Assignment-Discussion
Named Entity Identification

Hemendra Meena(200050052,4th,CSE)

Nagesh Kumar(200050082,4th,CSE)

Vir Wankhede(210050166,3rd,CSE)

30 November, 2023

Problem Statement

- Perform Named-Entity Identification using SVM classifier with appropriate feature engineering
- **Technique to be used**: SVM classifier
- **Dataset**: CoNLL NER Data; https://paperswithcode.com/dataset/conll-2003 and https://huggingface.co/datasets/conll2003 (they are same data, but have common and distinct information)

Problem Statement

- **Input**: A sentence

- **Output**: Name-No Name tagged for each word in the sentence

- **Example**:
  - **Input**: Delhi is the capital of India.
  - **Output**: Delhi_1 is_0 the_0 capital_0 of_0 India_1 ._0

Data Processing Info (Pre-processing)

- The dataset was already in required format, so just converted data in Pandas dataframe
- Also there were issue like the columns were not in list format, like ['SOCCER' '-' 'JAPAN' 'GET' 'LUCKY' 'WIN' ',' 'CHINA' 'IN' 'SURPRISE' 'DEFEAT' '.'], where they were separated by uneven spaces and different quotations(double and single), so we used different operation to separate them.

Feature Engineering

- For features we used :-
  - Capitalisation - In this we distinguished words in three types :-
    - Only first letter as capital
    - All letter capital
    - Others
- POS Tag of word
- POS Tags for next two words
  - If word is last word word of sentence, we assigned its next two POS tags as ‘end’ POS tag
  - If word is second last word word of sentence, we assigned its next second POS tags as new category as ‘end’ POS tag
- POS Tags of previous two words
  - If word is first word word of sentence, we assigned its previous two POS tags as ‘begin’ POS tag
  - If word is second word word of sentence, we assigned its previous second POS tags as new category as ‘begin’ POS tag

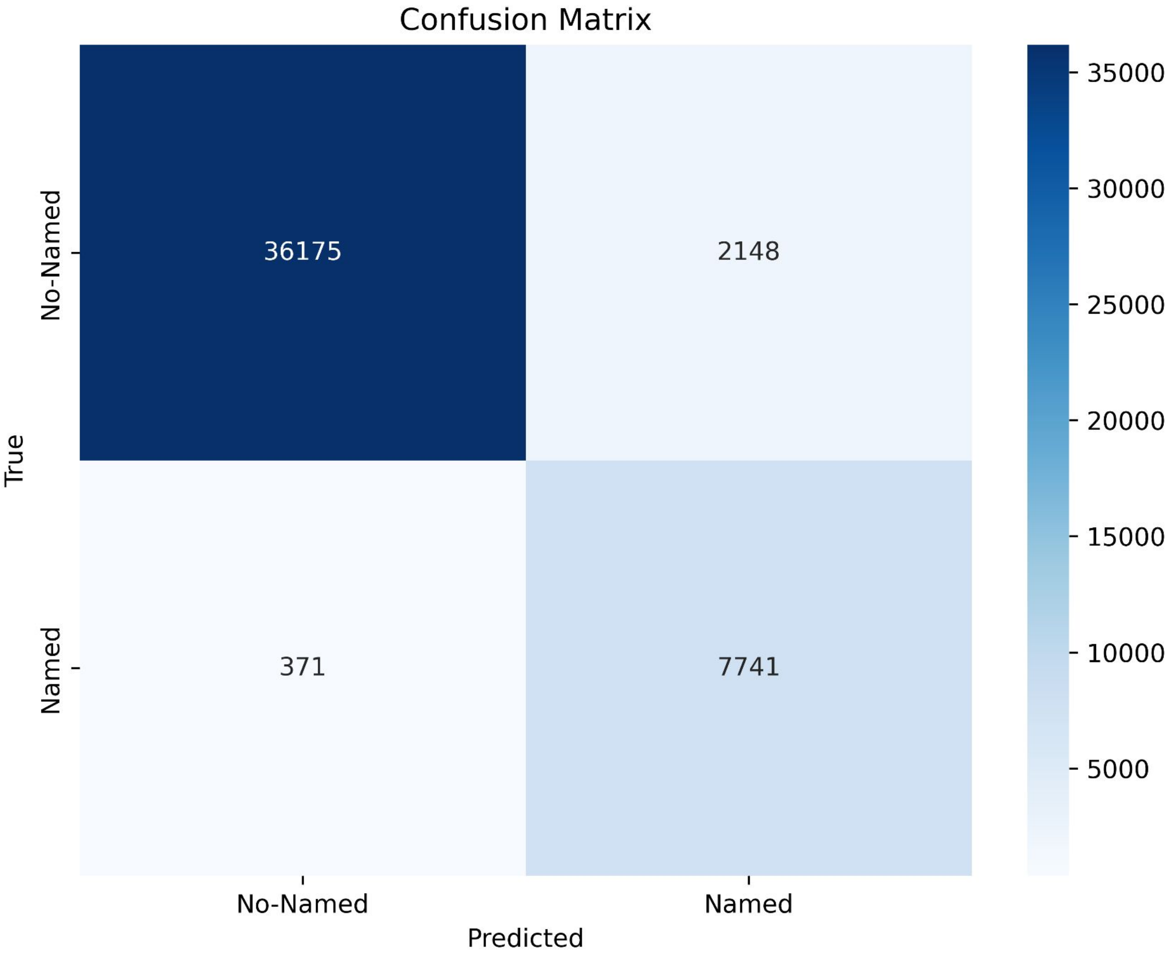
SVM Modelling

- We used ‘sklearn.svm.SVC’ Support Vector Classification.
- We trained it on three different kernels, viz. Linear, RBF and Polynomial.

Overall performance

	Precision	Recall	F1-Score	Support
No-Named	0.99	0.94	0.97	38323
Named	0.78	0.95	0.86	8112
Accuracy			0.95	46435
Macro avg	0.89	0.95	0.91	46435
Weighted avg	0.95	0.95	0.95	46435

Confusion Matrix



Error analysis

- In case of False Positives of Name Identity, there were cases of words with all capital letter like ‘I’, ‘A’, etc. words as Title having all capital letters.
- There were also cases of first word of sentence as no-named identity, which has first letter capital, being predicted as Named-Identity.
- In case of False Positives of No-Named Identity, there were words with initial letter as small like ‘trans-Atlantic’, ‘pro-Israeli’, etc.
- Most of the False positive seem to be due to the **capitalisation** feature.

Learnings

During the assignment we gained valuable insight importance of different features for Named-Entity Recognition(NER)

- **POS Tags** : These help in capturing the syntactic pattern of the sentence. The next and previous POS tags helped model in capturing the contextual relation of word with neighbouring words and contributed in significantly increasing the accuracy of model.
- **Capitalisation** : Though it emerged as strong feature which helped in increasing the accuracy of model significantly but at the same time it resulted in problem in predicting the initial words of sentence.

Thank You