Personalized Medicine: Redefining Cancer Treatment

<DataGeeks/>

AGENDA

1. Problem Definition
2. Dataset Explanation
3. Methodology
4. Results
5. Discussion

Problem Definition

- Develop a robust and accurate model to automate the process of identifying carcinogenic genetic mutations from medical text records.
- Could be analyzed as a text classification problem.
- Possible to eliminate the time consuming manual efforts.
- A chance to predict the cancer causing mutations early and treat the patient’s tumor at a preliminary stage.

Dataset Explanation

- Four files in total - training_variants, test_variants, training_text and test_text.
- Variants Files:
  - Information about genetic mutations
  - CSV file with 4 columns - ID, Gene (gene where mutation is located), Variation (Amino acid change for this mutation), Class (1-9 class where this mutation is classified on)
  - The test_variants file does not contain the last column specifying the class.
- Text Files:
  - Clinical evidence (text) that human experts used to classify genetic mutations.
  - Double pipe (||) delimited file with 2 fields - ID, Text (clinical evidence obtained from various research papers)
- Both sets of training and test files are linked using the ID field.
- Some of the test data is machine-generated to prevent hand labeling. Kaggle ignores the results from machine-generated samples in the final result.

Dataset Glimpse

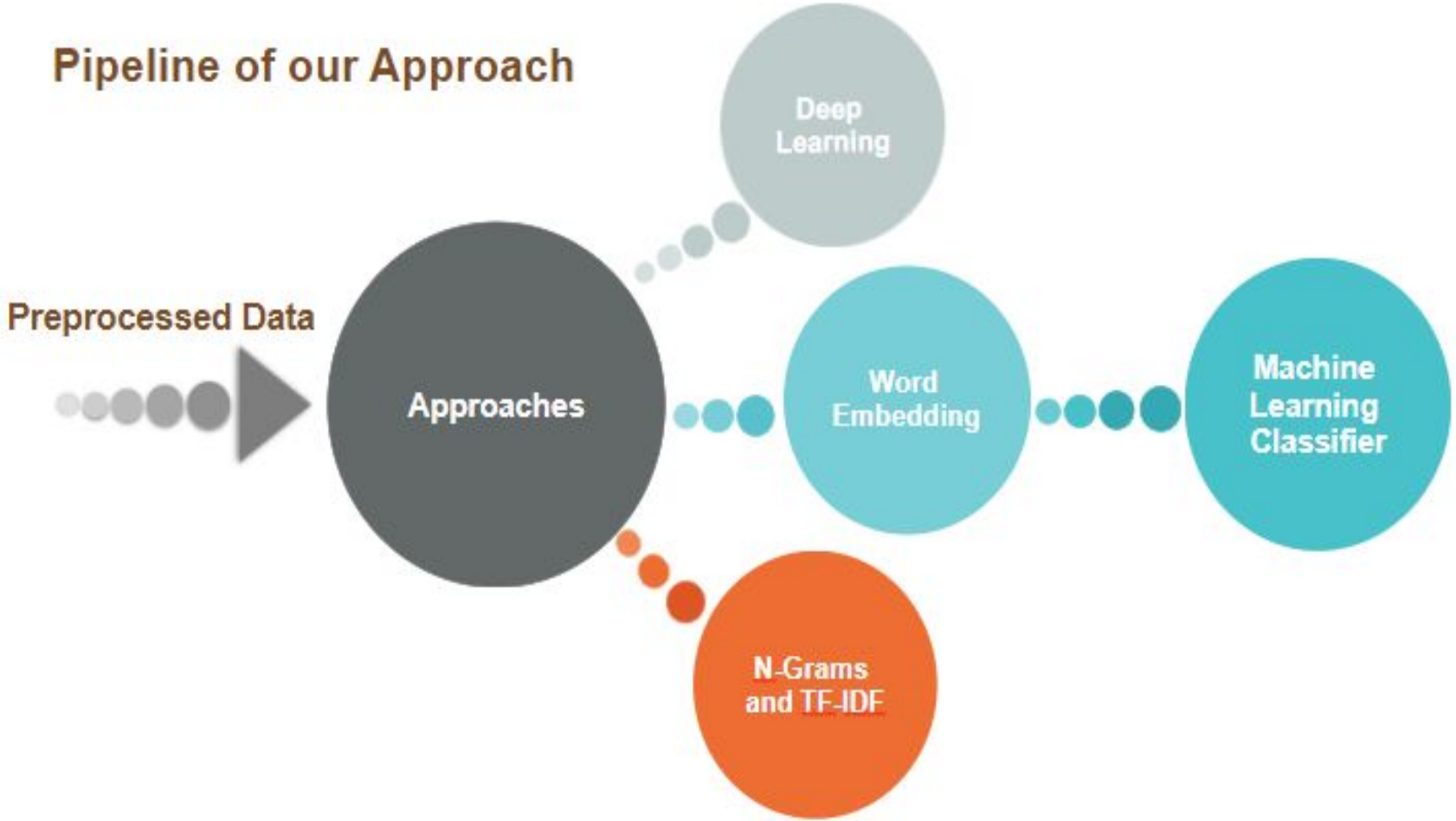
Training Variants :

ID	Gene	Variation	Class
0	FAM58A	Truncating Mutations	1
1	CBL	W802*	2
2	CBL	Q249E	2
3	CBL	N454D	3
4	CBL	L399V	4

Training Text :

ID	txt
0	Cyclin-dependent kinases (CDKs) regulate a variety of fu...
1	Abstract Background Non-small cell lung cancer (NSCLC)...
2	Abstract Background Non-small cell lung cancer (NSCLC)...
3	Recent evidence has demonstrated that acquired unipar...
4	Oncogenic mutations in the monomeric Casitas B-lineag...

Pipeline of our Approach



Feature Engineering

- TF-IDF
  - Text Data: Unigrams and Bigrams on word level.
  - Gene and Variants Encoding: 1-10 character level n-grams
  - Label Encoding for Gene and Variants
- Word Embeddings
  - Glove vectors trained on general domain web data.
  - Word2vec trained on PubMed medical text.
  - Word2vec trained on our data.
- Deep Learning
  - Long Short Term Memory Networks
  - Hierarchical LSTMs
  - Convolutional Neural Networks

Results

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$$

Algorithm	Private Score	Public Score
Hierarchical LSTMs	3.18	1.30
CNN	2.50	2.42
TFIDF + SVD + Gradient Boosting	2.41	2.11
Glove + Logistic Regression	2.80	1.44
Word2Vec (PubMed) + Gradient Boosting	2.45	1.35
Word2Vec (Training Data) + Gradient Boosting	**1.98**	**1.22**

Learned Embeddings

Mutation	Cosine Distance
substitution	0.269
variant	0.297
alteration	0.319
mutational	0.369
polymorphism	0.389

Disease	Cosine Distance
malignancy	0.260
tumor	0.395
illness	0.411
disorder	0.414
seizures	0.415

Public Leaderboard | Private Leaderboard

The private leaderboard is calculated with approximately 24% of the test data.

This comp

Refresh

Submission and Description	Private Score
ns_gbm_enc_latest.csv just now by sasank add submission details	1.98783

In the mo

#				Private Score	Entries	Last
1	▲ 303	ilmirashaim		2.03026	6	2mo
2	▲ 314	Waterpls		2.09095	11	2mo
3	▲ 189	Yang 3		2.12814	4	2mo
4	▼ 1	FourteenthTokyo		2.13316	21	2mo
5	▲ 320	Bcottman		2.13364	6	2mo
6	▲ 96	varstation		2.13613	9	2mo
7	▲ 60	NCTU_GoldX5		2.17964	14	2mo
8	▲ 61	DaXian		2.17964	11	2mo

Discussion

1. Deep Learning based approaches performed poorly → Overfitting the training data.
2. Glove < Word2vec (Pubmed) < Word2vec trained on the dataset. Embeddings trained on the dataset performed better than pretrained word2vec embeddings and Glove Embeddings.
3. Take home: Learning word vector embeddings on training data is extremely effective. Boosting techniques significantly improve the performance of the classifiers.
4. Visualizing the Embeddings: *http://projector.tensorflow.org/*

Thank You!!