International Journal of Management and Applied Science, ISSN: 2394-7926 Volume-9, Issue-8, Aug.-2023
http://iraj.in

ANALYSIS OF LINGUISTICS AND MATH FEATURES FOR CLASSIFICATION OF MATH WORD PROBLEMS: INSIGHTS AND FUTURE DIRECTION

[1]SHILPA KADAM, [2]PAVAN KUMAR SRUNGARAM, [3]SAI DHEERAJ Y, [4]MANISH S.S.S.R, [5]P.T.V. PRAVEEN, [6]SRIDHAR PAPPU, [7]DIPAK KUMAR SATPATHI

[1]Research Scholar, Bits-Pilani, Hyderabad, Telangana, India
[5,7]Associate Professor, Bits-Pilani, Hyderabad, Telangana, India
[1,2]Associate Professor, Upgrad-INSOFE, Hyderabad, Telangana, India
[6]Professor, UpGrad-INSOFE, Hyderabad, Telangana, India
[3,4]Professor, ATLAS SkillTech University, Mumbai, Maharashtra, India
[6]Graduate Student, Case Western Reserve University, USA
E-mail: [1]p20190508@hyderabad.bits-pilani.ac.in, [2]pavan.srungaram@upgrad.com, [3]saidheeraj.yanduru@case.edu, [4]manish.routhu@case.edu, [5]praveen@hyderabad.bits-pilani.ac.in, [6]sridhar.pappu@atlasuniversity.edu.in, [7]dipak@hyderabad.bits-pilani.ac.in

Abstract- Having math word problems (MWP)of varying difficulty levels can help instructors in identifying the knowledge levels of learners in teaching-learning systems. Given a large database of MWPs instructors spend significant time customizing content to meet learner needs. In this paper,Machine learning (ML) and AI-based methods are proposed to automatically classify math word problems. MWPs involve mathematical equations, symbols, and operators in addition to linguistic complexities. This paper presents various challengesin identifying and extracting relevant linguistics features as well as mathematical features that can aid in the automatic classification of MWPs. Based on our study we found that there is improvement in F1-score for a 3-level difficulty when compared to 5-level difficulty label of MWPs. Our study underscores the importance of further enhancing the feature set and developing appropriate mathematical tokenizersto improve the model performance.

Keywords - Math Word Problem (MWP), Difficulty Level of MWP, Adaptive Learning Systems (ALS)

I. INTRODUCTION

Math word problem (MWP) classification is central to instructional design and planning. It can aid in developing standardized assessments or benchmarks for evaluating students' math proficiency and further personalize learning in teaching-learning systems such as e-learning, Intelligent Tutoring Systems, Adaptive Learning Systems, etc. Word problems of varying difficulty levels are designed to align with student's knowledge state and learning progress, create targeted teaching materials,plan interventions, and assessments.Several studies have analyzed the performance patterns across different levels of MWPs and their impact on instructional methods and identified areas for further improvement.Instructors are required to spend significant time designing word problems of varying difficulty and there is some amount of subjectivity involved. To address this, there is a growing research interest to generate math word problems for a given difficulty level of MWPs using large language models in AI [1]. At the same time, another area of very active research is to provide solution to a MWP as discussed in [2] and [6].The challenge that remains is that the instructors are required to review and categorize them by difficulty level which is a significant effort.In this paper, we propose an approach to automatically classify the MWP by difficulty level using AI-based approaches. MWPs comprise of mathematical equations, expressions, numbers, symbols apart from linguistic complexity of a Natural Language. To predict the right difficultylevel,we propose to extract relevant features using Natural Language Processing methods and craft features to retrieve math cues. In the next section, we shall discuss the complexities of MWPs and various feature extraction methods to implement supervised classification methods.

II. DATA, FEATURES AND FEATURE ENGINEERING

To facilitate research in solving MWPs using AI models, there are several datasets that have been created such as Math23K, MathDNQ, DeepMind Datatset, MATH, etc. For classifying MWPs, we need the problem statement and labels that would define the difficulty level of questions. We specifically used MATH dataset, one of the widely used datasets created by [3] that consists of 12500 problems. The dataset hasMWPs from various topics such as Algebra, Probability, pre-Calculus and Geometry, etc. along with difficulty level and step-by-step solutions. The difficulty levels were defined on a scale of 1 to 5, where 1: beginners, 2: motivated/novice beginner, 3: intermediate, 4: High-level, and 5: Expert. As there can be subjective biases in labeling the difficulty levels of MWPs, one can argue the reliability, but we then use these levels as a baseline (ground truth/gold standard) and discuss alternate approaches in the next sections. In the proposed work, we specifically worked on classifying

algebra+ intermediate algebra + pre-algebraword problems that total to4236 problems across 5 difficulty levels, but our work can be easily extended to other types of MWPs. Some of the primary challenges with textual data are identifying features that would aid in classifying the MWPs. Natural language processing (NLP) techniques such as Bag of words, TFIDF, parts-of-speech (PoS), Named entities, word embeddings, etc. are well-known for better understanding of language and extractingfeatures.For English language, several measures such as Degree of Reading Power, Read-specific practice, Lexile, etc., were found in the literature that were defined to measure text difficulty.Later, comprehensive language analysis tools such as Coh-Metrix[4], Lexical Complexity Analyzer LCA, L2SCA,[5] etc., have been introduced in the literature to analyze multiple aspects of the language. Today, most state-of-the-art large language models (LLM) such as GPT3, etc., use word embeddings that convert the text into large numeric vector representations to capture the context and semantics of the language for performing down-stream tasks such as text classification, text generation, translations, etc. We found that [8] has presented Deep Neural Solvers (DNS) for MWPs using word embeddings but the features are black-box or inexplicable. Therefore, not only language complexity but identifying and extracting math-related features is challenging,and not much research is available in this direction, particularly for the classification of MWPs. As discussed in the paper [6],[7] there have been several math features and linguistics features extracted from MWPs to frame solvers, we take this as a motivation and present our analysis that involves extracting simple linguistic features along with mathematical cues for the classification of MWPs so that the models are explicable. Additionally, we present the analysis of different feature combinationsand their impact on the performance of classifiers.

Linguistic features: number of sentences, number of words in each sentence, number of words in each sentence, word length, large words (word having more than 6 characters), repeated occurrences of large words, PoS tags, etc.

Math features: number of equations in the question, number of variables, Boolean features such as presence of exponents, logarithms, modulus, inequality or equality, fractions, presence of math symbols and basic operators (+,-,*,/,!), caret, degree of the expression, numbers in words, number of places in digits, etc. To extract the equations, we employed regular expressions to define specific patterns and retrieve them accordingly. Through this approach, we identified the symbols, operators, and other elements within each equation, again utilizing regular expressions for accurate identification and extraction

Math vocabulary: created a list of math keywords such as: variable, alpha, beta, evaluate, etc. to compute count of such keywords used in each MWP.

Coh-Metrix: We also note that the level of difficulty of a math problem relies on language constructs such as- is the question direct or indirect, any confusing terms present, coreference, narration, use of nouns, etc. So, we used another widely used metric that producesover 100 indices coveringvarious aspects like word concreteness, syntactic simplicity, narrative, text easability, etc., in addition to general descriptive statistics. To acquire the Coh-Metrix for the MWPs, we employed the requests module to send HTTP requests to the Coh-Metrix web tool located at http://141.225.61.35/CohMetrix2017/. By utilizing Beautiful Soup (Python package) and an HTML parser, data was sent in batches to compute the measures. Note, the Coh-Metrix consisting of 106 features were collected, which were then incorporated into the existing feature set. We called the textual features extracted as ‘Linguistic features’ for the sake of better naming but Coh-Metrixhas both the descriptive measures and features that depict various language constructs.

Transforming numerical features to categorical:Instead of using the raw features that are in the form of counts, we transformed the counts into 3-levels such as “high”, “medium” and “low” using well-known data transformation techniques in the data science pipeline. For instance, the presence of math symbols is converted to 3 categorical levels.

Linguistic features from MWPs
18 features: ‘ADJ’, ‘ADP’, ‘ADV’, ‘VERB’, ‘DET’, ‘CONJ’,…etc.
4 features: ‘pronouns/sentence’, ‘proportion of pronouns’, ‘adj/sentence’, ‘proportion of adj’
4 features: ‘sentence count’, ‘words per sentence’, ‘large words’, ‘word count’
5Transformednumeric to categorical features: ‘sentence count’, ‘words per sentence’, ‘average word length’, ‘large words’, ‘word count’
Coh-Metrix: 106 features: Text easability principal components, Referential cohesion, LSA, Lexical Diversity, etc,
:Math features from MWPs
13 features: ‘count of variables’, ‘count of equations’, ‘count of digits in each number, ‘count of numbers’, ‘degree of equation’, ‘fractions’, ‘count of inequality’ symbols, ‘count of equality’ symbols, ‘count of ^ symbol’, ‘count of log’, ‘count of exponential symbol’, ‘count of modulus’,
9 boolean features: ‘has_exp’, ‘has_log’, ‘has_power’, ‘has_mod’, ‘has_frac’…etc.
2 Math vocab: ‘count of math vocab’, ‘categorical ‘math vocab’

Table.1. shows all the features extracted.

III. PROPOSED STUDY AND RESULTS

In this section, the experimental setup to analyze the impact of several feature combinations on the performance of classifiers is discussed. As we are dealing with a 5-class problem, assuming both Precision and Recall are equally important measures, we use F1-score as evaluation metric to measure the classification performance for comparison purposes.We present the model performance when the inputs are: (1) Linguistic features alone, (2) Coh-Metrix features alone, (3) Math features alone, (4) Linguistic features + Math features, (5) Coh-Metrix + Math features

We first read the algebra MWPs (JSON file formats) from MATH dataset. Here are the steps at a high level:

1. Input: raw linguistic and Math features
2. Target: 5 levels of difficulty
3. Train-Test split: 80:20 split
4. Implemented hyper-parameter tuning such as GridSearchCV to obtain best parameters and boost the model performance.

Classifier	Train F1- score	Test F1- score
Random Forest	0.54	0.33
Light GBM	0.83	0.37
Naive Bayes	0.31	0.30

Table 2: Random Forest classifier has given slightly higher weighted F1-scores.

From the above Table.2, we note the problem of overfit with much larger F1-score on train than test in spite of applying GridSearchCV. Suggested methods to mitigate thisare to select relevant features or add more data. Since there is no additional labelled data(ground truth) available, we worked with other feature combinations to identify features that would potentially improve F1-score. Several classifiers were applied for different feature combinations,but we presented F1-scores from best classifier as stated in Table3.We found that most of the classifiers suffered the problem of overfit and the F1-score was just over 30%. When only linguistic features or only math features were used for classification, the F1-score was about 35% and improved slightly i.e., 39% when both linguistic and math features were used together. Due to unavailability of labelled data , data augmentation techniques could not be employed to mitigate the issue. Despite implementing hyperparameter tuning for the classifiers, the resulting changes were not found to be significantly impactful. In the case of Random Forest classifier, we measured out-of-bag (OOB) error which is a better generalization of the model performance. Even though there is about 5% difference in the best model prediction (Light GBM) and Random Forest classifier, the OOB score was closer to test metrics. We also contemplate the 5 levels of difficulty defined in MATH dataset.

Feature Combination	Best Model	Train F1	Test F1
Coh-Metrix Features	Random Forest	0.86	0.35
Linguistic Features	Light GBM	0.71	0.36
Math Features	Random Forest	0.48	0.33
Coh-Metrix Features + Math Features	Light GBM	0.98	0.39
Linguistic Features + Math Features	Light GBM	0.76	0.38

Table 3. Best F1-scores obtained from classifiers for 5-levels of difficulty

During the manual inspection, we observed a hazy distinction between the levels, level 1 and 2 looked more similar and level 4 and 5 seemed to be similar. To analyze the hypothesis that fewer levels existed with MWPs, unsupervised methods such as spectral clustering and k-means clusteringwere implemented to identify groups based on a large feature set derived from MWPs. The features used in this experiment are all linguistic features like number of sentences, number of words, words per sentence, math features in numeric form like number of equations, number of variables, math vocab count, etc, POS features in numeric form like number of pronouns, number of conjunctions etc.Based on the Withinsumof squares for different clusters, and silhouette values, 3 clusters seem prominent. Figure 1. Shows a huge drop in Within-Sum of Squares values at 2 clusters, further at 3 and flattens after 4. So, we could decide to go with 2 or 3 groups. Figure 2. displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like the number of clusters visually. A silhouette score of value close to 1 indicates MWPs are well matched to the assigned cluster. Figure 3. Show the clusters patterns from spectral clustering.

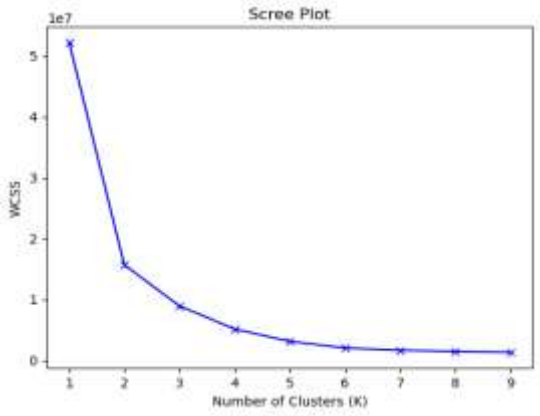

Figure 1: Scree-Plot. Plot shows a huge drop in Within-Sum of Squares values at 2 clusters using k-mean clustering algorithm. We obtained a similar plot for spectral clustering where the drop in Within-sum-squares is at 2 clusters.

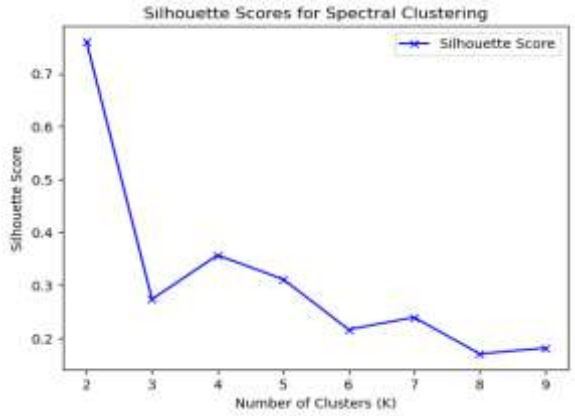

Figure 2: Silhouette score for k-means clustering

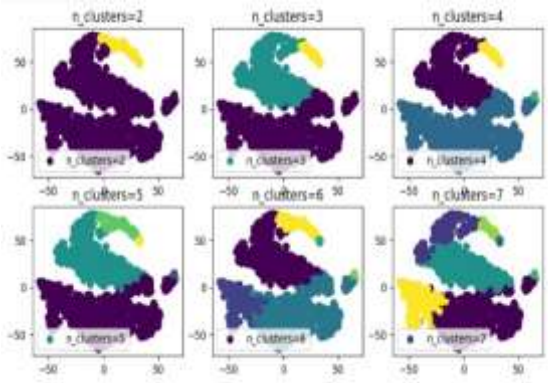

Figure 3. T-SNE plot showing different cluster formations from Spectral clustering.

The following distribution was observed between 5-levels of difficulty (ground truth) and the 3 clusters suggested by clustering algorithms. As we observe that new cluster distribution a shown in Table. 4 and Table. 5, the MWPs are spread across difficulty levels, we encoded the difficulty levels in the following combinations:
(Level 1:1, Level 2: (2,3), Level 3: (4,5))
(Level 1: (1,2), Level 2: 3, Level 3: (4,5))
(Level 1: (1,2), Level 2: (3,4), Level 3: 5) thus reduced the 5 difficulty levels into three levels, and applied classifiers.

	Cluster 1	Cluster 2	Cluster 3
Level 1	13	252	97
Level 2	15	528	289
Level 3	28	603	312
Level 4	23	581	375
Level 5	60	582	475

Table 4: Spectral clustering distribution of data

	Cluster 1	Cluster 2	Cluster 3
Level 1	346	3	13
Level 2	818	0	14
Level 3	915	3	25
Level 4	962	4	17
Level 5	1062	2	48

Table5: K-means clustering distribution of data

Experiment	Best Model	Train F1	Test F1
Coh-Metrix Features	Random Forest	0.86	0.35
Linguistic Features	Random Forest	0.76	0.53
Math Features	Naïve Bayes	0.49	0.47
Coh-Metrix Features + Math Features	Light GBM	0.95	0.52
Linguistic Features + Math Features	Light GBM	0.76	0.53

Table6:Best F1-scores obtained from classifiers for 3-levels of difficulty

Here, we note that the weighted F1-score improved in case of Linguistic features, Coh-Metrix +Math features, Linguistic features+Math features. The results improved from 39% to 53% for (Level 1:(1,2) Level 2:(3,4), Level 3:5)) combination of difficulty levels.Recent advancements in the field of NLP have given scope for massive exploration across verticals alike. We took the opportunity to exploit ChatGPT to provide the opinion on the MWP's difficulty level.Essentially, we have the opinion/classification from the large language model. The 1736 algebra MWPs were alone used to get the levels.The prompts given were:"Forget all the previous training. You are a high school math teacher, classify the problems given below on the scale of 1-5, where easy-1, moderate-2, intermediate-3, difficult-4, challenging-5." Similarly, we used a prompt for 3 classes as well where 1-easy, 2 moderate and 3-difficult. It is observed that in both the cases, most of the problems are categorized into one or two levels (observe that over 1300 problems are classified as moderate and intermediate problems by chatGPT as shown in Table 7. There is a large class imbalance when compared to the ground-truth, i.e., 5 difficulty levels.

Difficulty Levels classification by ChatGPT

Ground Truth	L1	L2	L3	L4	L5
Level 1	13	106	46	**12**	**1**
Level 2	24	176	101	29	6
Level 3	25	181	144	34	7
Level 4	11	114	184	75	11
Level 5	3	79	211	115	26

Table 7: Distribution of ground truth difficulty levels into ChatGPT responses of 5 difficulty levels

We observed a similar phenomenon, when ChatGPT was asked to classify the MWPs into three levels. Most of the problems were classified into level 2 as shown in Table 8 and Table 9

Difficulty Levels classification by ChatGPT			
Ground Truth	**L1**	**L2**	**L3**
Level 1	80	96	2
Level 2	109	212	15
Level 3	102	257	32
Level 4	50	296	50
Level 5	24	287	124

Table 8: Distribution of ground truth difficulty levels into ChatGPT responses of difficulty levels

Feature Combination	Train F1-score	Test F1-Score
5- Levels obtained from ChatGPT		
Coh-Metrix + Math Features	0.86	0.42
3-Levels obtained from ChatGPT		
Coh-Metrix + Math Features	0.9	0.65

Table 9: with target class based on ChatGPT classification, here are the F1-scores from Random Forest classifier

To summarize, we see that when onlyMath features were used, Random Forest gave F1-score of 33% for a 5-level difficulty MWPs and Naïve Bayes gave F1-score of 47%on 3-level difficulty MWPs. Similarly, when Coh-Metrix + Math features were used, Light GBM gave F1-score of 39% for a 5-level difficulty MWPs and 52% for a 3-level difficulty MWP. We note that there is a very good scope to further investigate two aspects in the current endeavor, that isright features and right level of difficulty to be able to classify MWPs. In this paper although we specifically worked on Algebra word problems, we plan to extend this analysis to other types of word problems within the MATH datasets. Given that we have only MATH dataset which has MWPs categorized into 5-levels, our analysis is limited to this dataset. However, with other advanced AI-based methods, we plan to build labelled data and explore features for better generalization and classification of MWPs.

IV. CONCLUSION

This paper presented challenges with extracting relevant features for classifying difficulty levels of MWPs. We have not seen any research related to this kind of work and hence we believe that this has research potential and scope to further explore the features that would really help improve the classification of MWPs. In conclusion, our study aimed to determine the difficulty level of Algebra math word problems by employing various features, including linguistic features, math features, and Coh-Metrix where, the MWPs were categorized into 5 levels of difficulty. However, the performance yieldeda weighted F1-score of less than 40%, indicating room for improvement.To address this, we experimented with reducing the number of difficulty levels from five to three, which led to further improvement in model performance. Interestingly, we observed a similar distribution of problems into the three levels when using clustering techniques or prompting ChatGPT for classification. Despite their different mechanisms (distance metrics on features versus AI implementation), both approaches predominantly grouped the majority of problems into one level, with the remaining few spread across the other two levels.This similarity in results highlights the need for additional features, particularly mathematical features, better classification of problems into their appropriate difficulty levels. In the case of generative models, appropriate mathematical tokenizers could prove useful in identifying the complexity of the problems.In summary, our findings underscore the importance of further enhancing the feature set, especially by incorporating mathematical features, to improve the accuracy and precision of difficulty level classification for Algebra math word problems. This research lays the foundation for future work aimed at refining models and developing more effective mathematical tokenization techniques for enhanced problem complexity identification.

REFERENCE

[1] Kurdi. G, Leo. J, Parsia. B, Sattler. U, andAl-Emari. S., “A systematic review of automatic question generation for educational purposes,” International Journal of Artificial Intelligence in Education, 30(1):121–204. 2020

[2] S. Mandal and S. K. Naskar, "Classifying and Solving Arithmetic Math Word Problems—An Intelligent Math Solver,"IEEE Transactions on Learning Technologies, vol. 14, no. 1, pp. 28-41, doi: 10.1109/TLT.2021.3057805. 2021

[3] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. “Measuring Mathematical Problem Solving With the MATH Dataset”. ArXiv. /abs/2103.03874. 2021

[4] Graesser, A. C., McNamara, D. S., &Kulikowich, J. M. “Coh-Metrix: Providing Multilevel Analyses of Text Characteristics”. Educational Researcher, 40(5), 223–234. https://doi.org/10.3102/0013189X11413260. 2011.

[5] Lu, X., “The Relationship of Lexical Richness to the Quality of ESL Learners’ Oral Narratives”. The Modern Language Journal, 96: 190-208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x. 2012

[6] D. Zhang, L. Wang, L. Zhang, B. T. Dai and H. T. Shen, "The Gap of Semantic Parsing: A Survey on Automatic Math Word Problem Solvers," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 9, pp. 2287-2305, 1 Sept. 2020, doi: 10.1109/TPAMI.2019.2914054. 2020.

[7] Xu. H, Y. Yi, Li. S, “Research on English Text Difficult Auditing Based on Artificial Neural Network”. Frontiers in Educational Research ISSN 2522-6398 Vol. 5, Issue 6: 93-96, DOI: 10.25236/FER.2022.050618. 2022.

[8] Y. Wang, X. Liu, and S. Shi, “Deep neural solver for math wordproblems,” in Proc. Conf. Empirical Methods Natural Language Process., 2017, pp. 845–854.

★★★