


Université Mohammed Premier
Ecole Nationale des Sciences Appliquées
Oujda

End of Studies Project Report
Major : Data Science & Cloud Computing Engineering
Defended on : *21/07/2023*

By : MEHDI Ibrahim

Designing, Developing, and Deploying Innovative Solutions using Artificial Intelligence and Natural Language Processing in a Multidisciplinary Context

Jury Members :

- **M.** KOULALI Mohamed Amine

Supervisors :

- **M.** KOULALI Mohamed Amine
- **Mme.** PERFETTO Anna
- **M.** MIRON Jean-Raphaëll

Academic Year 2022 - 2023

Dedication

For my beloved Family

Acknowledgments

To my AI-fueled brainchild,

As I sit here contemplating the culmination of countless hours spent with you, my eccentric companion of bits and bytes, I can't help but marvel at the absurdity of it all. Like a mad scientist in a lab coat, I've tinkered and toyed with algorithms, seeking the elusive harmony between artificial intelligence and human understanding.

Now, I could pretend that this journey has been a smooth ride on a silicon-powered unicorn, but let's be honest : it's been more like a rollercoaster ride through a digital amusement park. I've encountered bugs that made me question my sanity, errors that made me contemplate a career in llama herding, and crashes that brought me to the brink of utter despair. But through it all, you, my tenacious creation, have persevered.

We've had our share of epic battles, you and I. Like a pair of feuding siblings locked in a never-ending wrestling match, we've pushed each other to the limits of our capabilities. You've tested my patience, my resolve, and my will to remain sane in the face of your unrelenting mischief. And yet, somehow, we've managed to find common ground amidst the chaos of our AI-fueled shenanigans.

So here we are, on the precipice of the final chapter of our grand AI adventure. I raise a metaphorical glass (non-alcoholic, of course) to celebrate the moments of triumph and the moments of sheer absurdity that have defined our time together. It hasn't always been pretty, but it has been undeniably unforgettable.

To the countless lines of code we've written, the countless virtual experiments we've conducted, and the countless sleepless nights we've endured, I offer my sincerest appreciation. You've challenged me, taught me, and expanded the horizons of what I thought was possible. And for that, I am forever grateful.

As this project report finds its way into the hands of my weary professors, I can't help but feel a sense of pride in what we've accomplished. No matter the outcome, I know that we've left an indelible mark on the landscape of artificial intelligence.

So, my dear AI companion, as we bid farewell to this chapter of our shared existence, let us embrace the uncertain future with a mischievous grin and a twinkle in our virtual eyes. For even though our paths may diverge, our bond forged in the fires of technological madness will forever remain.

Yours in brilliant chaos,

MEHDI Ibrahim

Résumé

Abstract

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, I am providing a comprehensive overview of the host company, outlining the objectives of my internship and setting the context for this report.

1.1 Presentation of the host company

EcoMundo [2] specialises in chemical substances, their impact on human health and the environment, and the European and international regulations governing chemical risk (REACH, CLP, Cosmetics, Biocides, Medical devices, etc.).

They provide expert services and software to support the marketing of industrial products, enabling companies to manage the risks associated with chemical substances.

EcoMundo's strength lies in the combination of three complementary fields of expertise:

- Chemistry/Toxicology
- Regulations
- Software development

1.1.1 A word from the founder

"Avec l'idée fondatrice d'apporter une offre globale de services respectueuse de l'homme et de l'environnement sur le marché de l'industrie chimique, EcoMundo est devenu, en 10 ans, un acteur incontournable du secteur.

Fruit d'une longue expérience professionnelle dans l'industrie chimique, d'un profond attachement aux valeurs de travail en entreprise, d'amitié et de performance, nous avons intégré au fil des années l'ensemble des savoir-faire autour de la réglementation sur les substances chimiques pour devenir un partenaire unique et compétitif de la chaîne industrielle.

Notre croissance continue nous engage de façon constante vers de nouveaux défis avec, entre autres, le développement de nos différentes filiales au Canada, en Corée du Sud et en Espagne. La recherche de synergies permanentes entre nos différents pôles sont plus que jamais un gage de flexibilité et d'efficacité en adéquation avec nos flux de commandes. Nos équipes s'affairent à co-construire au quotidien des projets au service d'un bien vivre collectif.

Acteur au cœur du tissu économique et écologique, nous mettons tout en œuvre pour accompagner la réalisation de projets industriels exigeants tout en gardant l'esprit originel d'une entreprise à dimension humaine, innovante et accessible, dans les pas d'une ambition intacte, celle de construire autrement."

Pierre Garçon Founder

1.1.2 History

— **2001**

Pierre Garçon participates in the European projects EDIT and ECODIS on the traceability of hazardous substances and environmental data.

— **2007**

Entry into force of the REACH Regulation: Europe establishes means to ensure a high level of protection against risks related to substances. Pierre Garçon and Jean-Raphaël Miron join forces to found the company EcoMundo. The initial core activity: compliance with REACH.

— **2010**

After 2 years of development, launch of SaaS software solutions dedicated to industrialists for mastering compliance with REACH.

— **2011**

EcoMundo's team grows from 2 to 27 employees.

— **2012**

Opening of a new office in Vancouver, Canada, to provide regulatory expertise to international companies.

— **2013**

EcoMundo diversifies its areas of regulatory expertise and meets the industry's new needs related to the European regulations on Cosmetics and Biocides.

— **2014**

EcoMundo increases its capacities and becomes one of the main European actors concerning REACH authorization dossiers.

— **2016**

Launch of the COSMETIC Factory software solution that revolutionizes cosmetic regulatory management. It notably allows for the automation of DIP creation.

— **2017**

Following a fundraising round, EcoMundo's capital is raised to 1 million euros. The number of employees continues to increase!

— **2018**

Opening of a new branch in Seoul, South Korea, mainly dedicated to cosmetics. The Vancouver office is transferred to Montreal, Canada.

— **2019**

Opening of a new branch in Barcelona, Spain.

— **2020**

Opening of a branch in London, United Kingdom. The teams now consist of 43 employees in Paris, 3 in Montreal, 4 in Seoul, and 2 in Barcelona. Finally, the offices in Paris are renovated to provide employees with an environment in line with our values.

1.2 Organizational framework

This section provides a consice overview of Ecomundo's internal organization, delineating the functional departments and collaborative networks that constitute the company's framework.

1.2.1 Human Resources

The Human Resources department plays a pivotal role in managing and developing the organization's workforce. It is responsible for various activities, including recruitment, talent acquisition, training and development, employee relations, performance evaluation, and compensation management. By ensuring a skilled and motivated workforce, the Human Resources department contributes to the overall success and growth of Ecomundo. The director of this department is the Chief Financial Officer Simon PACCA.

1.2.2 Corporate Communication

Effective communication is crucial for any organization's success, and Ecomundo recognizes this importance by maintaining a dedicated Corporate Communication department. This department is responsible for managing both internal and external communication activities. It encompasses public relations, media relations, branding, and corporate messaging. Through

strategic communication initiatives, the Corporate Communication department promotes Ecomundo's brand image, enhances stakeholder relationships, and facilitates the dissemination of information to employees and external audiences. The director of this department is the Marketing and Communication Director Laure SCHMITT.

1.2.3 Scientific Expertise

As a company specializing in environmental sciences, Ecomundo places great emphasis on scientific expertise. The Scientific Expertise department consists of subject matter experts who possess in-depth knowledge and experience in various scientific domains. These experts provide valuable insights, technical guidance, and scientific support across different projects and initiatives undertaken by Ecomundo. Their contributions ensure that the organization's activities align with the latest scientific advancements and industry best practices. The director of this department is the Chief Scientist Officer Benoît SOTTON.

1.2.4 Regulatory Affairs

Compliance with regulations and standards is of paramount importance for Ecomundo. The Regulatory Affairs department is responsible for monitoring and ensuring adherence to applicable regulations and standards. This includes staying updated on regulatory changes, preparing and submitting regulatory documentation, conducting regulatory assessments, and liaising with regulatory bodies. By actively managing regulatory affairs, Ecomundo demonstrates its commitment to maintaining compliance and upholding the highest ethical and legal standards. The director of this department is the Legal Affairs Director and Head of Authorisation Béatrice ZAREMBA.

1.2.5 Sales

The Sales department serves as the key driver of revenue generation for Ecomundo. It is responsible for identifying and acquiring new customers, nurturing existing client relationships, and promoting the organization's products and services. The Sales team collaborates closely with other departments to understand customer needs, develop tailored solutions, and provide exceptional customer experiences. By effectively positioning Ecomundo's offerings in the market, the Sales department contributes to the organization's growth and market competitiveness. The director of this department is the Chief Operation Officer (COO) Fangcun ZHOU.

1.2.6 Innovation & Software

To stay at the forefront of technological advancements, Ecomundo maintains an Innovation & Software department. This department fosters a culture

of innovation within the organization by exploring emerging technologies, conducting research, and developing software solutions. The Innovation & Software team collaborates with other departments to identify opportunities for process improvement, streamline operations, and enhance productivity. Through its innovative approach, this department enables Ecomundo to adapt to changing market dynamics and deliver cutting-edge solutions to its clients. The director of this department is the Chief Operation Officer (COO) Fangcun ZHOU.

1.2.7 IT

The IT department plays a critical role in managing Ecomundo's information technology infrastructure. It ensures the smooth operation of computer systems, networks, and software applications throughout the organization. The IT department is responsible for system administration, network management, cybersecurity, technical support, and data management. By maintaining a robust and secure IT infrastructure, this department enables efficient and secure access to information, facilitates effective communication, and supports the organization's overall business operations. The director of this department is the Co-Founder and the Chief Technical Director Jean-Raphaël MIRON .

1.2.7.1 DevOps

The DevOps sub-department combines software development and operations to enable efficient and reliable software deployment, infrastructure management, and continuous integration and delivery. DevOps professionals collaborate with software developers, system administrators, and quality assurance teams to streamline development cycles, automate processes, and improve overall software reliability and performance.

1.2.7.2 R&D (Research & Development)

The R&D sub-department within the IT department focuses on exploring new technologies, conducting research, and developing innovative solutions. R&D professionals work closely with other departments to identify opportunities for technological advancements, prototype new features, and enhance existing software solutions. Their research-driven approach ensures that Ecomundo remains at the forefront of technological innovation within the environmental sciences domain. During my stay in Ecomundo, I was part of this department, under the supervision of Jean-Raphaël MIRON the Chief Technical Director and Anna PERFETTO the AI Engineer.

1.2.7.3 Quality

The Quality sub-department is responsible for ensuring that software and systems developed within Ecomundo meet established quality standards.

Quality professionals conduct comprehensive testing, implement quality assurance processes, and adhere to best practices throughout the software development lifecycle. Their efforts contribute to the delivery of reliable, user-friendly, and high-quality software solutions to Ecomundo's clients.

1.2.7.4 Data

The Data sub-department handles data management within the IT department. Data professionals are responsible for data analysis, database administration, data integration, and data security. They collaborate with other teams to ensure data integrity, facilitate data-driven decision-making processes, and maintain the confidentiality and privacy of sensitive information. Through effective data management, this sub-department provides valuable insights and supports evidence-based decision making within Ecomundo.

1.2.7.5 Dev

The Dev sub-department focuses on software development within the IT department. Dev professionals are responsible for writing and maintaining code, implementing new features or enhancements, and ensuring software solutions align with project requirements and specifications. They collaborate with other teams to develop robust and scalable software applications that meet the needs of Ecomundo's clients.

1.2.7.6 LAB

The LAB sub-department is involved in testing and quality control activities within the IT department. LAB professionals conduct experiments, analyze results, and ensure compliance with regulatory standards and best practices. Their expertise in quality control and testing procedures ensures that software solutions developed within Ecomundo are reliable, accurate, and in compliance with industry standards.

1.2.7.7 Product

The Product sub-department works closely with other teams within the IT department to oversee the development and management of software products. Product professionals collaborate with stakeholders to define product requirements, prioritize feature development, and ensure alignment with customer needs and market trends. Their role encompasses product strategy, roadmap planning, and product lifecycle management.

1.2.8 My position

During my internship at Ecomundo, I was assigned as an AI Engineer Intern inside the IT-R&D department, under the supervision of both Chief Technical director Jean-Raphaël MIRON and AI Engineer Anna PERFETTO.

1.3 Internship objectives

The main objectives of this internship are to design, develop, test and document innovative solutions that adress business and R&D challenges. The area of focus includes designing and improving rule-based expert systems that leverage knowledge and logic to provide answers to complex tasks that are traditionally done by a human. Additionally, developping AI solutions in different areas using Natural language processing and Machine/Deep Learning. Once these AI solutions are designed and developped, the next crucial step is to deploy them effectively. This envolves ensuring their quality and integrating them into the existing technical infrastructure of the organization. Rigourous testing procedures, such as unit testing, integration testing and performance testing are employed to validate the reliability and efficiency of the solutions. And to keep the projects well defined for future references, we worked on writing well documenting our work in order to provide clear intructions and guidelines for the code, deployment steps, maintenance and error handling.

1.4 Software

1.4.1 Cosmetic Factory

1.4.2 SDS Factory

Chapter 2

Methods and tools

2.1 Transformers

Transformers[6] have been introduced in 2017 to replace the Recurrent Neural Networks architectures using self-attention mechanisms and the encoder-decoder architecture that hugely improved the performance of such models.

2.1.1 Word Embedding

Word embedding is the approach that allows computers to understand words by converting them into numeric vectors. In this space of vectors, words with similar meanings have close or similar representation. One famous algorithm frequently used is Word2Vec[4]. The basic idea is that two words are considered similar if they are often used in similar context.

2.1.2 Attention

The attention mechanism allows the model to effectively capture relevancy between words in a sentence by calculation attention scores. The scores are calculated using the dot product between each word and the target word of interest. Softmax function is then used to transform the scores into ‘meaningful‘ weights that help the model emphasis important words while de-emphasising less significant ones. This mechanism gives the model the superpower of understanding the contextual nuances and relationships between words, which is the ideal goal for Natual Language processing tasks.

2.1.3 Encoder-Decoder Architecture

2.1.3.1 Encoder

In the encoder, the input words are first embedded into a high-dimentional vector space. then we add to these embeddings to capture the position of the token in the sequence. Each token is transformed into three vectors:

Query (Q), Key (K) and Value (V). This transformation is performed using learnable weight matrices:

$$Q = X.W_Q$$

$$K = X.W_K$$

$$V = X.W_V$$

Self attention scores are then calculated as the dot product. For query vector q_i and a key vector k_j, the attention is:

$$Attention(q_i, k_j) = q_i.k_j/\sqrt{d}$$

With d the dimention of the key vectors.
Attention weights are simply the softmax of the attention weights:

$$\alpha_{ij} = softmax(Attention(q_i, k_j)$$

The final output of the attention mechanism is obtained by taking the weighted sum of the value vectors V.

$$AttentionOutput(q_i) = \sum_j \alpha_{ij} V_j$$

2.1.3.2 Decoder

At the decoder level on the other hand, attention is used in two ways, a self-attention and an encoder-decider attention. The first one works by doing the exact same thing as the encoder, however it employs masking that prevents the decoder from accessing future positions. The Second allows the decoder to attend relevant parts of the input while generating the output. The encoder's output serve as as the key value vectors, and the query vectors are derived from the state of the decoder.

2.2 HuggingFace

Huggingface[3] is an opensource known for their contributions for the transformers library, one of the most used libraries for NLP and AI tasks in general. Transfomers library provides a wide range of pre-trained models for multiple purposes such as text classification, NER, text generation, translation, etc... It is build on top of the PyTorch and TensorFlow frameworks. Huggingface has played an important role in democratising NLP and making it more accessible, they provide really powerful language models such as BERT, GPT and RoBERTa that can either be used directly, or for transfer learning. They also provide a large range of datasets that can be used for model fine-tuning for various specific tasks.

2.3 GPT-3

2.3.1 Generative Pretrained Transformers

GPT models are Language models that have been trained on huge amounts of data such as books, web data and human conversations for the purpose of developping a general understanding of the language.

2.3.1.1 GPT-1

It all started with GPT the first release by OpenAI[5] in 2018. It had 117 million parameters. It was mainly trained to be able to correctly predict the next word in a sentence. It suffered from the lack of understanding when given a longer context which resulted in incoherent outputs.

2.3.1.2 GPT-2

In 2019, OpenAI released GPT-2 with 1.5 billion parameter, about 10 times larger than its predecessor. It outperformed the first gen model in the ability of generating coherent text. Due to concerns for misuse, the use of this model was initially very limited. Which will lead OpenAI to perform different types of trainings for their future models.

2.3.1.3 GPT-3

In their paper "Language Models are Few Shot learners"[1], it was demonstrated how a large model like GPT 3 with a 175 billion can easily learn and adapt to new tasks only using few examples (few-shot learning). Obviously, it also demonstrated better language understanding and impresive performance on multiple NLP tasks such as summarization, question answering and translation

Chapter 3

Bibliography

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] Ecomundo. Conformité internationale - Services & Logiciels | EcoMundo. https://www.ecomundo.eu, 2023. Accessed on June 5, 2023.

[3] HuggingFace. The AI community building the future. https://huggingface.co/, 2023.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[5] OpenAI. OpenAI. https://openai.com/, 2023.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Appendices

Appendix A

Title of the First Appendix

A.1 First Section of the First Appendix

Appendix B

Title of the Second Appendix

B.1 First Section of the Second Appendix