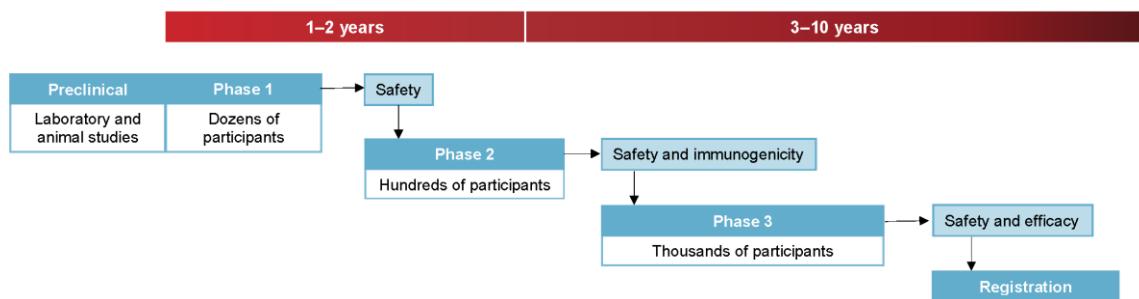


Phase Classification of COVID-19 Clinical Trials using NLP

Conventional pathway of vaccine development



@<https://ncirs.org.au/phases-clinical-trials>

ML question: Is it possible to predict the clinical trial phase using text information, using the features Title, Phase and Interventions?

Dataset Overview:

The dataset had many columns, which are 27 columns and 5783 records. The data is very much categorical, with only a numeric column, enrollment.

The columns required:

1. Title
2. Interventions
3. Outcome Measures
4. Phases
5. Study Type

The rest of the columns will not be used.

Data Cleaning:

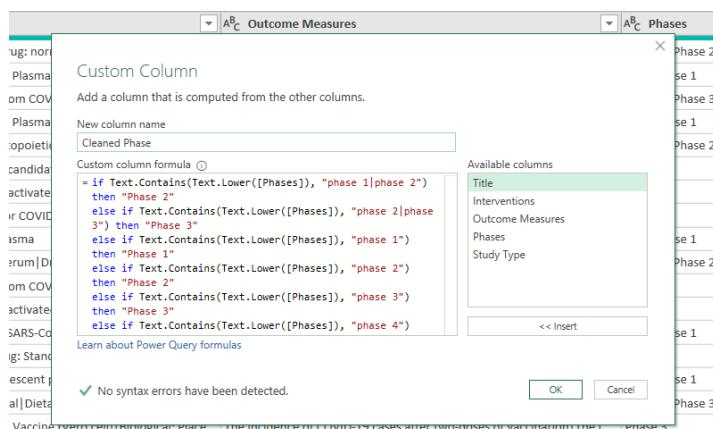
Microsoft Excel has been used to clean the data and remove the columns.

1. Removed Errors.
2. Removed unwanted columns.
3. Filtered: only Interventional in Study Type.

Why only Interventional studies?

It is because the clinical trial phases are formally defined only for interventional trials, and not for observational ones, as the observational studies are only for research purposes, and the phases are not applicable.

4. Unfiltered: Not applicable data in Phases.
5. Create a new custom column and rename it as Phase:



The newly added column has only 4 distinct categorical records in it, which are

- Phase 1

- Phase 2
 - Phase 3
 - Phase 4
6. Concatenating the columns Title, Interventions, and Outcome measure into a new column: text.
 7. Removing punctuations: Only letters, numbers, and spaces are kept.

A ^B _C text	A ^B _C text_clean
study to evaluate the efficacy of covid19-0001-usr in patients with mil...	study to evaluate the efficacy of covid190001usr in patients with mild...
convalescent plasma for covid-19 patients biological: convalescent cov...	convalescent plasma for covid19 patients biological convalescent covi...
covid19-convalescent plasma for treating patients with active sympto...	covid19convalescent plasma for treating patients with active symptom...
covid-19 plasma in treatment of covid-19 patients biological: convalesc...	covid19 plasma in treatment of covid19 patients biological convalesce...
study evaluating the safety and efficacy of autologous non-hematopo...	study evaluating the safety and efficacy of autologous nonhematopoei...
clinical trial to evaluate the safety and immunogenicity of the covid-19...	clinical trial to evaluate the safety and immunogenicity of the covid19 ...
an effectiveness study of the sinovac's adsorbed covid-19 (inactivated)...	an effectiveness study of the sinovacs adsorbed covid19 inactivated va...
comparison of the efficacy of rapid tests to identify covid-19 infection ...	comparison of the efficacy of rapid tests to identify covid19 infection c...

8. Remove extra spaces: trim & clean
9. Remove all columns except Phase and text.

Queries [1]		= Table.RemoveColumns#"Cleaned Text",{"Title", "Interventions", "Outcome Measures", "Study Type"})
	ABC Phase	A ^B _C text
1	Phase 2	study to evaluate the efficacy of covid190001usr in patients with mild...
2	Phase 1	convalescent plasma for covid19 patients biological convalescent covi...
3	Phase 3	covid19convalescent plasma for treating patients with active symptom...
4	Phase 1	covid19 plasma in treatment of covid19 patients biological convalesce...
5	Phase 2	study evaluating the safety and efficacy of autologous nonhematopoie...
6	Phase 1	clinical trial to evaluate the safety and immunogenicity of the covid19 ...
7	Phase 4	an effectiveness study of the sinovacs adsorbed covid19 inactivated va...
8	Phase 1	comparison of the efficacy of rapid tests to identify covid19 infection c...
9	Phase 1	covid19 convalescent plasma ccp transfusion biological covid convales...
10	Phase 2	evaluation of equine antibody treatment in patients with covid 19 infe...
11	Phase 2	convalescent plasma in the treatment of covid19 biological convalesce...
12	Phase 3	efficacy safety and immunogenicity study of sarscov2 inactivated vacci...
13	Phase 1	convalescent plasma in icu patients with covid19induced respiratory f...
14	Phase 2	clinical trial of allogeneic mesenchymal cells from umbilical cord tissue...

Applied steps in Power Query editor:

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Removed Errors
- Removed Columns
- Removed Other Columns
- Filtered Rows
- Added Custom
- Removed Columns1
- Renamed Columns
- Added Custom1
- Lowercased Text
- Added Custom2
- Removed Columns2
- Renamed Columns1
- Trimmed Text
- Cleaned Text
- Removed Columns3

Rohith Vaddepally

In the data transformation process, the features Title, Interventions, and Outcome measures are combined into a text column. So, with the feature text, the other column is phase which contains phases 1-4.

Data Loading:

New data consists of 2 features and 1968 records.

Phase distribution:

```
Phase
Phase 2    877
Phase 3    650
Phase 1    280
Phase 4    161
Name: count, dtype: int64
```

Phases 2 & 3 are dominating, while Phase 4 is in the minority. This is because not all the trials may reach stage 4.

Label Encoding:

The target variable Phase is categorical, but it should be transformed into a numerical variable. The label encoder library is used to encode them.

	Phase	phase_label
1	Phase 1	0
0	Phase 2	1
2	Phase 3	2
6	Phase 4	3

So, Phase 1 is now 0, Phase 2 is 1, Phase 3 is 2, and Phase 4 is 3.

Feature and Target variable:

```
X= df['text']
y= df['phase_label']
```

TF-IDF Vectorization:

Term Frequency-Inverse Document Frequency: The model cannot understand the text data. TF-IDF is used to convert the text into numbers.

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer(ngram_range=(1,2), max_features=5000,
min_df=5, max_df=0.9)
X_train_tfidf = tfidf.fit_transform(X_train)
X_test_tfidf = tfidf.transform(X_test)
```

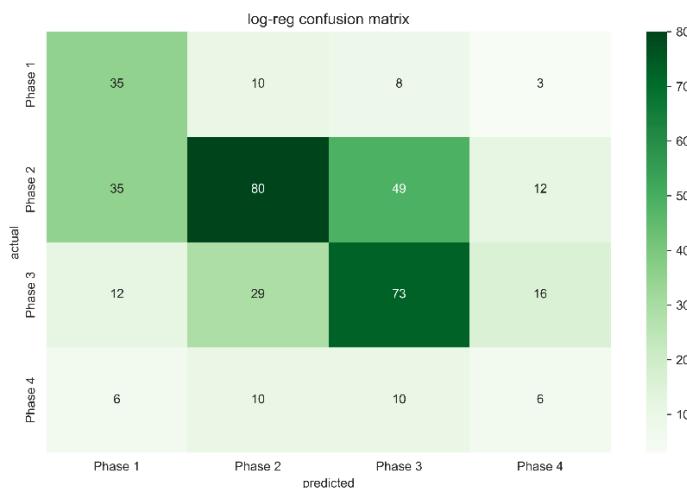
Now the text feature is numeric. X_train_tfidf is of 1574 trials, and 4119 learned textual features.

Logistic Regression Evaluation:

Logistic Regression learns which TF-IDF words and phrases push a clinical trial toward phase 1,2,3 or 4.

	precision	recall	f1-score	support
Phase 1	0.40	0.62	0.49	56
Phase 2	0.62	0.45	0.52	176
Phase 3	0.52	0.56	0.54	130
Phase 4	0.16	0.19	0.17	32
accuracy			0.49	394
macro avg	0.43	0.46	0.43	394
weighted avg	0.52	0.49	0.50	394

Phase 1 trials have a higher recall rate at 62%, which is 62%. Phase 1 is correctly identified and phase 4 has the lowest recall; this could be because of insufficient data for phase 4.



Phase 1 correctly predicted 35 keywords and misclassified a few words.

Phase 2 correctly predicted 80 and misclassified 35 words as 1 and 49 as 3.

Phase 3 correctly predicted 73 and misclassified 29 as phase 2

Phase 4 predicted only 6, due to a lack of data, which was expected.

```
Top terms for Phase 1:  
['form' 'stress' 'mesenchymal' 'abnormal' 'cells' 'response' 'enhance'  
'vaccine' 'biological' 'in healthy' 'plasma' 'safety' 'tolerability'  
'healthy' 'adverse']  
  
Top terms for Phase 2:  
['study' 'measured' 'alive and' 'of pulmonary' 'tofacitinib' 'subjects'  
'to' 'ruxolitinib' 'type' 'with covid19' 'requiring' 'respiratory' 'on'  
'viral' 'placebo']  
  
Top terms for Phase 3:  
['at' 'covid19' 'and safety' 'during' 'days' 'among' 'hospital'  
'participants who' 'of the' 'allcause' 'of patients' 'standard'  
'symptomatic' 'or' 'efficacy']  
  
Top terms for Phase 4:  
['tablets' 'hydrogen' 'dosedrug' 'rivaroxaban' 'drug dexamethasone'  
'influenza' 'vaccination' 'ivig' 'combined' 'cases' 'after' 'drug'  
'coronavac' 'dexamethasone' 'nitazoxanide']
```

Above are the words in which they strongly appear in the respective phases of the trial.

Evaluation by adding class weights to phase 4:

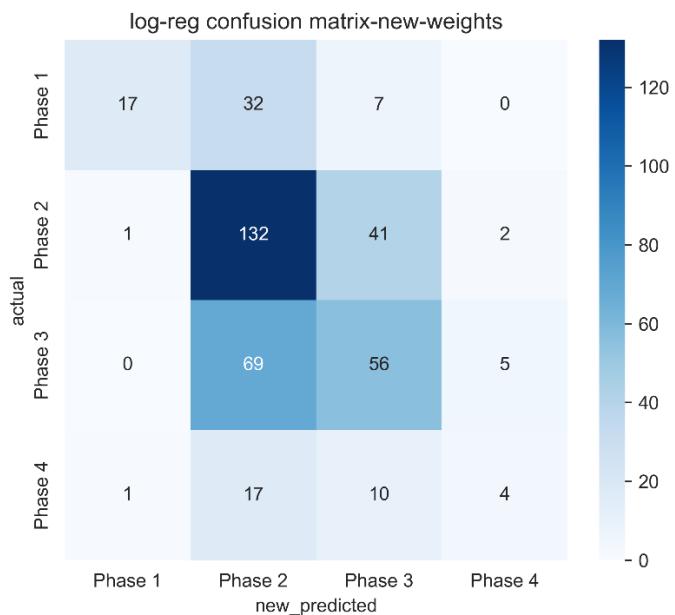
This is done to improve phase 4 recall.

	precision	recall	f1-score	support
Phase 1	0.89	0.30	0.45	56
Phase 2	0.53	0.75	0.62	176
Phase 3	0.49	0.43	0.46	130
Phase 4	0.36	0.12	0.19	32
accuracy			0.53	394
macro avg	0.57	0.40	0.43	394
weighted avg	0.55	0.53	0.51	394

With the new weights for phase 4, the recall dropped to 12% and previously with balanced weights it was 19%. Phase 4 is of a small size with a few unique terms and shares its terms with

When class weights increased for phase 4, Logistic regression increased the margin for phase 4 and adjusted the coefficients to reduce the phase 4 loss.

But phase 4 overlaps with phase 3, recall dropped.



Above is the confusion matrix with the added phase 4 weights=2.5. The reason for adding more weight only to phase 4 is to increase the recall value by penalising misclassification of phase 4. so that the model predicts phase 4 more often.

But that didn't work at all. As the recall dropped, it only predicts 4 keywords, which is fewer than before.

The predictions for phase 3 dropped.

The predictions for phase 2 increased as it is taking all the misclassifications from phase 3, phase 1, and phase 2.

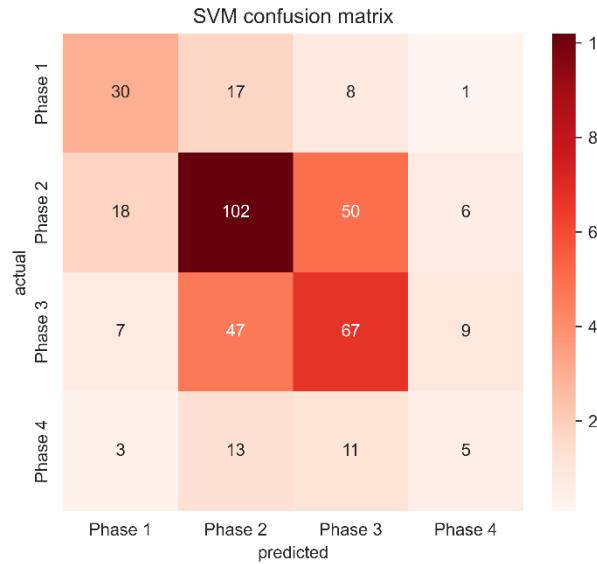
It can be concluded that logistic regression has done its best.

Linear Support Vector Machine Evaluation:

Linear SVM is better for the phase 3 to phase 4 problem, as it focuses on hard-to-separate boundary points.

	precision	recall	f1-score	support
Phase 1	0.52	0.54	0.53	56
Phase 2	0.57	0.58	0.57	176
Phase 3	0.49	0.52	0.50	130
Phase 4	0.24	0.16	0.19	32
accuracy			0.52	394
macro avg	0.45	0.45	0.45	394
weighted avg	0.51	0.52	0.51	394

Recall still did not improve for Phase 4.



Recall value of Phase 4 from the Linear-SVM model is 16%, which is less than the original logistic regression recall value. So, Linear-SVM did not improve the phase 4 recall value.

Improvements:

Phase 3: Although recall dropped for phase 3, errors of phase 3-> phase 4 decreased from 16 to 9.

Only a few Phase 3 samples collapsed into Phase 2.

Phase 2: Before with Log_reg, the correct predictions were 80, but they are 102, in which SVM did better, considering the samples Phase 2 initially had.

Phase 1: Recall dropped, but errors also dropped from phase 1--> phase 3, i.e., from 12 to 7.

So, both Logistic regression and Linear SVM models failed to improve phase 4 recall, and this is because of data scarcity.

Hierarchical Classification:

Level-1: As the models, logistic regression and linear SVM are not that great. It is better to have a hierarchy design, where

Early = Phase 1, Phase 2

Late = Phase 3, Phase 4

Linear SVM Model Evaluation:

	precision	recall	f1-score	support
Early	0.76	0.72	0.74	232
Late	0.63	0.67	0.65	162
accuracy			0.70	394
macro avg	0.69	0.70	0.69	394
weighted avg	0.70	0.70	0.70	394

Late recall: 67%, so in the late stage trials, 33% are being misclassified as early at Level 1 of the hierarchy.

But in a hierarchy, at a level-1, if a late-stage trial is misclassified as an early stage, it will reach level-2, i.e., phase 3 vs phase 4 classifier, and phase 4 recall will be completely lost. So, level-1 must over protect late stage.

Adding weights to the late stage – {2.0-4.0} :

I have added weight for the late stage from 2.0 to 4.0, but the recall doesn't go past 75%. So the TF-IDF text feature is not able to classify early and late stage trials. So, continuing hierarchical classification is not a good option as it is not possible to move to level-2.

Conclusion:

In this project, I was predicting COVID-19 clinical trial phases using text data. Starting with TF-IDF, I evaluated multiple linear classifiers, including Logistic Regression and Linear SVM. Detailed analysis of confusion matrices and model coefficients showed that there were misclassifications occurring in different phases of the trial.

Despite applying class weighting, switching classifiers, and reformulating the task as hierarchical classification, model performance, particularly recall for late-stage (Phase 4) trials, consistently stayed the same. Further investigation revealed that this limitation is in the clinical trial text data.

The key outcome: text-only TF-IDF features are insufficient to really classify the exact clinical trial phase, especially in late (phase 3, phase 4) stages. This finding highlights the importance of feature representation and problem formulation in applied machine learning and motivates future work incorporating structured metadata or contextual embeddings.