International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET229224

A Survey on Automatic Essay Evaluation System using Machine Learning

Nidhi Patel[1], Prof. Pradeep Gamit[2]
[1]PG Research Scholar, Software Engineering Department, Gujarat Technological University, Gujarat, India
[2]Assistant Professor, Computer Engineering Department, Gujarat Technological University, Gujarat, India

Article Info
Volume 9, Issue 2
Page Number : 160-166

Publication Issue :
March-April-2022

Article History
Accepted : 20 March 2022
Published: 30 March 2022

ABSTRACT

Manually assessment of descriptive answers in exam and assessment of an Essay requires more time and effort. In this era of E-Learning the Automated System for essay assessment is need of the time. There are many researches has been performed for this domain. In this paper we have reviewed some of the work related to this. Most of research uses Semantic Similarity Score and Sentimental Analysis for this purpose. Mostly NLTK and POS (Part of the Speech) is used. Various traditional algorithm of Machine Learning like SVM, Naïve Bayes, Random Forest etc. are used for performance classification. They have used measurement parameters in terms of KAPPA Statistics (QWK). We have also summarized methods related to Essay Evaluations with pros and cons in this paper. Feature selection methods and NLP attributes are also discussed.

Keywords: Machine Learning, Sentimental Analysis, NLP, Essay Evaluation

I. INTRODUCTION

With the ease of online education during pandemic, the need of automated essay and long answer evaluations are also in need. Large amount of enrolment in various courses from the students and submissions of their work requires a model that can be used for evaluation. As NLP has been used in so many applications related to long and short text, it can also be useful for Essay Evaluation also. NLP [1] has proven its efficiency in various applications like summarization of text, sentiment analysis, categorization of news etc. There are many options in machine learning that can be applied with NLP for essay evaluation also.

II. Various Methodologies

NATURAL LANGUAGE PROCESSING (NLP): In the field of technology, NLP makes systems to make communication toward users. This communication is done using human languages. We can define a language as a set of rules or we can also say a set of symbols called characters. In any human language, characters are combined in proper way, as a result there should be a way to convey the information for understanding the language to machine. The final

purpose for using NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable.

Following are the steps for NLP

- Audio data from Human Talk are captured by machine either by manual input or by recording devices.
- Recorded or input Audio is converted into text format for better understanding and meaning.
- Converted Text Data is pre-processed in which so many processes like stop word removal, double word removal is performed.
- The machine responds to the human by playing the audio.

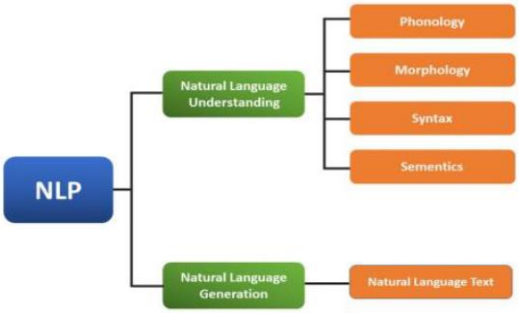

Figure 1 NLP Categories

Natural Language Understanding (NLU):

Phonology: Phonology is the branch of linguistics that describes the systematic arrangement of sounds. The term phonological comes from ancient Greek, the term phono denotes a voice or sound and the suffix - rosi denotes a word or speech. This level deals with the interpretation of phonetic sounds within words. There are three types of rules used in phonological analysis.

- Phonetic Rules - Used for sounds within words.
- Phoneme Rules - Used for pronunciation options when words are pronounced together.
- Rhyme Rule - Used to change the stress and intonation of a sentence.

Morphology : Other parts of a word are the smallest units of meaning known as morphemes. A morpheme is made up of the nature of a word that is started by the morpheme. The most important term in morphology is a morpheme, which is defined as "the smallest unit of meaning". For example, the word "misfortune". It can be divided into three morphemes (i.e., prefix, stem, and suffix), and each part of a word has a form of meaning. The prefix un stands for "non-existence" and the suffix ness stand for "the state of being".

Lexical: In lexical languages, the NLP system interprets the meaning of individual words in terms of lexical meaning and parts of speech. Language processing at this level uses the vocabulary of a language, which is a collection of individual vocabularies. A vocabulary is the basic unit of lexical meaning and an abstract expression unit of morpheme analysis, a set of forms or "meanings" that a morpheme takes. For example, "Duck" can take the form of a noun or verb, but it is a part-of-speech and lexical meaning and can only be obtained in context with other words used in a phrase/sentence.

Syntactic: The lexical analysis results can be used as input in this step. At this stage, the NLP system puts the words in the sentence into a grammatical form that is easy for humans to understand. You will need both a grammar and a parser at this level. There are many computer algorithms used to apply grammar rules to groups of words and extract meaningful words from them.

Semantic: Semantic processing interprets the possible meanings of a sentence by prioritizing the interactions between meanings at the word level of the sentence. This level focuses on interpreting the meaning of sentences rather than on individual word or phrase level analysis.

Discourse: The discursive level of NLP deals with text units longer than sentences. That is, it does not rely on

texts made up of multiple sentences, such as sequences of sentences that are part of a sentence that can be judged individually. Rather, discourse primarily focuses on the properties of the whole text that convey meaning by establishing connections between the components of a sentence.

Natural Language Generation (NLG):

The generation of Natural Language (NLG) is a process of meaningful, more easily and easier to understand, the production process of phrases, proposals and paragraphs. The NLG components are:

1) Speakers and generators - Generate a constructor or program that generates text that needs to use speakers or applications, and convert to a free syntax associated with a generator or generator or situation. In other words, it is converted into a natural language.
2) The training process of the component and level of the presentation includes the following path:
Select Contents: Information must be selected and must be included in the kit. You can delete block parts, depending on how this information is analysed in the representative unit of the device, but some others can be added by default.
Text Organization: You must organize information by text according to grammar. From the viewpoint of language relationships, such as modification, you must order continuously as well.
3) Application or Speaker - Only situation model is supported. Here, the speaker does not participate in language production, but simply initiates the process. Keep records, organize potentially important content, and distribute views of what you actually know. All of this forms a situation, highlighting the subset of sentences the speaker has. The only requirement is that the speaker understand the situation. The only requirement is that the speaker understand the situation.

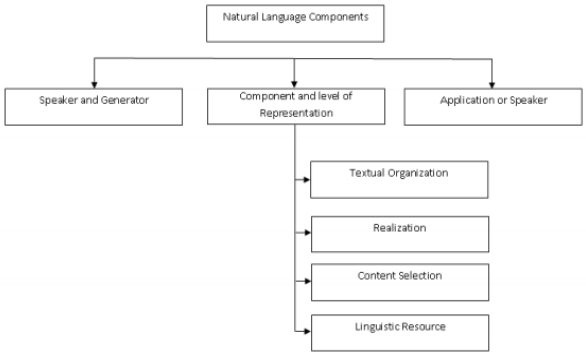

Figure 2 NLC Components

SOME COMMON PROBLEMS IN INFORMATION RETRIEVAL

Every automated system has its stumbling blocks, and information retrieval systems are no exception. Especially since we have to deal with the variability of human languages. Human language creates all kinds of complexity that is difficult to solve without resorting to the actual meaning of the words. Ambiguity is rampant if the system is unable to bring not only knowledge derived from the context of the text, but also knowledge of the real-world people carries with them. NLP can solve some common problems faced by information retrieval systems.

Too many synonyms.

We like to say the same thing in all sorts of different ways. The better the writer, the greater the variety. This creates a headache for searchers who have to guess how the author formulated the idea. We need a search engine that matches ideas, not words. Also, terms referring to the same thing may be used in different parts of the world or in different subject areas, such as truck and truck, elevator and hoist, pump and impeller, hypertension and hypertension. Exact matching systems skip critical operations unless you query for these other synonyms. The NLP system should be able to automatically complete queries with appropriate synonyms and place names.

Too many values

Most words have more than one value. In fact, Dr. Elizabeth Liddy, who created a search system for folding, is technically used in Napier Technical Services, and most words have an average of seven values. Think of how to use the word "table", "flight", "bank," or "seat". This phenomenon is called "polished". Information on information retrieval is a system that can determine the correct value of the word "edged". We use other words that mean the same thing, but we are also added to the chaos at the same time ("he can talk about the thick cat. When we investigate the word other than their context. For example, Barbara Quint frequently cites the search trap "Ask for terrorism and you get sports." If you are interested in African unrest, in the Boolean system (revolt, battle, rebellion, or skirmish) and (Southern (Africa), Kenya, Rwanda, Zambia, Zaire, or ...)

False Drops

Most modern commercial and web search technologies search for information without knowing its meaning. Matches a string (word) in a query to a document in the database to find an exact or best match. It's like trying to have a conversation with a parrot. Parrots can imitate words, but at best they associate words with cures or curses rather than with their own meanings. Boast ELIZA, a professional counselling system, also matches patterns and keywords, not meaning.

The result in both cases can be an inappropriate statement, a false positive. Boolean systems in particular suffer from this problem. If a searcher throws a fairly broad range with a simple AND query, if you search Japan's position on NAFTA you will find a Financial Times summary that discusses NAFTA in the first paragraph and Japan's problems with the yen's depreciation in the second paragraph. There are all search terms are in the document, but the document is unrelated. Both traditional and statistical systems suffer from this problem, but Boolean systems exacerbate the problem because they do not automatically find the proximity and frequency of terms.

Indexing Mismatch

Best practices give all documents of the same subject the same indexing condition. In fact, some studies have shown indexer consistency up to 50%. Given these facts of life, we need a system that can interpret and extract all variations of ideas. Spelling Variants and Errors What is the correct spelling for gray/grey or theatre/theater or aluminium/aluminum? What about spelling or typing errors that end up in print and in the database? This problem is exacerbated as we use optical character recognition to automatically add scanned text to a global database. Without careful checking, such scanned text can easily add 30 misspellings to any printed page. The purpose of the information retrieval system is to find the most relevant material to the user and exclude the least relevant material. We measure a system's ability to accomplish this feat with accuracy and recall. This is generally considered to be a "one of/or" sentence: more precise/less reproducible or more reproducible/less precise. Boolean systems are at one end of the spectrum. They find exactly what you are asking for. If you ask for what you want (which is not very common in the first request), you get what you want.

III. LITERATURE STUDY

We have reviewed some related works related to NLP and Sentimental Analysis for Essay Evaluation and other applications in various Indian Languages as well as English. Key points from various research work are summarized in Table 1 with their limitations.

In research work [1] they have used Graph Based Relationships with in the Essay Content. They have applied tokenize method and tested with SVM and RF.

Graph-based relationships within the essay's content and polarity of opinion expressions [1] is applied over it. They have used 23 salient features with high predictive power in their research work. Prediction is measured using Quadratic Weighted Kappa (QWK) in research work [1]. The system by researchers has produced a QWK of 0.793. They have applied algorithms SVM and RF.

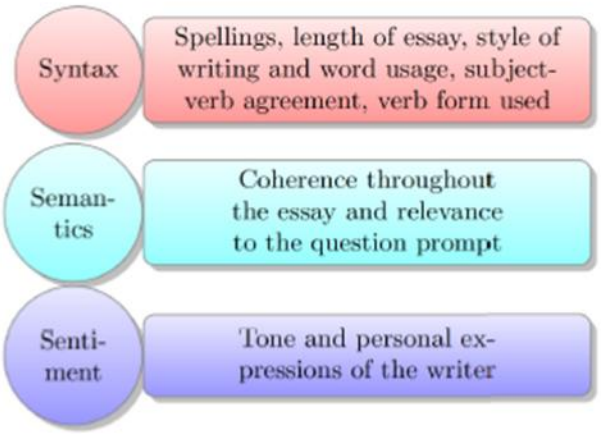

Figure 1 Overview of Syntax, Semantics and Sentiment [1]

Feature Set	Average QWK
Syntax	0.701
Semantic	0653
Sentiment	0.392
Syntax with Semantic	0.741
Syntax with Sentiment	0.744
Semantic with Sentiment	0.679
Syntactic, Sentiment and Semantic	0.793

Table 1 Result for Research work [1]

In research work [2] GLSA (Generalized Latent Semantic Analysis) method is applied for semantic analysis. They have also detected missing common words in compare to master data. Latent method has been applied in their work that improved satisfaction level. As a result, they got improved user satisfaction by 6.12%, from 74.44% to 80.56%.

In research work [3] well known neural network approach with LSTM is applied for this Decision Tree. They have measured their accuracy Quadratic Weighted Kappa and accuracy that is 85.74% and 70.80%. Transfer learning using Siamese dependency tree-LSTM is used in their system.

In research work [4] XGBoost algorithm is applied for boosting the result with 12 features. They have applied better feature selection method for their work. Accuracy of 66.87% achieved in the work.

Authors in research work [5] applied MO2E2 (Massive Open Online Essays Evaluations) model for essay evaluation. They have not applied any method to extract the meaning of words. They have performed statistical tests. Quadratic Weighted Kappa (QWK) parameter is used in their system.

In research work [6], researchers applied text mining and stemming process. Porter and cosine similarity algorithms are implemented over 35 essay data. As a result, they have achieved 97% accuracy.

In research work [7], they used 22 features with high predicting power over Dataset Kaggle's ASAP competition [7]. In their work they have applied Random Forest Algorithm.

In research work [8], Systematic literature review is done on automated essay scoring. They have studied various machine learning algorithms in review. according to their survey "Few researchers focused on Content-based evaluation, while many of them addressed style-based assessment"[8]. They have also studied feature extracting NLP libraries.

In research work [9], NLP (Natural language processing) is applied for feature selection. They have

generated 1592 linguistic indices. 93.9% is the figure that the researchers got in feature coverage.
In research work [10], Project Essay Grade (PEG), Intelligent Essay Assessor (IEA), Educational Testing service I, Electronic Essay Rater (E-Rater), C-Rater, BETSY, Intelligent Essay Marking System, SEAR and Paperless School are compared.

As per their review, only SAGrader and SAGE are the tools those use Semantics attributes. Most of tools are using Style and Content as their attributes. RNN has been tested with 97% accuracy with limited data, as per their work.

IV. COMPARATIVE STUDY

This section will go through the most commonly used machine learning algorithms for Essay Evaluation and Scoring. We have discussed well known traditional algorithms like SVM, Random Forest, XGBoost, LSA in our comparison work. We have mentioned general pros and cons of these methods. As each method has its own targeted type of data. So that we can say no any method can be declared as good or bad for all the types of contents like Small Text, Big Data etc. Selection of proper method for research work can be done based on aim of the research and nature of data to be tested.

Method	Advantages	Disadvantages
Support Vector Machine [1]	They are often more powerful and can scale to larger data sets. More efficient in spaces which are high dimensional. Suited for classes that have clear separation margins.	Doesn't work well when every data point's total features exceed total training data samples. Also for larger data sets, it's algorithm isn't suited. When target classes overlap or there's more noise in a data set, it starts to lag. Plus there's a lack of probabilistic explanation for support vector classifier's classification.
LSTM [3,8]	More Parameters like memory size, running time, gradient etc can be tuned for better performance.	LSTM Processing is more complex and required extra resources.
Random Forest [1,7]	Robust to outliers. Works well with non-linear data. Lower risk of over fitting.	Random forests are found to be biased while dealing with categorical variables. Slow Training
Latent Symantec Analysis [2]	LSA is capable of assuring decent results. In some cases, it is much better than plain vector space model. It works well on dataset with diverse topics.	It is a linear model, so not the best solution to handle nonlinear dependencies
XGBoost [4]	It works well on small data or it is also good when data with subgroups in big data or complicated data. Boosting process is used to improve the result.	It doesn't work so well on sparse data, though, and very dispersed data can create some issues.

Table 2 Comparative Study of Methods

V. CONCLUSION AND FUTURE DIRECTION

As we have researched over automated essay evaluation algorithm, we have founded some

challenging tasks that can improve overall performance of the system. Various researchers have applied different combination of algorithms.

As a part of future extension, we can decide the best methodology based on dataset need and size of the dataset for implementation. Here also some optimization for Hyper parameter can be applied to get better result.

VI. REFERENCES

[1]. H. K. Janda, A. Pawar, S. Du and V. Mago, "Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation," in IEEE Access, vol. 7, pp. 108486-108503, 2019, doi: 10.1109/ACCESS.2019.2933354.

[2]. J. Lemantara, M. J. Dewiyani Sunarto, B. Hariadi, T. Sagirani and T. Amelia, "Prototype of Automatic Essay Assessment and Plagiarism Detection on Mobile Learning "Molearn" Application Using GLSA Method," 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2019, pp. 314-319, doi: 10.1109/ISRITI48646.2019.9034652.

[3]. Wiratmo and C. Fatichah, "Assessment of Indonesian Short Essay using Transfer Learning Siamese Dependency Tree-LSTM," 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), 2020, pp. 1-5, doi: 10.1109/ICICoS51170.2020.9299044.

[4]. Y. Salim, V. Stevanus, E. Barlian, A. C. Sari and D. Suhartono, "Automated English Digital Essay Grader Using Machine Learning," 2019 IEEE International Conference on Engineering, Technology and Education (TALE), 2019, pp. 1-6, doi: 10.1109/TALE48000.2019.9226022.

[5]. J. Brito, J. Alves, C. Badue and E. Oliveira, "An Architecture for Massive Essays Evaluations," 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), 2021, pp. 1-6, doi: 10.23919/CISTI52073.2021.9476467.

[6]. N. D. Arianti, M. Irfan, U. Syaripudin, D. Mariana, N. Rosmawarni and D. S. Maylawati, "Porter Stemmer and Cosine Similarity for Automated Essay Assessment," 2019 5th International Conference on Computing Engineering and Design (ICCED), 2019, pp. 1-6, doi: 10.1109/ICCED46541.2019.9161090.

[7]. R. Bhatt, M. Patel, G. Srivastava and V. Mago, "A Graph Based Approach to Automate Essay Evaluation," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020, pp. 4379-4385, doi: 10.1109/SMC42975.2020.9282902.

[8]. Ramesh, D., Sanampudi, S.K. An automated essay scoring systems: a systematic literature review. Artif Intell Rev (2021). https://doi.org/10.1007/s10462-021-10068-2

[9]. Kumar, Vivekanandan and Boulanger, David, title=Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value, Frontiers in Education, 2020, doi=10.3389/feduc.2020.572367 https://www.frontiersin.org/article/10.3389/feduc.2020.572367

[10]. Srivastava, Kshitiz and Namrata Dhanda. “An Analysis of Automated Essay Grading Systems.” (2020). International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6

[11]. “NLP Processing”, https://www.ibm.com/topics/natural-language-processing.

[12]. “NLP Details”, https://www.javatpoint.com/nlp

[13]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6320437/

[14]. “Ontology”,https://riffyn.com/blog/the-importance-of-controlled-ontology

[15]. https://www.kaggle.com/aroraaman/quadratic-kappa-metric-explained-in-5-simple-steps/notebook

Cite this article as :

Nidhi Patel, Prof. Pradeep Gamit, "A Survey on Automatic Essay Evaluation System using Machine Learning ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 2, pp. 160-166, March-April 2022. Available at

doi : https://doi.org/10.32628/IJSRSET229224

Journal URL : https://ijsrset.com/IJSRSET229224