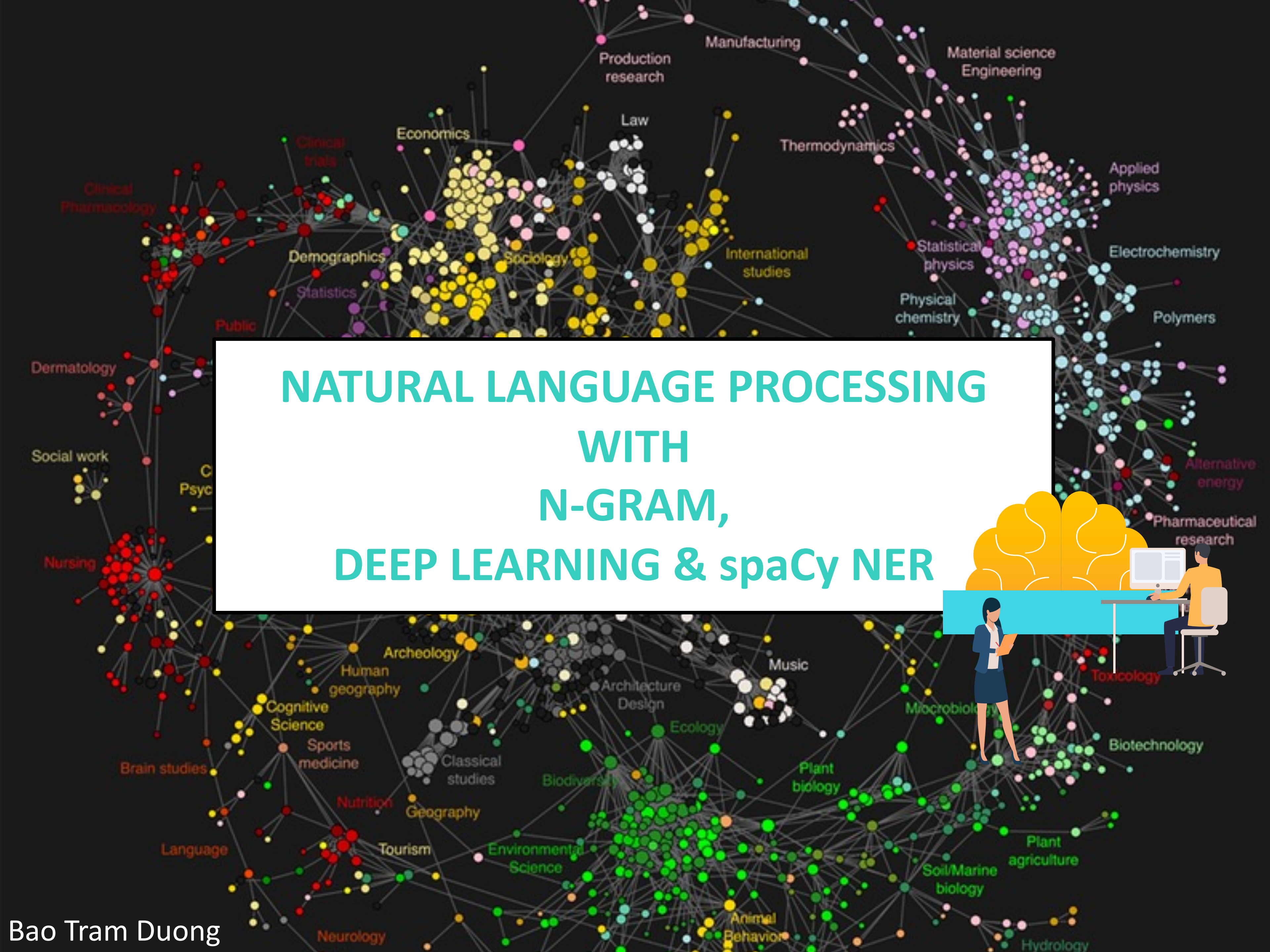
NATURAL LANGUAGE PROCESSING WITH N-GRAM, DEEP LEARNING & spaCy NER



Bao Tram Duong

Business Statement

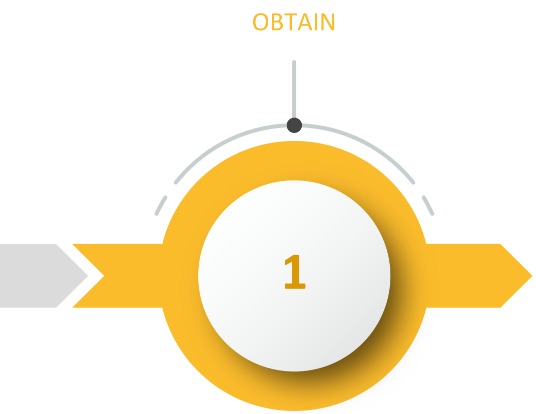
This Kaggle competition challenges data scientists to show how publicly funded data are used to serve science and society.

1. Can NLP find the hidden-in-plain-sight data citations?

2. Can ML find the link between the words used in research articles and the data referenced in the article?

Evidence-Based Policymaking (EBP)

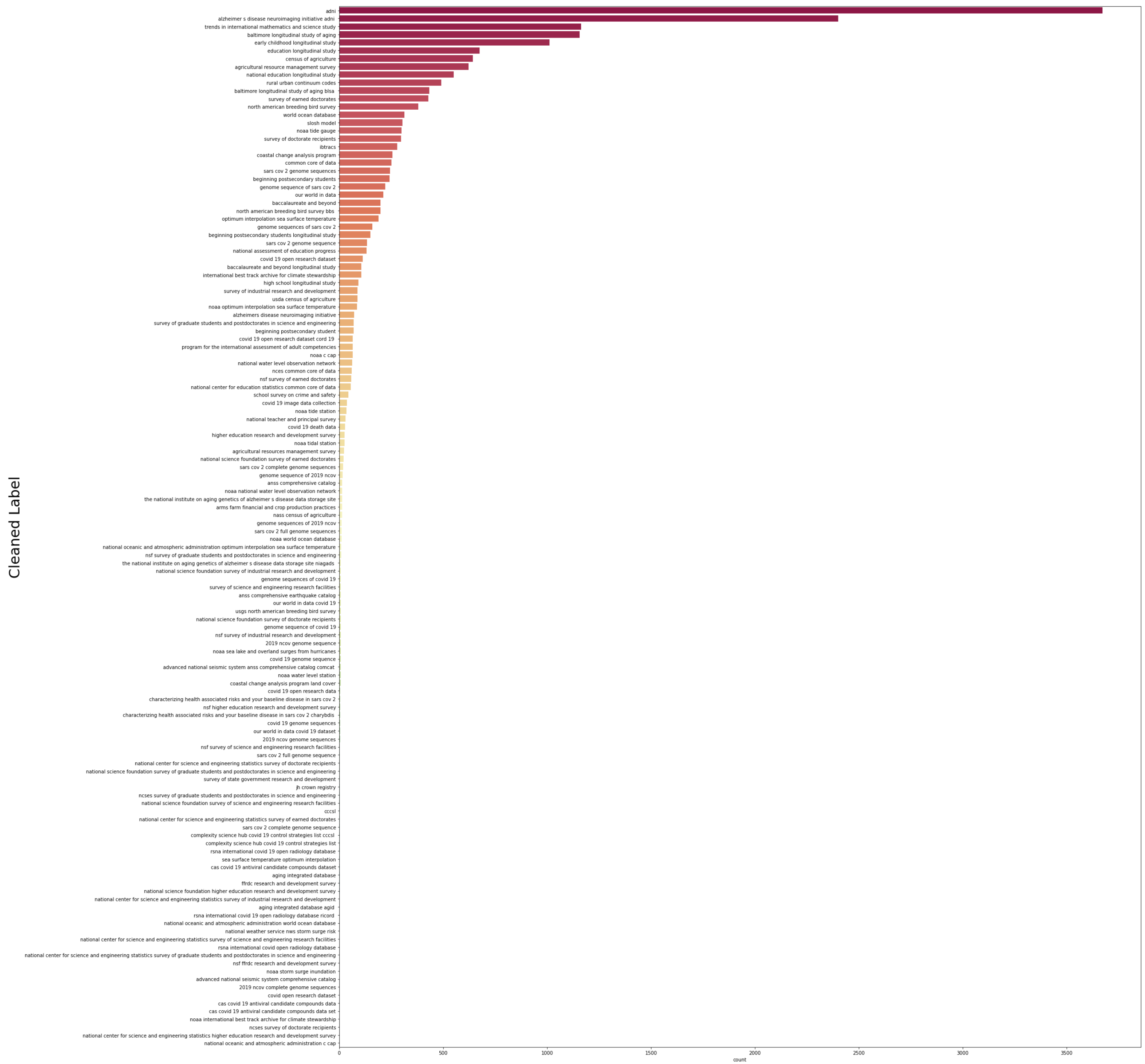
- Foundations of Evidence-based Policymaking Act (2016) requires all federal agencies to show how their data are being used
- Evidence through data is critical if government is to address the many threats facing society: pandemics, climate change, Alzheimer’s disease, child hunger, increasing food production, maintaining biodiversity, and addressing many other challenges.
- Automated NLP tool will enable government agencies and researchers to quickly find the information they need:
  - What datasets are being used to solve problems
  - What measures are being generated
  - Which researchers are the experts

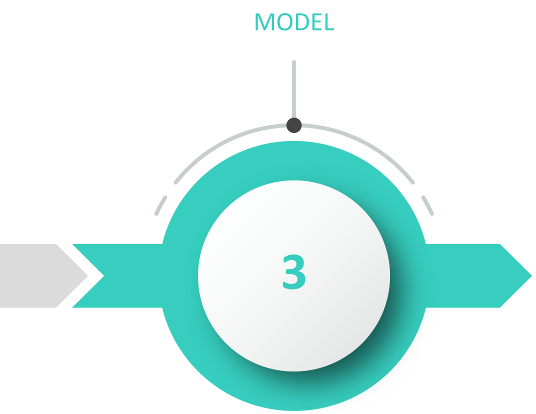


This project uses data from the Kaggle competition sponsored by Coleridge Initiative where scientific publications from numerous research areas are gathered from CHORUS publisher members and other sources

- There are 19,661 publications
- 130 labels


EXPLORE
2


Cleaned Label
count
0
500
1000
1500
2000
2500
3000
3500
adni
alzheimer s disease neuroimaging initiative adni
trends in international mathematics and science study
baltimore longitudinal study of aging
early childhood longitudinal study
education longitudinal study
census of agriculture
agricultural resource management survey
national education longitudinal study
rural urban continuum codes
baltimore longitudinal study of aging blsa
survey of earned doctorates
north american breeding bird survey
world ocean database
slosh model
noaa tide gauge
survey of doctorate recipients
ibtracs
coastal change analysis program
common core of data
sars cov 2 genome sequences
beginning postsecondary students
genome sequence of sars cov 2
our world in data
baccalaureate and beyond
north american breeding bird survey bbs
optimum interpolation sea surface temperature
genome sequences of sars cov 2
beginning postsecondary students longitudinal study
sars cov 2 genome sequence
national assessment of education progress
covid 19 open research dataset
baccalaureate and beyond longitudinal study
international best track archive for climate stewardship
high school longitudinal study
survey of industrial research and development
usda census of agriculture
noaa optimum interpolation sea surface temperature
alzheimers disease neuroimaging initiative
survey of graduate students and postdoctorates in science and engineering
beginning postsecondary student
covid 19 open research dataset cord 19
program for the international assessment of adult competencies
noaa c cap
national water level observation network
nces common core of data
nsf survey of earned doctorates
national center for education statistics common core of data
school survey on crime and safety
covid 19 image data collection
noaa tide station
national teacher and principal survey
covid 19 death data
higher education research and development survey
noaa tidal station
agricultural resources management survey
national science foundation survey of earned doctorates
sars cov 2 complete genome sequences
genome sequence of 2019 ncov
anss comprehensive catalog
noaa national water level observation network
the national institute on aging genetics of alzheimer s disease data storage site
arms farm financial and crop production practices
nass census of agriculture
genome sequences of 2019 ncov
sars cov 2 full genome sequences
noaa world ocean database
national oceanic and atmospheric administration optimum interpolation sea surface temperature
nsf survey of graduate students and postdoctorates in science and engineering
the national institute on aging genetics of alzheimer s disease data storage site niagads
national science foundation survey of industrial research and development
genome sequences of covid 19
survey of science and engineering research facilities
anss comprehensive earthquake catalog
our world in data covid 19
usgs north american breeding bird survey
national science foundation survey of doctorate recipients
genome sequence of covid 19
nsf survey of industrial research and development
2019 ncov genome sequence
noaa sea lake and overland surges from hurricanes
covid 19 genome sequence
advanced national seismic system anss comprehensive catalog comcat
noaa water level station
coastal change analysis program land cover
covid 19 open research data
characterizing health associated risks and your baseline disease in sars cov 2
nsf higher education research and development survey
characterizing health associated risks and your baseline disease in sars cov 2 charybdis
covid 19 genome sequences
our world in data covid 19 dataset
2019 ncov genome sequences
nsf survey of science and engineering research facilities
sars cov 2 full genome sequence
national center for science and engineering statistics survey of doctorate recipients
national science foundation survey of graduate students and postdoctorates in science and engineering
survey of state government research and development
jh crown registry
ncses survey of graduate students and postdoctorates in science and engineering
national science foundation survey of science and engineering research facilities
cccsl
national center for science and engineering statistics survey of earned doctorates
sars cov 2 complete genome sequence
complexity science hub covid 19 control strategies list cccsl
complexity science hub covid 19 control strategies list
rsna international covid 19 open radiology database
sea surface temperature optimum interpolation
cas covid 19 antiviral candidate compounds dataset
aging integrated database
ffrdc research and development survey
national science foundation higher education research and development survey
national center for science and engineering statistics survey of industrial research and development
aging integrated database agid
rsna international covid 19 open radiology database ricord
national oceanic and atmospheric administration world ocean database
national weather service nws storm surge risk
national center for science and engineering statistics survey of science and engineering research facilities
rsna international covid open radiology database
national center for science and engineering statistics survey of graduate students and postdoctorates in science and engineering
nsf ffrdc research and development survey
noaa storm surge inundation
advanced national seismic system comprehensive catalog
2019 ncov complete genome sequences
covid open research dataset
cas covid 19 antiviral candidate compounds data
cas covid 19 antiviral candidate compounds data set
noaa international best track archive for climate stewardship
ncses survey of doctorate recipients
national center for science and engineering statistics higher education research and development survey
national oceanic and atmospheric administration c cap



- N-GRAM CLASSIFIERS
- RECURRENT NEURAL NETWORK (RNNs)
- CONVOLUTIONAL NEURAL NETWORK (CNNs)
- SPACY NER

Data Annotation

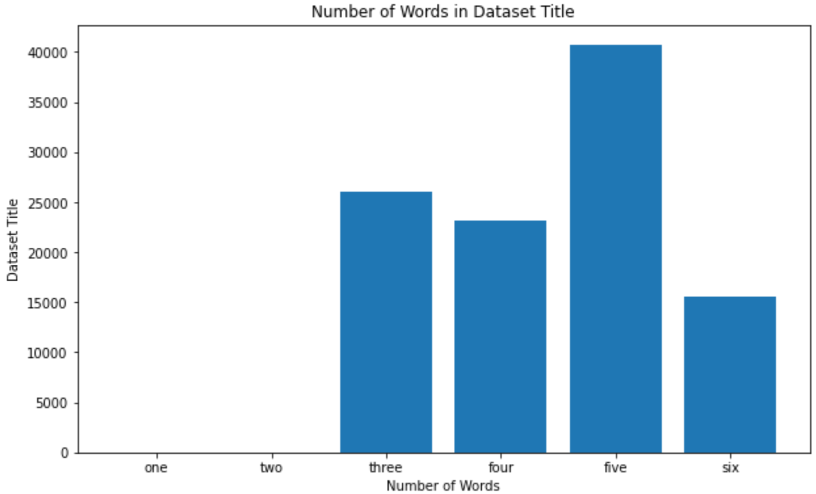
	Sentence	Label
64502	the international community came together to b...	rsna international covid open radiology database
64503	then several 3cl pro ligand complexes are used...	cas covid 19 antiviral candidate compounds dat...
64504	mccs was then applied to carry out the virtual...	cas covid 19 antiviral candidate compounds dat...
64505	section title repurposing cas covid 19 antivi...	cas covid 19 antiviral candidate compounds dat...
64506	using this data we applied a variety of machin...	cas covid 19 antiviral candidate compounds dat...
64507	nearly 50 000 substances from the cas covid 19...	cas covid 19 antiviral candidate compounds dat...
64508	the model was then also applied to the cas cov...	cas covid 19 antiviral candidate compounds dat...
64509	using suitable binary classifiers we were able...	cas covid 19 antiviral candidate compounds dat...
64510	after data cleaning and chemical structure sta...	cas covid 19 antiviral candidate compounds dat...
64511	after data cleaning and chemical structure sta...	cas covid 19 antiviral candidate compounds data

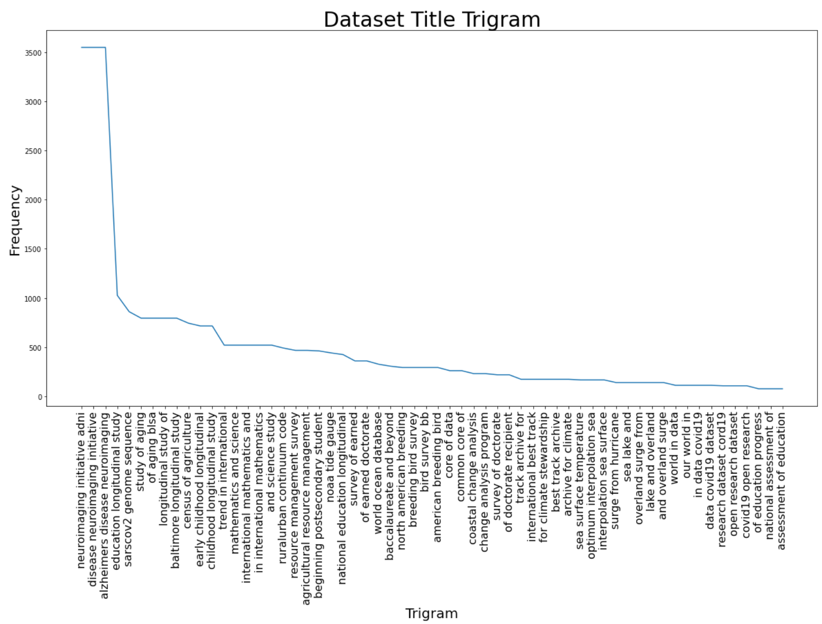
	Sentence	Entities
114805	nearly 50 000 substances from the cas covid 19...	{'entities': [(34, 84, 'DATASET')]}
114806	some predicted molecules of these models were ...	{'entities': []}
114807	the model was then also applied to the cas cov...	{'entities': [(39, 89, 'DATASET')]}
114808	the model predicted that 970 of these chemical...	{'entities': []}
114809	using suitable binary classifiers we were able...	{'entities': [(117, 167, 'DATASET')]}
114810	through these screenings we identified many po...	{'entities': []}
114811	after data cleaning and chemical structure sta...	{'entities': [(243, 294, 'DATASET')]}
114812	these searches led to the identification of st...	{'entities': []}
114813	after data cleaning and chemical structure sta...	{'entities': [(243, 290, 'DATASET')]}
114814	these searches led to the identification of st...	{'entities': []}

Sequence Labeling takes a sequence of input instance and learn to predict an optimal sequence of labels.

1. Collect a set of representative training sentences that has dataset title in them
2. Label each sentence
3. Train a classifier to predict the labels of each annotated training sentences
4. Test to see if the classifier appropriately output the recognized label

N-Grams



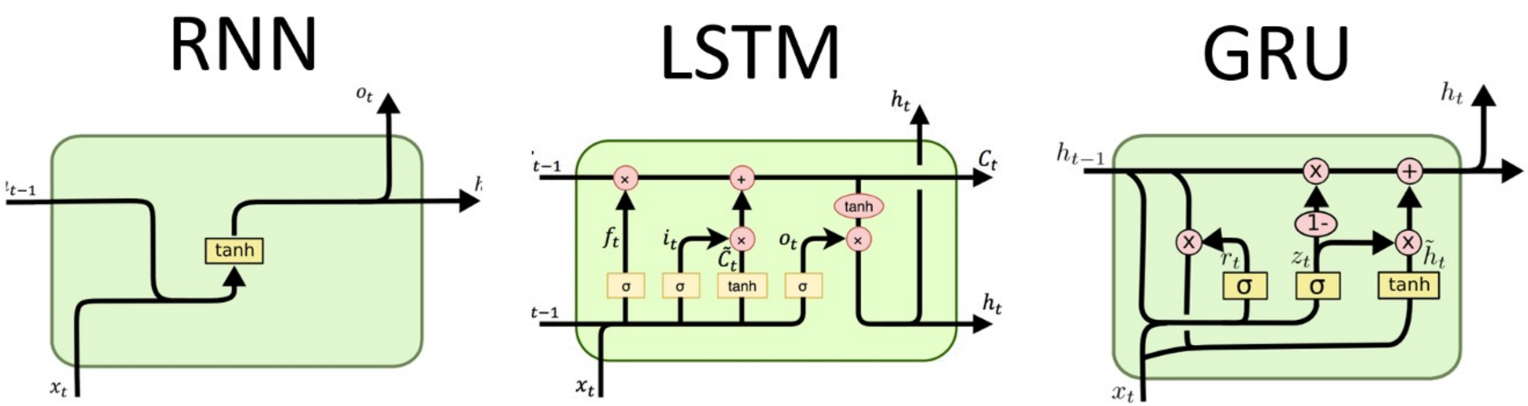


An n-gram model is the simplest model that assigns probabilities to sequences of words i.e. dataset label (3-6 word sequence), without considering the word order.

Tri-gram = how often three particular words occurs together

Few common classifiers: random forest, support vector machine (SVM) and naive Bayes.

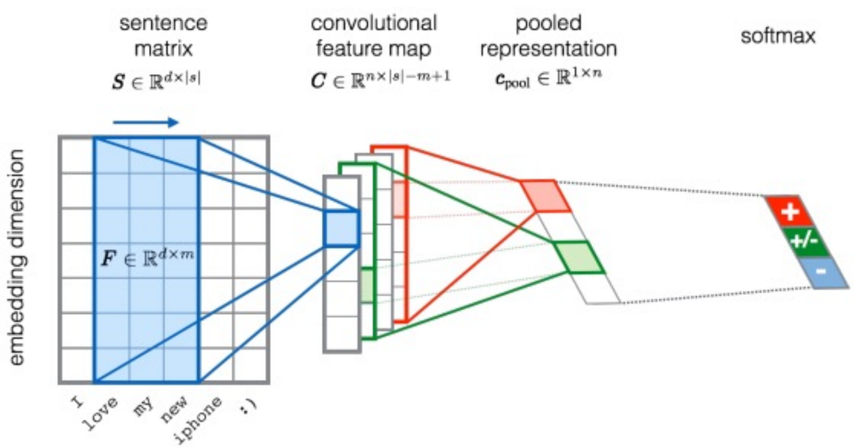
Recurrent Neural Networks (RNNs)



Gated units allow the network to pass or block information from one time step to the other: capable of keeping long-term dependencies effectively while handling the vanishing/exploding gradient problems.

- Input gate regulates how much of the new cell state to keep.
- Forget gate regulates how much of the existing memory to forget.
- Output gate regulates how much of the cell state should be exposed to the next layers of the network.

Convolutional Neural Network (CNNs)



Besides Computer Vision, CNNs can also be applied to NLP. Instead of image pixels, sentences or documents represented as a matrix are the input.

CNNs can identify special pattern of an n-gram in the sentence regardless of their position.

While the RNN computes a weighted combination of all words in the sentence, the CNN extracts the most informative n-grams.

CNNs is much faster than RNNs and much less computationally expensive than n-grams

spaCy Named Entity Recognition

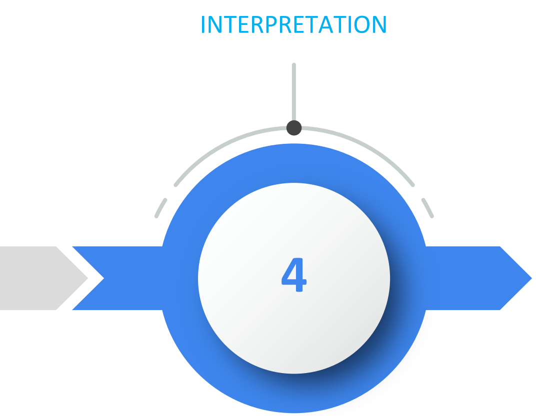
This study used data from the National Education Longitudinal Study **DATASET** (NELS:88) to examine the effects of dual enrollment programs for high school students on college degree attainment.

the international community came together to build the rsna international covid open radiology database **DATASET** ricord 1 which will be accessible to all investigator to further scientific knowledge and facilitate the development of rapid quantitative assessment tool

spaCy is an open-source library for advanced NLP.

Name Entity Recognition (NER) is used to identify entity DATASET mentions in sentences.

Summary of Key Findings



#	Model	Accuracy	CV	Precision	Recall	F1
0	CLF RandomForestClassifier	0.75	0.78	0.44	0.45	0.44
1	CLF Linear Support Vector Machine	0.62	0.62	0.21	0.09	0.11
2	CLF MultinomialNB	0.64	0.62	0.25	0.10	0.12
3	DL GRU	0.83	-	0.29	0.33	0.3
4	DL Bidirectional LSTM	0.83	-	0.28	0.31	0.28
5	DL sep-CNN	0.48	-	0.0	0.01	0.01
6	spaCy NER	0.81	-	-	-	-

- Highly imbalanced dataset
- Number of training samples is not enough
- Missing focus on tweaking the hyper-parameters

Future Work

1st Place Winning Notebook (0.576):
https://www.kaggle.com/dathudeptrai/biomed-roberta-scibert-base
https://www.kaggle.com/suicaokhoailang/submit-gpt-spacy?scriptVersionId=66488765

- Context Similarity via Deep Metric Learning
  - A shared Bert model for extract Context Embedding and Sequence Tokens Embedding
  - An ArcFace Loss for training Mask/NoMask Embedding
  - A BCE loss for training NER model to detect dataset citation in the input string
- Text extraction model with CLM backbone and beamsearch GPT

2nd Place Winning Notebook (0.575):
https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data/discussion/248296

- Search for named entities using the Schwartz-Hearst algorithm
  - Filter candidates using a fine-tuned Roberta-base binary classifier
  - Threshold and propagate candidates

Recommendations

1. With supervised learning algorithms, large annotated data for training is required which are expensive and often take a lot of time. Future efforts could be dedicated on providing more effective deep transfer learning models and exploring appropriate data augmentation techniques (He et al., 2020).
2. Most neural models for Sequence Labeling do not scale well because when the size of data grows, the parameters of models increase exponentially, leading to the high complexity of back propagation. There exists need for developing approaches to balance model complexity and scalability (He et al., 2020).
3. Increasing access to confidential data presumed significantly increasing privacy risks. However, the U.S. laws and practices are not currently optimized to support the use of data for evidence building, nor in a manner that best protects privacy. We need to improve data security and privacy protections beyond what exists today (US CEP, 2017).

THANK YOU

APPENDIX

Reference

Banerjee, I., Ling, Y., Chen, M. C., Hasan, S. A., Langlotz, C. P., Moradzadeh, N., Chapman, B., Amrhein, T., Mong, D., Rubin, D. L., Farri, O., & Lungren, M. P. (2019). Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artificial intelligence in medicine, 97, 79–88. https://doi.org/10.1016/j.artmed.2018.11.004

Britz, D. (2016, January 10). Understanding Convolutional Neural Networks for NLP. WildML. http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/.

Chollet, F. (2017). Chapter 6. Deep learning for text and sequences. Deep Learning with Python. · Deep Learning with Python. https://livebook.manning.com/book/deep-learning-with-python/chapter-6/18.

Google. (n.d.). Step 4: Build, Train, and Evaluate Your Model. Google. https://developers.google.com/machine-learning/guides/text-classification/step-4.

He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., Jiang, S. (2020). A Survey on Recent Advances in Sequence Labeling from Deep Learning Models. arXiv:2011.06727v1 [cs.CL]. Cornell University.

Meparlad. (2020, December 11). Text Classification in Natural Language Processing. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example/.

Natural Language Toolkit. Natural Language Toolkit — NLTK 3.6.2 documentation. (n.d.). https://www.nltk.org/.

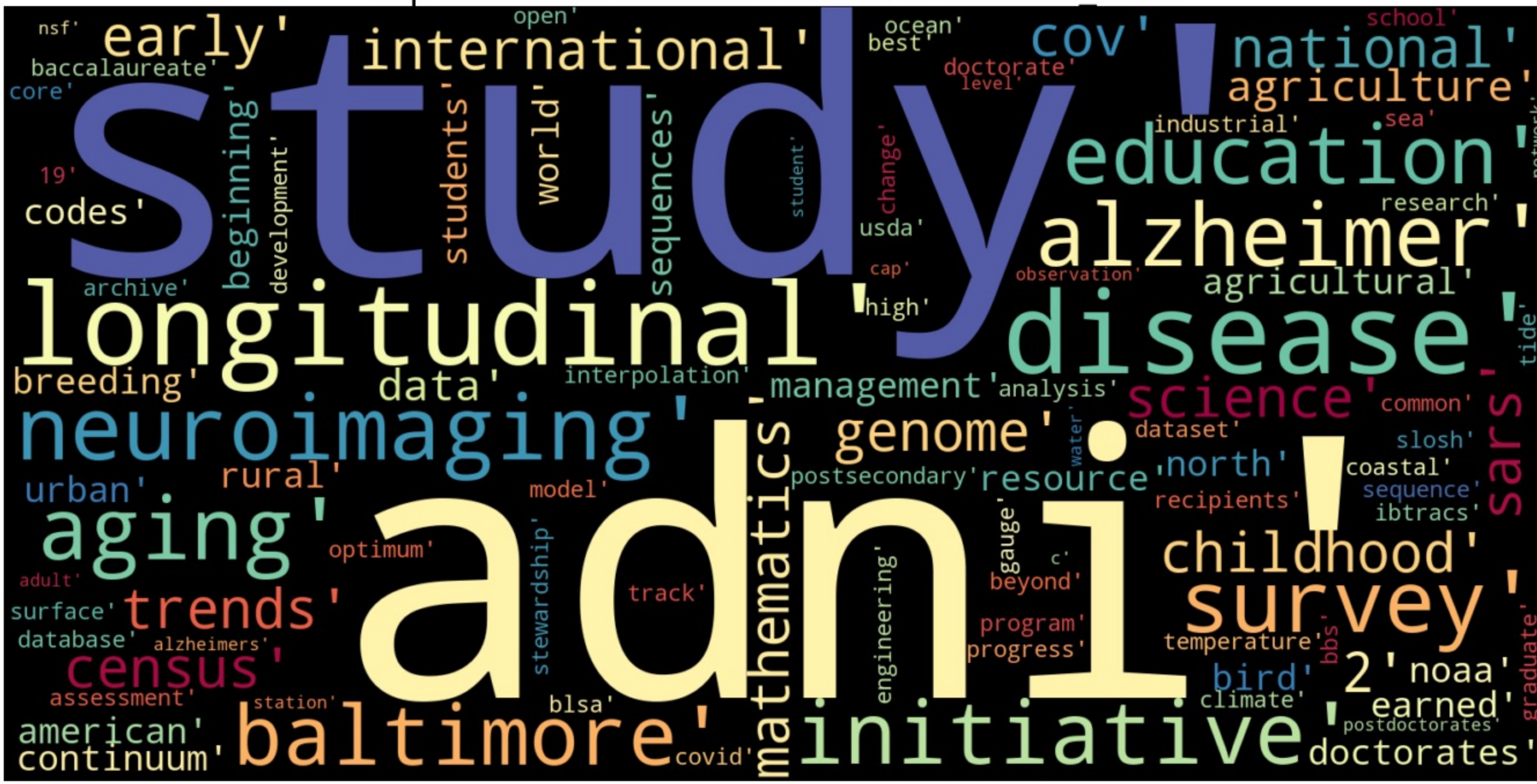
Pennington, J. (n.d.). GloVe: Global Vectors for Word Representation. https://nlp.stanford.edu/projects/glove/.

Phi, M. (2020, June 28). Illustrated Guide to LSTM's and GRU's: A step by step explanation. Medium. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

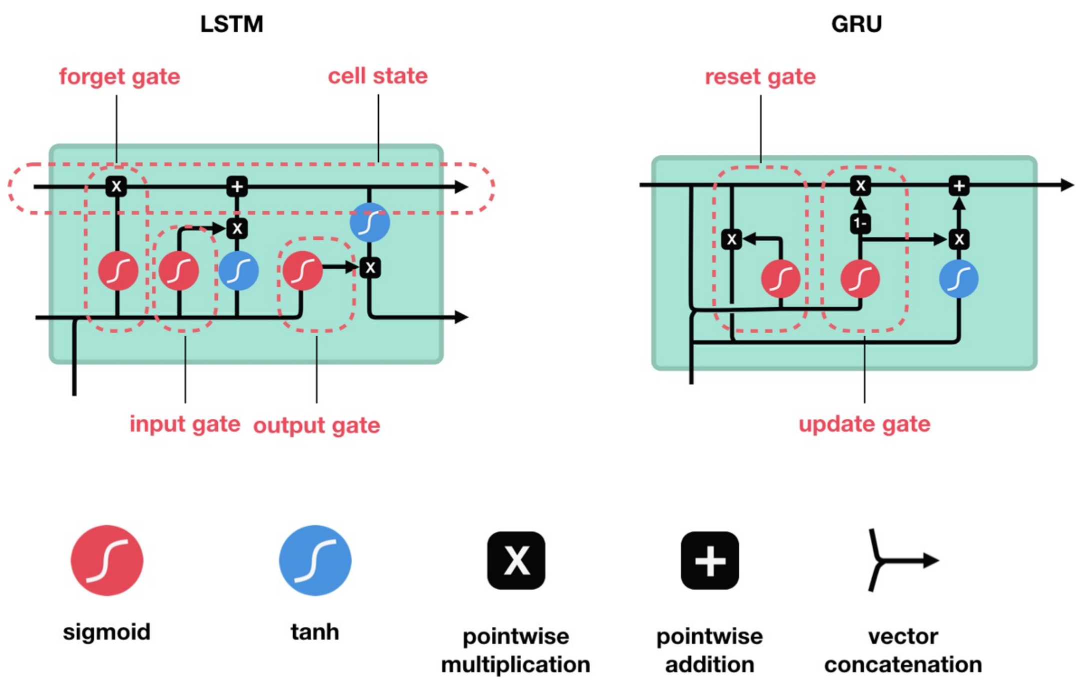
spaCy 101: Everything you need to know · spaCy Usage Documentation. spaCy 101: Everything you need to know. (n.d.). https://spacy.io/usage/spacy-101.

United States Commission on Evidence-Based Policymaking. (2017). The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking.

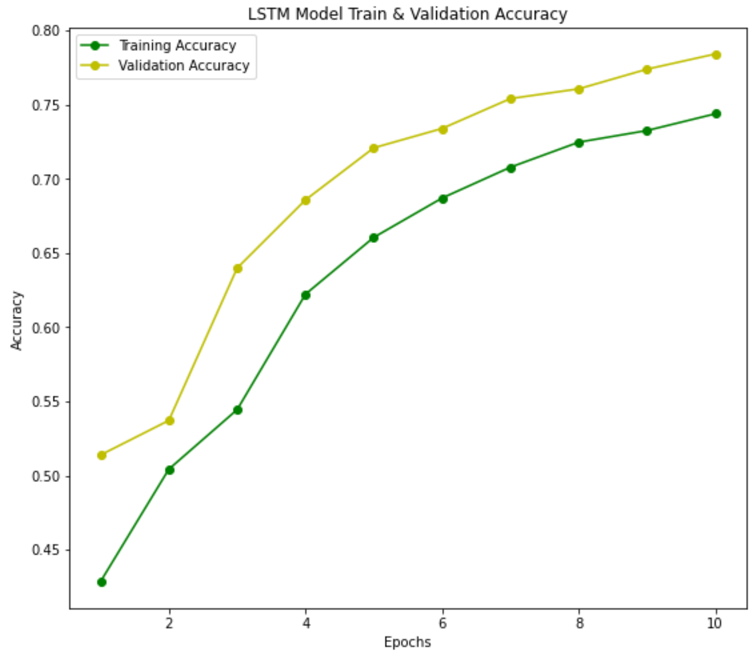
Top 100 Most Common Words in cleaned_label

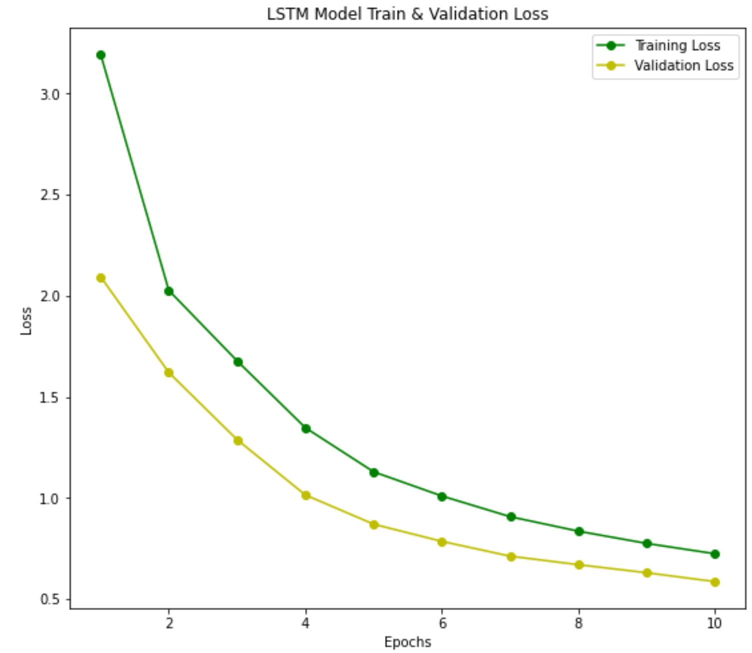


Bidirectional LSTM vs. GRU

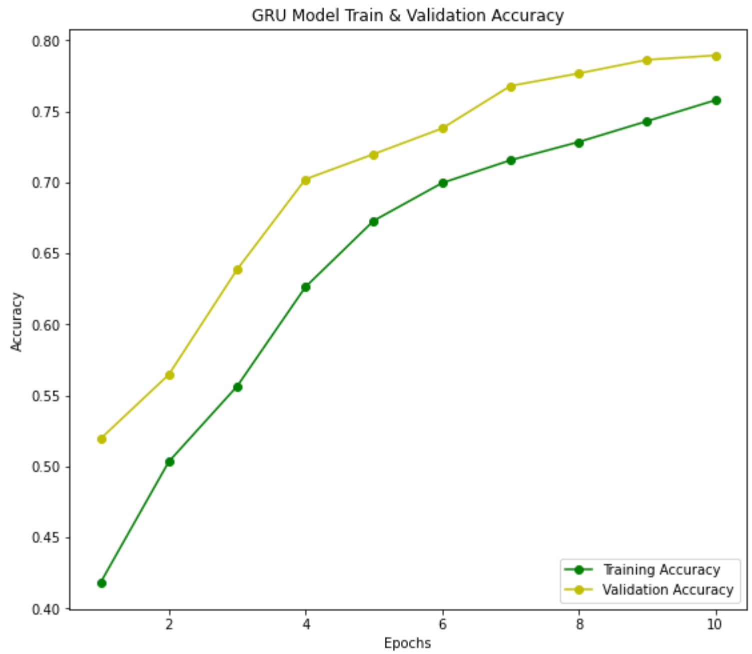


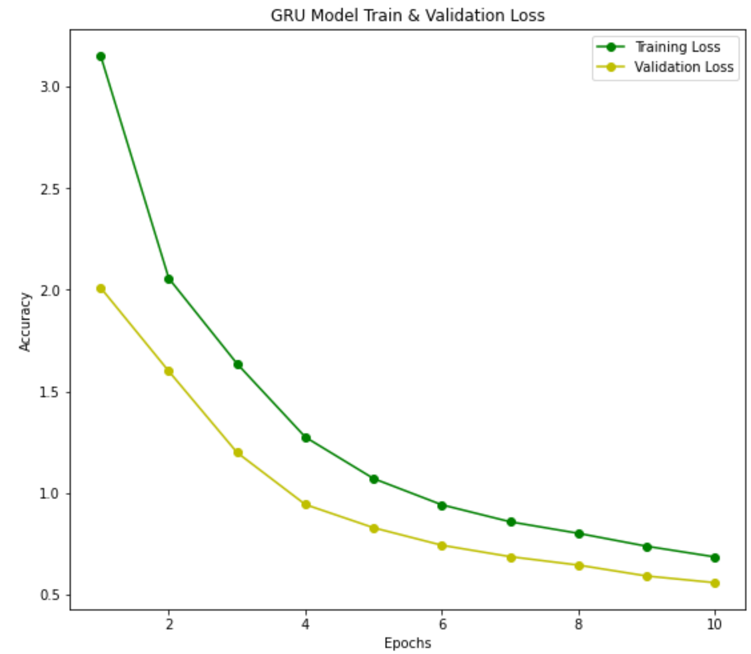
- GRU has two gates (reset and update gates) whereas an LSTM has three gates (input, output and forget gates)
- LSTM remember longer sequences than GRU
- GRU is simpler and trains faster than LSTM



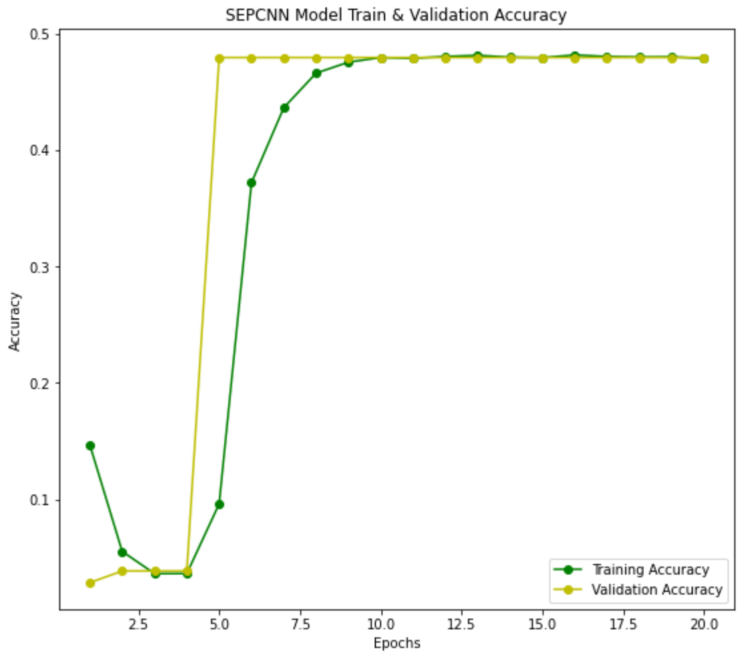


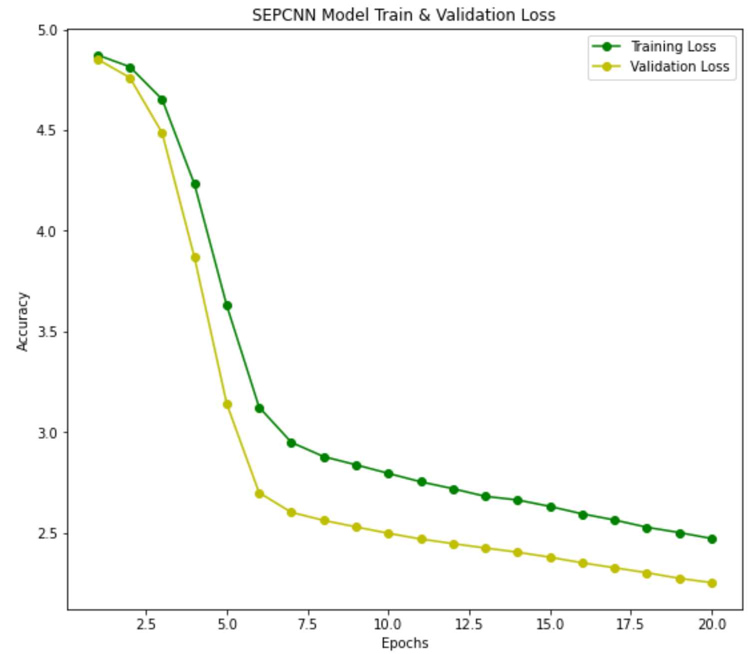

```
1613/1613 [==============================] - 346s 215ms/step - loss:
0.3470 - acc: 0.8369
Train loss & accuracy: [0.34700754284858704, 0.8368695378303528]


404/404 [==============================] - 88s 218ms/step - loss:
0.3859 - acc: 0.8283
Test loss & accuracy: [0.3859459459781647, 0.8282570242881775]
```






```
1613/1613 [==============================] - 437s 271ms/step - loss:
0.5258 - acc: 0.7959
Train loss & accuracy: [0.5257872343063354, 0.7959076762199402]


404/404 [==============================] - 109s 269ms/step - loss:
0.5597 - acc: 0.7894
Test loss & accuracy: [0.5597025156021118, 0.7893512845039368]
```






```
1613/1613 [==============================] - 24s 15ms/step - loss:
2.3333 - acc: 0.4805
Train loss & accuracy: [2.333292007446289, 0.4804975986480713]


404/404 [==============================] - 6s 15ms/step - loss:
2.3373 - acc: 0.4797
Test loss & accuracy: [2.3373019695281982, 0.4796558916568756]
```

```
cleaned label: {'adni', 'alzheimer s disease neuroimaging initiative
adni '}
RandomForestClassifier_label: {'adni'}
MultinomialNB_label: {'adni'}
SGDClassifier_label: {'adni'}
lstm label: {'adni'}
gru label: {'adni'}
sepcnn label: {'adni'}
spacy label: adni


cleaned label: {'common core of data', 'trends in international
mathematics and science study', 'nces common core of data'}
RandomForestClassifier_label: {'adni', 'common core of data',
'census of agriculture', 'trends in international mathematics and
science study', 'ibtracs', 'program for the international assessment
of adult competencies', 'baccalaureate and beyond'}
MultinomialNB_label: {'beginning postsecondary student', 'adni',
'trends in international mathematics and science study', 'early
childhood longitudinal study'}
SGDClassifier_label: {'common core of data', 'adni', 'trends in
international mathematics and science study'}
lstm label: {'adni'}
gru label: {'our world in data'}
sepcnn label: {'adni'}
spacy label: trends in international mathematics and science study


cleaned label: {'slosh model', 'noaa storm surge inundation', 'sea
lake and overland surges from hurricanes'}
RandomForestClassifier_label: {'noaa tide gauge', 'slosh model',
'adni', 'ibtracs', 'noaa storm surge inundation'}
MultinomialNB_label: {'slosh model', 'adni'}
SGDClassifier_label: {'adni'}
lstm label: {'ibtracs'}
gru label: {'ibtracs'}
sepcnn label: {'adni'}
spacy label: slosh model


cleaned label: {'rural urban continuum codes'}
RandomForestClassifier_label: {'adni', 'census of agriculture',
'rural urban continuum codes', 'ibtracs'}
MultinomialNB_label: {'adni', 'rural urban continuum codes'}
SGDClassifier_label: {'adni', 'rural urban continuum codes'}
lstm label: {'adni'}
gru label: {'adni'}
sepcnn label: {'adni'}
spacy label: rural urban continuum codes
```