Harvard John A. Paulson
School of Engineering
and Applied Sciences

Case Law Citation Prediction

Jiahui Tang, Yingchen Liu, Yujie Cai, Xin Zeng

Agenda

1. Introduction
2. EDAs - PCA analysis
3. Modeling
    a. Improved Baseline - LexNLP, Topic Modeling, Word2Vec
    b. Text Classification - FFNN / LSTM / BERT
    c. Graph Embedding - DeepWalk
    d. Our Proposed Model - LawPairBERT Model
4. Conclusion
5. Future Work

Introduction

In this project, we focused on processing case law and legal precedent data and performed an automatic relevant law citations generator.

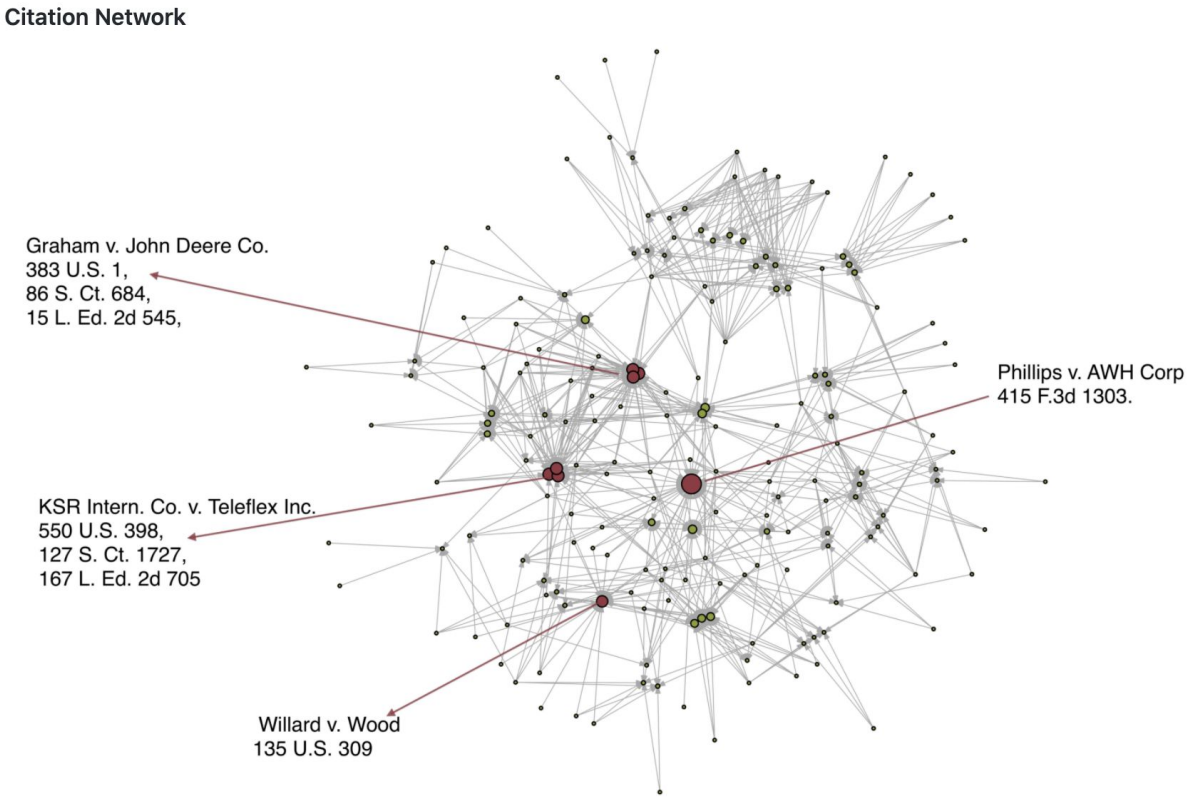


Image Source: https://github.com/YiAlpha/auto-law-review

Introduction - the dataset

	id	name	decision_date	court	citation	cites_to	author	type	text
0	8521088	JAMES L. OLIVER, JR., and CRAVEN VENTURE MANAG...	1980	North Carolina Court of Appeals	[{'type': 'official', 'cite': '49 N.C. App. 31...	[{'cite': '295 N.C. 733', 'case_ids': [8568681...	ERWIN, Judge.	majority	ERWIN, Judge.\nThe question presented for our ...
1	8564338	BRIAN FLIPPIN, by his Guardian ad Litem, MELVI...	1980	Supreme Court of North Carolina	[{'type': 'official', 'cite': '301 N.C. 108'}]	[{'cite': '116 N.W. 98', 'case_ids': [2600442]...	EXUM, Justice.	majority	EXUM, Justice.\nThis appeal presents two quest...

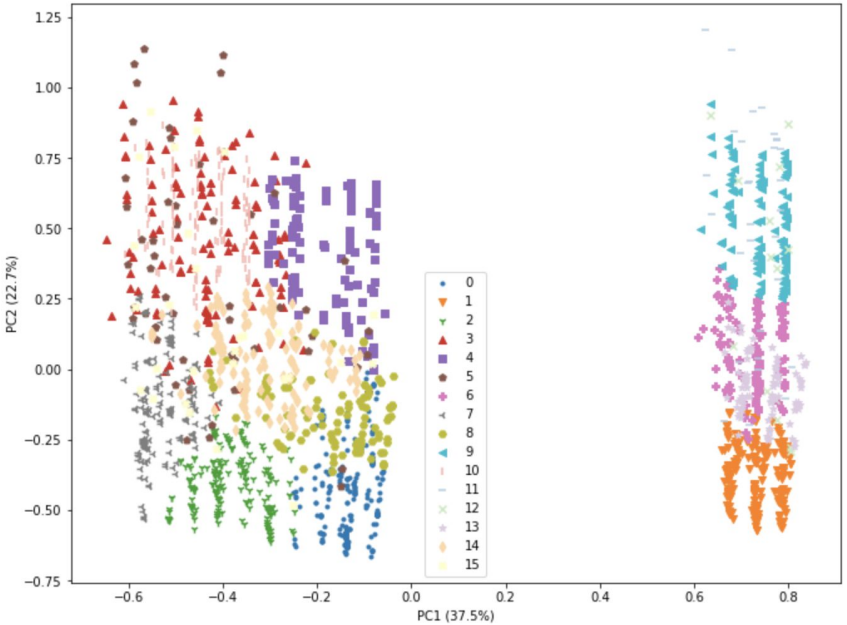
Case ID

The cases **cited by** the specific case

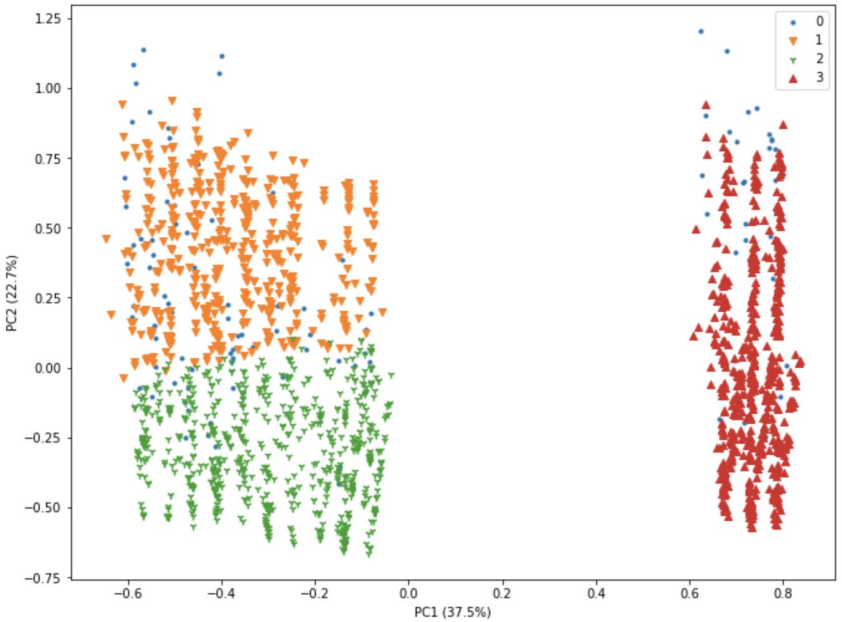
The opinion

PCA analysis

16 Clusters



4 Clusters



Information: Header contains some keywords including the names of judges, decision date, court, author and type

Baseline: Word2Vec

- Explored LexNLP extract and NLP features to pre-process data
- Build Topic Modeling and Clustered Legal Cases into Subcategories of Topics using *genism* library, NLTK and LexNLP, based on Bag Of Words
- Word2Vec Based on Law2Vec word embedding
    - find the most similar/closest law case document, use its citation result as our prediction result
    - self defined accuracy metrics (intersection of citation list/ union of citation list)

Topic Modeling

```
Topic: 0
Words: 0.010*"petition" + 0.009*"zone" + 0.008*"confess" + 0.007*"mortgag" + 0.006*"ordin" + 0.005*"treatme
```

Mortage

```
Topic: 1
Words: 0.009*"distribut" + 0.008*"contempt" + 0.008*"equit" + 0.006*"marit" + 0.005*"licens" + 0.005*"prisc
```

Marriage

```
Topic: 2
Words: 0.018*"petition" + 0.010*"leas" + 0.008*"default" + 0.006*"incom" + 0.005*"certif" + 0.005*"water" +
```

Lease, Default, Credit

```
Topic: 3
Words: 0.011*"disabl" + 0.010*"murder" + 0.008*"juror" + 0.005*"diseas" + 0.004*"prosecutor" + 0.004*"mitig
```

Disease and Disability

```
Topic: 4
Words: 0.007*"identif" + 0.006*"penalti" + 0.006*"intest" + 0.006*"juror" + 0.005*"truck" + 0.005*"photogra
```

Identification

```
Topic: 5
Words: 0.005*"murder" + 0.005*"conspiraci" + 0.004*"prison" + 0.004*"juror" + 0.004*"physician" + 0.004*"di
```

Murder, Prisoner

```
Topic: 6
Words: 0.013*"coverag" + 0.008*"testat" + 0.005*"loan" + 0.005*"dedic" + 0.005*"easement" + 0.005*"mutual"
```

Loan

```
Topic: 7
Words: 0.016*"search" + 0.010*"juvenil" + 0.009*"arbitr" + 0.005*"leas" + 0.004*"tenant" + 0.004*"warranti"
```

Lease and Tenant

Harvard John A. Paulsor
School of Engineering
and Applied Sciences

```
# select a sample
num = 7624
unseen_document = test_data.text[num]
print("====sample doc =====")
print(unseen_document)
```

```
====sample doc =====
TIMMONS-GOODSON, Judge.
Louis Ridgeway, Jr. (defendant) appeals from a judgment imposed upon his convictions of assault with a dea
The State's evidence presented at trial tended to show that in the early morning hours of 3 February 1996,
Angered that White had retrieved part of the money, defendant told Johnson to give him his gun. Johnson co
Unable to break White's hold on the gun, defendant called out to Johnson for help. Johnson, who had remair
Meanwhile, defendant was leaning against the car door holding the gun. He taunted White by repeatedly cock
At approximately 3:00 a.m. on the same morning, Officer Jessie Devane of the Fayetteville Police Departmer
White was brought into the emergency room shortly after defendant's arrival. Officer Michael Murphy, the c
At the close of the State's evidence, defendant moved to dismiss the charges against him. The court denied
By Ms first assignment of error, defendant contends that the trial court improperly admitted hearsay testi
Where, as here, a criminal defendant fails to object to the admission of certain evidence, the plain error
[T]he plain error rule ... is always to be applied cautiously and only in the exceptional case where, afte
,  (quoting United , 1002 (4th Cir. 1982)). Therefore, if after thorougMy examining the record, we are not
Defendant challenges the testimony offered by Officer Murphy wherein he stated that during his investigati
With his next assignment of error, defendant argues that the trial court erred in denying his motions to d
Upon a motion to dismiss, the question for the trial court is whether the State presented substantial evic
The defendant is guilty of an assault with a deadly weapon with intent to kill inflicting serious injury i
The essential elements of robbery with a dangerous weapon are: "(1) the unlawful taking or attempted takir
Viewing the evidence in the light most favorable to the State and drawing all reasonable inferences in its
For the foregoing reasons, we conclude that defendant received a fair trial, free of prejudicial error.
No error. ■
Judges GREENE and WALKER concur.
```

Prediction of Topics

- Prediction Result: Assault, Murder related

```
# Data preprocessing step for the unseen document
bow_vector = dictionary.doc2bow(preprocess(unseen_document))

for index, score in sorted(lda_model[bow_vector], key=lambda tup: -1*tup[1]):
    print("Score: {}\t Topic: {}".format(score, lda_model.print_topic(index, 10)))
```

```
Score: 0.9004489779472351        Topic: 0.012*"parent" + 0.007*"alimoni" + 0.006*"assault" + 0.006*"divorc"
Score: 0.07002376765012741       Topic: 0.020*"murder" + 0.009*"check" + 0.007*"sexual" + 0.006*"deliber" +
Score: 0.026460332795977592      Topic: 0.005*"murder" + 0.005*"conspiraci" + 0.004*"prison" + 0.004*"juror
```

Word2Vec

- Use the most similar/close document's citation list as our prediction result
- Accuracy Result not promising

```
============================
Min Distance is 0.026757875906723714
The case that matches most with current case is 18
predicted citation
['30 L.Ed. 2d at 433', '92 S.Ct. at 499', '357 A. 2', '434 U.S. 893', '417 U.S. 933', '235 N.W. 2d 581', '3
actual citation
['95 L. Ed. 2d 697', '652 A.2d 874', '518 S.E.2d at 215', '490 S.E.2d 569', '127 N.C. App. 426', '508 S.E.2
accuracy
0.0
============================
Min Distance is 0.11704468512102706
The case that matches most with current case is 75
predicted citation
['105 N. C., 411', '81 N. C., 106']
actual citation
['95 L. Ed. 2d 697', '652 A.2d 874', '518 S.E.2d at 215', '490 S.E.2d 569', '127 N.C. App. 426', '508 S.E.2
accuracy
0.0
============================
```

Text Classification

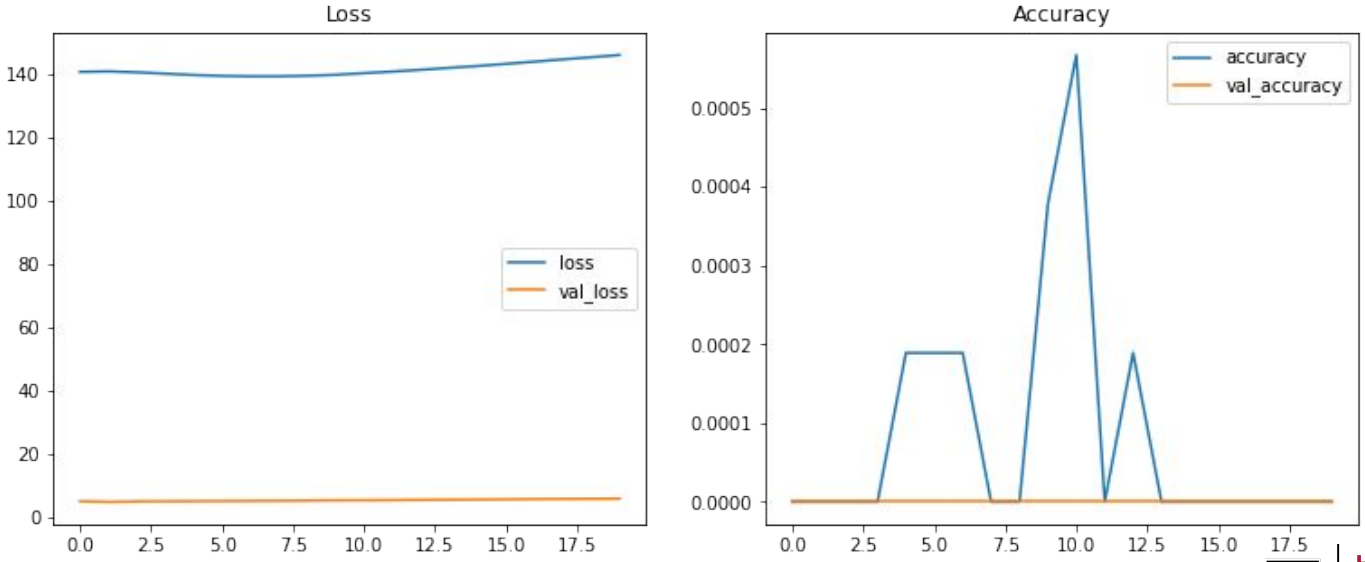
- Idea: text→ target (each citation in list) = supervised learning

- Including the following models:
  - **F**eed-**F**orward **N**eural **N**ets
  - **L**ong **S**hort-**T**erm **M**emory
  - **BERT**

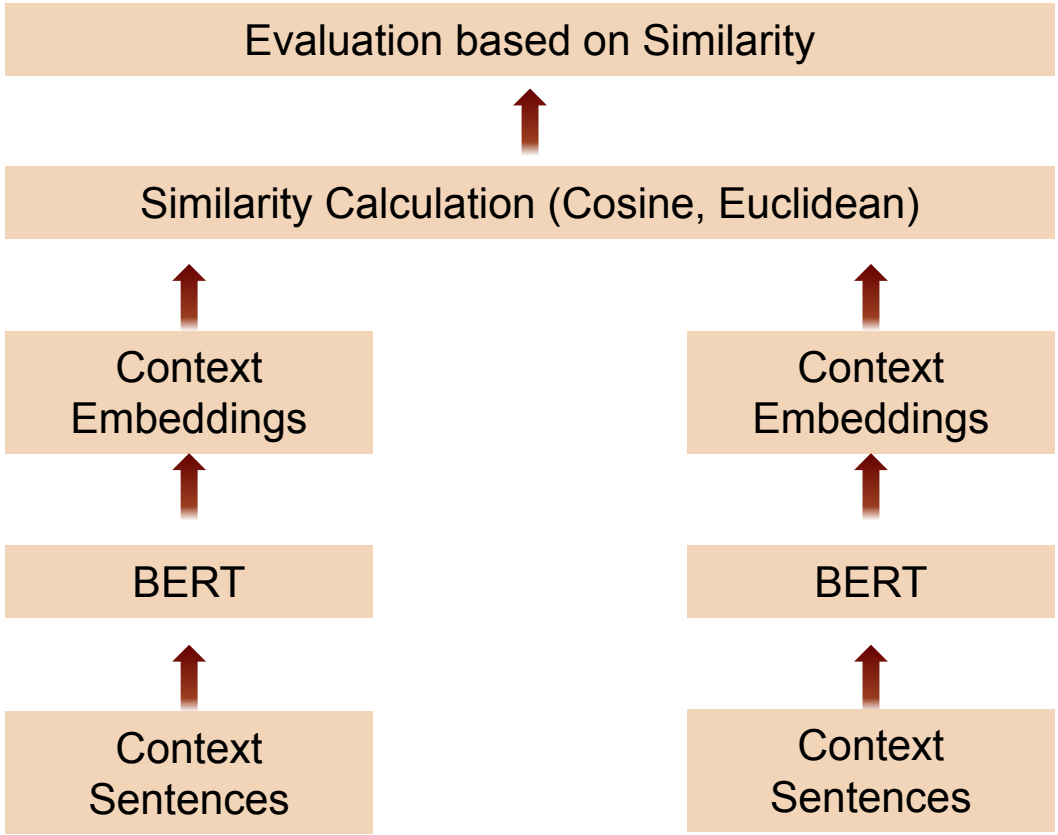
BUT WAIT ! 42534 different categories ?

Text Classification

Almost no accuracy and Computationally costly..



Architecture - BERT for Document Similarity



Concern

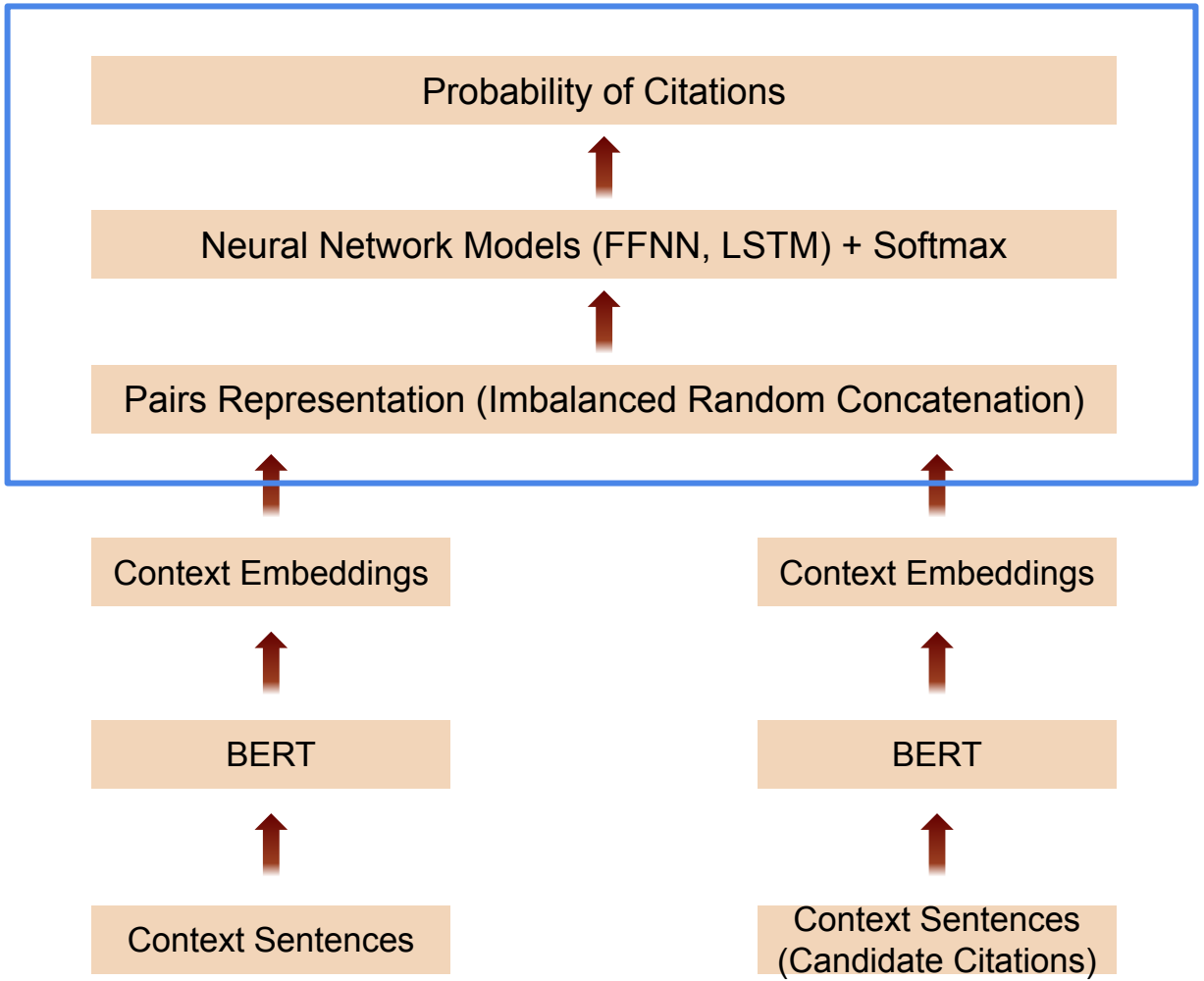
- Similar Document => Same Citations (Not a Valid Assumption)
- Low Accuracy
    - Top 1 similarity - 3.41%
    - Top 2 similarity - 5.71%
    - Top 3 similarity - 7.33%
- Limited Application
    - Vague boundary for classification
    - No difference between citations
    - High computational cost

Our Proposed Model - LawPairBERT

- BERT Embedding & Graph Embedding

- Advantages
  - Directly captures the citation relationship among each pair of law cases
  - High accuracy in predicting both true and false condition of citation
  - Reduce computational cost while remaining high accuracy

LawPairBERT

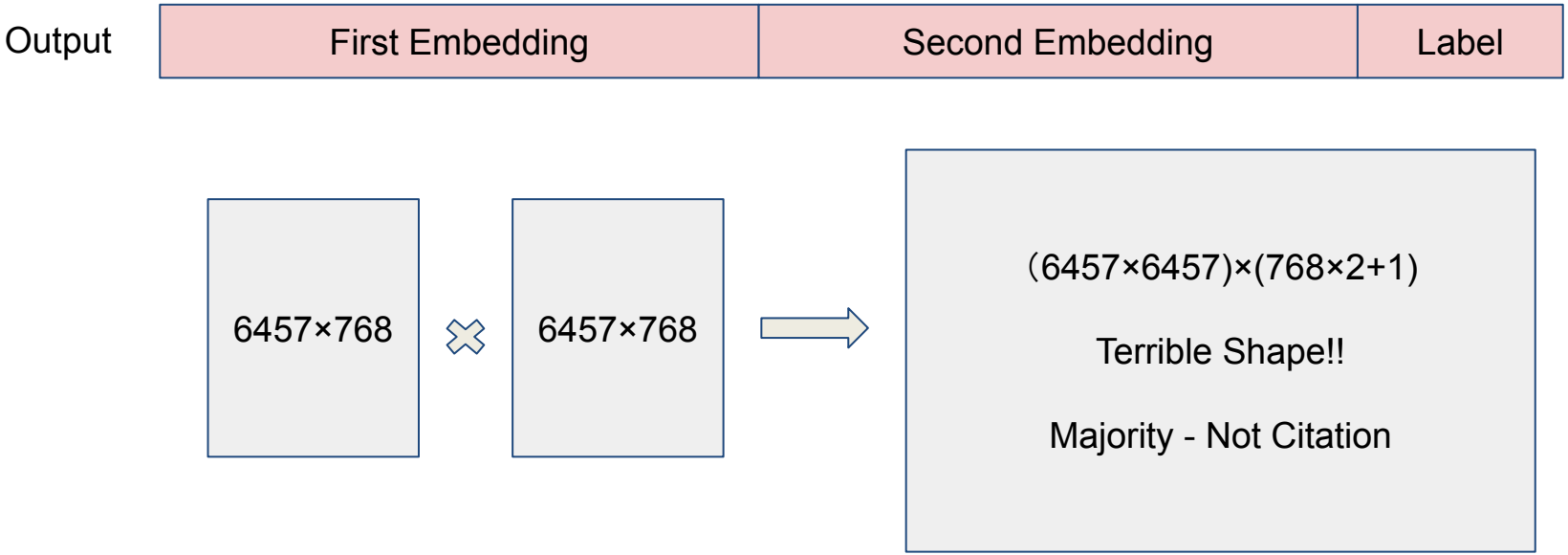


Pairs Representation (Imbalanced Random Concatenation)

Output	First Embedding	Second Embedding (Candidate Case)	Label

- Capture the citation relationship among any pairs of law cases
- Label: Whether the candidate case is the citation of the first case

Pairs Representation (Imbalanced Random Concatenation)



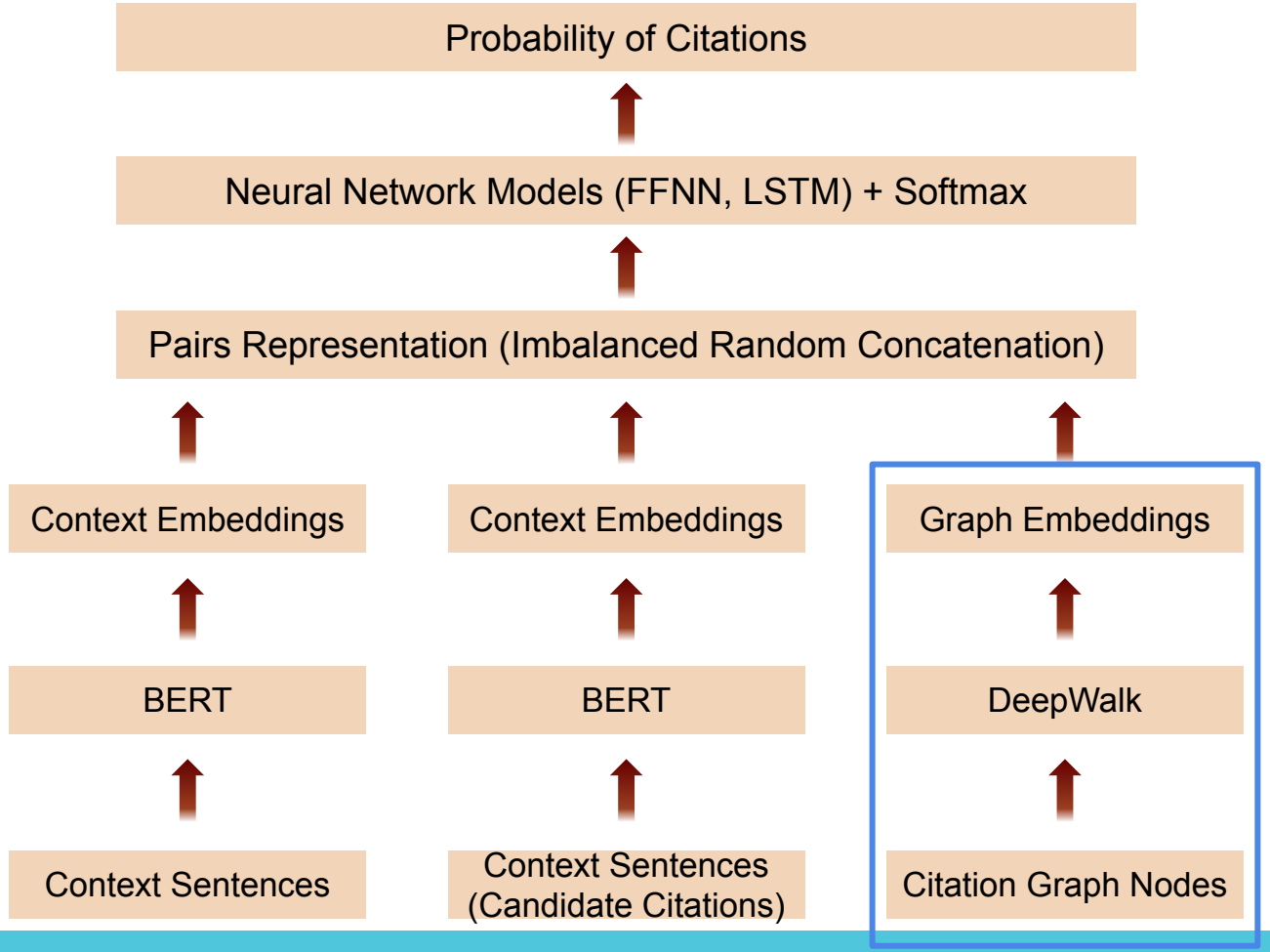
Pairs Representation (Imbalanced Random Concatenation)

- Include all pairs with citation relationship (label 1)
- Randomly select pairs without citation relationship (label 0)
    - Representativeness and variety
- Adjust size flexibly - 15000

Neural Network Models

- FFNN
- FFNN with self-defined early-stopping condition
    - Recall - improve accuracy of predicting citations correctly
- LSTM

LawPairBERT



Graph Embedding

- Idea: Obtain **Graph** representation of citation list
  - Edge and adjacency list shows relationship of citation
  - get embedding for further modeling

- DeepWalk (https://github.com/phanein/deepwalk)

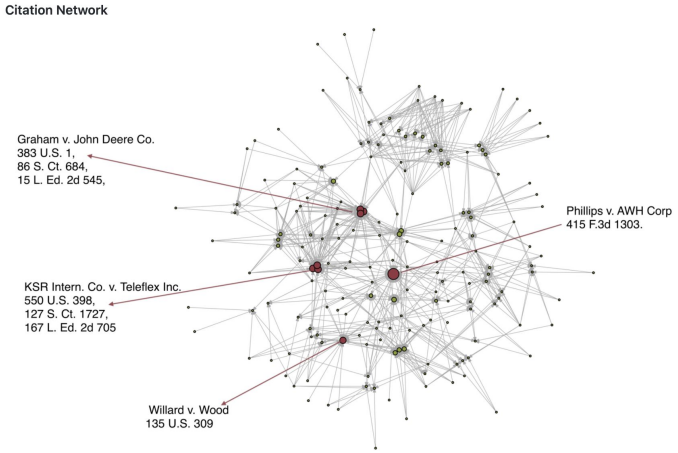
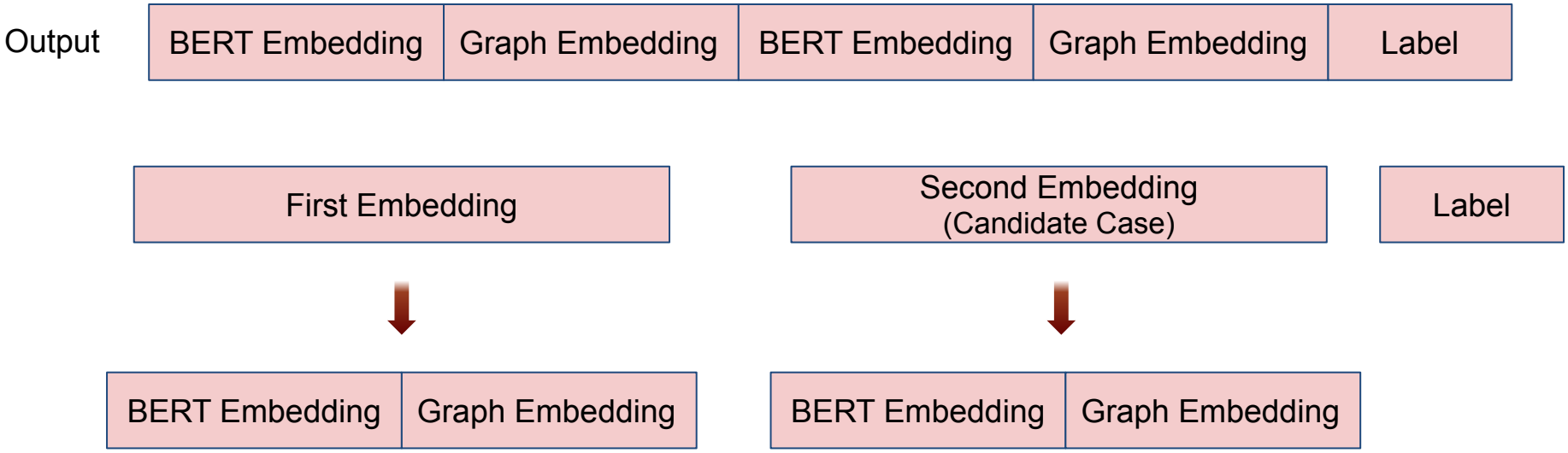


Image Source: https://github.com/YiAlpha/auto-law-review

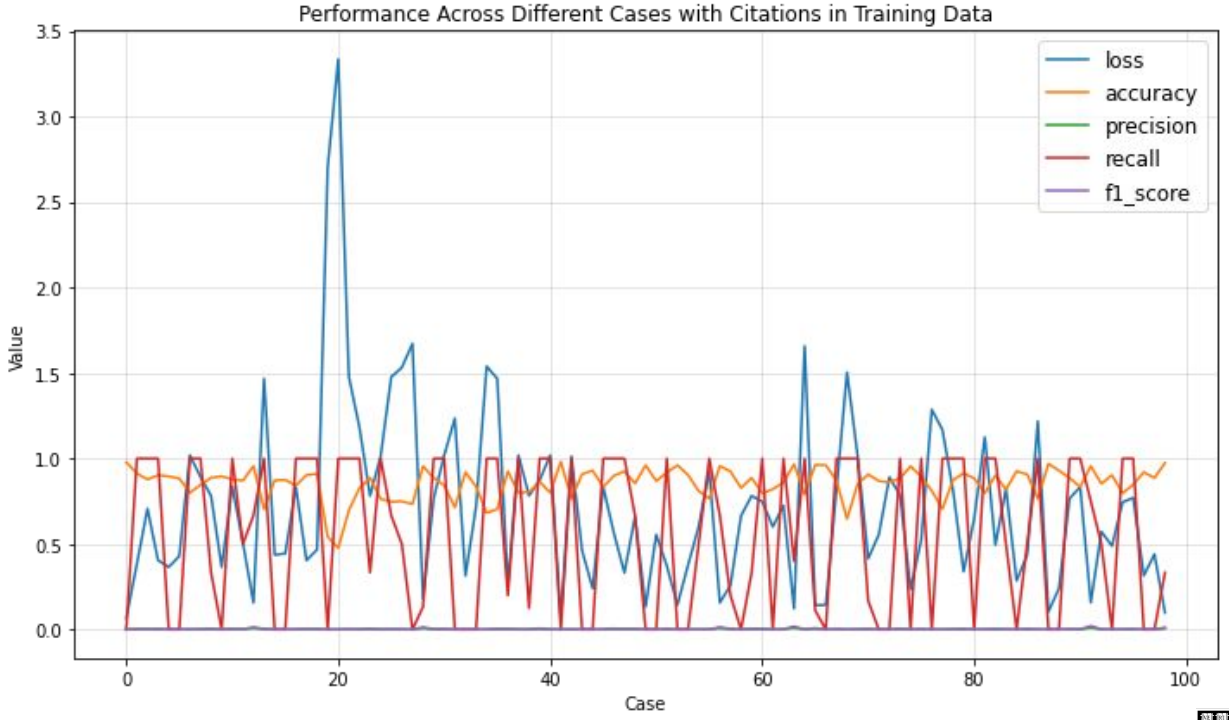
Pairs Representation (Imbalanced Random Concatenation)



LawPairBERT Model Performance

Model	Accuracy - Only True Citation (Test Data)	Overall Accuracy - Case 817 (4 true Citation)	Overall Recall - Case 817 (4 true Citation)
BERT - FFNN	0.4717	0.9178	0.5000
BERT - FFNN Recall	0.6488	0.8606	1.0000
BERT - LSTM	0.4449	0.9088	0.7500
BERT+Graph - FFNN	0.0967	0.9633	0.0000
BERT+Graph - FFNN Recall	0.0997	0.9382	0.0000
BERT+Graph - LSTM	0.1577	0.9572	0.2500

BERT - FFNN Recall Performance



Conclusion

- Different performance among different representations
  - Word - low accuarcy, high computational cost, hard to capture pair representation
  - LawPairBERT with BERT Embedding - high accuracy in true citation relationship
  - LawPairBERT with BERT and Graph Embeddings - high accuracy in true citation relationship

Future Work

- Try more Graph Embedding methods (GCN, SDNE)
- Validate and test LawPairBERT model across more states

Harvard John A. Paulson
School of Engineering
and Applied Sciences

The End

Thank you!