

BMS: Alignment of Attribute Names for Stability Testing

Amrutha Karuturi, Rohit Vernekar

1. Problem statement

The task is to efficiently clean and label each of the 7,730 entries from the dataset **Attributes** (with dimensions 7730 x 2) to one of the 32 standard names from the **standard_names** dataset (32 x 3), using various matching techniques. The primary objective was to assign the appropriate standard name to each entry based on its attributes and analysis information and organize them into three distinct datasets: **Accurate**(with accurate predictions), **Indecisive**(need human review), and **Trash**(junk data).

	Analysis	Attribute	Standard names
0	CE-SDS (NON-REDUCED)	HHL	CE-SDS (NON-REDUCED) HHL
1	CE-SDS (NON-REDUCED)	PURITY	CE-SDS (NON-REDUCED) PURITY
2	CE-SDS (REDUCED)	PURITY	CE-SDS (REDUCED) PURITY
3	SDS-PAGE (NON-REDUCED)	PURITY	SDS-PAGE (NON-REDUCED) PURITY
4	SDS-PAGE (REDUCED)	PURITY	SDS-PAGE (REDUCED) PURITY
5	RP-HPLC	PURITY (MAIN PEAK)	RP-HPLC PURITY
6	IEF	ACIDIC PEAKS	IEF ACIDIC PEAKS
7	IEF	BASIC PEAKS	IEF BASIC PEAKS
8	IEF	MAIN PEAK	IEF MAIN PEAK
9	ICIEF	ACIDIC PEAKS	ICIEF ACIDIC PEAKS
10	ICIEF	BASIC PEAKS	ICIEF BASIC PEAKS
11	ICIEF	MAIN PEAK	ICIEF MAIN PEAK
12	CEX	ACIDIC PEAKS	CEX ACIDIC PEAKS
13	CEX	BASIC PEAKS	CEX BASIC PEAKS
14	CEX	MAIN PEAK	CEX MAIN PEAK
15	AEX	ACIDIC PEAKS	AEX ACIDIC PEAKS
16	AEX	BASIC PEAKS	AEX BASIC PEAKS
17	AEX	MAIN PEAK	AEX MAIN PEAK
18	SE-HPLC	HMW	SE-HPLC HMW
19	SE-HPLC	LMW	SE-HPLC LMW
20	SE-HPLC	MONOMER	SE-HPLC MONOMER
21	SE-UPLC	HMW	SE-UPLC HMW
22	SE-UPLC	LMW	SE-UPLC LMW
23	SE-UPLC	MONOMER	SE-UPLC MONOMER
24	HIAC OR UV	PARTICULATE-MATTER >= 10-UM	PARTICULATE-MATTER >= 10-UM
25	HIAC OR UV	PARTICULATE-MATTER >= 25-UM	PARTICULATE-MATTER >= 25-UM
26	CELL-BASED BIOASSAY	POTENCY	POTENCY BY CELL-BASED BIOASSAY
27	ELISA	BINDING ACTIVITY OR POTENCY	POTENCY BY BINDING ELISA
28	SPR/BIOCORE	BINDING ACTIVITY	SPR BINDING ACTIVITY
29	PH	PH	PH
30	A280	PROTEIN CONCENTRATION	PROTEIN CONCENTRATION (A280)
31	POLYSORBATE 80	POLYSORBATE 80	POLYSORBATE 80

Standard names Dataset(32 x 3)

	analysis	Attribute_name
0	D_250475	IL2 INHIBITION ASSAY
1	D_95007196	PH
2	D_M00003744	ABATACEPT MAJOR BAND (REDUCED)
3	Y_SM_95011468_R	BIOASSAY
4	250684_CE_SDS_REDUC	SUM HEAVY AND LIGHT CHAIN
5	250580_TOTAL_PROT	PAAD
6	250580_TOTAL_PROT	SAMPLE 1 MASS
7	250684_CE_SDS_REDUC	SS RM 3 PURITY HC AND LC PEAKS
8	250683_CE_SDS_NON_RE	MAIN PEAK
9	Y_SM_95007441_R	B7 BINDING SPR

Attributes Dataset(7730 x 2)

2. EDA

In this task, we aimed to label each of the 7,000 entries from the dataset **Attributes** (with dimensions 7k x 2) to one of the 32 standard names from the **standard_names** dataset (32 x 3), using various matching techniques. The primary objective was to assign the appropriate standard name to each entry based on its attributes and analysis information.

We began by employing **direct matching with regular expressions (regex)**. In this approach, we searched for exact matching terms between the attributes and analysis fields in the **Attributes** dataset and the corresponding fields in the **standard_names** dataset. Specifically, we focused on matching the "analysis" and "attribute" values from **standard_names** with both the "analysis" and "attributes" from **Attributes**, as the two were found to be interchangeable in some cases.

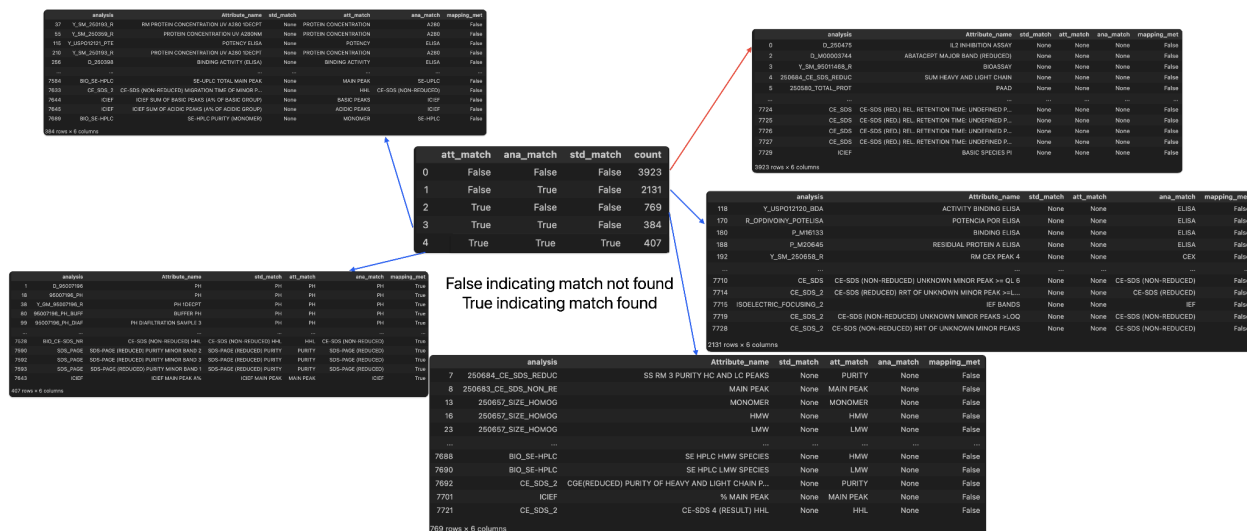


Figure: Showing the count of direct analysis and attribute matches found using regex.

Initial Results

From this initial matching approach, we were able to label 791 entries accurately. These matches were further validated manually to ensure their correctness and subsequently added to the training data.

Challenges Encountered

During this process, we encountered several challenges:

- Partial Matches:** Many of the analysis and attribute names in the test data were represented partially, causing difficulties in performing direct matches. For example:
 - The standard name "CE_SDS_REDUCED" appeared in the test data as variants like "CE_SDS_REDUC," "CE_SDS_R," and "CE_SDS(R)," making it challenging to match directly.
- Full Form Variations:** Some entries in the test data appeared in full forms or expanded versions, which were not an exact match to the entries in the **standard_names** dataset. Examples include:
 - "CEX" being represented as "CATION_EXCHANGE"
 - "IEF" being represented as "ISOELECTRONIC_FOCUSING"
 - "HMW" being represented as "High Molecular Weight"

3. These variations required more advanced matching strategies, such as fuzzy matching or semantic comparison(similarity between embeddings), to ensure accurate labeling.

3. Problems faced with data

1. Small Dataset

The initial dataset provided for training was very limited, containing only **one example per class**. However, training a large language model typically requires a significantly larger dataset, with around **2,000 to 10,000 labeled examples**. To address this challenge and enhance our dataset, we undertook the following steps:

1. **Manual Labeling:**

- We manually labeled **851 input samples** to expand the initial dataset.

2. **Data Augmentation (Abbreviation Substitution):**

- We augmented the data by substituting abbreviations with their corresponding full forms.
- This step increased the dataset size to **2,915 samples**.

3. **Handling Class Imbalance:**

- The dataset exhibited significant class imbalance, where certain classes had fewer examples than others.
- To address this, we applied **oversampling** to ensure an equal number of examples for all classes.
- This resulted in a balanced dataset with **20,520 samples**.

4. **Token Deletion for Robustness:**

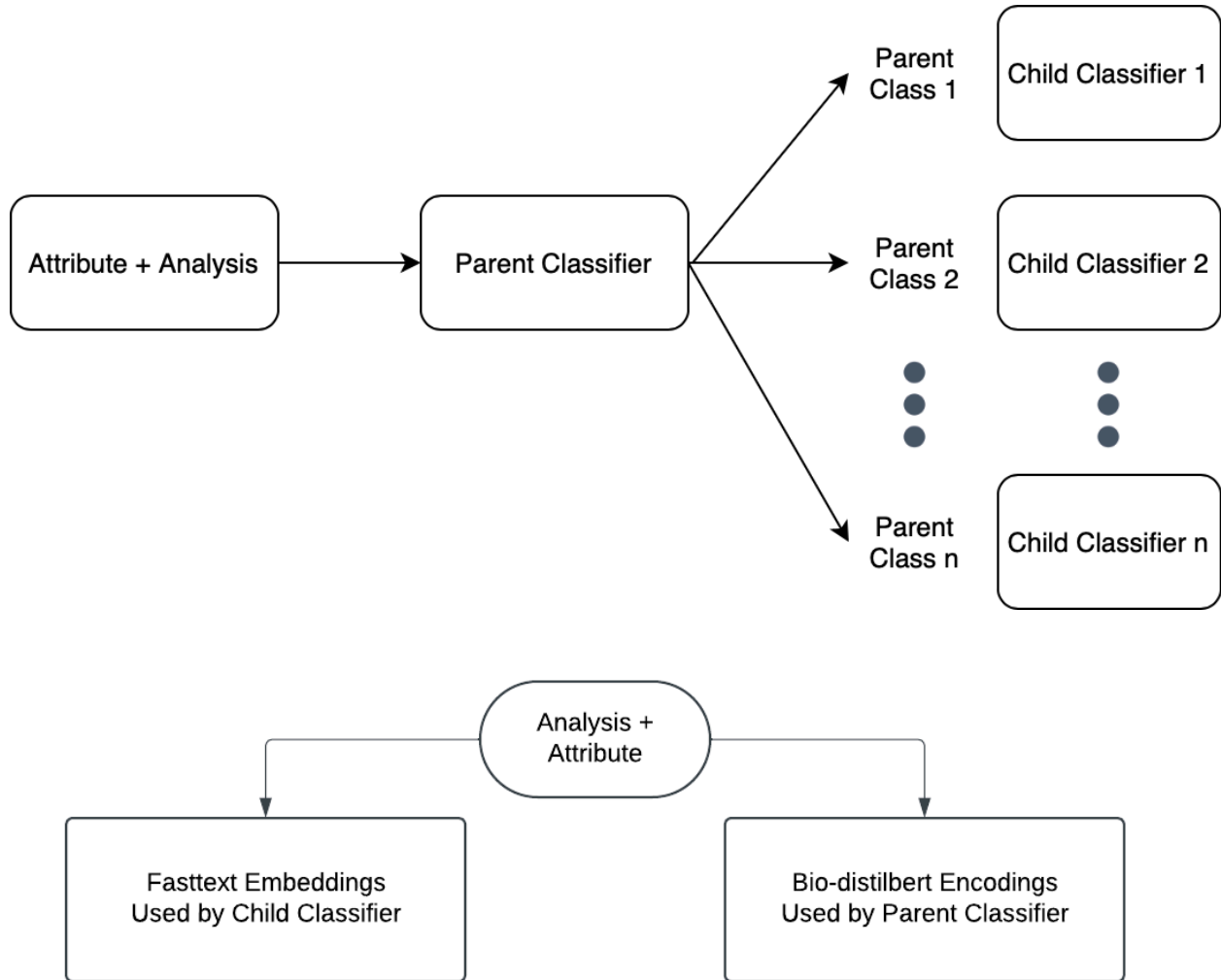
- To make the model more **robust and generalized**, we randomly deleted a few tokens from the input examples (attribute-analysis pairs).
- This augmentation further increased the dataset size to **45,456 samples**.

Through these steps, we significantly expanded and balanced the dataset, ensuring it was better suited for training a large language model.

2. Similarity between the Attribute-Analysis Pairs:

Standard names with similar attribute-analysis pairs (highlighted in the same color) can pose challenges for the model in accurately distinguishing between them during initial training. To address this, it is beneficial to first group these similar standard names together and then train a separate classifier to learn the finer distinctions among them.

4. Modeling architecture:



Parent classifier:

- Split 32 classes into 8 semantically cohesive groups.
- Grouping ensured classes in each group were closely related.
- Utilized BioBERT to get meaningful contextual embeddings of attribute-analysis pairs.
- Enabled precise classification into the appropriate class.

Child classifier:

- Leveraged FastText encoder for generating embeddings.
- Chosen for its ability to handle unknown words and misspellings effectively.
- Trained a custom neural network for each parent classifier.
- Enhanced precision by focusing on the unique characteristics of each group.

5. Results

The task is to classify inputs into one of 32 predefined classes and organize them into three distinct datasets: **Accurate**, **Indecisive**, and **Trash**. These datasets are defined based on the probabilities assigned to both the *parent class* and the *child class* for each input. The criteria for categorization are as follows:

1. **Accurate Dataset:**
 - Inputs with **high parent class probability** and **high child class probability**.
 - These predictions are considered accurate and reliable.
2. **Indecisive Dataset:**
 - Inputs with **high parent class probability** but **low child class probability**.
 - These predictions are uncertain and require manual labeling for resolution.
3. **Trash Dataset:**
 - Inputs with **low parent class probability**.
 - These data points are deemed irrelevant to any of the 32 classes.

This classification process ensures that data is appropriately segmented based on prediction confidence and relevance, facilitating further analysis or manual review where necessary.

Standard names	Accurate	Indecisive	Trash
CE-SDS (Non-Reduced) HHL	2000	2473	0
CE-SDS (Non-Reduced) Purity			
CE-SDS (Reduced) Purity			
SDS-PAGE (Non-Reduced) Purity			
SDS-PAGE (Reduced) Purity			
IEF Acidic Peaks	368	264	136
IEF Basic Peaks			
IEF Main Peak			
ICIEF Acidic Peaks			
ICIEF Basic Peaks			
ICIEF Main Peak			
CEX Acidic Peaks			
CEX Basic Peaks			
CEX Main Peak			
AEX Acidic Peaks			
AEX Basic Peaks			
AEX Main Peak			
RP-HPLC Purity	378	238	103
SE-HPLC HMW			
SE-HPLC LMW			
SE-HPLC Monomer			
SE-UPLC HMW			
SE-UPLC LMW			
SE-UPLC Monomer			
Particulate-Matter >= 10-um	0	495	35
Particulate-Matter >= 25-um			
Potency by Cell-Based Bioassay	202	211	191
Potency by Binding ELISA			
SPR Binding activity			
pH	77	0	120
Protein Concentration (A280)	220	0	201
Polysorbate 80	17	0	1
Overall	3262	3681	787

Parent class	Parent_class_prob threshold(Observed/ Suggested)	Child_class_prob threshold(Observed/ Suggested)
0	-	-
1	0.998	0.984
2	0.98	0.9659
3	0.9899	-
4	0.995	0.9999999
5	0.99	-
6	0.99	-
7	0.95	-

Above are the examples of threshold values derived from analyzing the data.

6. Future Work

- Incorporate domain knowledge to improve training data (for example Main Peak ~ Purity).
- Introduce irrelevant samples to help the model identify and classify irrelevant data effectively.
- FastText does not encode symbols (e.g., <, >, =). Develop a strategy to handle these symbols manually for improved accuracy.
- Address edge cases for more accurate and reliable classifications.