University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Szymon Kozłowski
Student no. 448304

Gustaw Blachowski
Student no. 448194

Kamil Dybek
Student no. 448224

Natalia Junkiert
Student no. 448267

Innovative methods of processing data coming from mobile devices for market and scientific research

Bachelor's thesis
in COMPUTER SCIENCE

Supervisor:
Jacek Sroka PhD
Institute of Informatics

Warsaw, February 27, 2025

Abstract

In an era of rapidly evolving digital applications, traditional scraping techniques face increasing challenges in maintaining reliable data collection pipelines. Commissioned by Murmuras, a company specializing in commercial and scientific data analysis [1], this project presents a novel approach to processing phone screen content, such as social media posts and website advertisements. Our solution leverages Large Language Models (LLMs) running locally on the user's device to handle diverse data formats while ensuring that sensitive information remains protected. The primary application explored in this study is the extraction of discount coupons, demonstrating the feasibility of our method in identifying and structuring valuable content from varying digital sources. Furthermore, the system is designed to be easily adaptable to other use cases, such as analyzing users' political views. The results highlight the potential of LLM-driven content analysis as an alternative to conventional scraping techniques.

Keywords

LLM, NLP, BERT, Android, Edge-device, Fine-Tuning

Thesis domain (Socrates-Erasmus subject area codes)

11.4 Artificial Intelligence

Subject classification

I.2.7: Natural Language Processing
H.3.3: Information Search and Retrieval

Tytuł pracy w języku polskim

Innowacyjne metody przetwarzania danych pochodzących z urządzeń mobilnych na potrzeby badań rynkowych i naukowych

Contents

1. Introduction 5
1.1. Project background and motivation 5
1.2. The definition of a coupon 6
1.3. Project goals 6
1.4. Potential applications of the project 6
1.4.1. Assessing coupon effectiveness 6
1.4.2. Market analysis and competitor monitoring 6

2. Machine learning and the dangers associated with it 7
2.0.1. Benchmark 7

3. Overview of the existing solutions 9
3.1. Murmuras' existing solution 9
3.2. Scapegraph AI 10

Bibliography 11

Chapter 1

Introduction

1.1. Project background and motivation

With the rapid advancement of information technology, the internet has become one of the most crucial facets for many businesses to perform marketing activities [4]. One of the key marketing tools in business-to-consumer (B2C) e-commerce is the digital coupon (also referred to as an electronic coupon) [5]. In comparison to paper coupons, digital coupons are characterized by their wide reach, rapid distribution, and low spread costs. Furthermore, a key advantage of digital coupons is their ability to facilitate targeted marketing by offering personalized discounts to different customers, thereby increasing sales [4]. To maximize the benefits of digital coupons, it is essential for businesses to assess the effectiveness of their coupon campaigns, evaluate their reach, and analyze their competitors' strategies. By tracking key performance metrics such as redemption rates, customer engagement, and sales impact, businesses can refine their marketing approaches to optimize results. Additionally, studying competitors' digital coupon strategies enables businesses to identify market trends, adjust their promotional tactics, and maintain a competitive edge in the evolving digital marketplace.

Machine learning has rapidly become a central focus in computer science research, offering powerful capabilities in pattern recognition and information extraction from unstructured data. This advancement has led to the development of models that can learn relevant features from large datasets, reducing reliance on heuristic-based algorithms that require extensive parameter tuning and handcrafted rules. Such models are particularly effective in handling the variability inherent in real-world data [7], including diverse coupon designs.

Recent statistics underscore the significance of mobile devices in this domain. For example, studies have shown that over 90% of digital coupon users access their vouchers via smartphones [8], and similar figures are reported by other industry sources [9]. This high rate of mobile usage creates a pressing need for coupon analysis tools that are optimized for mobile platforms, ensuring that consumers receive timely and personalized offers regardless of their location or device.

In light of these trends, the company Murmuras has tasked us with developing a machine learning model that can be deployed as a mobile application. This model will process input representing the user's onscreen view and extract digital coupons along with their relevant data. This solution must be capable of running locally on the device, ensuring efficient processing without relying on external servers. By leveraging advanced machine learning techniques, the app will handle the diverse formats and layouts of digital coupons, thereby facilitating the collection of data regarding coupons.

1.2. The definition of a coupon

A coupon is a physical piece of paper or digital voucher that can be redeemed for a financial discount when purchasing a product [2]. A coupon is characterized by a name, expiration date, and a discount type (e.g. '20% off', 'buy 1 get 1 free', etc.), however, not every coupon contains each of these features. Furthermore, coupons may contain numerous other features such as images and eligibility requirements. Henceforth, the term 'coupon' will refer exclusively to a digital coupon.

1.3. Project goals

1. A tool to process the data extracted from the device into a format suitable for use by the model.
2. A machine learning tool for extracting the data that is of interest to us, such as the coupon name, expiration dates, prices, etc. The model should be capable of handling various coupon formats and layouts with high accuracy.
3. An optional tool for post-processing the output data from the tool mentioned in the previous point into a common format.
4. An application that runs the above three tools on a mobile device. (Optional)
5. A key requirement is that the machine learning model must be deployable on the mobile device itself to guarantee data privacy.

1.4. Potential applications of the project

1.4.1. Assessing coupon effectiveness

Our solution will aid businesses in analyzing consumer behaviour and optimizing their marketing strategies accordingly. By facilitating the collection of data on coupon characteristics and their redemption rates, businesses will be able to assess the effectiveness of their coupon campaigns—determining whether they reach the intended audience and achieve the desired results. Additionally, large-scale analysis of coupon data can reveal valuable insights into purchasing patterns, preferred discount types, and the most appealing products or services for different customer segments. With this information, businesses can refine their promotional strategies, tailor offers to specific demographics, and enhance overall customer engagement.

1.4.2. Market analysis and competitor monitoring

The aforementioned gathering of data can also be utilized to monitor competitors' coupon strategies, their effectiveness, and whether they provide better discounts. Using machine learning to identify and analyze competitors' strategies is more cost-effective compared to exhaustive web scraping or mystery shopping [6]. This will enable businesses to make better-informed decisions about their own marketing campaigns and provide a comprehensive understanding of the competitive landscape.

Chapter 2

Machine learning and the dangers associated with it

Note: this chapter is a work in progress, bullet points aim to provide guidance when writing this section

(1) What is the difference between machine learning, artificial intelligence, and deep learning?
(1a) Provide the definitions/a brief explanation of each of the above.
(1b) Explain what a benchmark is and what it is used for.

2.0.1. Benchmark

Benchmarking is the process of running a set of, among others, computer programs against a set of tests to assess their relative performance or precision [3].

(2) Understanding ML models
(2a) Explain what a model is
(2b) Explain how a model works, how it is trained, datasets, linear regression, supervised vs unsupervised learning (?), federated learning (?), computer vision (?)
(2c) What is quantization and why it is this of interest to us

(3) What is NLP
(3a) Explain what NLP is and why it is of interest to us for this project
(3b) BERT, Llama, ChatGPT and other models (briefly explain their differences, advantages and disadvantages, paramters, memory usage (?))

(4) Should we be scared of AI?
(4a) https://www.youtube.com/watch?v=yh1pF1zaauc. (from our mentor)
(4b) Privacy and ethics of data collection and processing (present the problem, why people are concerned about this, then later on in the document we say that we resolved this issue because we are processing the data locally etc)
(4c) Adversarial attacks (I'm not sure this is particularly relevant to our project but it mught be worth mentioning)
(4e) Accuracy concerns, how can we be sure that our model is correct? Lack of human oversight
(4f) Environmental concerns // HF tutorial: env concerns => fine tune not training

Chapter 3

Overview of the existing solutions

To our knowledge, as of writing this thesis, there are no publicly available solutions that directly address this problem. The most comparable approaches involve existing multimodal models. While widely used models like ChatGPT and Gemini offer some relevant capabilities, they are not highly precise for this specific task. A major limitation of such models is their large size—for instance, GPT-3 has 175 billion parameters[10]—making them impractical for mobile deployment [16].

Alternatively, Computer Vision models exist for extracting text and bounding boxes from screen images. Microsoft's OmniParser [12], for example, performs well in this area but still requires preprocessing similar to our approach. Moreover, our experiments running OmniParser locally indicate that it depends on CUDA technology, making it unsuitable for mobile deployment [**?**].

3.1. Murmuras' existing solution

Murmuras' current approach relies on fixed scripts tailored to specific applications, making it inflexible and difficult to generalize across diverse coupon formats. This lack of adaptability limits its usefulness in real-world scenarios where coupon structures vary widely. Since our goal is to develop a solution that is easily adaptable for processing diverse mobile content, this method is not well-suited for our needs.

In contrast, Murmuras' most-recent proof of concept involves basic preprocessing of the extracted data before sending it to GPT-4o-mini for further processing. This approach leverages an LLM to interpret the data to extract relevant coupon details. However, the reliance on an external server means the solution does not run locally on the mobile device, leading to potential privacy concerns, latency issues, and a dependence on internet connectivity.

Additionally, the accuracy of this method is suboptimal. According to their own benchmarks, the average similarity ratio is only 56.49%, indicating significant inconsistencies in the extracted data. This benchmark measures the accuracy of extracted coupons by comparing them to expected ground-truth values across five key attributes: product name, discount percentage, old price, new price, and coupon validity.

The evaluation process involves using difflib's SequenceMatcher, a text similarity algorithm that computes the ratio of matching sequences between two strings [17]. For each extracted coupon, the similarity scores of its five attributes are individually compared against the ground-truth values. The results are then averaged to generate an overall similarity percentage. A low similarity ratio indicates significant variations between the extracted and expected coupon details, highlighting challenges in precise text interpretation and extraction.

3.2. Scapegraph AI

ScrapeGraphAI is an open-source Python library that streamlines data extraction from websites and local documents by utilizing LLMs and graph logic to construct efficient scraping pipelines. This approach automates data extraction, reducing the need for extensive manual coding. The library supports integration with various LLMs, including local models [11]. For instance, users have configured ScrapeGraphAI to work with local models like those served through vLLM [13] or Ollama [14].

However, this solution does not address the issue of deploying such models directly on mobile devices. This presents significant challenges since mobile devices typically have limited processing power and memory compared to desktop computers or servers [15].

Bibliography

[1] *Murmuras website*. https://murmuras.com/. [Accessed 2025-02-11].

[2] *Britannica Dictionary definition of COUPON*. https://www.britannica.com/dictionary/coupon. [Accessed 2025-02-03].

[3] *Computer Benchmark*. https://bhatabhishek-ylp.medium.com/benchmarking-in-computer-c6d364681512. [Accessed 2025-02-03].

[4] Xiong Keyi, Yang Wensheng *Research on the Design of E-coupons for Directional Marketing of Two Businesses in Competitive Environment*. https://www.sciencepublishinggroup.com/article/10.11648/j.ijefm.20200801.16. [Accessed 2025-02-04].

[5] Li Li, et. al. *Targeted reminders of electronic coupons: using predictive analytics to facilitate coupon marketing*. https://link.springer.com/article/10.1007/s10660-020-09405-4. [Accessed 2025-02-04].

[6] Bernhard König, et. al. *Analysing competitor tariffs with machine learning*. https://www.milliman.com/en/insight/analysing-competitor-tariffs-with-machine-learning. [Accessed 2025-02-04].

[7] Iqbal H. Sarker *Machine Learning: Algorithms, Real-World Applications and Research Directions*. https://link.springer.com/article/10.1007/s42979-021-00592-x. [Accessed 2025-02-05].

[8] Sara Lebow *How consumers access digital coupons*. https://www.emarketer.com/content/how-consumers-access-digital-coupons. [Accessed 2025-02-05].

[9] *Unveiling IT Coupons Trends and Statistics*. https://www.go-globe.com/unveiling-it-coupons-trends-statistics/. [Accessed 2025-02-05].

[10] *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* // Would be great to get the link and change this into APA

[11] Satyam Tripathi *ScrapeGraphAI Tutorial - Getting Started with LLMs Web Scraping* https://scrapingant.com/blog/scrapegraphai-llms-web-scraping

[12] *Yadong Lu, Jianwei Yang, Yelong Shen, and Ahmed Awadallah. Omni- parser for pure vision based gui agent, 2024.*

[13] *Can not Set Model Tokens to Local Model with OpenAI API Format #810* https://github.com/ScrapeGraphAI/Scrapegraph-ai/issues/810

[14] *Can't load tokenizer for 'gpt2' #752* https://github.com/ScrapeGraphAI/Scrapegraph-ai/issues/752

[15] Xiang Li, et. al. *Large Language Models on Mobile Devices: Measurements, Analysis, and Insights* https://dl.acm.org/doi/10.1145/3662006.366205

[16] Junchen Zhao, et. al. *LinguaLinked: A Distributed Large Language Model Inference System for Mobile Devices* https://arxiv.org/pdf/2312.00388

[17] *difflib — Helpers for computing deltas* https://docs.python.org/3/library/difflib.html