NER – Named Entity Recognition

Github URL- https://github.com/TheDevCarnage/NLP-Project-NER.git

My Github (my contributions/progress)- https://github.com/asb1996/NER-Project_NLP.git

What is NER?

Named entity recognition (NER) is a fundamental task in Natural Language Processing (NLP) and one of the first stages in many language understanding tasks.

Problem We are Solving?
We are using eBay listing titles for NER. A few examples of NER labeling of listing titles are shown below (these examples are in English to illustrate the concept, the challenge data will have German language listing titles).

About Data

- 10 million randomly selected unlabeled item titles from eBay Germany, “Athletic Shoes” categories.
  - 10,000 labeled item titles provided.
- Context
  - "New" tagged as "No Tag" in "New shoes", but "Marke" (brand) in "New Balance".
- Misspellings
- No Tag
  - Punctuation, and words adding no meaning (black & white), prepositions.
- Obscure
  - Un-deciphered

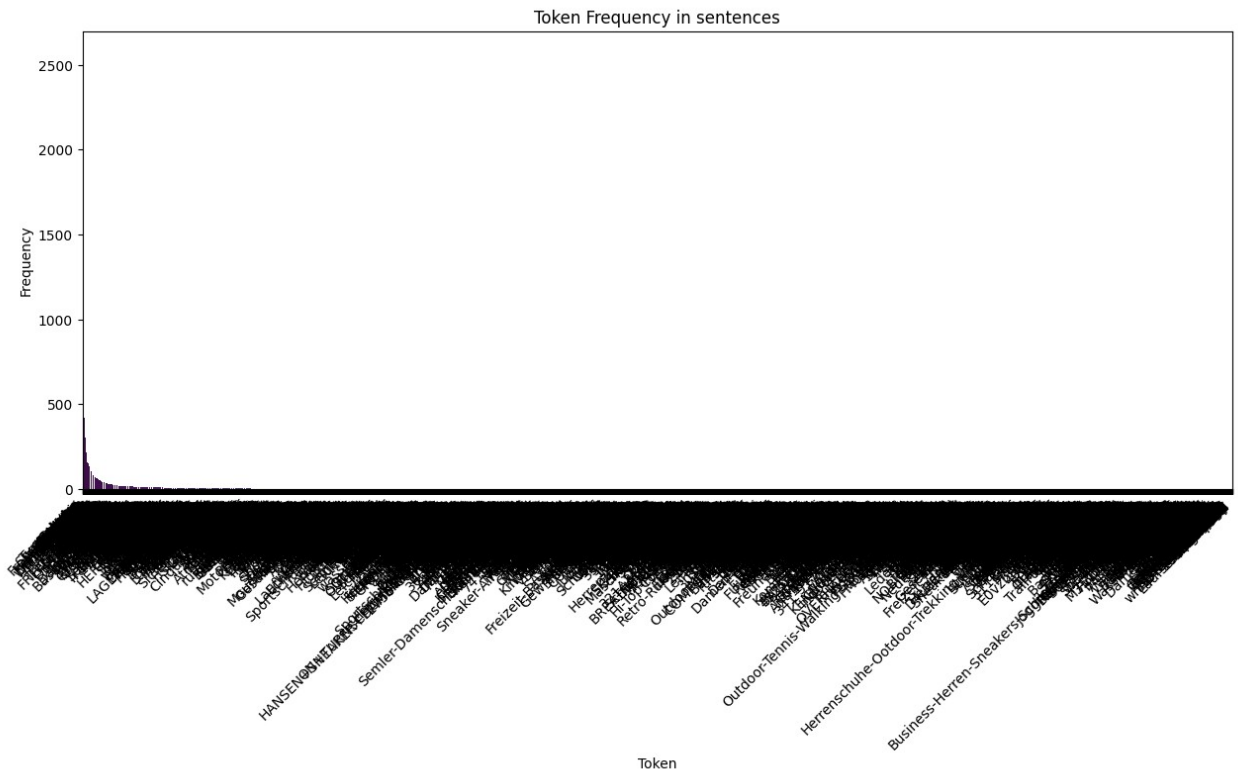
***You *can* refer to the Annexure.pdf file for more information on the dataset.**

Sample

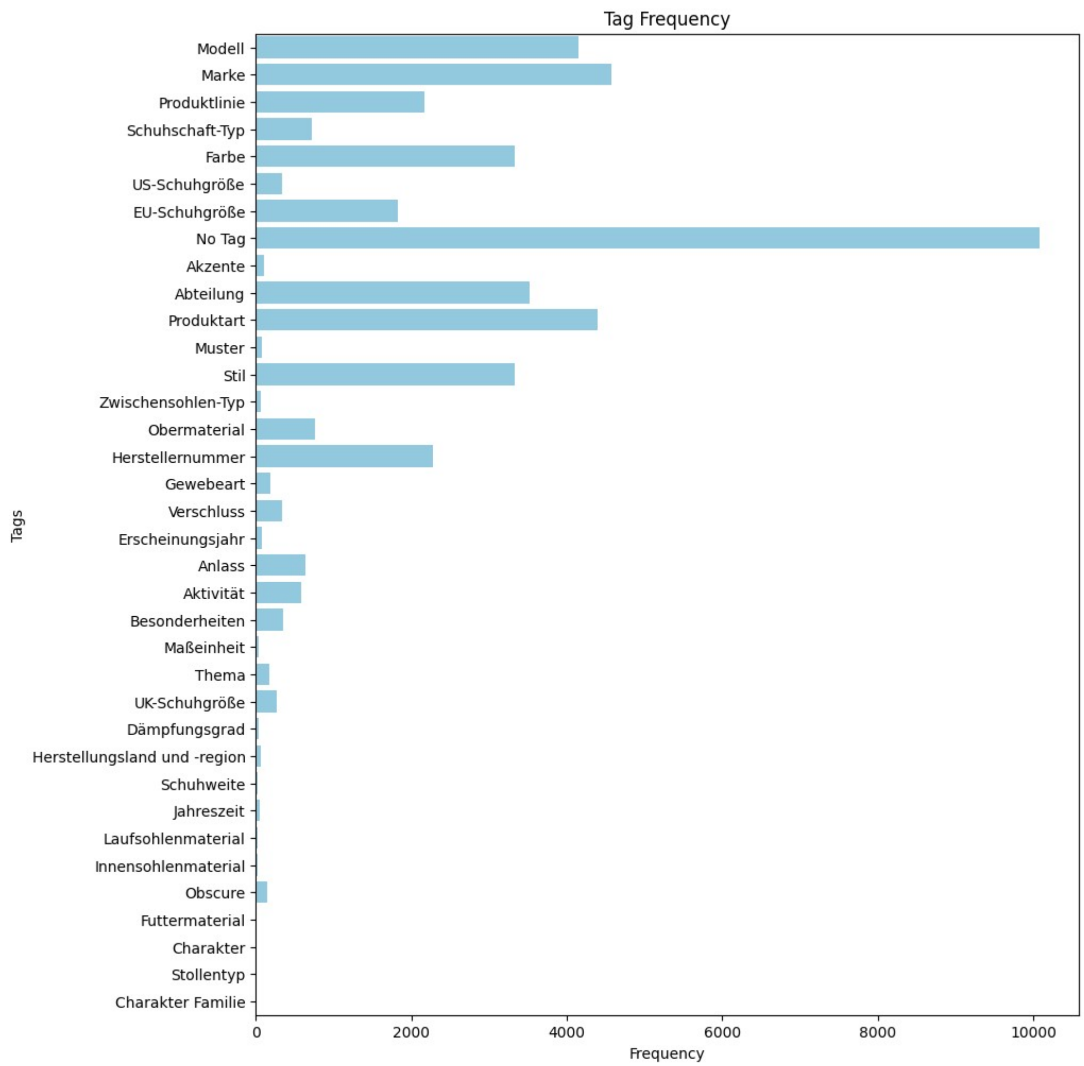
Record Number	Title	Token	Tag
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Supreme	Modell
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Nike	Marke
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	SB	Produktlinie
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Dunk	
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	High	Schuhschaft-Typ
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	By	Modell
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	any	
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Means	
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Red	Farbe
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	US10	US-Schuhgröße
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	EU44	EU-Schuhgröße
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Supreme	No Tag
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Box	No Tag
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Logo	Akzente
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Air	Produktlinie
1	Supreme Nike SB Dunk High By any Means Red US10 EU44 Supreme Box Logo Air Force	Force	

Data Visualization

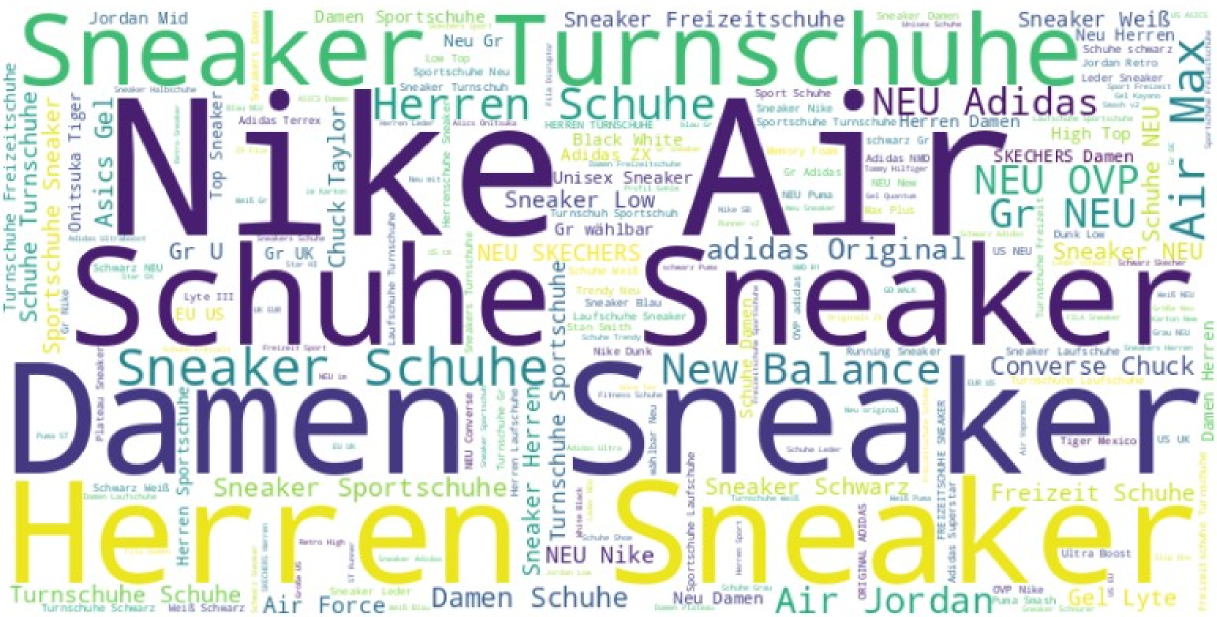
We visualized data to understand what features need to be extracted. Some techniques were better than others. Below a frequency chart for tokens was a bad approach so we pivoted towards word cloud. Following it you will find the word cloud and frequency chart for tags and after that the word cloud for all the tokens in the title.


Token Frequency in sentences
Frequency
2500
2000
1500
1000
500
0
Token
HANSEN
Semler-Damenschuhe
Sneaker-
Freizeit
Retro-
Outdoor-Tennis-Walking
Herrenschuhe-Outdoor-Trekking
Business-Herren-Sneakers


Sneaker
Herren
Schwarz
NEU
Gr
Herren Sneaker
Nike Air
Schuhe Sneaker
Sneaker Turnschuhe
Damen Sneaker
Damen
Weiß
Sneaker Schuhe
Adidas
Leder
Black
Laufschuhe
adidas Original
New Balance
Herren Schuhe
Air Max
Schuhe
Grau
Blau
White
Gr NEU
Puma
Turnschuhe Schuhe
Air Jordan
Herrenschuhe
Freizeitschuhe
Onitsuka Tiger
Herren Sportschuhe
Air Force
Fila
Converse Chuck
Halbschuhe
NEU Adidas
Gel Lyte
Sneaker Sportschuhe
Chuck Taylor
Classic
Converse
Damenschuhe
Damen Schuhe
Retro


Tag Frequency
Modell
Marke
Produktlinie
Schuhschaft-Typ
Farbe
US-Schuhgröße
EU-Schuhgröße
No Tag
Akzente
Abteilung
Produktart
Muster
Stil
Zwischensohlen-Typ
Obermaterial
Herstellernummer
Gewebeart
Verschluss
Erscheinungsjahr
Anlass
Aktivität
Besonderheiten
Maßeinheit
Thema
UK-Schuhgröße
Dämpfungsgrad
Herstellungsland und -region
Schuhweite
Jahreszeit
Laufsohlenmaterial
Innensohlenmaterial
Obscure
Futtermaterial
Charakter
Stollentyp
Charakter Familie
Tags
0
2000
4000
6000
8000
10000
Frequency


Modell nan
nan nan
Tag
Produktart nan
Stil Produktart
Tag EU
Marke Produktlinie
Tag Marke
EU Schuhgröße
nan Modell
Abteilung Produktart
Produktlinie nan
Marke Modell
Abteilung Stil
Schuhgröße nan
nan Abteilung
Tag Farbe
Farbe nan



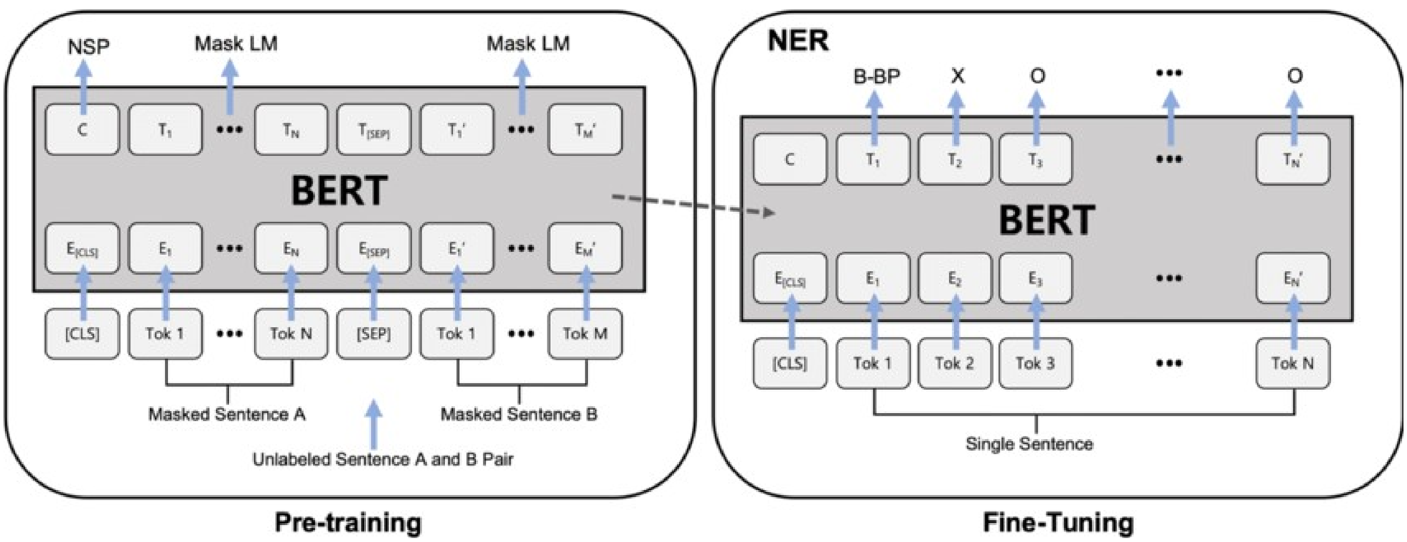
(Tokens in Title Word Cloud)

Pre-Processing Steps

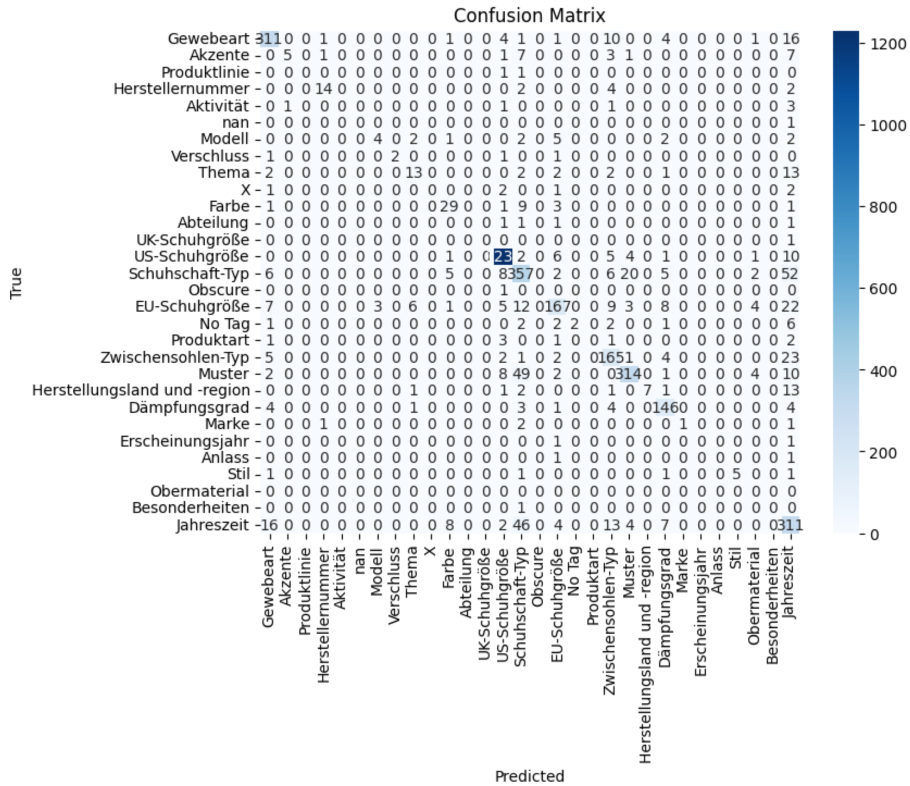
- Tokenization
- Attention Mask (Specific to Bert)
- Padding/Truncation
- Punctuation
- Label Encoding
- BERT Specific Input Format

```
tokenizerbert = BertTokenizer.from_pretrained('bert-base-uncased')
tokensbert= tokenizerbert(''.join(modelsteptokens), return_tensors='pt')
input_ids = tokensbert['input_ids']

attention_mask = [1] * len(input_ids)

label_ids = [tokenizerbert.convert_tokens_to_ids(tag) for tag in modelsteptags]
label_attention_mask = [1 if label_id != tokenizerbert.pad_token_id else 0 for label_id in label_ids]

input_ids_tensor = input_ids.unsqueeze(0)
attention_mask_tensor = torch.tensor([attention_mask])
label_ids_tensor = torch.tensor([label_ids])
label_attention_mask_tensor = torch.tensor([label_attention_mask])
```

Modeling with BERT



We used 'bert-base-german-cased' to tokenize and pre train the model with our dataset from Train_Tagged_Titles.tsv. Following is the confusion matrix and the F1 scores. Also attached with the submission will be a folder I wrote to to store the accuracy/recall/f1_score information. I have updated the confusion matrix from the project presentation day therefore it is different. However due to lack of time I was unable to plot the validation loss graph.


Confusion Matrix
True
Predicted
Gewebeart
Akzente
Produktlinie
Herstellernummer
Aktivität
nan
Modell
Verschluss
Thema
X
Farbe
Abteilung
UK-Schuhgröße
US-Schuhgröße
Schuhschaft-Typ
Obscure
EU-Schuhgröße
No Tag
Produktart
Zwischensohlen-Typ
Muster
Herstellungsland und -region
Dämpfungsgrad
Marke
Erscheinungsjahr
Anlass
Stil
Obermaterial
Besonderheiten
Jahreszeit
0
200
400
600
800
1000
1200

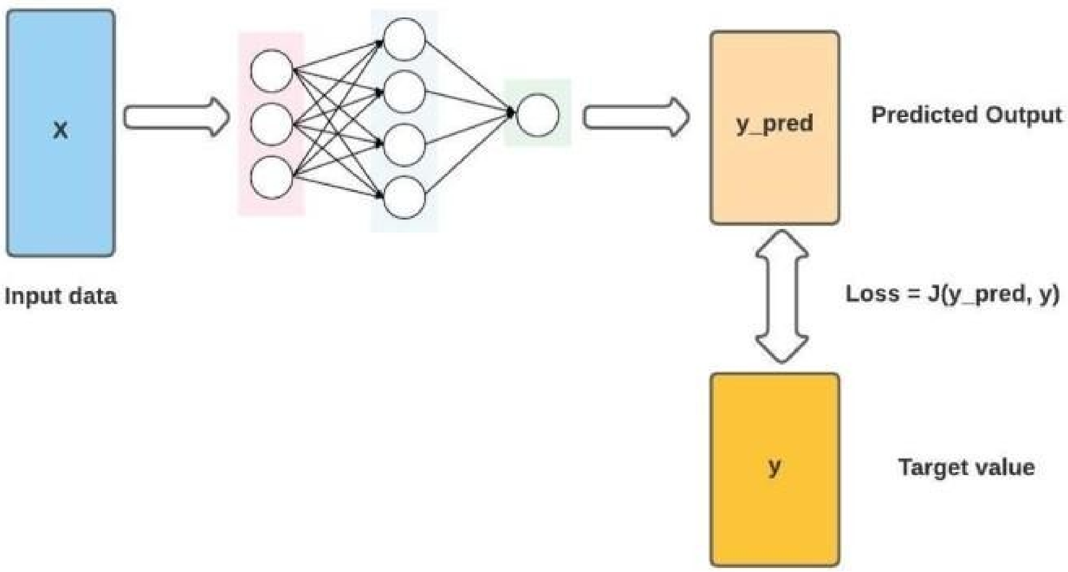
***** Eval results *****

	precision	recall	f1-score	support
Schuhgröße	0.5714	0.1818	0.2759	22
Typ	1.0000	0.2593	0.4118	27
_	0.0000	0.0000	0.0000	0
ahreszeit	0.0000	0.0000	0.0000	1
an	0.5796	0.6354	0.6062	384
arbe	0.5714	0.3429	0.4286	35
arke	0.9652	0.9706	0.9679	1257
bteilung	0.8592	0.8592	0.8592	348
erschluss	1.0000	0.5556	0.7143	9
erstellernummer	0.6000	0.5455	0.5714	44
esonderheiten	0.0000	0.0000	0.0000	6
ewebeart	0.0000	0.0000	0.0000	6
hema	1.0000	0.2000	0.3333	5
ktivität	0.8333	0.2000	0.3226	25
kzente	0.0000	0.0000	0.0000	2
nlass	0.8235	0.6364	0.7179	22
o Tag	0.7771	0.6239	0.6921	218
odell	0.7038	0.7251	0.7143	462
region	0.0000	0.0000	0.0000	4
roduktart	0.7214	0.7143	0.7178	203
roduktlinie	0.9012	0.7969	0.8458	389
rscheinungsjahr	1.0000	0.4000	0.5714	5
til	0.8022	0.8957	0.8464	163
uster	0.0000	0.0000	0.0000	1
ämpfungsgrad	0.0000	0.0000	0.0000	1
micro avg	0.8262	0.7994	0.8126	3639
macro avg	0.5484	0.3817	0.4239	3639
weighted avg	0.8247	0.7994	0.8063	3639

f1 socre: 0.812570
Accuracy score: 0.827429

Cross entropy loss is given in BERT Model is given by

$$\ell(x, y) = L = \{l_1, \dots, l_N\}^\top, \quad l_n = -w_{y_n} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})} \cdot 1 \cdot$$



Below is a picture of our poster from project presentation: