REAL-TIME CARBON NEUTRALITY MANAGEMENT AND OPTIMIZATION USING NATURAL LANGUAGE PROCESSING

Mathanika Mannavarasan

(IT19005218)

BSc (Hons) in Information Technology

Specialising in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2022

REAL-TIME CARBON NEUTRALITY MANAGEMENT AND OPTIMIZATION USING NATURAL LANGUAGE PROCESSING

Mathanika Mannavarasan

(IT19005218)

Dissertation Submitted in Partial Fulfillment of the Requirements for the BSc (Hons) in Information Technology Specialising in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

September 2022

Declaration

I declare that this is my own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or institute of higher learning, and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to the Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic, or other mediums. I retain the right to use this content in whole or part in future works (such as article or books).

Signature: Date: 09/09/2022

The above candidate has carried out this research for the undergraduate dissertation under my supervision.

Signature of the supervisor: Date:

Abstract

Carbon emission reduction is a worldwide priority. Businesses that refuse to change will face problems in the future. Reduced greenhouse gas emissions should be a key priority for every large, medium, or small firm. Governments also enforce many rules to control GHG emissions. Companies, on the other hand, tend to limit their carbon emissions. Collecting and keeping emission factors is a vital responsibility for every firm. A single business analyst (BA) or a small BA team is generally in charge of this. Collecting data about emission activities from various sources is a time-consuming effort for a business analyst, and it can sometimes be inaccurate. They usually capture emission data after the emission process has been finished for a more extended period, and most of these procedures are done manually. Therefore, there will be no real-time data on the organization's emissions and no real-time data on the organization's emissions. The solution of text input is implemented in a mobile application that takes the emission details from the employee's text. From the text emission factors, named entity recognition techniques will be extracted. The extracted factors will be forwarded to the search system to search for emission factors and provide ranked results.

Keywords: Carbon emission, Business Analysist, emission factors, named entity recognition, search system

Acknowledgment

I want to sincerely thank my parents for having faith in me and helping me pursue the degree by providing both financial and emotional support for me during the challenging times I faced while earning the degree. I want to express my appreciation to my colleagues Sathees P, Vishakanan S, and Vithurshan M for their collaboration and ongoing support, which were very helpful to me in completing this project.

My sincere gratitude goes out to both our project supervisor, Ms. Anjalie Gamage, who guided us toward success and believed in the quality of our work, and our co-supervisor, Ms. Sanjeevi Chandrasiri, for his support and counsel during the research disputes. I'd like to express my gratitude to the Sri Lanka Institute of Information Technology for providing research materials and assisting us as we conducted our studies. Without Mr. Jagath Wikramarathne, the current and former instructor in charge of the Research Project module, who led us through each stage of the project, it wouldn't be successful.

Finally, I want to express my gratitude to everyone who gave their time, responded to our questionnaires, advised and suggested adjustments to our system, regardless of how big or small.

Mathanika Mannavarasan
Department of Computer Science and Information Technology,
Faculty of Computing,
Sri Lanka Institute of Information Technology.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
CMS	Carbon Management Systems
GHG	Greenhouse Gases
EF	Emission Factor
BI	Business Intelligence
API	Application Programming Interface
ER	Entity Relation
UI	User Interface

1. INTRODUCTION

1.1. Background and Literature Survey

The total amount of greenhouse gas emissions—primarily carbon dioxide—that an individual, group, or country as whole causes are referred to as its "carbon footprint." Carbon dioxide (CO_2) or carbon dioxide equivalent is a common way to represent a carbon footprint (CO_2e). The majority of greenhouse gases, including CO_2 and methane (CH_4), are released when land is cleared, fossil fuels are burned, and a variety of goods and services are produced and consumed.

greenhouse gas emissions (GHGs). On a global scale, state emissions are considerable. The principal GHGs linked to global warming are CO_2 and CO. Currently, coal accounts for 30–40% of global CO emissions from fossil fuels. Carbon assessments may be a key component of a plan to reduce carbon dioxide emissions and increase revenue.

Other greenhouse gases include methane, nitrous oxide, and fluorinated gases. The individual volumes are all converted into CO_2 equivalents to determine the total amount of greenhouse gas emissions and assess the overall human effect (CO_2e).

Both globally and per person, the amount of these gases has significantly grown since the beginning of the industrial revolution. These greenhouse gases contributed to an annual CO_2e emission total of over 50 billion metric tons in 2016.

Nearly all of the increase in atmospheric CO_2 concentration since 1990 is attributable to human activities, which also accounts for the observed global warming.

The amount of carbon we now emit into the environment each year—10 gigatons—is raising the atmospheric CO_2 levels. We can split out the contribution from various fossil fuel kinds since we know how much of each type of fuel is used annually. The usage of coal, which is the dirtiest fuel, has recently sharply increased, which is not good for reducing carbon emissions in the future.[1]

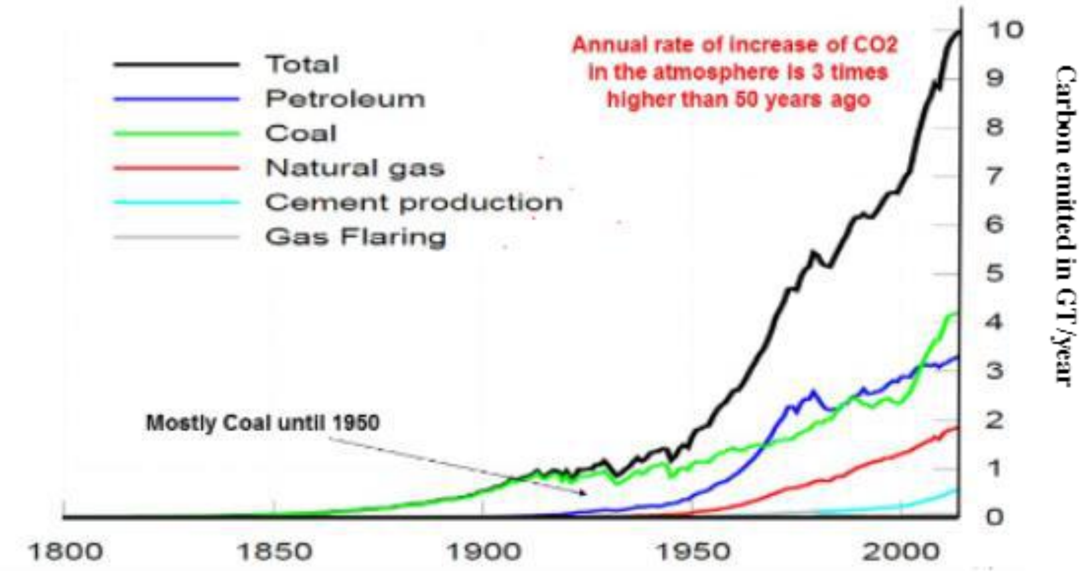

Figure 1-1: Annual rate of increase of CO2 because of fuel

Actual measurements of atmospheric CO2 have been carried out at various laboratories since 1958. Historic levels are derived from gases trapped in ice cores in Greenland and Antarctica. These observations suggest that 46% of the CO2 put into the atmosphere by human activity stays in the atmosphere, increasing its concentration. The rest is absorbed by the oceans and terrestrial biosphere.[1]

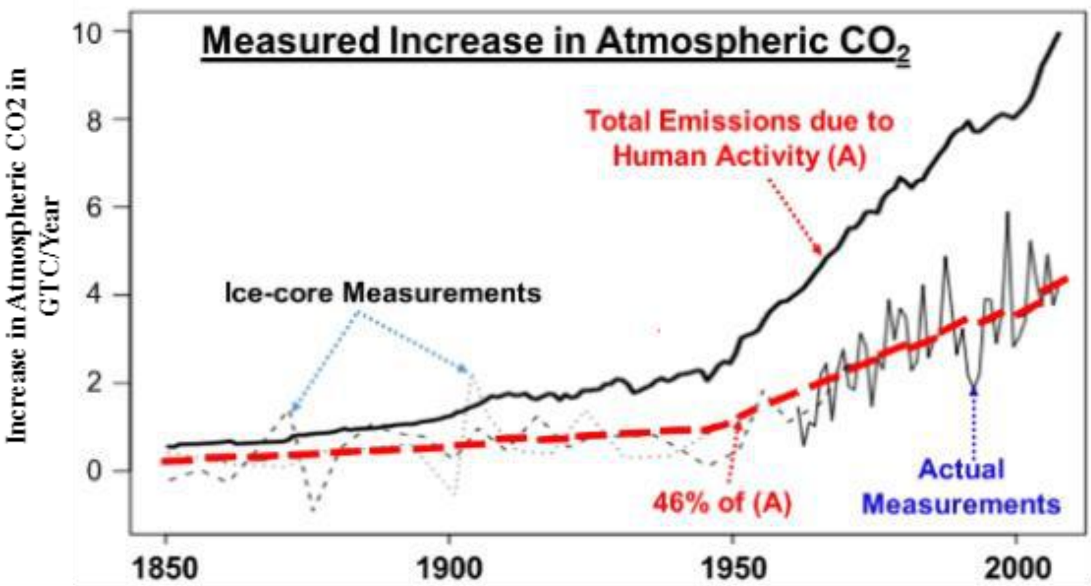

Figure 1-2: Annual rate of increase of CO2 measured in atmospheric CO2

In recent years, worries regarding the effects of greenhouse gas (GHG) emissions on climate change have greatly increased [2]. Reduced carbon emissions help to combat the consequences of climate change, enhance public health, stimulate the global economy, and preserve biodiversity. By reducing carbon emissions, we can ensure that future generations will enjoy access to healthier air, water, and food. Many nations and supranational organizations have established laws and agreements to promote decreasing GHG emissions in response to these worries [3]. The 1997 Kyoto Protocol and the 2016 Paris Agreement were two of these important initiatives [2]. Sri Lanka, one of the signatories to the Paris Agreement, has taken several measures to attain carbon neutrality by the year 2060, such as lowering emissions associated with power generation by utilizing renewable energy [4].

Artificial Intelligence (AI) has dramatically developed in recent years, expanding its potential. Natural Language Processing is one of AI's applications (NLP). In addition, emerging technologies such as virtual and augmented reality are modifying people's interactions with the world and altering digital experiences. Thanks to advancements in cloud computing, artificial intelligence (AI), and the Internet of Things, entity recognizers are the next step in human-machine connection (IoT). The term" Named Entity (NE)," which is widely used in Information Extraction (IE), Question Answering (QA), and other Natural Language Processing (NLP) applications, originated in the Message Understanding Conferences (MUC), which influenced IE research in the United States in the 1990s [Grishman and Sundheim 1996] (to be precise, it was first used in MUC-6 in 1995). MUC's concentration at the time was on IE tasks, which involved extracting structured information about the company and defense-related activities from unstructured text, such as newspaper articles. Every day, the amount of text generated in various fields, such as health care, news articles, scientific publications, and social media, skyrockets. The International Data Corporation (IDC) has forecast that by 2020, the volume of data will have increased 50-fold to 40 billion gigabytes [2]. Textual data is relatively frequent in most disciplines, but due to its unstructured nature, automated comprehension is difficult, which has led to the development of numerous text mining (TM) algorithms in the recent decade [2].

1.2. Research Gap

As was indicated previously in the literature study, there are a number of distinct literary resources that belong to a similar conceptual ideology of emission calculation. Several studies have been conducted in this area, and several products are currently in use by various organizations and governments. Furthermore, all the programs rely on employee surveys and manually maintained reports to obtain emission data [5], [6]. The most essential point is that none of the applications use real-time data. They use past emission data to calculate the emission value. This suggested system's main goal is to determine emission values utilizing real-time emission data. Employees can input their emission tasks through the text into the program. The emission factors will be extracted from that text and sent to the search system.

Before we begin integrating system features, we must do a thorough examination of relevant systems or goods currently on the market. Implementing a new application with the same features will be a waste of time. It is most beneficial to reduce the workload by studying the applications that are already accessible on the market.

The proposed structure will serve as a form of viewpoint for programs that use text to calculate carbon neutrality. This will eliminate the requirement for a large number of people as well as the time it takes to collect data.

Table 1-1: feature Comparison with other products

	Research 1 [1]	Research 2 [2]	Research 3 [3]	Research 4 [4]	Our System (Carbonis)
Gather emission activity daily	No	No	No	No	Yes
Calculate real-time emission value	No	No	No	No	Yes
Emission data collection from employees	No	No	Yes	No	Yes

The confirmation of a text extractor model in the illuminating zone is shown in this research. The fundamental explanation includes an arrangement of specific designs, a model for regulating correspondence, and outfitting the appropriate responses to the understudy [7].

Extracting emission variables from the employee’s text is a difficult problem that has received a lot of attention. We use the Names Entity Recognition technique in this study to extract the exact emission components from the employee's text.

These characteristics have generated a research gap, which has allowed us to develop a fashioned application employing Natural Language Processing.

2. RESEARCH PROBLEM

This study aims to calculate emission value from employee's text in real-time so that companies are better able to maintain the emission values and can maintain the carbon neutrality in the environment.

In the existing circumstances Global CO2 emissions from fossil fuels have increased dramatically in recent decades, generating a variety of issues such as global warming, health effects, and climate change, all of which have economic and environmental implications. Scientists believe that if the world's population continues to release carbon dioxide at its current rate, the globe will warm to unacceptable levels within 20 years [8], posing an environmental risk to everyone. Carbon emissions have climbed by roughly 90% since 1970, and according to recent study, carbon emission footprints will increase by 20% by the end of 2020 [9].

Limitation of carbon emission from the organizations is the most wanted thing these days. Most of the countries' governments that part take in the Paris agreement [9] release quite an extensive emission factor document for various activities, annually [10]. These factor values are used to calculate final emission values for the activities performed over the reporting period [6]. but most of the governments/ companies using historical data, bills, and surveys to calculate the emission value. This approach takes more time to calculate emission value and these calculations not doing for real-time emission data. so, making decisions using these results may not be very useful.

Research is actively involved in the development of custom named entity extraction application. This is currently in use on a variety of company websites around the globe. Real time emission calculation is rarely used, especially from employee's input. Organizations can calculate their emission value in real time using this method. As a result, they can maintain their emission value current. They can readily make decisions with this information. These decisions will be more precise and dependable as a result

Research question: How can we get the emission details of an organization in real-time?

Using custom named entity recognition, we can rectify this problem. There user/employee can just input the daily emission activity when the time emission activity is happened, through the system. Then the system will do the extraction process.

Below picture shows a sample input and output of custom named entity recognition process of a sample emission activity

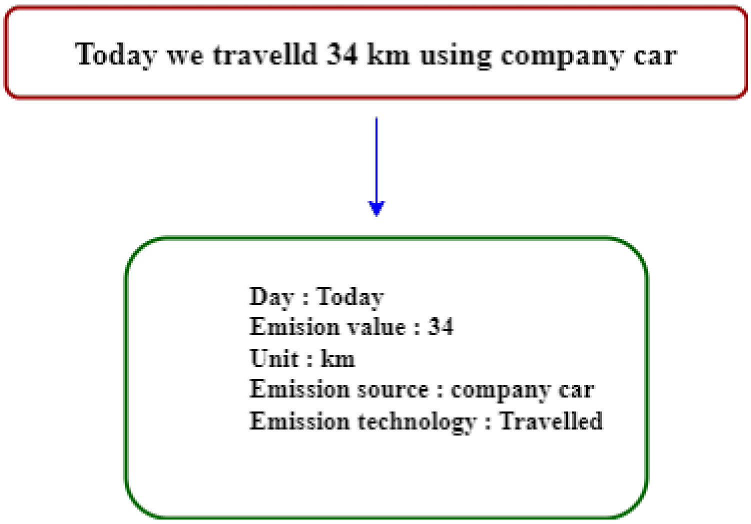

Figure 2-1: Sample Input and output for better understanding

3. RESEARCH OBJECTIVES

3.1. Main Objectives

The primary goal is to provide real time emission calculation system and create a progress plan. The main objective of implementing the carbon neutrality management system from the real-time text data is, there is substantial evidence that anthropogenic activities are responsible for the majority of this warming. Calculating carbon emission is an important first step toward quantifiable emission reduction since it shows how individuals, organizations, countries, and the entire planet react to global warming. Universities, fire departments, rescue services, food businesses, hotels, and hospitals have all calculated their CFP in recent years.

We provide an intelligent, user-friendly carbon neutrality management system for employees to input their daily carbon emission duties through their text in this research. This study is primarily concerned with calculating emission calculations and optimization systems using natural language processing.

3.2. Specific Objectives

For the achievement of this work’s primary objective of implementing an real-time emission calculation system, NER provide custom emission activity extraction feature using custom model.

• Extraction of emission factors from the employee’s input

NER is also simply known as entity identification, entity chunking and entity extraction. Here we use Custom NER for extract emission factors from the employee’s input. This input could be speech or text. If it is speech system will convert to readable text. From the text NER will extract the necessary emission factors for the emission calculation [11].

4. METHODOLOGY

The term "Named Entity (NE)," which is widely used in Information Extraction (IE), Question Answering (QA), and other Natural Language Processing (NLP) applications, originated in the Message Understanding Conferences (MUC), which influenced IE research in the United States in the 1990s [Grishman and Sundheim 1996] (to be precise, it was first used in MUC-6 in 1995). MUC's concentration at the time was on IE tasks, which involved extracting structured information about company activities and defense-related activities from unstructured text, such as newspaper articles [12].

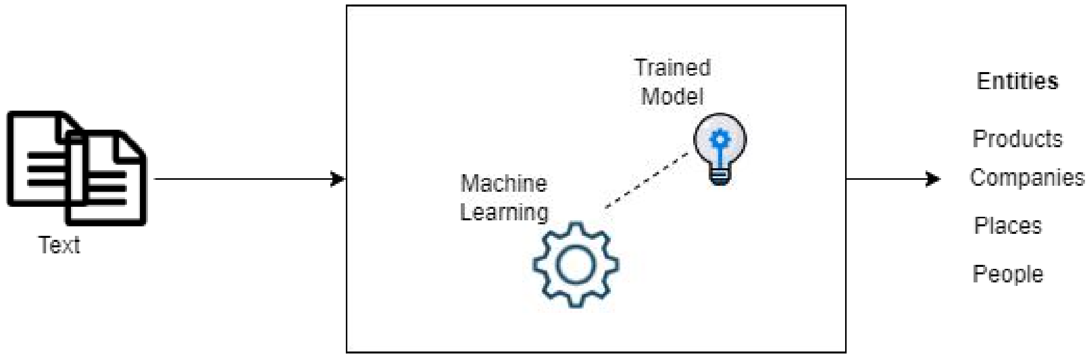

Figure 4-1: Action of Custom Named Entity Recognition

Every day, the amount of text generated in various fields such as health care, news articles, scientific publications, and social media skyrockets. The International Data Corporation (IDC) has forecast that by 2020, the volume of data will have increased 50-fold to 40 billion gigabytes [13]. Textual data is relatively frequent in most disciplines, but due to its unstructured nature, automated comprehension is difficult, which has led to the development of numerous text mining (TM) algorithms in the recent decade [14].

When it comes to custom named entity recognition in order to extract domain-specific entities from unstructured text, such as contracts or financial documents, users of custom NER can create their own AI models. Before making a model usable, developers can iteratively label data, train, assess, and enhance model performance by developing a Custom NER project. Performance of the model is heavily influenced by the quality of the labelled data. We require a source data of the entities with words when it comes to custom named entity recognition. only after that can we use our own words or entities to train the NER model, which will allow us to forecast the appropriate entity [15].

Here is the custom NER development cycle,

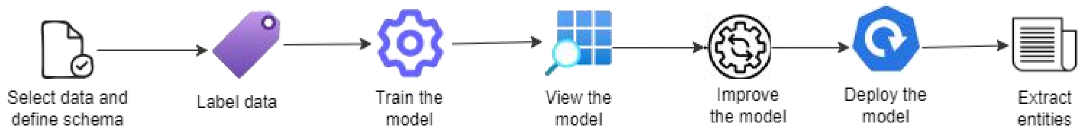

Figure 4-2: custom NER development cycle

1. Define your schema: know about your entities and words
2. Label your data: label your data with the suitable entity. You can do this using annotation tools
3. Model training: train the model using the annotated/labelled data
4. Validate the model: use some validation metrics (precision, recall) to validate your model
5. Deploy the model: deploying your model makes it available for anyone and anytime
6. Extract entities: you can use your deployed model for custom entity recognition

We are extracting the following emission parts: emission technology, emission date, emission source, emission value, and emission unit. NER tools are widely used by industries and are widely available for free. Hugging face, spaCy, Stanford NER, and Natural Language Toolkit (NLTK) The evaluation's goal was to determine whether the tool could identify names' borders and their proper types. Based on exact boundary and type matching, we graded NER systems. The objective of named entity recognition is to assign a specific class to each token (word) in a phrase. A person, a location, an organization, etc. can all be identified by the most popular NER systems that are freely accessible online as per the above mentioned reading it was crystal clear that the custom named entity recognition in real time emission calculating concept will be of many advantages. Although there are researchers prevailing in the area of the emission calculation concept but there is no research prevalent regarding the real time emission calculating using real time employee's emission data.

The Natural Language Toolkit (NLTK) can only extract • VBD: Verb, Past Tense (DT: Determiner, JJ: Adjective, NN: Npun) • IN: Stanford, Preposition, and Conjunction The only things that NER can extract are people, organizations, locations, money, and time. However, in our situation, we must extract specific entity extractors in order to discover emission components from the text of the employee. We may make use of the Hugging and spaCy models for this. In both cases, we may train the model using our own data set and then get the emission factors from it.

4.1. Preliminaries

This section gives a theoretical review of the core concepts used for the custom named entity recognition. This thesis implementation uses spaCy for the NER model. Therefore, this section briefly explains spaCy and Named entity recognition techniques.

4.2. Complete System Architecture

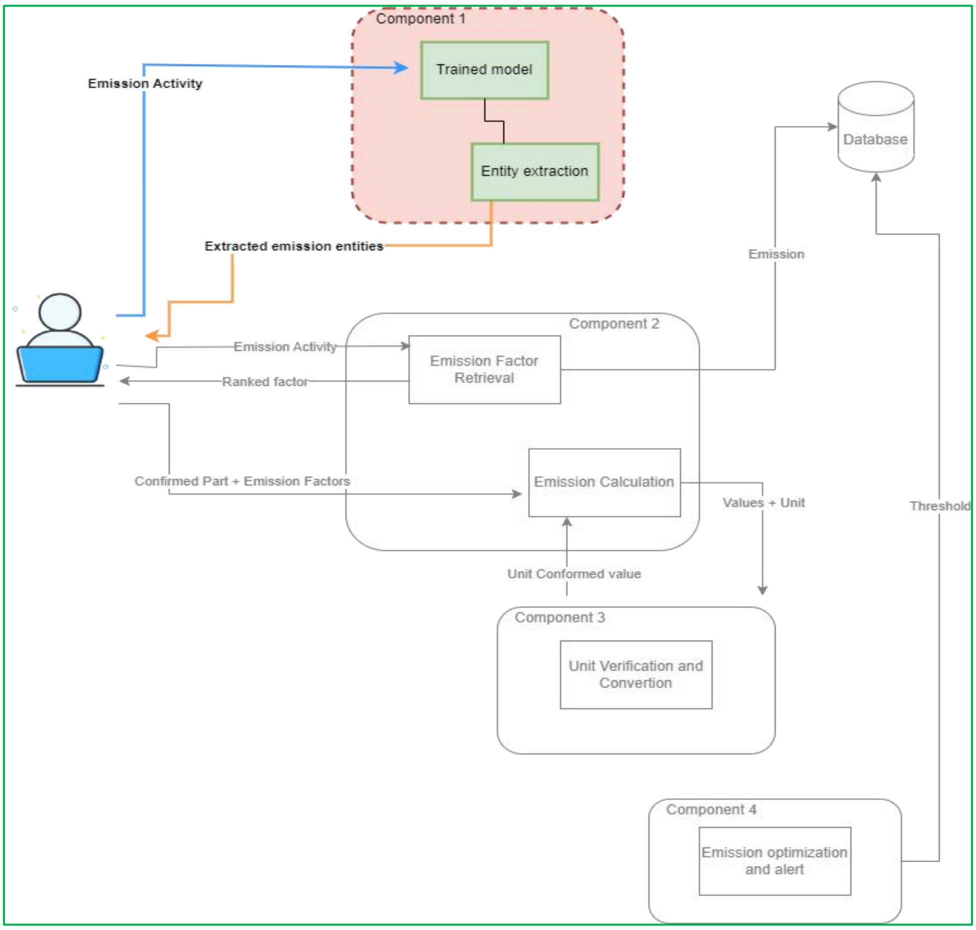

Figure 4-3: Complete system architecture with all research components. The components illustrated in color are for this thesis

As shown in the figure 4-3, this thesis is for component 1 of all four components of the proposed complete system architecture. Component 1 included emission entity extraction from the user/employee's text. NER takes natural language (in this case, English) emission activity data (query) as input from the employee [16] of that organization. Then it will extract the necessary emission relate entities from the input and send back to the user. After that EF retrieval part (component 2) will select suitable EF from the ranked results, the system will calculate emission and store it in the database and data warehouse. Emission values are sent to next component (component 3) to verify and convert units during emission calculation. Business users can generate Business Intelligence (BI) reports from mobile or off-the-shelf BI tools with the emission data stored in the data warehouse.

4.3. Data Collection

We require high-quality data to train the NER model before we can develop it. We gathered information for our emission extraction model through surveys and some common documents. We prepared surveys and distributed them to various client groups, including construction engineers, IT professionals, government officials, and municipality workers. We received about 100 responses.

Here are some responses we got through the surveys

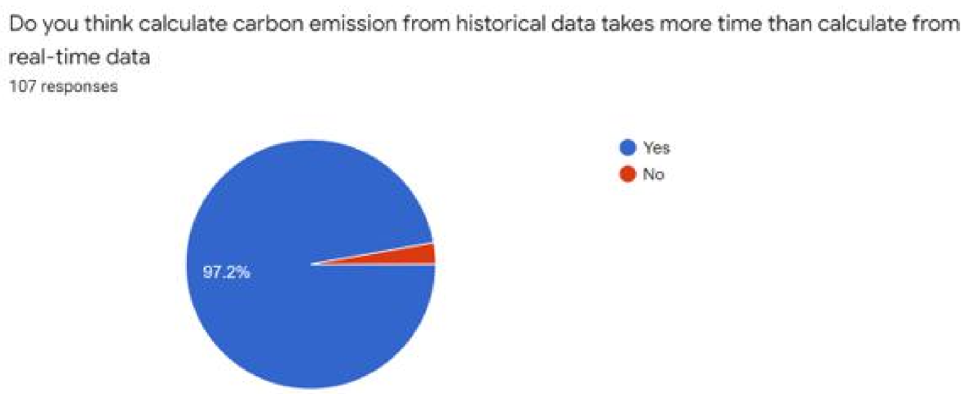

Figure 4-4: Survey result 1

The pie chart above depicts the response that the majority of employees (about 97 percentage) believed to be the most efficient and helpful: compute emission real-time. And below result shows that more percentage of employees (about 94 percentage) thinks that, calculate emission from real time data is more reliable than calculate from historical data.

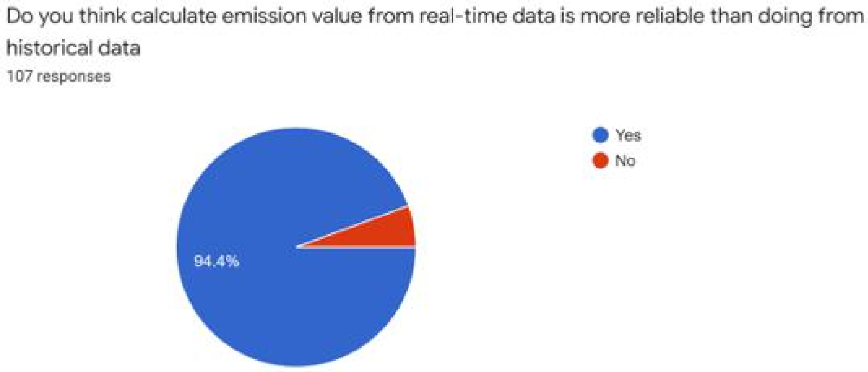

Figure 4-5: Survey result 2

Additionally, we obtained data from certain emission standard publications to train the NER model. those are DEFRA, CRIS, EPA, NGA, IPCC. We have gathered over 500 different types of emission technologies and emission units from these emission standard publications.

There is not enough information gathered from surveys and papers to train the emission extraction model. In order to make the data obtained useable, some pre-processing is required. Each model requires a distinct type of data source in order to be trained.

To train our model, extraction models require annotated data formats. On the internet, there are several free programs that may be used for annotation. The emission factors are included in the annotated data source along with the entity.

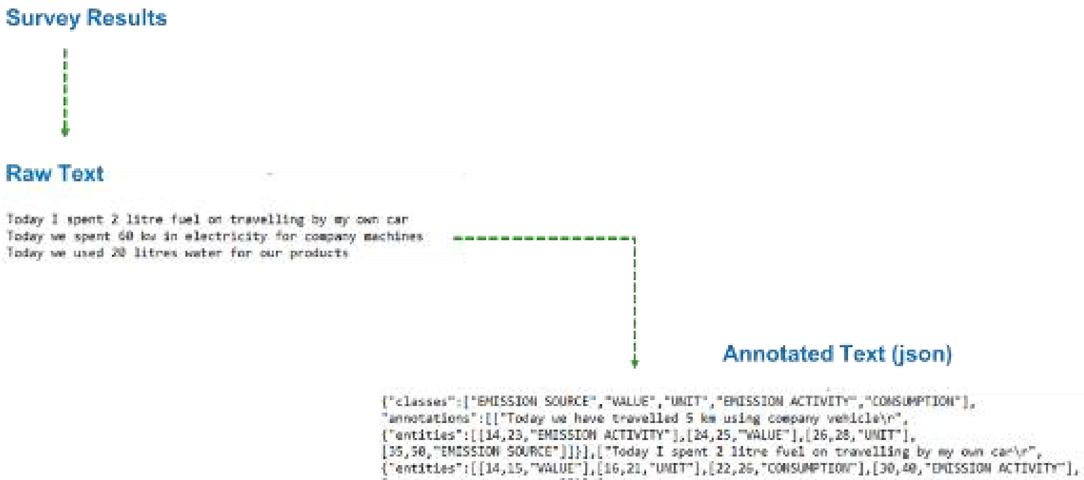

Figure 4-6: Annotation process

4.4. System Architecture

This section gives a practical point of view on how components of the NER system communicate to achieve the entity extraction.

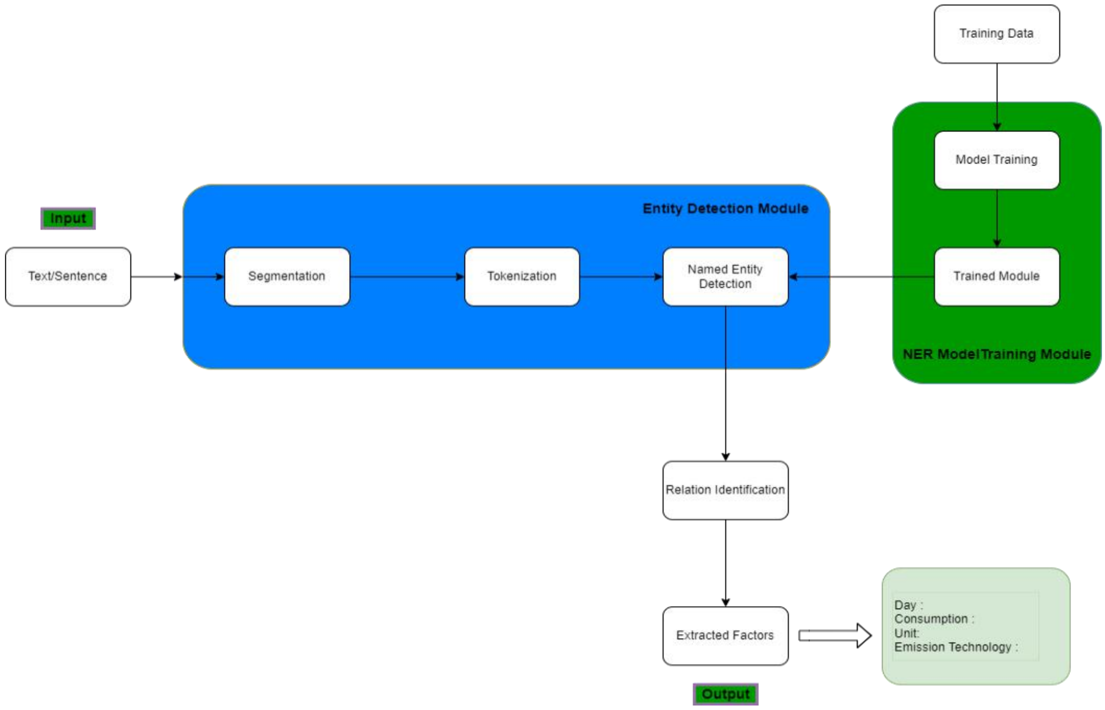

Figure 4-7: Component system architecture

4.4.1. Pre-computations

1. Data preparation

   Using online tools created annotated dataset for the model training. Since we have custom entities (emission related) to train our model. This annotated data will be store as a json file. We can use that file to train our model. Below picture shows the annotator's view.

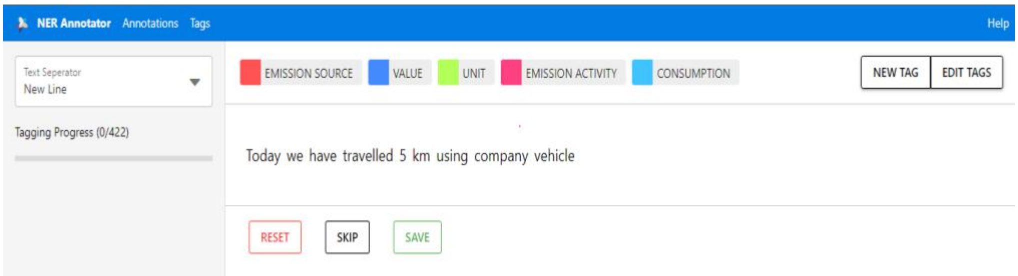

Figure 4-8: Tool used to annotate the data for NER

2. Source data normalization

   Tokens from the input text field were formed during the preparation of the gathered data by using the required linguistic operations, and their tokens were then saved in a database along with a normalized json file. before this process

   Word tokenization: the process of breaking up text strings into token lists

   Eliminating punctuation marks

   Eliminating stop words: Stop words include often used but low-contributing words, such as "a" and "the." Were done.

4.4.2. Application backend implementation

Below picture 4-9 shows the application database's Entity Relation (ER) diagram. It shows the relations between "Emission" with "User"," Emission source" and "Division". Figure 4-10. shows the physical diagram of the backend data warehouse model. This Data warehouse uses a star model. Business users can use this data warehouse to generate BI reports on their carbon emissions. 4-11shows the backend Django

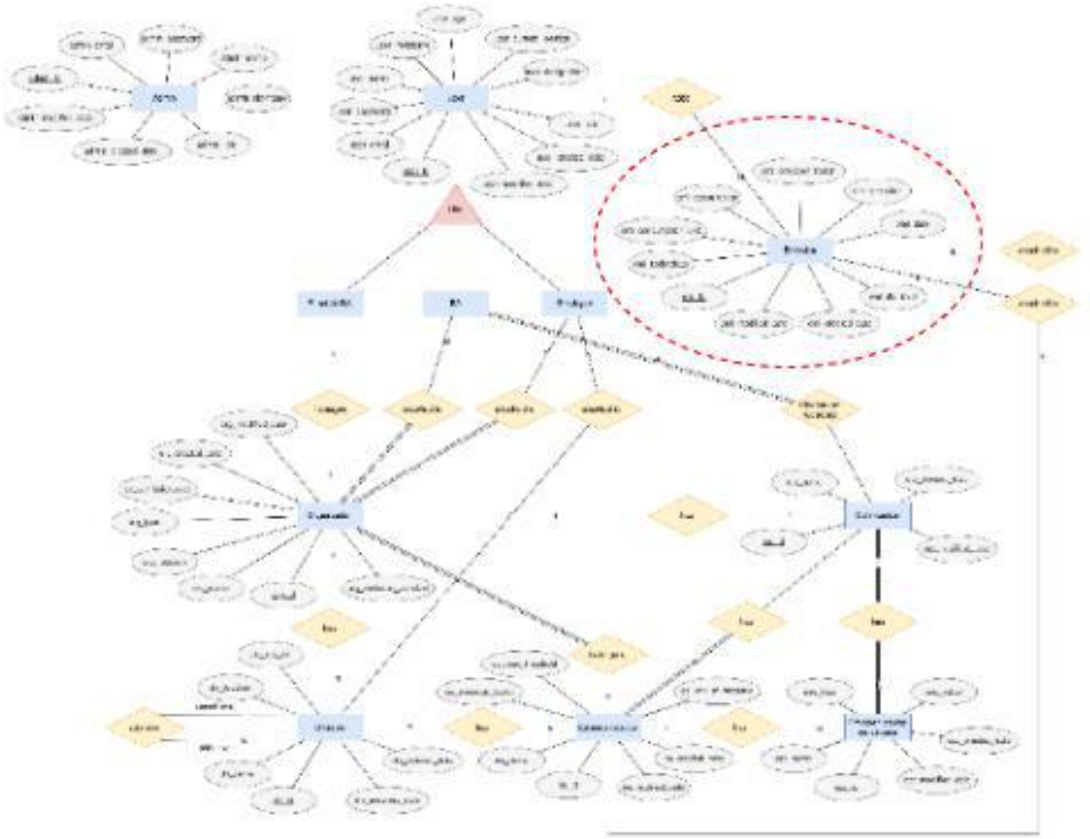

Figure 4-9: Complete system ER diagram

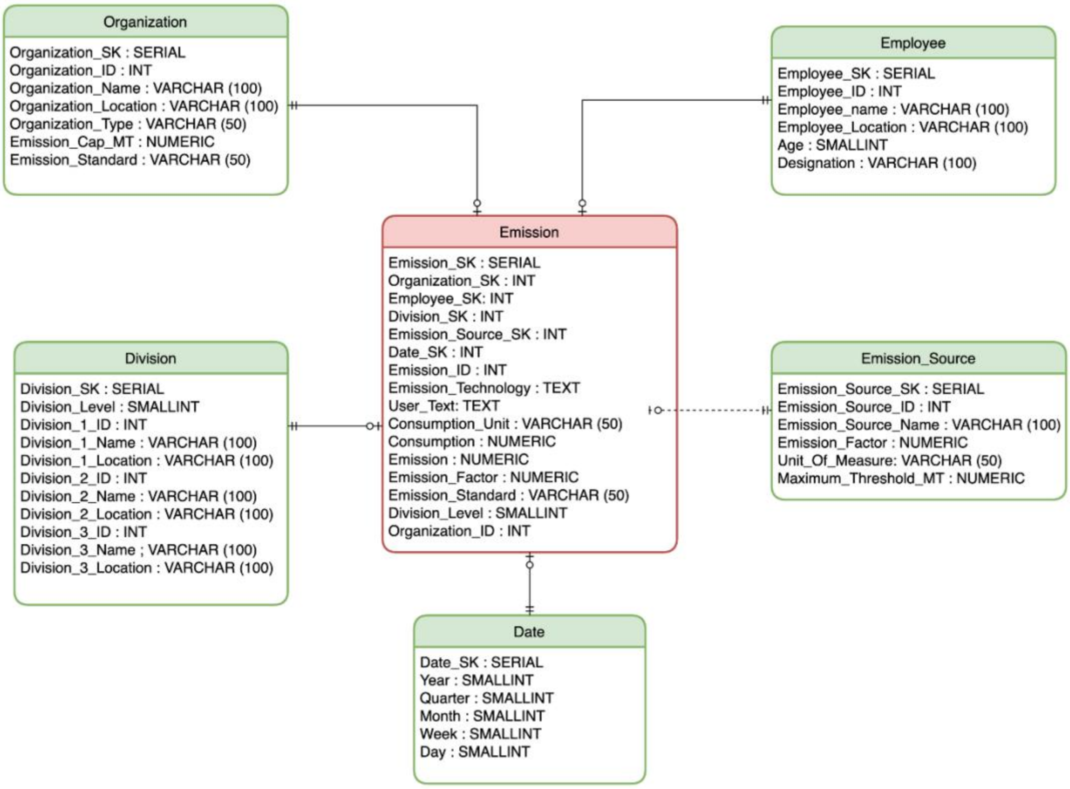

Figure 4-10: Complete system warehouse design

Above picture shows the complete system's warehouse design. This design confirms that emission activity plays a major role in the overall system/Application. It plays as the fact table and have connection between all other tables. And here we can see a star schema.

Technologies used for back-end development and designing:

- Back end – Django
- Development IDE - PyCharm
- Language – Python
- API testing – Postman

4.4.3. Application frontend implementation

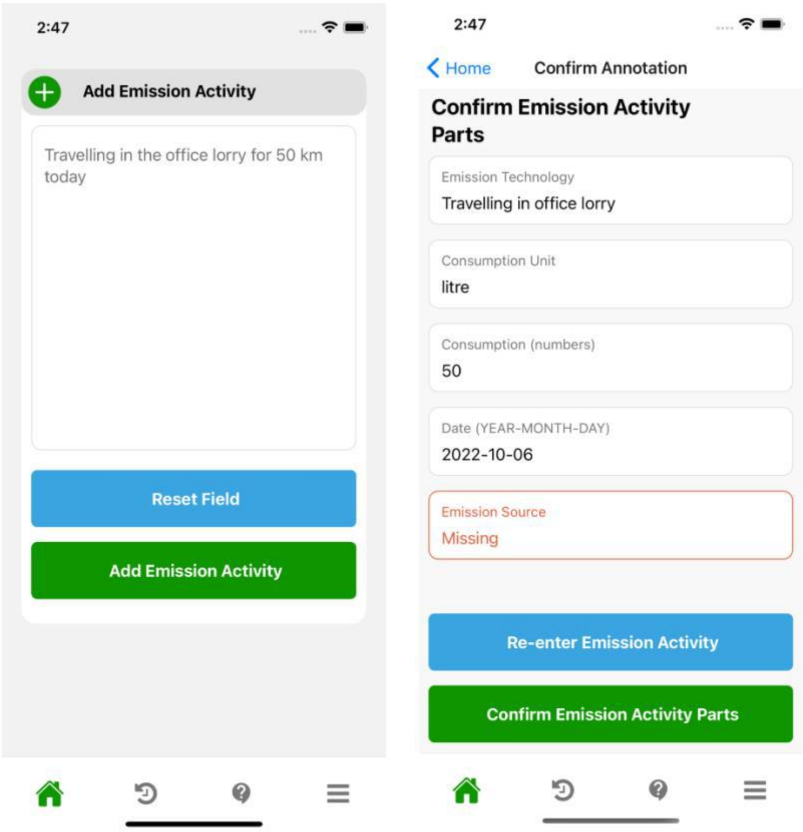

Figure 4-11: Emission activity collection UI

Error! Reference source not found. shows the frontend mobile screens implemented f or emission activity gathering and emission entity extraction

Technologies used for front end development and designing:

- UI design (prototype) – Figma
- Language – JavaScript
- Mobile App development – React-Native

4.4.4. Experimentation criteria

According to the thesis's hypothesis, implementing a novel emission entity extraction system with the NER approach should have better usability and scalability. [13] Therefore, the experimentation criteria are the NER system's usability and scalability. In addition, user satisfaction and extraction speed can contribute to the overall usability criteria of the system. Furthermore, NER scalability and system resource utilization can contribute to the overall scalability criteria. Therefore, experimentations should measure four metrics, such as user satisfaction, query speed, NER scalability, and system resource utilization, to test the thesis hypothesis.

4.4.5. User satisfaction measurement

The user's happiness with the results is a key factor in determining a system's performance, and there are numerous metrics you can use to gauge it, such as accuracy, recall, F1 score, mean average precision (MAP), mean reciprocal rank (MRR), and precision.

4.4.6. Entity extraction speed measurement

Another important factor for a bespoke NER system's usability is entity extraction from the input text or phrase speed. In order to assess extraction speed, tests calculated the average Extraction time (average time). High extraction speed is shown by a short average time, and higher extraction speed improves usefulness.

4.5. Commercialization

The stages involved in commercialization and upcoming marketing plans for our system are covered in this section.

1. Market analysis

Below image shows our basic business model canvas. Here we defined key partners, key activities, value propositions, customer relationships, customer segments, key resources, channels, cost structures and revenue streams for our complete system.

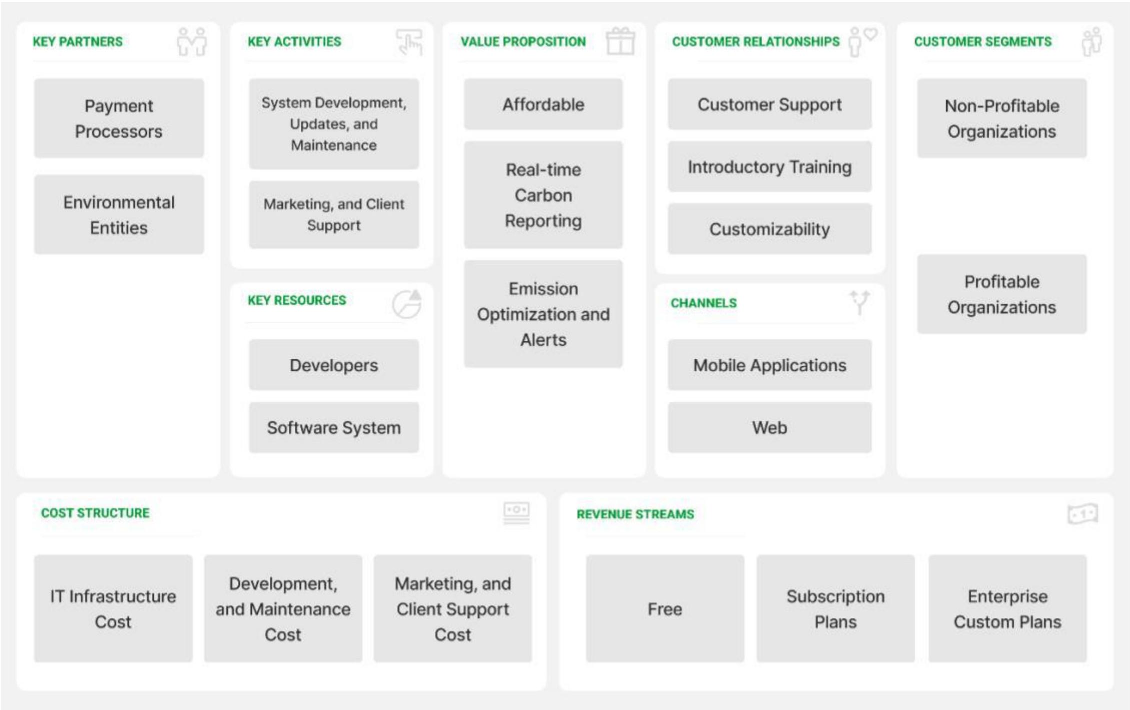

Figure 4-12: Business model canvas

2. Market Attractive opportunities in the carbon footprint management market (Source: Secondary research, markets and market analysis)

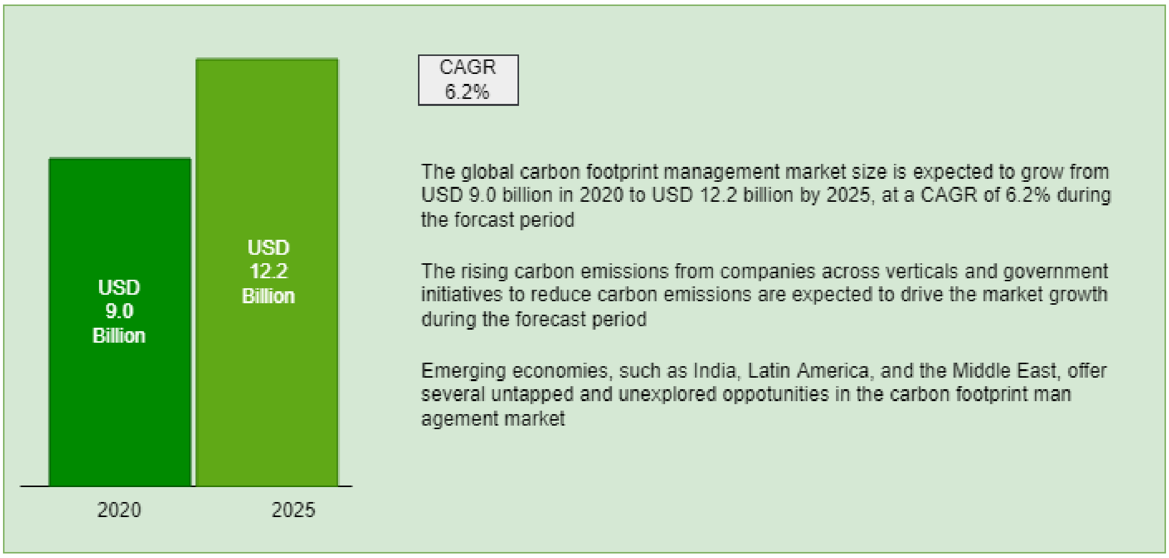

Figure 4-13: Market opportunities

3. Pricing plan

   Below picture shows our pricing plan for the future. We would like to start a basic plan with limited offers for now. But in the future, we are planning to expand the features and that will be a paid one.

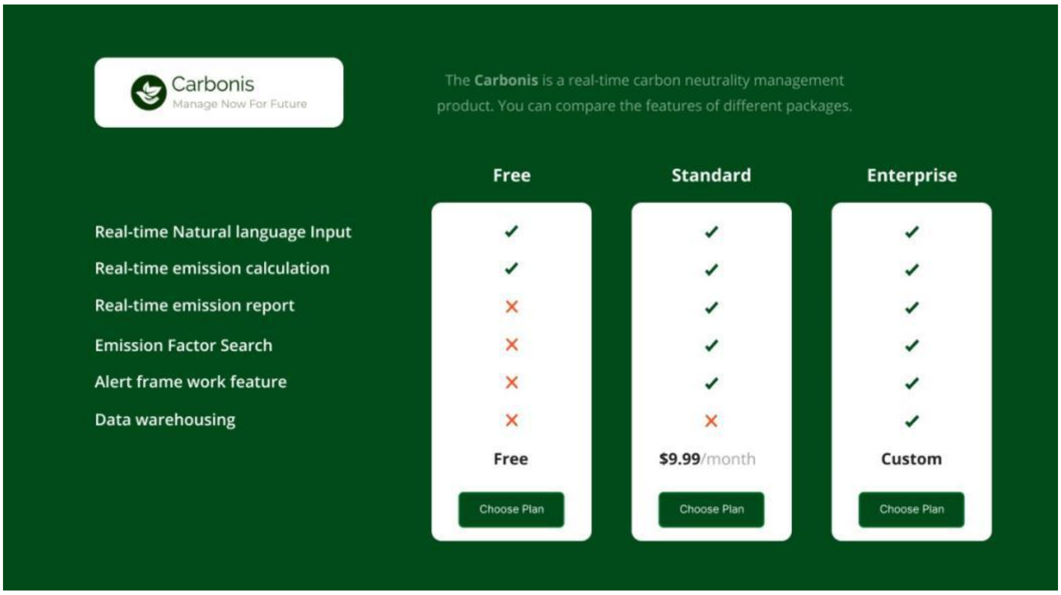

Figure 4-14: Pricing plan

4. Product landing page

Below picture shows our product landing page of our application in web

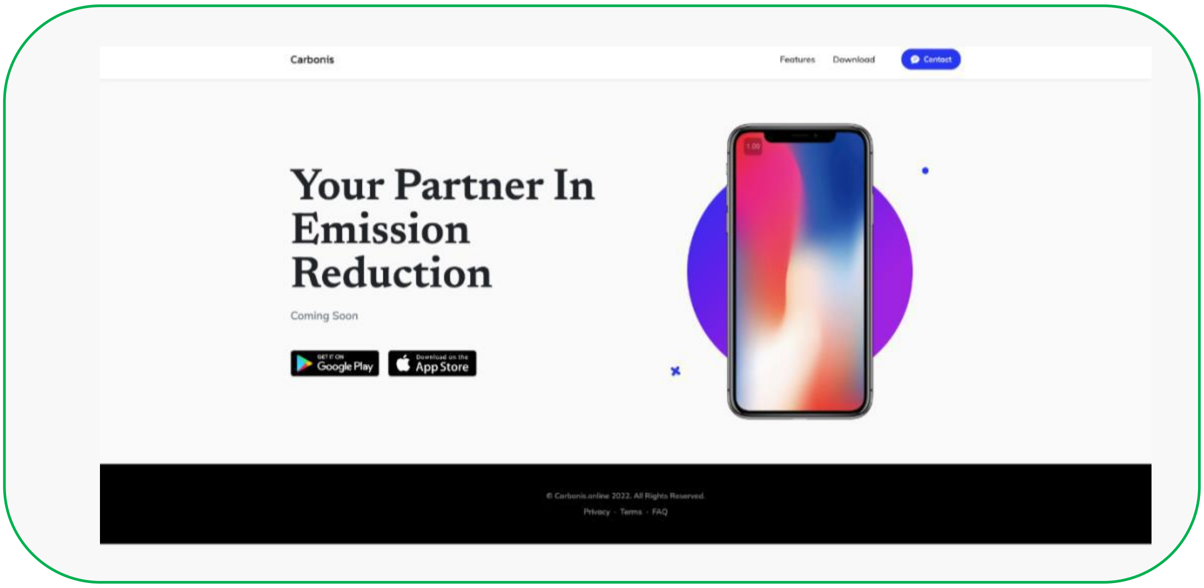

Figure 4-15: Carbonis landing page

5. Product promotion pamphlet

Figure 4-16: Carbonis promotion pamphlet

5. RESULTS AND DISCUSSION

This chapter presents experiment results in the results section and the findings of those experiments in the research finding section. Finally, the discussion section summarizes the finding and the reasoning behind these findings.

5.1. Results

User satisfaction: We tested this component with some random groups. Below is the result we gathered from them. Even this NER worked well. When it is comes to normal users, they need some basic training about this app. In below picture shows that, nearly 60% users feel good about this application.

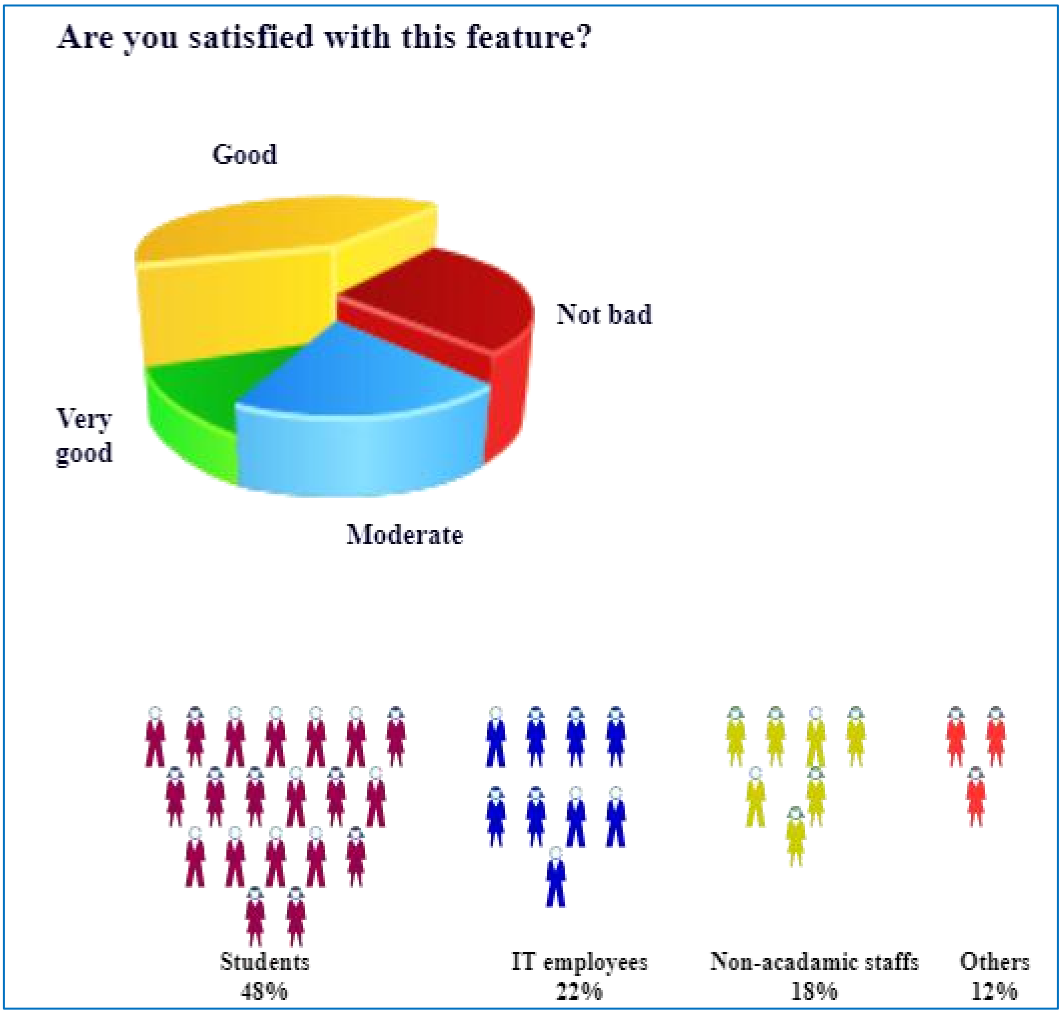

Figure 5-1: User satisfaction result

Average speed: Even When it is comes to speed, for a mobile application speed is one of the most important things. The application should be able to load fast. We gave this application to some set of users and the below graph shows the speed of the application pages and the loading speed.

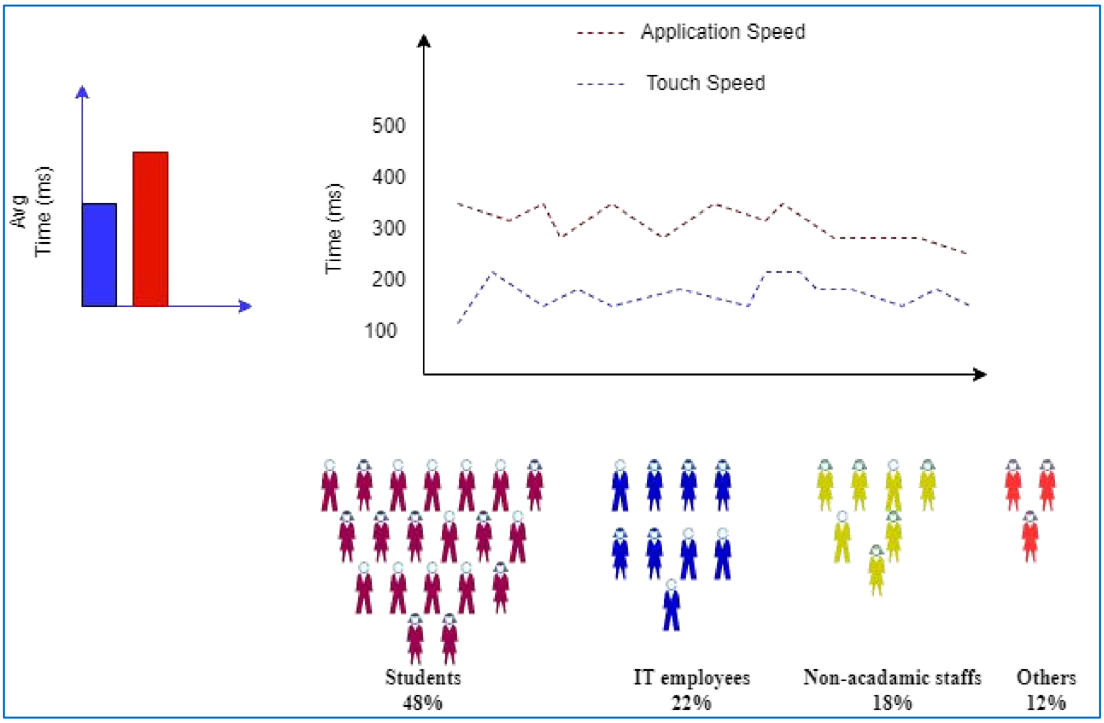

Figure 5-2: Application and Extraction average speed

5.2. Discussion

In this Emission parts extraction - As per our problem, our requirement is to extract the emission factors from the employees' text/input. In this paper we gave several solutions/models to solve. But only some models worked well for our requirement. Three different measures calculated for each NER models. The best model selected based on the performance. However, with more data we can build a most accurate model so that we can get emission activities from any kind of (with spelling errors) sentences/inputs.

6. CONCLUSIONS

Maintaining carbon neutrality is the most crucial factor in today's globe. If we refuse, we shall reach a critical juncture. As a result, all governments should take action to make the planet a healthier place, every nation should set limits on carbon emissions of every organization. Our proposed system will be quite useful in solving this problem. Every organization can calculate their daily emissions and regulate the number of emissions within their control utilizing our technology

REFERENCES

[1] D. Chakrabarty, “The climate of history: Four theses,” *Utopia y Praxis Latinoamericana*, vol. 24, no. 84, 2019, doi: 10.5281/zenodo.2653175.

[2] M. Roelfsema *et al.*, “Taking stock of national climate policies to evaluate implementation of the Paris Agreement,” *Nat Commun*, vol. 11, no. 1, 2020, doi: 10.1038/s41467-020-15414-6.

[3] A. Gulliver, P. E. Carnell, S. M. Trevathan-Tackett, M. Duarte de Paula Costa, P. Masqué, and P. I. Macreadie, “Estimating the Potential Blue Carbon Gains From Tidal Marsh Rehabilitation: A Case Study From South Eastern Australia,” *Front Mar Sci*, vol. 7, 2020, doi: 10.3389/fmars.2020.00403.

[4] B. Tranberg, O. Corradi, B. Lajoie, T. Gibon, I. Staffell, and G. B. Andresen, “Real-time carbon accounting method for the European electricity markets,” *Energy Strategy Reviews*, vol. 26, 2019, doi: 10.1016/j.esr.2019.100367.

[5] F. Mahmoudian, J. Lu, D. Yu, J. A. Nazari, and I. M. Herremans, “Inter-and intra-organizational stakeholder arrangements in carbon management accounting,” *British Accounting Review*, vol. 53, no. 1, 2021, doi: 10.1016/j.bar.2020.100933.

[6] J. Lask, J. Kam, J. Weik, A. Kiesel, M. Wagner, and I. Lewandowski, “A parsimonious model for calculating the greenhouse gas emissions of miscanthus cultivation using current commercial practice in the United Kingdom,” *GCB Bioenergy*, vol. 13, no. 7, 2021, doi: 10.1111/gcbb.12840.

[7] M. Honnibal and I. Montani, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,” 2017.

[8] N. Stern, “We must reduce greenhouse gas emissions to net zero or face more floods,” *The Guardian*, 2018.

[9] S. Thakur and A. Chaurasia, “Towards Green Cloud Computing: Impact of carbon footprint on environment,” in *Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016*, 2016. doi: 10.1109/CONFLUENCE.2016.7508115.

[10] C. Chantrapornchai and A. Tunsakul, “Information extraction on tourism domain using SpaCy and BERT,” *ECTI Transactions on Computer and Information Technology*, vol. 15, no. 1, 2021, doi: 10.37936/ecti-cit.2021151.228621.

[11] M. Konkol and M. Konopík, “Segment representations in named entity recognition,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9302. doi: 10.1007/978-3-319-24033-6_7.

[12] N. Alshammari and S. Alanazi, “The impact of using different annotation schemes on named entity recognition,” *Egyptian Informatics Journal*, vol. 22, no. 3, 2021, doi: 10.1016/j.eij.2020.10.004.

[13] EMC, “IDC Digital Universe Study: Big Data, Bigger Digital Shadows and Biggest Growth in the Far East,” *December 12*, 2012.

[14] R. Phadke, C. Manning, and S. Burlager, “Making it personal: Diversity and deliberation in climate adaptation planning,” *Clim Risk Manag*, vol. 9, 2015, doi: 10.1016/j.crm.2015.06.005.
[15] J. Townsend and J. Barrett, “Exploring the applications of carbon footprinting towards sustainability at a UK university: Reporting and decision making,” *J Clean Prod*, vol. 107, 2015, doi: 10.1016/j.jclepro.2013.11.004.
[16] A. Damm, J. Köberl, P. Stegmaier, E. Jiménez Alonso, and A. Harjanne, “The market for climate services in the tourism sector – An analysis of Austrian stakeholders’ perceptions,” *Clim Serv*, vol. 17, 2020, doi: 10.1016/j.cliser.2019.02.001.

APPENDICES

Appendix A. Model files

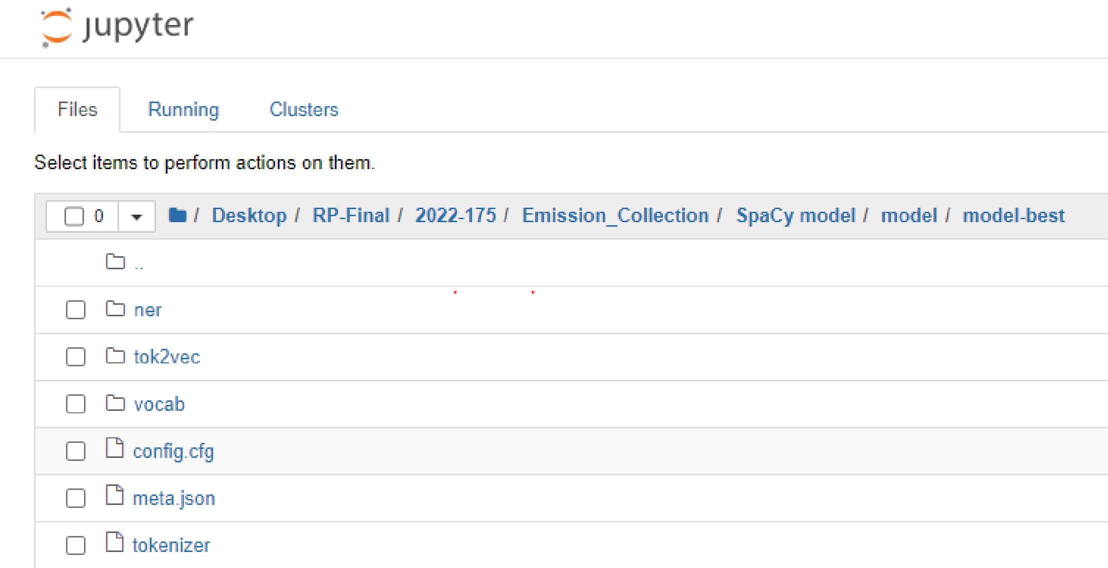

Figure 0-1: Trained model files (Model-best)

Appendix B. Pipeline creation to model training

```
In [5]: !python -m spacy init config --lang en --pipeline ner "config.cfg" --force

[!] To generate a more effective transformer-based config (GPU-only), install
the spacy-transformers package and re-run this command. The config generated now
does not use transformers.
[i] Generated config template specific for your use case
- Language: en
- Pipeline: ner
- Optimize for: efficiency
- Hardware: CPU
- Transformer: None
[+] Auto-filled config with all values
[+] Saved config
config.cfg
You can now add your data and train your pipeline:
python -m spacy train config.cfg --paths.train ./train.spacy --paths.dev ./dev.spacy
```

Figure 0-2: pipeline creation

Appendix C. Model training

```
[+] Created output directory: model
[i] Saving to output directory: model
[i] Using CPU

=========================== Initializing pipeline ===========================
[+] Initialized pipeline

============================= Training pipeline =============================
[i] Pipeline: ['tok2vec', 'ner']
[i] Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     90.44    8.33    4.76   33.33    0.08
  1      10          1.37    820.62    0.00    0.00    0.00    0.00
  2      20          5.50    617.19   23.53   14.29   66.67    0.24
  4      30         11.30    446.30   42.86   27.27  100.00    0.43
  5      40          3.94    205.04   46.15   30.00  100.00    0.46
  7      50          3.86     70.60   46.15   30.00  100.00    0.46
  8      60          5.71     30.70   46.15   30.00  100.00    0.46
 10      70          4.54     19.86   46.15   30.00  100.00    0.46
 11      80          6.31     27.19   46.15   30.00  100.00    0.46
 12      90          4.55     24.22   46.15   30.00  100.00    0.46
 14     100          3.56     20.99   46.15   30.00  100.00    0.46
[+] Saved pipeline to output directory
model\model-last
```

Figure 0-3: Model training

Appendix D. NER output (POS tagging)

```
In [9]: nlp1 = spacy.load(R"model/model-best") #load the best model
        doc = nlp1(sample_text) # input sample tex

        spacy.displacy.render(doc, style="ent", jupyter=True) # display in Jupyter
```

Today DAY we have travelled ET 5 CON km UNIT using company vehicle ET

Figure 0-4: NER output (POS)

Appendix E. API Output (Extracted parts)

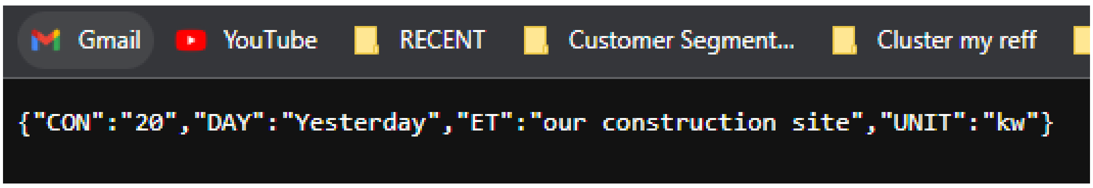

Figure 0-5: API Output

Appendix F. Mobile UI flow for the emission activity extraction

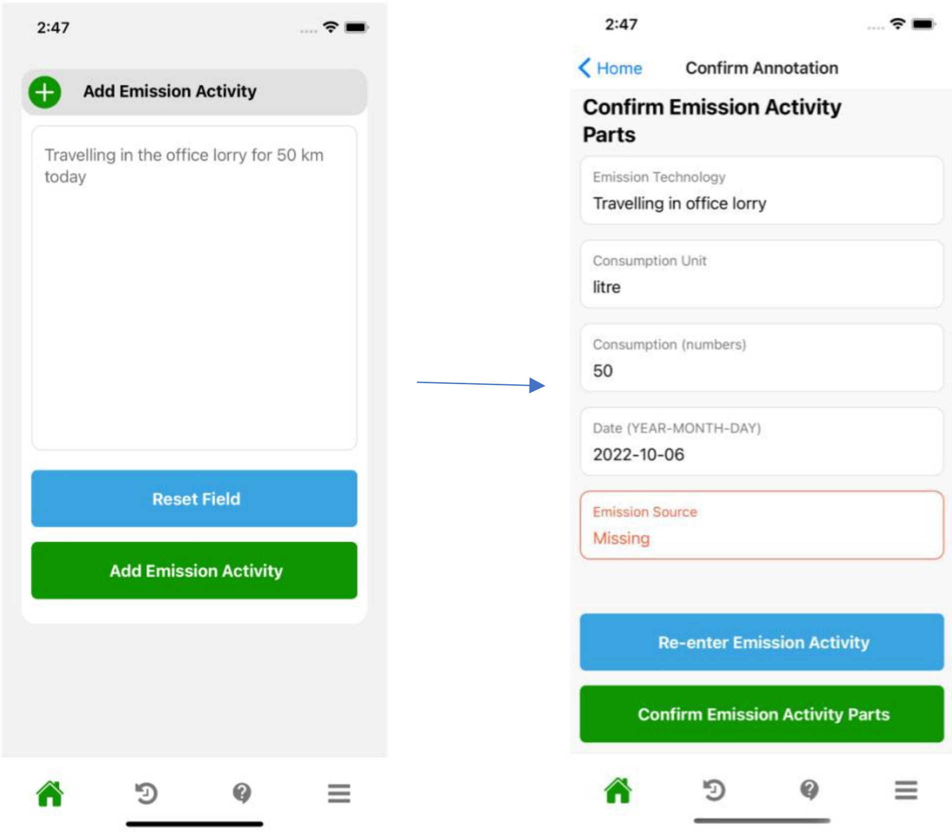

Figure 0-6: Emission parts extraction UI