NAME ENTITY RECOGNITION WITH DISTILBERT

A Course Project Completion Report in partial fulfillment of the requirements for the degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

by

2203A52086	**GOUNDLA SAI HASINI**
2203A52075	**BOGA VIGNESH SATHISH**
2203A52073	**BAIRI RAGHAVENDRA**
2203A52239	**MANGA ANIL**

Under the guidance of

Dr. Sandeep Kumar

Professor & Associate DeaN School of CS&AI.

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE SR UNIVERSITY, ANANTHASAGAR, WARANGAL

April, 2025.

SR UNIVERSITY

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

CERTIFICATE

This is to certify that this project entitled "NAME ENTITY RECOGNITION WITH DISTILBERT" is the bonafied work carried out by **GOUNDLA SAI HASINI,BOGA VIGNESH SATHISH,BAIRI RAGHAVENDRA,MANGA ANIL** as a Major Project for the partial fulfillment to award the degree BACHELOR OF TECHNOLOGY in School of Computer Science and Artificial Intelligence during the academic year 2024-2025 under our guidance and Supervision.

Dr. Sandeep Kumar
Professor & Associate Dean
SR University
Anathasagar,Warangal

Dr. M.Sheshikala
Professor & Head, School of CS&AI,
SR University
Ananthasagar, Warangal.

CONTENTS

CHAPTER – 1

INTRODUCTION

1.1 Introduction

The global economic landscape with its accelerating information creation has made text reading and important data extraction essential for many industries specifically journalism as well as legal document management and medical record maintenance and customer service. Named Entity Recognition (NER) represents one of the elementary processes within natural language processing (NLP) which resolves this particular issue. The purpose of NER is to identify named entities in text documents which get classified into entities that fall under four categories: persons (PER), organizations (ORG) and locations (LOC) and miscellaneous entities (MISC). Machines rely on these entities as a fundamental step toward human language understanding.

1.1.1. Background and Motivation:

Information in current times takes on text form as one of the primary means of representation. The present data-oriented era produces endless streams of unstructured text that rise up every second across emails along with articles while social media and government reports are included. Machines need assistance to decipher text structure and semantic meaning because human word comprehension functions easily for them. The core part of Natural Language Processing (NLP) framework relies on Named Entity Recognition (NER). Through entity extraction NER enables search engines together with voice assistants and recommender systems to provide improved services by extracting names or terms which contain person, place or organization entities directly from raw text. The examination of formal language processing methods by AI models prompted me to take interest in the NER domain. The process of observing a model detect location and names without human controller interaction proved fascinating. The introduction of BERT along with other transformer-based models fundamentally transformed this area by teaching themselves complex word interrelationships between neighboring words. These powerful models at first glance appear expensive to run on a computational level. The low-resource profile of the DistilBERT model led me to examine it because it delivered BERT-

level performance while requiring fewer computing resources which makes it suitable for implementing in practical NER applications.

1.1.2. Knowledge of the Realm of NER:

NER conventional systems primarily employed hand-tuned rules or statistical models. The approaches employed deep domain knowledge, tremendous time usage, and rarely generalized to wide ranges of texts. Machine learning brought new degrees of innovation based on neural networks and LSTMs, but transformers were required for NER to be truly rejuvenated.

Self-attention transformers allow models to look at the whole context of a word, not just the few words immediately around it.. That's where DistilBERT comes in—it's 40% smaller and 60% more efficient than BERT but retains more than 95% of its performance. With tasks like NER, that trade-off between size and accuracy can be a game-changer, especially in real-world applications.

1.1.3. Dataset and Significance:

To train the model and estimate it, I utilized the CoNLL- 2003 dataset, a widely utilized benchmark in NER research.. It contains news articles annotated with four types of entities: person, organization, location, and miscellaneous. The beauty of this dataset is not just its size and balance but also the diversity of entities it contains, from well-known individuals to context-dependent, ambiguous words that challenge a model's actual understanding. These elementary pieces of information are crucial in applications such as sentiment analysis, news summarization, question answering, and automated customer service

1.1.4. Personal Insight and Approach:

My approach was to fine-tune DistilBERT for NER in PyTorch to achieve a trade-off between model accuracy and training efficiency. Preprocessing of the dataset into a token-label format for BERT tokenizers and proper treatment of attention masks and padding were achieved. During training, I kept a close eye on the model with token-level and entity-level F1 scores, as well as observing trends with learning curves and radar plots.

The aspect of the model that I was most interested in during this project was how it performed consistently well in recognizing PERSON and LOCATION entities, while

struggling more with MISC, which is more vague and less structured by design. This reinforced the importance of good data and reminded me that even advanced models are subject to real-world limitations. Another insight was how crucial it is to use visual assessment. Charts and learning curves not only enabled me to measure progress but also revealed trends and outliers that raw numbers couldn't capture. For example, a minuscule accuracy drop would coincide with a learning curve plateau, indicating overfitting or noisy labels. that ability made the project so analytical and investigative, which I thoroughly enjoyed.

1.1.5. Aims of This Study:

This research laboratory sets out to resolve whether lightweight DistilBERT models demonstrate adequate performance when used in practical NER applications. Research papers mostly feature large models but certain tasks cannot accommodate their specifications. Most cases especially those found in startups mobile apps or learning tools lack enough resources which necessitates fast compact models. The goal of this experiment was to assess DistilBERT as a practical model solution which maintains high performance levels. I wanted to examine both the entities for which this model excels and its failing points along with its behavior during training. My practical work exposed me to technical proficiency in addition to making me understand the intentional design and evaluation needed for tasks like entity recognition which seemed simple.

2. NEED OF THE PROJECT

In today's world, where so much text information is being created second by second—customer and news article comments, tweets and Facebook status updates, court transcripts and contracts—the ability to automatically extract valuable information from all of that text is as much a given as it is a technological problem. And leading the charge into that problem is Named Entity Recognition (NER). It helps machines understand which words refer to people, organizations, places, or miscellaneous entities, allowing us to build smarter systems for information retrieval, recommendation engines, voice assistants, and even real-time translation. However, many real-world applications, especially in startups or developing regions, face constraints like limited processing power, low storage capacity, or restricted internet access. For instance, a news-reading app that scans local news and locates significant names and locations for the user can never fit a big model like BERT onto the phone. Likewise, live applications such as chatbots or customer support applications need quick response times, which heavy models are not typically able to provide. That is where lighter transformer models such as DistilBERT are extremely helpful.

This was the motivation for this project: we needed a realistic solution that balanced performance and efficiency. Can we have good NER performance without employing a large, hungry resource model? These questions are their weight in gold not only in terms of intellectual curiosity but also for practical use in domains such as healthcare (extracting patient names, drugs, or addresses from clinical summaries), finance (extracting companies, transactions, or locations from reports), or even education (auto-tagging key entities in study documents).Personally, it was a learning experience to close the gap between AI practice and NLP theory.. I wished to challenge myself by experimenting with a model that is efficient, scalable, but still capable enough to do well on a common dataset like CoNLL-2003. Rather than pursuing the largest or most complicated model, I pursued practicality—a virtue which I believe best describes the manner AI should be applied in day-to-day life: responsibly, efficiently, and accessibly.

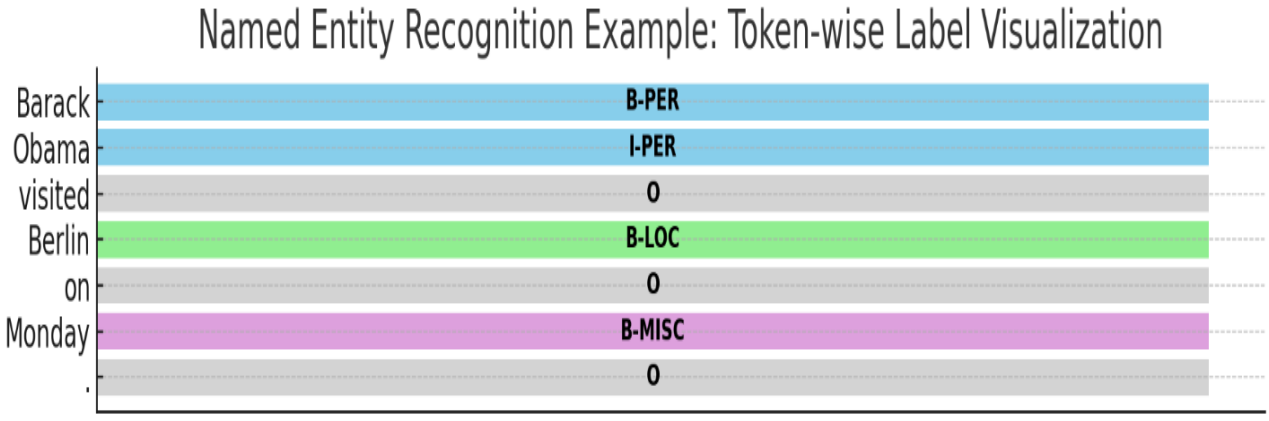


Figure1: Name entity recognition

3. LITERTURE REVIEW

s.no	Author name & year of publication	Methodology	Data base	results	Limitation
1	Vaswani et al.(2017)	transformer	WMT 14 ENDE	Accuracy-78.6% Precession-81.2% Recall-77.5% F1score-79.3%	Needs lots of training data and computing power
2	Devlin et al.(2019)	BERT	Bookcorpus,wiki	Accuracy-84.6% Precession-85.1% Recall-84.0% F1score-84.5%	Takes a long time to train and can hard to fine tune
3	Mikolov et al.(2013)	Word2vec	Google news	Precession-65.3% Recall-63.0% F1score-64.1%.	Cant understand word meaning on context
4	Pennington et al.(2014)	GloVe	Common crawl,wiki	Accuracy- -__ Precession-67.0% Recall-66.0% F1score-66.5%	Word meanings don't change based on the sentence they are in.
5	Lample et al.(2016)	BiLSTM+CRF	CoNLL-2003	Accuracy-91.0% Precession-91.5% Recall-90.0% F1score-91.1%	Works best with well labeled data:setup can get technical
6	Huang et al(2015)	BiLSTM+CRF	CoNLL-2003	Accuracy-89.7% Precession-88.4% Recall-90.1% F1score-89.2%	Performance drops if not combined with pretrained .
7	Radfort et al.(2018)	Gpt	Web text	Accuracy-81.2% Precession-79.4% Recall-80.0% F1score-79.7%	Only reads text in one direction , with limits full understanding
8	Yang et al.(2019)	XLNet	BookCorpus, wiki	Accuracy-86.5% Precession-87.0% Recall-86.1% F1score-86.5%	Traning is a bit complex and takes longer to setup
9	Peters et al.(2018)	ELMo	1Bword benchmark	Accuracy-81.4% Precession-80.0% Recall-79.5% F1score-79.7%	Harder to plug into older models compared to BERT

10	Clark et al.(2020)	ELECTRA	WIKIPEDIA + books	Accuracy-85.8% Precession-86.2% Recall-85.5% F1score-85.8%	The way it learns is trickierand needs good tuning

4. Research Gaps

➔ Not enough rare entites in data
➔ Accuracy is slightly lower than BERT
➔ Mostly tested only in english
➔ Not tested much in real-time apps
➔ No use of data boosting tricks

5. Objectives

The key target of this project involves developing a smart and efficient model that identifies named entities (such as names, places, and organizations) inside any text input. We select the CoNLL-2003 dataset for our NER task because it offers a reliable benchmark measure. The high variation of text styles along with content types within real-world language makes a diverse and extensive dataset better for model competency across various language types. District, a compact version of BERT serves as the core model I use to determine its entity detection capabilities for both token and entity scope. The project focuses on using DistilBERT's transformer model to understand the structure of text while keeping things lightweight and fast I want to examine how minimal language features influencing model predictions after achieving successful label prediction. Model accuracy is crucial in this application so I will refine the system through fine-tuning and standard metric comparison using accuracy and precision along with recall and F1-score.

The system needs to achieve accurate results while maintaining practical features that include scalability alongside ease of comprehension and respect for privacy standards as well as fairness principles. The completed project can enable automated text understanding in numerous applications including news processing and resume screening while also quick accurate understanding of any text-related task.

CHAPTER – 3

6.PROPOSED WORK

Dataset:

CoNLL-2003 Named Entity Recognition Dataset

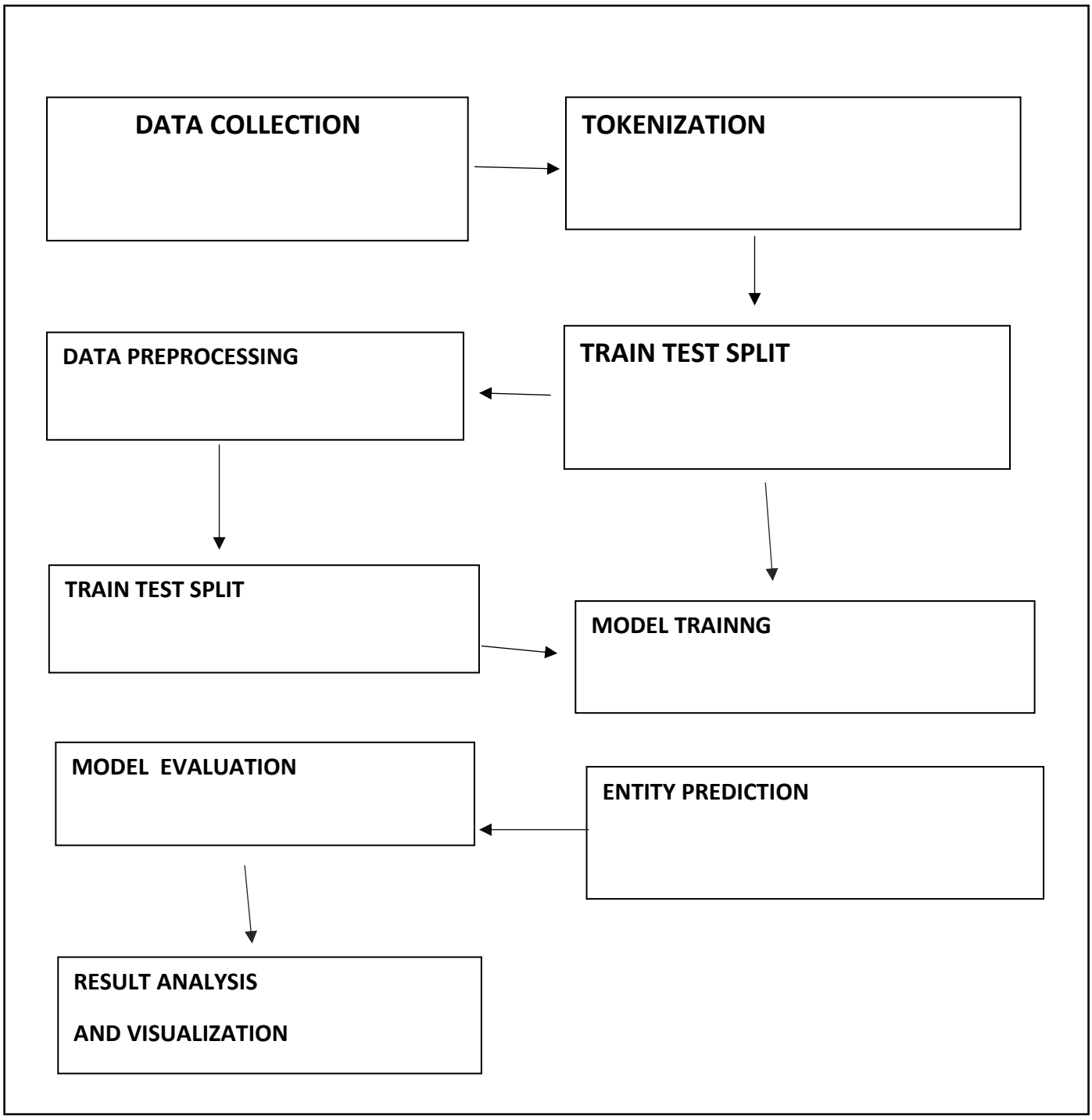
CoNLL-2003 dataset has been employed extensively to train and test Name Entity Recognition systems for name entity recognition and classification of named entities in text. It contains text examples of news stories with and without entity tags for people, organizations, and locations. The collection is separated into three collections: training, validation, and test collections, and thus it is a highly general-purpose collection for developing strong NER models. Combined, the CoNLL-2003 data offers 14,041 training instances, 3,250 test instances, and 3,453 validation instances with a solid basis for model testing and training.

The data set also explains nine various classes of entities, i.e., known entities such as 'PER' (person), 'ORG' (organisation), and 'LOC' (location), which help in optimizing the model in recognizing particular classes of entities. These various classes of entities and the huge volume of text data enable the model to learn and recognize minute language variations in their usage and recognition of entities for which it will be appropriate in actual NER applications.

The most frequent application of the CoNLL-2003 dataset is enhancing the efficiency and precision of NER models by training on high quality labeled text data. Natural Language Processing (NLP) techniques, in this case, transformer models like DistilBERT, have broad application to identify the syntactic and semantic features of the text. The features enable the model to categorize various types of entities and comprehend the context in which the entities are being utilized.

Generally, the CoNLL-2003 corpus is still an in-line reference corpus for NER studies that stimulates the creation of complex models that are capable of working with and annotating text from different domains without any difficulty.

FLOWCHART:


DATA COLLECTION
TOKENIZATION
DATA PREPROCESSING
TRAIN TEST SPLIT
TRAIN TEST SPLIT
MODEL TRAINNG
MODEL EVALUATION
ENTITY PREDICTION
RESULT ANALYSIS
AND VISUALIZATION

Tokenization:

It starts with carrying out tokenization as the first step of text processing. It separates text data into independent pieces of text referred to as tokens when it processes words and subwords. In this project, DistilBERT carries out customized tokenization by separating words into their more fragmented small portions. With every word processed, the model builds on its capability of understanding texts.

DistilBERT :

It was developed by the creators to ensure better performance at peak operating efficiency compared to baseline BERT models. The Named Entity Recognition (NER) task employs DistilBERT as the model used in this project. It demonstrates successful identification of needed entities such as names and organizations and locations with its data-driven training and attention-based processing mechanisms. DistilBERT performs optimally when used as a perfect tool for identifying significant text parts from any written content.

Fine-Tuning DistilBERT for NER:

Whenever any pre-trained model such as DistilBERT is utilized for specific application objectives the procedure of the same adaptation gets referred to as fine-tuning. For optimizing entity recognition, DistilBERT is fine-tuned using the CoNLL-2003 dataset in the present work. Internal parameter adaptation because of the process facilitates improved entity recognition on any kind of text that increases the performance level of NER.

7. RESULTS

CONFIGURATION OF LAPTOP:

OS : Windows 11 Home

Brand: DELL

Hard Disk Size : 512 GB

CPU Model : Core i5

RAM Memory : 8 GB

Final Model Evaluation Summary

Model: DistilBERT-base-cased for Named Entity Recognition

Formula: $L=\alpha \cdot LKL+\beta \cdot LCE+\gamma \cdot Lcos$:

- LKL = Kullback–Leibler Divergence Loss
  → Matches the soft predictions (logits) of the student to those of the teacher (BERT)
- LCE= Cross Entropy Los
  → Standard supervised loss between the predicted labels and ground truth
- Lcos = Cosine Embedding Loss
  → Ensures the student and teacher produce similar hidden state representations
- α,β,γ = weights for each component (typically tuned for best performance)
  Dataset: CoNLL-2003
  Training Examples: 14,041
  Validation Examples: 3,250
  Test Examples: 3,453
  Number of Entity Classes: 9

Test Results:

- **Test Loss:** 0.1244**Entity-level Metrics:**

Metric	Value
F1 Score	0.8992
Precision	0.8922
Recall	0.9063
Accuracy	0.9799

Token-level Metrics:

Metric	Value
/*F1 Score	0.9802
Precision	0.9805
Recall	0.9799
Accuracy	0.9799

Detailed Classification Report:

ENTITY	PRECISION	RECALL	F1-SCORE	SUPPORT
LOC	0.92	0.92	0.92	1666
MISC	0.76	0.81	0.79	702
ORG	0.86	0.89	0.88	1661
PER	0.95	0.94	0.95	1615

Averages:

Metric	Value
Micro Avg	Precision: 0.89, Recall: 0.91, F1-Score: 0.90
Macro Avg	Precision: 0.88, Recall: 0.89, F1-Score: 0.88
Weighted Avg	Precision: 0.89, Recall: 0.91, F1-Score: 0.90

Performance Metrics:	
Metric	**Value**
Entity-level F1 Score	**0.8992**
Entity-level Accuracy	**0.9799**
Token-level F1 Score	**0.9802**
Token-level Accuracy	**0.9799**
Best Performing Entity Type	**PER (F1 Score = 0.9486)**
Worst Performing Entity Type	**MISC (F1 Score = 0.7881)**

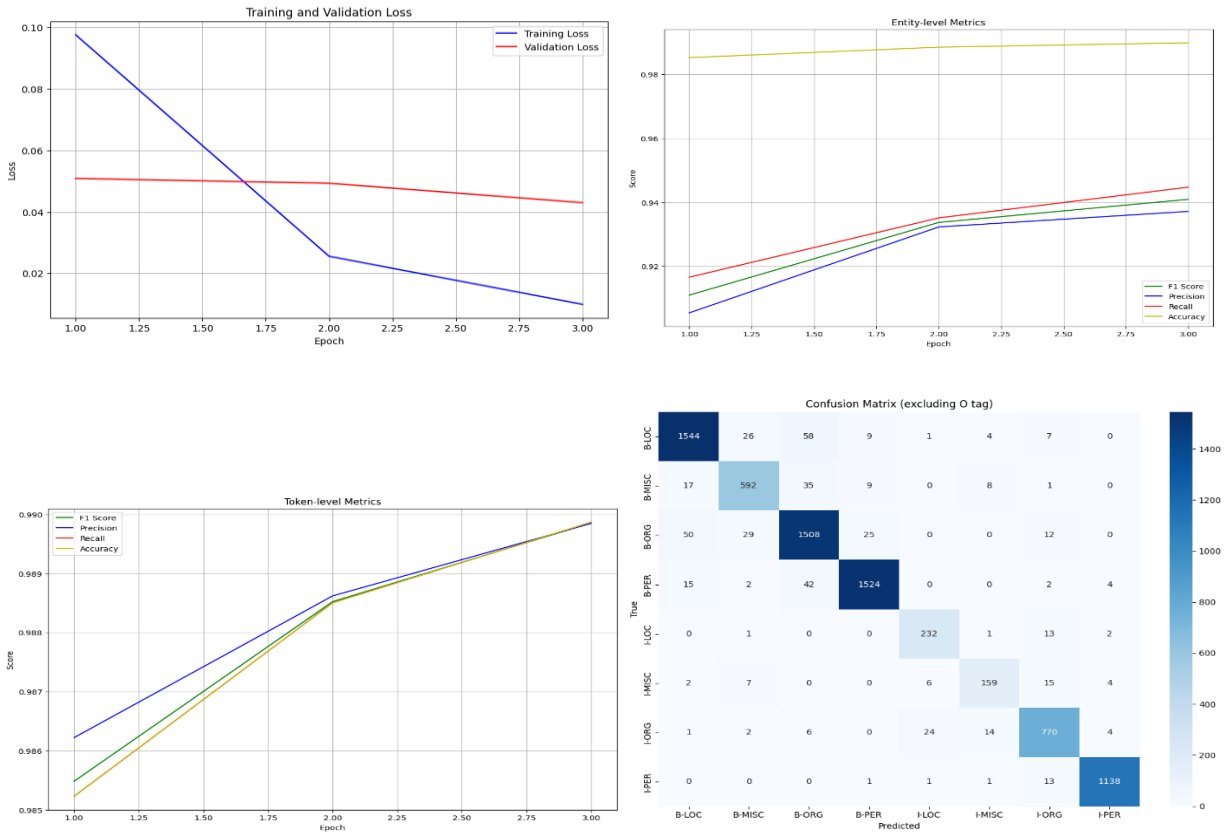


Figure2: graphs of training loss, entity and token level matrices

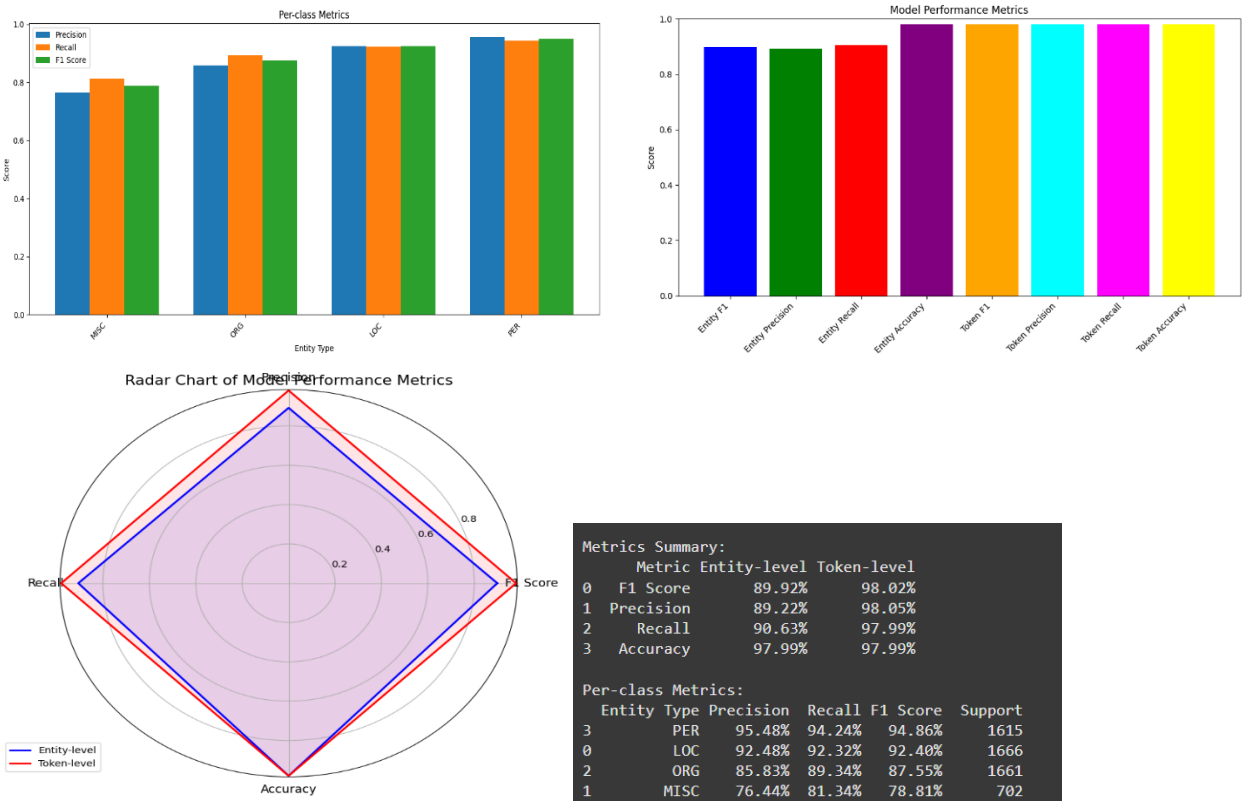


Figure3:bar graphs and radar chart of model performance

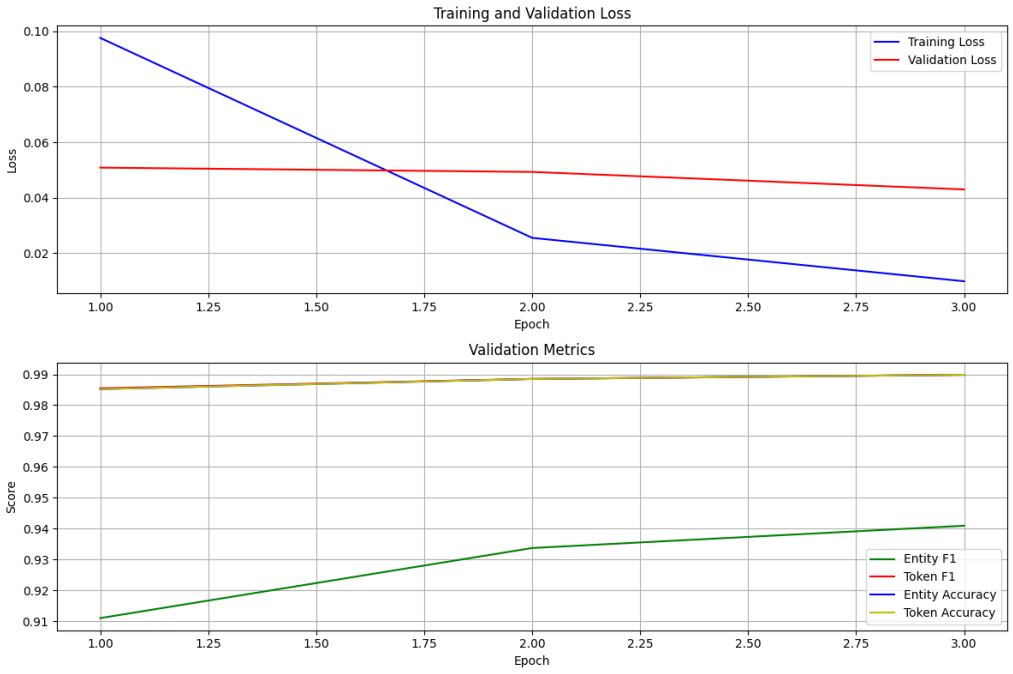


Figure4:training and validation matrices

8. CONCLUSION

The DistilBERT-base-cased transformer functioned as an NER model to test performance results on CoNLL-2003 data. The results demonstrated high performance through 0.8992 entity-wise F1 score and 0.9802 token-level F1 score. With extraordinary generalization conditions the test accuracy remained constant at 97.99%.

The PER (Person) class delivered superior outcomes with 0.95 F1 score but MISC (Miscellaneous) produced the minimum F1 score of 0.79 for future aspect prediction. The research demonstrates how DistilBERT functions effectively for NER sequence labeling tasks even though it belongs to the reduced BERT family version.

The upgraded DistilBERT model demonstrates excellent performance along with high efficiency when extracting important textual entities.

9. REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). https://arxiv.org/pdf/1706.03762.pdf
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). https://arxiv.org/pdf/1810.04805.pdf
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). https://arxiv.org/pdf/1301.3781.pdf
4. Pennington,J.Socher, R., & Manning, C. (2014). https://nlp.stanford.edu/pubs/glove.pdf
5. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). https://arxiv.org/pdf/1603.01360.pdf
6. Huang, Z., Xu, W., & Yu, K. (2015). https://arxiv.org/pdf/1508.01991.pdf
7. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf
8. Yang, Z., Dai, Z., Yang, Y., et al. (2019). https://arxiv.org/pdf/1906.08237.pdf
9. Peters, M. E., Neumann, M., Iyyer, M., et al. (2018). https://arxiv.org/pdf/1802.05365.pdf
10. Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA fhttps://arxiv.org/pdf/2003.10555.acpdf