Text Mining Report

Galante Fabio 851242
Balbin Canchanya Gianni Eduard 901609

Master's Degree in Data Science
University of Milano-Bicocca
2023/2024 Academic Year

Contents

1 Introduction

The objective of this project is to perform multi-label text classification on a dataset consisting of research articles from the PubMed repository. Each article is annotated with multiple Medical Subject Headings (MeSH) labels, covering various biomedical domains. Additionally, topic modeling is conducted to uncover latent topics within the abstracts.

1.1 Dataset Description

This dataset downloaded from Kaggle includes around 50,000 research articles gathered from the PubMed repository.
Within this Dataset, there are numerous labels categorized as MeSH major, leading to a significant output space. In order to address this problem, the dataset creates 14 major root categories, namely:

- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic, and Therapeutic Techniques, and Equipment [E]
- Psychiatry and Psychology [F]
- Phenomena and Processes [G]
- Disciplines and Occupations [H]
- Anthropology, Education, Sociology, and Social Phenomena [I]
- Technology, Industry, and Agriculture [J]
- Information Science [L]
- Named Groups [M]
- Health Care [N]
- Geographicals [Z]

2 Data Quality and Exploration

This stage involves data cleaning, where we conducted various checks such as examining for duplicate to ensure uniqueness and checking for any missing value in the `abstractText` and label columns.
Then we illustrated the distribution of words for every article in the dataset:

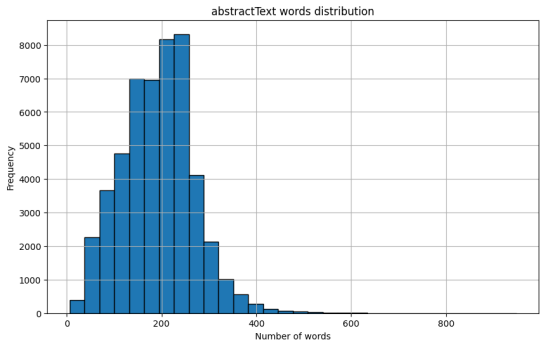

Figure 1: Distribution of the number of words per abstractText

The majority of articles range from 100 to 300 words. The distribution is asymmetric with a long right tail, indicating some abstracts much longer than average.

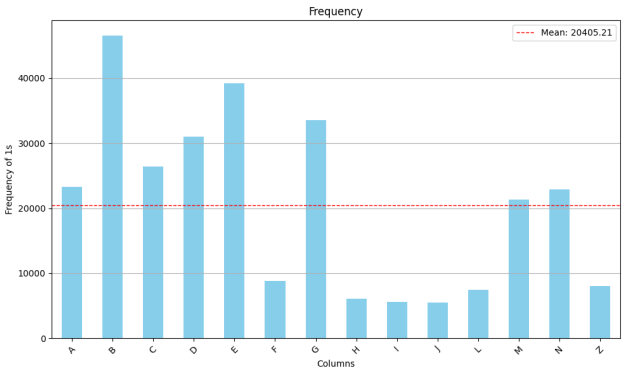

Figure 2: Frequency of 1 in MeSH labels

This plot shows how many articles present each MeSH label.

B have the highest frequency of about 45,000 occurrences, while the others are similarly distributed, some over and some below the mean of 20405.

2.1 Data Preprocessing

This phase involved different methods, including:

- **Tokenization**: we split the `abstractText` into individual tokens (words) for analysis.
- **Normalization**: converted all text to lowercase to ensure uniformity and removed punctuation and numerical characters from the text.
- **Stop-word removal**: we omit words which are not in a predefined dictionary to focus on meaningful words.
- **Lemmatization**: reduced words to their base form to standardize different forms of the same word.

Once all of these processes were executed, the result was integrated into a new column then defined as `Tokens`.
The following output shows the first 5 rows of this column:

```
0    [paraffin, embedded, tissue, section, patient,...
1    [present, study, conducted, determine, vitamin...
2    [occurrence, individual, amino, acid, dipeptid...
3    [lim, sun, introduced, microcapsule, coated, c...
4    [substantially, improved, hydrogel, particle, ...
Name: Tokens, dtype: object
```

Figure 3: First 5 rows of `Tokens` column

3 Multi-Label Text Classification

3.1 Problem Definition

The task is to classify each `abstractText` into one or more of the 15 MeSH categories. This is a multi-label classification problem due to the possibility of each abstract belonging to multiple categories.

3.1.1 Model Selection and Training

- Various machine learning models were considered, including logistic regression, support vector machines, and deep learning models like neural networks.
- Techniques such as one-vs-rest (OvR) and binary relevance were employed to handle the multi-label aspect.
- The dataset was split into training and test sets to evaluate model performance.

3.1.2 Evaluation Metrics

- **Accuracy:** The fraction of correct predictions among the total predictions.
- **Precision, Recall, and F1-Score:** Used to measure the performance of the model in terms of handling multiple labels per abstract.

First of all, we create a TF-IDF Matrix to establish a representation for the data we are using. By limiting the features to the top 1000 most frequent, we reduce the computational cost and improve the model's performance. We select the **abstractText** column as X and all major root categories as Y, inserting them into the matrix after being converted into an array format. We have separated the data into training and testing sets by splitting them with a ratio of 80-20.

3.2 Logistic Regression

Logistic regression is a binary classification algorithm that predicts the probability of an instance belonging to a class using a logistic function. In multi-label text classification, where each document can belong to multiple classes simultaneously, OvR extends logistic regression to handle multiple classes.

We show the classification report about this type of model:

Classification Report:

	precision	recall	f1-score	support
A	0.81	0.76	0.78	4657
B	0.95	0.99	0.97	9323
C	0.87	0.83	0.85	5332
D	0.91	0.90	0.91	6195
E	0.81	0.96	0.88	7796
F	0.86	0.62	0.72	1751
G	0.83	0.90	0.86	6707
H	0.66	0.13	0.22	1223
I	0.78	0.47	0.59	1127
J	0.78	0.29	0.42	1122
L	0.78	0.39	0.52	1462
M	0.87	0.87	0.87	4229
N	0.81	0.77	0.79	4516
Z	0.76	0.55	0.64	1608
micro avg	0.86	0.82	0.84	57048
macro avg	0.82	0.67	0.72	57048
weighted avg	0.85	0.82	0.83	57048
samples avg	0.86	0.83	0.83	57048

Figure 4: Classification Report for Logistic Regression

The model showcases impressive overall results, with specific classes (such as B, D) displaying high f1-score. Nevertheless, there are certain classes (such as H and I) that the model finds challenging, suggesting the necessity of additional tuning or data in order to enhance predictions for these particular classes.

3.3 Support Vector Machine

A Support Vector Machine (SVM) is a robust tool in machine learning that categorizes data into various groups based on an analysis of data features. Even in this case, we use the same combination and ratio for train and test sets. We report again the classification matrix for this model:

Classification Report:

	precision	recall	f1-score	support
A	0.79	0.70	0.74	4657
B	0.95	0.99	0.97	9323
C	0.86	0.81	0.83	5332
D	0.91	0.87	0.89	6195
E	0.78	1.00	0.88	7796
F	0.80	0.56	0.66	1751
G	0.81	0.89	0.85	6707
H	0.00	0.00	0.00	1223
I	0.76	0.39	0.52	1127
J	0.00	0.00	0.00	1122
L	0.79	0.23	0.36	1462
M	0.86	0.86	0.86	4229
N	0.81	0.75	0.78	4516
Z	0.72	0.51	0.60	1608
micro avg	0.84	0.80	0.82	57048
macro avg	0.70	0.61	0.64	57048
weighted avg	0.81	0.80	0.79	57048
samples avg	0.84	0.81	0.81	57048

Figure 5: Classification Report for SVM

The SVM model typically shows slightly inferior performance compared to the logistic regression model, demonstrated by lower macro and weighted averages.
The logistic regression model shows better overall performance (precision, recall, and F1-score) than the SVM model, suggesting more stable results across various classes.
The SVM model performs well overall but is slightly less effective than the logistic regression model in this dataset.
Furthermore this method is much slower, because it is really computationally heavy.

3.4 Random Forest

A Random Forest model involves multiple decision trees to make predictions in machine learning. Every tree is constructed using a random selection of the data and characteristics. The end result is determined by the most common prediction among the trees.. We set for the forest a number of decision trees equals to 100 and a fixed random state for reproducibility. Next, we describe its classification report:

Classification Report - Random Forest:

	precision	recall	f1-score	support
A	0.77	0.66	0.71	4657
B	0.94	1.00	0.97	9323
C	0.82	0.80	0.81	5332
D	0.86	0.85	0.86	6195
E	0.79	0.99	0.88	7796
F	0.83	0.26	0.40	1751
G	0.79	0.90	0.84	6707
H	0.50	0.00	0.00	1223
I	0.79	0.18	0.30	1127
J	1.00	0.00	0.01	1122
L	0.83	0.10	0.18	1462
M	0.83	0.87	0.85	4229
N	0.81	0.71	0.75	4516
Z	0.75	0.27	0.40	1608
micro avg	0.83	0.77	0.80	57048
macro avg	0.81	0.54	0.57	57048
weighted avg	0.83	0.77	0.76	57048
samples avg	0.83	0.78	0.79	57048

Figure 6: Classification Report for Random Forest

The logistic regression model outperforms the random forest in weighted average with a higher F1-score (0.83 versus 0.76), indicating a more balanced performance in terms of precision and recall.
Both models show comparable accuracy, with the logistic regression model slightly outperforming the random forest model in weighted average precision with a score of 0.85 versus 0.83.
In weighted average, logistic regression shows better identification of true positives with a recall of 0.82, surpassing the random forest's 0.77.
Once more, logistic regression demonstrates a more even performance across various metrics and categories.

3.5 Bigrams

Using bigrams instead of single words allows to capture local context and can provide more meaningful insights by considering word pair relationships. This time we need a different TF-IDF matrix, so we created it using bigrams not just single words. Then, we split the data into train and test sets again,

trained a logistic regression classifier, made predictions on the test data, and printed a classification report to evaluate the model's performance.
So, its classification report shows:

Classification Report:

	precision	recall	f1-score	support
A	0.72	0.59	0.65	4657
B	0.94	1.00	0.97	9323
C	0.81	0.71	0.76	5332
D	0.78	0.86	0.82	6195
E	0.78	0.99	0.87	7796
F	0.72	0.27	0.39	1751
G	0.76	0.90	0.82	6707
H	0.55	0.02	0.03	1223
I	0.67	0.20	0.31	1127
J	0.60	0.03	0.06	1122
L	0.71	0.12	0.21	1462
M	0.83	0.73	0.78	4229
N	0.77	0.65	0.71	4516
Z	0.66	0.30	0.42	1608
micro avg	0.80	0.74	0.77	57048
macro avg	0.74	0.53	0.56	57048
weighted avg	0.79	0.74	0.74	57048
samples avg	0.80	0.75	0.76	57048

Figure 7: Classification Report for Bigrams

In general, the logistic regression model tends to perform better than the bigrams model, especially in terms of recall and overall F1-score.
Both models have similar weighted average precision with the logistic regression model slightly ahead (0.85) compared to the bigrams model (0.79).
In weighted average, the logistic regression model outperforms the bigrams model with a higher recall rate of 0.82 compared to 0.74.
Both models face challenges when dealing with class H.

3.6 Classification with word embeddings

We wanted to try a model on a different text representation, we opted for words embeddings but this gave us some problems.
After creating the embeddings representation using Word2Vec, we applied a neural network, but it required too much time so we interrupted it.

We still wanted to see the results, so we tried to predict the labels using the training done till that moment.
Here you can see the classification report:

```
313/313 [==============================] - 174s 555ms/step -
loss: 0.3149 - accuracy: 0.4219
Validation Loss: 0.3149053156375885
Validation Accuracy: 0.4219329059123993
313/313 [==============================] - 170s 540ms/step
              precision    recall  f1-score   support

           0       0.78      0.78      0.78      4657
           1       0.95      0.99      0.97      9323
           2       0.85      0.85      0.85      5332
           3       0.90      0.91      0.91      6195
           4       0.79      0.98      0.88      7796
           5       0.78      0.70      0.74      1751
           6       0.82      0.89      0.85      6707
           7       0.62      0.01      0.01      1223
           8       0.68      0.50      0.58      1127
           9       0.63      0.20      0.30      1122
          10       0.78      0.25      0.38      1462
          11       0.85      0.91      0.88      4229
          12       0.80      0.76      0.78      4516
          13       0.70      0.65      0.68      1608

   micro avg       0.84      0.83      0.84     57048
   macro avg       0.78      0.67      0.68     57048
weighted avg       0.83      0.83      0.82     57048
 samples avg       0.84      0.84      0.83     57048
```

Figure 8: Classification Report using word embeddings

The results are actually really good, as you can see from the weighted averages it is similar or even better than the Logistic Regression model.
This suggests that this model could have been even better if it finished its training, but unfortunately it was too time consuming.

4 Topic Modeling

4.1 Objective

We have employed topic modeling two times:

- the objective was to find the topic, corresponding to MeSH Major at depth one.
- we aimed at discovering the subtopic, a higher level of detail, corresponding to the MeSH Major at depth two

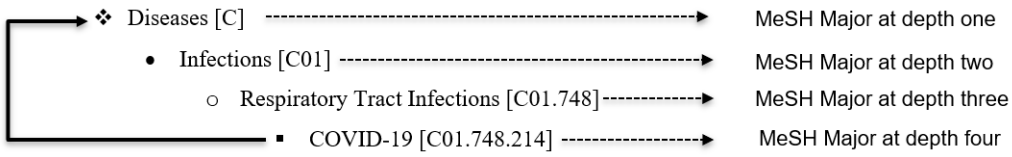

Figure 9: Classification Report using word embeddings

The techniques used for this task are:

- **Latent Dirichlet Allocation (LDA)**: every document is seen as a mixture of different topics, with each word in a document being selected randomly from the topics within that document.

  Topics are revealed by analyzing the joint distribution to calculate the conditional distribution of latent variables given the observed variables in documents.

- **Latent Semantic Analysis (LSA)**: the main concept is to break down the Document-Term matrix into a separate Document-Topic matrix and a Topic-Term matrix.

Topic Interpretation: Each topic was interpreted based on the top words contributing to it.

4.2 Main topic

To begin with, we created a dictionary of unique words using the 'tokens' column, filtering out extreme words and converting the documents into a bag-of-words format.

4.2.1 Latent Dirichlet Allocation

We've initialized and trained an LDA model on the provided corpus. The model was configured to discover the top 10 topics and provide topic distributions for individual words.

```
(0, '0.053*"cell" + 0.015*"expression" + 0.011*"mouse" + 0.010*"effect" + 0.009*"receptor" + 0.007*"activity" + 0.007*"response" + 0.007*"level" + 0.006*"human" + 0.006*"protein"')
(1, '0.042*"group" + 0.017*"p" + 0.016*"effect" + 0.016*"day" + 0.015*"treatment" + 0.013*"control" + 0.009*"response" + 0.009*"study" + 0.009*"significantly" + 0.009*"time"')
(2, '0.013*"health" + 0.011*"care" + 0.008*"study" + 0.008*"use" + 0.006*"medical" + 0.006*"research" + 0.006*"clinical" + 0.006*"quality" + 0.005*"practice" + 0.005*"data"')
(3, '0.008*"muscle" + 0.007*"bone" + 0.007*"mm" + 0.006*"left" + 0.005*"artery" + 0.005*"nerve" + 0.005*"eye" + 0.005*"tissue" + 0.005*"using" + 0.005*"area"')
(4, '0.074*"patient" + 0.014*"case" + 0.011*"disease" + 0.009*"clinical" + 0.009*"treatment" + 0.008*"cancer" + 0.007*"p" + 0.007*"year" + 0.006*"study" + 0.006*"diagnosis"')
(5, '0.025*"level" + 0.017*"blood" + 0.017*"serum" + 0.015*"p" + 0.015*"concentration" + 0.014*"plasma" + 0.009*"rat" + 0.008*"increased" + 0.008*"liver" + 0.008*"significantly"')
(6, '0.023*"gene" + 0.021*"protein" + 0.010*"dna" + 0.009*"binding" + 0.008*"sequence" + 0.008*"mutation" + 0.007*"two" + 0.007*"site" + 0.006*"strain" + 0.006*"region"')
(7, '0.013*"model" + 0.012*"method" + 0.009*"using" + 0.008*"data" + 0.007*"analysis" + 0.007*"result" + 0.007*"used" + 0.006*"system" + 0.006*"different" + 0.006*"study"')
(8, '0.009*"concentration" + 0.006*"specie" + 0.006*"water" + 0.006*"compound" + 0.006*"activity" + 0.006*"study" + 0.006*"result" + 0.005*"sample" + 0.005*"acid" + 0.005*"temperature"')
(9, '0.016*"study" + 0.014*"risk" + 0.013*"age" + 0.012*"child" + 0.011*"year" + 0.010*"woman" + 0.010*"associated" + 0.009*"ci" + 0.009*"among" + 0.008*"factor"')
```

Figure 10: Top 10 words for each topic found - LDA

We now provide the results of Perplexity and Coherence measures. The first measures how well a probabilistic model predicts a sample. In the context of LDA, lower perplexity indicates a better fit to the data.
The second measures the semantic similarity between high-scoring words in the topics, helping assess the interpretability of the topics.

- **Perplexity**: -8.263300084985667 - it suggests that the model has a reasonable fit to the data.
- **Coherence Score**: 0.5038943501035017 - it indicates that the topics are moderately coherent and interpretable.

4.2.2 Latent Semantic Analysis

For this analysis, we created the TF-IDF matrix to represent the tokens of our dataframe and then retrieved the feature names, which are the words in the vocabulary. LSA is performed using TruncatedSVD with number of topics set to 10, decomposing the matrix into 10 topics. Even in this case, we report the top 10 words for each topic found:

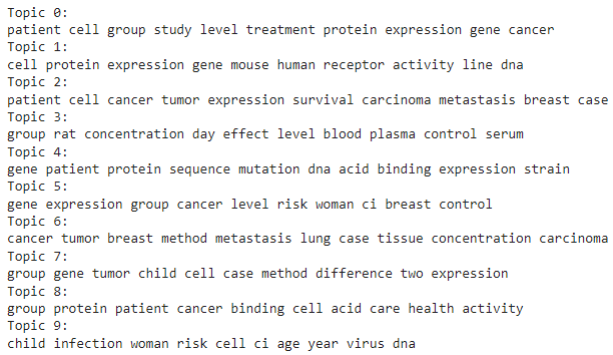

Figure 11: Top 10 words for each topic found - LSA

The Coherence score for this model is:

- **Coherence Score**: 0.5131372556287872

Even with this score, we decided to use the LDA model because the coherence is almost the same, but the topics we can infer from the top 10 words looks better, because for LSA many topics were too specific and too focused on cancer.
We now show the list of the 10 topic from LDA:

- **Topic 1: Cell Biology**: Focuses on cell-related terms like "cell," "expression," "mouse," "effect," "receptor," etc. This could be about cell behavior, cellular responses, or molecular biology.
- **Topic 2: Treatment Effects**: Includes words like "group," "effect," "day," "treatment," "control," etc., suggesting a focus on the effects of treatments on groups of people contolled day by day.
- **Topic 3: Healthcare and Medical Research**: Contains terms like "health," "care," "study," "use," "medical," "research," "clinical," suggesting a focus on healthcare studies, medical practice, and research quality.
- **Topic 4: Anatomy and Physiology**: Words like "muscle," "bone," "mm," "artery," "nerve," etc., indicate a focus on anatomical and physiological studies.
- **Topic 5: Clinical Cases and Patient Studies**: Focuses on "patient," "case," "disease," "clinical," "treatment," etc., which relates to patient studies, clinical cases, diseases, and treatments.

- **Topic 6: Blood and Biochemistry**: Includes "level," "blood," "serum," "concentration," "plasma," suggesting studies on blood components, serum levels, and biochemical analysis.
- **Topic 7: Genetics and Molecular Biology**: Words like "gene," "protein," "dna," "binding," "sequence," indicate a focus on genetic studies, protein interactions, and molecular biology.
- **Topic 8: Research Methods and Models**: Contains "model," "method," "using," "data," "analysis," etc., which suggests a focus on research methodologies, data analysis, and modeling.
- **Topic 9: Environmental Science and Chemistry**: Focuses on "concentration," "specie," "water," "compound," "activity," etc., related to environmental studies, chemistry, and ecological research.
- **Topic 10: Epidemiology and Risk Factors**: Includes "study," "risk," "age," "child," "year," "woman," "associated," "factor," which suggests a focus on epidemiological studies and risk factors.

4.2.3 Visualization of LDA

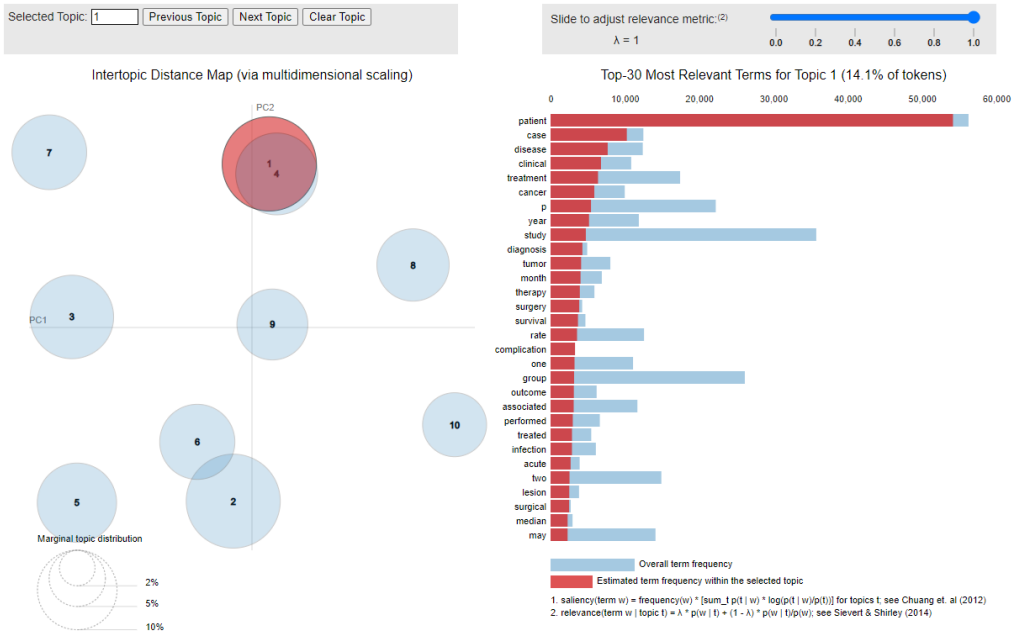

Figure 12: Use the interactive visualization on notebook

There is a scatterplot called "Intertopic Distance Map" on the left side that shows how topics are related in multidimensional scaling. The size of the circle directly reflects the topic's prevalence, and the distance between circles denotes the similarity or contrast between topics, such as topic 1 and topic 4.

A bar chart on the right side displays the top 30 terms associated with the chosen topic, arranged in ascending order. The estimated term frequency is shown by the red bar, whereas the total term frequency in the corpus is shown by the blue bar.

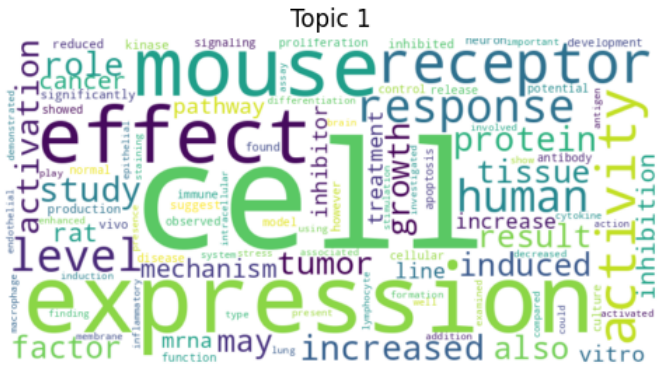

Figure 13: Word cloud for Topic 1

This visualization is a word cloud representing the most significant terms for Topic 1 from an LDA model. The size of each word indicates its relative importance or frequency within Topic.
We did one of those visualizations for each topic.

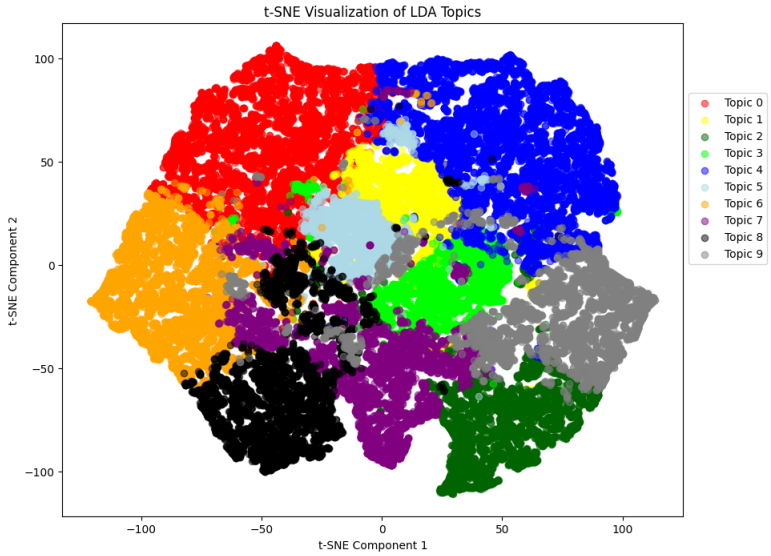

Figure 14: T-SNE Visualization of LDA Topics

The plot shows how topics are spread out in a two-dimensional area. The

T-SNE components are displayed on the two axes, with various topics represented by different colors and each document corresponding to a point on the plot. The grouping of points with matching colors shows that documents with similar topic distributions are clustered together.

4.3 Subtopic

With this application of topic modeling we want to find the subtopics, higher level of details, corresponding to the MeSH Major at depth two.
We confronted the topics previously obtained with the MeSH labels and we saw similarity between some of them, in particular we focused on topic 9 Environmental Science and Chemistry, that is similar to a combination of "Chemicals and Drugs [D]" and "Geographicals [Z]".
So we found all the articles containing D or Z, and extracted all the possible MeSH Major at depth two, finding 17 of them.
Then we proceeded employing topic modeling with the number of topics set to 17, on the new dataset containing only articles whose topic assigned was the Topic 9.

We provide the results of both models (LDA and LSA) and then we discuss which one is better.

```
(0, '0.029*"risk" + 0.027*"patient" + 0.021*"ci" + 0.019*"mortality" + 0.014*"disease" + 0.014*"study" + 0.014*"year" + 0.013*"age" + 0.011*"death" + 0.009*"associated"')
(1, '0.018*"health" + 0.011*"screening" + 0.010*"disability" + 0.010*"study" + 0.010*"woman" + 0.010*"care" + 0.010*"year" + 0.009*"disease" + 0.008*"older" + 0.008*"age"')
(2, '0.017*"association" + 0.017*"study" + 0.014*"p" + 0.014*"control" + 0.014*"polymorphism" + 0.013*"genotype" + 0.013*"gene" + 0.011*"patient" + 0.011*"associated" + 0.010*"genetic"')
(3, '0.018*"female" + 0.017*"male" + 0.017*"woman" + 0.012*"alcohol" + 0.011*"sexual" + 0.009*"men" + 0.009*"study" + 0.008*"difference" + 0.008*"effect" + 0.007*"relationship"')
(4, '0.020*"child" + 0.015*"infant" + 0.014*"birth" + 0.012*"pregnancy" + 0.012*"age" + 0.012*"group" + 0.011*"study" + 0.009*"month" + 0.008*"maternal" + 0.008*"woman"')
(5, '0.077*"cancer" + 0.038*"risk" + 0.031*"woman" + 0.029*"ci" + 0.021*"breast" + 0.020*"hiv" + 0.019*"among" + 0.015*"men" + 0.013*"hpv" + 0.011*"study"')
(6, '0.037*"infection" + 0.021*"prevalence" + 0.015*"vaccine" + 0.012*"transmission" + 0.011*"virus" + 0.011*"case" + 0.010*"influenza" + 0.010*"malaria" + 0.010*"vaccination" + 0.010*"antibody"')
(7, '0.025*"fracture" + 0.023*"twin" + 0.021*"bone" + 0.020*"hip" + 0.019*"bmd" + 0.011*"genetic" + 0.009*"factor" + 0.009*"sibling" + 0.009*"study" + 0.008*"fall"')
(8, '0.019*"health" + 0.013*"smoking" + 0.012*"use" + 0.011*"study" + 0.010*"among" + 0.009*"associated" + 0.008*"prevalence" + 0.008*"factor" + 0.007*"status" + 0.007*"year"')
(9, '0.041*"rate" + 0.024*"incidence" + 0.021*"year" + 0.014*"per" + 0.013*"case" + 0.013*"population" + 0.013*"country" + 0.012*"data" + 0.012*"trend" + 0.012*"mortality"')
(10, '0.020*"p" + 0.013*"volume" + 0.013*"pd" + 0.009*"cognitive" + 0.008*"effect" + 0.008*"associated" + 0.008*"â" + 0.008*"association" + 0.008*"correlation" + 0.008*"brain"')
(11, '0.022*"p" + 0.020*"diabetes" + 0.020*"bmi" + 0.019*"woman" + 0.015*"body" + 0.014*"weight" + 0.014*"obesity" + 0.012*"age" + 0.012*"study" + 0.011*"mass"')
(12, '0.020*"suicide" + 0.016*"schizophrenia" + 0.011*"population" + 0.011*"study" + 0.011*"group" + 0.010*"hcv" + 0.008*"factor" + 0.007*"suicidal" + 0.005*"family" + 0.005*"result"')
(13, '0.057*"patient" + 0.036*"symptom" + 0.031*"depression" + 0.027*"disorder" + 0.020*"scale" + 0.019*"score" + 0.019*"anxiety" + 0.012*"psychiatric" + 0.011*"depressive" + 0.011*"clinical"')
(14, '0.023*"adhd" + 0.022*"white" + 0.022*"memory" + 0.022*"black" + 0.018*"test" + 0.014*"function" + 0.012*"impairment" + 0.012*"stroke" + 0.011*"american" + 0.010*"ckd"')
(15, '0.060*"child" + 0.015*"adolescent" + 0.015*"parent" + 0.012*"behavior" + 0.011*"family" + 0.011*"study" + 0.010*"childhood" + 0.010*"cognitive" + 0.009*"age" + 0.009*"problem"')
(16, '0.012*"study" + 0.012*"injury" + 0.011*"group" + 0.009*"subject" + 0.008*"exposure" + 0.008*"year" + 0.007*"age" + 0.007*"test" + 0.006*"one" + 0.005*"control"')
```

Figure 15: LDA results

- **Perplexity**: -8.289456547562331
- **Coherence Score**: 0.4222941220164357

```
LSA Model with 17 topics:
Topic 1: patients, risk, 95, women, ci, children, age, study, years, associated
Topic 2: ci, 95, cancer, risk, women, mortality, breast, confidence, ratio, odds
Topic 3: children, ci, 95, child, infants, mothers, birth, parents, age, maternal
Topic 4: women, hiv, men, sexual, pregnancy, health, use, pregnant, alcohol, infection
Topic 5: hiv, patients, infection, infected, children, ci, 95, art, transmission, treatment
Topic 6: cancer, ci, 95, depression, symptoms, children, health, anxiety, use, self
Topic 7: women, patients, infants, cancer, pregnancy, birth, breast, depression, mothers, maternal
Topic 8: cancer, breast, mortality, cases, lung, genetic, incidence, risk, smoking, population
Topic 9: mortality, years, health, rates, incidence, age, year, death, rate, older
Topic 10: infants, birth, exposure, maternal, alcohol, mothers, mortality, pregnancy, infant, preterm
Topic 11: hiv, infants, cancer, bmi, smoking, weight, mortality, obesity, cognitive, body
Topic 12: alcohol, smoking, patients, use, drinking, diabetes, consumption, children, smokers, risk
Topic 13: diabetes, risk, health, depression, mortality, type, factors, symptoms, disease, cardiovascular
Topic 14: males, infants, sexual, females, depression, male, weight, patients, bmi, alcohol
Topic 15: smoking, smokers, exposure, prevalence, symptoms, asthma, depression, anxiety, levels, tobacco
Topic 16: mortality, anxiety, depression, males, levels, men, symptoms, children, females, women
Topic 17: risk, sexual, factors, men, age, subjects, ad, cognitive, disease, sex
```

Figure 16: LSA results

- **Coherence Score**: 0.529751879710176

Based on the provided coherence scores of both models, the LSA model is better since it has a higher coherence score (0.530 vs. 0.422). Coherence is generally more aligned with human judgment of topic quality.
So here we show the subtopics found:

- **Topic 1: Patient Demographics and Risk Factors**
  - **Keywords:** patients, risk, 95, women, ci, children, age, study, years, associated
  - **Interpretation:** This topic likely discusses demographic characteristics of patients in studies, focusing on various risk factors associated with age, gender (women), and study parameters.
- **Topic 2: Cancer and Mortality**
  - **Keywords:** ci, 95, cancer, risk, women, mortality, breast, confidence, ratio, odds
  - **Interpretation:** This topic is centered around cancer research, specifically discussing mortality rates, risk factors (like gender), and statistical measures (confidence intervals).
- **Topic 3: Child Health and Maternal Factors**
  - **Keywords:** children, ci, 95, child, infants, mothers, birth, parents, age, maternal

  - **Interpretation:** Focuses on child health, including maternal influences (mothers, birth), developmental stages (infants, children), and associated factors studied with statistical confidence.

- **Topic 4: Women's Health and Infectious Diseases**
  - **Keywords:** women, hiv, men, sexual, pregnancy, health, use, pregnant, alcohol, infection
  - **Interpretation:** Discusses women's health issues such as HIV, sexual health, pregnancy-related concerns, and alcohol use, within the context of infectious diseases.

- **Topic 5: HIV Treatment and Transmission**
  - **Keywords:** hiv, patients, infection, infected, children, ci, 95, art, transmission, treatment
  - **Interpretation:** Focuses on HIV infection and treatment strategies, including transmission dynamics (art = antiretroviral therapy) among patients and children.

- **Topic 6: Cancer and Mental Health**
  - **Keywords:** cancer, ci, 95, depression, symptoms, children, health, anxiety, use, self
  - **Interpretation:** Explores the intersection of cancer research with mental health aspects, including depression, anxiety, and self-reported symptoms among patients.

- **Topic 7: Women's Health and Pregnancy Outcomes**
  - **Keywords:** women, patients, infants, cancer, pregnancy, birth, breast, depression, mothers, maternal
  - **Interpretation:** Focuses on women's health issues related to pregnancy outcomes, maternal health, and impacts on infants and breast cancer patients.

- **Topic 8: Cancer Epidemiology and Genetic Factors**
  - **Keywords:** cancer, breast, mortality, cases, lung, genetic, incidence, risk, smoking, population
  - **Interpretation:** Discusses cancer epidemiology, genetic predispositions, mortality rates, and the influence of smoking within specific populations.

- **Topic 9: Mortality Rates and Aging**
  - **Keywords:** mortality, years, health, rates, incidence, age, year, death, rate, older
  - **Interpretation:** Analyzes mortality rates across different age groups, health implications, and trends observed over years.
- **Topic 10: Infant Health and Maternal Influences**
  - **Keywords:** infants, birth, exposure, maternal, alcohol, mothers, mortality, pregnancy, infant, preterm
  - **Interpretation:** Focuses on infant health outcomes, maternal influences (alcohol use, pregnancy), and risks associated with preterm births.
- **Topic 11: HIV, BMI, and Lifestyle Factors**
  - **Keywords:** hiv, infants, cancer, bmi, smoking, weight, mortality, obesity, cognitive, body
  - **Interpretation:** Explores HIV impacts on BMI (Body Mass Index), smoking habits, weight issues, mortality risks, and cognitive functions.
- **Topic 12: Alcohol Use, Smoking, and Disease Risk**
  - **Keywords:** alcohol, smoking, patients, use, drinking, diabetes, consumption, children, smokers, risk
  - **Interpretation:** Discusses health risks associated with alcohol consumption, smoking habits, diabetes, and their impact on patients' health, including children and smokers.
- **Topic 13: Diabetes and Cardiovascular Health**
  - **Keywords:** diabetes, risk, health, depression, mortality, type, factors, symptoms, disease, cardiovascular
  - **Interpretation:** Focuses on diabetes-related health risks, including cardiovascular implications, symptoms, and associated factors influencing mortality rates.
- **Topic 14: Mental Health and Gender Differences**
  - **Keywords:** males, infants, sexual, females, depression, male, weight, patients, bmi, alcohol

  - **Interpretation:** Analyzes mental health issues across genders (males, females), including depression, sexual health, weight concerns, BMI, and alcohol use.

- **Topic 15: Smoking, Asthma, and Psychological Symptoms**
  - **Keywords:** smoking, smokers, exposure, prevalence, symptoms, asthma, depression, anxiety, levels, tobacco
  - **Interpretation:** Discusses smoking-related issues like exposure, prevalence, and associated psychological symptoms such as asthma, depression, and anxiety.

- **Topic 16: Mental Health Issues Across Ages**
  - **Keywords:** mortality, anxiety, depression, males, levels, men, symptoms, children, females, women
  - **Interpretation:** Explores mental health challenges across different age groups (children, adults, elderly), genders, and associated symptoms like anxiety and depression.

- **Topic 17: Risk Factors and Cognitive Health**
  - **Keywords:** risk, sexual, factors, men, age, subjects, ad, cognitive, disease, sex
  - **Interpretation:** Focuses on risk factors affecting cognitive health, including sexual health implications, demographic subjects, Alzheimer's disease (ad), and cognitive functions.

5 Conclusion

The multi-label classification on PubMed abstracts has shown that it is possible to accurately categorize biomedical research articles into MeSH categories using machine learning techniques.
Topic modeling initially identified topics similar to MeSH Major at depth one. However, when we explored the subtopics in more detail, they did not seem to accurately represent their main topics. This is likely because the articles considered, like most articles, have multiple MeSH Major at depth one, making it difficult to individuate the MeSH Major at depth two for just one of them.
Future Work:

- Enhancing the classification model using more advanced techniques such as transformer-based models (e.g., BERT, BioBERT).
- Exploring hierarchical classification to leverage the structure of MeSH categories.
- Integrating more sophisticated text preprocessing and feature engineering methods to improve model performance.
- Identify through classification or topic modeling all level of depth of the MeSH labels.

References

[1] PubMed *https://pubmed.ncbi.nlm.nih.gov/*

[2] MeSH (Medical Subject Headings) resources and documentation