International Journal of Future Generation Communication and Networking
Vol. 13, No. 3s, (2020), pp. 44–52

Online Score Prediction System for Descriptive Answers

Rutuja Jadhavrao[1], Akshata Kulkarni[2], Uma Deshpande[3]

Department of Computer Engineering, Sinhgad Institute of Technology and Science, Narhe

kulakshata14@gmail.com[1]

rutujadhavrao7677@gmail.com[2]

umadeshp@gmail.com[3]

Abstract

In today's era online learning is prevalent in the educational field. There are many online platforms which provide online objective tests. But, very few platforms which deliver descriptive answers test assessment. It is difficult to assess the scoring of these tests, quizzes or complex descriptive answers. The proposed system should be able to predict scores for descriptive answers and make use of data mining techniques. This paper presents text classification for scoring descriptive answers. The product can be used for evaluating the descriptive answers by using TF-IDF and kNN. The similarity will be calculated between actual and expected answers based on word matching and word order. If the pair of the text matches and has the same order, the similarity is high. To reduce human efforts to analyze descriptive answers, we are developing a system that will automatically evaluate the answers and predict scores using TF-IDF and kNN.

Keywords— Text matching, kNN (k-Nearest neighbor), TFIDF (term frequency-inverse document frequency)

I. INTRODUCTION

Nowadays E-learning is becoming popular in the educational field. E-learning adapts the use of computers and networks to improve the quality of education. The motive of E-learning is to allow the student to learn independently at any place at any time. In order to examine the skills and knowledge of the student, the test is important. The questions of testing can be classified into objective and subjective. Objective test questions or closed-response questions are the questions that provide many choices. The objective question is quickly and easily scored by a computer. In contrast, subjective test questions or open-response questions require humans to evaluate the answer. The answer is complex as it includes many words. In order to solve this problem, we are using an automatic scoring approach for descriptive answers using Machine Learning (ML) algorithm. Machine learning is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

TF-IDF and kNN algorithms use synonyms, pronominal reference in order to improve the accuracy of predictive score of the descriptive answer. The main process of the proposed algorithm includes three main steps of: generates supplied test set, split the answer text to the list of words/phrases, and finally apply TF-IDF for keyword extraction and k-nearest neighbor classifier for score prediction with the proposed similarity measure. The study defines the available similarity measures for KNN algorithms based on word matching and word order. The algorithm can be applied to short and long subjective answers.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

II. LITERATURE SURVEY

This section presents the related work carried out by the other researchers in the field of imbalanced data classification using various methods such as: (A) Text Matching, (B) TF-IDF, (C) k-NN. These methods of classification from different reference papers written by researchers are:

Zhenzhong Li et al. [1] has proposed a text classification algorithm to classify the disorder data from the news articles. In this paper, the author did research about the news text classification. The authors propose a news text classification model based on Latent Dirichlet Allocation (LDA). Due to the dimension of the news texts being too high, this model uses topic model to make text dimension reduced and get features. At the same time, the authors also did research on Softmax regression algorithm to solve multi-class of text problems in our life and make it as model's classifier. The authors evaluate proposed models on a real news dataset and the result of the experiment shows the improved model performs relatively well. The model can effectively reduce the features dimension of the news text and get good classification results.

In [2], the author has described the string vector based version of the kNN as an approach for text categorization. Traditionally, texts should be encoded into numerical vectors for using the traditional version of kNN, and encoding so leads to the three main problems: huge dimensionality, sparse distribution, and poor transparency. In order to solve these problems, in this paper, texts are encoded into string vectors, instead of numerical vectors. The similarity measure between string vectors is defined, and the kNN is modified into the version where string vector is given its input.

Soha M. Eid et al. [3] have discussed the importance of linguistic features in Automated Essay Scoring (AES) system. AES is the solution to a tedious and time consuming activity of manually scoring students essays. Authors focus on the importance of the linguistic features in AES. The AES system contains a set of 22 lexical features which are captured from different writing qualities. These features are : Richness of content, Complexity of term usage, Orthography, Text complexity and Essay Organization. There are three commercial approaches based on AES : Electronic Essay Rater, Intellimetric and Intelligent Essay Assessor(IEA).QWK(Quadratic Weighted Kappa) is a measuring attribute for this comparison. Disadvantage of this approach is discussed in the paper is that the number of nouns degrades AES system performance.

In [4] the author has combined the traditional feature words weight calculation method and analyzed the shortcoming of traditional TF-IDF algorithm. The author discusses the advantages and disadvantages of the traditional TF-IDF algorithm. Aiming at the shortage of the TF-IDF algorithm, the author modify the traditional TF-IDF algorithm formula. The modified algorithm excludes the inner impact to disturb characteristics, adding the concept of intra-class dispersion. Using TF-IDF algorithm and improved algorithm experiments were carried out, which shows superiority of the improved algorithm. But there are a lot of other aspects that can be improved like the reduction methods. Authors summarize the existing term weight calculation method. TF-IDF is an effective method used to reflect how important a word is to a document.

Chiang Rai and Kittakorn Sriwanna[5] have discussed the text classification for subjective scoring using k-Nearest Neighbor algorithm.Online testing is the main part in online learning, e-Learning. The modern problem is the answer scoring of assessments, tests, and quizzes. Subjective test is a complex question, which requires human judgement evaluation. It requires extensive time to score the answers. The proposed approach makes use data mining techniques of k-nearest neighbors (KNN) algorithm to predict the score. The proposed algorithm splits text of subjective answers to

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

many words/phrases using a dictionary, which can cope with Thai and English languages. After that, KNN algorithm is applied with the proposed similarity algorithm. The proposed similarity is based on word matching and word ordering. If the pair of text match words/phrases and have the same order, the similarity is high.

In [6] the author Fadi Yamout and Rachad Lakkis have discussed the TFIDF weighting techniques in document retrieval. In information retrieval, documents are usually retrieved using lexical matching which matches where words in a user's query with words found in a set of documents. A significant model used in information retrieval is the vector space model where these words are represented as a vector in space and are assigned weights using a favorite weighting technique called TFIDF (Term Frequency Inverse Document Frequency). In this paper, authors have devised three new weighting techniques to improve the TFIDF weighting technique. The first technique is Dispersed Words Weight Augmentation (DWWA) which gives more weight to the words distributed in most of the document's paragraphs; authors have considered that those words are more significant than words found in a few paragraphs. The second technique is called Title Weight Augmentation (TWA) which gives more weight to the words found in the document's title and first paragraph. The third technique is called First Ranked Words Weight Augmentation (FRWWA) which increments further the weight of the most frequent words in a document.

Text scanning approach for exact string matching approach was introduced by Muhammad Zubair, Fazal Wahab, Iftikhar Hussain and Junaid Zaffar et.al[7]. Exact String matching is an important subject in the domain of text processing and an essential component in practical applications of computer systems. In this research, authors have proposed a new algorithm to solve the problem of exact string matching by scanning text string for last and first characters of pattern in its preprocessing phase. The matching phase of TSPLFC (Test Scanning for Pattern Last and First Character) compares pattern with text window from both directions simultaneously.

Taeho Jo[8] has suggested a kNN approach for extracting keywords. The approach is concerned with the table based KNN as the approach to the keyword extraction task. The keyword extraction task is viewed as an instance of word classification, and it is discovered that encoding words into tables improved the word categorization performance. In this paper, words are encoded into tables and the correspondingly modified version of kNN is applied to the keyword extraction task. As the benefits from this research, like the case in the general word categorization, authors expect the better performance in the keyword extraction, as the special word classification.

Majed AbuSafiya[9] proposed a string matching algorithm based on letters' frequencies of occurrences. The varying frequencies of occurrence of letters in the words in natural languages to propose a new string matching algorithm. Unlike standard string matching algorithms that compare letters in the pattern against the text in fixed order, the proposed algorithm ranks the letters of the pattern according to their frequency of occurrence. The letters of patterns are then matched against text according to that ranking, starting with the letter with the least frequency of occurrence. The proposed algorithm significantly outperformed the KMP(Knuth-Morris-Pratt) algorithm in terms of number of comparisons.

Keyword Extraction using Table based k-Nearest Neighbors is introduced by Duke Taeho Jo et.al[10]. In this research, words are encoded into tables, instead of numerical vectors, as the approach to the keyword extraction. The keyword extraction is mapped into a binary classification task within a domain, and the task should be distinguished from the topic based word categorization. In this research, words are encoded into tables each of which consists of entries of text identifiers

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

and their weights, the KNN algorithm is modified by adopting the proposed similarity metric, and it is applied to the keyword extraction which is mapped into a binary classification. It is validated empirically that the proposed kNN version is better than the traditional version in extracting keywords from a text which is tagged with its own domain.

Zhau Yao, Fan Hangbo, Liu Lijun and Huang Qingsong discussed the fast string matching for very short patterns et.al[11]. Exact string pattern matching is the fundamental problem in computer science. The performance of existing string matching algorithms is poor. In this article, based on the most basic exact single pattern matching algorithm authors presented a serial improved algorithms by introducing the q-grams method, the loop unrolling method and modifying the smallest processing unit from byte to integer. Experimental results indicated that the new algorithm is obviously faster than very known algorithms for string patterns.

Table No.1

Reference No.	Algorithms used	Main Context	Strengths	Weaknesses
[1]	● LDA(Latent Dirichlet Allocation)	1. Automatic text classification methods.	1. Text processing 2. Topic model modelling 3. Feature Extraction.	1. Based on limited input data
[2]	● SVM(Support Vector Machine) ● kNN	1. kNN implementation using string vector. 2. Encode the text into numerical format.	1. Similarity matrix created 2. String vector 3. Semantic matrix created	1. Classifies the text only in specific domains
[3]	● TFIDF ● Latent Semantic Analysis	1. TF IDF algorithm 2. Simple 3. Higher accuracy rate	1. Eliminates the most common terms 2. Extracts only most relevant terms.	1. Based on limited input data
[4]	● TFIDF	1. Automated Essay	1. Eliminate text feature	1. Need to improve the

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

		Scoring(AES) 2. Captures aspects of writing qualities	separately 2. Feature set shortened without degrading the performance	reduction methods
[5]	● kNN	1. Online scoring for subjective answers using kNN	1. Text Matching 2. Apply kNN and evaluate score	1. Complex to implement
[6]	● TFIDF	1. Information retrieval in documents using TFIDF	1. Dispersed Words Weight Augmentation 2. Title Weight Augmentation 3. First Ranked Words Weight Augmentation	1. New weighing techniques may get complex in the next steps.
[7]	● BM(Boyer-Moore) ● BMH(Boyer-Moore Horspool)	1. Text scanning for exact string matching.	1. TSPLFC (Test Scanning for Pattern Last and First Character)	1. Comparatively slow method.
[8]	● kNN	1. kNN for extracting keywords.	1. Encoding words into tables. 2. Apply kNN for keyword extraction	1. Keyword extraction is complex process for table based kNN
[9]	● KMP(Knuth-Morris-Pratt)	1. String matching	1. Knuth-Morris-Pratt	1. Number of comparisons

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

		algorithm	(KMP) string matching algorithm 2. Frequency of occurrence based string matching algorithm	might make it a long process
[10]	● kNN	1. Keyword extraction using table based k-Nearest Neighbor algorithm	1. Mapping keyword extraction into binary classification 2. Domain dependant classification	1. Result may not be accurate
[11]	● BF(brute-force) ● HBF(Hybrid BF)	1. String matching algorithm 2. Short patterns matching	1. O-grams method 2. Loop unrolling method 3. Modifying the smallest processing unit from byte to integer	1. Complex to implement

III. TECHNIQUES FOR TEXT CLASSIFICATION

A. Text Matching

In the text matching the main task in natural language processing. It uses words, phrases, or sentences to produce the similarity score between the pair of texts. Text matching can start form word matching. It finds the same word between the pair of texts. In a high number of words, phrase matching, it finds the same phase using n-gram. In a high number of phrase, sentence matching, it matches the sentence between the pair of texts. It requires word and phrase matching to produce the matching score.In this study, text (subjective answer) can be written in Thai and English languages,

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

it is hard to use phrase matching and sentence matching. To find the similarity score, this study uses basic word matching.

B. TF-IDF Technique:

To compute the TF-IDF value for one document, we take that document and calculate the TF-IDF score for each unique word without stop words. For each unique word, or term,

1. Term frequency: This is used for measuring how frequently a particular term appears in documents.$(the\ term, document) = \frac{Number\ of\ times\ the\ term\ appears\ in}{a\ doctoral\ number\ of\ words\ in\ a\ doc}$.....eq1
2. Inverse Document Frequency: This is used for measuring how frequently a particular term appears in all documents.
3. On equation 1 calculate overall TFIDF of the word, TF-IDF= TF * ID

C. k-Nearest Neighbors Algorithm (KNN):

KNN algorithm is one of the simplest classification algorithms and it is one of the most used learning algorithms.KNN is in the top 10 algorithms in data mining. K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. The Algorithm finds the answer of a sample using the majority vote of it's neighbor's samples.

IV. PROPOSED SYSTEM

The proposed system has two actors-teacher and student. Both the teachers and students will create their accounts. The teachers will create the tests. The test(s) will be in the form of descriptive answers. The teachers will store the descriptive answers of these tests in the database. The students will attempt the test(s). Answers given by students will be compared to the answers set by teachers in the database. Using the Term Frequency-Inverse Document Frequency(TF-IDF) algorithm the keywords will be extracted from the students answer and compared with the corresponding answer set by the teacher. Based on the similarity of the matched keywords, the score of the test will be predicted using the k Nearest Neighbour(kNN) algorithm. The proposed system will be useful for the teachers and will reduce their efforts tremendously. Figure 1 gives the system architecture of the system and following points state that how the answers will be compared with actual answers.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

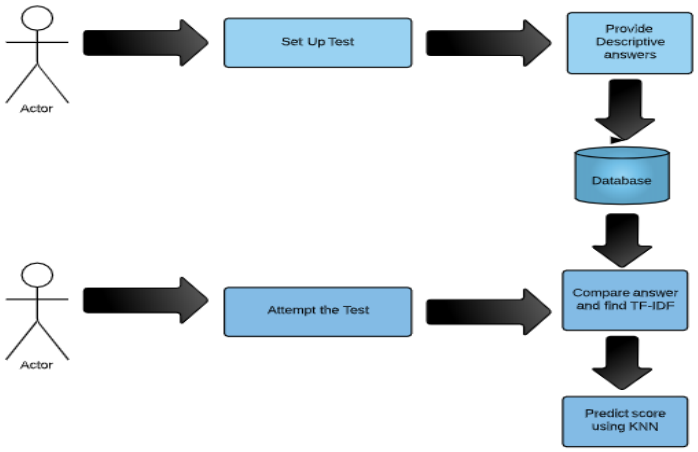

Figure 1. System Architecture

A. Document word frequency:

In the initial stage all the stopwords in answers will be removed and remaining words will be considered for the frequency count. Next step is to calculate the frequency of the whole document which is known as inverse document frequency. The overall TFIDF of the word is calculated using the formula,

TF-IDF= TF * ID

Now, each term will consist of some frequency weight.

B. Text categorization based on TF-IDF:

For kNN algorithm, documents will be mapped in a vector space format which will express the actual TF-IDF weights from the database.For the test results, the goal is to search the k-nearest neighbors from the database answers. So each database answer is also viewed as vector in the TF-IDF weight. The similarity between actual answer in the database and answer given by student is computed and the accurate score is given as an output.

V. CONCLUSIONS

This paper presents text classification for scoring descriptive answers using TF-IDF and kNN algorithm.Thus using TF-IDF and kNN,accurate score of the descriptive answers is predicted. Manual checking of descriptive answers can be quite cumbersome. Using the proposed online platform can help resolve this issue. The descriptive answers of the users can be scored accurately with the help of this online platform.

REFERENCES

[1] W. Horton, *E-learning by design*,pp. 38-40,2011

[2] L. Claudia and M. CC-rater, "Automated scoring of short-answer questions," *Computers and the Humanities*, vol. 37, pp. 92–96, 2003.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

[3] J. Z. Sukkarieh, "Using a maxent classifier for the automatic content scoring of free-text responses," in *AIP Conference Proceedings*, vol. 1305, no. 1. AIP, pp. 41–48.

[4] L. Bin, L. Jun, Y. Jian-Min, and Z. Qiao-Ming, "Automated essay scoring using the knn algorithm," in *Computer Science and Software Engineering, 2008 International Conference on*, vol. 1. IEEE, 2008, pp. 735–738,ISSN: 2349-7300.

[5] L. M. Rudner and T. Liang, "Automated essay scoring using bayes' theorem," *The Journal of Technology, Learning and Assessment*, vol. 1, no. 2, 2002.

[6] L. Kang, B. Hu, X. Wu, Q. Chen, and Y. He, "A short texts matching method using shallow features and deep features," in *Natural Language Processing and Chinese Computing*. Springer, pp. 150–159,2014.

[7] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text matching as image recognition." in *AAAI*, 2016, pp. 2793–2799.

[8] C. Haruechaiyasak and S. Kongyoung, "Tlex: Thai lexeme analyzer based on the conditional random fields," in *Proceedings of 8th International Symposium on Natural Language Processing*, 2009.

[9] C. Haruechaiyasak and A. Kongthon, "Lextoplus: A Thai lexeme tokenization and normalization tool," *WSSANLP-2013*, p. 9, 2013.

[10] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[11] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[12] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, pp. 137–142, 1998.

[13] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[14] T. C. Smith and E. Frank, *Statistical Genomics: Methods and Protocols*. New York, NY: Springer, 2016, ch. Introducing Machine Learning Concepts with WEKA, pp. 353–37

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC