Multilabel Classification of Research Articles

Teo Zhan Rui

29th April 2021

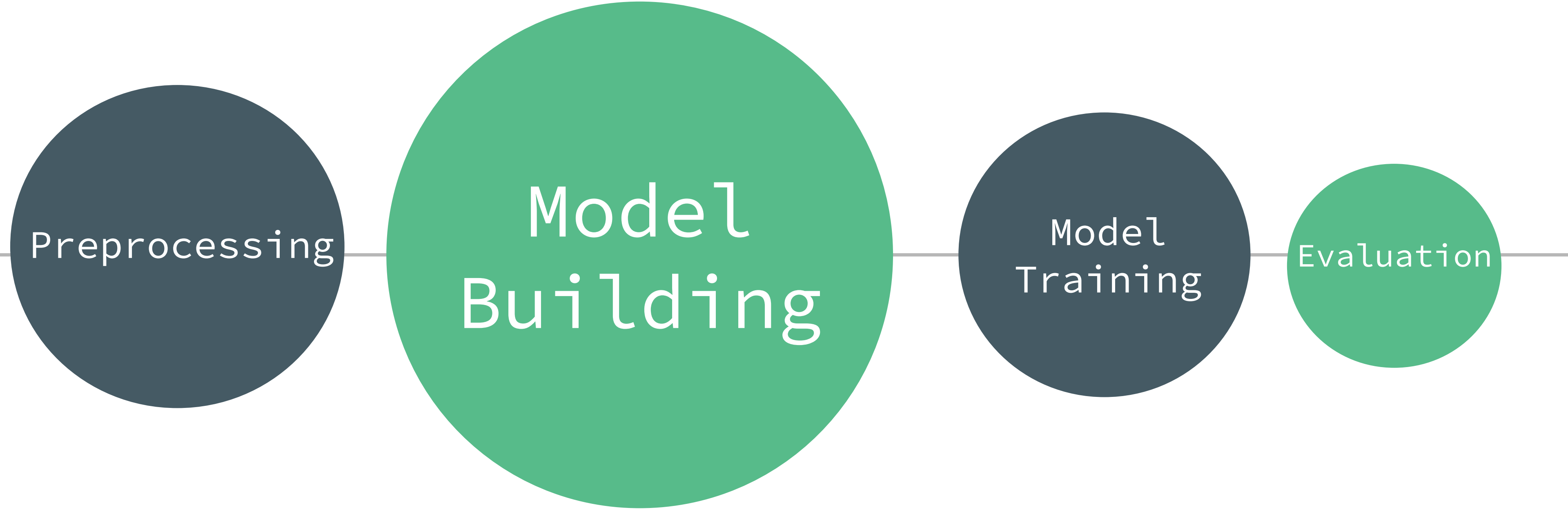
Problem Statement

1) What are the challenges of multilabel classification?

2) What are the suitable metrics for measuring performance?

3) Which model produced the best scores and are suitable for deployment?

Workflow



Data

number of articles

20792

maximum length of text

492

total number of labels

6

What is Multilabel Classification?

	Com Sci	Physics	Math	Statistics	Quant. Biology	Quant. Finance
article 1	1	0	1	1	0	0
article 2	0	1	1	0	0	0
article 3	0	0	0	0	1	0

Which metrics to use?

$$\textbf{micro F1} = \frac{2 \text{ x precision x recall}}{\text{precision + recall}}$$

$$\left.\begin{array}{r} \text{precision} = \dfrac{TP}{TP + FP} \\[2em] \text{recall} = \dfrac{TP}{TP + FN} \end{array}\right\}$$

Sum up TP, FP and FN
for each class.

Which metrics to use?

	Com Sci	Physics	Math
ground truth	**1**	**0**	**1**
prediction	**0**	**1**	**1**

hamming loss = 2/3

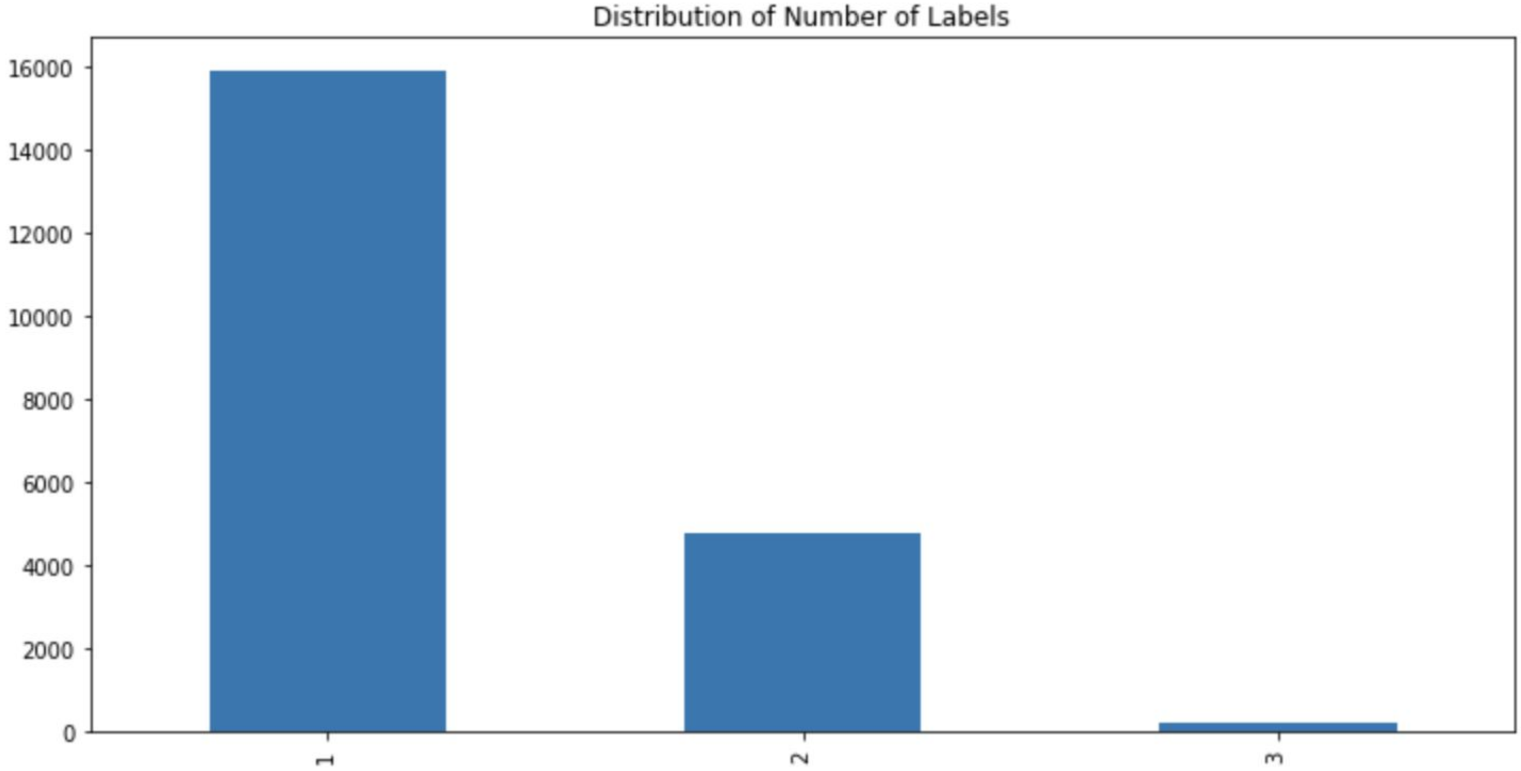
all labels correct :

hamming loss = 0/3

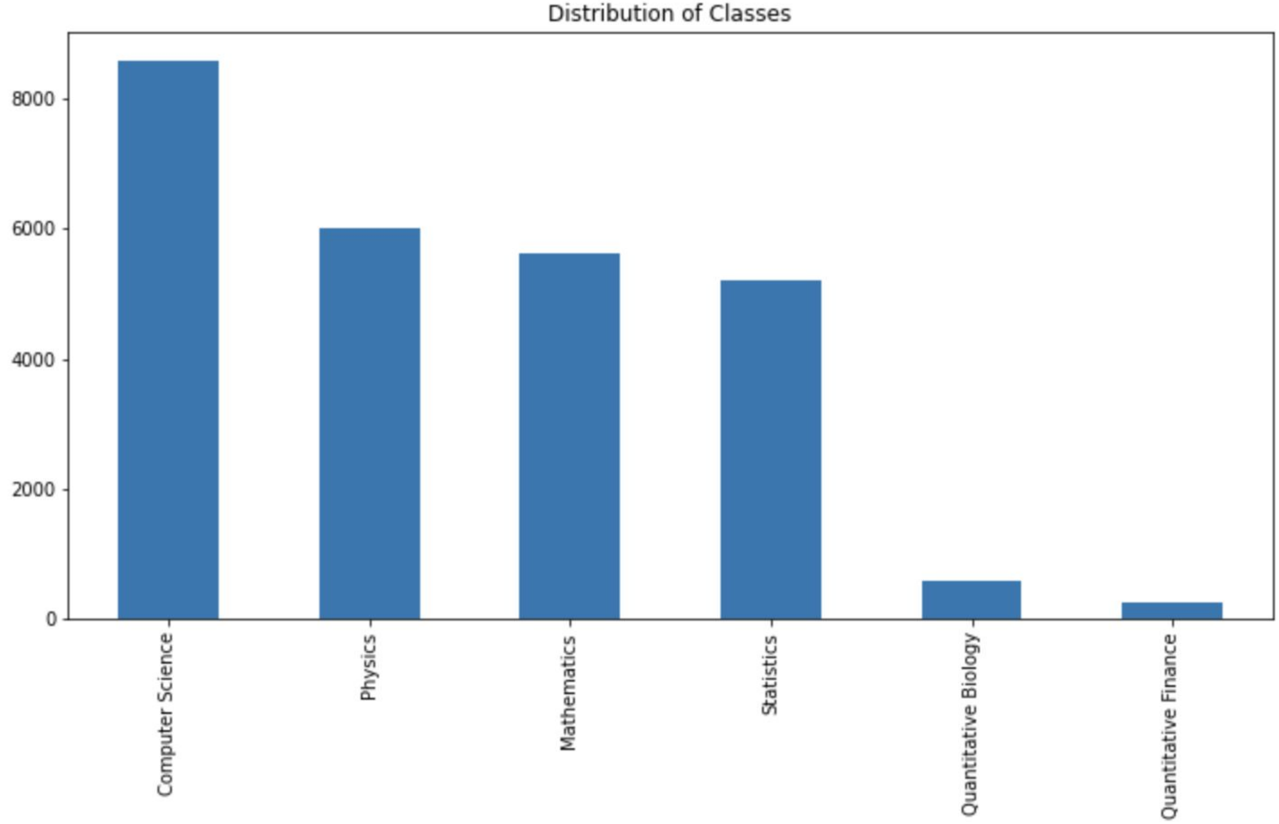
all labels wrong :

hamming loss = 3/3

How many labels do the articles have?

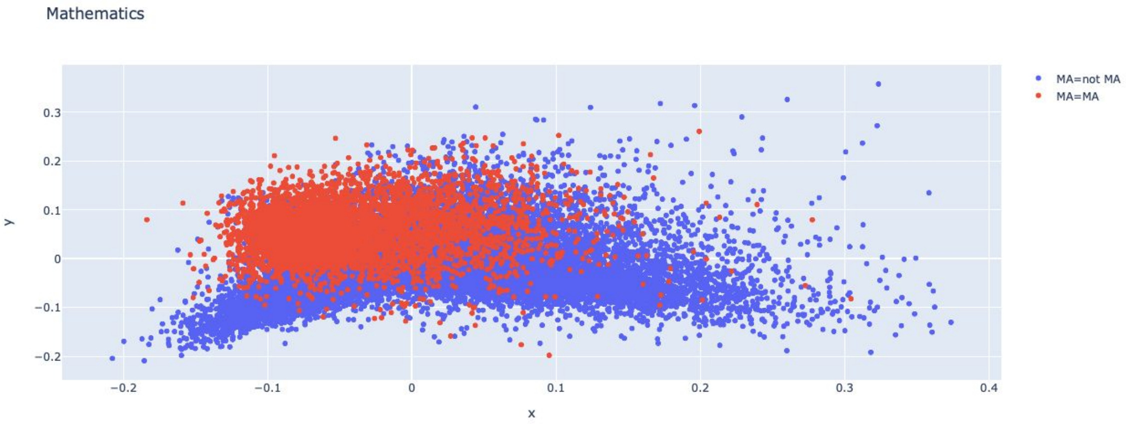


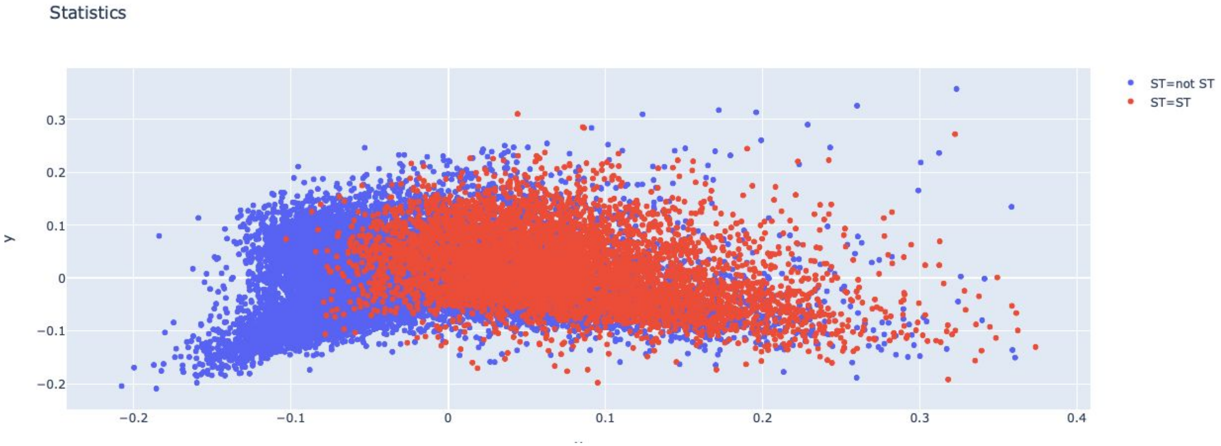
How many observations for each class?



Visualizing in 2D

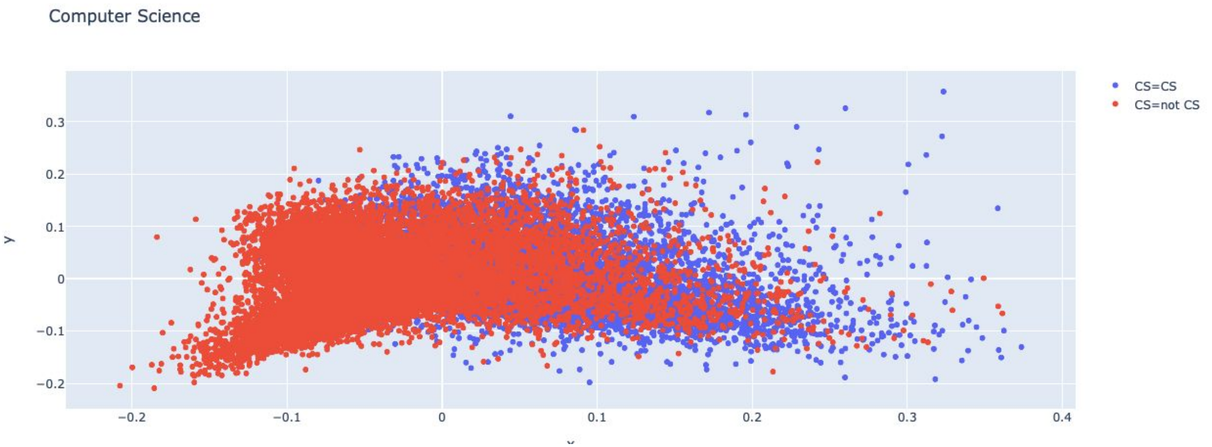
Transform TFIDF into 2 dimension vectors using PCA

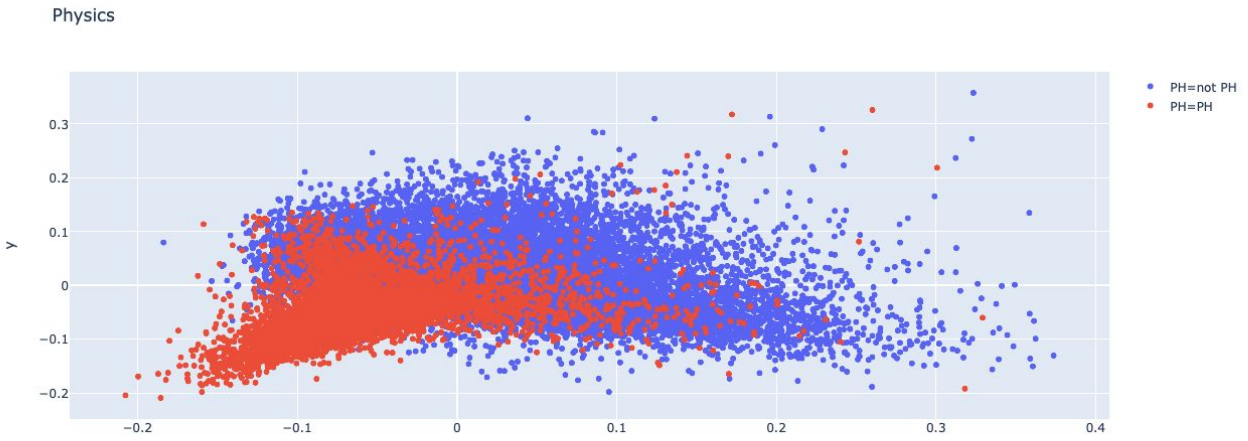




Visualizing in 2D

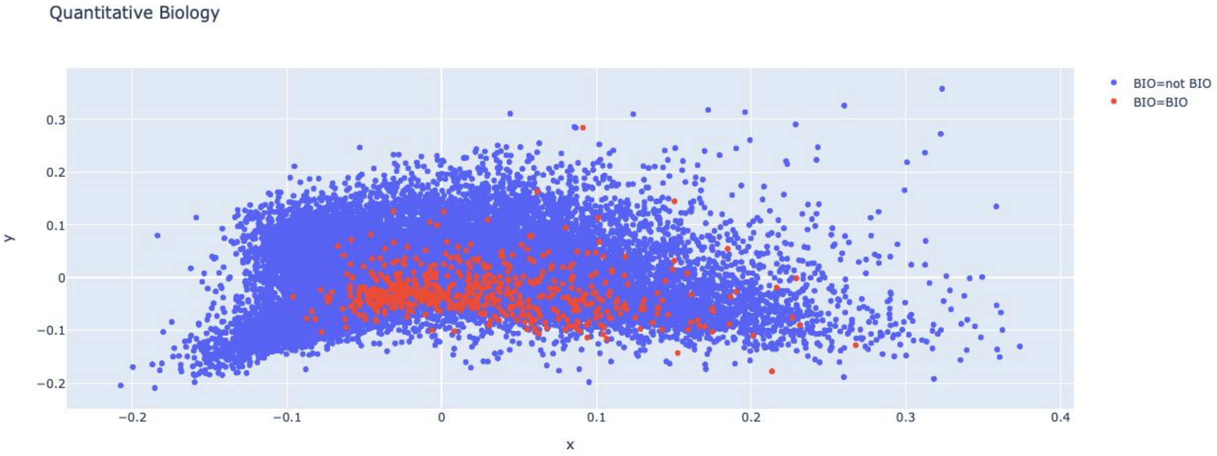
Transform TFIDF into 2 dimension vectors using PCA

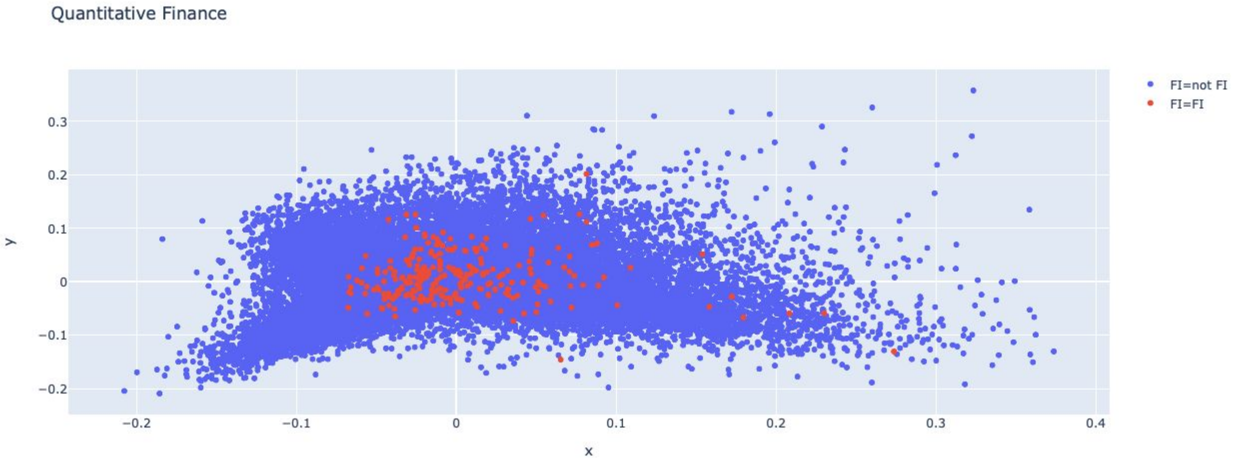

Computer Science


Physics

Visualizing in 2D

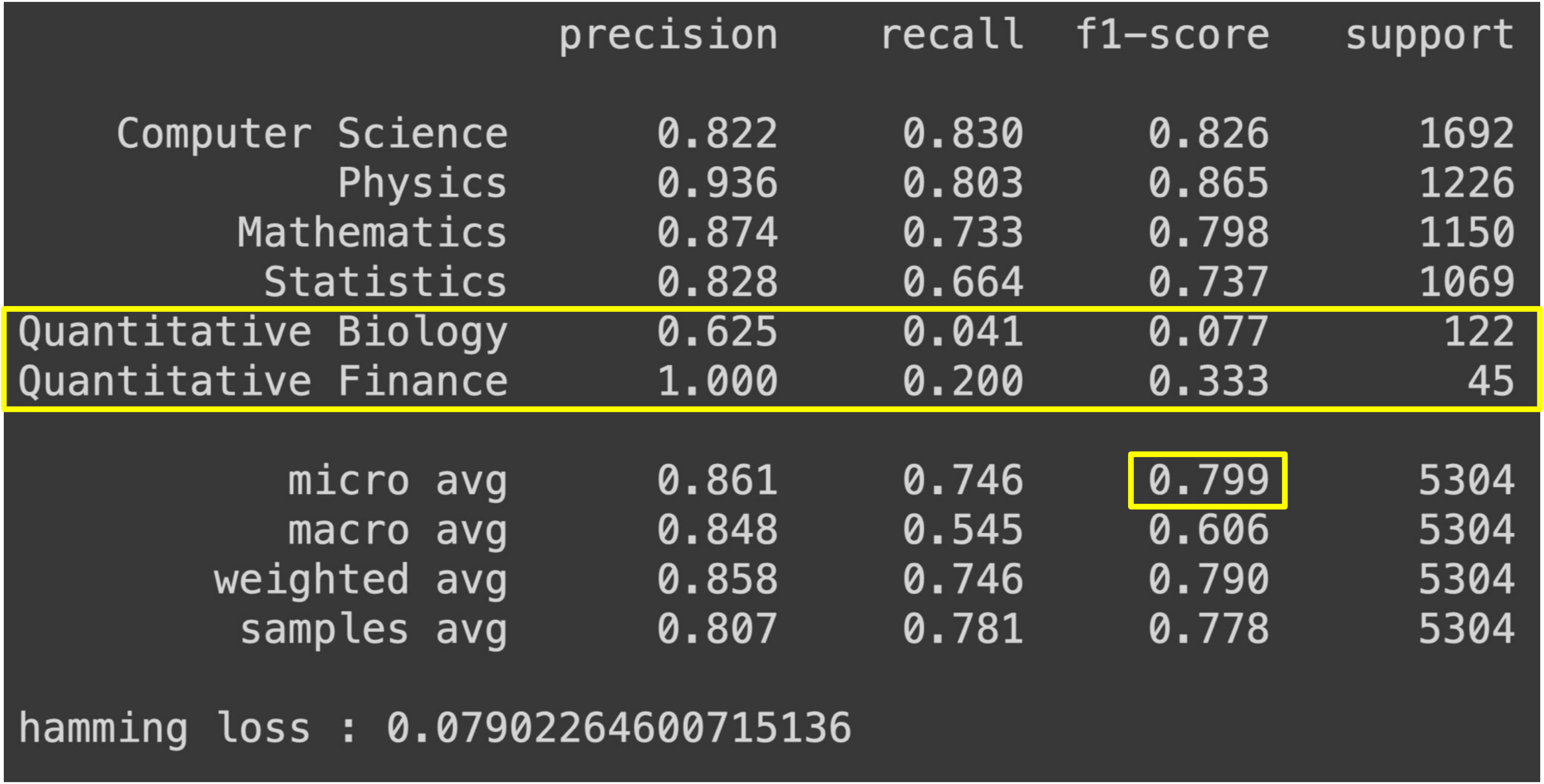
Transform TFIDF into 2 dimension vectors using PCA





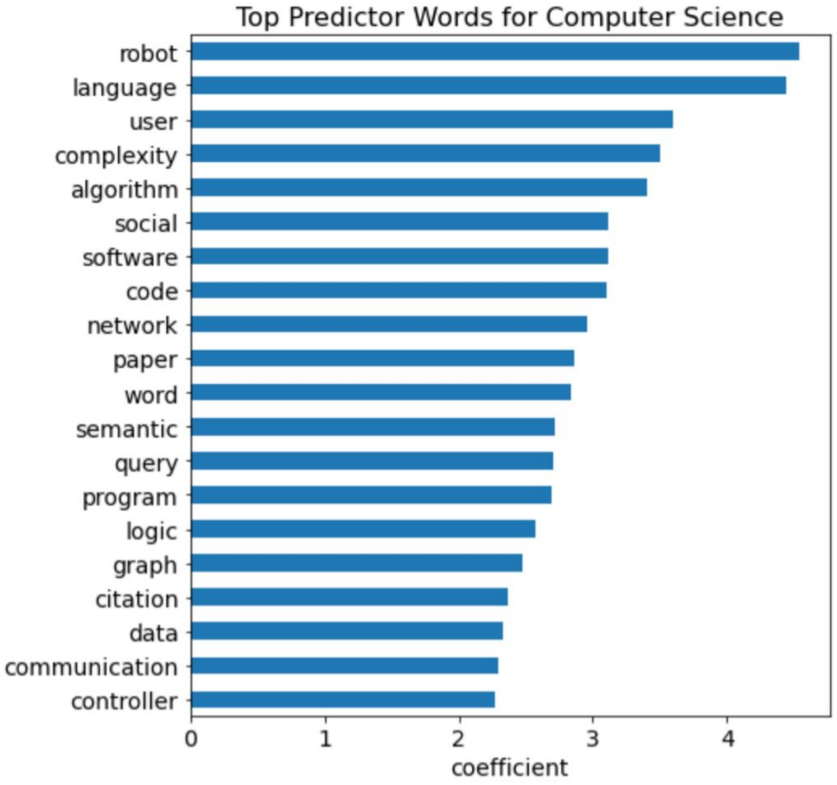
Sklearn

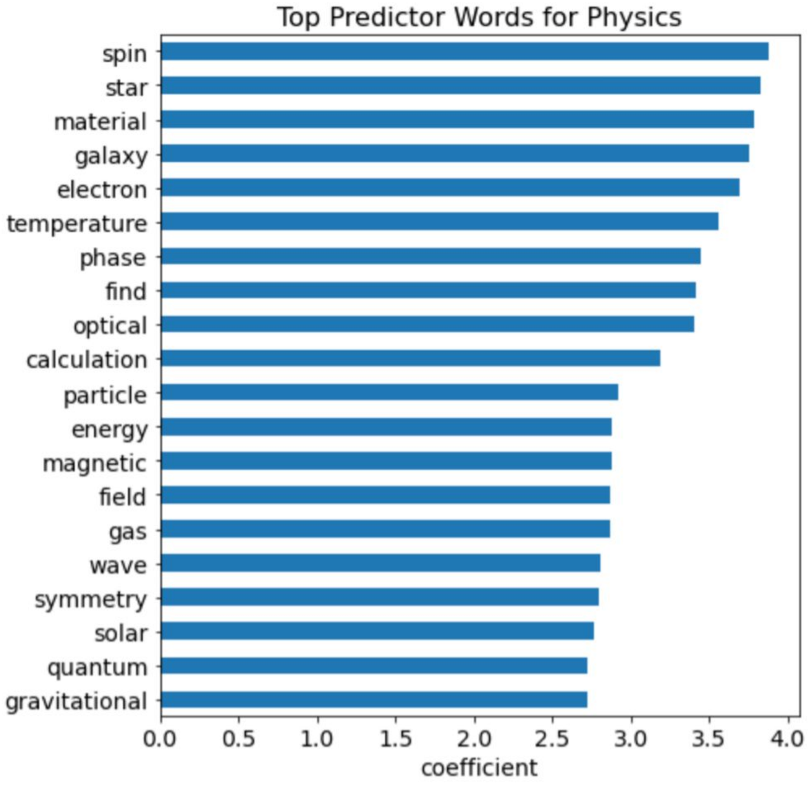
Logistic Regression - Benchmark



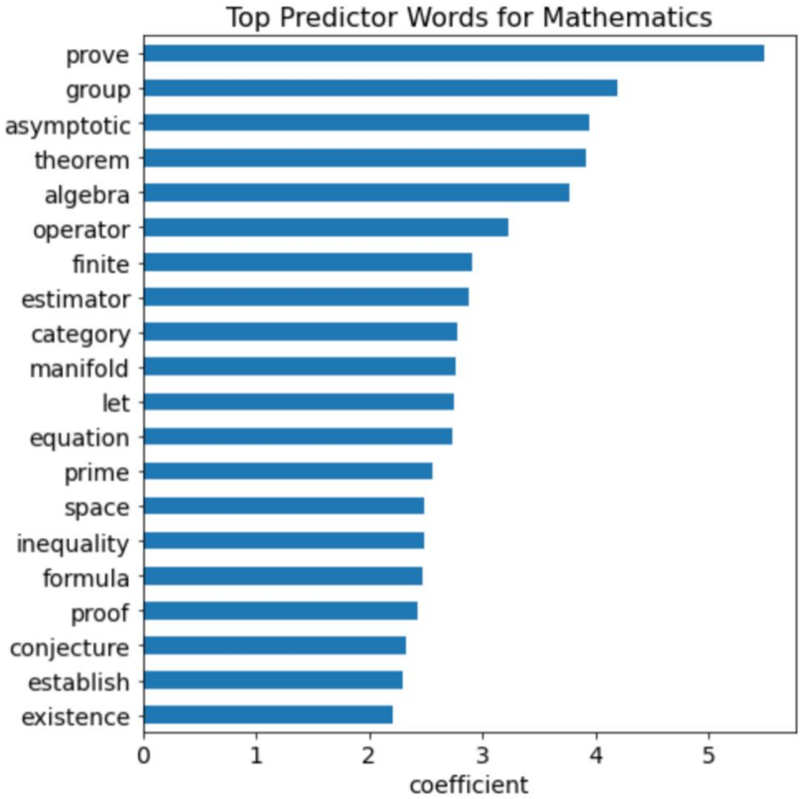
```
                        precision    recall  f1-score   support

      Computer Science      0.822     0.830     0.826      1692
               Physics      0.936     0.803     0.865      1226
           Mathematics      0.874     0.733     0.798      1150
            Statistics      0.828     0.664     0.737      1069
  Quantitative Biology      0.625     0.041     0.077       122
  Quantitative Finance      1.000     0.200     0.333        45

             micro avg      0.861     0.746     0.799      5304
             macro avg      0.848     0.545     0.606      5304
          weighted avg      0.858     0.746     0.790      5304
           samples avg      0.807     0.781     0.778      5304

hamming loss : 0.07902264600715136
```

Logistic Regression

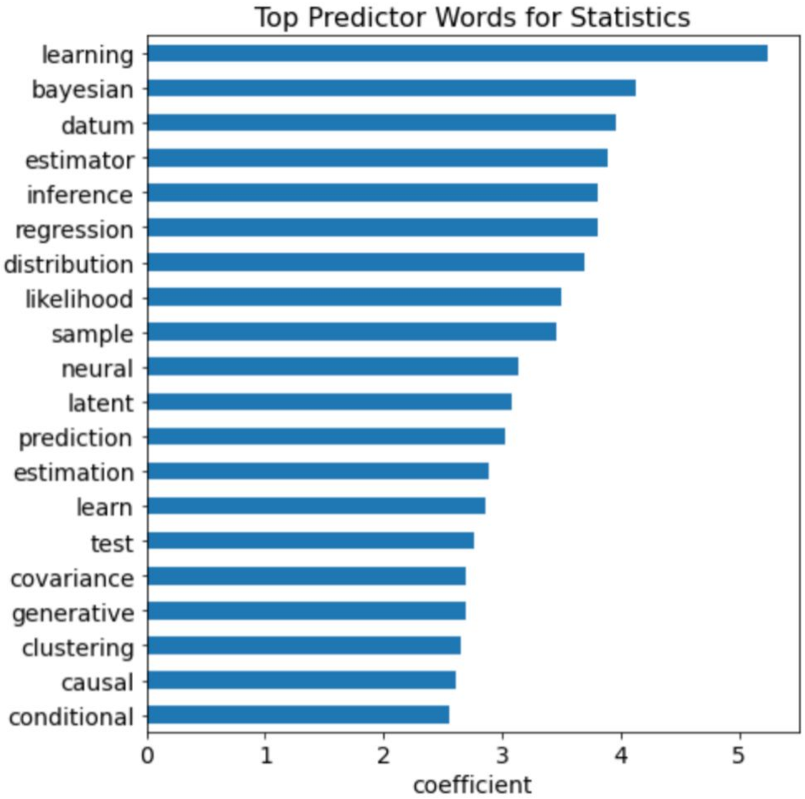




Logistic Regression

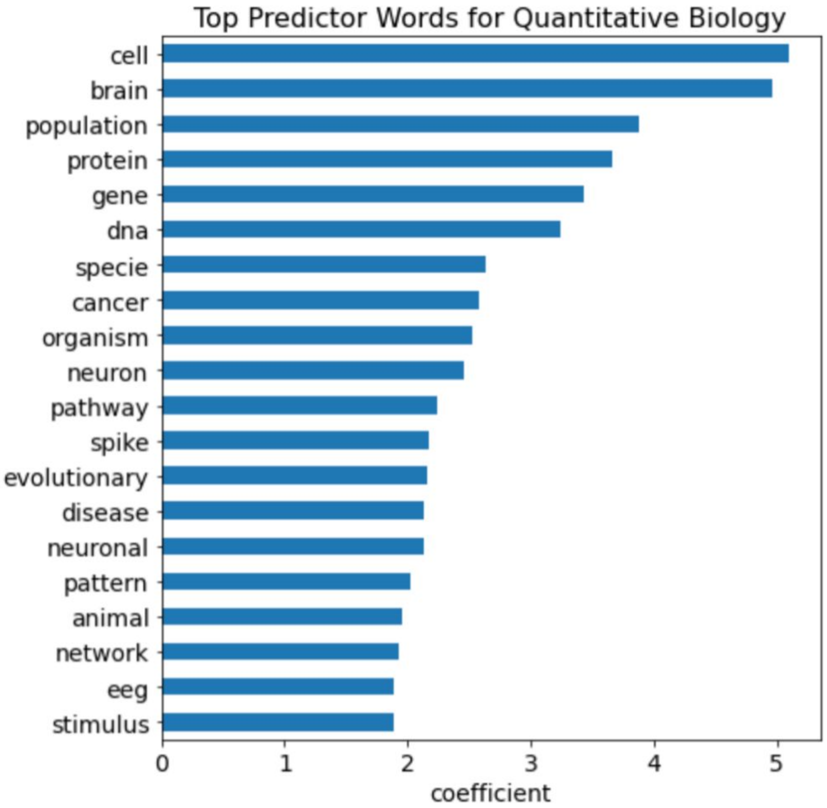


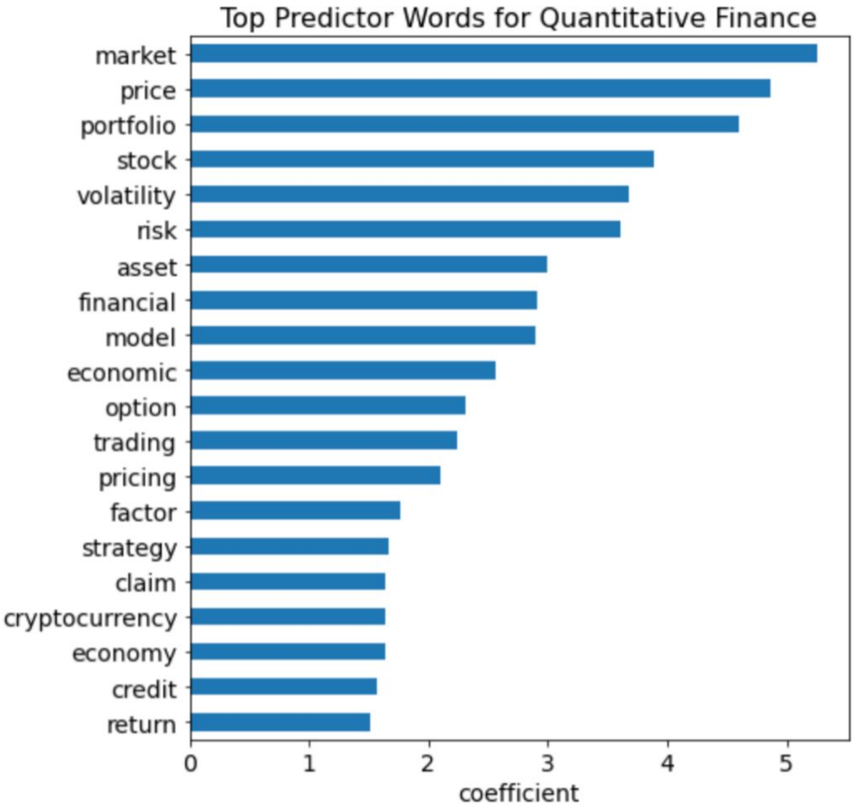
Top Predictor Words for Mathematics



Top Predictor Words for Statistics

Logistic Regression


Top Predictor Words for Quantitative Biology

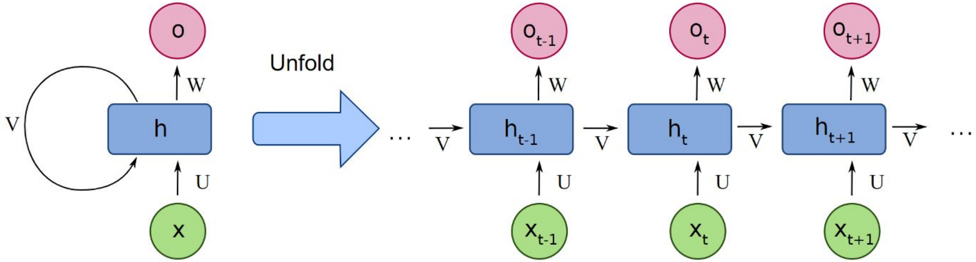

Top Predictor Words for Quantitative Finance

Are sklearn models suitable?

1) does not support multilabel
2) One vs Rest classifier - 6 different models
3) SVM - 2 minutes to complete prediction on 4000 validation articles (X_test)

	Com Sci	Physics	Math	Statistics	Quant. Biology	Quant. Finance
article 1	1	0	1	1	0	0
article 2	0	1	1	0	0	0
article 3	0	0	0	0	1	0

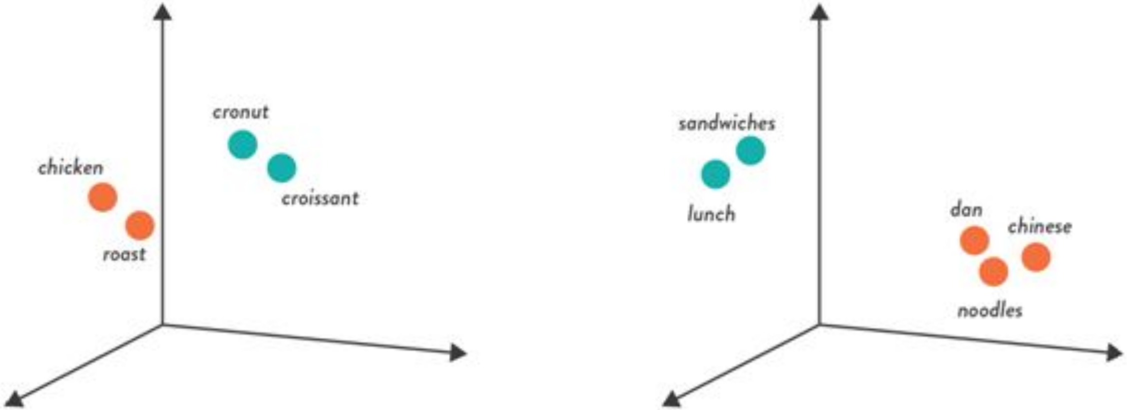
Recurrent Neural Networks



Source:
https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010

Word2Vec and LexVec

- representation of words in 300 dimension vectors
- compare similar words
- self training vs pre-trained



Source: https://www.gsitechnology.com/Leveraging-Word2vec-for-More-than-Text

Word2Vec and LexVec

‘bayes’ :

CBOW	Skipgram	LexVec
dirichlet distribution	gp sum	bayesian
selection procedure	sure convergence	regression
posterior sampling	hellinger distance	inference
sure convergence	entropy sgd	probabilistic
frequentist	likelihood bootstrap	probability

Word2Vec and LexVec

‘cat’ :

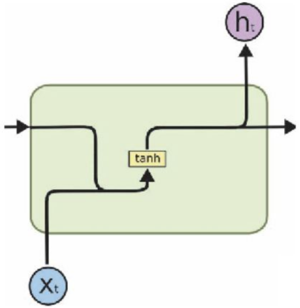
CBOW	Skipgram	LexVec
dog	defence	dog
traffic sign	debugger	cats
robot assist	curvature bound	feline
large annotate	dog	puppy
handwritten character	fully autonomous	kitten

My Neural Networks - Multilabel

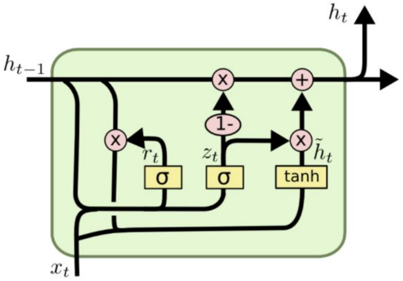
- output layer : 6
- activation : sigmoid
- optimizer : adam
- loss : binary crossentropy

	Com Sci	Physics	Math	Statistics	Quant. Biology	Quant. Finance
article 1	1	0	1	1	0	0
article 2	0	1	1	0	0	0
article 3	0	0	0	0	1	0

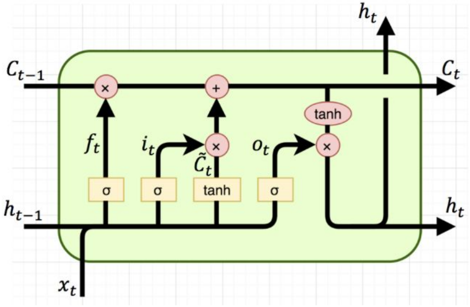
RNN



GRU



LSTM



- increasing complexity
- increasing train time
- slower convergence

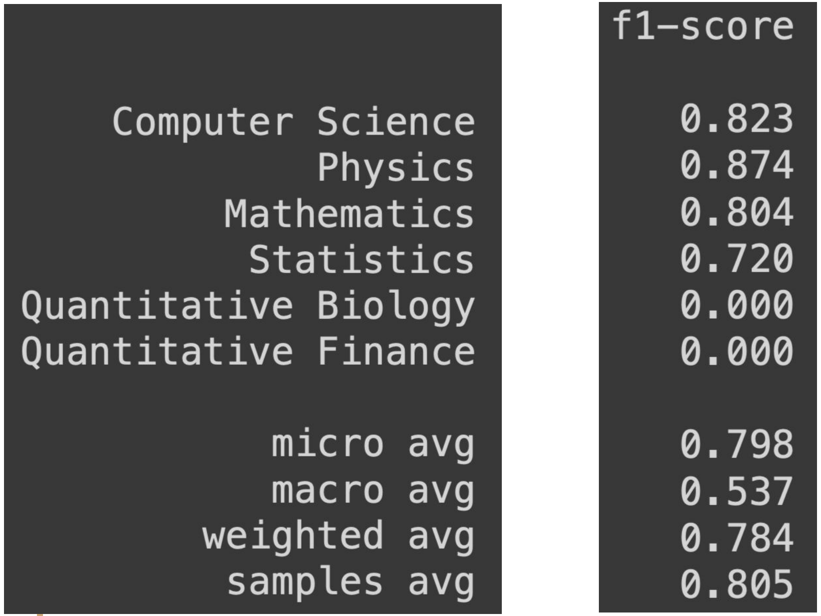
Source:
https://medium.com/@saurabh.rathor092/simple-rnn-vs-gru-vs-lstm-difference-lies-in-more-flexible-control-5f33e07b1e57

Simple RNN

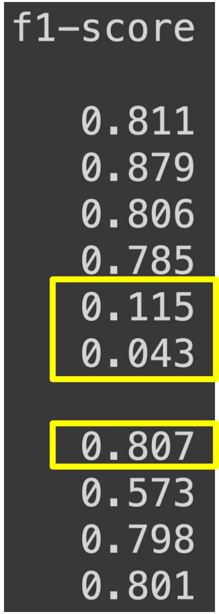
no pretrained word vectors

GRU + LSTM

LexVec



```
                              f1-score

      Computer Science           0.823
               Physics           0.874
           Mathematics           0.804
            Statistics           0.720
  Quantitative Biology           0.000
  Quantitative Finance           0.000

             micro avg           0.798
             macro avg           0.537
          weighted avg           0.784
           samples avg           0.805
```

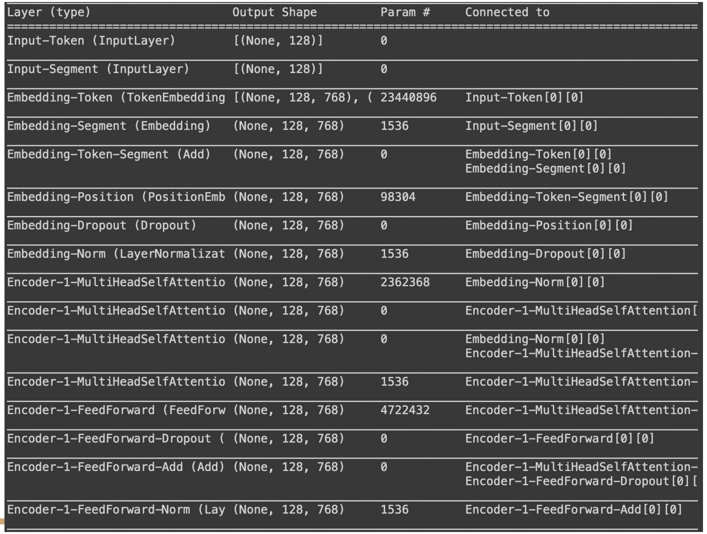


```
f1-score

   0.811
   0.879
   0.806
   0.785
   0.115
   0.043

   0.807
   0.573
   0.798
   0.801
```

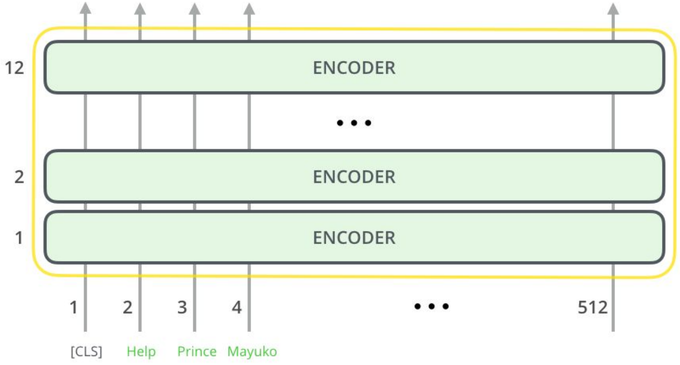
Bidirectional Encoder Representations from Transformers

BERT

1) 12 layers of encoder and decoder
2) attention mechanism



```
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
Input-Token (InputLayer)        [(None, 128)]        0
__________________________________________________________________________________________________
Input-Segment (InputLayer)      [(None, 128)]        0
__________________________________________________________________________________________________
Embedding-Token (TokenEmbedding [(None, 128, 768), ( 23440896    Input-Token[0][0]
__________________________________________________________________________________________________
Embedding-Segment (Embedding)   (None, 128, 768)     1536        Input-Segment[0][0]
__________________________________________________________________________________________________
Embedding-Token-Segment (Add)   (None, 128, 768)     0           Embedding-Token[0][0]
                                                                 Embedding-Segment[0][0]
__________________________________________________________________________________________________
Embedding-Position (PositionEmb (None, 128, 768)     98304       Embedding-Token-Segment[0][0]
__________________________________________________________________________________________________
Embedding-Dropout (Dropout)     (None, 128, 768)     0           Embedding-Position[0][0]
__________________________________________________________________________________________________
Embedding-Norm (LayerNormalizat (None, 128, 768)     1536        Embedding-Dropout[0][0]
__________________________________________________________________________________________________
Encoder-1-MultiHeadSelfAttentio (None, 128, 768)     2362368     Embedding-Norm[0][0]
__________________________________________________________________________________________________
Encoder-1-MultiHeadSelfAttentio (None, 128, 768)     0           Encoder-1-MultiHeadSelfAttention[
__________________________________________________________________________________________________
Encoder-1-MultiHeadSelfAttentio (None, 128, 768)     0           Embedding-Norm[0][0]
                                                                 Encoder-1-MultiHeadSelfAttention-
__________________________________________________________________________________________________
Encoder-1-MultiHeadSelfAttentio (None, 128, 768)     1536        Encoder-1-MultiHeadSelfAttention-
__________________________________________________________________________________________________
Encoder-1-FeedForward (FeedForw (None, 128, 768)     4722432     Encoder-1-MultiHeadSelfAttention-
__________________________________________________________________________________________________
Encoder-1-FeedForward-Dropout ( (None, 128, 768)     0           Encoder-1-FeedForward[0][0]
__________________________________________________________________________________________________
Encoder-1-FeedForward-Add (Add) (None, 128, 768)     0           Encoder-1-MultiHeadSelfAttention-
                                                                 Encoder-1-FeedForward-Dropout[0][
__________________________________________________________________________________________________
Encoder-1-FeedForward-Norm (Lay (None, 128, 768)     1536        Encoder-1-FeedForward-Add[0][0]
__________________________________________________________________________________________________
```

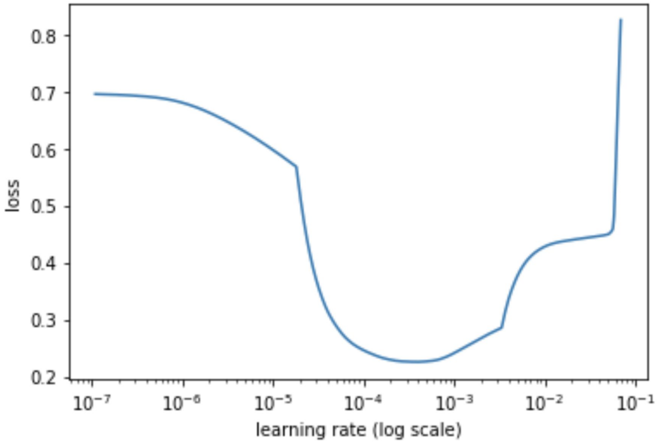


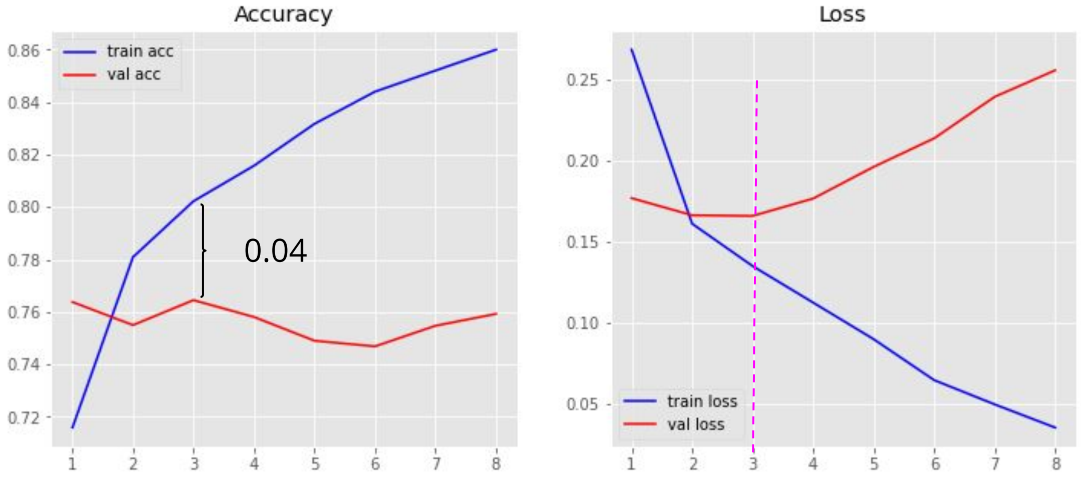
BERT

Size Matters

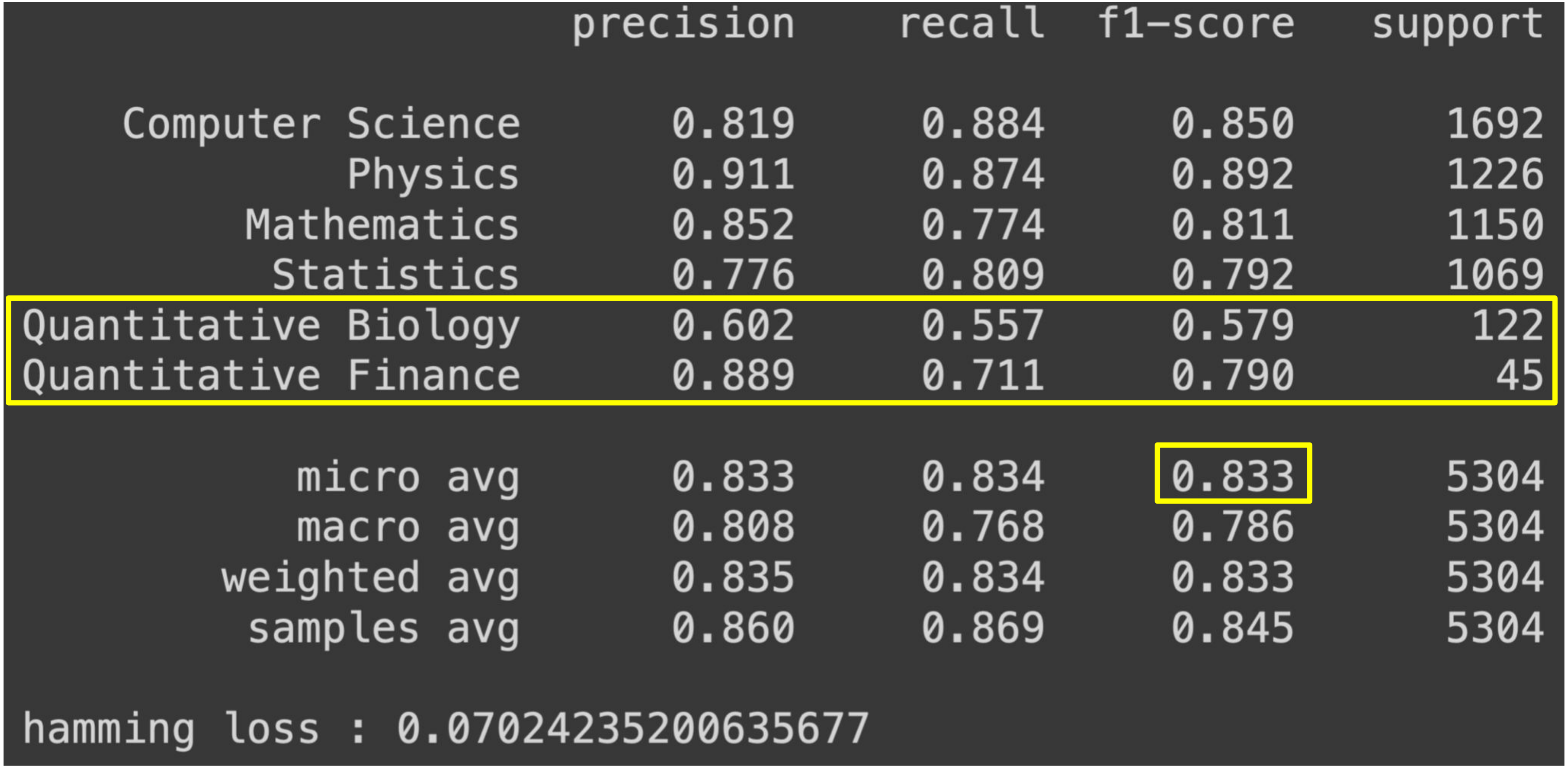
	BERT	**distilBERT**
max sequence	128	no limitation (300)
parameters	109,191,942	66,958,086
output model size	1.3 GB	250 MB

DistilBERT


loss
0.8
0.7
0.6
0.5
0.4
0.3
0.2
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
learning rate (log scale)

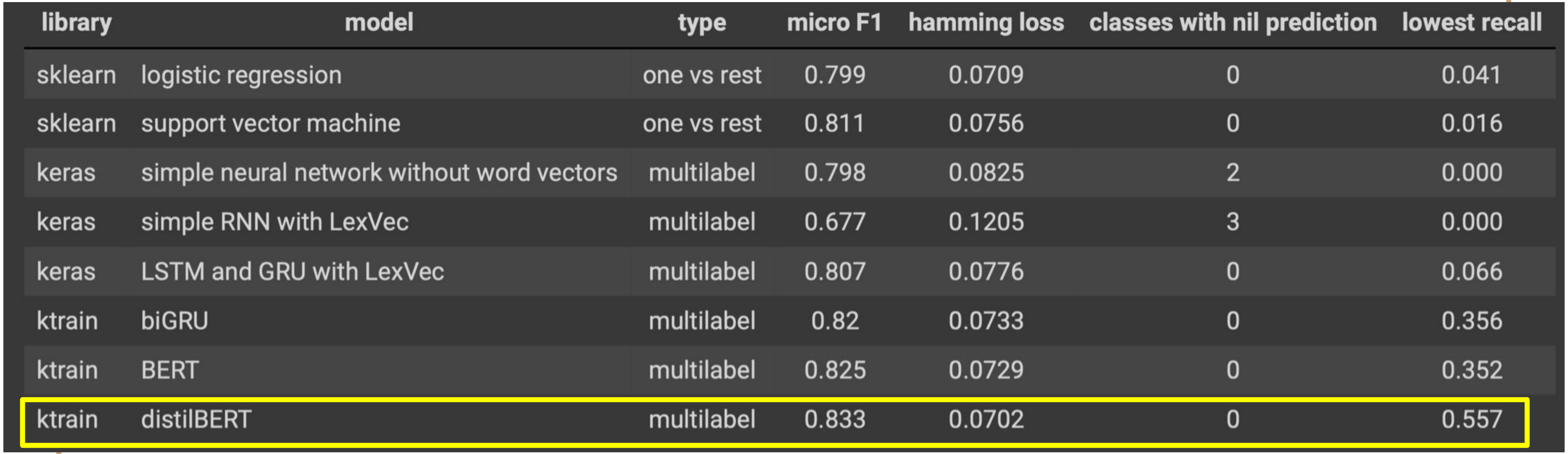

Accuracy
train acc
val acc
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.04
1
2
3
4
5
6
7
8
Loss
0.25
0.20
0.15
0.10
0.05
train loss
val loss
1
2
3
4
5
6
7
8

Final Model - distilBERT



```
                        precision    recall  f1-score   support

      Computer Science      0.819     0.884     0.850      1692
               Physics      0.911     0.874     0.892      1226
           Mathematics      0.852     0.774     0.811      1150
            Statistics      0.776     0.809     0.792      1069
Quantitative Biology      0.602     0.557     0.579       122
Quantitative Finance      0.889     0.711     0.790        45

             micro avg      0.833     0.834     0.833      5304
             macro avg      0.808     0.768     0.786      5304
          weighted avg      0.835     0.834     0.833      5304
           samples avg      0.860     0.869     0.845      5304

hamming loss : 0.07024235200635677
```

Summary of Results



library	model	type	micro F1	hamming loss	classes with nil prediction	lowest recall
sklearn	logistic regression	one vs rest	0.799	0.0709	0	0.041
sklearn	support vector machine	one vs rest	0.811	0.0756	0	0.016
keras	simple neural network without word vectors	multilabel	0.798	0.0825	2	0.000
keras	simple RNN with LexVec	multilabel	0.677	0.1205	3	0.000
keras	LSTM and GRU with LexVec	multilabel	0.807	0.0776	0	0.066
ktrain	biGRU	multilabel	0.82	0.0733	0	0.356
ktrain	BERT	multilabel	0.825	0.0729	0	0.352
ktrain	distilBERT	multilabel	0.833	0.0702	0	0.557

Is that the ground truth?

“Efficient and consistent inference of ancestral sequences in an evolutionary model with insertions and deletions under dense taxon sampling. In evolutionary biology, the speciation history of living organisms is represented graphically by a phylogeny, that is, a rooted tree whose leaves correspond to current species and branchings indicate past speciation events. Phylogenies are commonly estimated from molecular sequences, such as DNA sequences, collected from the species of interest.”

hamming loss = 4/6

	Com Sci	Physics	Math	Statistics	Quant. Biology	Quant. Finance
ground truth	**1**	**0**	**1**	**1**	**0**	**0**
prediction	**0**	**0**	**0**	**0**	**1**	**0**

Is that the ground truth?

“Cyclic Dominance in the Spatial Coevolutionary Optional Prisoner's Dilemma Game This paper studies scenarios of cyclic dominance in a coevolutionary spatial model in which game strategies and links between agents adaptively evolve over time. The Optional Prisoner's Dilemma (OPD) game is employed. The OPD is an extended version of the traditional Prisoner's Dilemma where players have a third option to abstain from playing the game. We adopt an agent-based simulation approach and use Monte Carlo methods to perform the OPD with coevolutionary rules. The necessary conditions to break the scenarios of cyclic dominance are also investigated. This work highlights that cyclic dominance is essential in the sustenance of biodiversity.”

hamming loss = 4/6

	Com Sci	Physics	Math	Statistics	Quant. Biology	Quant. Finance
ground truth	**1**	**1**	**1**	**0**	**0**	**0**
prediction	**0**	**0**	**0**	**0**	**1**	**0**

Suggested Further Studies

1) Tune other hyperparameters such as optimizer and learning rates.

2) Summarize text using BART or T5 to half the length and repeat training using BERT.

3) Relabel or remove the data with obviously wrong ground truths.

Thank you!

Teo Zhan Rui
teozhanrui@gmail.com

www.linkedin.com/in/teozhanrui

https://github.com/teozhanrui