ABSTRACT

Manual essay scoring can be a daunting process for human evaluators, assessing descriptive answers can present a huge overhead owing to their limited numbers and the proportional number of essays to be graded hence leading to an inefficient or an inaccurate score. The recent advance in natural language processing technology in artificial intelligence provides a way to score essays automatically.

Implementing an automated essay scoring system helps reduce manual workload and speed up learning feedback. However, Existing systems for AES are typically trained to predict the score of every single essay at a time without considering the rating schema. One of the reasons is that the various kinds of rating criteria are very hard to represent. To address this issue, we propose a machine learning framework. The automated essay scorer (AES) will provide consistent and objective assessments and can approach human rater's assessments when developed using the above method.

In this Project Report, we create an automatic essay scorer (AES) by using 2 -layered Long Short-Term Memory (LSTM) network and word embedding and take the Automated Student Assessment Prize (ASAP) dataset for training and evaluation. This shows an automated system that can rate essays in electronic text. We combine manually crafted features and Word2Vec embedding in training the model, which makes it more interpretable. We carefully tune the hyperparameter to improve the precision of our model. The LSTM network reaches a quadratic weighted Kappa score (QWK) of 0.92, which outperforms many other rating systems. Moreover, because of the major shift in paradigm from traditional classroom education to online education engendered by the COVID-19 pandemic. It seems plausible to infer that future assessment of education shall be online, making our solution of automatic essay scorer not only relevant but of paramount importance.

1. INTRODUCTION

The chapter of introduction deals with the introduction to terminology and defines the scope of the project. It deals with problem definition and methodologies required and outline of results.

1.1 Problem Definition including the significance and objective

Manual Grading essays can be a tough work for human evaluators; examining descriptive responses can be time-consuming due to their limited number and the proportional number of essays to be scored, resulting in an inefficient or erroneous score. As a result of the outbreak, there has been a significant shift in paradigm from traditional classroom education to online education. It appears reasonable to assume that future educational evaluations would be conducted online, making our automatic essay scorer solution not just relevant but significant. The use of specialized computer systems to award grades to essays written in an educational setting is known as automated essay scoring (AES). It is a type of educational assessment as well as a natural language processing application. Its goal is to divide a vast number of textual elements into a small number of discrete categories that correspond to the various grades.

1.2 Methodologies

Automatic Essay scoring is done with the help of natural language processing and deep learning. NLP is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised.

The proposed system's key strength is that it combines novel graph-based semantic characteristics with syntactic and emotive features to increase the accuracy of existing systems while reducing the overall number of features employed. Various prediction

models are put to the test in order to determine which one works best for predicting essay scores. To process the essay, we follow the steps below.

The order/arrangement of content is referred to as syntax. Syntactic categories are word classes that roughly correlate to traditional parts of speech (e.g., noun, verb, preposition, etc.). Phrasal categories (e.g., noun phrase, verb phrase, prepositional phrase, etc.) are also syntactic categories in phrase structure grammars. A submitted essay is divided into sentences and given full stops in order to examine the syntax. These sentences are then tokenized into words in order to examine each word in the essay. The purpose of syntactic analysis is to determine the structure of the input text. This structure consists of a hierarchy of phrases, the smallest of which are the basic symbols and the largest of 2 which is the sentence. To analyze the syntax of the essay we will extract 12 features which describe the syntactic structure of the essay.

The study of meaning is also known as semantics. Any language's primary goal is to allow people to transmit meaning to one another. The semantic properties of NLP are crucial indicators of its meaning. Semantic analysis is a technique for extracting and describing the context meaning of a single word or a group of words. To earn good grades, even a well-structured and grammatically correct essay must be well-coordinated. To examine the semantics, we must ensure that the essay's content accurately reflects the question prompt and that the content flow in the essay's sentences is meaningfully related. Every component of an essay must be semantically compatible. The incoherence of a section of an essay with other sub-sections suggests that that section is disconnected from the remainder of the essay. The essay's organization around a primary unifying theme is a key criterion for essay grading. To check for coherence, compute semantic similarities between an essay's sentences (comparisons between consecutive pairs are insufficient), and do a semantic comparison between each sentence in the essay. To assess the essay's overall coherence, it is proposed that semantic similarity be computed not only between successive claims but also between each pair of sentences. The similarities between all of the phrases allow us to create new Graph-based characteristics that show how distinct essay sentences are semantically connected and their patterns. To determine how semantically related the entire essay and the inquiry prompt are, semantic similarity is computed. Because each essay is produced in answer to a question, there should be some semantic overlap between them. To compute semantic similarity

between the essay response and the question prompt, the semantic similarity algorithm is proposed.

Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques . Sentiment Analysis is the study of people's opinions, attitudes, and emotions towards a specific person or thing. Each writer has a distinct or shifting tone when it comes to the issue at hand. If a writer wants to write about his disagreement with a scenario, for example, this analysis will tell us how negative the language employed is, as well as how positive or neutral the tone of the writer is. In AES, sentimental analysis is very significant.

1.3 Outline of the results

The final result of this application is the score of the student submitted answer. The application evaluates the student's answer by considering all these aspects the score is predicted by the prediction model.

The length of selected essays varies from 150 to 550 words per response. Some of the writings rely on information from sources, while others do not. All of the responses were written by children in grades 7 through 10, and they ranged in age from 7 to 10. All essays were double-scored and assessed by hand.

The deep learning prediction model takes the input as features that were calculated using the above mentioned analysis and extracts important features and the score is predicted. This model provides 0.92 kappa score using 5-Fold Cross Validation

1.4 Scope of the project

The main scope of this project is to focus on the automation of easy scoring, and to equip teachers and educational institutions to focus on automation of this task. Teachers all throughout the world devote a significant amount of time to marking their students' work. As a result, they must reduce the amount of time they can devote to their other responsibilities. Even yet, they may lack the time necessary to thoroughly assess the large number of students they have. As a result, many authors believe that this problem must be addressed, and some have proposed the computer as

a new assessment tool. These authors are not attempting to replace instructors with computers, but rather to assist teachers with computer software. Although the technological approaches taken by these tools are extremely varied, the purpose and concepts underpinning them are all the same. On the other hand, the thought of a computer judging human essays has always been met with skepticism.

There are still some skeptics among scholars who do not believe that automatic grading is feasible. However, developments in natural language processing, machine learning, and neural network techniques, a lack of time to provide proper feedback (despite the widespread belief that it is important), and the belief that MCQs are a poor evaluation tool all point to a change in this situation. Automatic assessment of students' texts can be thought of as the top level of a hierarchy that includes two subcategories: automatic assessment of short responses and automatic assessment of essays. Although the same tool can sometimes evaluate both types of writing, the distinctions between the two jobs are evident, and most computer automated evaluation programmes only analyze essays or short pieces of writing.

The second subcategory, automatic assessment of essay answers, is the focus of this research. Style refers to the syntax, mechanics, diction, and other aspects of the writing, whereas content refers to what the essay says. Some researchers are vehemently opposed to this classification because they believe that in order to evaluate a text, both content and style must be considered, and that the former should not be considered without the latter. These Computer automated evaluation software applications seek for direct aspects in the text, such as word number, word lengths, or adjective use, and transform them into more abstract measurements, such as variety, fluency, or quality, in order to grade the style. Several automated assessment systems for evaluating essay content have lately appeared, and some of them are even commercially available. Furthermore, numerous traditional examinations, such as the Graduate Management Admissions Test (GMAT), the Test of English as a Foreign Language (TOEFL), and the Graduate Record Examination (GRE), include open-ended questions delivered through computer, which may allow the use of automated scoring methods. As a result, this is a burgeoning topic that attracts the attention of scholars, educators, and business leaders, and it has a bright future.

1.5 Organization of the report

This introduction section is followed by the Literature Survey. The literature survey explains the current existing systems which are commonly based on automated essay evaluation. It also introduces domain specific terminology which forms the background to understand this project. It discusses in depth about some existing solutions' core aspects which also forms the basis for many other solutions. The section also discusses the drawbacks in all the solutions exhaustively. The literature survey section is followed by the Design of the Proposed System section. This section discusses the evolution and design of the proposed solution. Next section discusses the implementation of the design discussed in the previous section.

The Data Flow Diagrams and Flowcharts are discussed in this section of the project. The algorithm is also discussed in this section. The data set being used, the features of the data set, and their significance are mentioned. The testing process is also included. The next section deals with the result analysis. The system is executed over the test cases and the results are analyzed and discussed. The final section deals with limitations and recommendations. The references showing the base papers used in this project are then mentioned.

2. LITERATURE SURVEY

Literature survey deals with introduction to problem domain, the existing works and related work to the project.

2.1 Introduction to the problem domain terminology

There may be a recent surge of hobby in neural networks, that are primarily based on non-stop-area illustration of the enter and nonlinear features. Therefore, neural networks are able to model complicated patterns in records. Furthermore, when you consider that these techniques no longer depend on manual engineering of features, they can be applied to solve issues in a quit-to-stop style. Essay writing is usually part of the student assessment method. Several organizations, inclusive of instructional testing provider (ETS)1 , evaluate the writing competencies of college students in their examinations. Because of the big variety of college students participating in these tests, grading all essays may be very time consuming.

Accordingly, a few businesses were using AES systems to reduce the time and cost of scoring essays. Computerized essay scoring refers to the system of grading scholar essays without human interference.

An AES device takes as input an essay written for a given spark off, and then assigns a numeric rating to the essay reflecting its pleasantness, based on its content, grammar, and enterprise. Such AES structures are normally based on regression methods carried out to a hard and fast of carefully designed functions.

The method of feature engineering is the toughest part of constructing AES systems. Furthermore, it's hard for human beings to recollect all the factors which can be involved in assigning a rating to an essay.

Famous strategies include the use of word embeddings to seize semantic residences of words, and an increase in stop-to-end mastering of a better-level task (e.g., query answering) in preference to relying on a pipeline of separate intermediate responsibilities (e.g., part-of-speech tagging and dependency parsing).

In a few areas, this shift has entailed significant modifications in how NLP systems are designed, such that deep neural community-based total methods can be viewed as

a new paradigm wonderful from statistical natural language processing. As an example, the term neural system translation (NMT) emphasizes the truth that deep getting to know-based techniques to machine translation directly learn sequence-to-series modifications, obviating the need for intermediate steps together with phrase alignment and language modeling that was utilized in statistical machine translation (SMT).

2.2 Existing Systems

Existing systems describe the current models that are used for commercial or research purposes in the real world. They are as follows

2.2.1 ETS

The e-rater automated writing assessment engine is ETS's patented capability for automatic assessment of expository, persuasive and summary essays. more than one evaluation application uses the engine.

The engine is utilized in combination with human raters to attain the writing sections of the TOEFL iBT and GRE assessments. The e-rater engine is also used as the only rating in getting to know contexts, which includes formative use in a study room putting with ETS's Criterion® online essay assessment machine.

Inside the Criterion software, the engine is used to generate individualized remarks for students, addressing an increasingly more important want for automated essay assessment that is dependable, legitimate, rapid and flexible.

The e-rater engine functions associated with writing best encompass:

▪ Mistakes in grammar (e.g., subject-verb agreement)

▪ Utilization (e.g., preposition selection)

▪ Mechanics (e.g., capitalization)

▪ Fashion (e.g., repetitive word use)

▪ Discourse shape (e.g., presence of a thesis assertion, major factors)

▪ Vocabulary usage (e.g., relative sophistication of vocabulary)

▪ Sentence range

2.2.2 IntelliMetric

Analytical Writing Evaluation (AWA) portion of the Graduate management Admission check (GMAT(R)) with its patented IntelliMetric®(R) technology. IntelliMetric rankings candidates' responses to 2 essay prompts inside each GMAT(R) AWA exam. beneath a separate agreement with ACT, IncPearson VUE, and the Graduate control Admission Council (GMAC(R)), ACT will oversee GMAT AWA activate improvement and scoring.

ACT gives human expert essay scoring, which will be blended with Vantage mastering's IntelliMetric platform to ensure the most correct, reliable rankings.

2.2.3 C - Rater

C-Rater is an automated scoring engine that has been developed to attain responses to content material based essay answer questions. It isn't always in reality a string-matching software,as an alternative it makes use of predicate argument shape, pronominal reference, morphological analysis and synonyms to assign full or partial credit score to a short solution question. Crater has been utilized in two studies:

national evaluation for educational development (NAEP) and a statewide assessment in Indiana. In both research, C-Rater agreed with human graders approximately 84% of the time.

C-Rater’s goal is to map the student’s response onto the version, and in so doing to demonstrate the correctness of the reaction or, failing that, its incorrectness or inadequacy.

The model is built by hand however the mapping is fully computerized.due to the fact a version is needed, the query has to have an unmarried accurate answer or more than a few accurate solutions. because of this C Rater is not designed to score open-ended

questions, including ones that ask for examples taken from non-public experience, or for an opinion, or for revolutionary methods to resolving a war.

To attain essay answer responses, the scoring engine must be able to understand while an idea is expressed and when it isn't always. We consider the set of accurate responses as being paraphrases of the precise answer, and of the C-Rater scoring engine as a paraphrase recognizer that identifies the contributors of this set. it's vital to notice that C-Rater is not certainly matching words ,the paraphrases should obey syntactic constraints.

A question may be scored by C-Rater if there may be a finite range of standards that fulfill it. As a consequence, an open-ended query asking for an opinion or for examples from the student's personal experience is not a question for C-Rater.

2.2.4 CarmelTC

Another in advance system is CarmelTC by way of Carolyn P. Rose, Antonio Roque, Dumi sizwe Bhembe, and Kurt VanLehn. It has been designed as an element inside the hy2 educational talk gadget. Despite the fact that Rose et al role CarmelTC inside the context of essay grading, in their data, the average period of a student reaction is approx. 48 words. Their gadget is designed to perform textual content category on unmarried sentences within the scholar responses, where every magnificence of textual content represents one viable version reaction, plus an extra elegance for 'no suit'.

They integrate decision timber running on an automatic syntactic evaluation, a Naive Bayes textual content classifier, and a bag-of-phrases technique. In a 50-fold move validation experiment with one physics question, six lessons and 126 scholar responses, hand-tagged through annotators, CarmelTC reaches an F measure price of 0.85. They do not record on a baseline. Regarding the first-rate of the gold wellknown, they file that conflicts in the annotation have been resolved.

2.2.5 E-Rater (Electronic Essay Rater)

E-Rater was developed by means of Burstein and others (Burstein, Kukich, Wolff, Chi,& Chodorow, 1998; Burstein, Leacock, & Swartz, 2001). E-Rater makes use of the NLP device for parsing all sentences within the essay. E-Rater uses a mixture of statistical and NLP techniques to extract linguistic capabilities from the essays to be graded.

Essays are evaluated towards a benchmark set of human graded essays. With E-Rater, an essay that stays on the topic of the question, has a robust, coherent and nicely-organized argument shape, and presentations a diffusion of phrase use and syntactic shape will acquire a score at the better end of a six-point scale. E-Rater features include the evaluation of the discourse structure, of the syntactic structure and of the vocabulary utilization (domain evaluation). E-Rater adopts a corpus-based total approach to version building by using the use of actual essay statistics to analyze the capabilities of a sample of essay responses.

The utility is designed to become aware of functions in the text that reflect writing features laid out in human reader scoring standards and is presently composed of 5 most important independent modules.

3 of the modules perceive capabilities that may be used as scoring guide criteria for the syntactic variety, the agency of thoughts and the vocabulary usage of an essay. A fourth impartial module is used to pick and weigh predictive features for essay scoring. sooner or later, the final module is used to compute the final score.

E-Rater, consequently offering extra comments approximately features of writing related to subject matter and fluency handiest. E-Rater is far more complicated and requires more training than many others to have structures. Moreover, no demonstration and no downloadable trial version of E-Rater had been made available to the clinical network.

2.2.6 Intelligent Essay Assessor

Intelligent Essay Assessor (IEA) IEA become developed in the late Nineties (Hearst, 2000; JerramsSmith, Soh, & Callear, 2001) and is primarily based at the Latent Semantic analysis (LSA) technique that was at the start designed for indexing documents and textual content retrieval (Deerwester, Dumais, Landauer, Furnas, &Harshman, 1990. Key functions of IEA include exceptionally low unit cost, short customized feedback, and plagiarism detection.

Moreover, the authors declare that the gadget is very well applicable to analyze and rating expository essays on subjects together with technological know-how, social research, history, medicinal drug or commercial enterprise, but no longer appropriate to evaluate real expertise.

IEA automatically assesses and evaluations electronically submitted text essays, and represents a useful domain-independent tool. It supplies instant comments at the content material and the first-rate of the student's writing.

2.2.7 Indus Marker

Indus Marker is designed for genuine answers where there may be a clear criterion for solutions being right or wrong. The device is based on shape matching, i.e. matching a pre-detailed structure, advanced through a purpose-built structure editor, with the content of the scholar's answer text.

An examiner specifies the required shape of a solution in a simple purpose designed language referred to as query answer Markup Language (QAML). The language turned into initially referred to as QAL however afterward the authors redefined it as a sublanguage of XML and named it question answer Markup Language (QAML). The syntax and semantics of QAL is supposed to be suitable for educators with broadly differing computing talents, i.e QAL is deliberately easy and sufficient to be effortlessly understandable and consequently clean to study.

The language additionally embodies the important constructs to specific a shape for a herbal language text. The shape editor guarantees that students assemble correct required structures (with appreciation to QAML‟s syntax and casual semantics) in a form that is suitable for accurate computerized marking.

2.2.8 IAT

Statistics extraction templates shape the center of the sensible assessment technology machine. Those templates are created manually in a special-reason authoring device by way of exploring sample responses. They permit for syntactic variation, e.g., filling the difficulty slot in a sentence with extraordinary equivalent principles.

The templates corresponding to a question are then matched against the student answer. In contrast to other structures, IAT functions templates for explicitly invalid answers. They examined their technique with a development check that must be taken by way of remedial college students. about 800 college students plowed via 270 check gadgets.

The robotically graded responses then had been moderated: Human judges streamlined the solutions to reap an extra steady grading. This step had already been executed earlier than with exams graded by human beings. Mitchell et al state that their device reaches ninety nine.4% accuracy on the whole dataset after the guide adjustment of the templates via the moderation system.

Summarizing, they document a blunders of “among 5 and five.5%” in inter-rater settlement and an error of 5.eight% in automatic grading without the moderation step, though it isn't completely clear which data these facts correspond to. No facts on the distribution of grades or a random baseline is furnished.

2.2.9 WebLAS

One among the earlier structures is WebLAS, presented by Bachman et al. WebLAS addresses the issues of a P&P format.

It affords an included approach to assessing language ability for the purpose of creating selections approximately placement, analysis, development and success inside the East Asian language programs, as the content material specs of the assessment device for these languages are based directly at the direction content, as laid out in scope and sequence charts, and utilize obligations which are similar to the ones utilized in study room education. WebLAS is designed with the following advantages:

1. More administrative performance

2. More actual, interactive and valid checks of language ability along with integration of course and evaluation content and incorporation of slicing edge and multimedia generation for assessment.

A human mission creator feeds the system with rankings for version answers. everyday expressions are then created automatically from those version answers. Considering every ordinary expression is related to a score, matching the expression in opposition to a scholar solution yields a score for that answer. Bachman et al. (2002) do not provide an evaluation examination based totally on records.

2.3 Related Works

Some of the earliest structures of AES had been depending on hand made capabilities and characteristic engineering. In 1986, page(1986) evolved an AES tool referred to as Challenge Essay Grade(PEG) with the aid of the use of best linguistic surface functions. possibly an instance is E-Rater (Jill Burestein, 1998) that employed extra traditional strategies of natural language processing.

The equal mission was launched underneath version 2 within the year 2004 which utilizes a brand new set of capabilities to symbolize traits associated with employer and improvement, lexical complexity ,and so on. A lot of these methods shared a commonplace regression equation for essay assessment, consequently sharing a commonplace difficulty of being dependent on feature engineering.

The introduction of neural networks eliminated the need for handcrafted functions .Alikaniotis et al. (2016) and, Taghipour and Ng, (2016) offered scoring models based on lengthy-ShortTerm-memory networks. These shaped a number of the early examples of application of deep gaining knowledge in the computerized essay scoring procedure. In particular Taghipour and Ng, (2016) offered a method to extract word degree semantics by way of applying 1D convolution over vectors.

The foremost obstacle of the paper being utilization of 1-hot representations that don't extract members of the family as effectively as phrase embedding does. The use of unmarried layer LSTM additionally does not offer powerful semantic relation analysis. Apparently, Dong and Zhang,(2016)offered a model involving CNN's, eventually adding interest mechanism.(Dong et al., 2017).

In current years, we have visible charming neural architectures carried out to automated essay scoring structures. Zhang and Litman,(2018) proposed a version that offers supply articles for scoring the essay. Jiawei Liu et al., (2019) offered a two level getting to know technique leveraging each hand made capabilities and neural networks to calculate three specific ratings and giving a final rating based on those. Siamese Neural architecture was introduced through Liang G et al, (2018) where Bidirectional LSTM became used in a Siamese fashion to achieve ratings.

A Neural approach to automated essay scoring is one of the earliest papers that carried out the concepts of deep gaining knowledge of computerized essay scoring. The paper makes use of a protracted brief term based totally recurrent neural network to seize the semantics of the essay. The phrases of the essay have been represented through one-hot encoded vectors.

These vectors have been convoluted with 1D convolution to extract the critical statistics and perform a dimensional reduction. The contextual records extracted from the recurrent layer are then given to 'suggest through the years' which calculates the sum of expected scores over all the time instances and produces the very last score. The proposed model in this paper received an accuracy of 0.seventy four quadratic weighted Kappa score.

2.4 Tools and Technologies used

The venture is realized with the center structure of two Layered LSTM approaches to the manner of essay textual content assessment and scoring. In doing so, we've applied certain gear and technology

Python

Python has been used as the programming language for each version improvement. Python is an interpreted, object-oriented, high-stage programming language with dynamic semantics. Its excessive-level built in statistics structures, mixed with dynamic typing and dynamic binding, make it very appealing for fast software improvement, as well as to be used as a scripting or glue language to attach existing additives together.

TensorFlow

TensorFlow is an open source software library for excessive performance numerical computation. Its flexible structure allows smooth deployment of computation across a variety of structures (CPUs, GPUs, TPUs), and from computers to clusters of servers to cell and edge gadgets. TensorFlow turned out to be utilized in growing the deep neural network version.

NLTK

NLTK (Natural Language Toolkit) is a set that includes libraries and applications for statistical language processing. It is one of the most effective NLP libraries, which incorporates applications to make machines apprehend human language and respond to it with the best response. NLTK is used to perform pre-processing of textual content statistics to fit the model.

3. DESIGN OF THE PROPOSED SYSTEM

Design of proposed system includes block diagram description, modules used and the related algorithm foundations.

3.1 Block Diagram

Figure 3.1 depicts the block diagram of an automated essay scoring device which essentially includes five levels including essays which can be written by using the student; those essay sentences are combined with phrase embeddings to generate the phrase vectors.

Those word vectors are then handled through a deep neural community to carry out the syntactic and semantic take a look at. The very last text illustration acquired is given to a classification or regression algorithm that predicts the very last score.

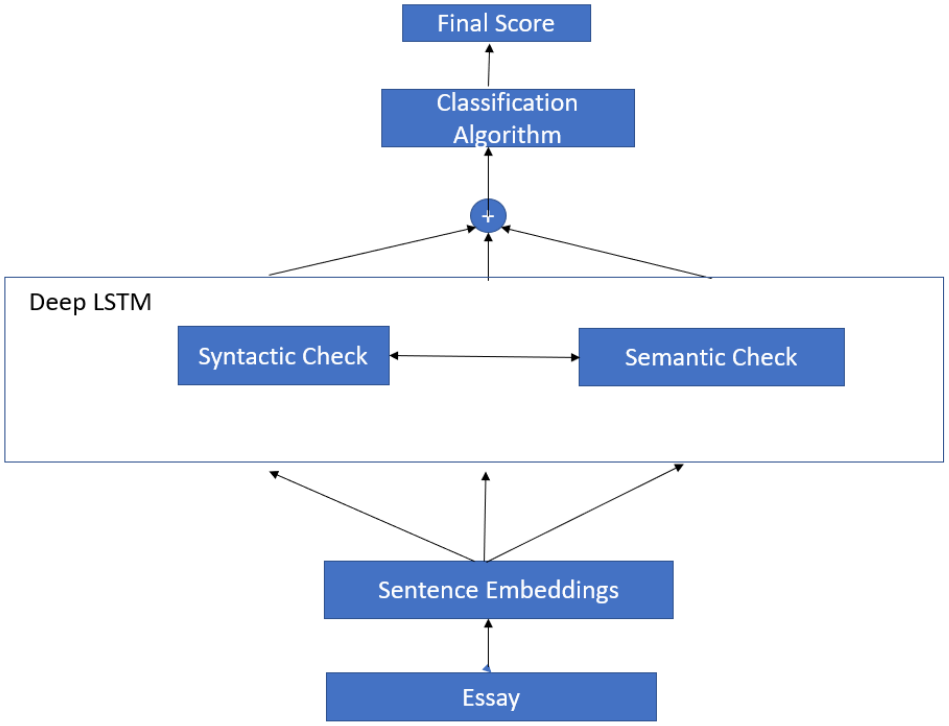


Fig 3.1 Block Diagram of the Proposed System

3.2 Module Description

3.2.1 Essay

This module consists of essays written by the student as a response to a particular prompt. The essay is not restricted by size or structure. The student will be evaluated by the system based on the content that they have written.

3.2.2 Sentence Embeddings

Sentence embeddings are embeddings formed on conversion of words to word vectors. These embeddings are pretrained models such as those of Glove or Word2vec. Sentence and word embeddings from the first layer of neural network.

Syntactic check or syntactic evaluation deals with the matter whether the essay written by adheres to the grammar rules of english. This includes and is not limited to word spellings and grammatical rules. This is done by the encoder part of architecture

Semantic check or semantic evaluation pertains to word relations and sentence relations. The decoder part of the architecture is used for this particular step. The essay is analyzed to find relations and word collocations that contribute to a high score for the essay.

3.2.3 Regression Algorithm

The final text representation obtained is treated as an array that can be classified or regressed to predict the final score. The activation function varies according to the model we choose.

3.3 Theoretical Foundation/Algorithm

Theoretical Foundation lays the groundwork of our project. It gives a glimpse of the various theories found in the advanced concepts of deep learning and artificial intelligence. It includes the intricacies of various neural networks and attention mechanisms used in the project.

3.3.1 Convolutional Neural Network

A Convolutional Neural community (ConvNet/CNN) is a Deep gaining knowledge of a set of rules that could absorb an entered image, assign significance (learnable

weights and biases) to various components/items inside the picture and be able to distinguish one from the opposite. The pre-processing required in a ConvNet is lost compared to different type algorithms. While in primitive methods filters are hand-engineered, with sufficient education, ConvNets have the capability to study those filters/characteristics.

In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, usually carried out to investigate visual imagery. They are additionally referred to as shift invariant or space invariant synthetic neural networks (SIANN), based at the shared-weight architecture of the convolution kernels or filters that slide alongside input features and provide translation equivariant responses known as function maps.

Counter-intuitively, maximum convolutional neural networks are handiest equivariant, as opposed to invariant, to translation. they've applications in photo and video popularity, recommender systems, photograph type, image segmentation, medical photo analysis, herbal language processing, mind-pc interfaces, [7] and financial time collection. CNNs are regularized variations of multilayer perceptrons. Multilayer perceptrons generally imply fully linked networks, that is, each neuron in one layer is hooked up to all neurons within the subsequent layer.

The "full connectivity" of those networks cause them to be vulnerable to overfitting records. normal methods of regularization, or preventing overfitting, encompass: penalizing parameters during training (such as weight decay) or trimming connectivity (skipped connections, dropout, etc.)

CNNs take a one of a kind technique in the direction of regularization: they take gain of the hierarchical sample in records and collect patterns of growing complexity using smaller and less complicated patterns embossed in their filters. Consequently, on a scale of connectivity and complexity, CNNs are on the decrease severely.

Convolutional networks were stimulated by biological approaches in that the connectivity pattern between neurons resembles the agency of the animal visual cortex. Person cortical neurons respond to stimuli handiest in a confined place of the sight view referred to as the receptive area.

The receptive fields of various neurons partially overlap such that they cowl the complete visual field. CNNs use fairly little pre-processing as compared to different picture class algorithms.

Because of this the network learns to optimize the filters (or kernels) through automatic studying, while in conventional algorithms these filters are hand-engineered. This independence from previous knowledge and human intervention in feature extraction is a chief benefit.

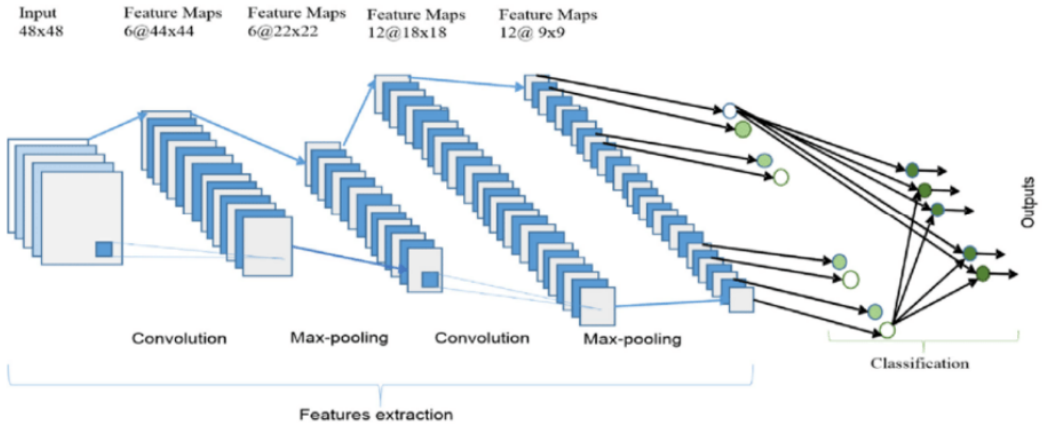


Fig. 3.2 CNN Architecture

3.3.2 Long Short Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. They are an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feed forward neural networks, LSTM has feedback connections.

Long short-term memory(**LSTM**) is an artificial recurrent neural network (RNN) architecture used in the discipline of deep studying. Unlike general feed-ahead neural networks, LSTM has comment connections.

It can not handle unmarried information points (including photos), however also complete sequences of data (including speech or video). As an instance, LSTM is relevant to obligations which includes unsegmented, linked handwriting reputation,

speech recognition and anomaly detection in network visitors or IDSs (intrusion detection systems).

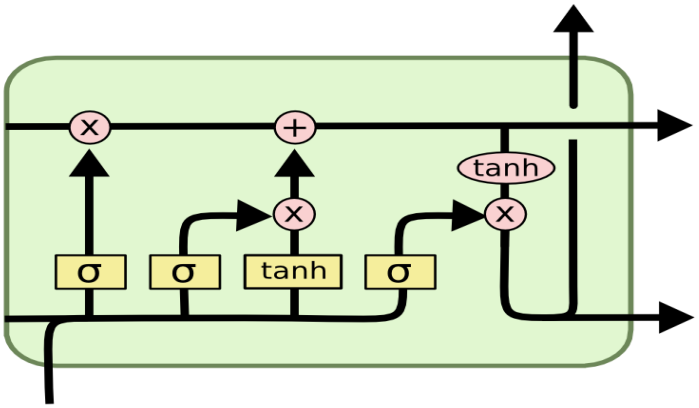


Fig. 3.3 LSTM Cell

A common LSTM unit consists of a mobile, an input gate, an output gate and a forget gate. The cellular recollects values over arbitrary time durations and the 3 gates regulate the float of facts into and out of the mobile.

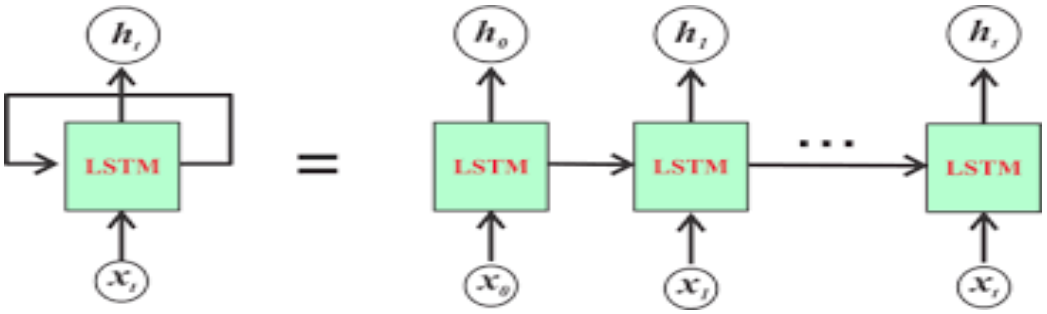


Fig 3.4 LSTM Network

LSTM networks are nicely-suitable to classifying, processing and making predictions primarily based on time series statistics, considering that there can be lags of unknown duration between crucial events in a time series.

LSTMs had been advanced to cope with the vanishing gradient trouble that can be encountered when schooling conventional RNNs. Relative insensitivity to gap length is a bonus of LSTM over RNNs, hidden Markov fashions and different series getting to know techniques in several applications.

4. IMPLEMENTATION OF THE PROPOSED SYSTEM

4.1 UML and DFD

UML and DFD includes the flow and construction of the proposed system.

4.1.1 DFD

Data Flow Diagrams (DFD) is the simplest structural diagram that depicts the flow of data through the system. In the system that we propose, the essay is the input data that gets transformed into vectors and flows through the semantic analysis phase and predicts the score output.

Level 0

DFD Level 0 is also called a Context Diagram. It's a basic overview of the whole system or process being analyzed or modeled. It's designed to be an at-a-glance view, showing the system as a single high-level process, with its relationship to external entities. It should be easily understood by a wide audience, including stakeholders, business analysts, data analysts and developers.

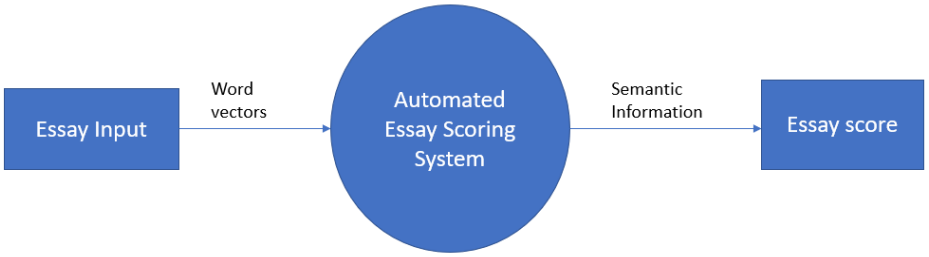


Fig. 4.1 DFD Level-0 of AES

Level 1

DFD Level 1 provides a more detailed breakout of pieces of the Context Level Diagram. You will highlight the main functions carried out by the system, as you break down the high-level process of the Context Diagram into its sub processes.

The input essay presented to the model is embedded into a vector form to perform a grammatical check and a deep semantic analysis extracting the intent of the essay, sent further to classify into a score range.

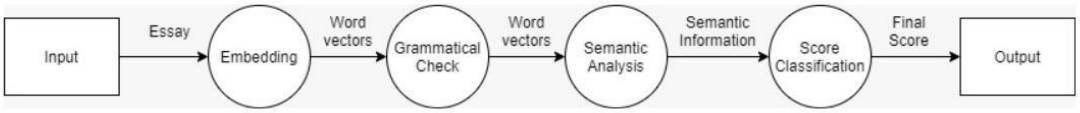


Fig. 4.2DFD Level-1 of AES

4.1.2 Flowchart

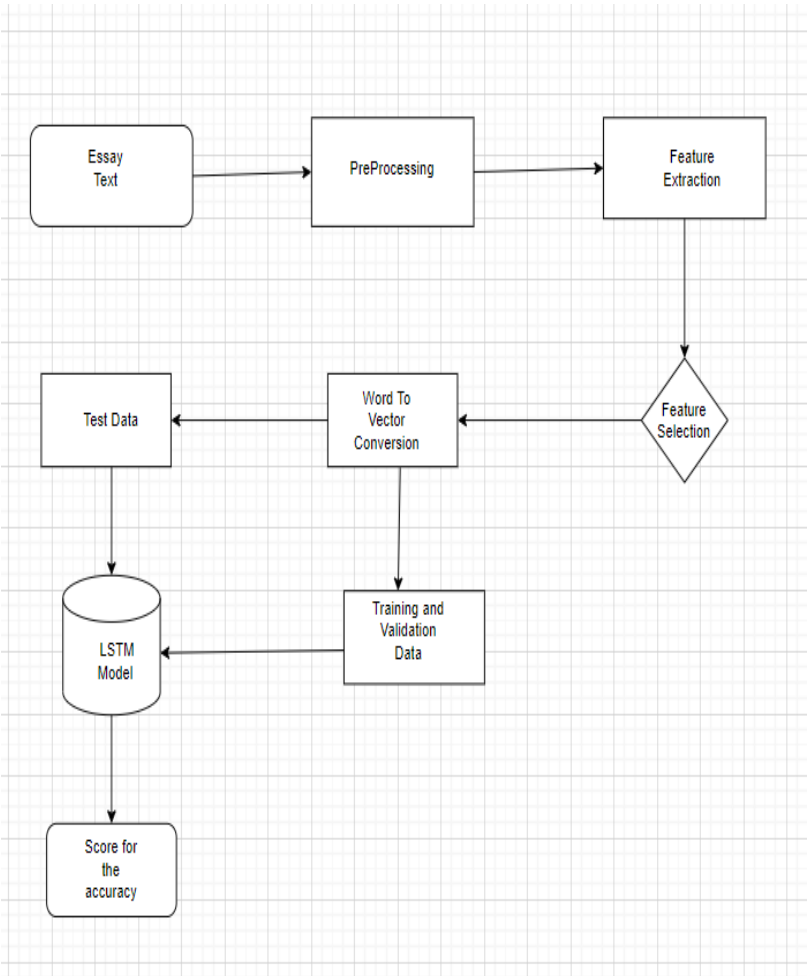


Fig 4.3 Flowchart of AES

4.1.3 Use Case

A **use case diagram** is a graphical depiction of a user's possible interactions with a system. A use case diagram shows various use cases and different types of users the system has and will often be accompanied by other types of diagrams as well. The use cases are represented by either circles or ellipses. The actors are often shown as stick figures.

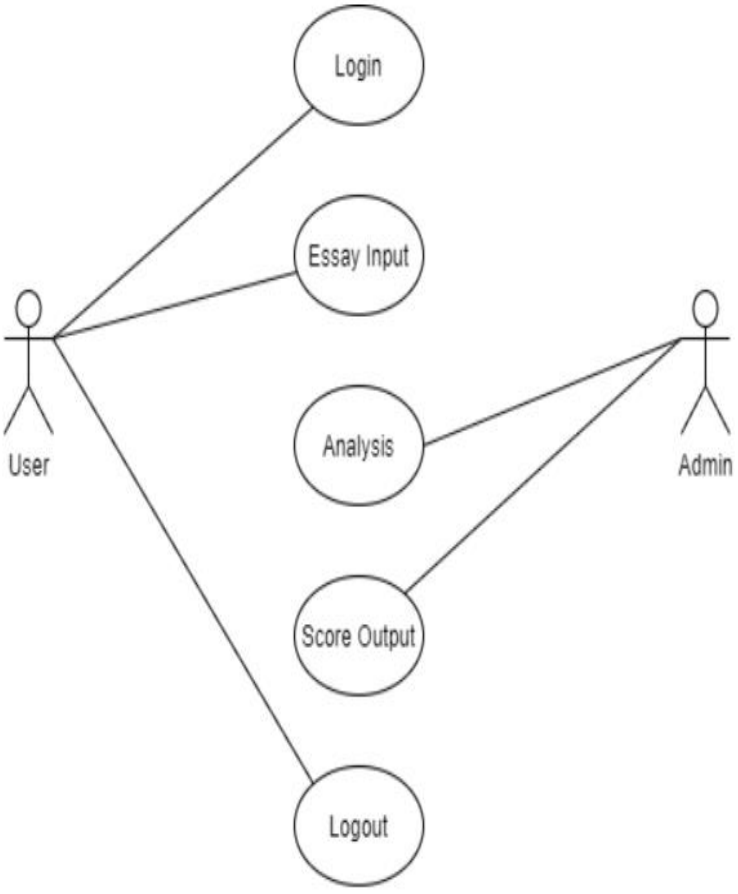


Fig. 4.4 Use Case Diagram for the Proposed System

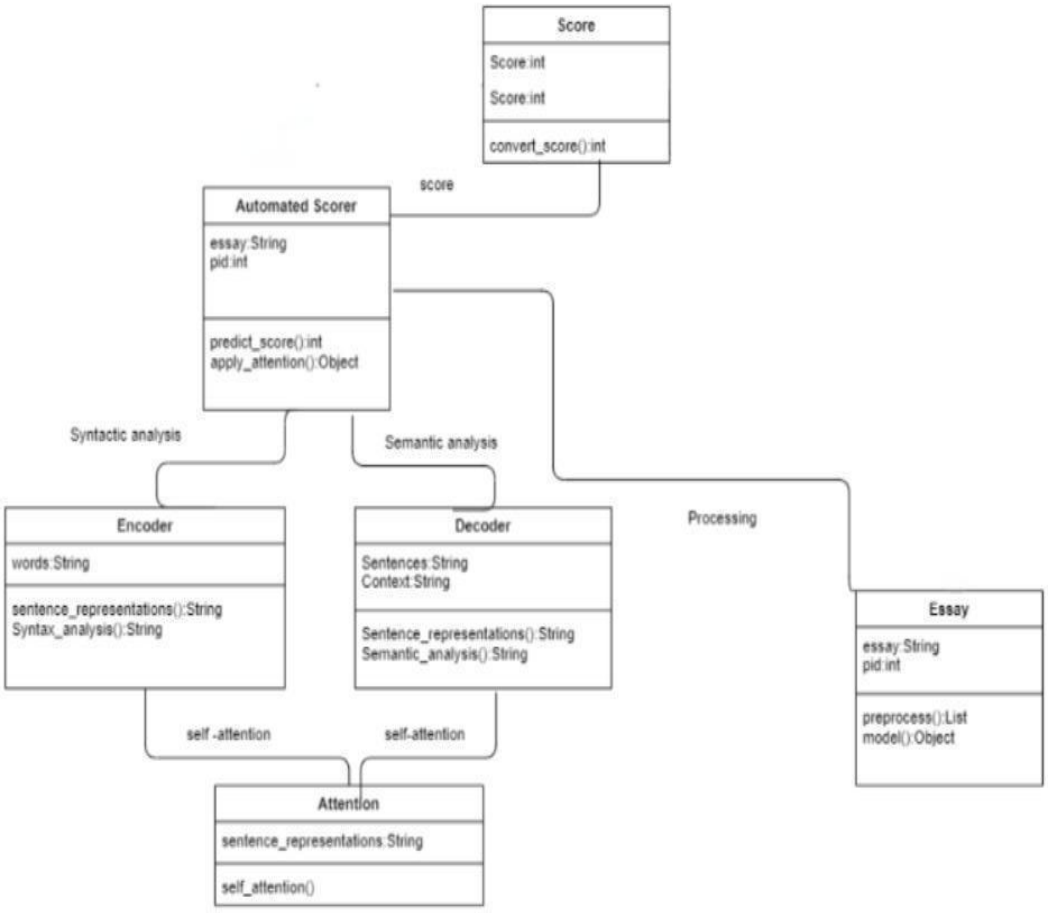
Use Case Diagram for the Proposed System The actors being a user and an admin navigate through the use cases of a login and inputting essay for the user and analysis followed by score output for the admin.

4.1.4 Class Diagram

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application.

Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modeling of object oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages.

Class diagram shows a collection of classes, interfaces, associations, collaborations, and constraints. It is also known as a structural diagram.



\

Fig 4.5 Class Diagram of the model

The class diagram in **Figure 4.1.4**, the main classes for the model are encoder and decoder that perform syntactic check and semantic check respectively.

The attention class is utilized in both of these classes, that is, encoder and decoder for construction of better text representations by learning the association between specific words and the corresponding scores.

The user class pertains to the student who can be an old or a new student. The student can login and submit the essay which is handled by the essay class. The essay class is again given to an automated scorer which pre-process the essay and predicts the score to score class. This score can be accessed by both the student and the teacher.

4.2 Design and Test Steps

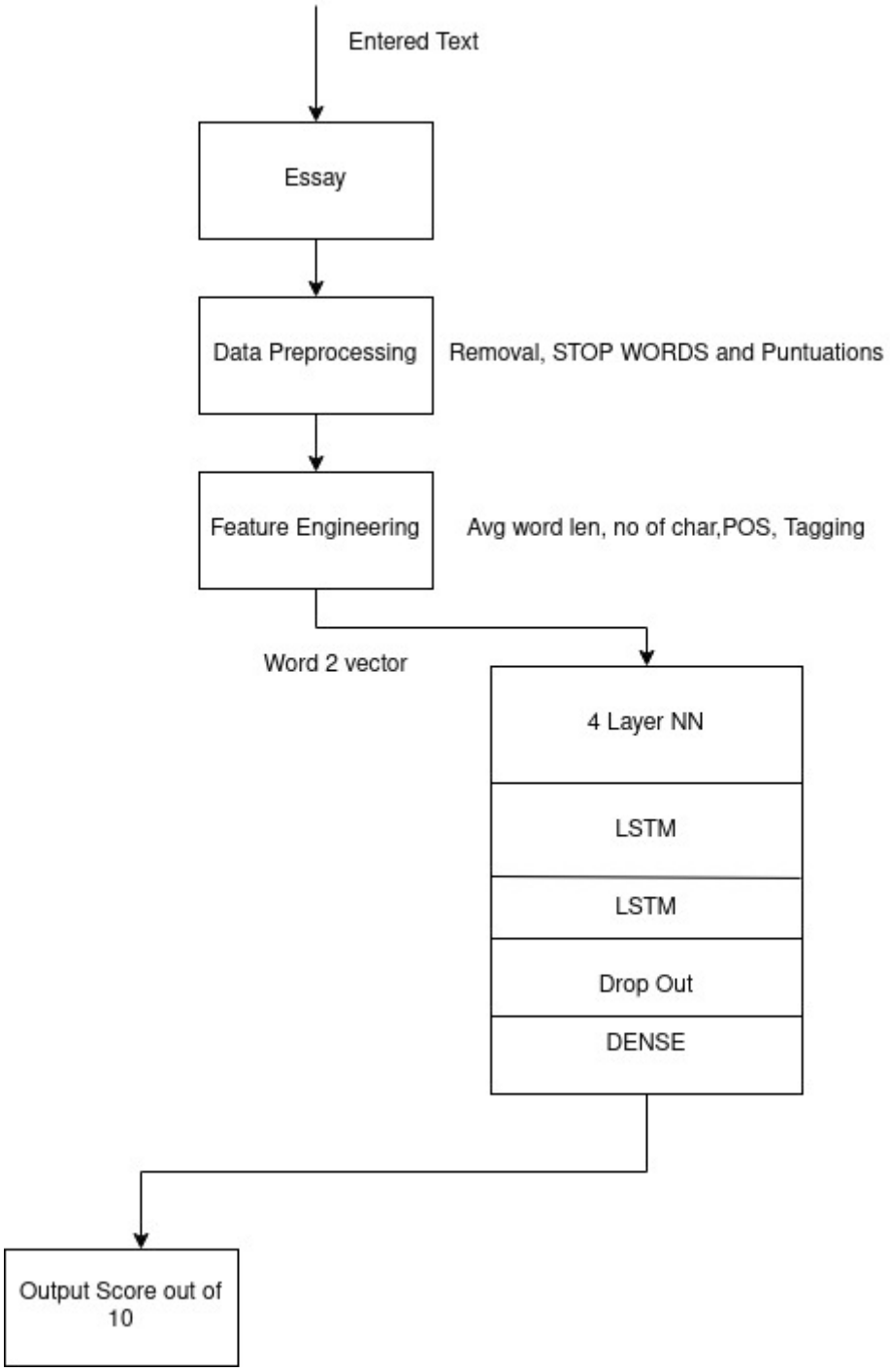


Fig 4.6 Design and test steps flow chart

Our model is inspired by the neural architecture presented by Dong et al.,(2017). Their entire model is divided into three sections mainly, they used a convolution layer and attention pooling to capture sentence representations.

Secondly, they used a LSTM for document representation. Lastly, they used a sigmoid layer for score prediction. We have introduced a 2 layer into the network architecture, influenced by the performance of multi layers presented by Robert Susik, (2020).

By doing so, our model extracts meaningful semantic relationships between sentences in the essay written by the student.

4.2.1 LSTM Architecture

Architecture consists of 4 layers namely, word embedding layer, 2 LSTM layers and dense layer.

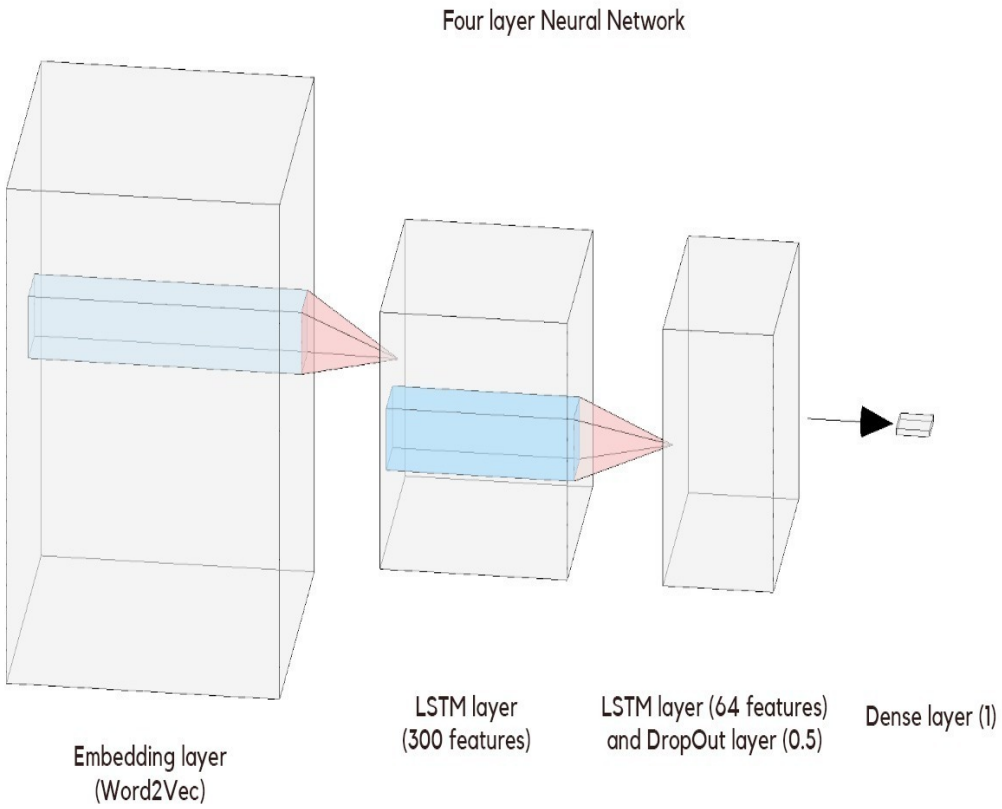


Fig 4.7 LSTM Architecture layer of the model

4.2.2 Word Embedding Layer

Word embeddings are used to map a word to a specific dimensional vector. We have used Word2Vec embeddings (Pennington et al., 2014) to obtain word embeddings. Word2Vec is a method to construct such an embedding. It can be obtained using two methods (both involving Neural Networks):

Skip Gram and Common Bag Of Words (CBOW).Preprocessing steps for neural networks are different from preprocessing steps for machine learning algorithms. Our training data is fed into the Embedding Layer which is Word2Vec. Word2Vec is a shallow, two-layer neural network which is trained to reconstruct linguistic contexts of words.

It takes as its input a large corpus of words and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

Word2Vec is a particularly computationally-efficient predictive model for learning word embeddings from raw text.

4.2.3 LSTM layer (300 features)

Features from Word2Vec are fed into LSTM. LSTM can learn which data in a sequence is important to keep or away. This largely helps in calculthrowating scores from essays. Finally the Dense layer with output 1 predicts the score of each essay.

4.2.4 LSTM layer (64 features)

LSTM(64), takes the 3x128 input from Layer 1 and reduces the feature size to 64. Since return_sequences=False, it outputs a feature vector of size 1x64.LSTM (64), and Layer 5, LSTM (128), are the mirror images of Layer 2 and Layer 1 to which the essay output is displayed.

4.2.5 Dense Layer

Dense Layer is used to classify images based on output from convolutional layers.Dense prediction is concerned with predicting a label for each of the input units, such as pixels of an image. Accurate dense prediction for time-varying inputs finds applications in a variety of domains. with this the essay score of each part is predicted.

4.2.6 Evaluation and Testing Criteria

The scores generated by AES systems need to be compared to ratings assigned by human - annotators. While there are many correlation metrics such as Pearson's correlation, Spearman's correlation, we have chosen Quadratic Weighted Kappa(QWK) score to be our evaluation metric.

The main reason for this choice is because this metric is useful when it's necessary to evaluate the possible impact of random selection in computation of standard accuracy. (Giuseppe Bonaccorso,2017)

IN QWK, a weighted matrix is calculated as follows

$$\boldsymbol{W}_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

Where i and j are the reference ratings and hypothesis rating respectively. N is the number of possible ratings. A matrix O is calculated where O(i,j) denotes the number of essays that received a rating i from human annotators and rating j from AES. An expected count matrix E is constructed as the outer product of histogram vectors of two(reference and hypothesis) ratings.

After normalization of E such that sum of elements of E and O are same, QWK is calculated as follows

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} o_{i,j}}{\sum_{i,j} w_{i,j} o_{i,j}}$$

In our experiments, we compared the QWK scores of our model to the chosen baseline and performed paired t-test analysis to test the improvement obtained.

4.3 Algorithm/Pseudo code

Algorithm and pseudo code deals with algorithms of neural network and preprocessing.

4.3.1 NLTK Pre-Processing

```
import nltk
from nltk.corpus import stopwords

english_stopwords = set(stopwords.words('english'))


def process_text(text, remove_stopwords=False):
    tokens = []
    sentences = text.decode('utf8').lower().split('.')
    for sentence in sentences:
        ttokens = [token for token in nltk.word_tokenize(sentence) if token.isalpha()]
        if remove_stopwords:
            ttokens = [token for token in ttokens if not token in english_stopwords]
        tokens += ttokens
    return tokens
```

Fig.4.7 Tokenization and Stop Word Removal

Tokenization is the manner of tokenizing or splitting a string, textual content into a list of tokens. one could think about a token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

In Figure 4.1 while a token magnificence represents multiple viable lexemes, the lexer often saves sufficient statistics to breed the original lexeme, in order that it is able to be utilized in semantic evaluation.

The parser typically retrieves these facts from the lexer and stores it within the abstract syntax tree. That is important so that you can keep away from statistics loss in the case of numbers and identifiers. Tokens are identified based totally on the particular regulations of the lexer.

Some strategies used to perceive tokens include: normal expressions, specific sequences of characters termed a flag, precise separating characters known as delimiters, and explicit definition by using a dictionary. unique characters, together with punctuation characters, are typically used by lexers to perceive tokens because of their herbal use in written and programming languages.

Tokens are frequently labeled by way of man or woman content material or via context within the facts stream. classes are described with the aid of the rules of the lexer. categories regularly involve grammar factors of the language used inside the facts movement.

Programming languages regularly categorize tokens as identifiers, operators, grouping symbols, or via information kind. Written languages normally categorize tokens as nouns, verbs, adjectives, or punctuation. classes are used for postprocessing of the tokens both by the parser or by using different features within the program.

In computing, forestall phrases are words that are filtered out before or after processing of natural language facts.

Though "stop phrases" usually refers back to the most not unusual words in a language, there is no single customary list of stop phrases used by all herbal language processing equipment, and certainly no longer all gear even use this sort of list. a few gear specially keep away from disposing of these forestall phrases to guide phrase search.

Any organization of words can be selected because they prevent words for a given cause. For a few search engines like google and yahoo, these are some of the maximum not unusual, brief function phrases, consisting of the, is, at, which, and on. In this situation, forestall phrases can cause problems when trying to find terms that consist of them, in particular in names including "The Who", "The The", or "Take That".

Other search engines like google and yahoo cast off some of the most commonplace words—consisting of lexical phrases, consisting of "need"— from a query which will improve performance.

Fig. 4.8 Stemming

```
# Porter Stemmer: commonly used
from nltk.stem.porter import *
stemmer = PorterStemmer()

tokens = process_text(text)
stem_tokens = [stemmer.stem(token) for token in tokens]

# Snowball Stemmer: not so commonly used
from nltk.stem.snowball import SnowballStemmer
stemmer = SnowballStemmer("english")

tokens = process_text(text)
stem_tokens = [stemmer.stem(token) for token in tokens]
```

In linguistic morphology and data retrieval, stemming is the system of decreasing inflected (or occasionally derived) words to their word stem, base or root shape—usually a written word form.

The stem does not want to be the same to the morphological root of the phrase; it also includes sufficient that related words map to the equal stem, even if this stem isn't always in itself a valid root. Algorithms for stemming have been studied in laptop technology since the 1960s.

Many search engines like google and yahoo treat phrases with the identical stem as synonyms as a sort of question enlargement, a technique referred to as conflation. A computer program or subroutine that stems word can be called a stemming program, stemming algorithm, or stemmer. Stemmers for English running at the stem cat have to become aware of such strings as cats, catlike, and catty.

A stemming set of rules may additionally lessen the phrases fishing, fished, and fisher to the stem fish. The stem no longer be a word, as an example the Porter set of rules reduces, argue, argued, argues, arguing, and argus to the stem argu.

```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()

tokens = process_text(text)
lemma_tokens = [lemmatizer.lemmatize(token) for token in tokens]
```

Fig. 4.9 Lemmatization

Lemmatization (or lemmatization) in linguistics is the procedure of grouping together the inflected styles of a word so that they may be analyzed as an unmarried item, diagnosed by the word's lemma, or dictionary shape.

In computational linguistics,(four.four.1.three) lemmatization is the algorithmic process of determining the lemma of a word based on its supposed meaning. Unlike stemming, lemmatization depends on effectively figuring out the meant part of speech and that means of a phrase in a sentence, as well as in the large context surrounding that sentence, such as neighboring sentences or maybe a whole record. As a result, developing green lemmatization algorithms is an open location of studies.

Lemmatization commonly refers to doing matters well with the usage of a vocabulary and morphological evaluation of phrases, generally aiming to eliminate inflectional endings only and to go back the base or dictionary shape of a word, that is known as the lemma .

If faced with the token noticed, stemming might go back simply s, while lemmatization could try and return either see or noticed relying on whether using the token became as a verb or a noun. The two may differ in that stemming most typically collapses derivationally associated phrases, while lemmatization commonly collapses the unique inflectional varieties of a lemma. Linguistic processing for stemming or lemmatization is often executed by means of a further plug-in aspect to the indexing procedure, and a number of such additives exist, each commercial and open-supply.

4.3.2 Model Algorithm and Training Model

Algorithms pertain to the order and set of commands of the layers of the neural community. schooling refers back to the iterative method of predicting the output and learning with assist of loss feature the discrepancy between predicted output and the real output.

4.3.2.1 Model Algorithm

start

1. Pre-technique the essay and generate tokens

2. For- every essay:

2.1 Generate Embeddings with help of word2vec.

2.2 An LSTM Layer is applied to study sentence representation and context.

2.3 follow the LSTM layer to research the final representation.

3. A dense layer with sigmoid activation is used for the very last rating prediction.

for. score output.stop

Loss:

MSE(suggest square error) calculates the average cost of distinction between gold widespread ratings yi * and prediction rankings yi. MSE is applied ubiquitously to regression obligations. consequently we've determined to adopt this loss characteristic for our AES device. The following equation defines MSE, given N is the full range of samples.

$$MSE(y, y^*) = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^*)^2$$

Optimization:

In this paper, we followed the relu optimizer (Ba et al., 2017) owing to its efficiency. the studying price is about to zero.001, momentum to zero.9 for education our complete model. We have set the Dropout price to 0.five to save you from overfitting.

4.4 Dataset Description

The records that we used for our education is the one published by Hewlett basis for the 2012 opposition titled ‘computerized pupil assessment Prize’(ASAP)1 on Kaggle.

There are eight essay gadgets. every unit of essays was generated from a single spark off. Those encompass eight prompts with 3 special sorts of essays: persuasive, supply-structured and narrative.

The essays have extraordinary rating levels, being scored on average through three raters across domain names. selected essays variety from a median length of one hundred fifty to 550 phrases consistent with reaction.

All essays have been hand graded and were double-scored. Each of the 8 statistics sets has its very personal precise tendencies. The variety is supposed to check the bounds of your scoring engine’s abilities.

table 1: Description of the ASAP AEG dataset. The Avg. The duration column offers the average length of the essay, in terms of a wide variety of phrases. The score range column lists the scoring range of the diverse attributes that we rating.

We use the same rating range as the overall rating variety of the essays. The remaining column tells us the prompts whose attribute scores we contribute. all the essays were written with the aid of local English talking youngsters from instructions 7 to ten.

The training data is furnished in three formats: a tab-separated price (TSV) report, a Microsoft Excel 2010 spreadsheet, and a Microsoft Excel 2003 spreadsheet. The cutting-edge release of the training statistics consists of essay sets 1–6. units 7–8 might be released on February 10, 2012. each of those documents consists of 28 columns:

- essay_id: a completely unique identifier for each man or woman student essay
- essay_set: 1–eight, an identity for each set of essays
- essay: The ascii textual content of a scholar's response
- rater1_domain1: Rater 1's domain 1 rating; all essays have this
- rater2_domain1: Rater 2's domain 1 score; all essays have this
- rater3_domain1: Rater 3's area 1 score; only a few essays in set 8 have this.
- domain1_score: Resolved rating among the raters; all essays have this
- rater1_domain2: Rater 1's domain 2 rating; most effective essays in set 2 have this
- rater2_domain2: Rater 2's area 2 score; only essays in set 2 have this
- domain2_score: Resolved score among the raters; only essays in set 2 have this rater1_trait1 score — rater3_trait6 score: trait ratings for units 7–8

we've got used the training_set_rel3.tsv document in our code so one can educate our version.

essay_id	essay_set	essay	rater1	rater2	rater3	domain1	rater1_	rater2_	domain2	rater1_tra	rater1_tra	rater1_tra	rater1_tra	rater7
19237	7	One time I was going to see my frie	11	12		23				3	3	2	3	
19238	7	One @CAPS1, I was very patient wh	9	11		20				2	2	2	3	
19239	7	Being patient! Patience is a good le	8	11		19				2	2	2	2	
19240	7	Being patient is a very hand thing fo	9	7		16				2	2	2	3	
19241	7	At one @CAPS2 in left you would h	8	8		16				2	2	2	2	
19242	7	I remember one time I was impatie	9	8		17				2	3	2	2	
19243	7	Patience in @CAPS1: Understandin	12	12		24				3	3	3	3	
19244	7	It was my first game. I was pumped	10	9		19				2	3	2	3	
19245	7	Hunting when you are turkey hunti	8	8		16				2	2	2	2	
19246	7	One day, two years ago I was going	7	8		15				2	2	2	1	
19247	7	In my opinion being patient is whe	5	8		13				0	1	2	2	
19248	7	Have you ever had to be patient be	9	12		21				2	3	2	2	
19249	7	You will say that you are board very	6	9		15				0	2	2	2	
19250	7	One time I was patient was when I	11	12		23				3	3	2	3	
19251	7	one time I was in patient. We was a	7	4		11				2	2	1	2	
19253	7	One time I had to use patience whe	8	8		16				2	2	2	2	
19254	7	Being patient is hard. I hate waiting	4	6		10				0	1	1	2	
19255	7	When I was @NUM1 yrs old my mo	9	6		15				2	2	3	2	
19256	7	One time when I was patient was w	8	8		16				2	2	2	2	
19257	7	When I was patient I was sitting at a	8	8		16				2	2	2	2	
19258	7	We were going to @CAPS1. Me and	9	12		21				2	2	2	3	
19259	7	Patience is a hard thing to have, bu	7	8		15				2	2	2	1	

Fig 4.4.1 training_set_rel3.tsv

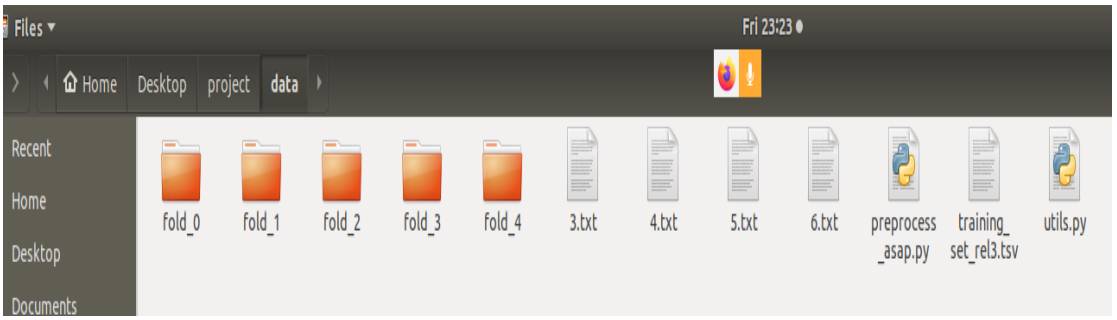


Fig. 4.4.2 Dataset Folder

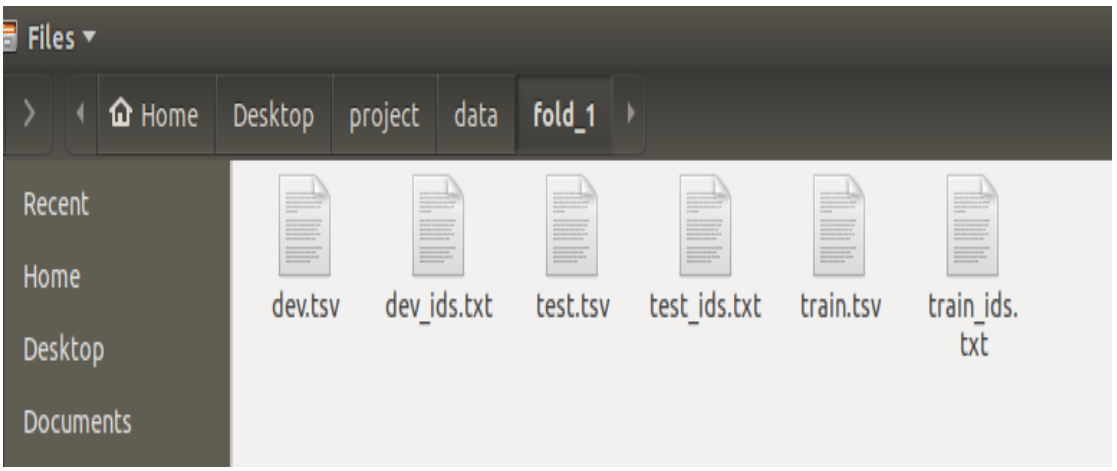


Fig. 4.4.3 Train, Validation and Test Division in a Fold

Figure 4.3.2 offers a concept approximately the dataset department into five folds, in each fold is again divided into test, train and validation datasets as proven in figure four.3.3 The test set that's given is used for trying out tactics.

4.5 Testing Process

We have designed our experiments to test three hypotheses:

H1: Our model will outperform or at least perform equally well as our baseline model on ASAP essay corpora in holistic score prediction.

H2: Our model will outperform or perform equally well as the non-neural network baselines.

H3: Our model will have a better or at least equal semantic score as our baseline model

Text pre-processing is done with NLTK (Steven Bird, 2009), and the vocabulary size is limited to 4000 words, with all additional words being handled as unknowns. The scores are on a scale of [0,1]. (Taghipour and Ng, 2016).

During model evaluation, the projected scores are transformed back into original score ranges for model training and prediction assessment. To do 5-fold cross validation, we partitioned the dataset into five folds. Sixty percent of the data in each fold is utilized for training, twenty percent for development sets, and twenty percent for testing. Table 2 summarizes the hyperparameters that were utilized to train the models.

5. RESULTS AND DISCUSSIONS

The performance of the model with the architecture of LSTM+LSTM is better than our baseline. When using a paired t-test, the QWK scores derived for the model were shown to be significantly improved (p0.05).

Finer text representations achieved by the architecture of a two-layered LSTM surface and the use of word2vec at both the phrase and sentence degrees can be attributed to this high performance.

Table 5.1 Different Models and their average kappa score. The scores that improved statistically significantly (p0.05) are noted with a "*." The prompts with the highest scores are shown in bold.

Model Layers	STUDENT PROMPT								
	1	2	3	4	5	6	7	8	Avg QWK
LSTM+LSTM	0.808	0.648	0.686	0.761	0.811	0.823	0.786	0.702	0.753
GRU+GRU	0.784	0.620	0.612	0.771	0.757	0.791	0.802	0.675	0.722
CNN+LSTM	0.796	0.644	0.593	0.752	0.761	0.782	0.762	0.699	0.719
EASE(SVR)	0.781	0.621	0.630	0.749	0.7/8 2	0.771	0.727	0.534	0.699
EASE(BLRR)	0.761	0.606	0.621	0.742	0.784	0.775	0.730	0.617	0.705

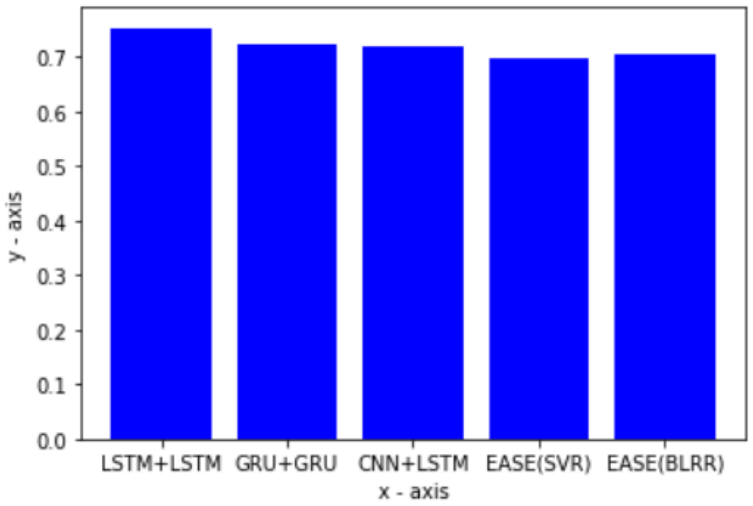


Fig 5.1: Comparing Different models with their Average QWK scores

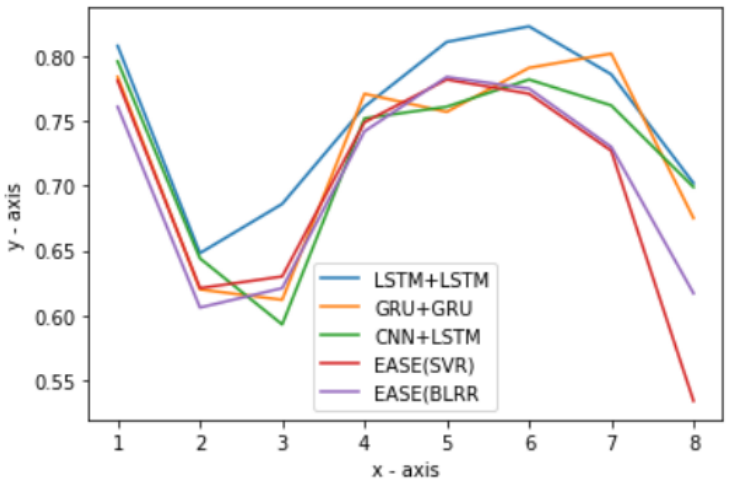


Fig 5.2: Comparing Different models with their respective QWK scores on each prompt

The proposed design outperforms or performs similarly well throughout all non-neural architectures, in step with QWK ratings for every layer, proving this premise. One clarification is that the very last representation in neural architectures has greater semantic records than records recorded by manually.

Table 5.2: For prompt 5 responses, the accuracy score and attention visualisations were calculated. The intensity of red is proportionate to the value of attention assigned.

System-layers	Essay	Avg. Attention Score
LSTM+LSTM	The mood created by the author in this memoir is gratitu de Narciso Rodriguez , grateful for way his cuban parent s brought him up when they had so little to begin with @ CAPS thing that are traditions family passed on . One of would be there rich culinary skills and a love cooking m other father came country give better life even though it meant	0.492
GRU+GRU	The mood created by the author in this memoir is gratitu de Narciso Rodriguez , grateful for way his cuban parent s brought him up when they had so little to begin with @ CAPS thing that are traditions family passed on . One of would be there rich culinary skills and a love cooking m other father came country give better life even though it meant	0.501
LSTM+BILSTM	The mood created by the author in this memoir is gratitu de Narciso Rodriguez , grateful for way his cuban parent s brought him up when they had so little to begin with @ CAPS thing that are traditions family passed on . One of would be there rich culinary skills and a love cooking m other father came country give better life even though it meant	0.533
CNN +LSTM	The mood created by the author in this memoir is gratitu de Narciso Rodriguez , grateful for way his cuban parent s brought him up when they had so little to begin with @ CAPS thing that are traditions family passed on . One of would be there rich culinary skills and a love cooking m other father came country give better life even though it meant	0.353

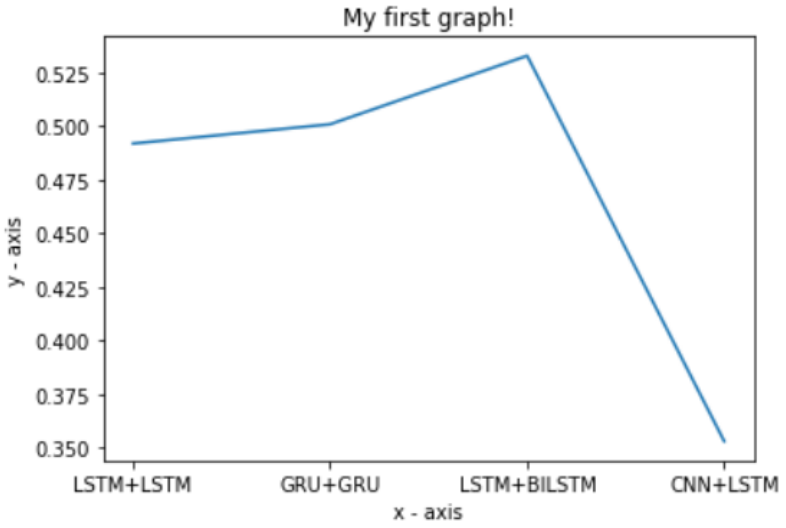


Fig 5.3: Comparing Different models with their respective accuracy

6. CONCLUSIONS AND FUTURE SCOPE

6.1 Conclusion

In this paper, we propose a recurrent-based 2 layered LSTM model that outperforms state-of-the-art LSTM-based models for automated essay scoring.

This model is capable of representing both local and contextual usage of information by essay scoring. This model yields score-specific word embeddings used later by a recurrent neural network in order to form essay representations. We have shown that this kind of architecture is able to surpass similar state-of-the-art systems.

We also introduced a novel way of exploring the basis of the network's internal scoring criteria and showed that such models are interpretable and can be further explored to provide useful feedback to the author.

It was satisfying that our neural network model using 300-dimensional LSTM as initialization to the embedding layer was our most successful model. Our model generates better sentence representations than existing models, resulting in a more thorough semantic analysis. In terms of quadratic weighted Kappa score, empirical results on the ASAP dataset show that our model outperforms strong existing baselines. The model reaches a Quadratic Weighted Kappa score of 0.92.

We believe that a more extensive hyperparameter search with our LSTM based models could outperform this result. There are many ideas moving forward. Trying out the models in Ensemble mode is also an extension we wish to try out in the near future.

6.1.1 Limitations

The limitation is that student essay answers must be in computer-readable format. In the real world, this isn't the case; the majority of exams are still conducted on paper. To transform paper written replies into digital format, greater effort and sophisticated technology such as Optical Character Recognition (OCR) are required in such scenarios.

It is vital to note that only answers were evaluated for grading in this project text, and complex mathematical formulas, graphs, tables, and other elements were not taken into account. The majority of pupils like to communicate themselves not only through writing but also through diagrams, graphs, and formulae. As a result, the scope is restricted to text-only responses.

6.2 Future Scope

The future scope of the system is to make the task of scoring essays prompt agnostic. The present model caters only to the English essays.

An extension of the system to other languages would further elevate the utility of the model. A multi-modal model which can take audio, video, etc as input and process it, would be useful to scale the system to consumers from all walks of life including the disabled and differently abled.