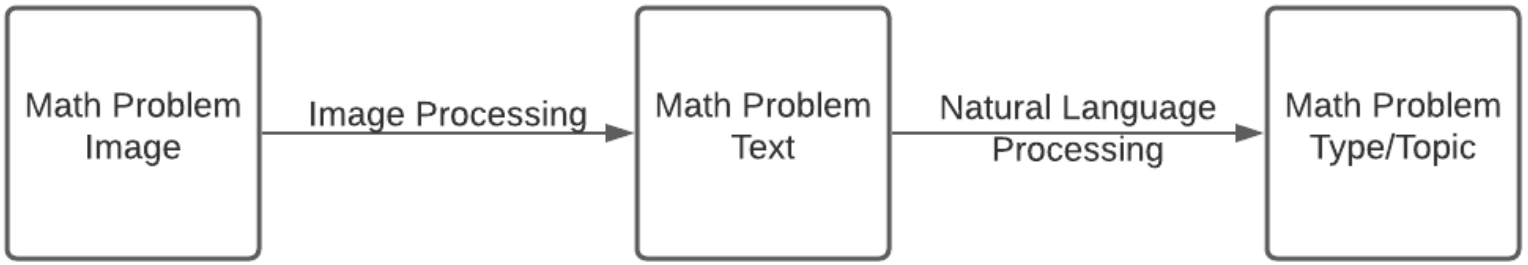
Classification for Math Problems

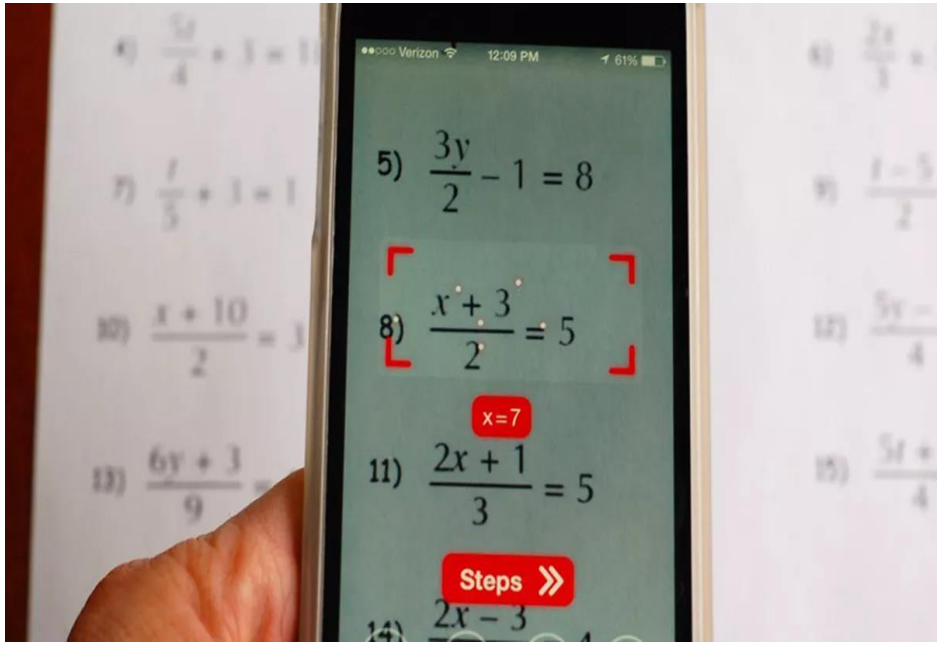
Problem Statement

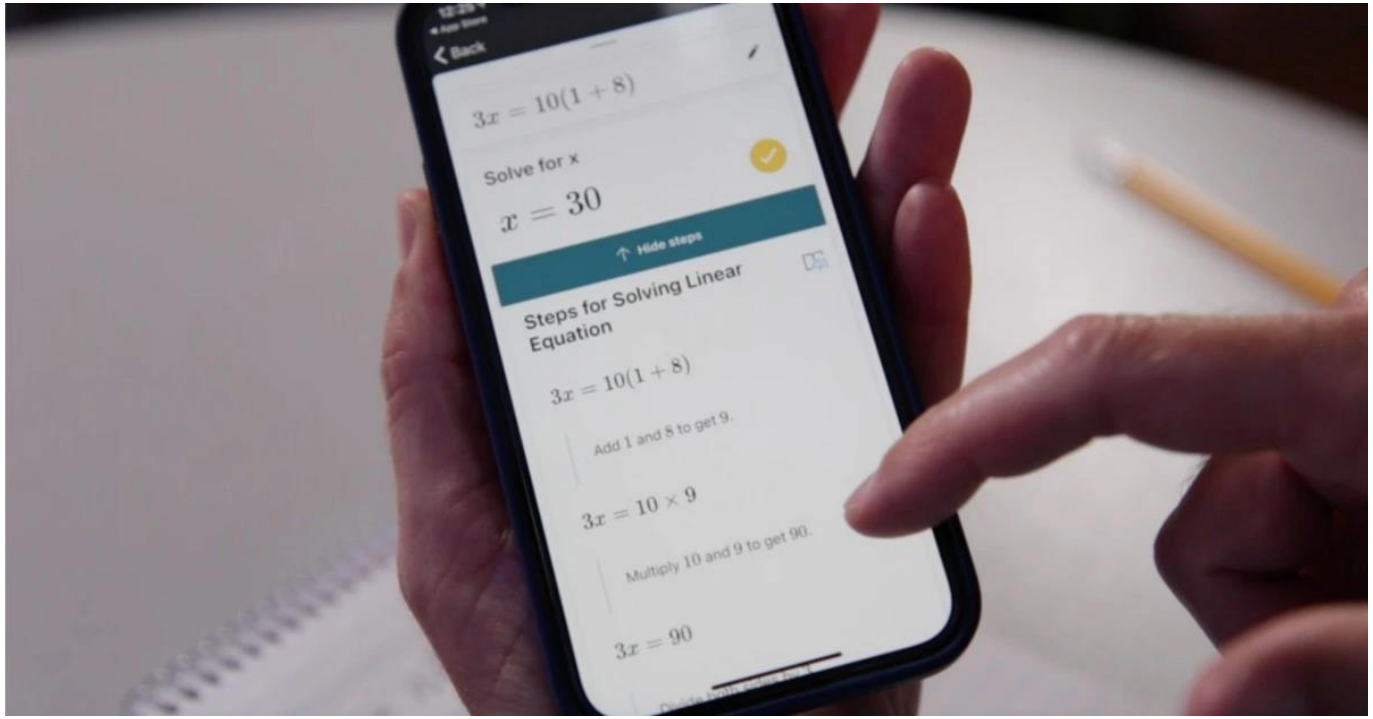
- We are looking to classify math problems by problem type



Processing Flow


(9x + 6)/3 = 5
x=1


Verizon
12:09 PM
61%
5) $\frac{3y}{2} - 1 = 8$
8) $\frac{x+3}{2} = 5$
x=7
11) $\frac{2x+1}{3} = 5$
Steps


Back
$3x = 10(1+8)$
Solve for x
$x = 30$
Hide steps
Steps for Solving Linear Equation
$3x = 10(1+8)$
Add 1 and 8 to get 9.
$3x = 10 \times 9$
Multiply 10 and 9 to get 90.
$3x = 90$

Use of Classification

- Recommend learning resources

- Future Considerations:
  - Provide answer check
  - Gamification?

Analytic Approach

- Muli-output/Multilabel classification problem

- We trained a Decision Tree algorithm to perform multi-output classification of text into categories of math problems.

Data Processing

- Sourced from GitHub community called Hugging Face
- Subset of larger data set
  - 56 categories of math problems
  - We took 10,000 questions from each category
- 56 categories of math problems
  - Examples: Algebra Linear 1d, Arithmetic Add or Sub, Calculus differentiate, Numbers is prime, Polynomials expand
- Problems are synthetically created

Some examples of math problems

	question
0	Find the first derivative of 2*d**4 - 35*d**2 - 695 wrt d.
1	Find the third derivative of -a**3*g**3*t**3 + 642*a**3*g*t**3 + 16*a**3*g*t**2 - 5*a**2*t**2 + a*g**3 wrt t.
2	What is the second derivative of 12518*f**3 + 3760*f?
3	What is the third derivative of -t**4 - 880*t**3 + 152*t**2 wrt t?
4	What is the second derivative of 2*c*n**2*z**3 + 30*c*n**2 + 2*c*n*z**2 - 2*c + n**2*z**2 - 3*n*z**3 - 2*n*z wrt n?
5	Find the first derivative of -1373*u**3 + 81 wrt u.
6	What is the derivative of -2612*k - 37?
7	Find the third derivative of w**6 - 2*w**5 + 579*w**4 - 5032*w**2.
8	What is the third derivative of -70*f**5 + 6*f**2?
9	Differentiate 745*b**4 - 287 with respect to b.
10	What is the second derivative of -600*l**4 - 5*l**3 - 6*l - 529?

	question
0	What is -5 - 110911?
1	What is -0.188 + -0.814?
2	Sum 259 and -46.
3	Sum -10 and -52539.
4	What is the difference between -2 and 251860?
5	-9259432 + 1
6	What is 1141.09 less than 1?
7	What is -5 - 72726?
8	Total of 0.3 and 170.7.
9	Work out 29.8 + -0.18.
10	What is 19450 minus -0.8?

	question
0	Solve 0 = 4*b + b + 15 for b.
1	Solve -3*d = -0*d + 3 for d.
2	Solve -4*h + 9 = 41 for h.
3	Solve 2514*m = 2508*m - 24 for m.
4	Solve -7*a + 6*a = 4 for a.
5	Solve 288*w - 298*w = -70 for w.
6	Solve -14*h = -4*h - 10 for h.
7	Solve 5*w + 3 = -2 for w.
8	Solve -15*f + 21*f - 12 = 0 for f.
9	Solve -22 = 6*c - 4 for c.
10	Solve 13*z - 7*z + 30 = 0 for z.

Model Form and Fitting

- Decision Tree algorithm:
  - LGBM Classifier (Light Gradient Boosted Machine Classifier)
- Two methods:
  - Classifier Chain and Multi Output Classifier
- Used TF-IDF measure
  - Determines how relevant a word is to a question and a category
  - Counts appearances of words in questions (Term Frequency) and across categories (Inverse Document Frequency)
  - Higher the TF-IDF score, the more relevant the word is

Model Results and Analysis

- Used Three-fold Cross Validation
- Calculated F-scores to measure accuracy of the fitted models

- F-score for LGBM with Multi Output Classifier: .949699
- F-score for LGBM with Classifier Chain: .948426

- Chose LGBM with Multi Output Classifier as better model

Model Results and Analysis

Classification Report:

- Most Categories have 1.00 Precision, Recall, & F score
- Some categories with poor scores likely due to related categories

	precision	recall	f1-score	support
algebra__linear_1d	0.95	1.00	0.97	1938
algebra__linear_1d_composed	0.84	0.85	0.84	1973
algebra__linear_2d	1.00	1.00	1.00	1949
algebra__linear_2d_composed	0.86	0.85	0.85	2084
algebra__polynomial_roots	0.93	0.47	0.62	2099
algebra__polynomial_roots_composed	0.68	0.91	0.78	1915
algebra__sequence_next_term	1.00	1.00	1.00	1980
algebra__sequence_nth_term	1.00	1.00	1.00	2049
arithmetic__add_or_sub	1.00	0.79	0.88	2056
arithmetic__add_or_sub_in_base	1.00	1.00	1.00	1989
arithmetic__add_sub_multiple	0.85	0.99	0.91	1955
arithmetic__div	1.00	1.00	1.00	1970
arithmetic__mixed	0.95	0.67	0.78	1981
arithmetic__mul	0.94	1.00	0.97	1904
arithmetic__mul_div_multiple	0.80	0.93	0.86	1999
arithmetic__nearest_integer_root	1.00	1.00	1.00	2007
arithmetic__simplify_surd	1.00	1.00	1.00	2091
calculus__differentiate	0.91	0.45	0.60	2046
calculus__differentiate_composed	0.64	0.96	0.77	1937

comparison__closest	1.00	1.00	1.00	2097
comparison__closest_composed	1.00	1.00	1.00	1994
comparison__kth_biggest	1.00	1.00	1.00	2027
comparison__kth_biggest_composed	1.00	1.00	1.00	1996
comparison__pair	1.00	1.00	1.00	1979
comparison__pair_composed	1.00	1.00	1.00	1980
comparison__sort	1.00	1.00	1.00	2010
comparison__sort_composed	1.00	1.00	1.00	2013
measurement__conversion	1.00	1.00	1.00	1966
measurement__time	1.00	1.00	1.00	1966
numbers__base_conversion	1.00	1.00	1.00	1960
numbers__div_remainder	1.00	1.00	1.00	1973
numbers__div_remainder_composed	1.00	1.00	1.00	2002
numbers__gcd	1.00	1.00	1.00	1977
numbers__gcd_composed	1.00	1.00	1.00	2104
numbers__is_factor	1.00	1.00	1.00	1937
numbers__is_factor_composed	1.00	1.00	1.00	2055
numbers__is_prime	0.99	1.00	1.00	1996
numbers__is_prime_composed	1.00	0.98	0.99	1993
numbers__lcm	1.00	1.00	1.00	2088

numbers__lcm_composed	1.00	0.99	1.00	1970
numbers__list_prime_factors	1.00	1.00	1.00	2058
numbers__list_prime_factors_composed	1.00	1.00	1.00	2042
numbers__place_value	0.99	1.00	1.00	2007
numbers__place_value_composed	1.00	0.99	0.99	2082
numbers__round_number	1.00	1.00	1.00	1990
numbers__round_number_composed	1.00	1.00	1.00	1991
polynomials__add	0.87	0.82	0.84	2017
polynomials__coefficient_named	1.00	1.00	1.00	1932
polynomials__collect	1.00	1.00	1.00	2008
polynomials__compose	0.87	0.80	0.83	1968
polynomials__evaluate	0.99	0.92	0.96	1962
polynomials__evaluate_composed	0.91	0.92	0.92	2012
polynomials__expand	1.00	1.00	1.00	1938
polynomials__simplify_power	1.00	1.00	1.00	1934
probability__swr_p_level_set	1.00	1.00	1.00	2026
probability__swr_p_sequence	1.00	1.00	1.00	2028

Conclusion

- Using a LGBM Classifier Algorithm to classify math problem text into concept categories
- The model worked well with some limitations
- Could improve the model and use case of the model by making changes to the dataset
  - Changing the categories and how they are comprised
  - Adding more questions from other sources to introduce diversity of math problems

Citations

Hugging Face Datasets.
Github: https://github.com/huggingface/datasets/tree/master/datasets

Disaster Message NLP Pipeline.
Github: https://github.com/ChristopherCochet/Disaster-Message-NLP-Pipeline/blob/master/notebooks/ML%20Pipeline%20Preparation.ipynb

Text Classification in Python by Miguel
Fernandez Zafra. Url: https://towardsdatascience.com/text-classification-in-python-dd95d264c802

A gentle introduction to OCR. Url: https://towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201aa.

Text Detection CTPN
by Shaohui Ruan. Github: https://github.com/eragonruan/text-detection-ctpn

Understanding Classification Report. Url: https://muthu.co/understanding-the-classification-report-in-sklearn/