INFORMATION EXTRACTION FROM EMAIL USING BERT

Presented By:

Tika Sah

THA079MSISE017

Project Supervisor

Er. Kiran Chandra Dahal

Institute of Engineering, Thapathali Campus

Presentation Date: 21-08-2024
Institute of Engineering, Thapathali Campus

PRESENTATION OUTLINE

- Motivation
- Background
- Problem Statement
- Objective of Project
- Scope of Project
- Originality of Project
- Potential Applications
- Literature Review
- Methodology
- Results
- Discussion and Analysis
- Future Enhancements
- Conclusion
- Timeline
- Reference

MOTIVATION

- In the world of automation some task are still done manually.
- Electronic Data Interchange (EDI) for automated transaction processing in business environments.
- Many organization relay on email mechanism for transactional communication.
- Data extraction from email is very necessary in the business world.
- Manual extraction of data from the email is laborious and error-prone.
- Automating data extraction form email is necessary for accurate transaction processing in business operations.

BACKGROUND

- Despite advancements in Electronic Data Interchange (EDI) systems, many businesses still use email for transactional.
- Massive email volume makes information management difficult. Manual extraction is slow and inaccurate.
- Developments in natural language processing (NLP), particularly models like BERT offer potential solutions for automating information extraction from text.
- Existing Natural Language Processing (NLP) techniques can extract specific data points (names, dates, locations).

PROBLEM STATEMENT

- **Manual Processing Challenges:** Manually processing transactional information from emails is time-consuming and error-prone, leading to inefficiencies and potential delays in business operations.

- **Lack of Automated Solutions:** Existing methods for extracting transactional data from emails often involve manual data entry .

- **Complexity of Email Content:** Emails often contain nonstandard formatting, language ambiguity, and context-dependent information, which makes automating the extraction of transactional data challenging.

OBJECTIVE OF PROJECT

- To develop an automated information extraction system utilizing the BERT model.
- To extract information from email content and attachments of type .doc, with the aim of automating the information extraction process.

SCOPE OF PROJECT

- The system aims to accurately identify and extract key entities from diverse and complex email communications.
- The focus will be on extracting specific information from email content and attachments, particularly in DOCX format.

ORIGINALITY OF PROJECT

- BERT for email data extraction along with attachments.
- Solution for automatically extracting transactional information from email .
- Focus on specific challenges in transactional information extraction.

POTENTIAL APPLICATIONS

- Extraction of information will be automated.
- It will boost customer service by providing fastest response by automatically extracting information.
- Workflows will be streamlined via Automated Transactional Email Processing.
- Manual entry of data will be reduced so error will be reduced.

LITERATURE REVIEW[1]

Paper	Topic	Methods	Contribution	Remarks
1	Performance Study on Extractive Text Summarization Using BERT Models	Statistical methods and Deep learning methods	Novel compressed models DistilBERTSum and SqueezeBERTSum with comparable performance to BERTSum ROUGE evaluation metric	Outlines future research directions like hyperparameter tuning, domain adaptation, abstractive summarization with SqueezeBERT, quantization and pruning
2	BERT- and BiLSTM-Based Sentiment Analysis of Online Chinese Buzzwords	Hybrid model combining BERT and BiLSTM	Outperformed other methods on OCB dataset for F1, recall, precision	Lacks comparison to recent models

LITERATURE REVIEW[2]

Paper	Topic	Methods	Contribution	Remarks
3	BART Model for Text Summarization : An Analytical Survey and Review	BART (Bidirectional and Auto-Regressive Transformers) denoising autoencoder	Improved BLEU and ROUGE scores Competitive/superior to models like RoBERTa, GPT	Limited discussion on evaluation metric nuances Insufficient exploration of model limitations Potential data biases
4	Explore BiLSTM-CRF-Based Models for Open Relation Extraction	BiLSTM-CRF with embeddings	Handles overlapping relations High recall and predicate matching score NTS-BERT-BiLSTM-CRF model accurate for multiple relations	Limited comparison to other Open RE methods
5	Natural Language Processing for Knowledge Discovery and Information Extraction from Energetics Corpora	LDA, Word2Vec, Transformer models	Trained on 80,000 energetics documents , Captured energetics concepts well across models	LDA interpretability issues Transformer hyperparameter tuning computationally costly Need further enhancements

METHODOLOGY[1]

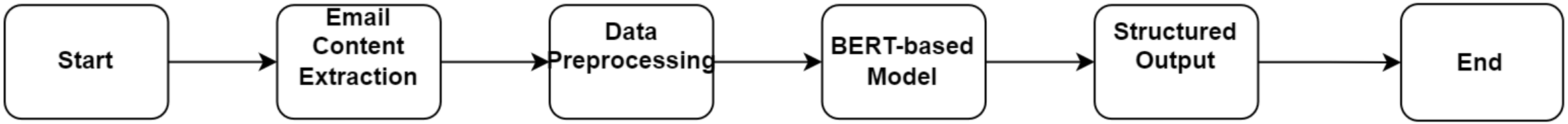


Fig: System Block Diagram

METHODOLOGY[2]

Email Content Extraction:

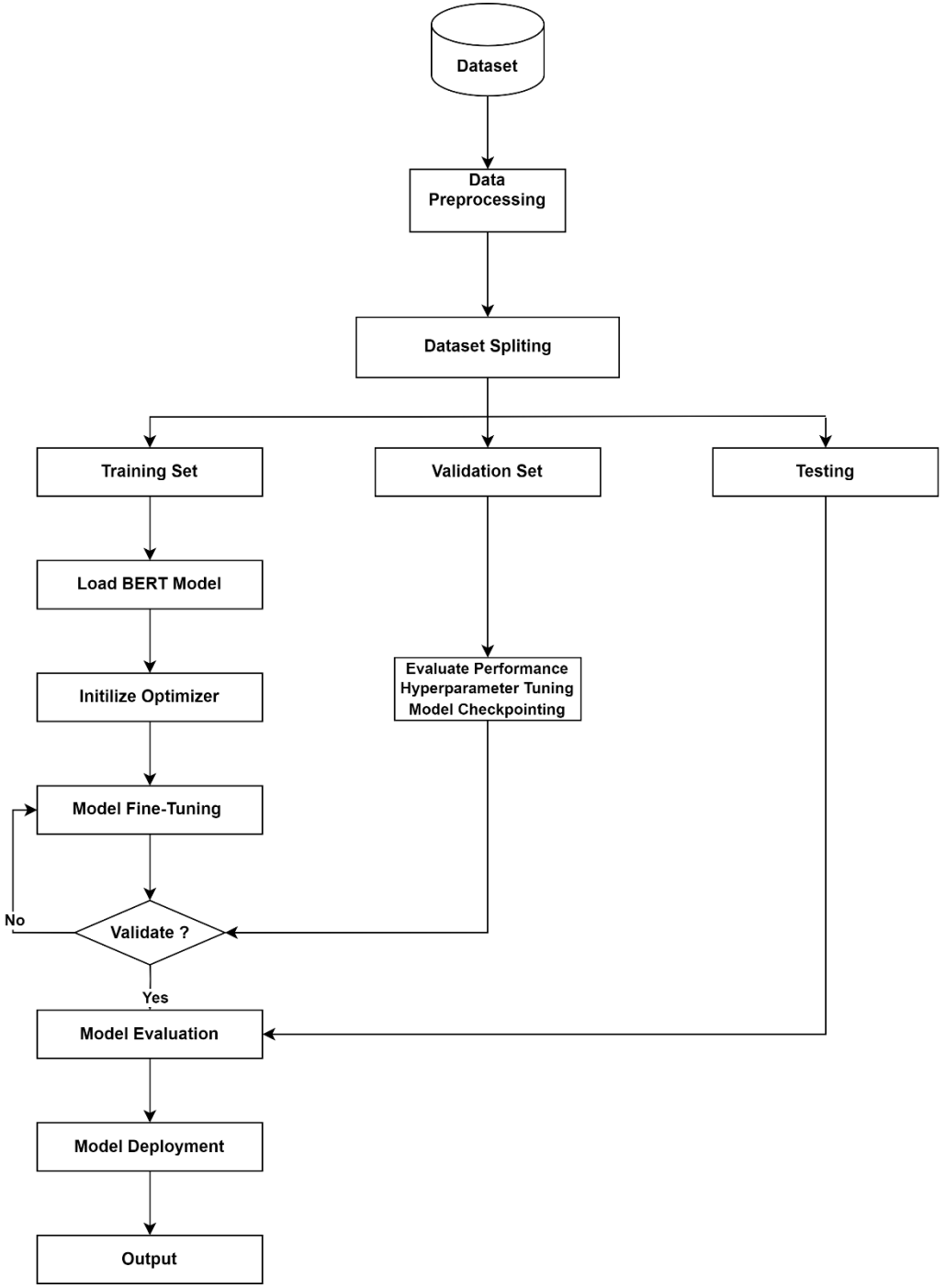
- Parsing raw email data based on RFC 5322 standards to efficiently benchmark key components like headers, body content, and attachments.
- Handling attachments of type docx also extracting content from it.
- Email content are cleaned using Beautisoap and stored in database.

METHODOLOGY[3]

Preprocessing:

- During tokenization, alignment with the respective word labels is preserved to ensure consistency.
- "[CLS]" and "[SEP]" tokens are prefixed and postfixed to tokenized sentences with "O" (Outside) labels inserted for these special tokens.
- Tokenized sentences and their labels are truncated if they exceed the maximum length or padded with "[PAD]" tokens if they are shorter, with labels also padded to align with the sentences.
- The attention mask is a binary list marking each position with 1 for real tokens and 0 for padding tokens, guiding the BERT model to focus on actual content.
- Specific token IDs are used: 101 for "[CLS]", 102 for "[SEP]", 0 for "[PAD]", and other numbers for token IDs.

METHODOLOGY[4]



METHODOLOGY[5]

Model Generation:

- The dataset is split into training (70%), validation (20%), and testing (10%) sets to balance learning, fine-tuning, and unbiased evaluation.

- Data is fed to the model in which model capture semantic meaning by converting tokened to dense vectors call embedding.

- self-attention mechanisms to weigh the importance of each token relative to others.

- In classifier layer , transformer outputs are converted to logits, indicating label likelihood for each token.

- The model is trained over 5 epochs, updating the model performance after each epoch.

METHODOLOGY[6]

- After each epoch model performance is evaluated on validation set using metrics such as loss and accuracy.
- After model is trained , final evaluation is done on test set to assess real-world performance and identify overfitting or underfitting.
- Visualizing training and validation loss curves to monitor model performance over epochs.
- Confusion matrix plotted to actually view the data correctly predicted and which are not correctly predicted.

METHODOLOGY[7]

- Google Colab was used for training the model on GPU and mysql server for data management.
- Redis server used for running background task to receive unseen email.
- Around 20000 data collected from business email and annotation was done for creating dataset for the project.

RESULTS[1]

- Extraction of data from email content along with multiple attachment content and saving in database is working fine .

Mail Details

SN	Subject	Body	Attachments	Sender
1	Test order from docx	Greetings, We are pleased to inform you that we have received a purchase order from RCG Of North Carolina ,...	['test_850_order_.docx']	Tika sah
2	Re: 855 Order Status Question	Hello Christina I thought that approach might work. I was mainly concerned we might get messages about duplicates. I will...		Burney Gibson

Fig : Extracted data from mail

RESULTS[2]

- **Best case scenario:** In the Best case scenario model successfully extract all required entities from the model .

“Hello random We must regrettably cancel the order numbered SRM3720109 involving 32 units of Self Lubricating Wear Plates Item ID 03221413 priced at 609 per Each We apologize for any inconvenience this may cause and ask for your understanding Please confirm the cancellation at your earliest convenience and inform us if further steps are necessary Sincerely NextGen”

Predicted Result:

{

"Sender": nextgen,

"Item ID": 03221413

"Quantity": 32,

"Unit of Measure": each,

"Order Number": srm3720109

“product Name”:plates

"Cost": 609

}

RESULTS[3]

- **Worst case scenario:** In the worst case scenario model unable to extract all entities from data , however some of the entities can be extracted.

‘*Dear Emily Brown,I hope you are doing well. I am writing to place an order for 120 units of Ergonomic Office Chair with Item ID having EOC45678 under Order Number ORD54321. The total cost for this order is \$12,000, and the items should be delivered to 456 Corporate Blvd, Suite 300, Los Angeles, CA 90017. Please ship the chairs in Pallets Unit of Measure: Pallet and ensure that the shipment is handled with care to avoid any damage. The order is being placed by Michael Lee, who can be contacted at michael.lee@company.com for any clarifications.Best regards,Michael Lee.’*

Predicted Result:

{

"Sender": Not Found,

"Item ID": "03535132",

"Quantity": 120,

"Unit of Measure": Not found,

“Product Name":Not found

"Order Number": ord54321,

"Cost": 1200

}

RESULTS[4]

- Classification Report from validation set of data 20% of total data .
- On 4000 data Classification report is generated.
- On all entities product name entities have low score .

	precision	recall	f1-score	support
Cost	1.00	1.00	1.00	12221
Item_ID	0.99	1.00	1.00	18589
Product_Name	0.90	0.94	0.92	8516
Unit_Of_Measure	1.00	1.00	1.00	4015
order_num	1.00	1.00	1.00	22229
quantity	1.00	0.98	0.99	4187
receiver	1.00	1.00	1.00	4000
sender	1.00	1.00	1.00	9180
micro avg	0.99	0.99	0.99	82937
macro avg	0.99	0.99	0.99	82937
weighted avg	0.99	0.99	0.99	82937

DISCUSSION AND ANALYSIS[1]

- Training loss consistently decreases across epochs , indicating that the model is learning and improving during training.
- Validation loss is relatively stable and lower than the training loss, suggesting that the model is generalizing well to unseen data.
- The gap between the training and validation loss curves indicates that overfitting is not an issue.

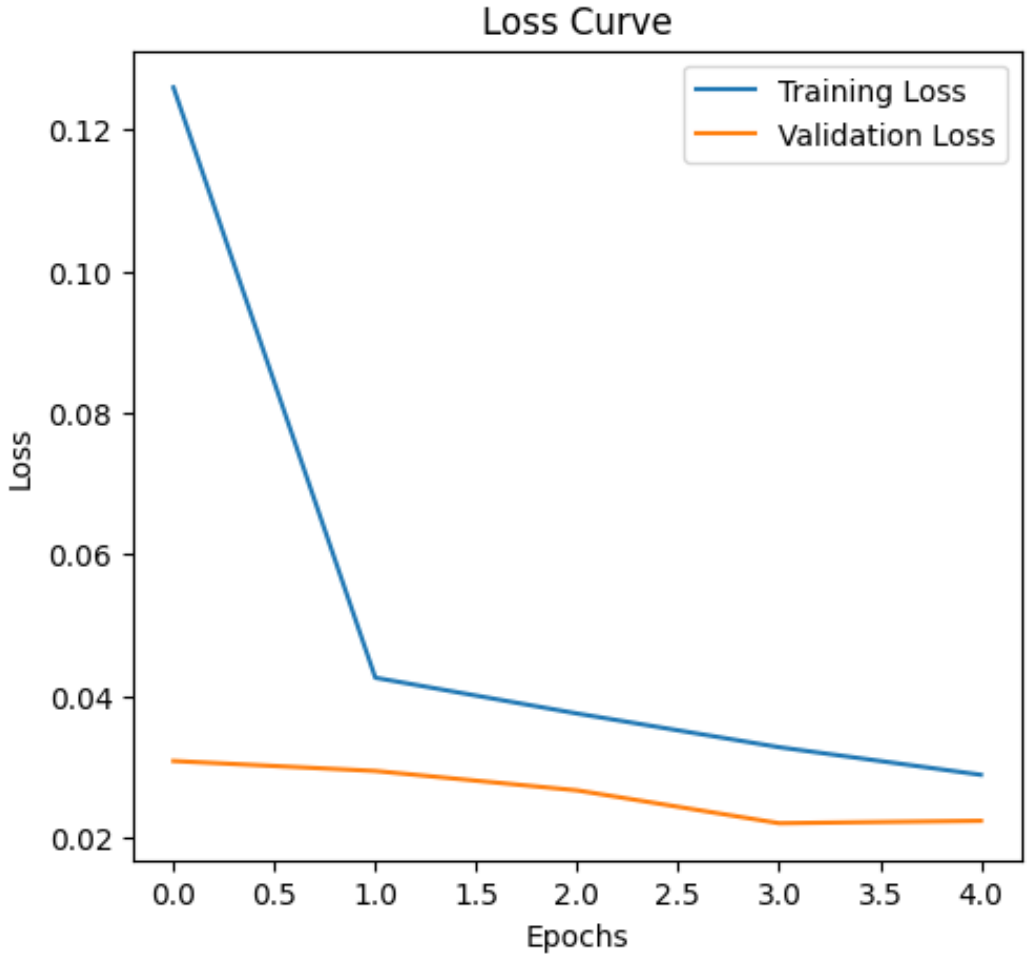


Fig: Loss Curve

DISCUSSION AND ANALYSIS[2]

- The plot shows both training and validation accuracy across 5 epochs, with the training accuracy increasing rapidly.
- Both training and validation accuracies reach over 99%, indicating a well-trained model.
- The slight difference between training and validation curves suggests monitoring for potential overfitting in future epochs..

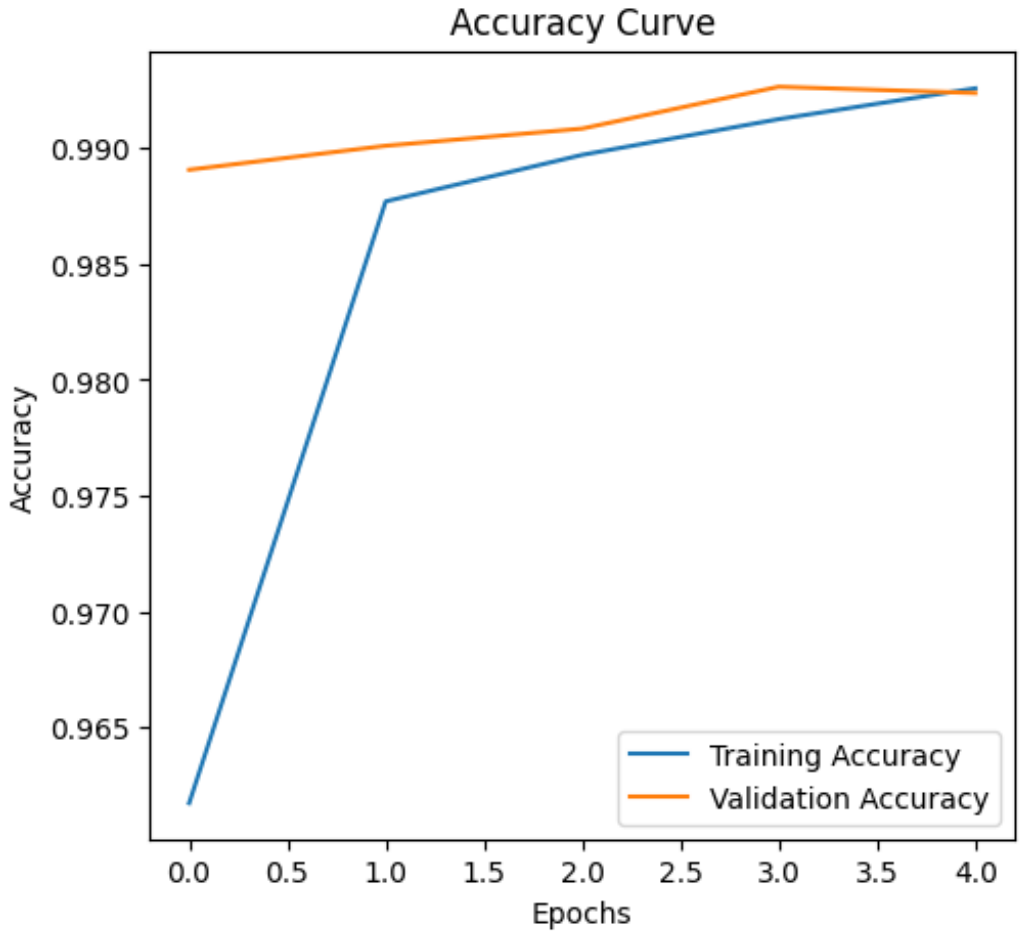


Fig: Accuracy Curve

DISCUSSION AND ANALYSIS[3]

- The project meet the objective of extracting email content smoothly.
- Even the mail contain multiple attachments of docs type it will extract the text content from it .
- Output of model is satisfactory , for normal mail it will extract key entities smoothly even for complex only few entities remain to extract.
- Error in result due to non-standard formatting and multiword entity disambiguation

DISCUSSION AND ANALYSIS[4]

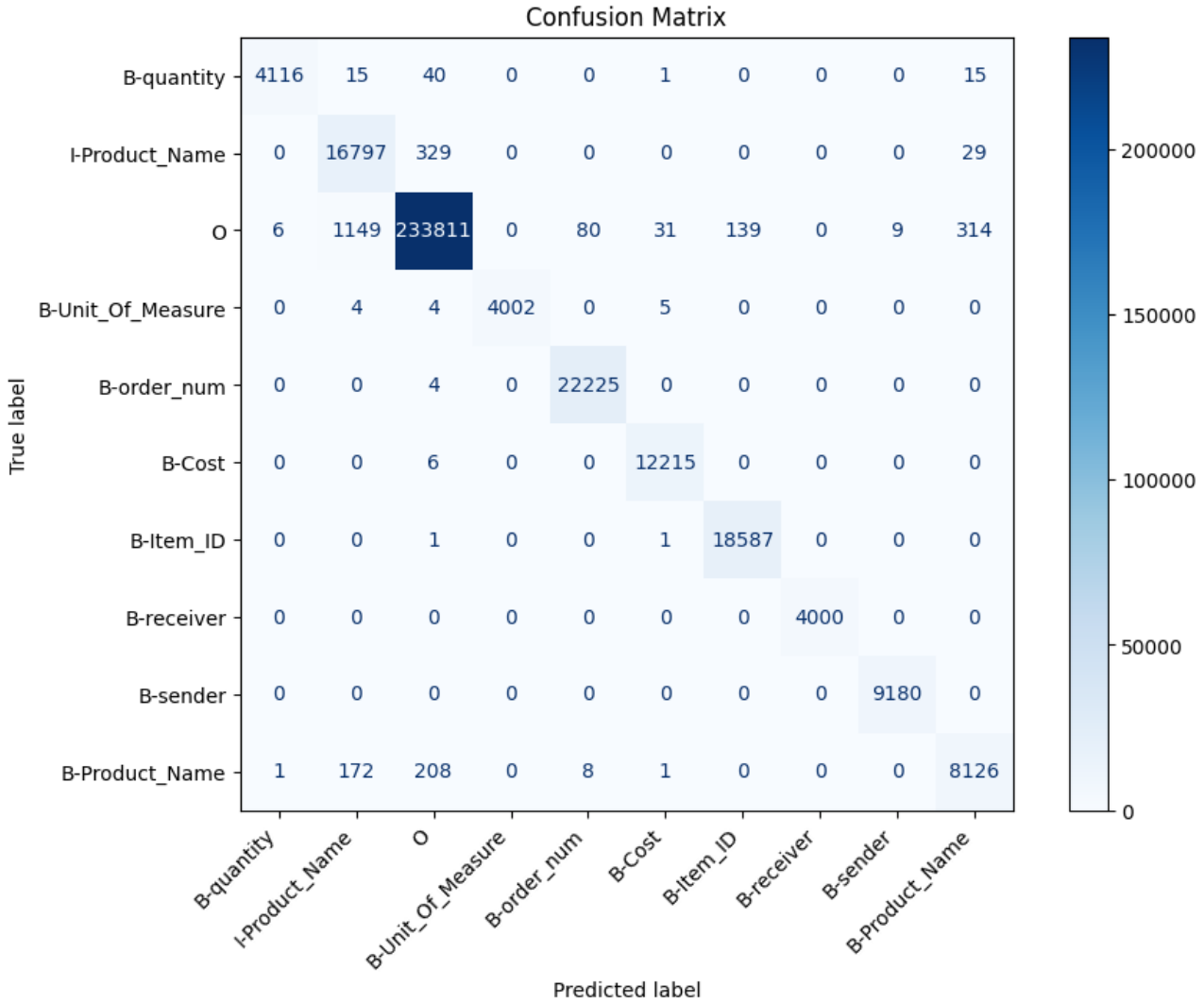


Fig: Confusion Matrix

FUTURE ENHANCEMENTS

- Include a more diverse range of email samples to improve the model's Performance.
- Focus on thorough data cleaning and consistent annotation to ensure enhancing model performance.
- Explore different NER models, such as Conditional Random Fields, to improve entity recognition, especially in complex email contexts.
- Implement ensemble modeling and active learning techniques to boost prediction accuracy and reliability in information extraction tasks.

CONCLUSION

- Celery background tasks allows the system to continuously fetch and process unseen emails with minimal manual intervention.
- The system effectively extracted key from emails and .docx attachments, demonstrating the BERT model's high accuracy in real-life applications.
- The BERT model achieved perfect score showcasing its effectiveness in identifying critical information.
- Challenges in extracting "Product Name" and "Cost" entities for this model and extracting other entities from complex email indicating opportunities for future enhancements in the model's performance.

TENTATIVE TIMELINE (GANTT CHART)

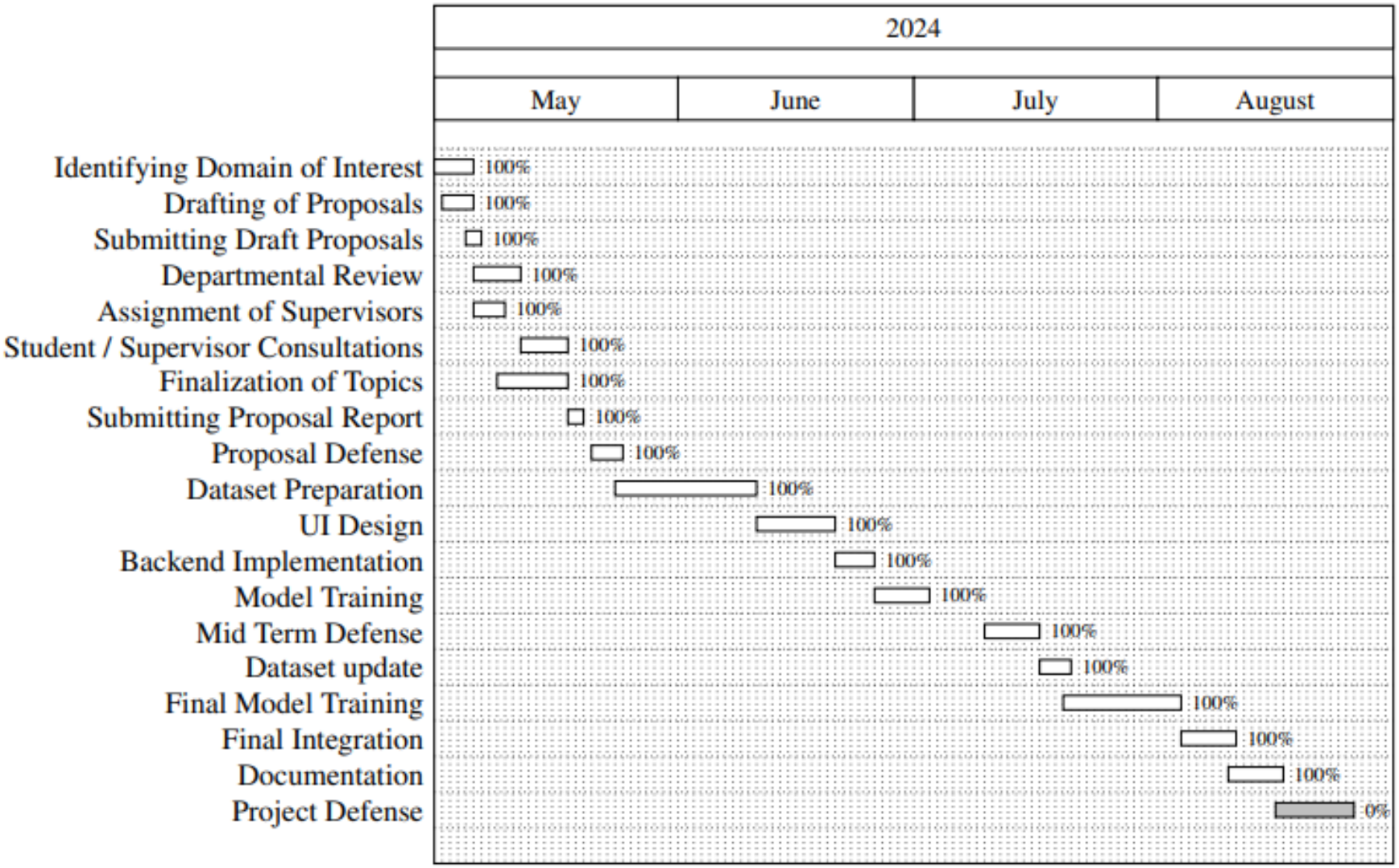


Fig: Gantt Chart

REFERENCES

[1] Mario Franco Sarto, Lorenzo Morselli, Alessia Campi, and Fausto Giunchiglia.Knowledge-based management of patients in intensive care units: An ontology- driven approach. Information, 13(2):67, 2022.

[2] X. Li, Y. Lei, and S. Ji. Bert- and bilstm-based sentiment analysis of online Chinese buzzwords. Future Internet, 14(11):332, 2022.

[3] C. Mankar et al. Bart model for text summarization: An analytical survey and review. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 2(5), may 2022.

[4] T. Ni, Q. Wang, and G. Ferraro. Explore bilstm-crf-based models for open relation extraction. https://www.researchgate.net/publication/XXXXXXX, apr 2021.

[5] Kayal Padmanandam, K. V. N. Sunitha, B. M. Jafari, A. Jafari, M. Zhao, and N. Pitla, "Customized named entity recognition using BERT for the social learning management system platform CourseNetworking," *Journal of Computer Science*, vol. 20, no. 1, pp. 88–95, 2023.

[6] F. G. VanGessel, E. Perry, S. Mohan, O. M. Barham, and M. Cavolowsky. Natural language processing for knowledge discovery and information extraction from energetics corpora. Prep, oct 2023. First published: 06 October 2023

THANK YOU!