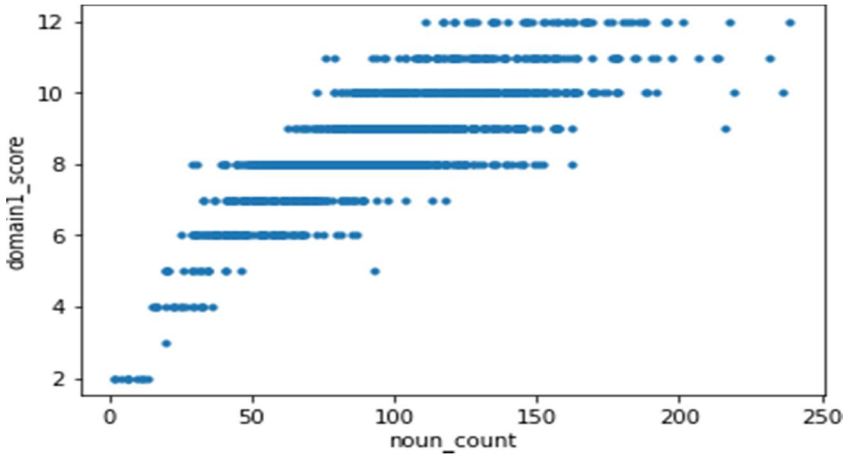


Fig 4.8 Noun Count Scatter Plot

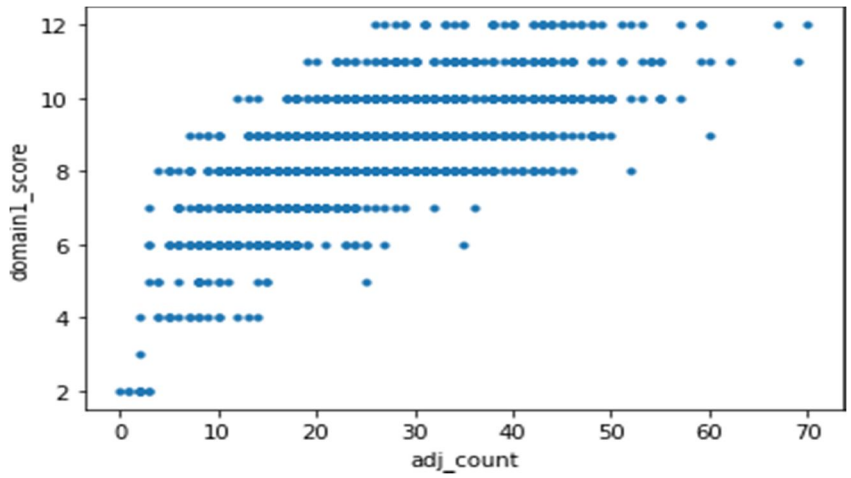


Fig 4.9 Adjective Count Scatter Plot

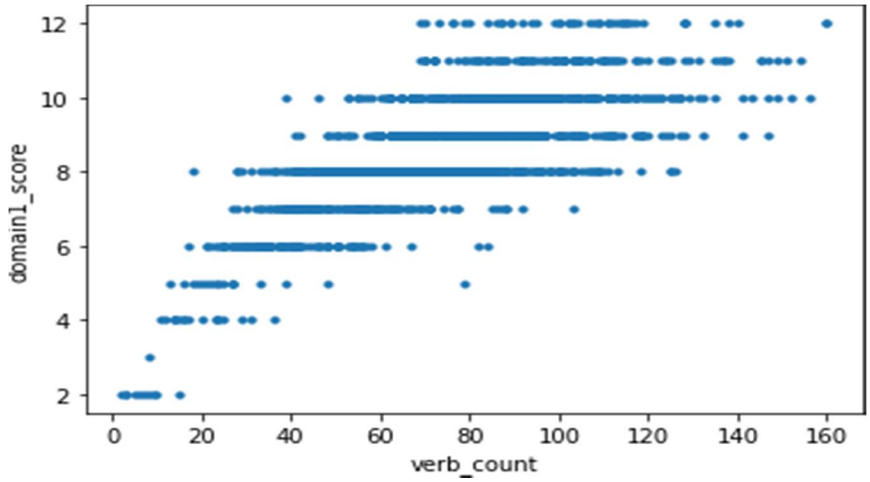


Fig 4.10 Verb Count Scatter Plot

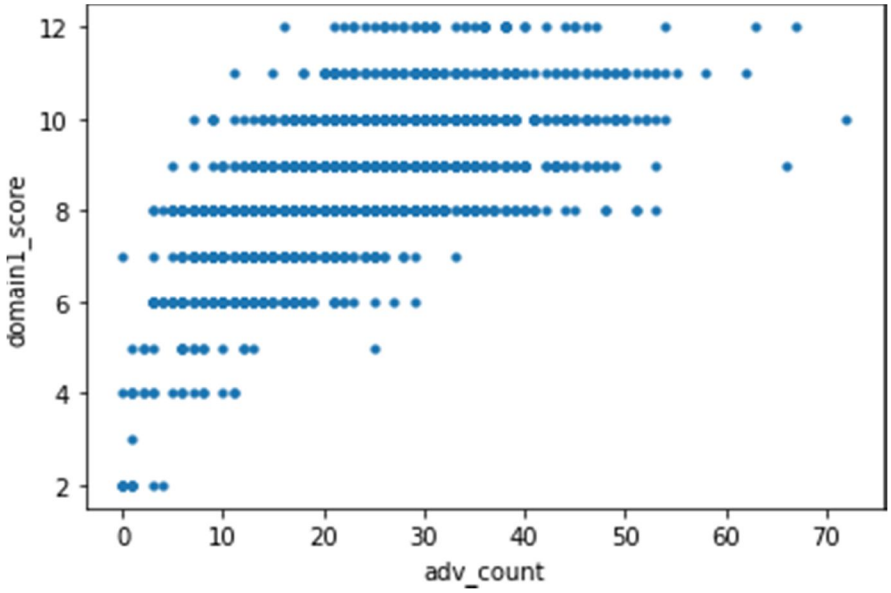


Fig 4.11 Adverb Count Scatter Plot

V. CHAPTER V

A. *Result Analysis*

We successfully trained different machine learning models and observed their performances to see which one had a better performance.

Feature	Models	Mean Squared Error	Cohen Kappa Score
Using Only Bag of Words	Linear Regression	1.92	0.14
	Lasso Regression	1.11	0.25
Using Bag of Words + All Features	Lasso Regression	0.67	0.4
	Support Vector Regression	1.26	0.24
	Ridge	1.23	0.22
	LSTM	0.86	0.93

Fig 5.1 Result Analysis

B. *Conclusion*

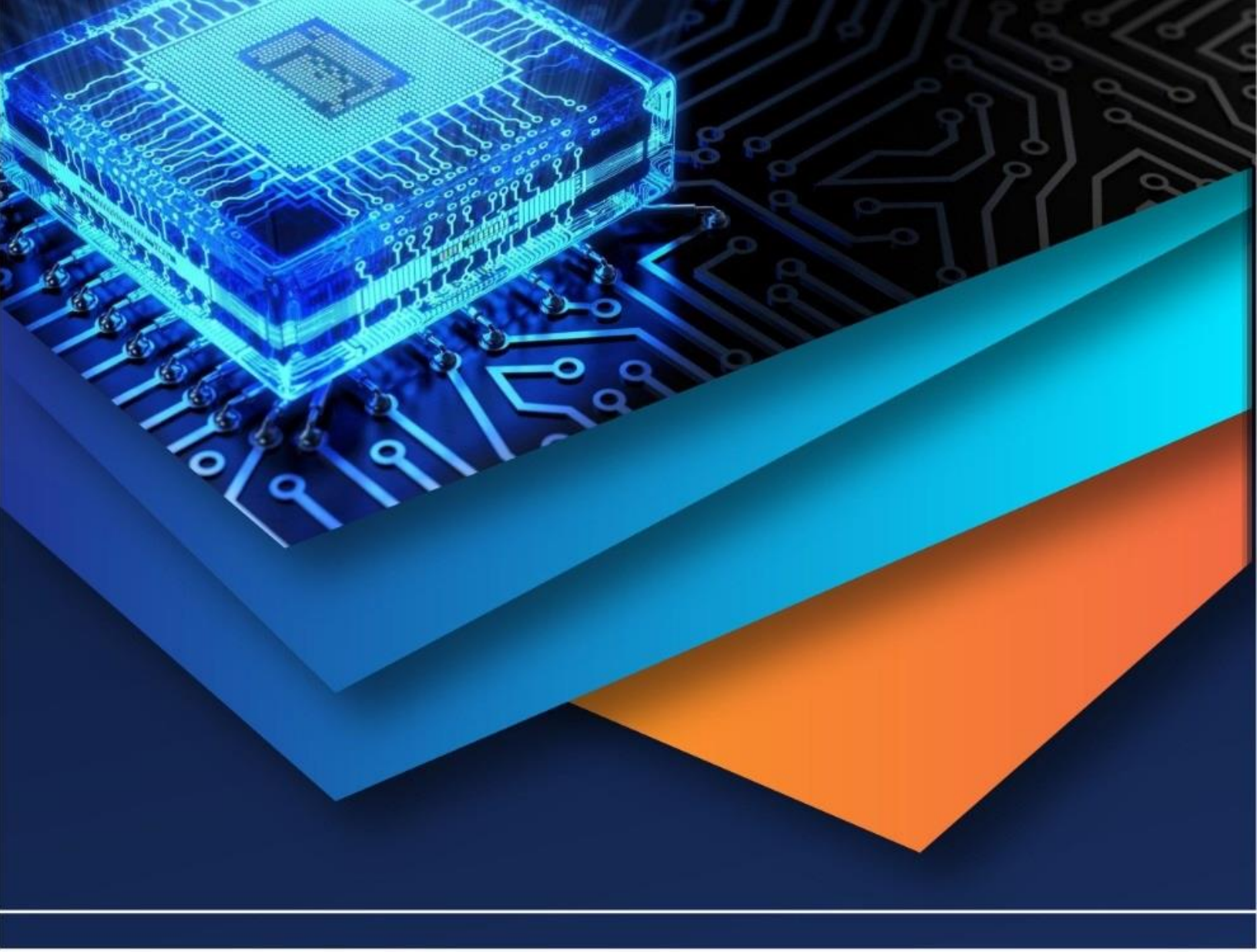
After observing these models, we were able to conclude that the Long Short-Term Memory Model using BOW + all features gave the best results as it has the highest Cohen Kappa Score and a comparatively low mean squared error.

C. *Future Scope*

The given problem can extend in various dimensions. One such area is to search and model good semantic and syntactic features. For this, various semantic parsers etc. can be used. Other area of focus can be to come up with a better tool than linear regression with polynomial basis functions like neural networks etc.

REFERENCES

[1] S. Dikli, “Automated Essay Scoring”, Florida State University, Tallahassee.
[2] Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998.
[3] V. Salvatore, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading.", Journal of Information Technology Education: Research 2.1 (2003):319-330, 2003.
[4] Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 ofa series on evaluation of scientific publications. Deutsches Arzteblatt international, 107(44), 776–782.
[5] Ranstam, J., and J. A. Cook. "LASSO regression." Journal of British Surgery 105, no. 10 (2018): 1348-1348.
[6] Awad, Mariette, and Rahul Khanna. "Support vector regression." In Efficient learning machines, pp. 67-80. Apress, Berkeley, CA, 2015.
[7] Dr. Samira Muhammad Salh, “Using Ridge Regression Model to solving Multicollinearity Problem.” International Journal of Science and Engineering Research, Volume 5, Issue 10, October-2014, ISSN 229-5518.
[8] Alex Sherstinsky, “Fundamentals of Recurrent Neural Networks (RNN) and Long Short- Term Memory (LSTM) Network.” in the Elsevier journal “Physica D: Nonlinear Phenomena”, Volume 404, March 2020: Special Issue on Machine Learning and Dynamical Systems.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY