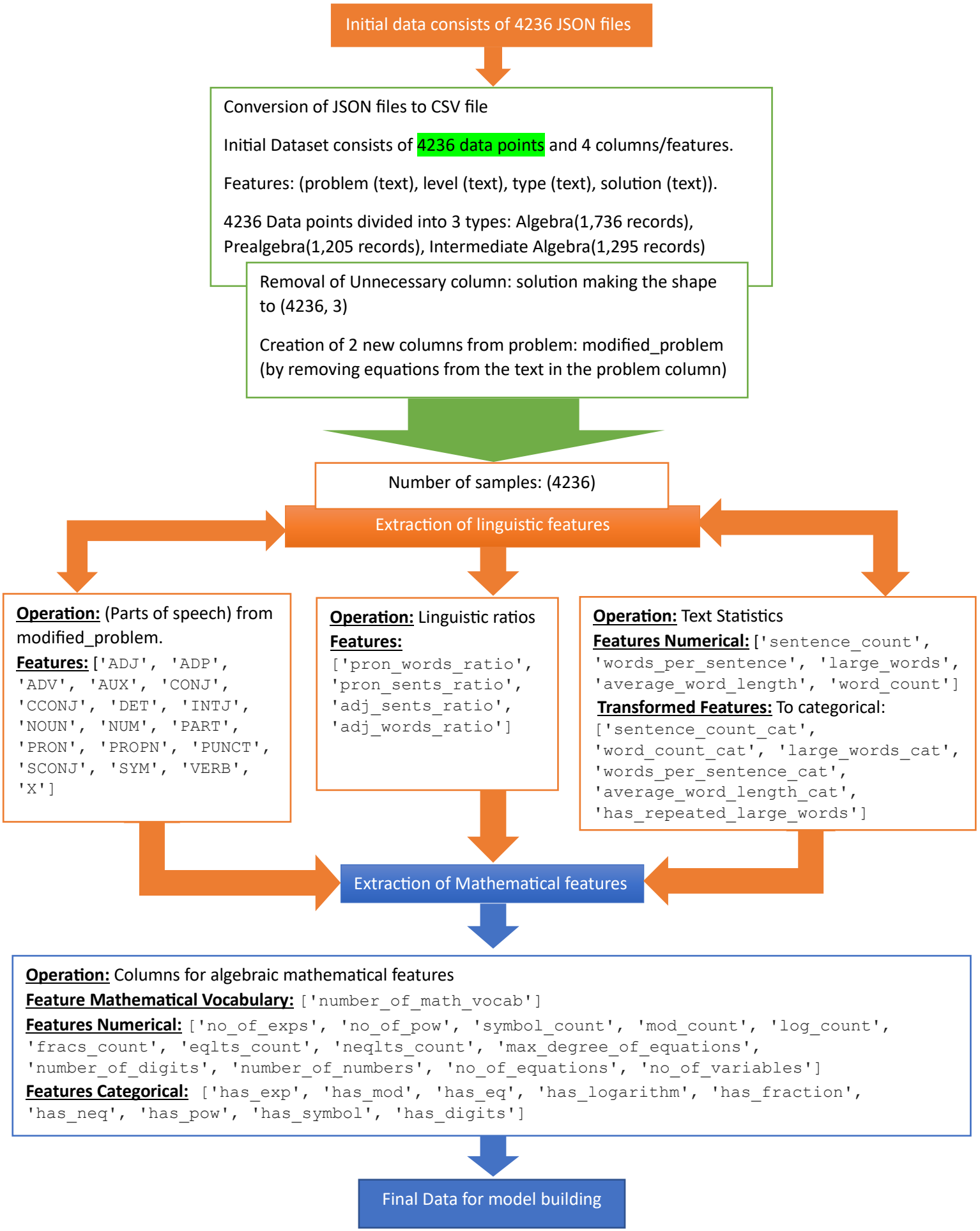
Initial data consists of 4236 JSON files
Conversion of JSON files to CSV file
Initial Dataset consists of 4236 data points and 4 columns/features.
Features: (problem (text), level (text), type (text), solution (text)).
4236 Data points divided into 3 types: Algebra(1,736 records), Prealgebra(1,205 records), Intermediate Algebra(1,295 records)
Removal of Unnecessary column: solution making the shape to (4236, 3)
Creation of 2 new columns from problem: modified_problem (by removing equations from the text in the problem column)
Number of samples: (4236)
Extraction of linguistic features
Operation: (Parts of speech) from modified_problem.
Features: ['ADJ', 'ADP', 'ADV', 'AUX', 'CONJ', 'CCONJ', 'DET', 'INTJ', 'NOUN', 'NUM', 'PART', 'PRON', 'PROPN', 'PUNCT', 'SCONJ', 'SYM', 'VERB', 'X']
Operation: Linguistic ratios
Features:
['pron_words_ratio', 'pron_sents_ratio', 'adj_sents_ratio', 'adj_words_ratio']
Operation: Text Statistics
Features Numerical: ['sentence_count', 'words_per_sentence', 'large_words', 'average_word_length', 'word_count']
Transformed Features: To categorical:
['sentence_count_cat', 'word_count_cat', 'large_words_cat', 'words_per_sentence_cat', 'average_word_length_cat', 'has_repeated_large_words']
Extraction of Mathematical features
Operation: Columns for algebraic mathematical features
Feature Mathematical Vocabulary: ['number_of_math_vocab']
Features Numerical: ['no_of_exps', 'no_of_pow', 'symbol_count', 'mod_count', 'log_count', 'fracs_count', 'eqlts_count', 'neqlts_count', 'max_degree_of_equations', 'number_of_digits', 'number_of_numbers', 'no_of_equations', 'no_of_variables']
Features Categorical: ['has_exp', 'has_mod', 'has_eq', 'has_logarithm', 'has_fraction', 'has_neq', 'has_pow', 'has_symbol', 'has_digits']
Final Data for model building