Medical Text Classifier

Built Differently by: “The Natural Language Prodigies”

Table of contents

01

Business Problem

02

Approach

03

Modeling

04

Evaluation

05

Conclusions

Business Problem

- Physicians overworked and overwhelmed
- Lots of research to read
- Our model will point them to the ***most relevant*** research papers by field of medicine:
    - Cancer
    - Digestive
    - Neuro
    - Heart

26.7 hours

Of clinical work per day physicians would need to do to provide guideline-recommended care[1]

68%

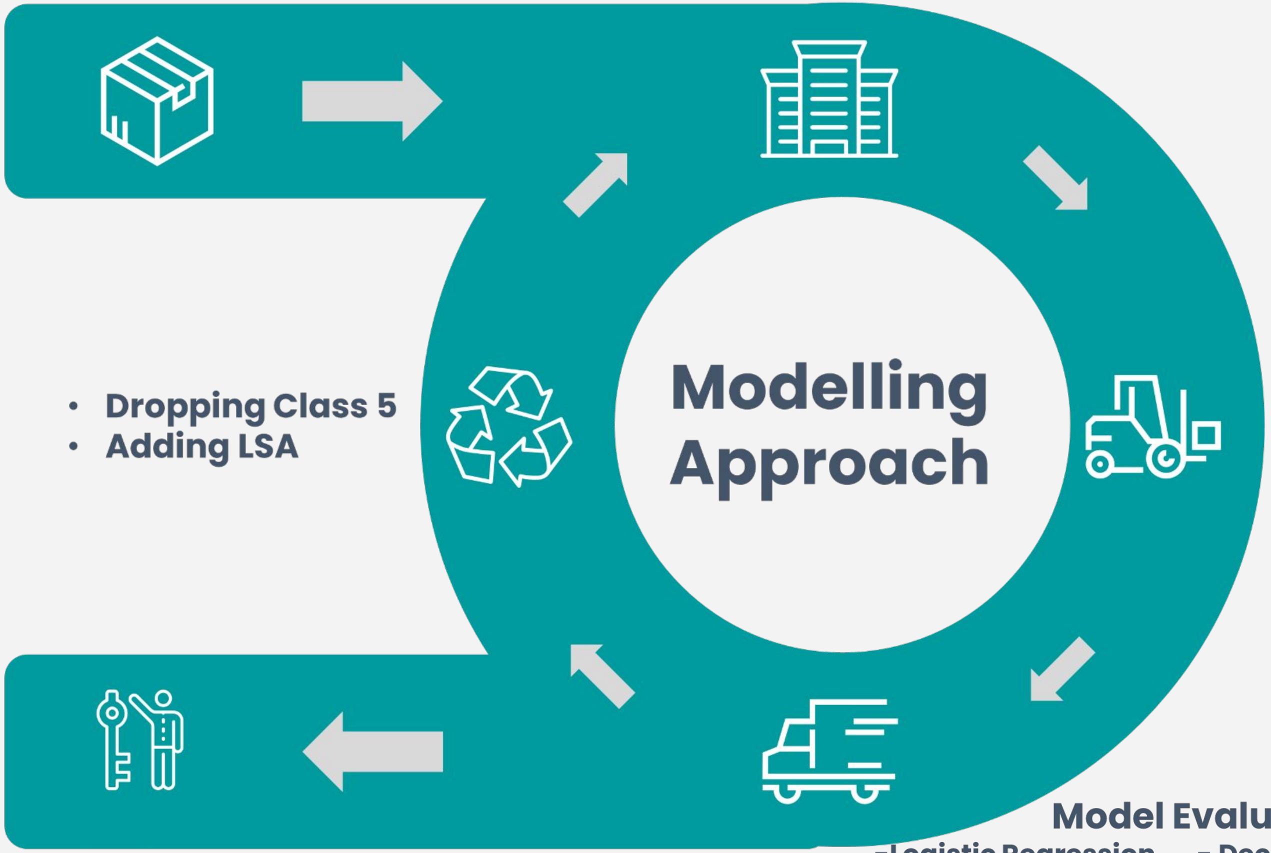
Of physicians feel overwhelmed by the amount of information they need to review to stay current[1]

3.5 years

The rate at which medical knowledge is doubling[1]

[1] *HealthcareFinanceNews.com (link)*

Modelling Approach



Text Cleaning Process

Set up Base Model

Run Machine learning Models

Model Evaluation

-Logistic Regression
-Gradient Boost
-Stacking Classifier
-Neural Networks
- Decision Tree
- Random Forest
- xgboost

Finalize the best Model

- Dropping Class 5
- Adding LSA

Cleaning Process

01 **Tokenize**

Define function to remove punctuation, change to lowercase and separate words

02 **Filter**

Use NLTK’s stopwords list to filter out superfluous words from our lists of tokens

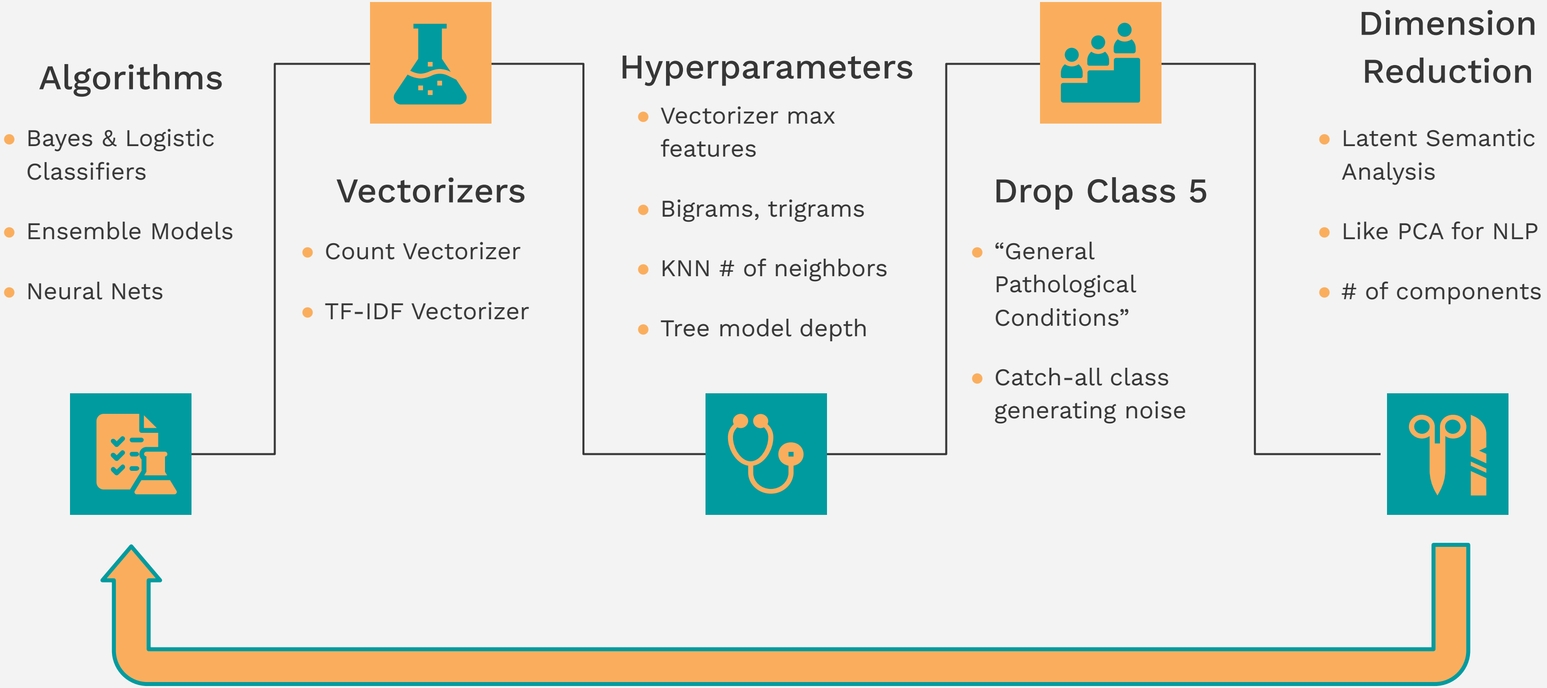
03 **Lemmatize**

Lemmatize rather than stem to prioritize accuracy and dataset is not huge

04 **Common Word Removal**

Identified frequent words common to all classes and removed them

Iterative Process



Algorithms

- Bayes & Logistic Classifiers
- Ensemble Models
- Neural Nets

Vectorizers

- Count Vectorizer
- TF-IDF Vectorizer

Hyperparameters

- Vectorizer max features
- Bigrams, trigrams
- KNN # of neighbors
- Tree model depth

Drop Class 5

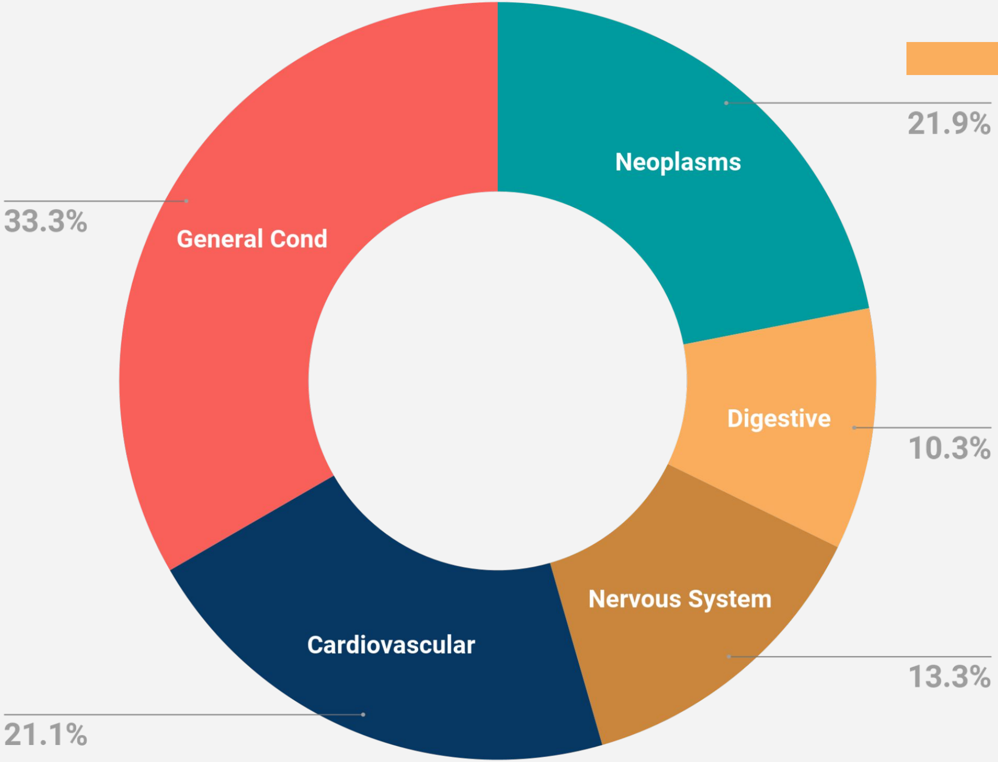
- “General Pathological Conditions”
- Catch-all class generating noise

Dimension Reduction

- Latent Semantic Analysis
- Like PCA for NLP
- # of components

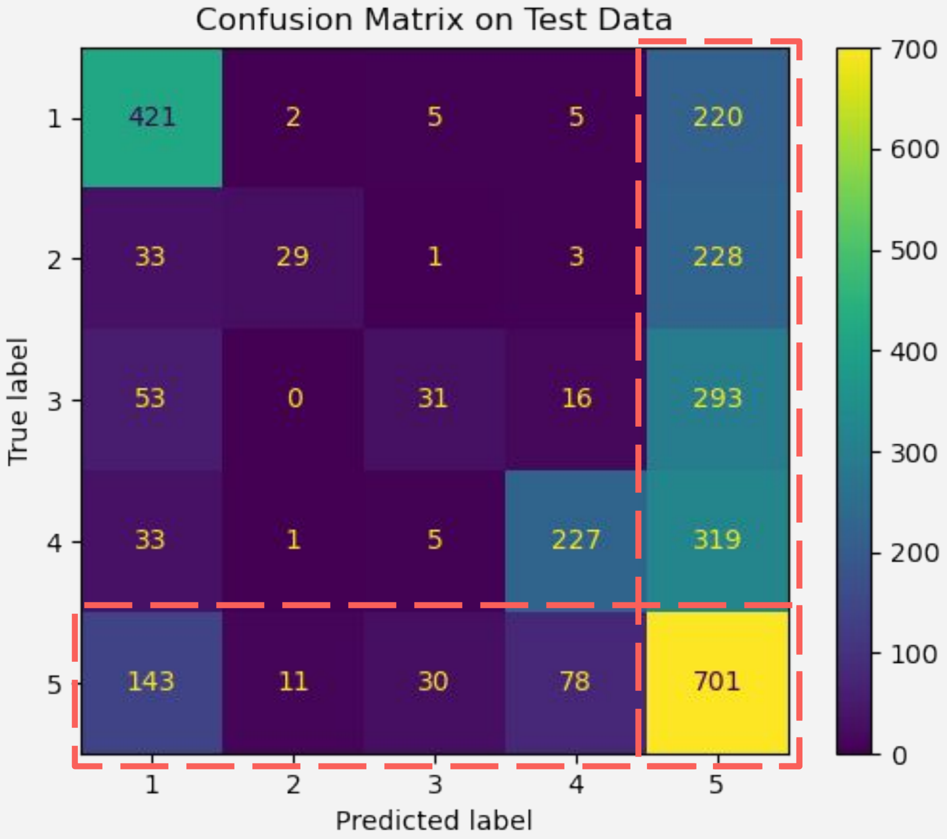
Class Distribution

Neoplasms	3,163
Digestive	1,494
Nervous System	1,925
Cardiovascular	3,051
General Conditions	4,805



Confusion Matrix for Five Class KNN

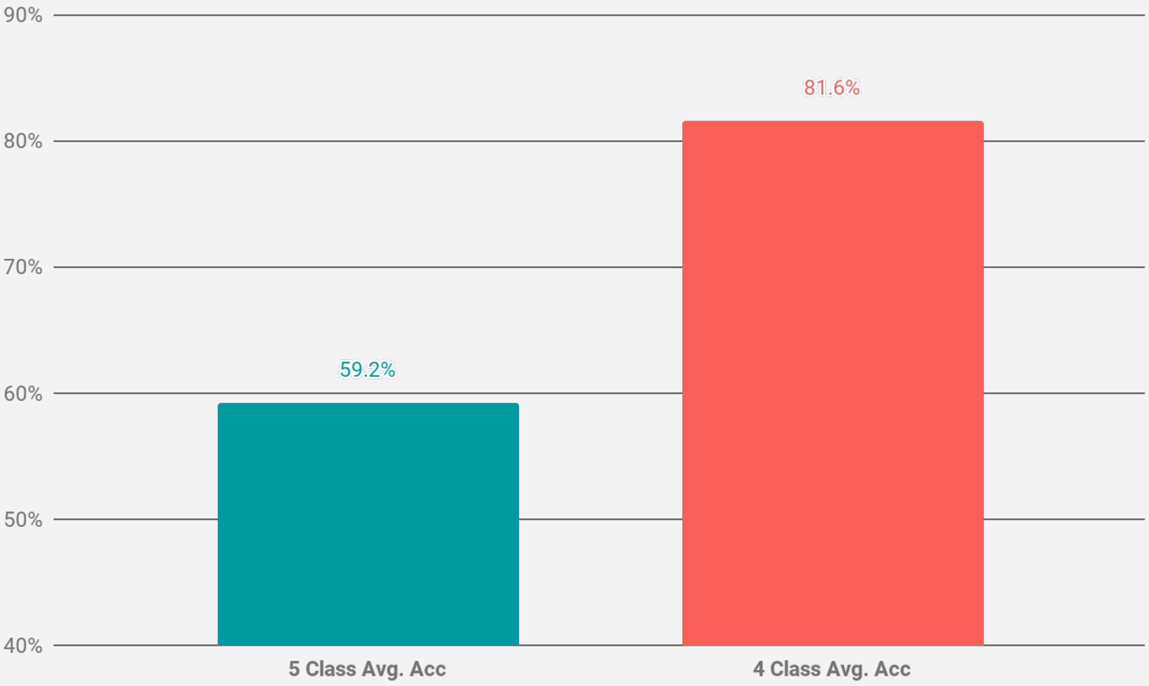
1. *Neoplasms (Cancer)*
2. *Digestive diseases*
3. *Nervous system diseases*
4. *Cardiovascular diseases*
5. ***General pathological conditions***



Confusion Matrix on Test Data

True label \ Predicted label	1	2	3	4	5
1	421	2	5	5	220
2	33	29	1	3	228
3	53	0	31	16	293
4	33	1	5	227	319
5	143	11	30	78	701

5 Class vs. 4 Class



22.4%

Increase in Accuracy

Hyperparameter Tuning

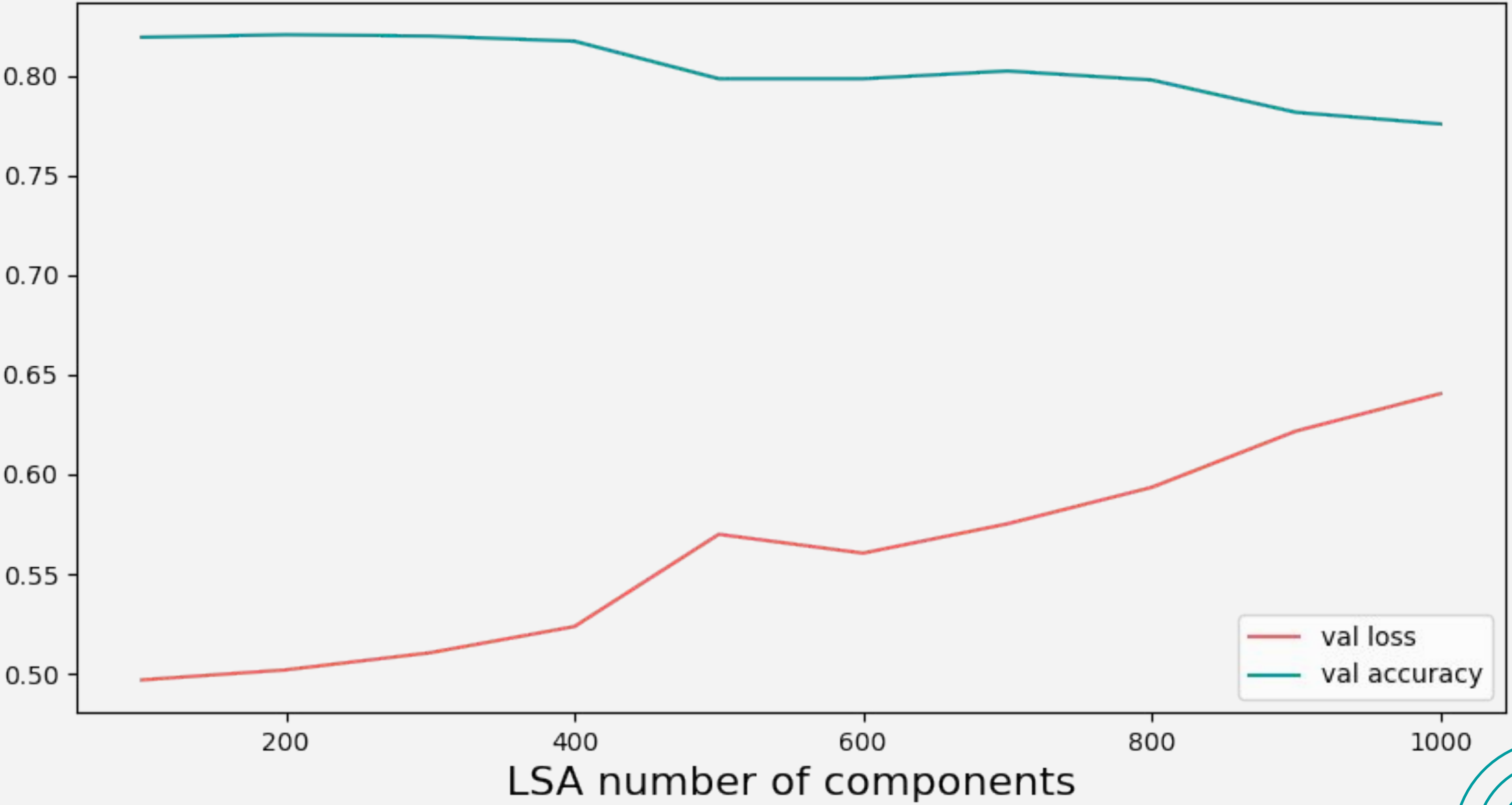
Vectorization

- Count
- TF-IDF

Ngram range

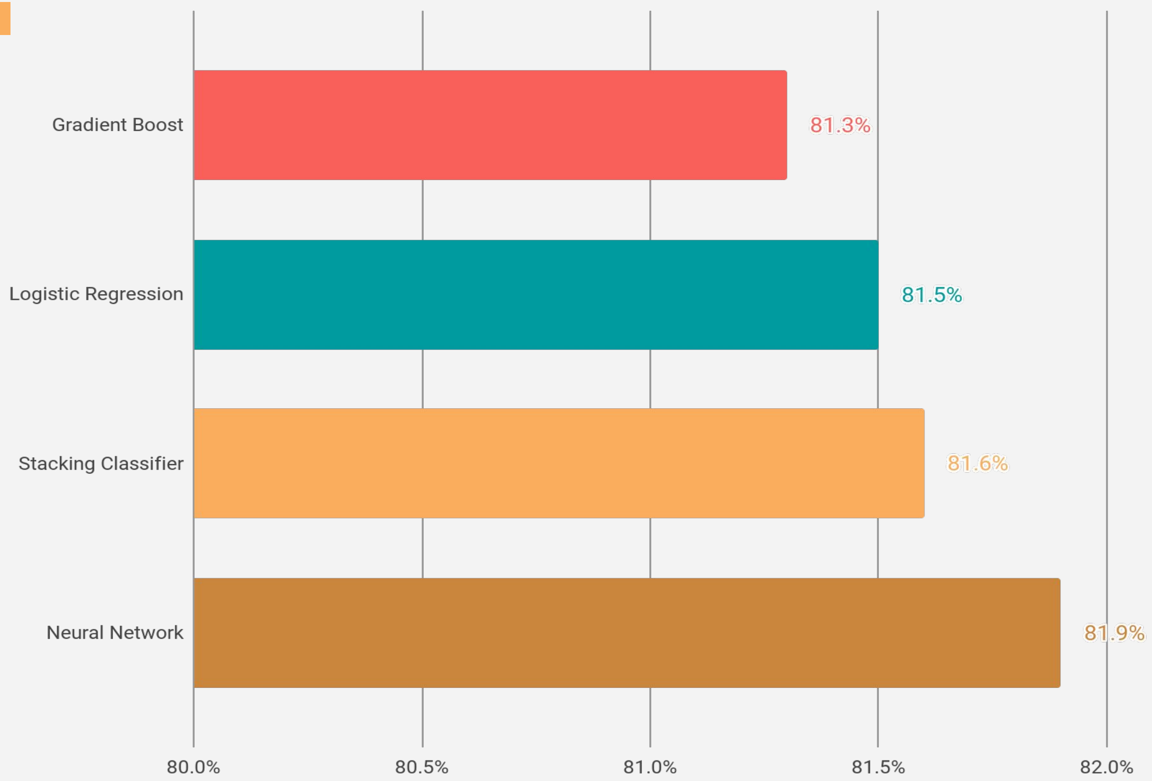
- Unigrams
- Bigrams
- Trigrams

Latent Semantic Analysis[1]



[1] *Neural Network performance with varying degrees of LSA components*

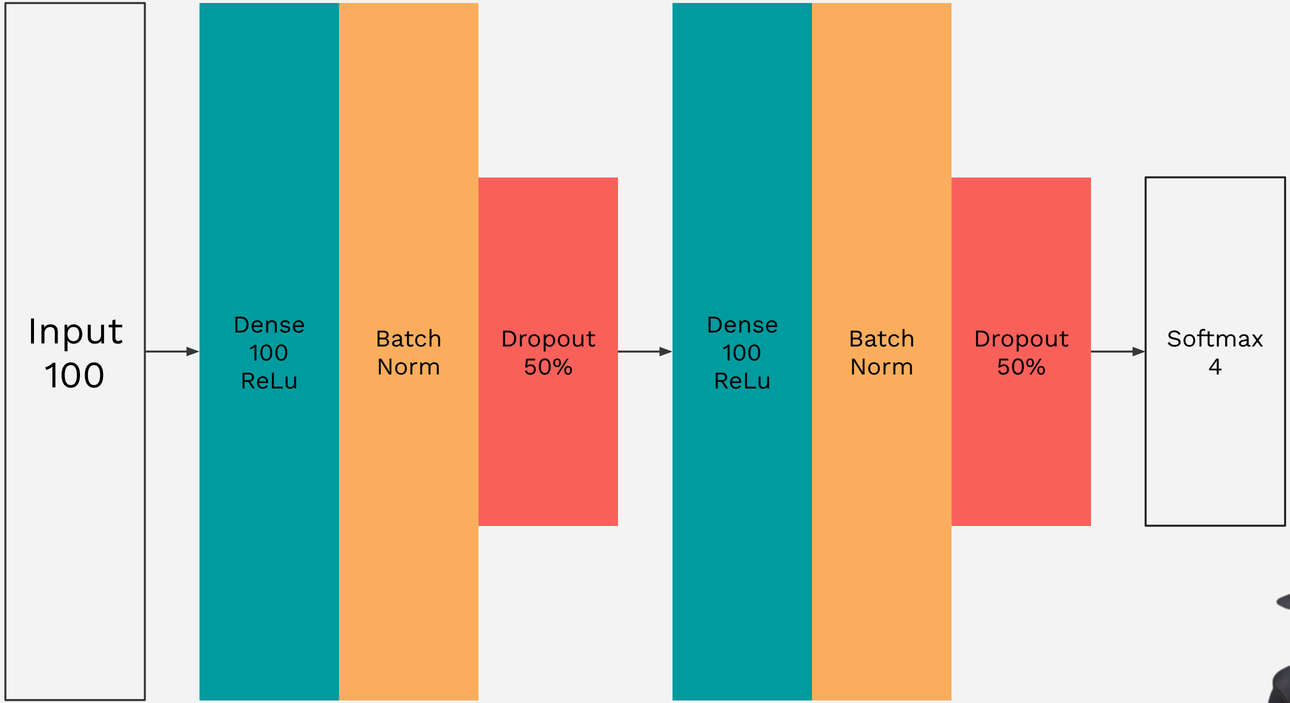
Top 4 Models



81.9%

Neural Network Accuracy

The Disease Deducer



Input 100

Dense 100 ReLu

Batch Norm

Dropout 50%

Dense 100 ReLu

Batch Norm

Dropout 50%

Softmax 4

Params:

- TF-IDF
- LSA = 100
- Adam optimizer
- Batch size = 32
- Early stopping with patience of 5

Conclusion

- Neural Nets are good
- We need more data
- We're ready to sell our model to Google or whatever

Next Steps

- Other possible applications
- Training on more data
- More classes
- Improve accuracy using Word2Vec
- Sequential modeling

Thanks! Questions?

Tristan Trechsel
tristantrechsel@gmail.com
Github: @ttrechsel
LinkedIn: /in/trechsel

Nick Kai
nhknicholas@gmail.com
Github: @nihkai
LinkedIn: /in/Nihkai

Heath Jones
jimmyhj9@gmail.com
Github: @heefjones
LinkedIn: /in/heefjones

Yasitha De Alwis
ydealwis@gmail.com
Github: @yasiSriLanka
LinkedIn: /in/yasitha-de-alwis