1 Statement of model problem

1.1Statement of essay grading problem

Automatic Essay Evaluation is an important research area in Natural Language Processing to automate evaluation and scoring. Unlike multiple-choice questions and short question answers, an essay is an open-ended question. There is no fixed format; students can write an essay in multiple ways. Manually grading essays is a resource-intensive task requiring time and labor. Teachers have to spend their valuable time grading the essays of students. If teachers have an automated essay grading system, then it will be helpful for them to speed up the evaluation process. They can devote more time to teaching and mentoring students. And for our website students from all over the world can take an exam made by their instructors and answer in **multiple languages** and should take a corresponding grade. There are parts in the answer that is needed be matched or just have a similar meaning. With no free available data, the problem become more challenging.

1.2 Statement of plagiarism problem

Plagiarism is a pressing problem for educational and research institutions, and publishers to counteract plagiarism, many institutions employ text-matching software. These tools reliably identify duplicated text yet are significantly less effective for paraphrases, translations, and other concealed forms of plagiarism

2 Literature review for the models:

2.1 Grading model:

- **Language Models and Automated Essay Scoring** [1]
  Method: Fine-tuning BERT and XLNet with a classification layer
  Data: Automated Student Assessment Prize (ASAP)
- **Article Automated Essay Scoring Using Transformer Models** [2]
  Method: Fine-tuning BERT with a classification layer
  Data: Automated Student Assessment Prize (ASAP)
- **A NLP Approach for Automatic Test Evaluation System** [3]
  Method: Preprocessing and tokenization with NLTK
  Bag of words, TF-IDF, Gensim, avg. similarity for the words tokens
  The proposed system takes about 15 seconds to evaluate a response (one student)
- **An Automated System for Essay Scoring of Online Exams in Arabic based on Stemming Techniques and Levenshtein Edit Operations**. [4]
  Method: Heavy Stemming then for each word get the similarity using the edit distance then compute the formula

$$FinalMark \cdot = wordj \cdot weightj, weight = \frac{1}{totalwords\ in\ correct\ answer}$$

- **NLP-based Automatic Answer Script Evaluation** [5]

  Method:

  - OCR
  - text summarization: take the average frequent words that have been selected as keywords where the most frequent and less frequent words are ignored. Then the weight of each sentence in the text is calculated based on the number of keywords in the sentence squared and divided by the window size. The window size is the maximum distance between two significant words in a sentence. Then sort the sentence in descending order based on their weight value and finally take the first n sentence as a summary of the long text.
  - C) tokenization with NLTK
  - D) four similarities measures
    - D.1) cosine for TF-IDF between student answer and the model answer
    - D.2) Jaccard similarity between the two-word list
    - D.3) structural similarity using Bigram similarity between the two-word list
    - D.4) Synonym Similarity with NLTK wordnet match every word in students with the model then divide by average word length of two documents.

    E) To count the spelling and grammar errors, a python package language check is used weighted sum of all measures for the predicted value

  Data: SemEval-2013 dataset

  Table 1 NLP-based AEE results

Precision	Recall	F-score
0.9	0.83	0.86

- **Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature**. [6]

  Method: SBERT “paraphrase-xlm-r-multilingual-v1” with 64 neurons single layer classifier

  Data: The Ukara dataset 2019

Table 2 IAES results

Model	SBERT+ NN	fastText + Stacking	Bidirectiona; LSTM	TF-IDF + Random forest, Logistic regression
F1-score	0.829	0.821	0.811	0.81

- **Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring** [7]

  Method: Fine-tuning BERT with a classification layer
  - ConvNet + attention pooling for the embeddings
  - special LSTM + attention over the hidden layers
  - Sigmoid

  Data: Automated Student Assessment Prize (ASAP)

Table 3 RNN for AES results

model	LSTM-CNN-attention
Avg. QWK	**0.764**

- **Automated essay scoring using efficient transformer-based language models** [8]

  Method: using BERT, ALBERT, Electra, Reformer or MobileBERT With a classification layer

  Data: Automated Student Assessment Prize (ASAP)

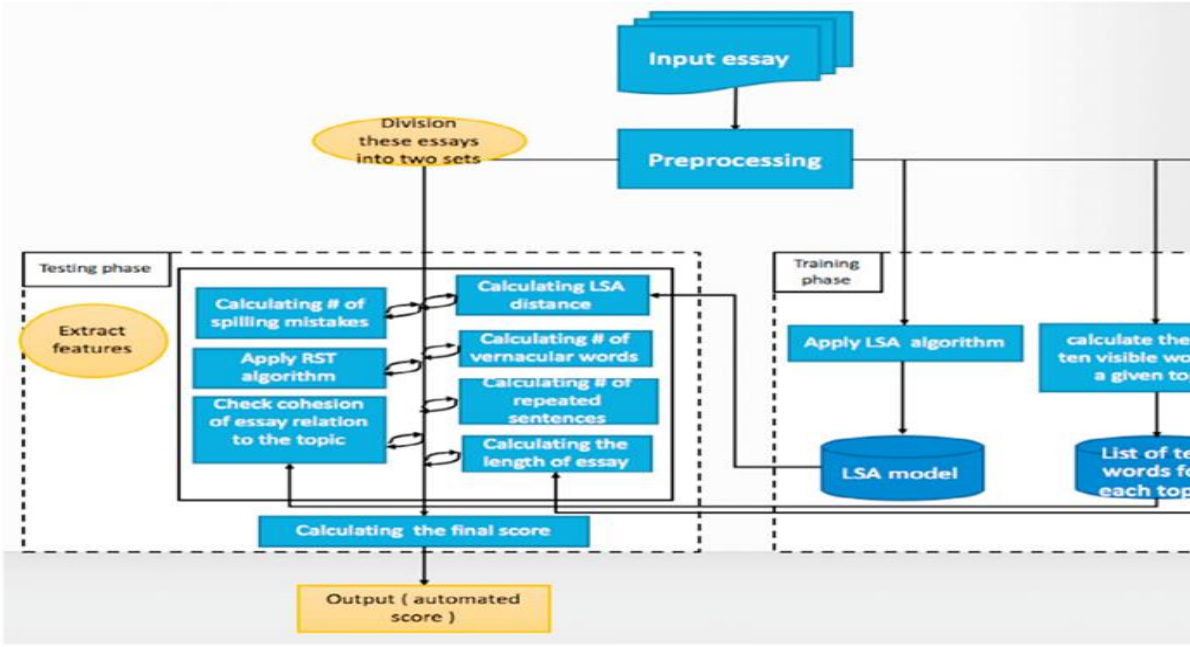
Table 4 transformers AES results

model	BERT	Electra	AlBERT	Mobile-BERT	Reformer
Avg. QWK	0.758	0.759	0.763	0.762	0.713

- **Automated Evaluation of School Children Essays in Arabic. Procedia Computer Science** [9]

  Method: Based on surveys done in Saudi Arabia the criteria are: spelling and grammar mistakes, the coherence and organization of the essay, the essay should be related to the topic, and sticking to Modern Standard Arabic (MSA) words.

Figure 1 AAEE method



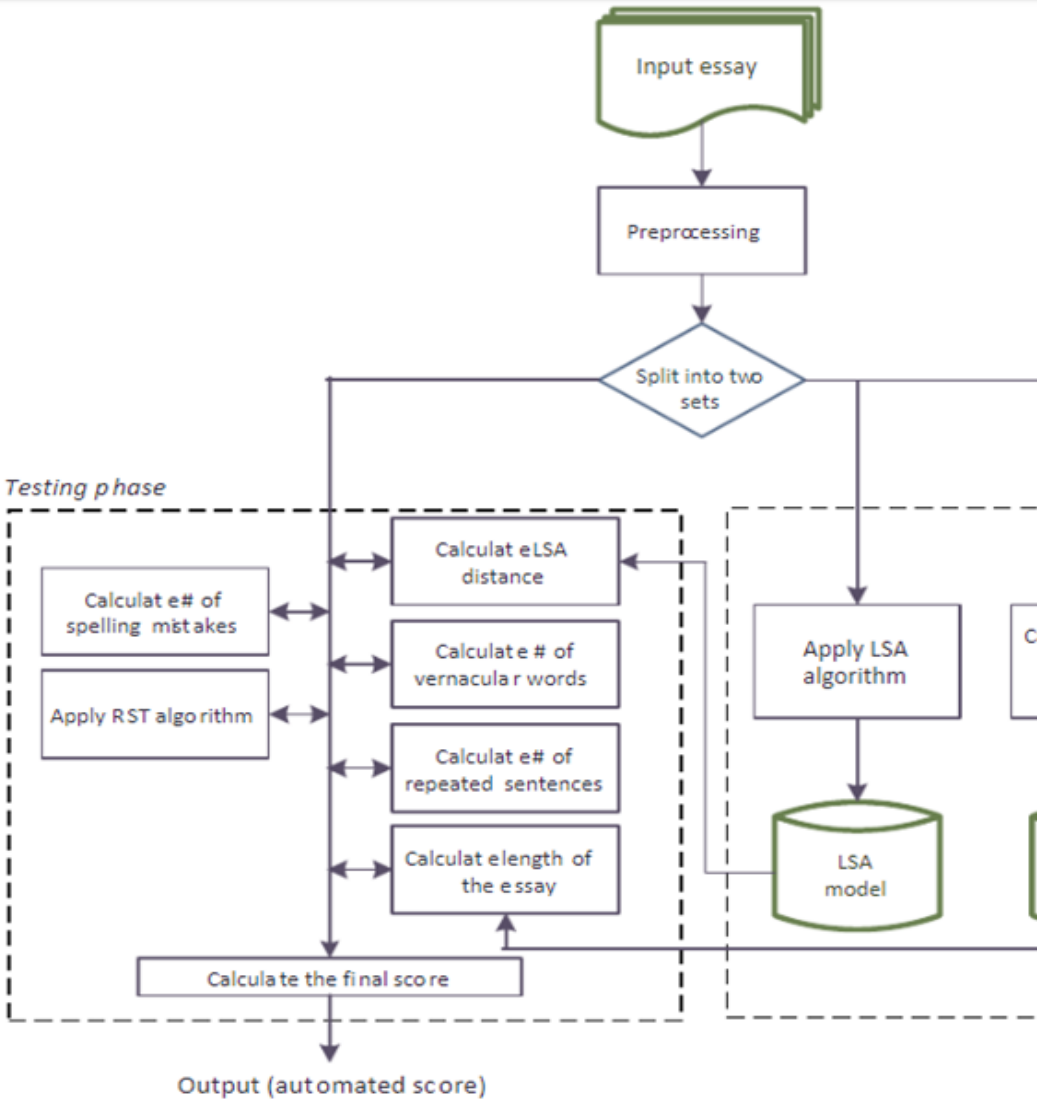
Data: the authors collected around 600 essays written by university students in Saudi Arabia.

The essays were part of a test in Arabic language course. The length of an essay ranged between 100- 200 words

- Updated paper from the same authors of the previous one [10]

  Method:

Figure 2 AAEE method



- **Automated Evaluation of Telugu Text Essays Using Latent Semantic Analysis** [11]

  Method: Using Latent Semantic Analysis

Figure 3 TAEE method

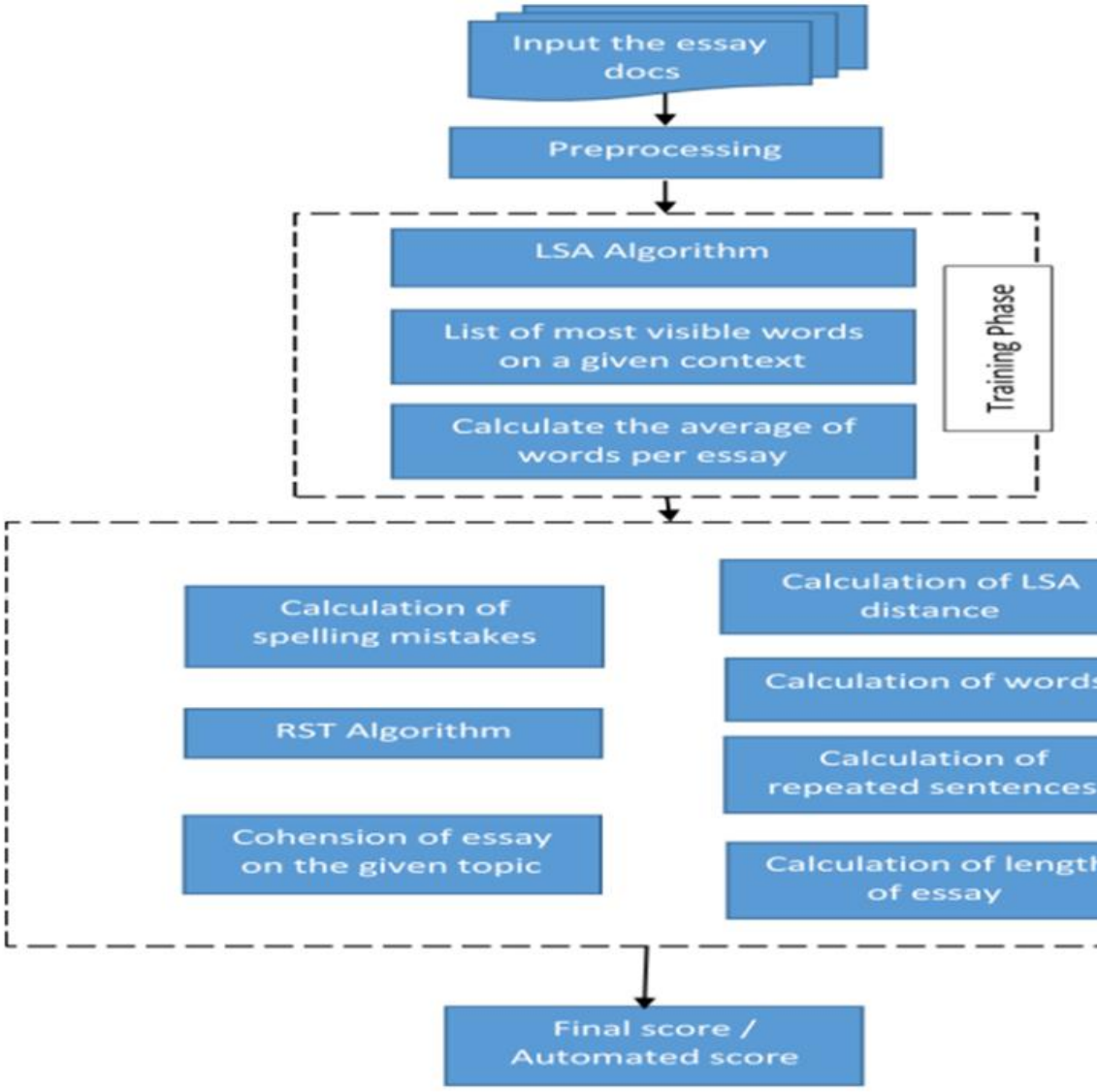


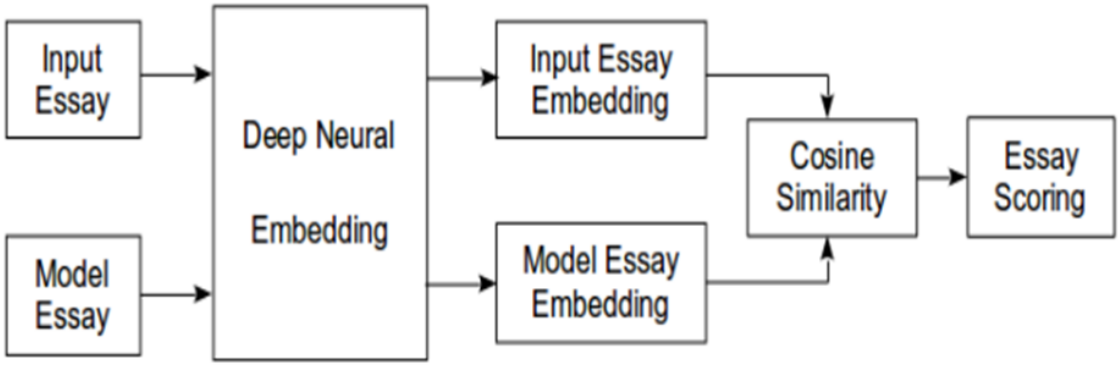
Figure1: Automated Evaluation of Essays

Data: private from schools

- **Efficacy of Deep Neural Embeddings based Semantic Similarity in Automatic Essay Evaluation** [12]

Method: embeddings from Google Sentence Encoder or ELMo. Then Cosine similarity

Figure 4 Semantic Similarity in AEE



Data: ASAP++

- **Automated Arabic Essays Grading System based on F-Score and Arabic WordNet** [13]

Method: Feature selection with F-score of model answer and student's answer then Cosine similarity

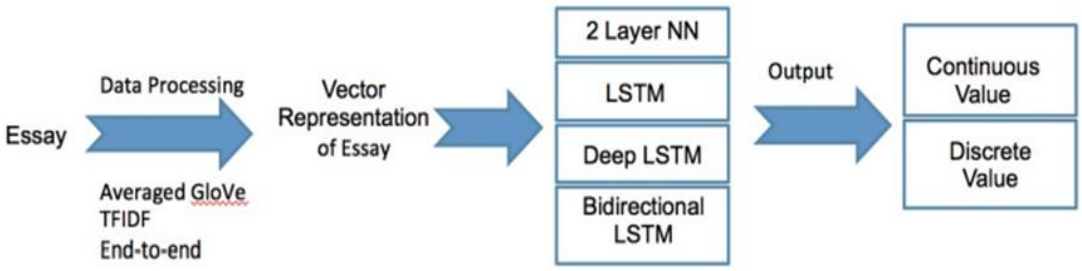
Data: private from a school in Jordan

- **Neural Networks for Automated Essay Grading Evaluation** [14]

Method: There are two kinds of models we are building: regression models and classification models.

For regression models, our output will be a continuous value between 0 and 12. For classification models, our output will be a discrete value between 0 and 12.

Figure 5 NN in AEG method



The best score achieved was 0.9447875, using a 2-layer neural network that trains word vectors together with the weights. Word vectors were initialized to the GLoVe word vectors

- **Improving short answer grading using transformer-based pre-training.** [15]

Method: finetuning BERT with a classification layer on 40% of the data

Data: SemEval-2013 the scientsBank-3way

Table 5 improving trasformers AES results

Accuracy	Micro-averaged-F1	weighted-averaged-F1
86.7	79.1	87.5

- **Using BERT and XLNET for the Automatic Short Answer Grading Task.** [16]
  Method: finetuning BERT with a classification layer
  Data: SemEval-2013 the scientsBank-2&3&5way
  Result: three test sets: test of unseen-answers (TUA), test of unseen-questions (TUQ), and
  test of unseen-domains (TUD) scenarios
  3-way: TUA sung model is better [16] in TUA, TUQ, TUD
  5-way & 2-way: BERT and XLNET are the best in TUA, TUQ, TUD
- **An Expert System for Automated Essay Scoring (AES) in Computing using Shallow NLP Techniques for Inferencing** [17]
  Method: match keywords from the student's answer to the model answer or in the Domain Specific Dictionary
  Limitations: is not suitable to assess free-text answers where the word order is important. It is also, more effective in applications to short text answers rather than bulky texts

2.2 plagiarism model:

- **Identifying Machine-Paraphrased Plagiarism** [18]
  Method: Embedding of the text using transformer-based models like BERT then pass the results to machine learning algorithms like Logistic regression, SVM, and naïve bayes.
  Data form arXiv, Wikipedia, theses datasets.
- **Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT** [19]
  Method: Siamese Bert model
  Data: Japanese data
- **Plagiarism Detector Using Machine Learning** [20]
  Text vectorizer then cosine similarity

3 the data:

For Grading model We used The Hewlett Foundation: Automated Essay Scoring or Automated Student Assessment Prize (ASAP)

Selected essays range from an average length of 150 to 550 words per response. All responses were written by students ranging in grade levels from Grade 7 to Grade 10. All essays were hand graded and were double scored. Final score is score 1. Score 2 is for inter-rater reliability purposes.

For plagiarism model data is Quora Question Pairs from Kaggle.

Validation data is survey data, web scrapped data about science and literature, and data from the website for both grading and plagiarism models.

3.1 exploratory analysis

The training data contained 10 essay sets. Each essay set had a description and a rubric for the score, all students were in grade 10 for all except essay set 10 were in grade 7

Essay set 1 is about a science lab procedure called “Acid Rain” total of 1672 essays and 558 for the test set. With an average length of 50 words per student’s response, and given seven model answers, type open-ended questions are scored on a scale (0–3)

Essay set 2 is about a science lab procedure called “Polymer Investigation” total of 1278 essays and 426 for the test set. With an average length of 50 words per student’s response. The answer contains two parts Conclusion and experimental design improvement we did each possible combination of them as model answers. type open-ended questions are scored on a scale (0–3)

Essay set 3 and 4 are a reading passage in English language arts about koalas/pandas, Invasive species respectively. total of 1891, 1738 respectively essays and 631, 580 for the test set. With an average length of 50 words per student’s response. And not given model answers as they are open-ended responses are scored on a scale of (0–2)

Essay set 5 and are in Biology about Protein synthesis, Cell membrane respectively. total of 1795, 1797 respectively essays and 599 for the test set. With an average length of 60, 50 respectively words per student’s response.

Given key elements for the model answer, required to mention at least four of them for the full mark and scored on a scale of (0–3)

Essay set 7 and 8 are reading comprehension passages in English language arts about “Trait of Rose”, and “Mr. Leonard” respectively. total of 1799 essays and 601 for the test set for each set alone. With an average length of 50 words per student’s response. And not given model answers as they are open-ended responses are scored on a scale of (0–2)

Essay set 9 is a reading comprehension passages in English language arts about “Organization of article”. total of 1798 essays and 600 for the test set. With an average length of 40 words per student’s response. And not given model answers as they are open-ended responses are scored on a scale of (0–2)

Essay set 10 is about a science lab procedure called “Doghouse”. total of 1640 essays and 548 for the test set. With an average length of 60 words per student’s response. And given sample model answers are scored on a scale of (0–2)

There are two scores for inter rater reliability (IRR). where IRR is a way to measure the level of agreement between multiple raters
Kolmogorov-Smirnov test

Table 6 K-S test results

	statistic	P-value
K-S test result	0.002150287673621243	0.9999999999992726

The kl divergence:
For the data as a whole = 0.00585

Table 7 KL divergence scores for each set

Essay set	1	2	3	4	5	6	7	8	9	10

Kl score	4e-05	7e-05	0.00321	0.00103	0.00039	0	0.00001	6e-05	0.00018	0.00076

Very ageable scores from both raters

Figure 6 essay set 1 word cloud sample for each student answer in Score j

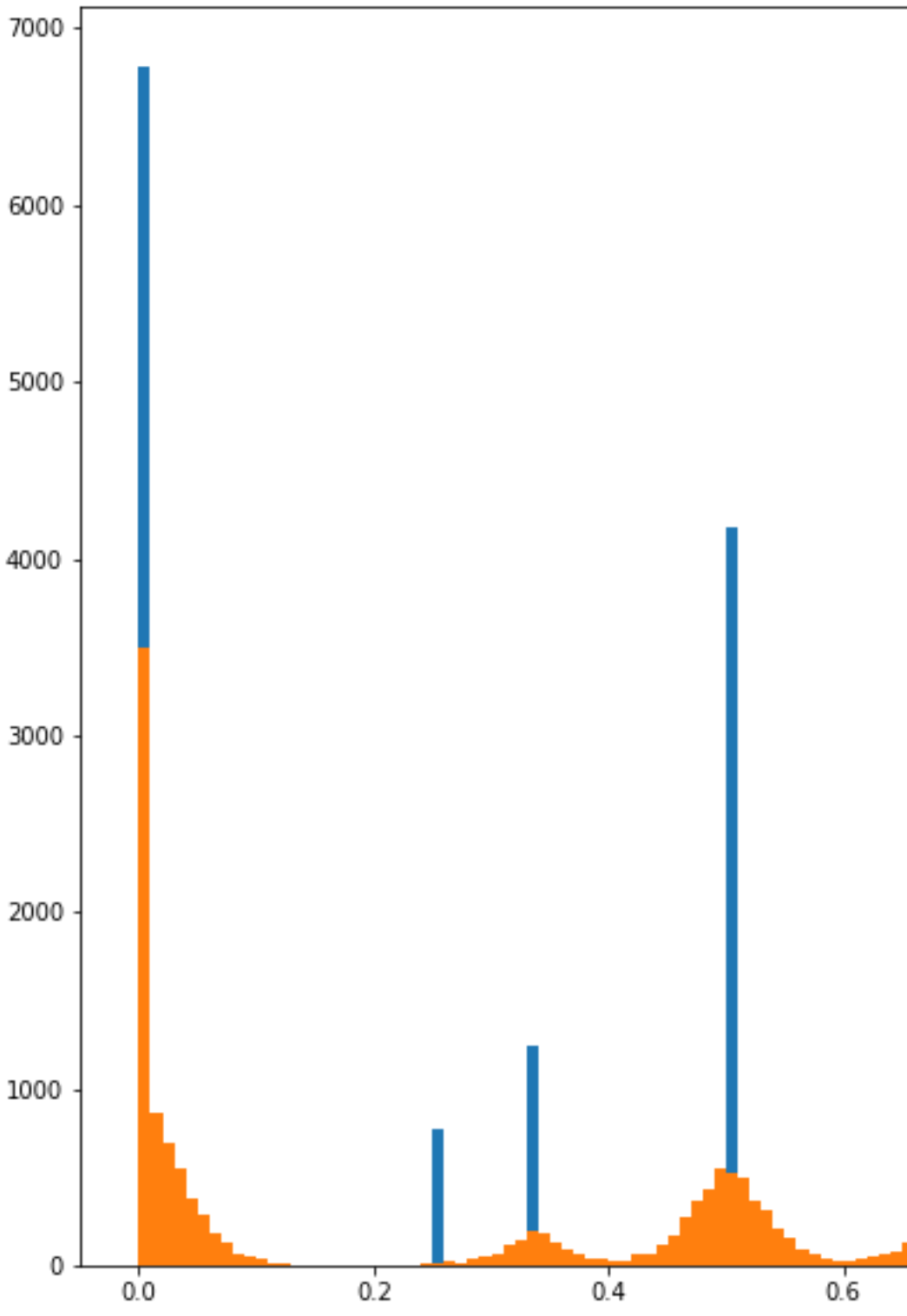
Here is a word cloud for essay set 1 seeing how the words change depending on the grade and as a who

3.2 Data model answers synthetization

Data model answers synthetization: the data set is not meant to be as model answer against students answers so we had to synthetize model answer to meet our system's inputs get 22 model answers for each essay set.

For essay set one given seven short model answers and random sampling the rest to get to 22 model answers.
For essay set two it's a two parts' answer mentioning from answers one and two, so we did a combination. And the rest is randomly sampled.
For essay set three and four are comprehension paragraphs so all was randomly sampled w.r.t answer length like so randomly sampling 100 answers then take the longest five and shortest five and the rest of 22 is varying length from between.
For essay set five and six are required to mention at least four, and three parts, respectively to get the full mark. And of course, students can mention more than four and three parts. So, we had to make a random permutation with varying length of at least four and three. Then the rest is random sampled w.r.t length three of the longest and 6 from the shortest and rest is random sampling.
For essays sets seven, eight, and nine are comprehension paragraphs so all was randomly sampled w.r.t answer length like so randomly sampling 100 answers then take the longest five and shortest five and the rest of 22 is varying length from between.
For essay set ten was explaining the chosen color between [white, dark gray, light gray, black] and querying the 22 model answers that starts with a color from the sampled data.
The label scores are of values [0, 1, 2, 3] so we had to tweak the distribution to allow other values, so we toke the mean of score1 and score2 and add gaussian noise.

Figure 7 scores distribution before and after modification


7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6

4 The model:

4.1 grading model:

Consists of: NER, manual keywords, KeyBERT automatic keywords extraction and grading, Siamese, and an output layer for the result.

- For the NER model used from Spacy
- Manual keywords follow a pseudo grammar of: “”keyword”” enclosing the keyword by double quotes to parse from the model answer or ””+keyword+”” for a similar keyword.

  Both are O(n) and they are instant as for parallel processing.
- Siamese for semantic similarity.
- Key-BERT for auto keywords extraction [21] and hyper parameter tuning [22]

Student answer length = k, model answer length = L, number of students answers = n

Encoding complexity = O(n) were n < 10,000 and O(n^2) were n > 10,000

And number of operations = O (n^2 d)

Complexity = top_n * L % n_grams + n^2 * top_n * n_grams * k * n * n_grams * (k % n_grams) * Encoding complexity

= top_n * L % n_grams + n^3 * n_grams^2 * k * (k % n_grams) * top_n * Encoding complexity

= n^3 * top_n * n_grams^2 * k * (k % n_grams) * Encoding complexity

Then done some optimization by taking the unique values of n_grams become = a value between 1 and 2 reducing the complexity to

= 2 * n^3 * k * (k % n_grams) * Encoding complexity

Then done some optimization by saving the encodings in a dictionary per word reducing the complexity to

= n^3 * k * (k % n_grams) * Encoding complexity

Then done some optimization by comparing only keywords of both the model answer and student answer reducing the complexity to

= n^2 * k * Encoding complexity

Output of keywords grading:

- Is dividing number of keywords similar against a tuned threshold over number of keywords
- A neural network to weighted averaging each keyword

- A regression, gaussian regression, decision tree models and toke the best performance model of them

For all encoding model we used [BERT, RoBERT, LongFormer, LineFromer, Distil-BERT, xlm-r- Distille-RoBERTa, paraphrase-multilingual-MiniLM-L12-v2] Then chose the last one “paraphrase-multilingual-MiniLM-L12-v2” as meeting the time, space, processing, and accuracy criteria. And better than xlm-r model in dissimilarity measures

Support grading and plagiarism for 50+ languages

4.2 plagiarism model:

- Siamese model
- Cosine similarity with a tuned threshold on the training data
- Neural network train on the network

For all encoding model we used [BERT, RoBERT, LongFormer, LineFromer, Distil-BERT, xlm-r- Distille-RoBERTa, paraphrase-multilingual-MiniLM-L12-v2] Then chose the last one “paraphrase-multilingual-MiniLM-L12-v2” as meeting the time, space, processing, and accuracy criteria. And better than xlm-r model in dissimilarity measures

All experiments results are available in the main project repository [23]

The code is professionally documented in the repository with user guides.

Conclusions

In short terms, we saw how we could work around the limitations of social distancing strategies with the discussed solutions, and even more by facilitating the educational examination system in either the international systems (schools, universities) or casual individuals like course instructors. And we saw how combining cutting-edge technologies in web technologies, Computer vision, and natural language processing can bridge so many gaps in today’s world. For future work. Keywords provided by the instructor manually not automated. Ability to Separate grading models. Train a model for times indicated grading. Higher accuracy approach for multi-parts answers grading for answer that contains multiple parts and meanings

References

[1] P. Rodriguez, A. Jafari and C. Ormerod, Language Models and Automated Essay Scoring. arXiv, arXiv:1909.09482, 2019.

[2] C. M. C. H. K. E. a. S. B. Sabrina Ludwig, Article Automated Essay Scoring Using Transformer Models Sabrina Ludwig, keil Germany : Psych. 3. 897-915. 10.3390/psych3040056. , 2021.

[3] M. &. K. R. &. B. V. &. T. A. Agarwal, AutoEval: A NLP Approach for Automatic Test Evaluation System, india: 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 2021.

[4] E. Al-shalabi, An Automated System for Essay Scoring of Online Exams in Arabic based on Stemming Techniques and Levenshtein Edit Operations, Al-Balqa': ijcsi International Journal of Computer Science, 2016.

[5] M. &. S. F. Rahman, NLP-based Automatic Answer Script Evaluation, india: DUET Journal, 2018.

[6] R. A. Rajagede, Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature, Indonesia: KINETIK, 2021.

[7] F. &. Z. Y. &. Y. J. Dong, Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring, YangSingapore: CoNLL 2017, 2017.

[8] A. M. A. J. Christopher M Ormerod, Automated essay scoring using efficient transformer-based language models, usa: -, 2021.

[9] M. &. A. A. Al-Jouie, Automated Evaluation of School Children Essays in Arabic, ksa: Procedia Computer Science, 2017.

[10] A. &. A.-J. M. &. H. M. Azmi, AAEE - Automated evaluation of students' essays in Arabic language, ksa: Information Processing & Management, 2019.

[11] V. &. R. K. &. K. S. &. M. G. &. A. A. Mangu, Automated Evaluation of Telugu Text Essays Using Latent Semantic Analysis, turkey: Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021.

[12] P. M. Hendre, Efficacy of Deep Neural Embeddings based Semantic Similarity in Automatic Essay Evaluation, India: International Journal of Computing and Digital Systems, 2021.

[13] S. &. A.-S. B. &. A.-R. T. Awaida, Automated Arabic Essays Grading System based on F-Score and Arabic WordNet, Jordan: Jordanian Journal of Computers and Information Technology, 2019.

[14] P. M. Hendre, Efficacy of Deep Neural Embeddings based Semantic Similarity in Automatic Essay Evaluation, india: International Journal of Computing and Digital Systems, 2021.

[15] C. D. T. I. a. M. N. Sung, improving short answer grading using transformer-based pre-training, -: In International Conference on Artificial Intelligence in Education, 2019.

[16] H. A. A. Z. a. M. C. D. Ghavidel, Using BERT and XLNET for the Automatic Short Answer Grading Task, Canada: -, 2020.

[17] A. &. C. J. Olu, An Expert System for Automated Essay Scoring (AES) in Computing using Shallow NLP Techniques for Inferencing, -: International Journal of Computer Applications, 2012.

[18] J. R. T. F. T. M. N. G. B. Wahle, Identifying Machine-Paraphrased Plagiarism, -: Springer,, 2022.

[19] F. W. M. S. Y. S. W. E. Aramaki, Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT, -: -, 2021.

[20] M. T. R. K. N. C. Hiten Chavan, Plagiarism Detector Using Machine Learning, india: -, 2021.

[21] M. Grootendorst, KeyBERT: Minimal keyword extraction with BERT, -: Zenodo, 2020.

[22] N. K. N. K. N. Giarelis, A Comparative Assessment of State-Of-The-Art Methods for Multilingual Unsupervised Keyphrase Extraction, -: Springer, 2021.

[23] h. ashraf, project repository, cairo: -, 2022.