LDSI W2021 Project Report

Name: Devansh Sharma
Gloss ID: tum_ldsi_67

Summary

The project deals with the classification of standard annotation types in BVA decision cases using and formulating best feature vectors via various NLP techniques focusing on legal text data and using Machine Learning models to perform classification with help of these features and observing the results.

Dataset Splitting

Document wise split:

Train set docs	Dev set docs	Test set docs
113	14	14

Annotation wise split:

Train set annotations	Dev set annotations	Test set annotations
12236	1410	1703

Test set document ids:

Granted Decisions:
61aea55f97ad59b4cfc41332, 61aea55f97ad59b4cfc41336, 61aea55f97ad59b4cfc41334
61aea55f97ad59b4cfc41337, 61aea55f97ad59b4cfc41335, 61aea55f97ad59b4cfc4133b
61aea55f97ad59b4cfc4133c,

Denied Decisions:
61aea56f97ad59b4cfc41342,61aea56f97ad59b4cfc41343,61aea56f97ad59b4cfc41344,
61aea56f97ad59b4cfc41347, 61aea56f97ad59b4cfc41349
61aea56f97ad59b4cfc4134b, 61aea56f97ad59b4cfc4134c

Development set document ids:

Granted Decisions:
61aea55c97ad59b4cfc41290, 61aea55c97ad59b4cfc41297, 61aea55c97ad59b4cfc41299
61aea55c97ad59b4cfc4129b, 61aea55c97ad59b4cfc4129d, 61aea55c97ad59b4cfc4129e
61aea55c97ad59b4cfc4129f,

Denied Decisions:
61aea56f97ad59b4cfc4134d, 61aea57097ad59b4cfc41351
61aea57097ad59b4cfc41352, 61aea57097ad59b4cfc41355, 61aea57097ad59b4cfc41358
61aea57097ad59b4cfc4135a, 61aea57097ad59b4cfc4135b

Sentence Segmentation

Error Analysis on standard spacy segmenter:

True Positives - 7564
False Positives - 5154
False Negatives - 4672
Precision - 0.595
Recall - 0.618
F1 Score - 0.606

Examining three different documents with lowest F1 scores here are the findings:

- It has difficulty in segmenting headings since I believe punctuation marks are the most basic attribute for it to mark sentence ends.
- Various keywords like: Vet. App. DOCKET NO. Supp. Stat. Fed. etc. mark end of sentence as they have a full stop in their end.
- Case header has difficulty in being segmented because of its double parentheses ")\n)" to mark its end.

Error Analysis on improved spacy segmenter:

True Positives - 8626
False Positives - 3906
False Negatives - 3610
Precision - 0.688
Recall - 0.704
F1 Score - 0.697

Improvements and additions to standard spacy segmenter:

- Using spacy's add_pipe() feature to add some custom rules for segmentation. Certain rules were added to make sure most headings always get correctly segmented. The fact that most headings are in all CAPS("ORDER", "REPRESENTATION" etc.) was exploited here. Such as
- By manually observing the certain keywords and phrases were added via add_special_case() function so over segmentation is prevented. The words added: 'Vet. App.' 'DOCKET NO.' 'Supp.' 'Stat.' 'Pub. L. No.' 'Fed.' 'Reg.' 'Fed. Cir.' 'in-' 'Cf.'

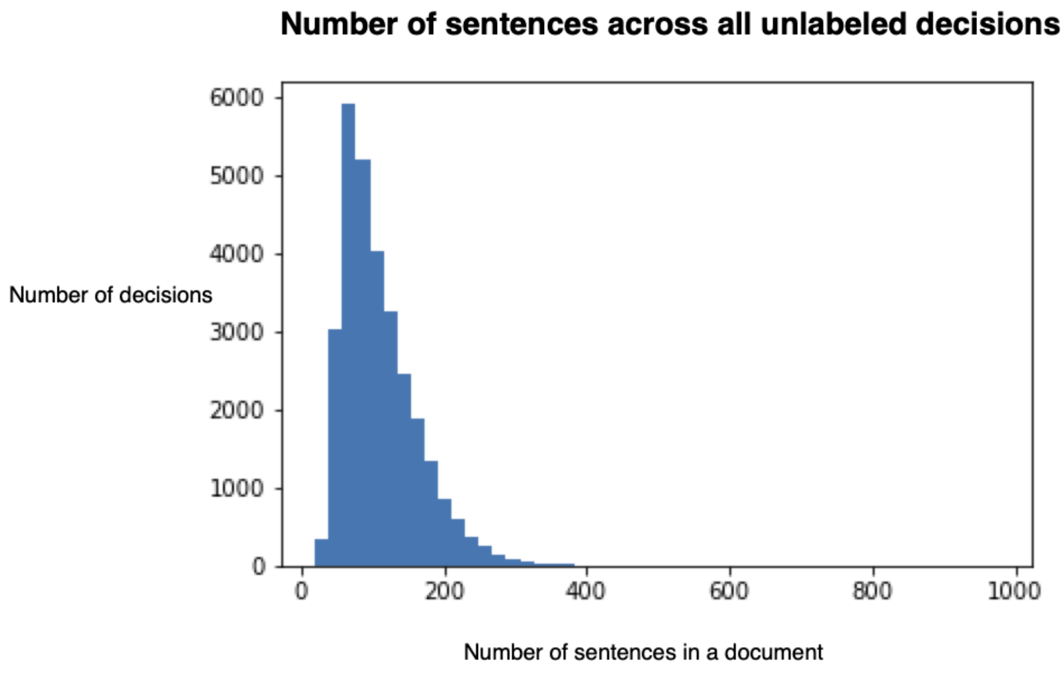
Error Analysis on Savelka's law-specific sentence segmenter (Savelka et al.) :

True Positives - 12148
False Positives - 2359
False Negatives - 88
Precision - 0.837
Recall - 0.992
F1 Score - 0.908

- It is pretty clear from the metrics that using the law-specific segmenter is very helpful.
- It accurately segments most annotations including the headers and ignores most law specific keywords to prevent over segmentation.
- Although it was observed the case header(if space left add example I guess) was over segmented and some citations got over segmented too but overall there was no significant difference between true and Savelka's segmentation.

Considering the observations, using Savelka's law-specific sentence segmenter seems a better option.

Preprocessing



Tokenizer:

- The spacy tokenizer has been used which ignores tokens with pos tag PUNCT being ignored and tokens with pos tag NUM being appended as <NUM(len(n))> .

Improvements to tokenizer:

- It was observed that some punctuations were included in the tokens. These were the full stops followed by the keywords like ‘Vet.’ or ‘Cir.’ or ‘App.’ Spacy for some reason tokenizes ‘Vet’ and ’.’ as separate tokens with pos tag PROPN.
- To avoid this regex was used in the tokenizer function to remove all non-alphanumeric characters and also lowercase all the tokens for uniformity.

Example:
example_mixed_1 = 'In Dingess v. Nicholson, 19 Vet. App. 473 (2006), the U.S. Court of Appeals for Veterans Claims held that, upon receipt of an application for a service-connection claim, 38 U.S.C.A. � 5103(a) and 38 C.F.R. � 3.159(b) require VA to provide the claimant with notice that a disability rating and an effective date for the award of benefits will be assigned if service connection is awarded. '

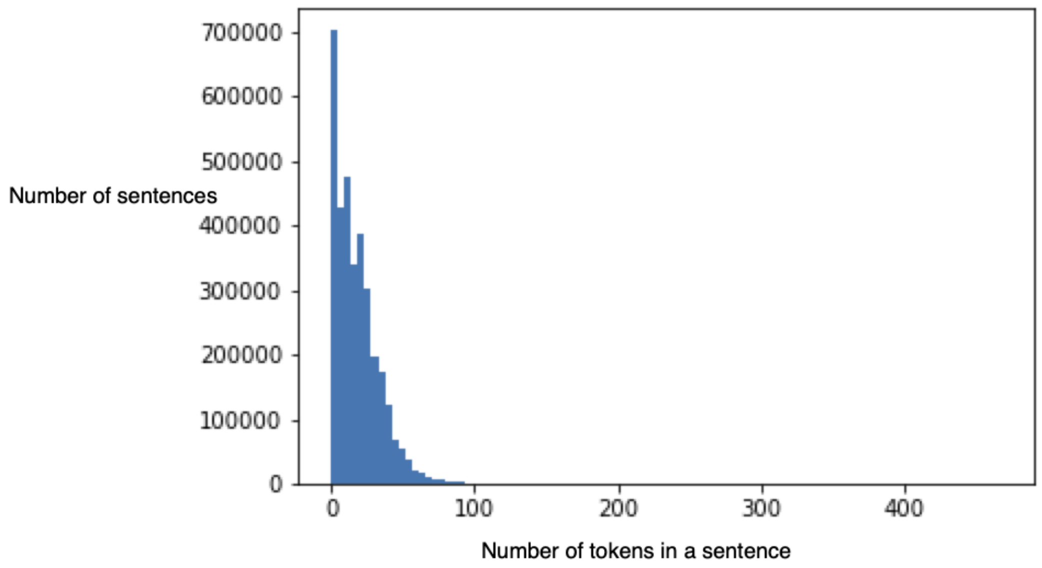
Tokens without improvement:

['in', 'Dingess', 'v.', 'Nicholson', '<NUM2>', 'Vet', '.', 'App', '<NUM3>', '<NUM4>', 'the', 'U.S.', 'Court', 'of', 'Appeals', 'for', 'Veterans', 'Claims', 'hold', 'that', 'upon', 'receipt', 'of', 'an', 'application', 'for', 'a', 'service', 'connection', 'claim', '<NUM2>', 'U.S.C.A.', '�', '<NUM6>', 'and', '<NUM2>', 'C.F.R.', '�', '3.159(b', 'require', 'VA', 'to', 'provide', 'the', 'claimant', 'with', 'notice', 'that', 'a', 'disability', 'rating', 'and', 'an', 'effective', 'date', 'for', 'the', 'award', 'of', 'benefit', 'will', 'be', 'assign', 'if', 'service', 'connection', 'be', 'award']

Tokens after improvement:

['in', 'dingess', 'v', 'nicholson', 'num2', 'vet', 'app', 'num3', 'num4', 'the', 'us', 'court', 'of', 'appeals', 'for', 'veterans', 'claims', 'hold', 'that', 'upon', 'receipt', 'of', 'an', 'application', 'for', 'a', 'service', 'connection', 'claim', 'num2', 'usca', 'num6', 'and', 'num2', 'cfr', '3159b', 'require', 'va', 'to', 'provide', 'the', 'claimant', 'with', 'notice', 'that', 'a', 'disability', 'rating', 'and', 'an', 'effective', 'date', 'for', 'the', 'award', 'of', 'benefit', 'will', 'be', 'assign', 'if', 'service', 'connection', 'be', 'award']

Number of tokens in each sentence for unlabled data



Custom Embeddings

Exploring nearest neighbors of custom FastText (Bojanowski et al.) embeddings for certain strings:

“veteran” - appellant, he, his, the, kettelle, additionally, leboff, she, regan, coyle

The word appellant here is caught well as veteran here is the appellant.
The pronouns like he,his and she are also caught in as they are usually the pronouns of the veteran with he,his having more signal which can be inferred as more veterans being male.
The names Kettelle,Leboff and Regan seemed to be the most frequent attorneys for the boards. Since this is always followed by the INTRODUCTION heading and the first sentence mostly starts with “The veteran…” that signal is caught on too.
The word additionally appears many times in the vicinity of word veteran in Evidence, EvidenceBasedReasoning etc.

“v” - vet, app, brown, nicholson,supra, f3d, principi, shinseki, num3, curiam

I am using “v” here instead of “v.” as in pre-processing all punctuations are removed regardless.
All of these neighbors can be seen as part of various citations in the documents and since v. occurs in the vicinity it has been captured from the corpora.

“argues” - argue,ihp,argument, agree, holbrook, reassert, arguendo,highlight, richards,guess

ihp - informal hearing presentation is the informal hearing you take for appealing to VA decision. Holbrook,richards are citation related words. We can see other similar words too

“ptsd” - mdd,depressive,pstd,bipolar, dysthymic, dysthymia, schizoaffective, schizophrenia anxiety,depression

mdd - Major depressive disorder, many other mental disorder related keywords found.

“granted” - grant, unwarranted, bardwell, grantham, granting, perman, routen, connection, brock, dismiss

“Korea” - korean,vietnam,germany,panama,okinawa,iraq,dmz ,saudi,overseas,kuwait

Various countries where the veterans were deployed were found similar.
Dmz- Demilitarized zone areas nearby frontiers where soldiers are usually deployed

“Holding” - ruling,clemons,kent,supra,dingess,mandate,hartman,clemon,directive,arneson

Various legal keywords found similar including the names occurring in various citations

“Also”-additionally,furthermore,far,lastly,addition,although,nevertheless,moreover,nonetheless, 5103aac

Various synonyms found and the citation keyword 5103aac also has slight correlation

“Board” - boards,bva,appeals,appeal,the,decision,veterans,court,remand,ro

We get all necessary important neighbors like bva, the full form etc. also the word ro which refers to regional office is identified

“Representative” - accredit, accredited,supplemental,attorney,dav,furnish,appoint,spouse ,opportunity,representation

dav- disabled american veterans found as neighbor, successfully correlates to potential representative like attorney,spouse,dav, and other meaningful words like accredit

“Duty” - assist acdutra, active, guard, reserves, fulfill, reserve, adt, obligation,statutory

Various relatable words like active duty, fulfill duty, obligation, statutory found.
acdtura - active duty training, adt - active duty training, important acronyms related

Training & Optimizing Classifiers

TFIDF featurization:
2961 features formed after fitting on 12236 annotations from the training set.

Classification report for best TFIDF Model(RandomForest) on Train and Dev:

TRAIN:

```
                                   precision    recall  f1-score   support

CaseFooter                             1.00      1.00      1.00       112
CaseHeader                             1.00      0.99      1.00       115
CaseIssue                              0.99      1.00      1.00       113
Citation                               1.00      1.00      1.00      1963
ConclusionOfLaw                        1.00      0.97      0.98       273
Evidence                               0.86      1.00      0.92      3592
EvidenceBasedOrIntermediateFinding     1.00      0.85      0.92      1185
EvidenceBasedReasoning                 1.00      0.66      0.79       842
Header                                 0.98      1.00      0.99      1191
LegalRule                              0.99      0.96      0.97      1567
LegislationAndPolicy                   1.00      0.81      0.89       130
PolicyBasedReasoning                   1.00      0.71      0.83        21
Procedure                              1.00      0.95      0.98      1130
RemandInstructions                     1.00      0.50      0.67         2

accuracy                                                   0.95     12236
macro avg                              0.99      0.89      0.92     12236
weighted avg                           0.95      0.95      0.95     12236
```

DEV:

```
                                   precision    recall  f1-score   support

CaseFooter                             1.00      1.00      1.00        14
CaseHeader                             0.93      0.93      0.93        14
CaseIssue                              1.00      1.00      1.00        14
Citation                               0.99      0.97      0.98       204
ConclusionOfLaw                        0.80      0.71      0.75        34
Evidence                               0.68      0.97      0.80       466
EvidenceBasedOrIntermediateFinding     0.79      0.37      0.50       140
EvidenceBasedReasoning                 0.75      0.03      0.07        87
```

```
Header                              0.95      0.97      0.96       146
LegalRule                           0.79      0.76      0.77       165
LegislationAndPolicy                0.50      0.15      0.24        26
PolicyBasedReasoning                0.00      0.00      0.00         6
Procedure                           0.96      0.88      0.92        93
RemandInstructions                  0.00      0.00      0.00         1

accuracy                                                0.80      1410
macro avg                           0.72      0.62      0.64      1410
weighted avg                        0.80      0.80      0.76      1410
```

Classification report for best Word Embedding Model(RandomForest) on Test set:
TRAIN:

```
                                  precision    recall  f1-score   support

CaseFooter                          1.00      1.00      1.00       112
CaseHeader                          1.00      0.98      0.99       115
CaseIssue                           0.99      1.00      1.00       113
Citation                            1.00      1.00      1.00      1963
ConclusionOfLaw                     1.00      0.82      0.90       273
Evidence                            0.78      1.00      0.88      3592
EvidenceBasedOrIntermediateFinding  0.99      0.69      0.81      1185
EvidenceBasedReasoning              1.00      0.56      0.72       842
Header                              0.97      1.00      0.99      1191
LegalRule                           0.99      0.91      0.95      1567
LegislationAndPolicy                1.00      0.84      0.91       130
PolicyBasedReasoning                1.00      0.71      0.83        21
Procedure                           0.99      0.92      0.96      1130
RemandInstructions                  0.00      0.00      0.00         2

accuracy                                                0.91     12236
macro avg                           0.91      0.82      0.85     12236
weighted avg                        0.93      0.91      0.91     12236
```

DEV:

```
                                  precision    recall  f1-score   support

CaseFooter                          1.00      1.00      1.00        14
CaseHeader                          0.93      0.93      0.93        14
CaseIssue                           1.00      1.00      1.00        14
Citation                            0.97      0.99      0.98       204
ConclusionOfLaw                     0.89      0.50      0.64        34
Evidence                            0.63      0.97      0.76       466
EvidenceBasedOrIntermediateFinding  0.62      0.09      0.16       140
EvidenceBasedReasoning              0.33      0.01      0.02        87
Header                              0.95      0.99      0.97       146
LegalRule                           0.74      0.53      0.61       165
LegislationAndPolicy                0.50      0.12      0.19        26
PolicyBasedReasoning                0.00      0.00      0.00         6
Procedure                           0.73      0.95      0.83        93
```

```
RemandInstructions                        0.00      0.00      0.00         1

accuracy                                                      0.74      1410
macro avg                                0.66      0.58      0.58      1410
weighted avg                             0.72      0.74      0.68      1410
```

I tried two models, the LinearSVM from the linear models and Random Forest Classifier(RF) from non-linear models using both TFIDF and word embeddings separately. RF performs better in both cases and TFIDF outperforms word embeddings model. Three hyper parameters were focused on in RF n_estimators, max_depth and min_samples_split. Three variations were tried and at the end with n_estimators = 300 for both models, max_depth = 30,16 respectively for TFIDF and word embeddings and min_samples_split =2 final results were obtained. Increasing the estimators or depth beyond this leads to overfitting. Also LinearSVM with C = 1e-4 obtained comparable F1-score to the best model with TFIDF but RF had much higher precision.

TFIDF outperformed word embeddings everywhere but it was especially shown in the Evidence, EvidenceBasedReasoning and EvidenceBasedOrIntermediateFinding classification. RF was able to predict many EvidenceBasedOrIntermediateFinding correctly and differentiate it from Evidence. On the other hand, the word embedding model predicted all three of them as Evidence mostly. Both models struggle with EvidenceBasedReasoning.

Test Set Evaluation

Classification report for best TFIDF Model on Test set:

```
TEST:
                                    precision    recall  f1-score   support

CaseFooter                              1.00      1.00      1.00        14
CaseHeader                              1.00      1.00      1.00        14
CaseIssue                               1.00      0.87      0.93        15
Citation                                0.96      0.98      0.97       263
ConclusionOfLaw                         0.80      0.80      0.80        30
Evidence                                0.67      0.97      0.79       583
EvidenceBasedOrIntermediateFinding      0.79      0.32      0.46       148
EvidenceBasedReasoning                  0.50      0.01      0.02       122
Header                                  0.95      1.00      0.98       142
LegalRule                               0.85      0.72      0.78       210
LegislationAndPolicy                    0.43      0.16      0.23        19
PolicyBasedReasoning                    0.00      0.00      0.00         2
Procedure                               0.96      0.76      0.85       141

accuracy                                                      0.79      1703
macro avg                                0.76      0.66      0.68      1703
weighted avg                             0.79      0.79      0.75      1703
```

Classification report for best Word Embedding Model on Test set:

```
TEST:
```

```
                                     precision    recall  f1-score   support

CaseFooter                               1.00      0.93      0.96        14
CaseHeader                               1.00      1.00      1.00        14
CaseIssue                                0.93      0.93      0.93        15
Citation                                 0.95      0.98      0.97       263
ConclusionOfLaw                          0.94      0.50      0.65        30
Evidence                                 0.64      0.96      0.77       583
EvidenceBasedOrIntermediateFinding       0.50      0.05      0.10       148
EvidenceBasedReasoning                   1.00      0.02      0.03       122
Header                                   0.97      0.99      0.98       142
LegalRule                                0.76      0.60      0.67       210
LegislationAndPolicy                     0.60      0.16      0.25        19
PolicyBasedReasoning                     0.00      0.00      0.00         2
Procedure                                0.78      0.88      0.83       141

accuracy                                                     0.75      1703
macro avg                                0.77      0.62      0.63      1703
weighted avg                             0.77      0.75      0.69      1703
```

Error Analysis

As stated before model has difficulty in differentiating between Evidence, EvidenceBasedReasoning and EvidenceBasedOrIntermediateFinding

sentence # 950 / case 0828636.txt / @1475
pred: Evidence / true: EvidenceBasedOrIntermediateFinding A private health care professional has linked pes planus of the right foot to the veteran's active service.

- This can very well be annotated as Evidence without the context of previous or upcoming sentences, attention based models might be helpful to overcome this.

sentence # 804 / case 0721357.txt / @7034
pred: Evidence / true: EvidenceBasedReasoning
As stated, the veteran never returned the two PTSD questionnaires that were sent to
him by the RO in June 2003 and December 2004, so these questionnaires are not on record.

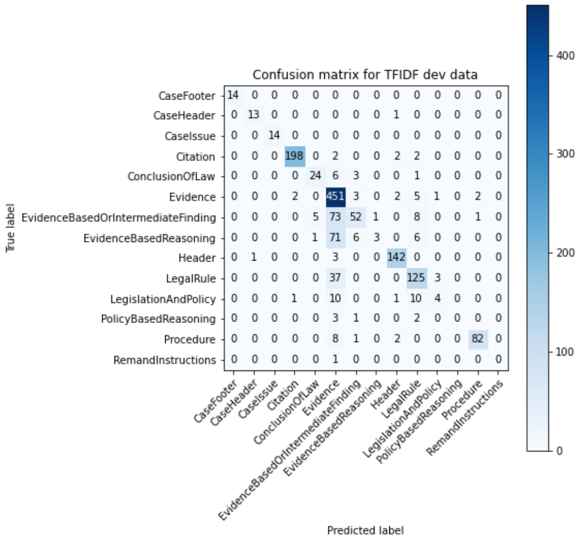
- The keyword here is ‘so’ without that last part of sentence even a human can’t infer this as EvidenceBasedReasoning. Making features combining such linguistic keywords that infer the essence of annotations can be helpful and differentiating character.

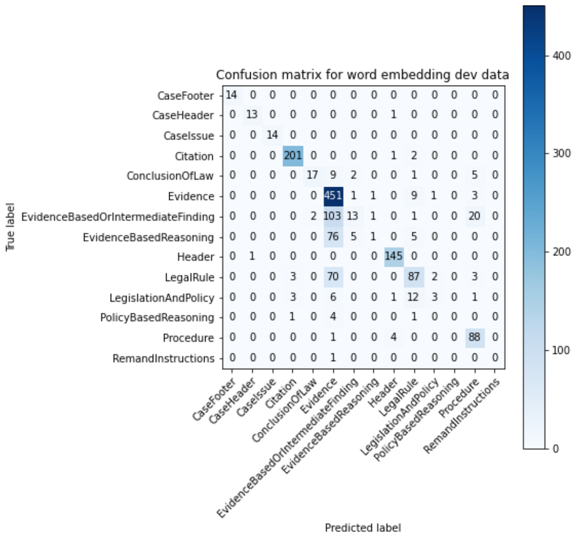
sentence # 427 / case 1317248.txt / @2947
pred: LegalRule / true: LegislationAndPolicy
As part of the notice, VA is to specifically inform the claimant and the claimant's representative, if any, of which portion, if any, of the evidence is to be provided by the claimant and which part, if any,

- LegislationAndPolicy are rare and difficult to annotate even as an expert as it requires acute knowledge as it can be seen as a LegalRule by a rookie and our model here acts like one.

sentence # 317 / case 1309811.txt / @1734
pred: Procedure / true: RemandInstructions
Those issues are referred to the RO for the appropriate development.

- As there are only 2 samples in the train data set it's impossible to learn about RemandInstructions if there were more samples I am sure the model would pick up on tokens like RO, referred etc.





Discussion

Segmenter:

- Using a law specific segmenter boosts performance.

Word Embedding:

- FastText, especially using the skip-gram model seems most beneficial.

Classifier:

- We get decent performance on a relatively small dataset with basic feature engineering, but struggle with very similar types like Evidence, EvidenceBasedReasoning etc. and Due to very low data on some types like PolicyBasedReasoning nothing can be inferred.

Recommendations:

- Maybe form stricter rules for annotating Evidence and other Evidence based types or not separate them if not necessary from a legal point of view.
- Obtaining more annotated data to improve model performance.
- More manual pre-processing especially with help of legal experts would help create and test much better embeddings.
- Deep learning models like LegalBERT etc. can be used to obtain better performance given large dataset
- Include a number of tokens in each sentence normalized as a feature with TFIDF features.

Lessons Learned:

- The course covered a wide variety of topics and served as a good way to get introduced to Legal NLP and AI
- Though without much Legal expertise it was difficult to make sense of the literature provided throughout the course and relate it to the project task sometimes.

Code Instructions

- Run analyze.py as intended by the project
- I have included the unlabeled dataset to access the BVA decisions from within the zip file I submitted to access the path to text file from within the particular directory of my submission
- PLEASE enter the path name in quotations
- Example run:
- Open command prompt
- Enter the following command: python analyze.py “unlabeled\0600090.txt”
- Code runs yay!

References

Bojanowski, Piotr, et al. “Enriching Word Vectors with Subword Information.” *arXiv*, 15 July 2016, https://arxiv.org/abs/1607.04606. Accessed 9 March 2022.

Savelka, Jaromir, et al. “Sentence Boundary Detection in Adjudicatory Decisions in the United States.” *Github*, 2017, https://github.com/jsavelka/luima_sbd.