Label Analysis and Classification Using Random Forest for Named Entity Recognition

Nirel Amoyaw

Introduction

Trade Offs

Traditional machine learning models, like Random Forest, rely on handcrafted features and may struggle to generalize for complex datasets.

Transformer-based models, such as BERT, capture contextual relationships more effectively but are computationally expensive.

SAUDI TECH FUND
$40 BILLION PUSH INTO A.I. MARKET
PAGE 7 | BUSINESS

PAST PURVEYOR
GO-TO ACTOR FOR HISTORICAL ROLES
PAGE 13 | CULTURE

BASEBALL DIPLOMACY
SOUTH KOREANS EMBRACE DODGERS' JAPANESE STAR
PAGE TWO

The New York Times

INTERNATIONAL EDITION | THURSDAY, MARCH 21, 2024

One path forward for Gaza peace

Bret Stephens

OPINION

Sooner or later, the war in Gaza will end. Hamas's leaders hope that when it does, they will emerge from their tunnels to raise their green banners over the rubble — a symbolic victory for "Resistance" in the face of the misery they sowed on Oct. 7.

Israel's security leaders hope that hen it does, Gaza will be temporarily ided into a patchwork of subregions inistered by local clans known to li security services. The Israeli y will then operate in the territory an indefinite period on a counism mission, assume greater control along the border with Egypt and deradicalize the population.

President Biden hopes that "a revitalized Palestinian Authority" will return to govern the territory from which it was forcibly

How Paris keeps its spirit alive

New law may erode Hong Kong's status as hub

BEIJING

Security measures pass as economy is challenged and companies trickle out

BY KEITH BRADSHER

Paul Chan, the top finance official of Hong Kong, traveled to Paris, London, Frankfurt and Berlin last September to lure foreign investors. Last month, he abolished taxes on foreigners' purchases of Hong Kong real estate. And he is soon set to host an international art show, as well as conferences for big-money funds and advisers to wealthy families.

Mr. Chan's brisk work pace represents an attempt to shore up Hong Kong's role and image as the financial hub of Asia. But that effort is now colliding with a move by the city's Beijing-appointed leaders to further tighten their crackdown on the remaining political freedoms in the city.

Hong Kong's legislature approved broadly worded security legislation on Tuesday. City leaders described the law as necessary to stop foreign interference in local politics, but critics characterized it as a comprehensive effort to muzzle dissent.

Under its top leader

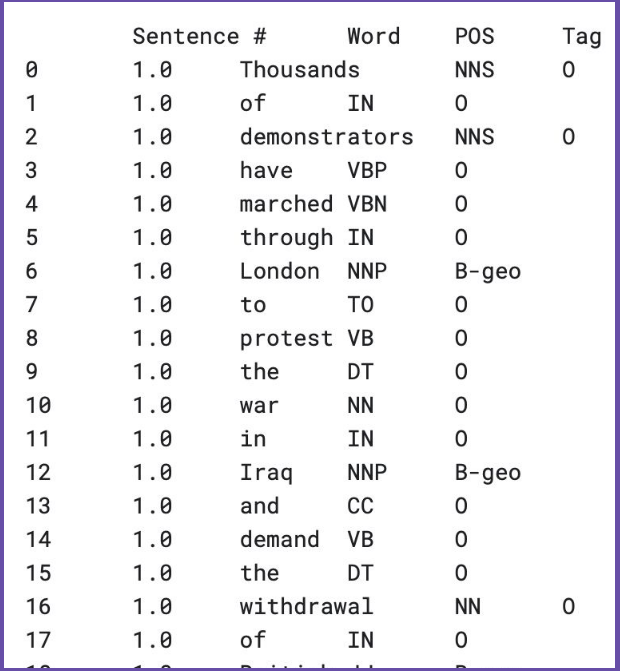
Dataset Composition

What does the dataset look like?

NER datasets use structured formats like BIO tagging (Begin, Inside, Outside) to indicate an entity's position in text.

The tagging is followed by the following classes:

geo = Geographical Entity
org = Organization
per = Person
gpe = Geopolitical Entity
tim = Time indicator
art = Artifact
eve = Event
nat = Natural Phenomenon

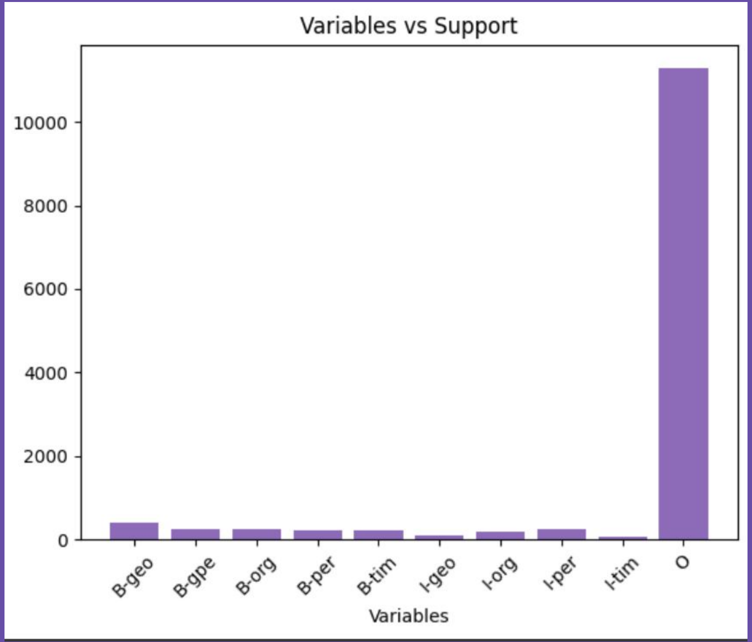


```
        Sentence #      Word    POS     Tag
0       1.0     Thousands       NNS     O
1       1.0     of      IN      O
2       1.0     demonstrators   NNS     O
3       1.0     have    VBP     O
4       1.0     marched VBN     O
5       1.0     through IN      O
6       1.0     London  NNP     B-geo
7       1.0     to      TO      O
8       1.0     protest VB      O
9       1.0     the     DT      O
10      1.0     war     NN      O
11      1.0     in      IN      O
12      1.0     Iraq    NNP     B-geo
13      1.0     and     CC      O
14      1.0     demand  VB      O
15      1.0     the     DT      O
16      1.0     withdrawal      NN      O
17      1.0     of      IN      O
```

Objective

Conduct a NER and **Random Forest** study to:

- Assess their performance on the **Annotated-GMB Corpus** for Named Entity Recognition.
- Identify strengths, weaknesses, and trade-offs in terms of **accuracy**, **computational efficiency**, and **scalability with varying features**.



Data

The Annotated GMB (Groningen Meaning Bank) Corpus

Location

Kaggle!

https://www.kaggle.com/datasets/shoumikgoswami/annotated-gmb-corpus/data)

The Annotated GMB (Groningen Meaning Bank) Corpus

Purpose

The dataset is specifically designed to train machine learning models for NER

The GMB corpus is a collection of annotated text data, where each word or token is labeled with its corresponding entity type or tag.

The Annotated GMB (Groningen Meaning Bank) Corpus

Size

The GMB corpus is considered fairly large (6,6162 rows and 5 columns), containing substantial annotated data.

The Annotated GMB Corpus

Limitations

The dataset is not perfect; it is considered a "silver standard" rather than a "gold standard."

Many annotations were initially generated by automated tools and later corrected by humans, meaning some inaccuracies might still remain.

Methods

Choosing Random Forest

- Random Forest serves as a non-deep learning baseline
- Random Forest relies on engineered features (e.g., word embeddings, character n-grams, POS tags), allowing me to evaluate the impact of feature design on NER performance.

- Random Forests tend to be more robust to slight variations in input data compared to deep learning models, which may rely heavily on context learned during pretraining

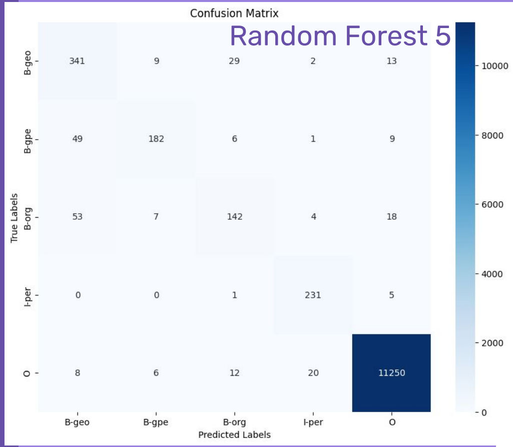
What did I do

- **I ran 4 instances of Random Forest**
  - **5 features**
  - **10 features**
  - **17 (all features)**
  - **16(without 0)**

How did I go about this?

- Words have non-linear relationships
  - 1. Evaluate Label distribution
  - 2. Multi-class classifier(random forest)
  - 3. Confusion Matrix Analysis
  - 4. Cluster Label Representations(Word2Vec word embeddings)

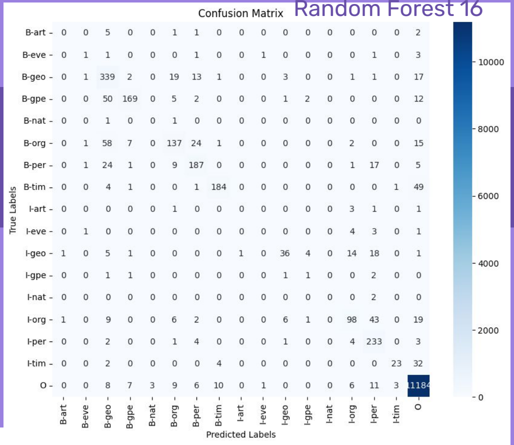
Confusion Matrices



Confusion Matrix

Random Forest 5

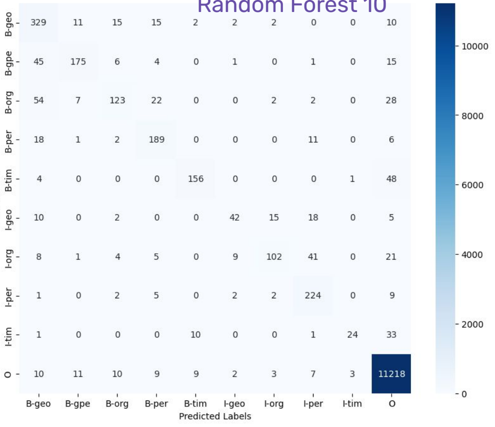
True Labels \ Predicted Labels	B-geo	B-gpe	B-org	I-per	O
B-geo	341	9	29	2	13
B-gpe	49	182	6	1	9
B-org	53	7	142	4	18
I-per	0	0	1	231	5
O	8	6	12	20	11250



Confusion Matrix

Random Forest 16

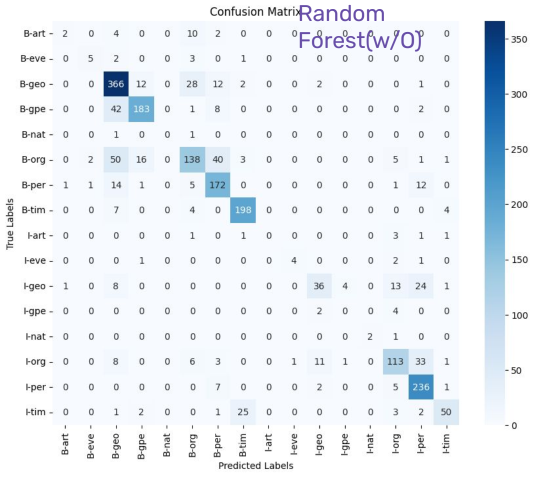
True Labels \ Predicted Labels	B-art	B-eve	B-geo	B-gpe	B-nat	B-org	B-per	B-tim	I-art	I-eve	I-geo	I-gpe	I-nat	I-org	I-per	I-tim	O
B-art	0	0	5	0	0	1	1	0	0	0	0	0	0	0	0	0	2
B-eve	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1	0	3
B-geo	0	1	339	2	0	19	13	1	0	0	3	0	0	1	1	0	17
B-gpe	0	0	50	169	0	5	2	0	0	0	1	2	0	0	0	0	12
B-nat	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
B-org	0	1	58	7	0	137	24	1	0	0	0	0	0	2	0	0	15
B-per	0	1	24	1	0	9	187	0	0	0	0	0	0	1	17	0	5
B-tim	0	0	4	1	0	0	1	184	0	0	0	0	0	0	0	1	49
I-art	0	0	0	0	0	1	0	0	0	0	0	0	0	3	1	0	1
I-eve	0	1	0	0	0	0	0	0	0	0	0	0	0	4	3	0	1
I-geo	1	0	5	1	0	0	0	0	1	0	36	4	0	14	18	0	1
I-gpe	0	0	1	1	0	0	0	0	0	0	1	1	0	0	2	0	0
I-nat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
I-org	1	0	9	0	0	6	2	0	0	0	6	1	0	98	43	0	19
I-per	0	0	2	0	0	1	4	0	0	0	1	0	0	4	233	0	3
I-tim	0	0	2	0	0	0	0	4	0	0	0	0	0	0	0	23	32
O	0	0	8	7	3	9	6	10	0	1	0	0	0	6	11	3	11184



Confusion Matrix

Random Forest 10

True Labels \ Predicted Labels	B-geo	B-gpe	B-org	B-per	B-tim	I-geo	I-org	I-per	I-tim	O
B-geo	329	11	15	15	2	2	2	0	0	10
B-gpe	45	175	6	4	0	1	0	1	0	15
B-org	54	7	123	22	0	0	2	2	0	28
B-per	18	1	2	189	0	0	0	11	0	6
B-tim	4	0	0	0	156	0	0	0	1	48
I-geo	10	0	2	0	0	42	15	18	0	5
I-org	8	1	4	5	0	9	102	41	0	21
I-per	1	0	2	5	0	2	2	224	0	9
I-tim	1	0	0	0	10	0	0	1	24	33
O	10	11	10	9	9	2	3	7	3	11218



Confusion Matrix

Random Forest(w/O)

True Labels \ Predicted Labels	B-art	B-eve	B-geo	B-gpe	B-nat	B-org	B-per	B-tim	I-art	I-eve	I-geo	I-gpe	I-nat	I-org	I-per	I-tim
B-art	2	0	4	0	0	10	2	0	0	0	0	0	0	0	0	0
B-eve	0	5	2	0	0	3	0	1	0	0	0	0	0	0	0	0
B-geo	0	0	366	12	0	28	12	2	0	0	2	0	0	0	1	0
B-gpe	0	0	42	183	0	1	8	0	0	0	0	0	0	0	2	0
B-nat	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
B-org	0	2	50	16	0	138	40	3	0	0	0	0	0	5	1	1
B-per	1	1	14	1	0	5	172	0	0	0	0	0	0	1	12	0
B-tim	0	0	7	0	0	4	0	198	0	0	0	0	0	0	0	4
I-art	0	0	0	0	0	1	0	1	0	0	0	0	0	3	1	1
I-eve	0	0	0	1	0	0	0	0	0	4	0	0	0	2	1	0
I-geo	1	0	8	0	0	0	0	0	0	0	36	4	0	13	24	1
I-gpe	0	0	0	0	0	0	0	0	0	0	2	0	0	4	0	0
I-nat	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0
I-org	0	0	8	0	0	6	3	0	0	1	11	1	0	113	33	1
I-per	0	0	0	0	0	0	7	0	0	0	2	0	0	5	236	1
I-tim	0	0	1	2	0	0	1	25	0	0	0	0	0	3	2	50

Results

Training Time

Random Forest 5: 47s

Random Forest 10: 9min 59sec

Random Forest 17: 10 min 0 sec

Random Forest 16(w/O): 33 sec

In-Depth Data

```
Random Forest Classification Report 5 features::
              precision    recall  f1-score   support

       B-geo       0.76      0.87      0.81       394
       B-gpe       0.89      0.74      0.81       247
       B-org       0.75      0.63      0.69       224
       I-per       0.90      0.97      0.93       237
           O       1.00      1.00      1.00     11296

    accuracy                           0.98     12398
   macro avg       0.86      0.84      0.85     12398
weighted avg       0.98      0.98      0.98     12398
```

```
Random Forest Classification Report (after filtering O):
              precision    recall  f1-score   support

       B-art       0.50      0.11      0.18        18
       B-eve       0.62      0.45      0.53        11
       B-geo       0.73      0.87      0.79       423
       B-gpe       0.85      0.78      0.81       236
       B-nat       1.00      0.00      0.00         2
       B-org       0.70      0.54      0.61       256
       B-per       0.70      0.83      0.76       207
       B-tim       0.86      0.93      0.89       213
       I-art       1.00      0.00      0.00         7
       I-eve       0.80      0.50      0.62         8
       I-geo       0.68      0.41      0.51        87
       I-gpe       0.00      0.00      0.00         6
       I-nat       1.00      0.67      0.80         3
       I-org       0.75      0.64      0.69       177
       I-per       0.75      0.94      0.84       251
       I-tim       0.85      0.60      0.70        84

    accuracy                           0.76      1989
   macro avg       0.74      0.52      0.55      1989
weighted avg       0.76      0.76      0.75      1989
```

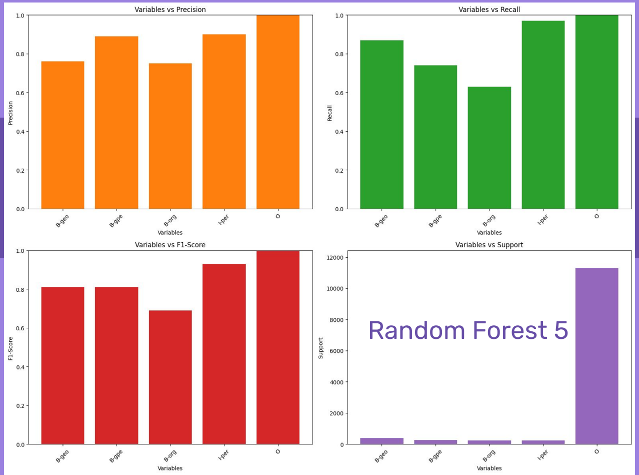
Random Forest 10 features

```
              precision    recall  f1-score   support

       B-geo       0.69      0.85      0.76       386
       B-gpe       0.85      0.71      0.77       247
       B-org       0.75      0.52      0.61       238
       B-per       0.76      0.83      0.79       227
       B-tim       0.88      0.75      0.81       209
       I-geo       0.72      0.46      0.56        92
       I-org       0.81      0.53      0.64       191
       I-per       0.73      0.91      0.81       245
       I-tim       0.86      0.35      0.49        69
           O       0.98      0.99      0.99     11282

    accuracy                           0.95     13186
   macro avg       0.80      0.69      0.72     13186
weighted avg       0.95      0.95      0.95     13186
```

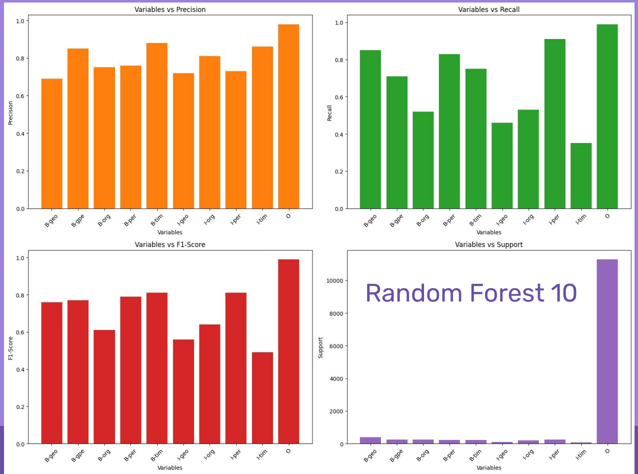
Random Forest 17 features

```
              precision    recall  f1-score   support

       B-art       0.00      0.00      0.00         9
       B-eve       0.20      0.12      0.15         8
       B-geo       0.67      0.85      0.75       397
       B-gpe       0.89      0.70      0.79       241
       B-nat       0.00      0.00      0.00         2
       B-org       0.72      0.56      0.63       245
       B-per       0.78      0.76      0.77       245
       B-tim       0.92      0.77      0.84       240
       I-art       0.00      0.00      0.00         6
       I-eve       0.00      0.00      0.00         9
       I-geo       0.75      0.44      0.56        81
       I-gpe       0.12      0.17      0.14         6
       I-nat       1.00      0.00      0.00         2
       I-org       0.74      0.53      0.62       185
       I-per       0.70      0.94      0.80       248
       I-tim       0.85      0.38      0.52        61
           O       0.99      0.99      0.99     11248

    accuracy                           0.95     13233
   macro avg       0.55      0.42      0.44     13233
weighted avg       0.95      0.95      0.95     13233
```

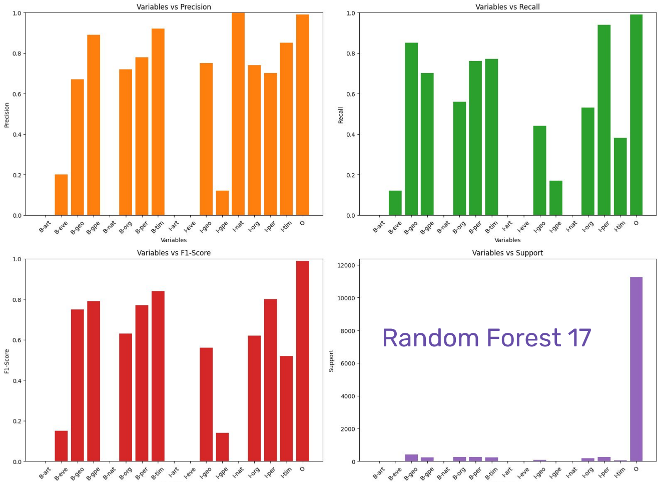
Precision, Recall, and F-1



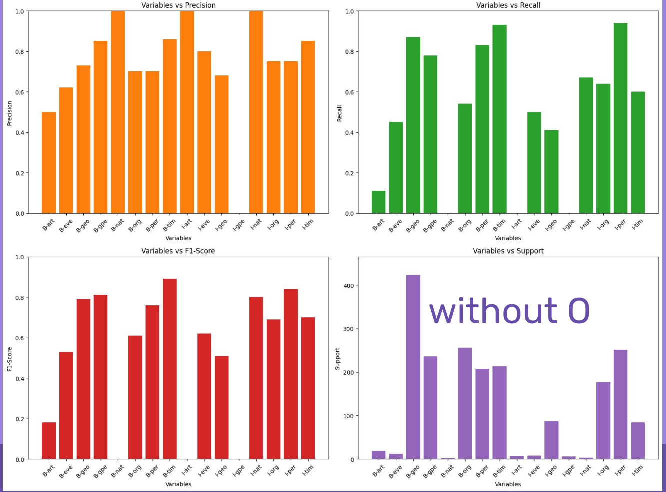
Random Forest 5



Random Forest 10

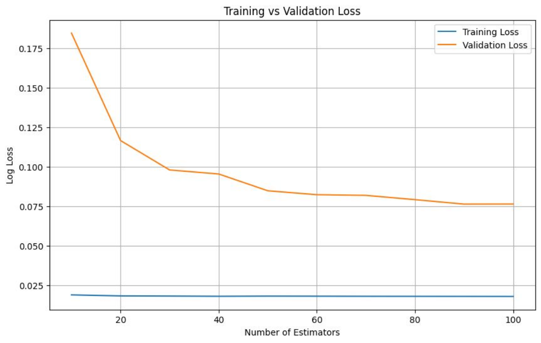


Random Forest 17

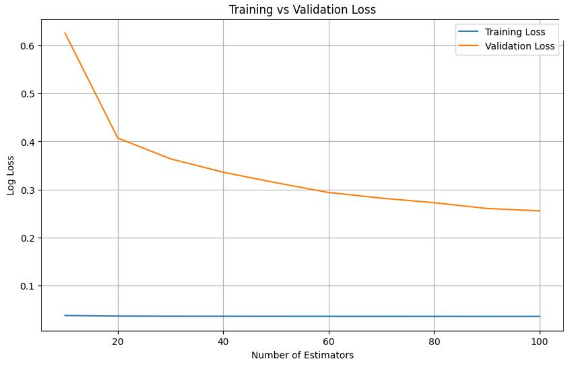


without O

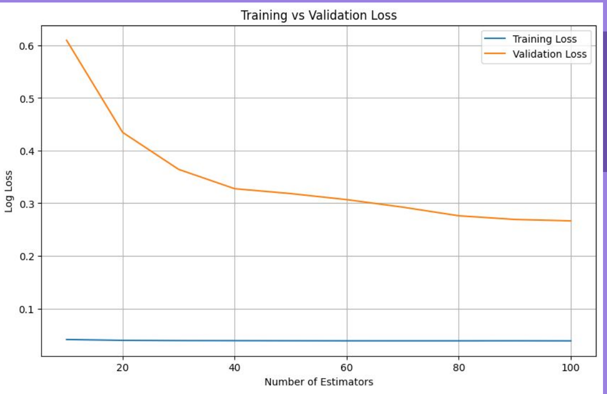
Training and Validation Loss



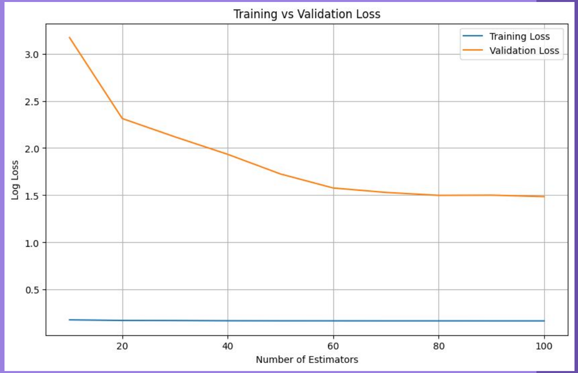
Random Forest 5



Random Forest 10



Random Forest 17



Random Forest 16 (w/O)

Learnings

What did I learn?

- What Named Entity Recognition is
- How to work with big datasets for NER
- Successfully completed a classification problem: word classification
- Implementing datasets from Kaggle
- How to Prepare sentences for NER
- word2features, sent2features, sent2labels
- Random forest on big dataset for NER
- Plotting statistical measures for analysis
- Being cautious of computational expense and maneuvering it tactically

What didn’t go as expected?

- Originally wanted to use BERT, but when trying to remove labels, I realized that the discrepancies I was seeing was because BERT is pretrained on a lot of data, so removing the labels isn’t doing much
- Ran out of CPU, had to give Google my $$$
- Lots of non-alignment errors (fixed by rerunning cells from scratch)

What would I do differently If I were to do this project again?

- I saw some other people online also using CRF and Bi-LTSM, maybe trying out those models
- Edit the data by taking account grouping words
- Try to improve accuracy by amending the ‘O’ column
- Work more with predicting sentences
- More different methods of choosing how to choose which labels to take into account

Limitations

- Needed more feature selection actions to account for non-linear features
- Dependence on the 'O' Label
- Computational Efficiency
- Contextual Dependencies
- etc.

Feature Count and Model Complexity

Performance Across Features:

- Models with 5 features had high accuracy (98%), but results were overly boosted by the "O" label.
- Adding more features (10 and 17) improved performance on common labels like B-geo and I-per, but rare labels (e.g., B-art, B-nat) still struggled.
- Removing the "O" label (16 features) lowered accuracy to 76% but improved representation for some rare labels like B-art and B-eve.

Interpretation of Results:

- Adding more features helps the model pick up subtle patterns, improving some rare labels.
- However, performance gains drop off after 10 features, showing that more complexity alone won't fix imbalance or mislabeling issues.

Scalability Considerations:

- Training time increases with features, from 47 seconds (5 features) to 10 minutes (17 features).
- Removing the "O" label reduced training time to 33 seconds, showing its impact on computation.

Class Specific Performance

Major Classes:

- Labels like B-geo and B-gpe consistently achieved higher F1 scores across all models due to their relatively larger representation in the dataset. However, even these classes saw a drop in recall when "O" was excluded, indicating that some instances of these entities were mislabeled as "O."

Minority Classes:

- Labels such as B-art, B-nat, and I-nat had near-zero recall and precision in all models, reflecting their severe underrepresentation in the dataset. Even when "O" was removed, these classes remained challenging to predict, suggesting that their features were not sufficiently distinct to be effectively learned by the model.

Inter-Class Confusion:

- Some entities, particularly those with overlapping features (e.g., B-org and I-org), had lower precision and recall. This indicates that the model struggled to distinguish between similar entity types, a challenge possibly due to the silver-standard nature of the dataset.

Conclusions

Conclusion

- **The "O" Label is a Double-Edged Sword**
- **Feature Selection Matters**
- **Class Imbalance Requires Targeted Solutions**
- **Contextual Features are Key**
- **Broader Implications**:
  - The challenges observed in this dataset are reflective of real-world NER applications in healthcare, legal documents, and social media, where imbalanced and noisy data are common. Insights from this work could inform better data preprocessing and model training practices in such domains

References

https://itrexgroup.com/blog/machine-learning-costs-price-factors-and-estimates/

https://www.geeksforgeeks.org/why-is-machine-learning-so-expensive-at-scale/

https://www.tomshardware.com/reviews/best-gpus%2C4380.html?utm_source=chatgpt.com

https://huggingface.co/google-bert/bert-base-cased

https://www.kaggle.com/code/shoumikgoswami/ner-using-random-forest-and-crf

random forest and NER

Thank You!