ABBREVIATION DISAMBIGUATION WITH NLP

Final Report

Table of Contents

1 Abstract

What is the possibility for a Data Scientist to create a model that can read Medical Text and correctly classify the Medical Abbreviations to enable readers to make sense of the Medical Text in correct context?

Acronyms and abbreviations within clinical text are widespread, and their use continues to increase. Several reasons for this ongoing growth include adoption of electronic health record (EHR) systems with increased volume of electronic clinical notes accompanied by the wide usage of acronyms and abbreviations, the time-constrained nature of clinical medicine encouraging the use of shortened word forms, and a longstanding tradition of commonly using acronyms and abbreviations in clinical documentation.

The process of understanding the precise meaning of a given acronym or abbreviation in texts is one of several key functions of automated medical natural language processing (NLP) systems and is a special case of word sense disambiguation (WSD).

This project uses the freely available Medical Dataset for Abbreviation Disambiguation (MeDAL) downloaded directly from Kaggle. This project employs the implementation of some of the well-known NLP libraries like spacy and gensim to understand the context of a medical text and based on that context predict what the expanded form of an Abbreviation (if any) is in that medical text.

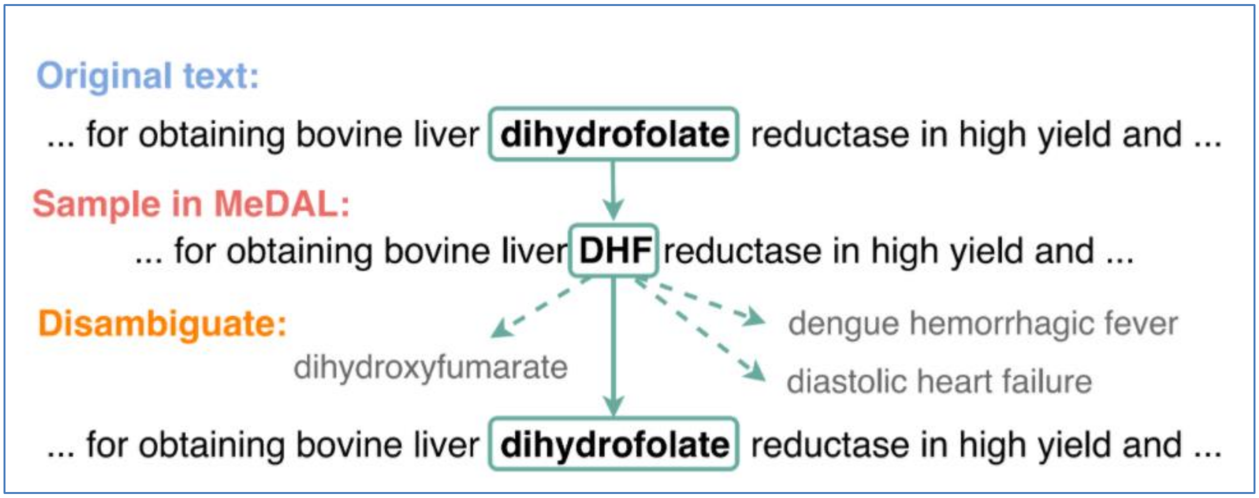

Original text:
... for obtaining bovine liver dihydrofolate reductase in high yield and ...
Sample in MeDAL:
... for obtaining bovine liver DHF reductase in high yield and ...
Disambiguate:
dihydroxyfumarate
dengue hemorrhagic fever
diastolic heart failure
... for obtaining bovine liver dihydrofolate reductase in high yield and ...

Figure 1. *Objective*

2 Introduction

Acronyms and abbreviations within clinical text are widespread, and their use continues to increase. Several reasons for this ongoing growth include adoption of electronic health record (EHR) systems with increased volume of electronic clinical notes accompanied by the wide usage of acronyms and abbreviations, the time-constrained nature of clinical medicine encouraging the use of shortened word forms, and a longstanding tradition of commonly using acronyms and abbreviations in clinical documentation.

The process of understanding the precise meaning of a given acronym or abbreviation in texts is one of several key functions of automated medical natural language processing (NLP) systems and is a special case of word sense disambiguation (WSD).

Acronyms and abbreviations each have a short form (the acronym or abbreviation) and a long form (the expansion of the acronym or abbreviation). In clinical documents, the expanded long form is rarely proximal to the short form of the acronym or abbreviation because clinical texts rarely conform to the formalism of enclosing the long form in parentheses after the first mention of the abbreviation, as is customary in scientific literature. This lack of the formalism is one of the significant barriers associated with using clinical text for NLP research, which has resulted in limited data resources for research. Because of this informality and the shortage of the available resources/research, while researchers have explored the use of supervised machine learning (ML) approaches for acronym and abbreviation WSD, some of the related issues with optimal window size and orientation and with training sample size minimization to reduce the associated cost and time to manually annotate training corpora remain open.

Unstructured clinical texts contain rich information, but this resource isn't quickly accessible. Mining this knowledge base using NLP can facilitate clinical research and improve patient care. WSD is a critical step towards building useful clinical NLP applications. Existing open-source machine learning libraries do not adequately represent clinical text and result in a drop in the performance.

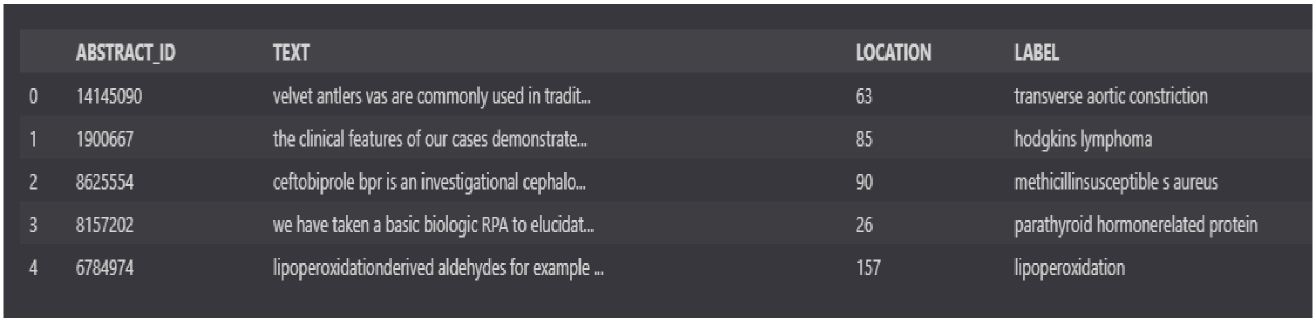
So there is a need to develop medicine specific machine learning toolkits. However, it has unique challenges because of the unavailability of a significant amount of annotated data.

The aim of this project is to differentiate between similar looking Acronyms and Abbreviations based on the context in which they are being used. The MeDAL dataset used here has around 10M Medical Extracts containing some of the most confusing Acronyms and Abbreviations. In this project, of the total data, I have used 70K extracts containing 20 unique Abbreviations with more than 2000 unique expanded forms to train the model.

3 Work Done

3.1 MeDAL Database

The MeDAL database used in this project contains more than 10M Medical text extracts available on Kaggle.com. Due to hardware limitations, this project will use only 70K such extract from Training, Validation and Testing of the model. Some sample data is shown below in Figure 2.



	ABSTRACT_ID	TEXT	LOCATION	LABEL
0	14145090	velvet antlers vas are commonly used in tradit...	63	transverse aortic constriction
1	1900667	the clinical features of our cases demonstrate...	85	hodgkins lymphoma
2	8625554	ceftobiprole bpr is an investigational cephalo...	90	methicillinsusceptible s aureus
3	8157202	we have taken a basic biologic RPA to elucidat...	26	parathyroid hormonerelated protein
4	6784974	lipoperoxidationderived aldehydes for example ...	157	lipoperoxidation

Figure 2. *Data from MeDAL Dataset.*

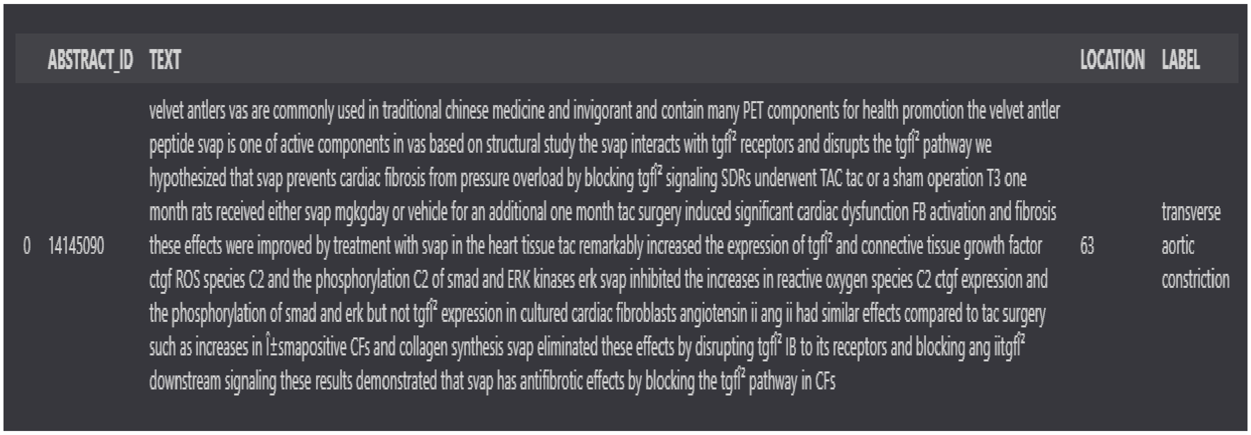
3.2 Data Cleaning

The MeDAL database contains more than 10M rows but due to Hardware limitations we will use only 70K rows for this project. The dataset here doesn't contain any Duplicate or Null values which need to be handled. Hence, the data is already clean and no further cleaning is required.

3.3 Exploratory Data Analysis (EDA)

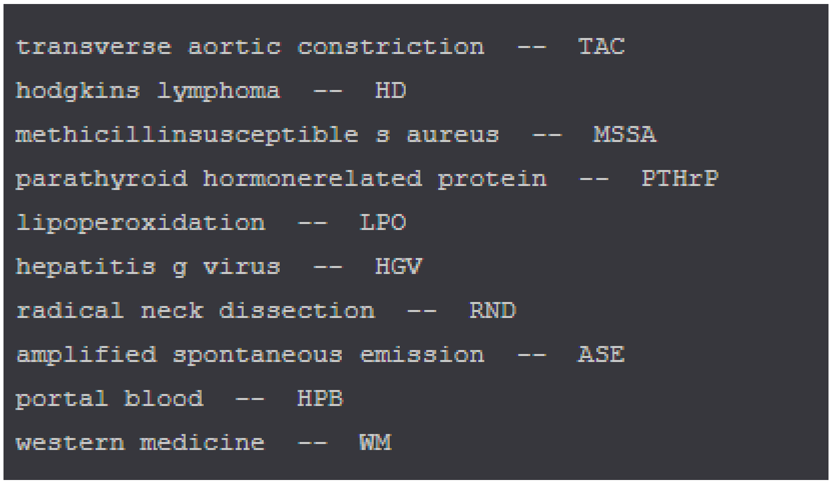
After Data Cleaning, I performed EDA on this clean dataset to find how the data looks like and if there is any kind of pattern in the Data. The dataset contains 4 columns (Figure 3a):

- **ABSTRACT_ID:** This column contains the unique Text Id.
- **TEXT:** This column contains the Medical text which has some Abbreviations.
- **LOCATION:** This column contains the word count at which the abbreviation is present (Figure 3b).
- **LABEL:** This column contains the Expanded form of the Abbreviation.



	ABSTRACT_ID	TEXT	LOCATION	LABEL
0	14145090	velvet antlers vas are commonly used in traditional chinese medicine and invigorant and contain many PET components for health promotion the velvet antler peptide svap is one of active components in vas based on structural study the svap interacts with tgfî² receptors and disrupts the tgfî² pathway we hypothesized that svap prevents cardiac fibrosis from pressure overload by blocking tgfî² signaling SDRs underwent TAC tac or a sham operation T3 one month rats received either svap mgkgday or vehicle for an additional one month tac surgery induced significant cardiac dysfunction FB activation and fibrosis these effects were improved by treatment with svap in the heart tissue tac remarkably increased the expression of tgfî² and connective tissue growth factor ctgf ROS species C2 and the phosphorylation C2 of smad and ERK kinases erk svap inhibited the increases in reactive oxygen species C2 ctgf expression and the phosphorylation of smad and erk but not tgfî² expression in cultured cardiac fibroblasts angiotensin ii ang ii had similar effects compared to tac surgery such as increases in î±smapositive CFs and collagen synthesis svap eliminated these effects by disrupting tgfî² IB to its receptors and blocking ang iitgfî² downstream signaling these results demonstrated that svap has antifibrotic effects by blocking the tgfî² pathway in CFs	63	transverse aortic constriction

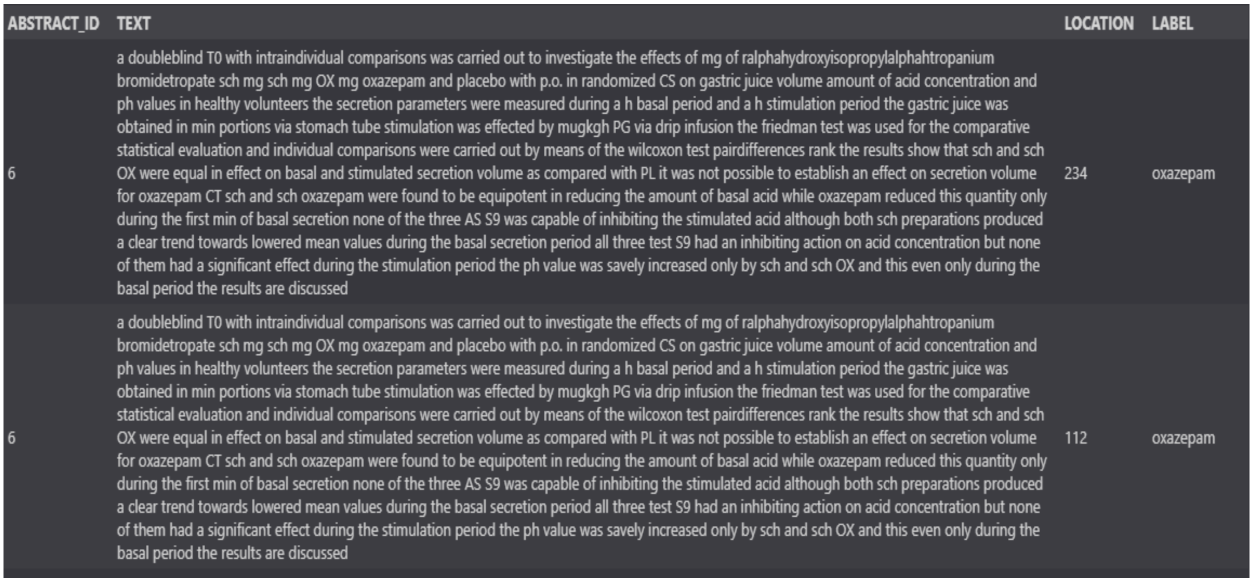
Figure 3a. *A row from the Dataset.*



```
transverse aortic constriction  --  TAC
hodgkins lymphoma  --  HD
methicillinsusceptible s aureus  --  MSSA
parathyroid hormonerelated protein  --  PTHrP
lipoperoxidation  --  LPO
hepatitis g virus  --  HGV
radical neck dissection  --  RND
amplified spontaneous emission  --  ASE
portal blood  --  HPB
western medicine  --  WM
```

Figure 3b. Abbreviation derived from LOCATION column and its Expanded form.

On further analysis, it was found that there were several rows which had the same TEXT but different LABEL. Thus, there are certain text extracts which contain more than one Abbreviation for disambiguation (Figure 4).



ABSTRACT_ID	TEXT	LOCATION	LABEL
6	a doubleblind T0 with intraindividual comparisons was carried out to investigate the effects of mg of ralphahydroxyisopropylalphahtropanium bromidetropate sch mg sch mg OX mg oxazepam and placebo with p.o. in randomized CS on gastric juice volume amount of acid concentration and ph values in healthy volunteers the secretion parameters were measured during a h basal period and a h stimulation period the gastric juice was obtained in min portions via stomach tube stimulation was effected by mugkgh PG via drip infusion the friedman test was used for the comparative statistical evaluation and individual comparisons were carried out by means of the wilcoxon test pairdifferences rank the results show that sch and sch OX were equal in effect on basal and stimulated secretion volume as compared with PL it was not possible to establish an effect on secretion volume for oxazepam CT sch and sch oxazepam were found to be equipotent in reducing the amount of basal acid while oxazepam reduced this quantity only during the first min of basal secretion none of the three AS S9 was capable of inhibiting the stimulated acid although both sch preparations produced a clear trend towards lowered mean values during the basal secretion period all three test S9 had an inhibiting action on acid concentration but none of them had a significant effect during the stimulation period the ph value was savely increased only by sch and sch OX and this even only during the basal period the results are discussed	234	oxazepam
6	a doubleblind T0 with intraindividual comparisons was carried out to investigate the effects of mg of ralphahydroxyisopropylalphahtropanium bromidetropate sch mg sch mg OX mg oxazepam and placebo with p.o. in randomized CS on gastric juice volume amount of acid concentration and ph values in healthy volunteers the secretion parameters were measured during a h basal period and a h stimulation period the gastric juice was obtained in min portions via stomach tube stimulation was effected by mugkgh PG via drip infusion the friedman test was used for the comparative statistical evaluation and individual comparisons were carried out by means of the wilcoxon test pairdifferences rank the results show that sch and sch OX were equal in effect on basal and stimulated secretion volume as compared with PL it was not possible to establish an effect on secretion volume for oxazepam CT sch and sch oxazepam were found to be equipotent in reducing the amount of basal acid while oxazepam reduced this quantity only during the first min of basal secretion none of the three AS S9 was capable of inhibiting the stimulated acid although both sch preparations produced a clear trend towards lowered mean values during the basal secretion period all three test S9 had an inhibiting action on acid concentration but none of them had a significant effect during the stimulation period the ph value was savely increased only by sch and sch OX and this even only during the basal period the results are discussed	112	oxazepam

Figure 4. Multiple rows with same TEXT but Different LOCATION

3.4 Data Preprocessing

As per EDA, we can see that there are a whole medical extracts in the TEXT column. At this point, we perform the following Data Preprocessing steps to make the data ready for the next step.

1. **Feature Engineering:** Created a new feature-ABV using the Location and Text columns. This feature contains the Abbreviation whose label is to be identified.
2. **Lowercase:** Convert data in Text column to lowercase.
3. **Punctuations:** Remove punctuations marks from data in Text column.

4. **Tokenization:** Convert the data in Text column into Tokens.
5. **Dropping Columns:** Dropping columns which are not required. In this case Abstract_Id and Location.
6. **Stop words:** Remove stop words from the Tokenized array.

Due to Hardware restrictions, I have extracted 20 unique Abbreviations having more than 250 unique expanded forms. This new dataset has an around 70K row which has further been divided into three parts:

1. **Train set:** having around 50K records.
2. **Validation set:** having around 10K records.
3. **Test set:** having around 10K records.

3.5 Paragraph Embeddings

Once done with data preprocessing, the data is ready to be converted to Vector form. Since the problem here is dependent on context of the data, I used paragraph embedding techniques so that the context is retained when the Tokenized array is converted into a vector form. For this I used Doc2Vec method of Gensim library with a minimum occurrence count of 2 and window size of 2.

3.6 Data Modeling

At this point, our test data was ready to be fed to different models to train them. For modeling, I used three of the most popular Classifiers, namely:

1. Logistic Classifier
2. Support Vector Classifier(SVC)
3. XG Boost Classifier

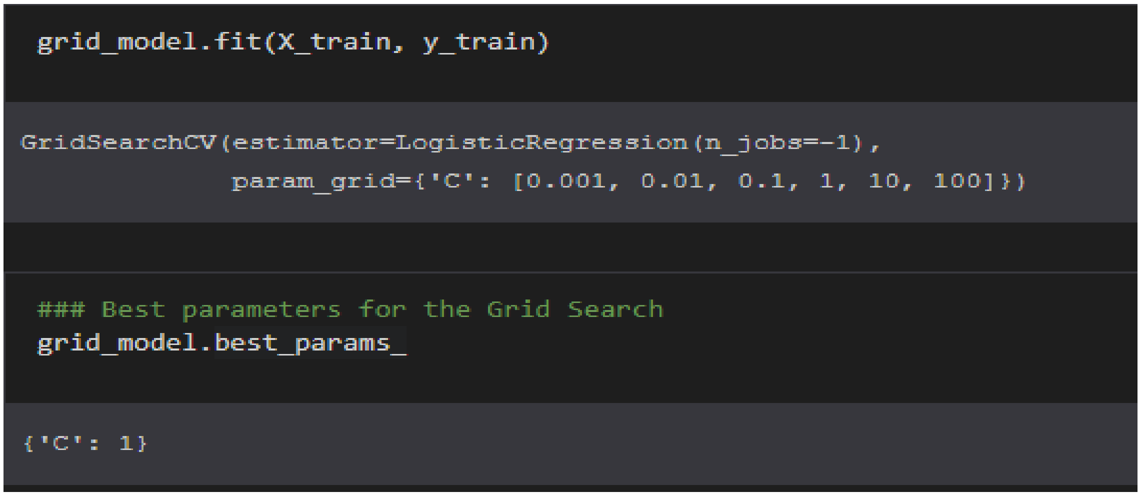
3.7 Training the Model

The dataset was divided into 3 proportions of 50K: 10K, Train to Validation data to train the model and validate it against the Validation Data. A separate set of 10K rows were kept for Prediction from the trained model (Test dataset).

The Models were first passed through GridSearchCV to find the best parameters. Once completed, the final model was created using the best parameters and validated using the Validation dataset.

3.7.1 Logistic Classifier

The Grid Search when executed with this classifier gave C = 1 as the best parameter (Figure 5) hence, this classifier was trained with the same value.



```
grid_model.fit(X_train, y_train)
```

```
GridSearchCV(estimator=LogisticRegression(n_jobs=-1),
             param_grid={'C': [0.001, 0.01, 0.1, 1, 10, 100]})
```

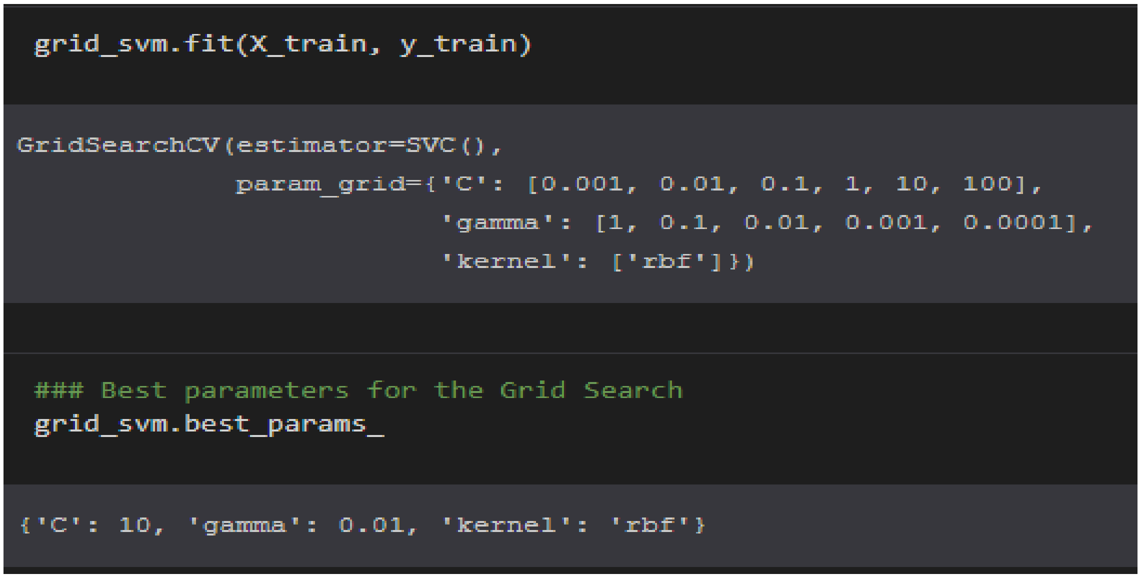
```
### Best parameters for the Grid Search
grid_model.best_params_
```

```
{'C': 1}
```

Figure 5. Grid Search with Logistic Regression

3.7.2 Support Vector Classifier (SVC)

The Grid Search when executed with this classifier gave C = 10, gamma = 0.01 as the best parameters. The kernel used was ‘rbf’. Hence, this classifier was trained with these parameters.



```
grid_svm.fit(X_train, y_train)
```

```
GridSearchCV(estimator=SVC(),
             param_grid={'C': [0.001, 0.01, 0.1, 1, 10, 100],
                         'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
                         'kernel': ['rbf']})
```

```
### Best parameters for the Grid Search
grid_svm.best_params_
```

```
{'C': 10, 'gamma': 0.01, 'kernel': 'rbf'}
```

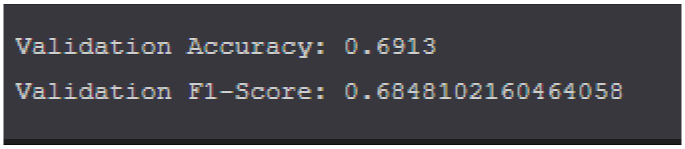
Figure 6. Grid Search with SVC

4 Results & Discussion

The trained models were used on Test dataset to predict the expanded forms and the accuracy metrics were evaluated for all the three models. The metrics used for this project are Accuracy Score, Average Precision Score, Average Recall Score and F1-Score.

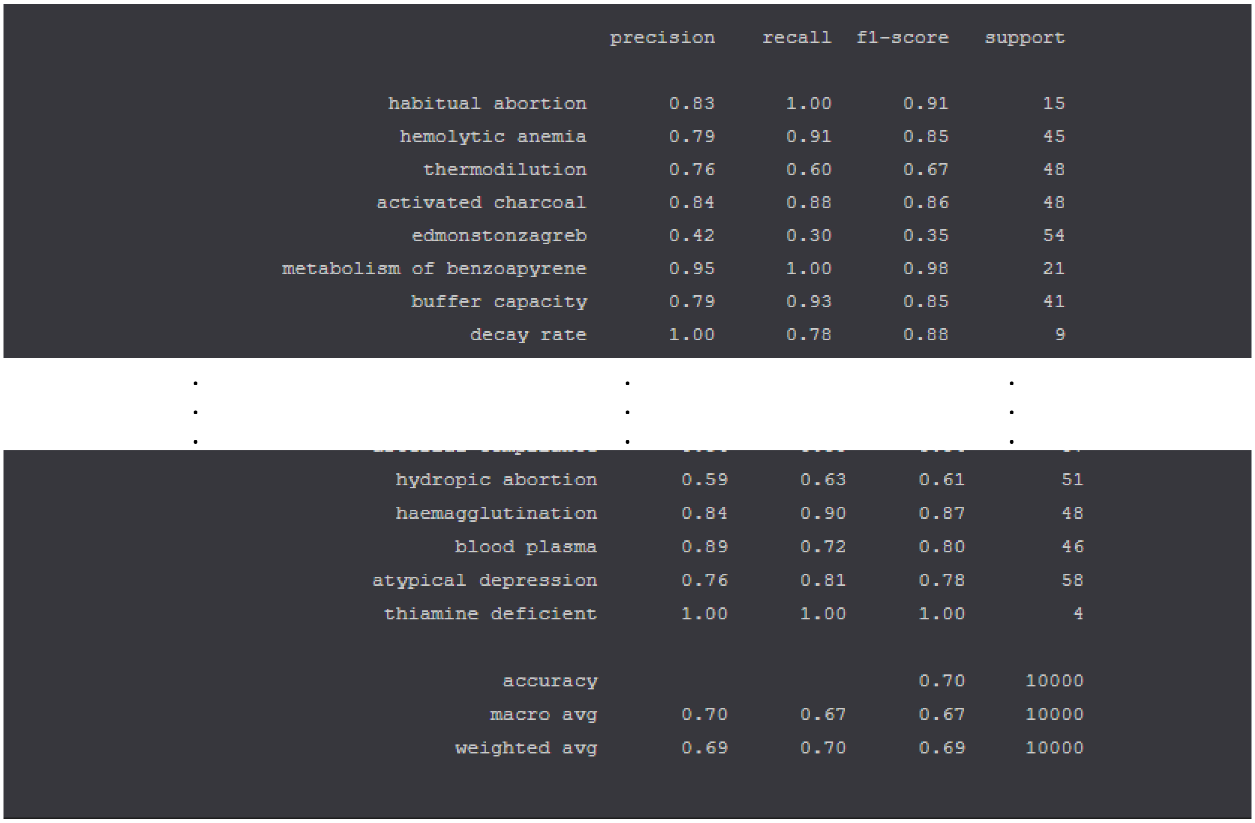
4.1 Logistic Classifier

Logistic Classifier was trained on the 50K rows of the train dataset and then validated using the validation dataset. The model gave an Accuracy score of 69% and an F1-Score of 0.68 on Validation dataset (Figure 7). This model was then used on Test dataset to predict the results. The model achieved an Accuracy score of 69.2% and F1-Score of 0.69 (Figure 8). Along with Accuracy and F1-Score, Average Precision and Recall values were also calculated for test dataset (Figure 8).



```
Validation Accuracy: 0.6913
Validation F1-Score: 0.6848102160464058
```

Figure 7.*Validation dataset result metrics*

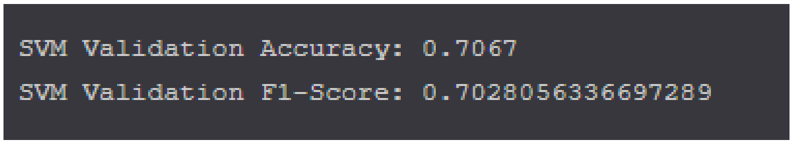


```
                              precision    recall  f1-score   support

           habitual abortion       0.83      1.00      0.91        15
             hemolytic anemia       0.79      0.91      0.85        45
              thermodilution       0.76      0.60      0.67        48
            activated charcoal       0.84      0.88      0.86        48
              edmonstonzagreb       0.42      0.30      0.35        54
     metabolism of benzoapyrene       0.95      1.00      0.98        21
              buffer capacity       0.79      0.93      0.85        41
                   decay rate       1.00      0.78      0.88         9
        .                        .                                 .
        .                        .                                 .
        .                        .                                 .
            hydropic abortion       0.59      0.63      0.61        51
            haemagglutination       0.84      0.90      0.87        48
                 blood plasma       0.89      0.72      0.80        46
          atypical depression       0.76      0.81      0.78        58
           thiamine deficient       1.00      1.00      1.00         4

                     accuracy                           0.70     10000
                    macro avg       0.70      0.67      0.67     10000
                 weighted avg       0.69      0.70      0.69     10000
```

Figure 8.*Test Dataset with Precision and Recall values*

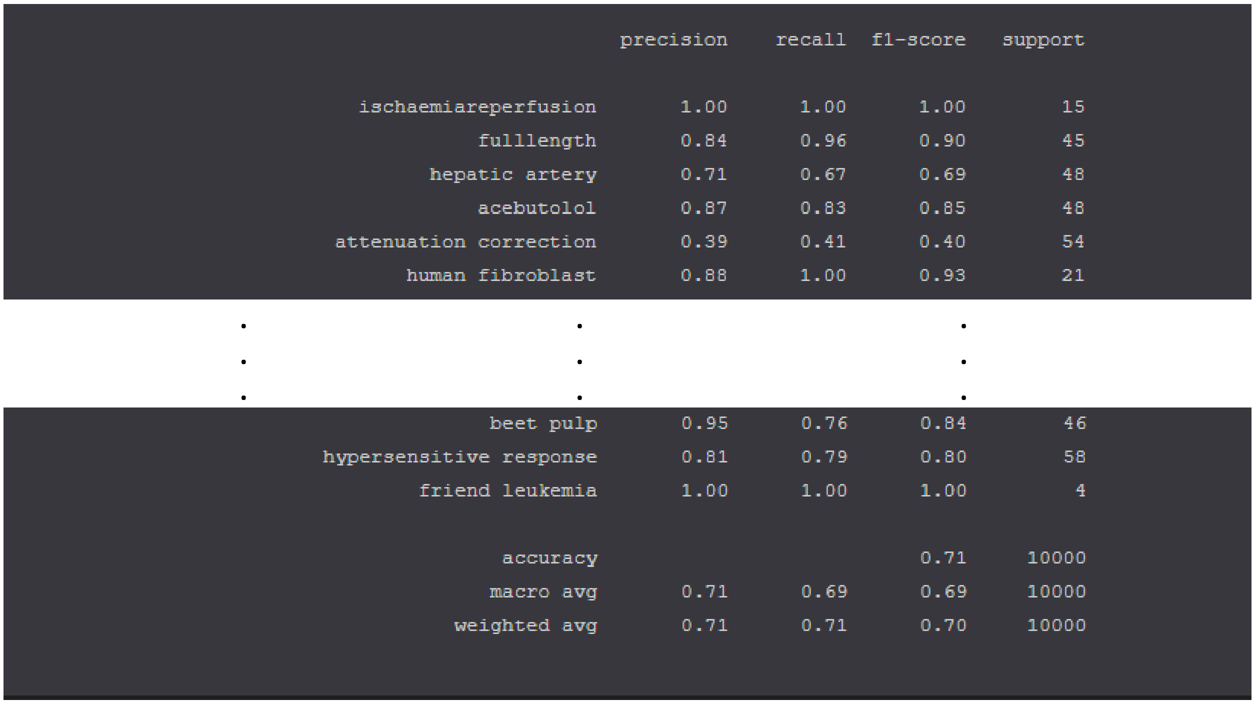
4.2 Support Vector Classifier

Support Vector Classifier was trained on the 50K rows of the train dataset and then validated using the validation dataset. The model gave an Accuracy score of 70.6% and an F1-Score of 0.70 on Validation dataset (Figure 9). This model was then used on Test dataset to predict the results. The model achieved an Accuracy score of 70.5% and F1-Score of 0.70 (Figure 10). Along with Accuracy and F1-Score, Average Precision and Recall values were also calculated for test dataset (Figure 10).



```
SVM Validation Accuracy: 0.7067
SVM Validation F1-Score: 0.7028056336697289
```

Figure 9. *Validation dataset result metrics*

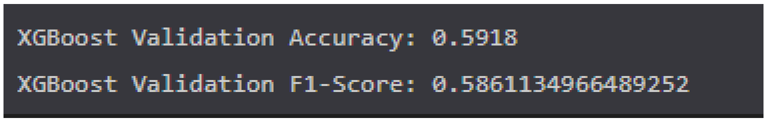


```
                           precision    recall  f1-score   support

     ischaemiareperfusion       1.00      1.00      1.00        15
               fulllength       0.84      0.96      0.90        45
           hepatic artery       0.71      0.67      0.69        48
               acebutolol       0.87      0.83      0.85        48
   attenuation correction       0.39      0.41      0.40        54
         human fibroblast       0.88      1.00      0.93        21
.                            .                        .
.                            .                        .
.                            .                        .
                beet pulp       0.95      0.76      0.84        46
  hypersensitive response       0.81      0.79      0.80        58
          friend leukemia       1.00      1.00      1.00         4

                 accuracy                           0.71     10000
                macro avg       0.71      0.69      0.69     10000
             weighted avg       0.71      0.71      0.70     10000
```

Figure 10.*Test Dataset with Precision and Recall values*

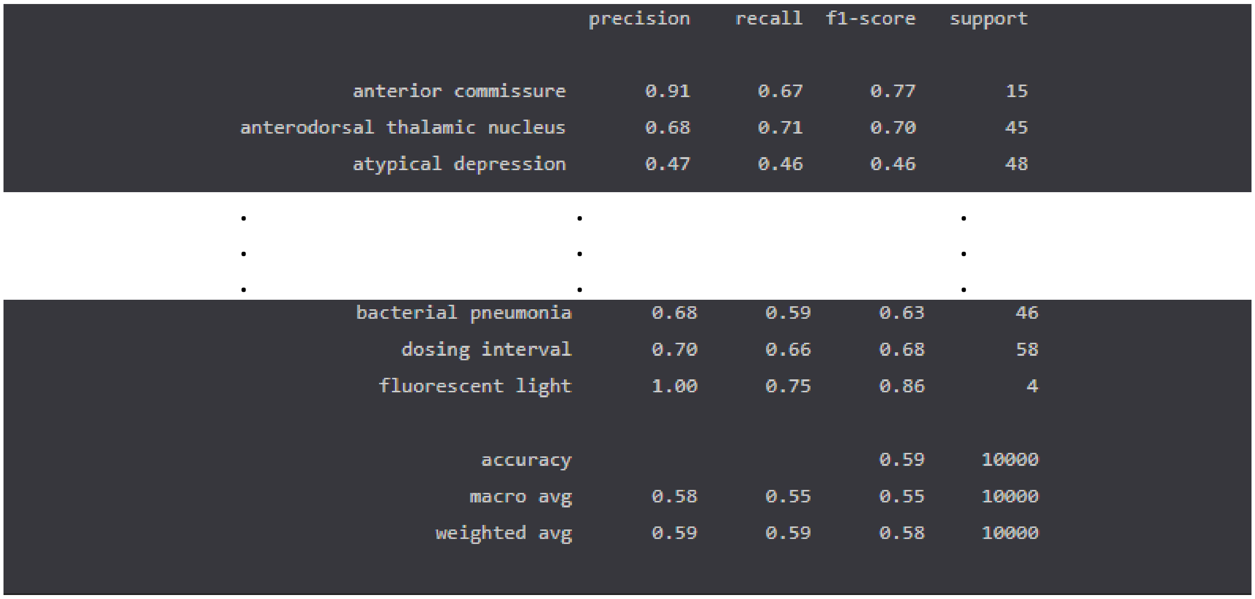
4.3 XG Boost Classifier

XG Boost Classifier was trained on the 50K rows of the train dataset and then validated using the validation dataset. The model gave an Accuracy score of 59.2% and an F1-Score of 0.59 on Validation dataset (Figure 11). This model was then used on Test dataset to predict the results. The model achieved an Accuracy score of 58.9% and F1-Score of 0.58 (Figure 12). Along with Accuracy and F1-Score, Average Precision and Recall values were also calculated for test dataset (Figure 12).



```
XGBoost Validation Accuracy: 0.5918
XGBoost Validation F1-Score: 0.5861134966489252
```

Figure 11. *Validation dataset result metrics*



```
                              precision    recall  f1-score   support


        anterior commissure       0.91      0.67      0.77        15
 anterodorsal thalamic nucleus       0.68      0.71      0.70        45
        atypical depression       0.47      0.46      0.46        48
      .                      .                      .
      .                      .                      .
      .                      .                      .
         bacterial pneumonia       0.68      0.59      0.63        46
             dosing interval       0.70      0.66      0.68        58
          fluorescent light       1.00      0.75      0.86         4

                   accuracy                           0.59     10000
                  macro avg       0.58      0.55      0.55     10000
               weighted avg       0.59      0.59      0.58     10000
```

Figure 12.*Test Dataset with Precision and Recall values*

Table 1 below shows a summary of results produced by all three the Models.

Metric	Logistic	SVC	XG Boost
Validation Accuracy	69.1%	70.7%	59.2%
Test Accuracy	69.6%	70.6%	58.9%
Precision	69%	71%	59%
Recall	70%	71%	59%
F1 Score	0.69	0.70	0.58

Table 1.*Comparison of results of three models*

5 Conclusion

This project employed the implementation of some of the well-known NLP libraries like spacy and gensim to understand the context of a medical text and based on that context predict what the expanded form of an Abbreviation (if any) is in that medical text. The classification was performed using as subset of MeDAL Dataset containing Medical Extracts having Abbreviations and there Location and Expanded forms. Paragraph Embeddings were created using Gensim library and then those were used as inputs to some of the well-known classifiers. All the models created, showed a satisfactory Accuracy score with Support Vector Classifier performing the best with an Accuracy of 71% on both Validation and Test Datasets.

Regarding further work, other well-known Classifiers can be implemented on the Dataset to see how there results fare with the results obtained above. A little Hyper-parameter tuning while generating Paragraph Embeddings can also be done to identify the Ideal window size for Text Context extraction. An Ensemble of all the above Models can also be implemented with a polling system to decide the outcome as we have already seen that an Ensemble of different Methods can lead to better prediction accuracies.

In future, this Model can also be generalized to predict expanded forms for Abbreviations for not only Medical Extracts but other domains as well, like Banking and Astronomy. The model can be trained to detect the Drugs and Adverse Reactions in the Medical Journals as well.

6 References

1. MeDAL: Medical Abbreviation Disambiguation Dataset for Natural Language Understanding Pretraining by Zhi Wen, Xing Han Lu and Siva Reddy.
2. Clinical abbreviations and acronyms - A word-sense disambiguation problem by Prathamesh Prabhudesai.
3. Automated Disambiguation of Acronyms and Abbreviations in Clinical Texts: Window and Training Size Considerations by Sungrim Moon, Serguei Pakhomov and Genevieve B. Melton.
4. Multi-Class Text Classification with Doc2Vec & Logistic Regression by Susan Li
5. Dataset Sources:
    a. https://www.kaggle.com/xhlulu/medal-emnlp