Report

Introduction

ProjectTitle: Genetic Mutation Classification

A lot has been said during the past several years about how precision medicine and, more concretely, how genetic testing is going to disrupt the way diseases like cancer are treated.
But this is only partially happening due to the huge amount of manual work still required.
Once sequenced, a cancer tumor can have thousands of genetic mutations. But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers).

Currently this interpretation of genetic mutations is being done manually. This is a very time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature.

For this , we have got dataset from a Kaggle competition. This dataset is an expert-annotated knowledge base where world-class researchers and oncologists have manually annotated thousands of mutations.

Description

In this project, we have used algorithms to classify genetic mutations based on clinical evidence (text).

There are 9 different classes a genetic mutation can be classified on.

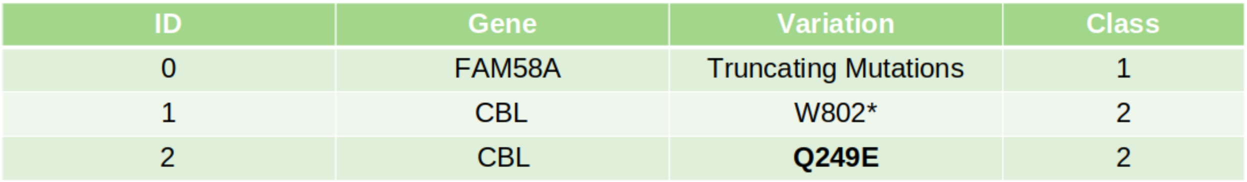
This is not a trivial task since interpreting clinical evidence is very challenging even for human specialists. Therefore, modeling the clinical evidence (text) will be critical for the success of our approach.

Both, training and test, data sets are provided via two different files. One (training/test_variants) provides the information about the genetic mutations, whereas the other (training/test_text) provides the clinical evidence (text) that our human experts used to classify the genetic mutations. Both are linked via the ID field.

Therefore the genetic mutation (row) with ID=15 in the file training_variants, was classified using the clinical evidence (text) from the row with ID=15 in the file training_text

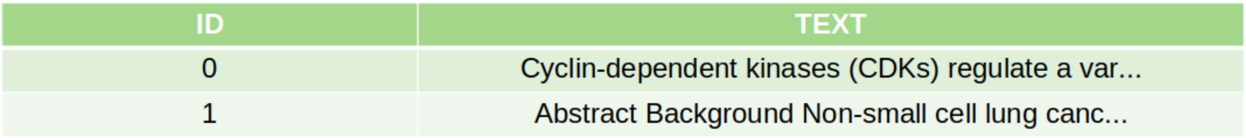
Data

The training data set contains 2124 records for genetic mutation classes i.e., 1 to 9. For Cross-Validation Data, 665 records of labelled data are there. For Test Data, 532 records of labelled data are there. Examples from the dataset -



ID	Gene	Variation	Class
0	FAM58A	Truncating Mutations	1
1	CBL	W802*	2
2	CBL	**Q249E**	2

Table 1. Examples from the dataset(Training_variants)



ID	TEXT
0	Cyclin-dependent kinases (CDKs) regulate a var...
1	Abstract Background Non-small cell lung canc...

Table 2. Examples from the dataset(Training_text)

NOTE : TEXT field is like a large paragraph. So, we have shown only some part of it above.

Methodology

As we have only 3234 inputs, we have used ML algorithms(other than DL) to make our models with.

We first merged ‘Training_variants’ and ‘Training_text’ files on ‘ID’ column. Due to only 87 rows in the ‘TEXT’ column, many ‘NAN’s were there in the merged data. We replaced ‘NAN' with the string formed by concatenating ‘Gene’ and ‘Variation’ values of the corresponding row.

Then we removed stopwords, special characters and multiple spaces from the ‘TEXT’ data.

After that we encoded the text data with different embeddings – Onehot, TF-IDF, Response Coding and Word2Vec.

And then we applied different ML algorithms on the preprocessed data and noted the result.

Evaluation

For evaluation, we considered ‘log loss’ as the metric as it was also used in Kaggle competition.After performing some data preprocessing and some visualizations, we applied the models on the data and got the following results.

We can see that (ResponseCoding + KNN) outperformed all other models.

Model	Test-log loss
1. Onehot + SGD Classifier(Logistic Regression) with class_weight = 'balanced'	1.124
2. TF-IDF + SGD Classifier(LR) with class_weight = 'balanced'	1.113
3. Onehot + SGD Classifier(Logistic Regression) with class_weight = 'None'(default)	1.115
4. TF-IDF + SGD Classifier(LR) with class_weight = 'None'	1.148
5. Onehot + Linear SVM	1.297
6. TF-IDF + Linear SVM	1.225
7. Onehot + Random Forest	1.307
8. TF-IDF + Random Forest	1.313
9. Response Coding + Random Forest	1.394
10. Word2Vec + Random Forest	1.268
11. *Response Coding + K-Nearest Neighbors*	***1.107***
12. Onehot + Naive Bayes	1.203
13. TF-IDF + Naïve Bayes	1.219
14. Maximum Voting Classifier	1.234