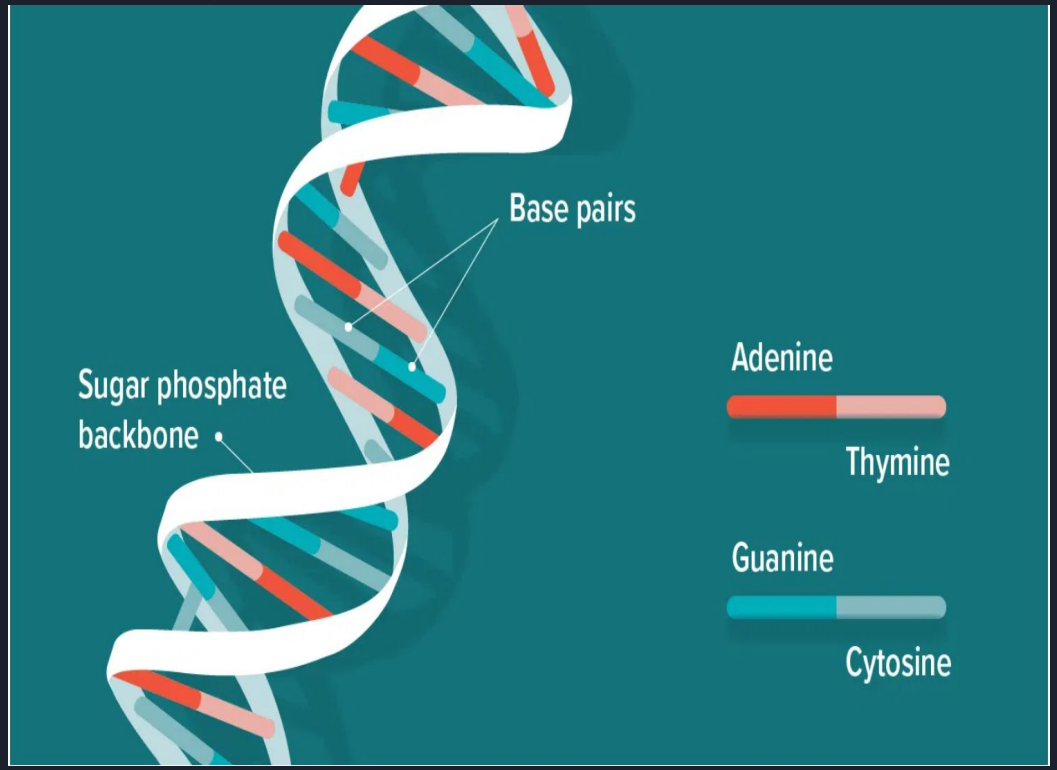
Genetic Defect Characterization for Disease Identification

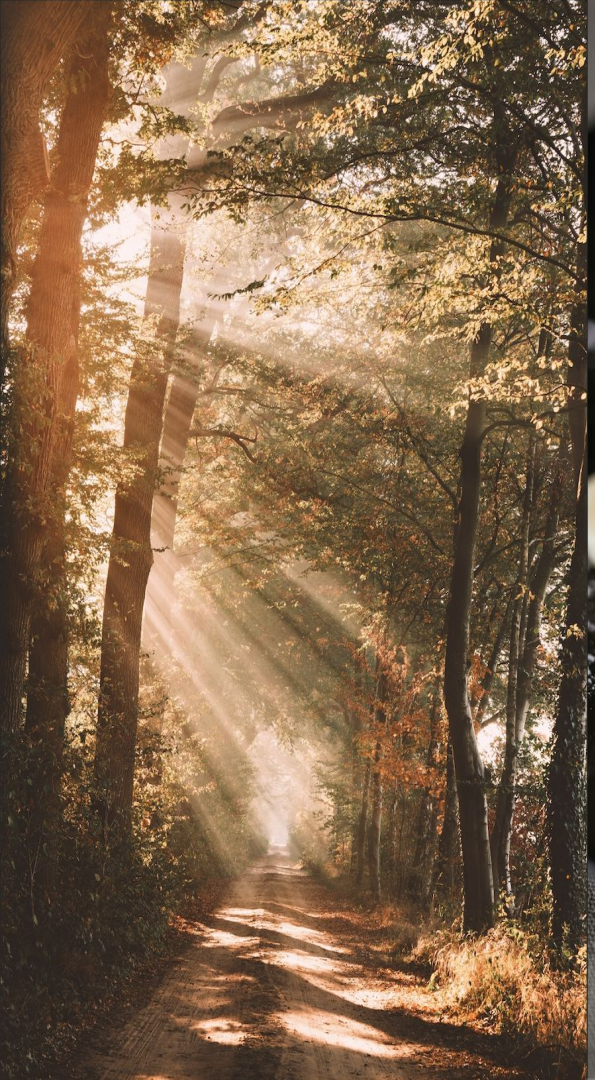
NLP Team:

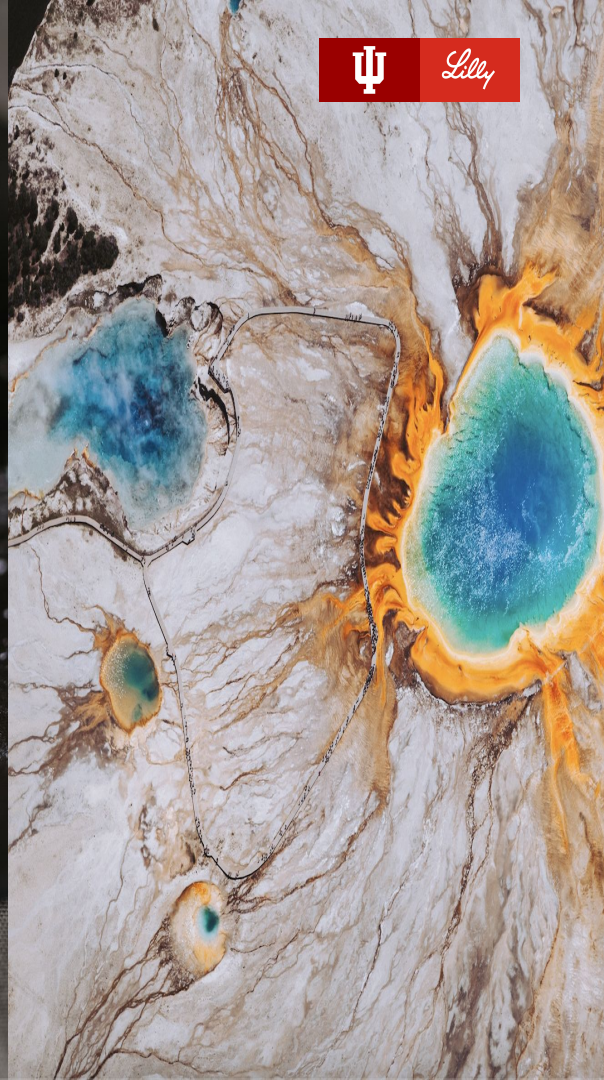
- Amanda Turney (Technical Coach)
- Prudhvi Vajja
- Sai Prasad Parsa
- Advait Save
- Vijay Iyer

DNA - DeoxyriboNucleic Acid



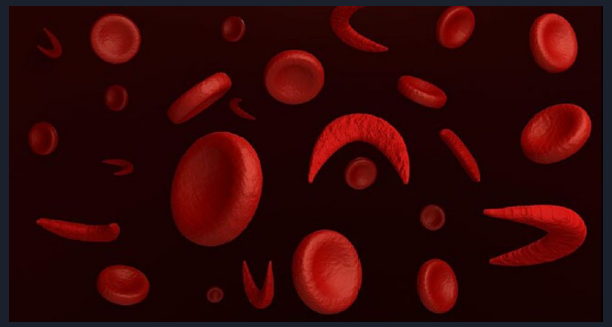
- Genetic code of all beings
- We are born with 2 sets of more than 3 billions letters of DNA from our parents.
- Point Mutation
- Missense
- Silent , Etc.

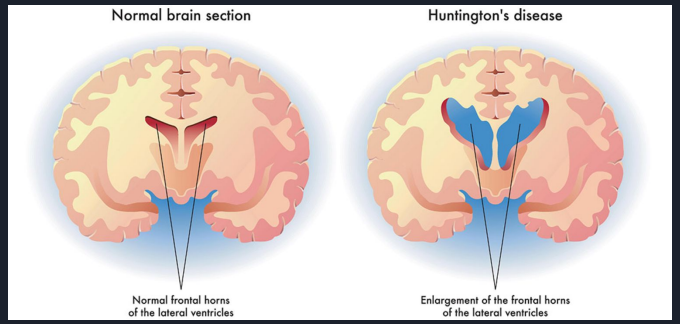




=> Cystic fibrosis

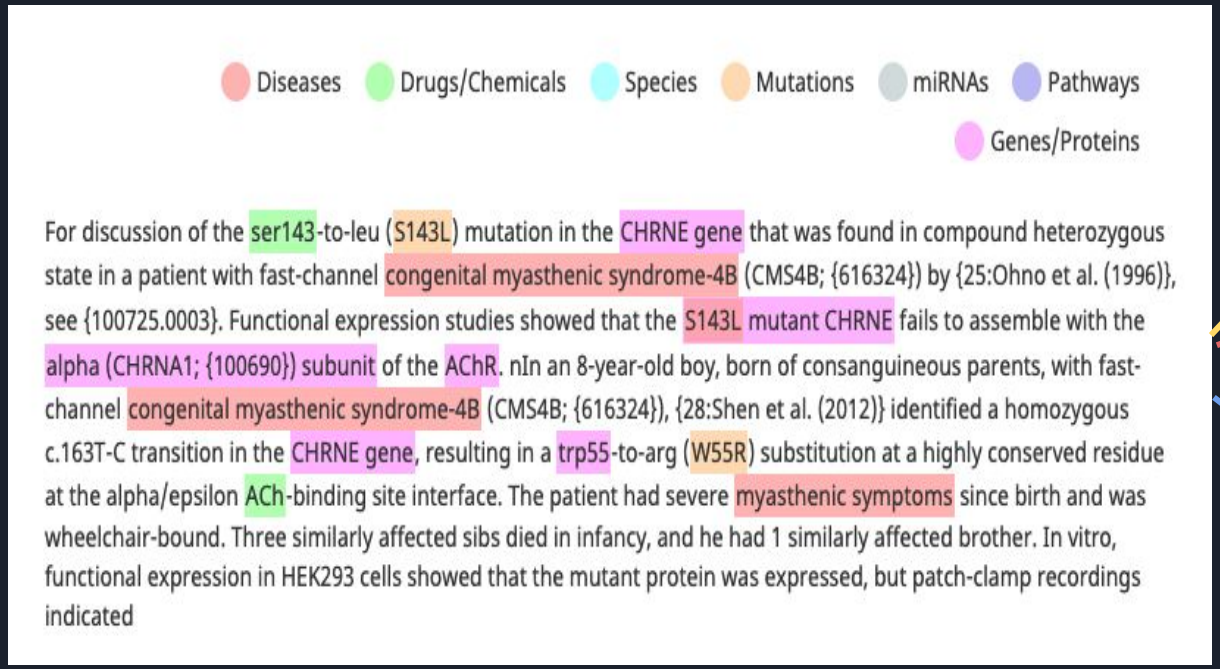
Sickle Cell Anemia <=





=> Huntington Disease

Overview of the Project



Loss of Function

Gain of Function

Dominant Negative

HaploInsufficiency

Project Objectives

Methodology

- Explore Data.
- Explore multiple NLP models.
- Making the models more robust.
- Evaluate the performance/metrics.
- Create a pipeline to automate the process.
- Improve Drug discovery.

Experimentation and Results

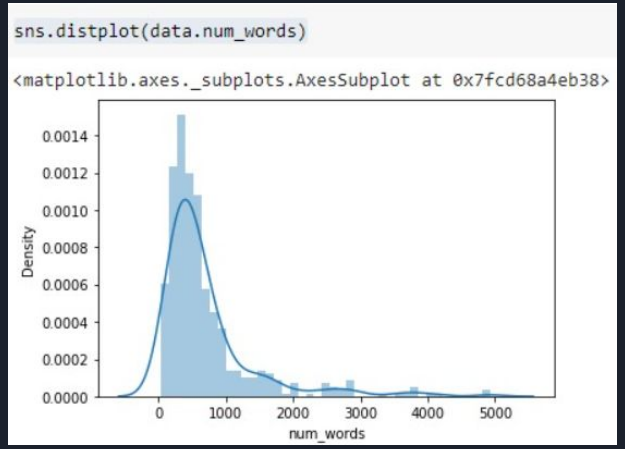
Learnings from Dataset

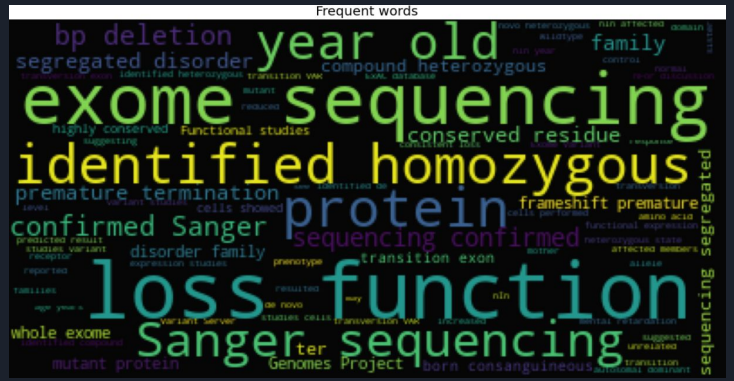
- Long aggregated literature per disease is a challenge for the NLP models

- Small dataset for a multiclass problem with 5 classes

  (Gain of Function, Loss of Function, Haploinsufficiency, Dominant Negative, None)

- Word Clouds reveal patterns (LOF wordcloud displayed)





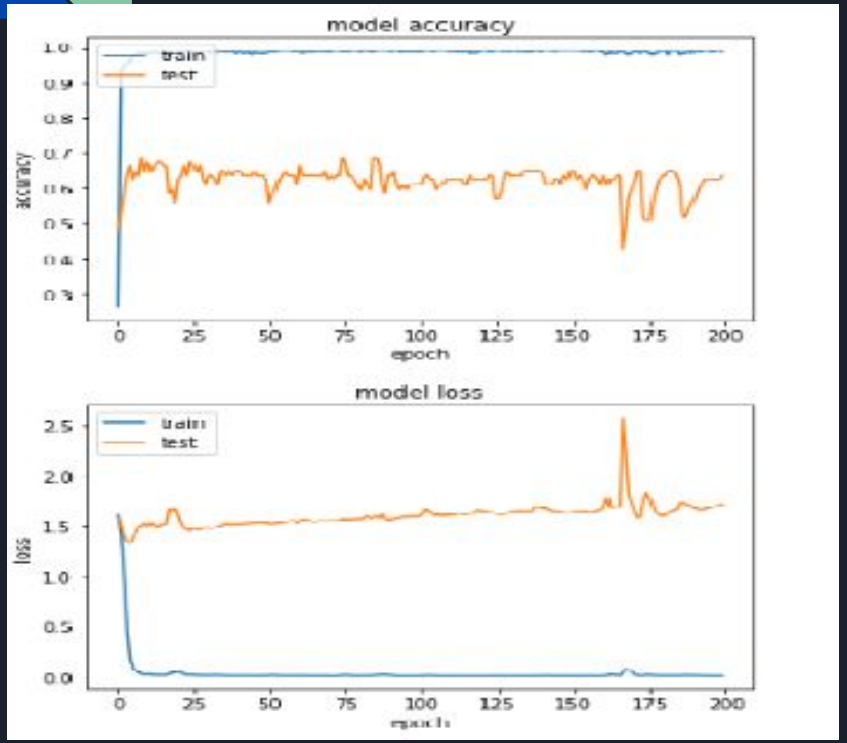
Baseline models overfit

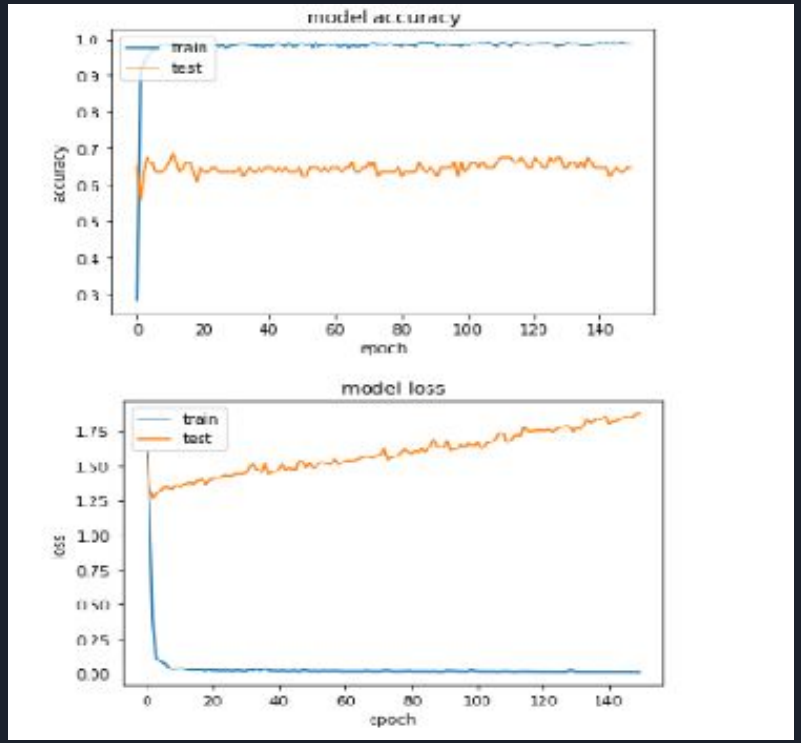
- Baseline Logistic models overfit on training dataset
- Cleaning text field improves performance slightly (removing genetic information)

	model_name	index	DN	GOF	HI	LOF	none	macro avg	accuracy	hamming_loss
1	TF-IDF + Basic Clean + Focused Text + GS LR ovr - Train	f1-score	0.925	0.957	0.940	0.947	1.000	0.954	0.952	0.048
3	TF-IDF bigram + Basic Clean + focused text + stopw_curated + without ids + GS LR ovr - Train	f1-score	0.986	0.985	0.993	0.985	1.000	0.986	0.985	0.015
5	TF-IDF + Basic Clean + Focused Text + GS LR ovr - Train	precision	0.895	0.943	0.955	0.984	1.000	0.956	0.952	0.048
7	TF-IDF bigram + Basic Clean + focused text + stopw_curated + without ids + GS LR ovr - Train	precision	0.986	0.985	1.000	1.000	1.000	0.986	0.985	0.015
9	TF-IDF + Basic Clean + Focused Text + GS LR ovr - Train	recall	0.958	0.971	0.926	0.913	1.000	0.954	0.952	0.048
11	TF-IDF bigram + Basic Clean + focused text + stopw_curated + without ids + GS LR ovr - Train	recall	0.986	0.985	0.986	0.971	1.000	0.986	0.985	0.015

	model_name	index	DN	GOF	HI	LOF	none	macro avg	accuracy	hamming_loss
0	TF-IDF + Basic Clean + Focused Text + GS LR ovr - Test	f1-score	0.708	0.815	0.577	0.652	0.746	0.679	0.681	0.319
2	TF-IDF bigram + Basic Clean + focused text + stopw_curated + without ids + GS LR ovr - Test	f1-score	0.758	0.889	0.755	0.678	0.821	0.780	0.778	0.222
4	TF-IDF + Basic Clean + Focused Text + GS LR ovr - Test	precision	0.636	0.848	0.714	0.593	0.847	0.698	0.681	0.319
6	TF-IDF bigram + Basic Clean + focused text + stopw_curated + without ids + GS LR ovr - Test	precision	0.876	0.923	0.909	0.714	0.742	0.793	0.778	0.222
8	TF-IDF + Basic Clean + Focused Text + GS LR ovr - Test	recall	0.793	0.786	0.484	0.516	0.850	0.692	0.681	0.319
10	TF-IDF bigram + Basic Clean + focused text + stopw_curated + without ids + GS LR ovr - Test	recall	0.862	0.857	0.645	0.645	0.920	0.786	0.778	0.222

Baseline models overfit

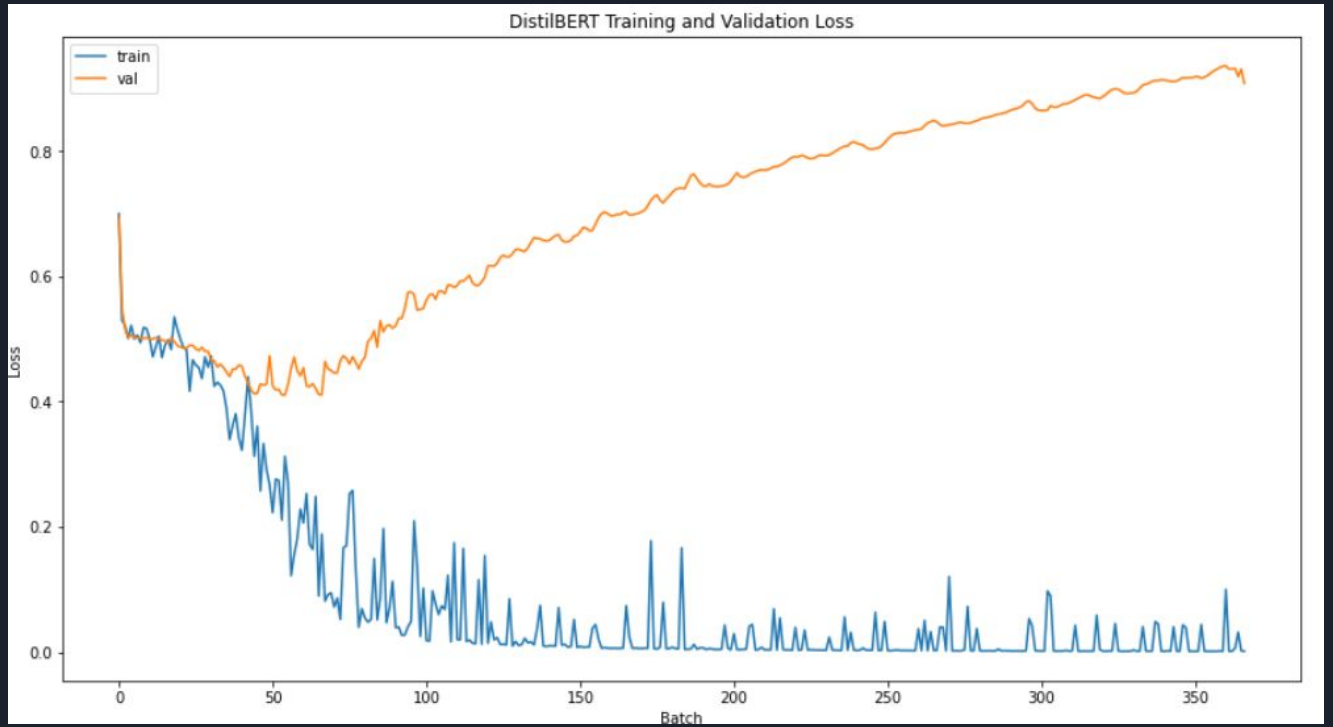




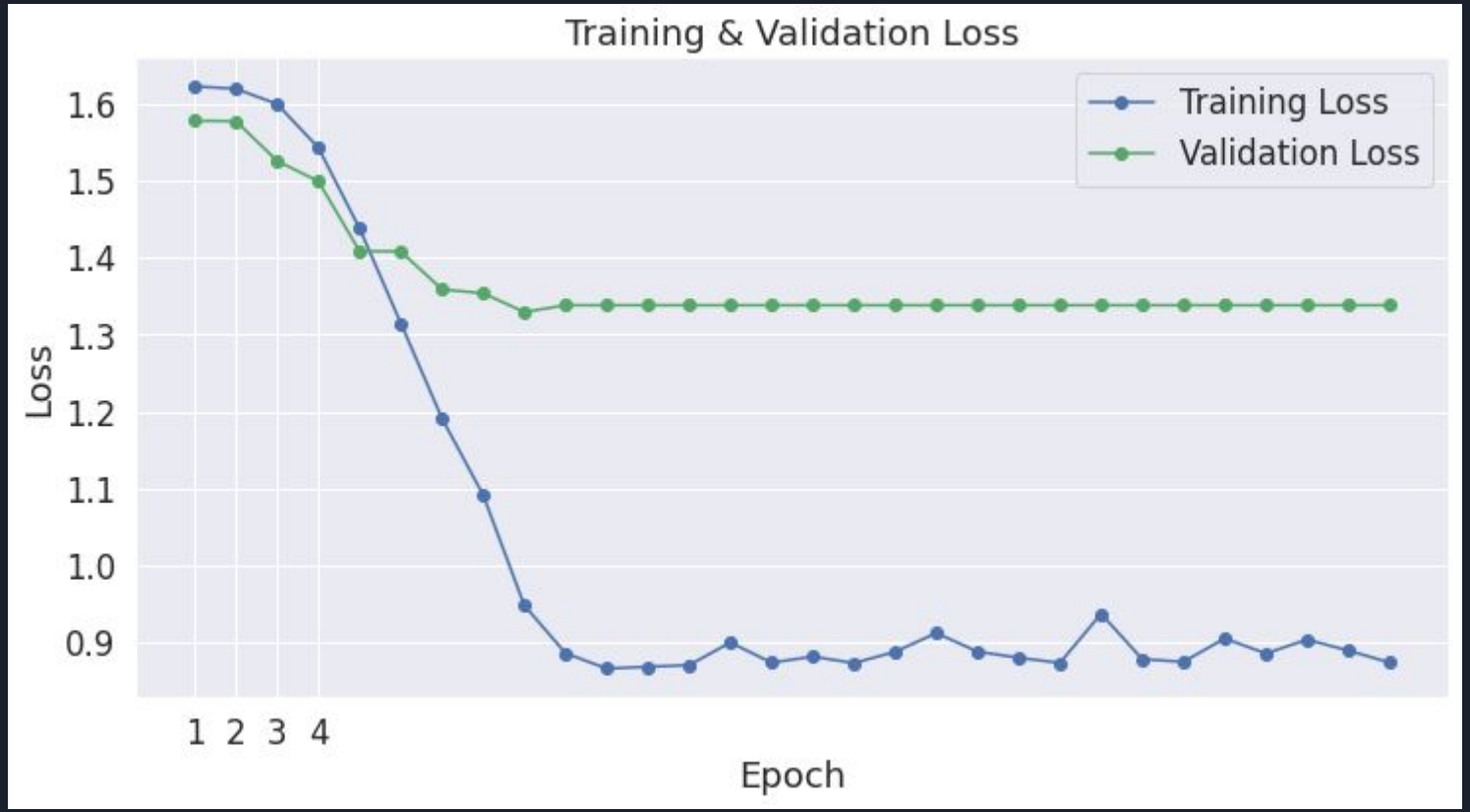
TFIDF + 2 LAYER FULLY CONNECTED NN

DistilBERT Fine-tuning Curve

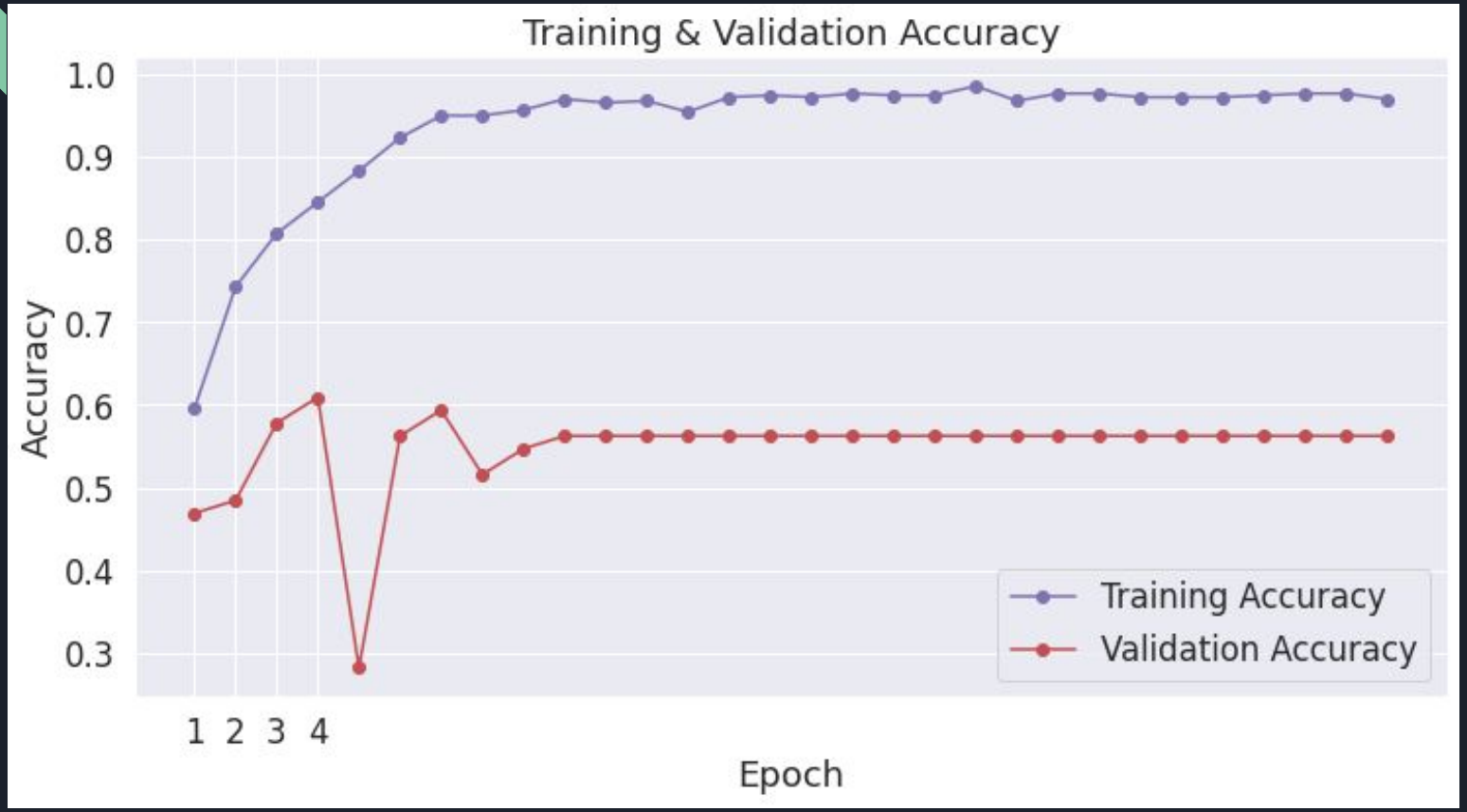
- DistilBERT model overfits after 5 epochs
- This suggests model is focusing on the wrong information in text



BlueBERT Classifier

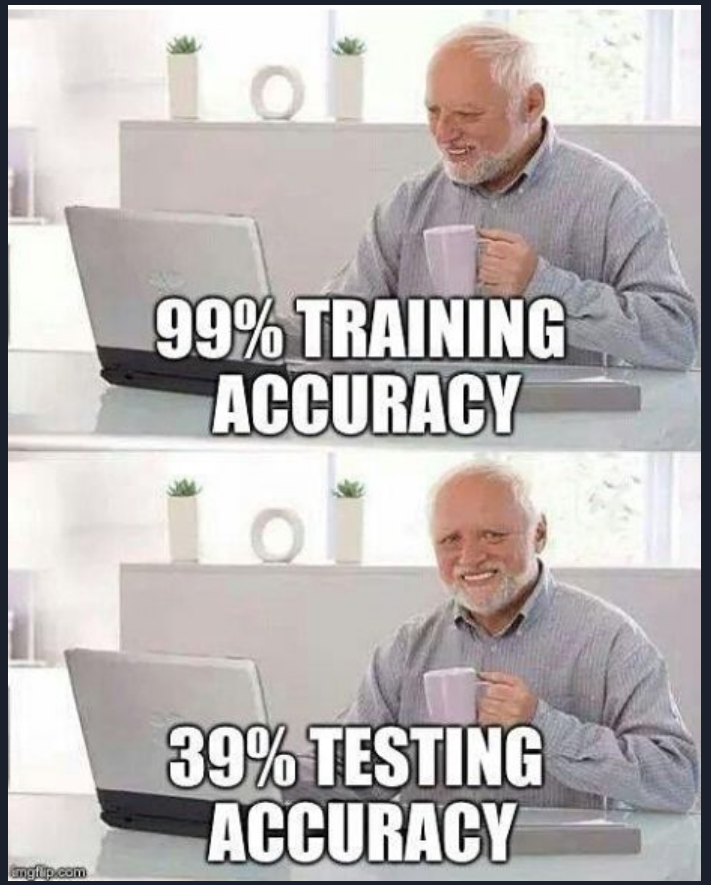


BlueBERT Performance



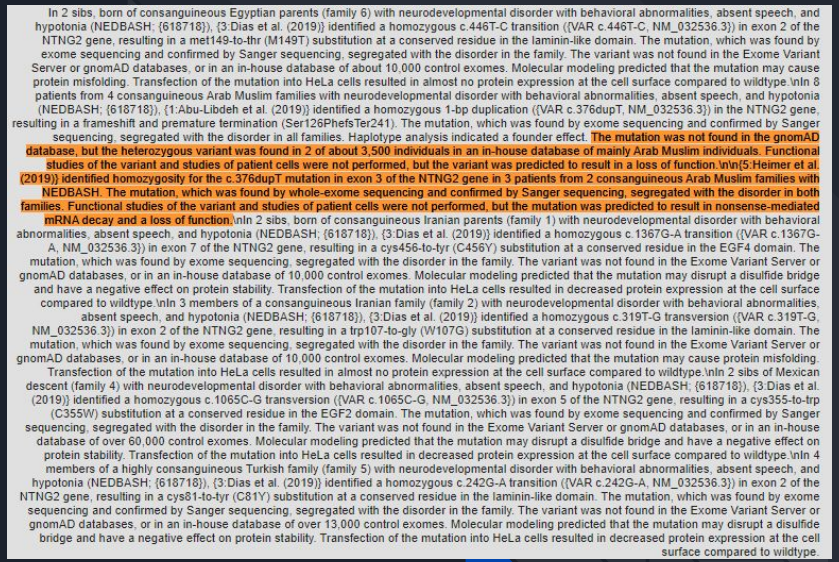
Challenges

- Small training dataset
- Overfitting
- Understanding misclassifications (lack of domain knowledge)



Improving Data Quality by Cleaning Text

- Only 2-3 sentences talk about relevant mutation defect effect
- Noise in long literature text leads to overfitting
- Reducing data to relevant text helps train NLP models

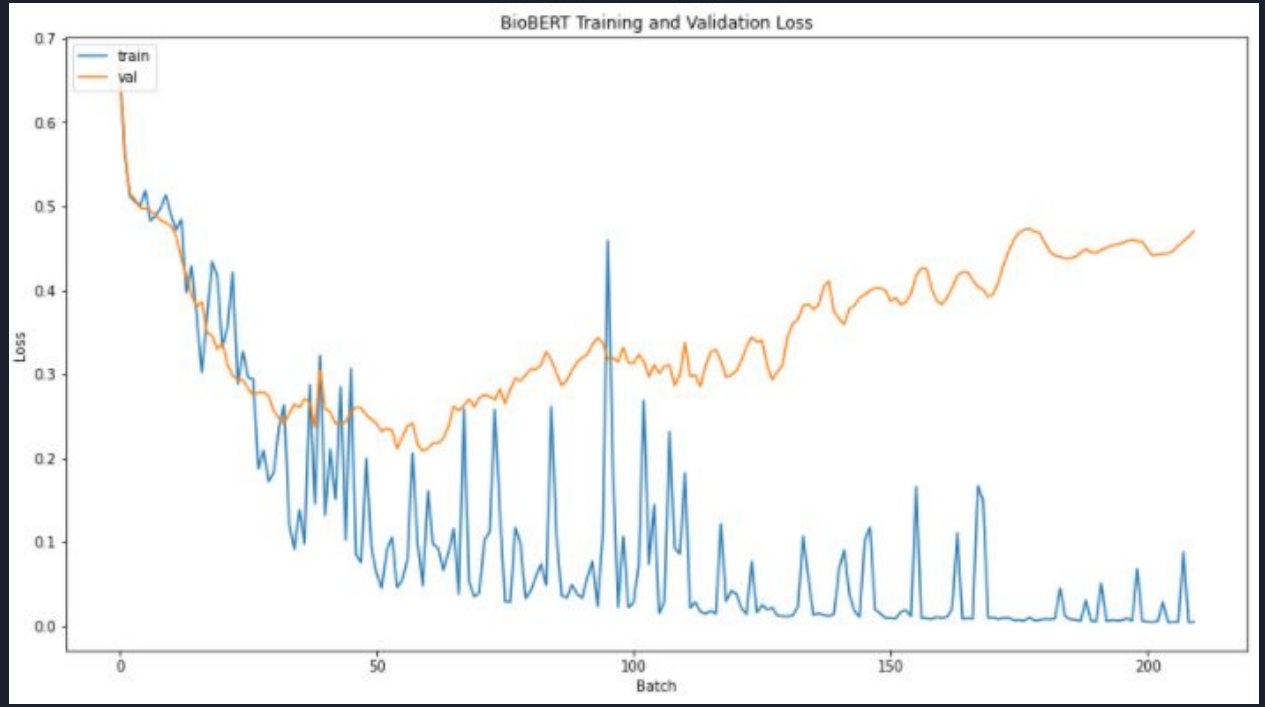


In 2 sibs, born of consanguineous Egyptian parents (family 6) with neurodevelopmental disorder with behavioral abnormalities, absent speech, and hypotonia (NEDBASH; {618718}), {3:Dias et al. (2019)} identified a homozygous c.446T-C transition ({VAR c.446T-C, NM_032536.3}) in exon 2 of the NTNG2 gene, resulting in a met149-to-thr (M149T) substitution at a conserved residue in the laminin-like domain. The mutation, which was found by exome sequencing and confirmed by Sanger sequencing, segregated with the disorder in the family. The variant was not found in the Exome Variant Server or gnomAD databases, or in an in-house database of about 10,000 control exomes. Molecular modeling predicted that the mutation may cause protein misfolding. Transfection of the mutation into HeLa cells resulted in almost no protein expression at the cell surface compared to wildtype.\n\n 8 patients from 4 consanguineous Arab Muslim families with neurodevelopmental disorder with behavioral abnormalities, absent speech, and hypotonia (NEDBASH; {618718}), {1:Abu-Libdeh et al. (2019)} identified a homozygous 1-bp duplication ({VAR c.376dupT, NM_032536.3}) in the NTNG2 gene, resulting in a frameshift and premature termination (Ser126PhefsTer241). The mutation, which was found by exome sequencing and confirmed by Sanger sequencing, segregated with the disorder in all families. Haplotype analysis indicated a founder effect. **The mutation was not found in the gnomAD database, but the heterozygous variant was found in 2 of about 3,500 individuals in an in-house database of mainly Arab Muslim individuals. Functional studies of the variant and studies of patient cells were not performed, but the variant was predicted to result in a loss of function.\n\n{5:Heimer et al. (2019)} identified homozygosity for the c.376dupT mutation in exon 3 of the NTNG2 gene in 3 patients from 2 consanguineous Arab Muslim families with NEDBASH. The mutation, which was found by whole-exome sequencing and confirmed by Sanger sequencing, segregated with the disorder in both families. Functional studies of the variant and studies of patient cells were not performed, but the mutation was predicted to result in nonsense-mediated mRNA decay and a loss of function.**\n\n 2 sibs, born of consanguineous Iranian parents (family 1) with neurodevelopmental disorder with behavioral abnormalities, absent speech, and hypotonia (NEDBASH; {618718}), {3:Dias et al. (2019)} identified a homozygous c.1367G-A transition ({VAR c.1367G-A, NM_032536.3}) in exon 7 of the NTNG2 gene, resulting in a cys456-to-tyr (C456Y) substitution at a conserved residue in the EGF4 domain. The mutation, which was found by exome sequencing, segregated with the disorder in the family. The variant was not found in the Exome Variant Server or gnomAD databases, or in an in-house database of 10,000 control exomes. Molecular modeling predicted that the mutation may disrupt a disulfide bridge and have a negative effect on protein stability. Transfection of the mutation into HeLa cells resulted in decreased protein expression at the cell surface compared to wildtype.\n\n 3 members of a consanguineous Iranian family (family 2) with neurodevelopmental disorder with behavioral abnormalities, absent speech, and hypotonia (NEDBASH; {618718}), {3:Dias et al. (2019)} identified a homozygous c.319T-G transversion ({VAR c.319T-G, NM_032536.3}) in exon 2 of the NTNG2 gene, resulting in a trp107-to-gly (W107G) substitution at a conserved residue in the laminin-like domain. The mutation, which was found by exome sequencing, segregated with the disorder in the family. The variant was not found in the Exome Variant Server or gnomAD databases, or in an in-house database of 10,000 control exomes. Molecular modeling predicted that the mutation may cause protein misfolding. Transfection of the mutation into HeLa cells resulted in almost no protein expression at the cell surface compared to wildtype.\n\n 2 sibs of Mexican descent (family 4) with neurodevelopmental disorder with behavioral abnormalities, absent speech, and hypotonia (NEDBASH; {618718}), {3:Dias et al. (2019)} identified a homozygous c.1065C-G transversion ({VAR c.1065C-G, NM_032536.3}) in exon 5 of the NTNG2 gene, resulting in a cys355-to-trp (C355W) substitution at a conserved residue in the EGF2 domain. The mutation, which was found by exome sequencing and confirmed by Sanger sequencing, segregated with the disorder in the family. The variant was not found in the Exome Variant Server or gnomAD databases, or in an in-house database of over 60,000 control exomes. Molecular modeling predicted that the mutation may disrupt a disulfide bridge and have a negative effect on protein stability. Transfection of the mutation into HeLa cells resulted in decreased protein expression at the cell surface compared to wildtype.\n\n 4 members of a highly consanguineous Turkish family (family 5) with neurodevelopmental disorder with behavioral abnormalities, absent speech, and hypotonia (NEDBASH; {618718}), {3:Dias et al. (2019)} identified a homozygous c.242G-A transition ({VAR c.242G-A, NM_032536.3}) in exon 2 of the NTNG2 gene, resulting in a cys81-to-tyr (C81Y) substitution at a conserved residue in the laminin-like domain. The mutation, which was found by exome sequencing and confirmed by Sanger sequencing, segregated with the disorder in the family. The variant was not found in the Exome Variant Server or gnomAD databases, or in an in-house database of over 13,000 control exomes. Molecular modeling predicted that the mutation may disrupt a disulfide bridge and have a negative effect on protein stability. Transfection of the mutation into HeLa cells resulted in decreased protein expression at the cell surface compared to wildtype.

Improved Results

BioBERT Fine-tuning Curve

- A pre-trained biomedical language representation model for biomedical text mining
- BioBERT achieves good test data performance after 3 Epochs of finetuning



Comparative Model Performance

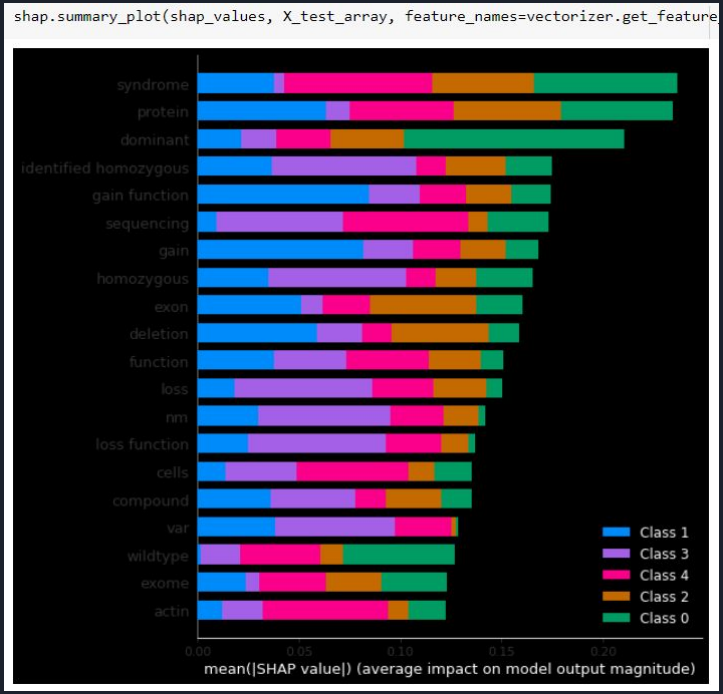
- Deep Learning achieves good performance on Test Data

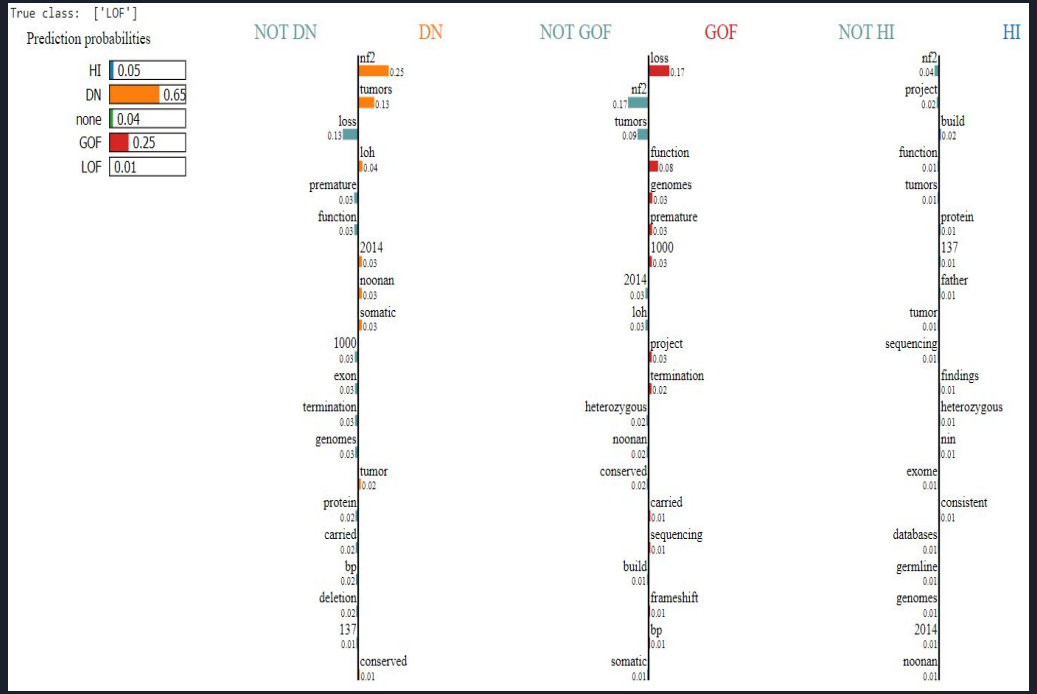
- Best Performance is observed from BioBERT

	Performance - F1 score					
Model & Dataset	**DN**	**GOF**	**HI**	**LOF**	**none**	**Macro**
TF-IDF + Logistic - Train	86.5%	95.2%	88.1%	90.2%	100.0%	92.0%
TF-IDF + Logistic - Test	64.8%	65.3%	53.8%	52.2%	73.5%	61.9%
DistilBERT - Train	89.5%	96.7%	88.2%	88.7%	99.3%	92.5%
DistilBERT - Test	77.4%	83.3%	71.0%	78.8%	66.7%	75.5%
BioBERT - Train	86.5%	93.9%	87.7%	85.5%	99.3%	90.6%
BioBERT - Test	81.3%	80.0%	75.9%	75.9%	88.0%	80.2%

Model Explainability

- Model Agnostic packages like Lime and Shap can help explain predictions





Future Work

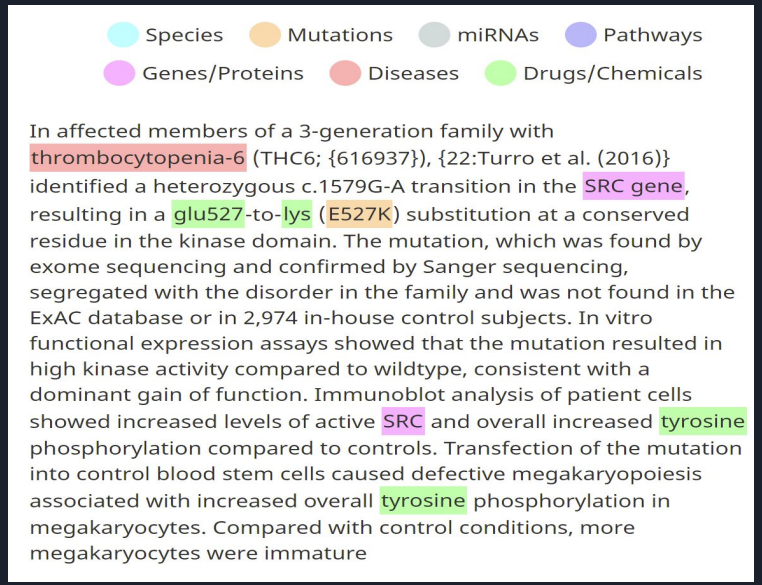
Next Steps

- Acquiring more data
- Trying different data cleaning techniques to increase the quality of the data
- Ground truth acquisition to reduce the bias in the data labels
- Using autoregressive language models like GPT-3

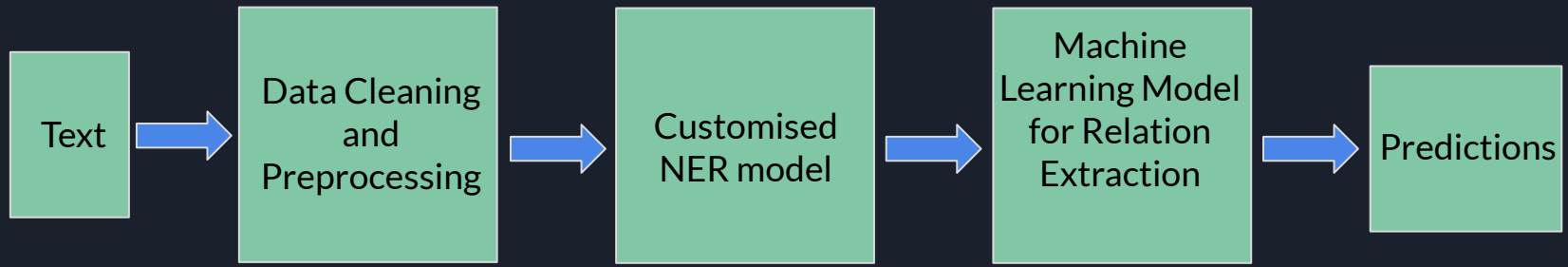
Named Entity Recognition

- Acts as building block in biomedical text mining
- To extract information of interactions and relations among biological entities
- BIO-NER system can be used in building text summarization system and question-answering system

Neural Biomedical Entity Recognition and multi-type Normalization (BERN)



Custom trainable relation extraction pipeline



Thank You!
&
Demo or Questions ?