DECLARATION

We VASU TOMER 20BCE2333 SIDDHARTH SARKAR 20BDS0202 ABHIJAY THAKUR 20BCE2375 hereby declare that the thesis entitled "smart taxation using natural language processing" submitted by us, for the award of the degree of *Bachelor of Technology in Computer Science* to VIT is a record of bonafide work carried out by me under the supervision of KANNADASAN R.

We further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place : Vellore

VASU TOMER
SIDDHARTH SARKAR
ABHIJAY THAKUR

Date :8/05/2024

Name & Signature of the Candidates

CERTIFICATE

This is to certify that the thesis entitled "smart taxation using natural language processing" submitted by VASU TOMER 20BCE2333, SIDDHARTH SARKAR 20BDS0202, ABHIJAY THAKUR 20BCE2375, SCOPE VIT, for the award of the degree of *Bachelor of Technology in Computer Science*, is a record of bonafide work carried out by them under my supervision during the period, 01. 02. 2024 to 8.05.2024, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore
Date :8/05/2024

9/5

Signature of the Guide

Internal Examiner

External Examiner

Head of the Department
Btech CSE(CORE,DATA
SCIENCE
)

SMART TAXATION USING NATURAL LANGUAGE PROCESSING

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Computer Science (CORE)

by

VASU TOMER 20BCE2333

SIDDHARTH SARKAR 20BDS0202

ABHIJAY THAKUR 20BCE2375

Under the guidance of

Prof. KANNADASAN R

SCOPE VIT,

Vellore.

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

MAY 2024

DECLARATION

We VASU TOMER 20BCE2333 SIDDHARTH SARKAR 20BDS0202 ABHIJAY THAKUR 20BCE2375 hereby declare that the thesis entitled "smart taxation using natural language processing" submitted by us, for the award of the degree of *Bachelor of Technology in Computer Science* to VIT is a record of bonafide work carried out by me under the supervision of KANNADASAN R.

We further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place : Vellore

VASU TOMER
SIDDHARTH SARKAR
ABHIJAY THAKUR

Date :8/05/2024

Name & Signature of the Candidates

CERTIFICATE

This is to certify that the thesis entitled “smart taxation using natural language processing” submitted by VASU TOMER 20BCE2333, SIDDHARTH SARKAR 20BDS0202, ABHIJAY THAKUR 20BCE2375, SCOPE VIT, for the award of the degree of *Bachelor of Technology in Computer Science*, is a record of bonafide work carried out by them under my supervision during the period, 01. 02. 2024 to 8.05.2024, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore

Date :8/05/2024

Signature of the Guide

Internal Examiner

External Examiner

Head of the Department

Btech CSE(CORE,DATA

SCIENCE

)

ACKNOWLEDGEMENTS

We extend our sincere gratitude to Prof.**KANNADASAN R** of the School of Computer Science Engineering and Technology at VIT University for his invaluable guidance and mentorship during the development of our Capstone Project, " smart taxation using natural language processing." His expertise, constructive feedback, and unwavering support have been instrumental in shaping the success of our project and enhancing our overall learning experience.

We also express our thanks to VIT University for providing the necessary infrastructure, resources, and a conducive learning environment. The university's commitment to academic excellence has played a crucial role in our project's successful execution.

Our appreciation extends to all the faculty members of the School of Computer Science Engineering and Technology for their valuable insights, encouragement, and continuous support. Their expertise has enriched our academic journey and contributed significantly to the success of smart taxation using natural language processing.

Special thanks to our peers and classmates for their collaboration, brainstorming sessions, and mutual support. The exchange of ideas and collaborative spirit within the department have been pivotal in overcoming challenges and achieving our goals.

We would also like to express our gratitude to our families and friends for their unwavering support, understanding, and encouragement throughout the course of our project. Their belief in our abilities has been a constant source of motivation.

This project would not have been possible without the collective effort and support from all these individuals and entities. We are grateful for the opportunity to undertake this endeavor and look forward to applying the knowledge and skills gained in our future endeavors.

Thank you.

Sincerely,

[VASU TOMER, SIDDHARTH SARKAR , ABHIJAY THAKUR]

[20BCE2333, 20BDS0202, 20BCE2375]

B.Tech in Computer Science and Engineering

VIT University

Student Names:

VASU TOMER

SIDDHARTH SARKAR

ABHIJAY THAKUR

Executive Summary

This paper introduces a groundbreaking method for automating and simplifying tax computation in India, leveraging advanced technologies like machine learning (ML), natural language processing (NLP), and optical character recognition (OCR). Our system is designed to accurately interpret and process financial data extracted from users' bank statements, streamlining tax calculation in line with the Indian tax framework. The proposed approach employs NLP techniques for efficient data extraction, cleaning, sorting, and categorization, facilitating accurate tax computations. This method significantly addresses the complexities and limitations of existing tax calculation systems, offering a more streamlined process. By automating tax computation, we aim to reduce human error, save time and effort for taxpayers, and deter fraudulent activities. Despite its promising outcomes, our system acknowledges certain limitations and areas for enhancement. These include improving OCR capabilities for more precise financial data extraction, adding multilingual support to cater to India's diverse linguistic landscape, and effectively handling complex tax scenarios, particularly for businesses. Additionally, the paper highlights the need for improved scalability to manage large volumes of data and stringent measures to safeguard privacy and security. Our ongoing research and development efforts are focused on overcoming these challenges. Enhancing OCR accuracy is pivotal, as it forms the foundation for reliable data extraction from bank statements. Addressing multilingual challenges is crucial, given India's vast linguistic diversity. This involves not only recognizing different scripts but also accurately interpreting financial terminologies in multiple languages. Handling complex tax scenarios is another critical aspect. The system must be adept at navigating the nuances of various business transactions and tax implications, which requires sophisticated algorithmic approaches and a deep understanding of the Indian tax system. Moreover, scalability is essential for the system to efficiently process the vast amount of data generated by millions of taxpayers. Privacy and security are paramount, considering the sensitive nature of financial data. Our system incorporates robust security measures to protect user data and ensure compliance with relevant data protection regulations. The potential impact of our tax computation system is significant. By automating and simplifying tax calculations, it aims to transform the tax filing experience in India, making it more efficient, accurate, and user-friendly. This can lead to increased compliance, reduced administrative burdens, and a more transparent tax system. Our work contributes to the broader field of financial technology and showcases the potential of AI-driven solutions in addressing complex regulatory challenges.

Keywords- Optical Character Recognition, Tax Calculation, Natural Language Processing, Machine Learning, Data Privacy, Scalability, Multilingual Support.

CONTENTS

Page No.

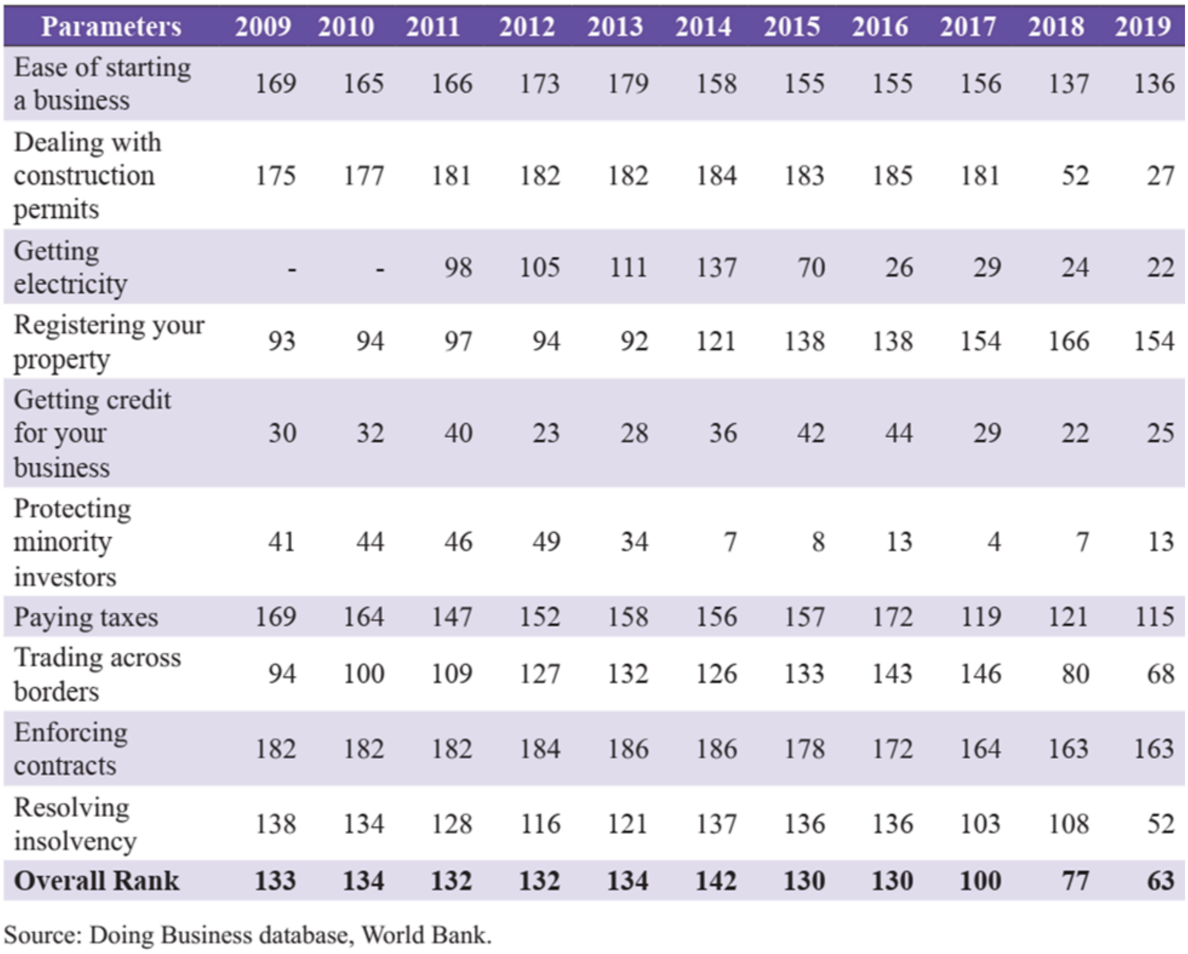
LIST OF FIGURES

1. INTRODUCTION

1. Background

India has emerged as the world's fastest expanding economy, and with this rapid expansion has come a painfully complicated tax system.Filing taxes is an integral part of financial planning for any individual or business.In India, navigating through the process of filing taxes can be very complex, with various forms to be filled ,documents to be submitted and rules to be followed. According to a yearly 'Ease of Doing Business' report by World Bank, India ranks 115th out of 190 countries [1] in the ease of paying taxes, highlighting the country's painfully complex tax system.The current landscape of taxation in India is characterized by a labyrinth of regulations and procedures, posing considerable challenges for individuals and businesses alike. The process of filing taxes involves navigating through an intricate web of forms, submitting an array of documents, and adhering to a multitude of rules. As a result, the country's taxpayers often find themselves grappling with the complexity of the system, leading to delays, errors, and a heightened sense of financial uncertainty.

The Indian government is very aware of this problem as it has been taking steps to simplify and ease the tax filing process. The Income Tax Department of India,keeping up with digitisation of the nation, launched the electronic tax filing system of Income Tax Returns because the lion's share of revenue of the country is generated by direct taxes. E-taxation scheme was one of the "action lines" introduced in Indian tax machinery in the A.Y. 2006-07 for all assessments for improving the Return filing system [2].The past year Indian government also increased the threshold for tax exemption, an attempt to reduce the burden on lower income households.Despite these efforts by the government to simplify the process, filing taxes remains a daunting task for many individuals and businesses alike. The process of filing taxes involves gathering and organizing various documents such as bank statements, rent receipts, and salary receipts. The task of categorizing and extracting relevant information from these documents can be time-consuming and prone to errors. Additionally, fraudulent activities such as misrepresentation of income or false deductions can lead to incorrect tax calculations. These challenges make tax filing a complicated task for individuals and businesses, leading to confusion and delays in the process. The impact of the convoluted tax system extends beyond bureaucratic headaches. Individuals and businesses experience the brunt of this complexity in the form of heightened stress, financial losses due to inadvertent errors, and delays in the processing of returns. As the World Bank's report indicates, the 115th position in ease of paying taxes reflects not only on administrative hurdles but also on the tangible impact on the financial well-being of taxpayers. For a massive economy as India has become today this is a major disadvantage as well as a huge disappointment.



Parameters	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Ease of starting a business	169	165	166	173	179	158	155	155	156	137	136
Dealing with construction permits	175	177	181	182	182	184	183	185	181	52	27
Getting electricity	-	-	98	105	111	137	70	26	29	24	22
Registering your property	93	94	97	94	92	121	138	138	154	166	154
Getting credit for your business	30	32	40	23	28	36	42	44	29	22	25
Protecting minority investors	41	44	46	49	34	7	8	13	4	7	13
Paying taxes	169	164	147	152	158	156	157	172	119	121	115
Trading across borders	94	100	109	127	132	126	133	143	146	80	68
Enforcing contracts	182	182	182	184	186	186	178	172	164	163	163
Resolving insolvency	138	134	128	116	121	137	136	136	103	108	52
Overall Rank	**133**	**134**	**132**	**132**	**134**	**142**	**130**	**130**	**100**	**77**	**63**

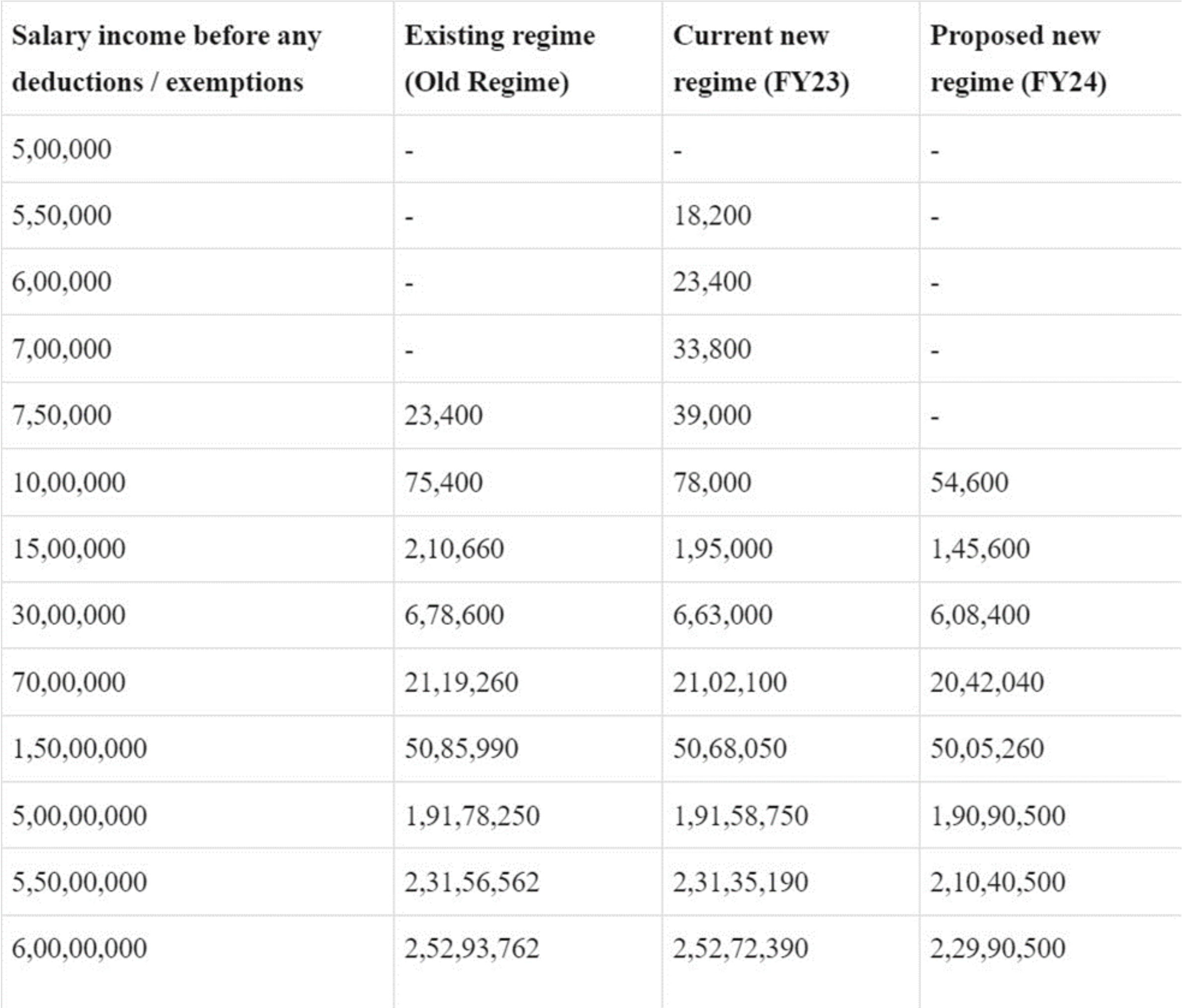
Source: Doing Business database, World Bank.

Fig 1: Doing Business Rankings of India over the years

Automating the process of tax calculation using technology can help alleviate the burden and make it easier for people to file their taxes. A study conducted for the IMF in 2014 provides some insightful analysis of the challenges faced by modern tax administrations in achieving optimal tax collections while minimising administration and compliance costs [3]. The authors of this working paper argued that self-assessment is the most effective means of achieving voluntary compliance, and that tax administrations need to consistently apply self-assessment principles in their income tax laws.This recommendation is consistent with a 2019 study conducted by Atlanta press, which also emphasised the importance of self assessment in the digital age of tax administration. [4] However, the authors note that many tax administrations continue to rely heavily on desk auditing, with risk management practices being largely underdeveloped or underutilised.The evolution of tax compliance technologies has been a dynamic journey marked by significant milestones. From manual data entry to the advent of automated tax systems, there's a clear trajectory toward efficiency and accuracy. Understanding this historical progression provides valuable insights into the challenges faced by existing solutions and underscores the need for innovative approaches, such as the one proposed in this project, to address the persisting complexities in tax computation.

Table 2:

Fig 2: Estimated Tax payable under new regime



Salary income before any deductions / exemptions	Existing regime (Old Regime)	Current new regime (FY23)	Proposed new regime (FY24)
5,00,000	-	-	-
5,50,000	-	18,200	-
6,00,000	-	23,400	-
7,00,000	-	33,800	-
7,50,000	23,400	39,000	-
10,00,000	75,400	78,000	54,600
15,00,000	2,10,660	1,95,000	1,45,600
30,00,000	6,78,600	6,63,000	6,08,400
70,00,000	21,19,260	21,02,100	20,42,040
1,50,00,000	50,85,990	50,68,050	50,05,260
5,00,00,000	1,91,78,250	1,91,58,750	1,90,90,500
5,50,00,000	2,31,56,562	2,31,35,190	2,10,40,500
6,00,00,000	2,52,93,762	2,52,72,390	2,29,90,500

Despite the growing popularity of automated tax computation systems, there are currently few real-world solutions that can accurately and swiftly compute taxes using relevant financial facts. Several modern solutions created by companies like Groww, Clear, and Turbotax rely on manually input data, which can be time-consuming and error prone. Automated tax systems or online tax systems have also been introduced in a number of nations; similar to previous examples they rely on manually input data and place a greater emphasis on data discrepancies and tax fraud detection than on simplifying taxpayers' tax filing processes [5] . Capturing India's decade long journey in the Doing Business rankings article Automated Tax Return Verification using Blockchain Technology suggests using Blockchain technology as a way to automate tax verification in Bangladesh, a nation with a tax filing system comparable to India's and one that has similar problems [6]. On the other end a paper Tax compliance and privacy rights in profiling and automated decision making argued about how automated decision making in tax matters are included in the broader public interest exception, safeguards to taxpayers' privacy rights need to be in place [7].

Google has also made an effort to address this issue by creating Document AI. Compared to conventional approaches, document AI has a number of benefits, including quicker processing times, more accuracy, and a lower chance of fraud. Invoices, receipts, and contracts are just a few examples of the sorts of data that may be extracted and categorised using document AI. This data can then be utilised for a variety of tasks, including tax compliance, auditing, and financial analysis. Nevertheless, there are also drawbacks to using Document AI, including the requirement for a sizable quantity of training data, the possibility of bias in the training data, and the possibility of OCR mistakes [8].The advent of Document AI highlights the industry's attempts to address tax complexity, but challenges such as the need for significant training data and the potential for biases underscore the necessity for a more comprehensive solution.

In order to overcome these difficulties, we combine machine learning (ML) and natural language processing (NLP) techniques to create a novel strategy for autonomously calculating taxes in India. We will start by using a method known as document scan categorization because we will be working with paper-based data. Document scan classification is a way of categorising scanned documents based on their content. Large numbers of scanned documents may be organised, saved, and identified using the classification process. We will utilise OCR, or optical character recognition, to implement this approach. OCR may be used to scan text documents and aid in text categorization by extracting text from the documents and applying it for text classification. With OCR technology, handwritten or printed text may be transformed into machine-encoded text. OCR has been used extensively for many years in a variety of businesses, primarily for digitising text documents. In one study, Javier Ferrando (2020) employed OCR to extract text from scanned documents, and the obtained text was subsequently used for document classification. The study showed an accuracy rate of 89.47 when classifying the text using EfficientNet models [9]. The incorporation of OCR technology represents a pivotal step in our proposed solution, allowing for the transformation of physical documents into machine-encoded text for efficient categorization and classification.

The required financial information may, however, be difficult to extract from these papers due to the possibility of unstructured writing. Using NLP approaches, this problem may be solved by automatically identifying the texts and extracting the pertinent financial data. Natural Language Processing (NLP) is an area of artificial intelligence that deals with the interaction of computers and humans through natural language. By utilising several strategies and techniques including Named Entity Recognition, Sentiment Analysis, and Information Extraction, NLP may be utilised to categorise documents and extract pertinent financial data that can aid in computing an individual's tax [10].The integration of NLP techniques addresses the challenge of unstructured data, allowing for the automatic identification and extraction of relevant financial information from text, thus facilitating more accurate tax computations.

In the midst of these challenges, our project takes a user-centric approach to redefine the tax computation experience in India. Understanding that tax filing can be a daunting task, we are developing a user-friendly website that leverages machine learning (ML) and natural language processing (NLP) techniques. The website aims to simplify the tax computation process for individuals and small businesses with limited financial resources. By offering an intuitive interface, clear instructions, and personalized features, our goal is to make the tax filing process more accessible and less burdensome. Additionally, the website will provide users with valuable insights into their financial situations, offering personalized advice based on each person's unique circumstances. This user-centric design philosophy extends to every aspect of our project, ensuring that technology serves the needs of the end-users effectively.

In India, particularly for individuals and small businesses with limited financial resources, we believe that our recommended method has the potential to significantly improve tax computation efficiency and accuracy. By automating the tax computing process and implementing fraud detection tools, we can decrease the likelihood of errors and fraud while simultaneously lowering the time and effort required by taxpayers. Furthermore, by combining OCR, supervised learning, natural language processing, and machine learning technologies, tax computations may be done more precisely, and personalised advise based on each person's unique financial situation may be offered.

2. Problem Statement

India's rapid economic growth is hindered by a complex tax system, particularly during the tax filing process for individuals and businesses. Despite digital initiatives, the existing landscape remains burdened by intricate regulations, leading to delays, errors, and financial uncertainty for taxpayers. Manual tasks, such as document categorization and data extraction, contribute to inefficiencies and potential errors. Current automated tax systems have limitations, and the adoption of innovative technologies like Blockchain and Document AI is sparse.

In response to these challenges, this project proposes a novel approach utilizing OCR, supervised learning, and NLP techniques. By harnessing the power of ML, the project aims to automate tax computation, minimize errors, and enhance accuracy. A user-centric website design further seeks to simplify the tax filing experience for individuals and small businesses, offering clear instructions and personalized features. The integration of technology aims to streamline the tax computation process, contributing to a more efficient and user-friendly tax system in India.

3. Objectives

The primary objective of this capstone project is to design and implement an innovative tax computation system tailored specifically for Indian citizens. Our aim is to revolutionize the current tax calculation process, transitioning from traditional manual methods to a more automated and simplified approach.

To achieve this, we will:

- **Integrate Advanced Technologies:** Utilize the capabilities of Machine Learning (ML), Natural Language Processing (NLP), and Optical Character Recognition (OCR) to develop a comprehensive solution that can accurately interpret and process financial data.

- **Focus on Data Interpretation:** Harness the power of OCR technologies to extract crucial financial data from a user's bank statement, ensuring that the system is not just automated but also precise in its interpretations.

- **Simplify the User Experience:** By leveraging the combined strengths of ML, NLP, and OCR, our system will aim to provide a seamless experience for the user, reducing the complexities involved in tax computations and ensuring a more straightforward approach to tax filing for Indian citizens.

In essence, this capstone project seeks to bridge the gap between the intricate tax system in India and the average taxpayer, providing a tool that is both technologically advanced and user-friendly.

4. Scope And Limitations

It's important to understand the scope of what we as a team can achieve . Here's a breakdown of what can be done and what might be more challenging:

What We Can Do:

- **Develop a Robust Website:** We will create a website to serve as the user interface for your tax calculation system. With our current web development skills, we can design an intuitive and user-friendly interface for users to interact with your software.

- **Create a Prototype:** We will develop a functional prototype of our tax calculation system using NLP, machine learning, and OCR technologies. This prototype can demonstrate the core functionalities of the system, including data extraction, categorization, and tax calculation.

- **Test and Iterate:** We will conduct extensive testing and iterative development to refine our prototype. This includes testing with different types of financial documents, data sources, and scenarios to improve accuracy and reliability.

- **Incorporate Basic Security Measures:** Implement basic security measures to protect user data, such as authentication, and access controls, to ensure the privacy and security of financial information.

- **Address Bias and Fairness:** We will take steps to identify and mitigate bias in our prototype's algorithms and data. Regular audits and adjustments can help improve fairness in categorization and taxation outcomes.

What Might Be More Challenging:

Scalability: Developing a system that can handle a high volume of financial documents and users in a real-world industry setting can be complex. Scaling up to meet industry-level demand often requires more extensive resources and infrastructure.

- **Production-Ready Software:** Developing a fully production-ready tax calculation software that meets industry standards and compliance requirements is a significant undertaking. It involves robust testing, security assessments, and compliance with relevant financial and tax regulations.

- **Integration with Tax Authorities:** Integrating our software with tax authorities' systems for real-time tax filing and compliance can be complex due to regulatory and legal considerations.

- **Data Privacy Compliance:** Ensuring full compliance with data privacy laws, such as India's data protection regulations, can be challenging. We will need legal and compliance expertise to navigate these requirements.

- **Ongoing Maintenance:** Maintaining and updating our software to adapt to changing tax laws and regulations is an ongoing task that requires dedicated resources and expertise.

As a small developer group, we can make significant progress by creating a functional prototype and a user-friendly website to showcase our tax calculation system's capabilities. However, for a full-scale, industry-ready solution, we will need to collaborate with experts in tax law, data privacy, scalability, and compliance to address the complex challenges involved in deploying such a system in a real-world industry setting.

CHAPTER 2: LITERATURE REVIEW

The pursuit of optimizing and automating tax systems has been a subject of numerous studies and innovations. This literature review collates key insights from various sources, elucidating the evolving narrative of technology's role in tax matters and the parallel importance of safeguarding taxpayers' privacy rights.

- **Blockchain and Tax Verification:**

  The potential of Blockchain in transforming tax verification processes is underlined in the study titled "Automated Tax Return Verification using Blockchain Technology." The study underscores Bangladesh as a case study—a country with tax intricacies akin to India. Blockchain's decentralized and immutable characteristics are heralded as game-changers for tax verification, offering enhanced transparency and minimizing opportunities for manipulation or fraud.

  However, amidst this optimism, it is crucial to recognize the counterpoint. “Challenges of Taxation and Blockchain Technology”[11]delves into the challenges posed by the increased use of technologies like Blockchain in tax systems. Traditional legal foundations, rooted in tangible proof and official documentation, may face conflicts with the decentralized and immutable characteristics of Blockchain. While the study applauds Blockchain for enhancing transparency and minimizing manipulation or fraud, the counterpoint suggests that the reliance on paper-based and officially recognized proofs, integral to established legal constants, may clash with the evolving landscape of digital transactions. As we explore the transformative potential of Blockchain, it becomes imperative to navigate and address the inherent tensions between traditional legal frameworks and the disruptive nature of emerging technologies in the realm of tax verification. This dynamic landscape requires careful consideration and adaptation of tax systems to accommodate both traditional legal constants and the advancements brought forth by technologies like Blockchain.

- **Automated Decision Making and Privacy:**

The nuanced perspective on tax compliance and privacy rights, as meticulously explored in the thought-provoking article 'Tax Compliance and Privacy Rights in Profiling and Automated Decision Making', adds a crucial layer to our understanding of the challenges associated with the implementation of automated tax systems. This insightful study not only acknowledges the efficiency gains brought about by automated decision-making in tax matters but also underscores the necessity of considering the broader public interest exception. By delving into the intricacies of profiling and automated decision-making, the paper raises awareness about potential privacy infringements and calls for the implementation of robust safeguards. These insights are particularly pertinent to our project, which aims to automate tax computation while carefully navigating the fine line between enhanced efficiency and the protection of individual privacy rights. Aligning with the concerns highlighted in the paper, our project takes inspiration from the need to strike a delicate balance, ensuring that the implementation of advanced technologies in tax computation remains ethically sound and respects the privacy rights of taxpayers.

- **Document AI and Tax Computation:**

In the realm of tech giants, Google's foray into this sector via Document AI stands out. Contrary to traditional methodologies, Document AI offers a suite of advantages that revolve around expeditious processing times, heightened accuracy, and a diminished scope for fraudulent activities. The technology is proficient in extracting and categorizing data from diverse documents like invoices, receipts, and contracts. Beyond just extraction, this data becomes pivotal for various operations, encompassing tax compliance, auditing, and even intricate financial analyses.

The evolving literature in the tax automation domain underscores a dual narrative. On one hand, there's a push for embracing cutting-edge technologies like Blockchain and AI to reform complex tax systems, as seen in countries like India and Bangladesh. On the other, there's an equally strong emphasis on ensuring that the very efficiencies brought about by automation do not infringe upon individual privacy rights. Balancing these two aspects will define the trajectory of future innovations in tax administration.

CHAPTER 3: APPROACH TO DESIGN/ METHODOLOGY

1. Methodology

A. Data Extraction:

In our journey to revolutionize tax computation, the first step is akin to assembling puzzle pieces - gathering data from bank statements. Most of these come in PDFs, which, let's be honest, can be a bit of a jumble to decode. That's where Tablo, our trusty tool, steps in. It's like having a skilled detective who can pick out the crucial tables from each PDF page.

Imagine each bank statement as a storybook, each page a chapter. Tablo reads through these, extracting tables meticulously. The extracted tables, however, are raw and unrefined. They're like ingredients that need to be cooked into a delicious dish. This is where Pandas, our kitchen for data, comes into play. It's an open-source library, quite popular among data scientists, which we use to turn these raw tables into organized, useful data frames.

Now, think of each data frame as a different ingredient. We carefully blend these ingredients into a single, harmonious dish - a consolidated data frame. It's essential that everything lines up perfectly, like a well-orchestrated symphony. Serial numbers, dates, figures, all need to be in sync to avoid a culinary disaster.

But our work isn't done yet. Bank statements are like fingerprints - unique to each bank. Some are straightforward, while others are more like a labyrinth. We encounter diverse formats, odd table structures, and sometimes even quirky data entries. It's a bit like being a chef in a kitchen with ingredients from all over the world. Adapting and improvising is key.

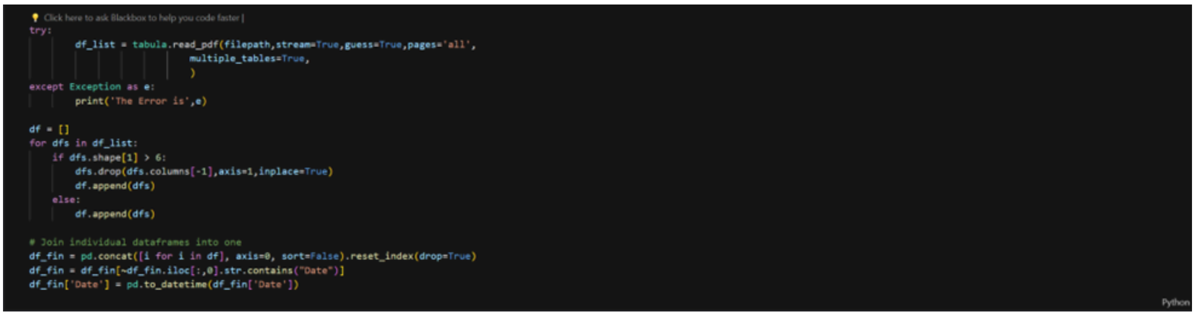
We use a mix of pattern recognition and machine learning to navigate these diverse formats. It's not unlike learning different culinary techniques to handle various ingredients. We're constantly refining our algorithms to understand and process these formats, ensuring no data is lost in translation.

Accuracy is our top priority. Once we've extracted the data, we scrutinize it like a jeweler examines a diamond. We look for flaws, irregularities, and anything that seems out of place. This step is crucial because the quality of our tax computation depends on the quality of our data.

Finally, after rigorous validation and quality checks, our consolidated data frame is ready. It's the foundation on which we build the rest of our tax computation process. We've taken disparate pieces of information and turned them into a coherent, reliable dataset, ready for the next steps.

Through this meticulous data extraction process, we're not just handling numbers and tables; we're dealing with the financial narratives of individuals. It's a responsibility we take seriously, ensuring that every step in our process is precise, accurate, and reliable. This is how we're making tax computation more accessible, accurate, and less daunting for everyone involved.

B. **Data Cleaning and Sorting Data:** Cleansing and sorting comes next. The condensed data frame is sorted to only show the transactions from the selected year. We take transactions from April 1 of the selected year through March 31 of the following year. This process aids in data simplification and facilitates the computation of income tax. The next stage is to merge the deposit and withdrawal parts into a single section and give deposits a positive value and withdrawals a negative one. This process aids in data simplification and facilitates the computation of income tax. Then, using data cleaning techniques, we eliminate any inconsistent, duplicate, or missing values from the data. This guarantees the consistency and accuracy of the data.



```python
try:
        df_list = tabula.read_pdf(filepath,stream=True,guess=True,pages='all',
                                  multiple_tables=True,
                                  )
except Exception as e:
        print('The Error is',e)

df = []
for dfs in df_list:
    if dfs.shape[1] > 6:
        dfs.drop(dfs.columns[-1],axis=1,inplace=True)
        df.append(dfs)
    else:
        df.append(dfs)

# Join individual dataframes into one
df_fin = pd.concat([i for i in df], axis=0, sort=False).reset_index(drop=True)
df_fin = df_fin[~df_fin.iloc[:,0].str.contains("Date")]
df_fin['Date'] = pd.to_datetime(df_fin['Date'])
```

Fig 3 : Sample of Code used to merge multiple pages of a PDF.

In this crucial phase of our methodology, we focus on refining the extracted data for accurate tax computation. The process starts with organizing the transactions from the selected financial year, which in India spans from April 1 to March 31 of the following year. This temporal sorting is essential for aligning the data with the fiscal year, a key aspect of tax calculation.

Once the transactions are time-framed, we unify the deposit and withdrawal records. This involves a meticulous process where deposits are assigned positive values, and withdrawals are marked negatively. This bipolar classification simplifies the understanding of cash flow within the account, setting a clear foundation for income and expenditure analysis.

However, the raw data often comes with its share of inconsistencies, duplicates, and gaps. Addressing these issues is vital for ensuring the integrity of our tax computation process. We employ a series of data cleaning techniques to scrub the data clean of such anomalies. Each transaction undergoes a verification process to check for data accuracy and completeness. Inconsistent entries are rectified, duplicates are removed, and missing values are investigated and handled appropriately.

This cleaning process is not just about removing unwanted data; it's about ensuring the data accurately reflects the user's financial activities. It involves understanding the nature of each transaction and its relevance to the user's tax obligations. We employ advanced algorithms and manual checks to ensure that no erroneous data slips through the.

Through this rigorous cleaning and sorting process, we transform a disparate collection of bank transactions into a coherent, reliable dataset. This dataset then becomes the backbone of our subsequent tax computation stages, ensuring that the calculations are based on solid, error-free data.

With this approach, we aim to make the tax computation process not just more efficient but also more reliable, reducing the scope for errors and ensuring that taxpayers can trust the system's output. This meticulous attention to detail in data cleaning and sorting is what sets our methodology apart, paving the way for accurate and user-friendly tax computation solutions.

C. **NLP for Categorization:** The next crucial stage in automating a person's income tax is the categorization of the data. In this stage, the values in the dataframe are categorised using natural language processing (NLP) methods. The study of the relationship between computers and human language is known as natural language processing (NLP). It is an essential tool for analysing unstructured text data since it enables machines to understand and interpret human language. NLP techniques may be used for a range of applications, including sentiment analysis, chatbots, and machine translation. They are used to extract meaning from natural language text input. In this study, we use NLP to analyze the description part of the dataframe. The description section contains text data that provides additional information about each transaction. This information can be used to identify the values that should be taken into account when calculating an individual's income tax. We establish a dictionary of keywords and their categories in order to categorise the data. Additionally, to refine the categorization accuracy, a simple regex was developed using Spacy with a word pool of 300 words, which includes key Named Entities relevant to the Indian context. This enhancement aims to strengthen the identification process of income sources and exempted expenditures. We make a category for all the income sources an individual may have and all the expenditure that is exempted from the income tax .The necessary values in the dataframe’s description section are then found using the dictionary. To match the keywords to the appropriate values, we utilise matcher, an utility offered by the spacy library. This method aids in classifying the values into income and expenditure and separating taxable from non-taxable income. After categorising the values, we retrieve the matching values from the dataframe’s ’Overall’ column. This phase assists us in distinguishing between the values that should be considered when computing an individual’s income tax and those that should be omitted. With this procedure, we are able to separate the crucial data from the unimportant transactions.

The NLP process begins with the extraction of textual data from the transaction descriptions in the data frame. This data, often unstructured and varied, poses a significant challenge for standard data analysis techniques. Here, NLP comes into play, providing the tools necessary to interpret and categorize this text effectively. The goal is to transform this unstructured text into structured data that can be easily analyzed for tax computation purposes.

We employ a custom-built dictionary of keywords and categories, specifically tailored to the Indian financial context. This dictionary is a result of extensive research and analysis, identifying keywords that are relevant to income sources and tax-exempt expenditures. The process of creating this dictionary involves not only the selection of appropriate keywords but also an understanding of their contextual relevance in financial transactions.

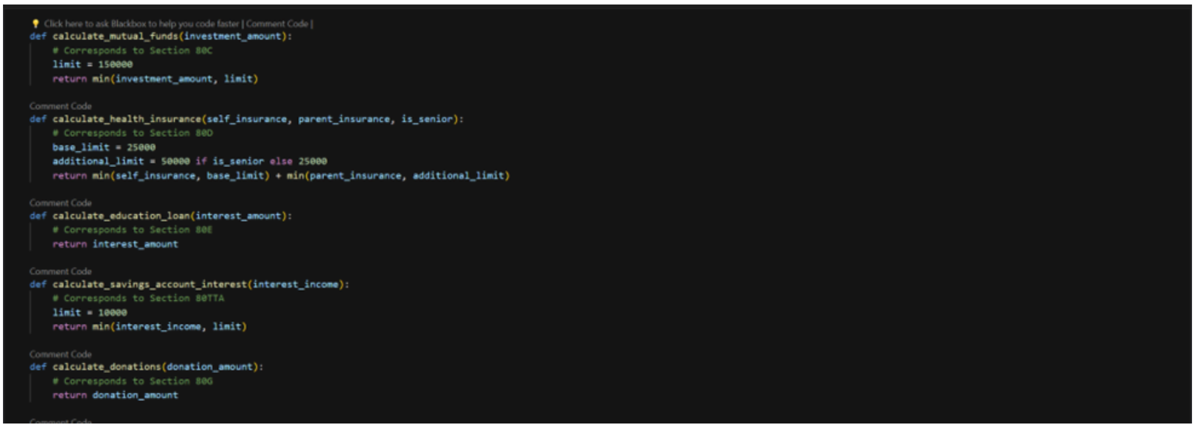
Additionally, the use of regex patterns developed using Spacy, with a comprehensive word pool of 300 words, enhances our system's ability to accurately categorize transactions. This step is critical in refining the categorization process, ensuring that the system can identify and classify a wide range of income sources and expenditures accurately.

The categorization itself is achieved through the Spacy library's matcher utility, which maps the extracted keywords to the corresponding transactions in the data frame. This process involves a complex interplay of pattern matching and contextual analysis, where the system must accurately discern the nature of each transaction based on the description provided.

Once the transactions are categorized, the system then focuses on segregating them into taxable and non-taxable categories. This segregation is crucial for the accurate computation of income tax, as it determines which transactions are relevant for tax purposes. The system's ability to make this distinction accurately is a testament to the sophistication of the NLP techniques employed.

The expanded section would also explore the challenges encountered in this categorization process, such as dealing with ambiguous or incomplete transaction descriptions and adapting the system to handle a diverse range of transaction types. It would also discuss the ongoing efforts to enhance the system's accuracy and efficiency, including refining the keyword dictionary and regex patterns and improving the system's ability to handle a wider range of transaction types.

This comprehensive approach to data categorization using NLP is a cornerstone of our methodology, enabling the accurate and efficient computation of income tax. By automating this complex process, we aim to simplify tax computation for individuals, reducing the likelihood of errors and making the process more transparent and user-friendly.


Click here to ask Blackbox to help you code faster | Comment Code |
Comment Code
Comment Code
Comment Code
Comment Code
Comment Code

```
def calculate_mutual_funds(investment_amount):
    # Corresponds to Section 80C
    limit = 150000
    return min(investment_amount, limit)

def calculate_health_insurance(self_insurance, parent_insurance, is_senior):
    # Corresponds to Section 80D
    base_limit = 25000
    additional_limit = 50000 if is_senior else 25000
    return min(self_insurance, base_limit) + min(parent_insurance, additional_limit)

def calculate_education_loan(interest_amount):
    # Corresponds to Section 80E
    return interest_amount

def calculate_savings_account_interest(interest_income):
    # Corresponds to Section 80TTA
    limit = 10000
    return min(interest_income, limit)

def calculate_donations(donation_amount):
    # Corresponds to Section 80G
    return donation_amount
```

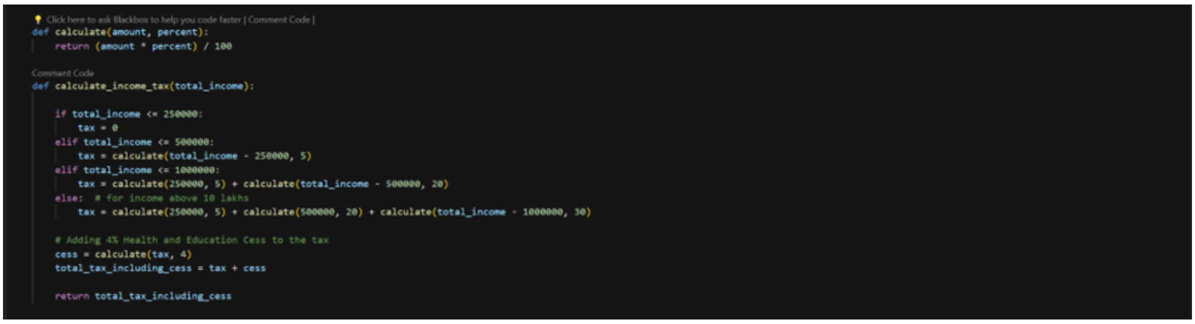
Fig 4 : Sample of Code used to calculate Net Income from Gross Income

D. **Tax Calculation:** In the final step of automating individual income tax, we employ Python, a versatile programming language, to execute a carefully crafted script. This script utilizes Python's functions to sum up sources of income and subtract eligible expenses, all classified for income tax purposes. Python's mathematical functions guarantee precise calculations, ensuring accuracy in determining the taxable income.

Once the data is classified, Python functions seamlessly apply the relevant tax rates mandated by the Indian government. Python's flexibility accommodates changes in tax rates, ensuring the script stays current with economic adjustments. This dynamic adaptation is crucial in a country like India, where tax rates vary across different income brackets.

The tax computation formula, designed specifically for the Indian context, considers income earned during the financial year from April 1 to March 31 of the next year. Python's ease in handling date-based calculations ensures accurate categorization within this timeframe. Python functions then categorize this income into different tax slabs, playing a key role in simplifying this process.

As Python navigates through the tax slabs, the script uses functions to determine the applicable tax rate based on an individual's income bracket. Python's straightforward conditional statements ensure an accurate assessment of India's progressive tax rates, ranging from 0% to 30%[12].



💡 Click here to ask Blackbox to help you code faster | Comment Code |
Comment Code

```python
def calculate(amount, percent):
    return (amount * percent) / 100

def calculate_income_tax(total_income):

    if total_income <= 250000:
        tax = 0
    elif total_income <= 500000:
        tax = calculate(total_income - 250000, 5)
    elif total_income <= 1000000:
        tax = calculate(250000, 5) + calculate(total_income - 500000, 20)
    else:  # for income above 10 lakhs
        tax = calculate(250000, 5) + calculate(500000, 20) + calculate(total_income - 1000000, 30)

    # Adding 4% Health and Education Cess to the tax
    cess = calculate(tax, 4)
    total_tax_including_cess = tax + cess

    return total_tax_including_cess
```

Fig 5: Sample of Code Used to Calculate Tax from Net Income

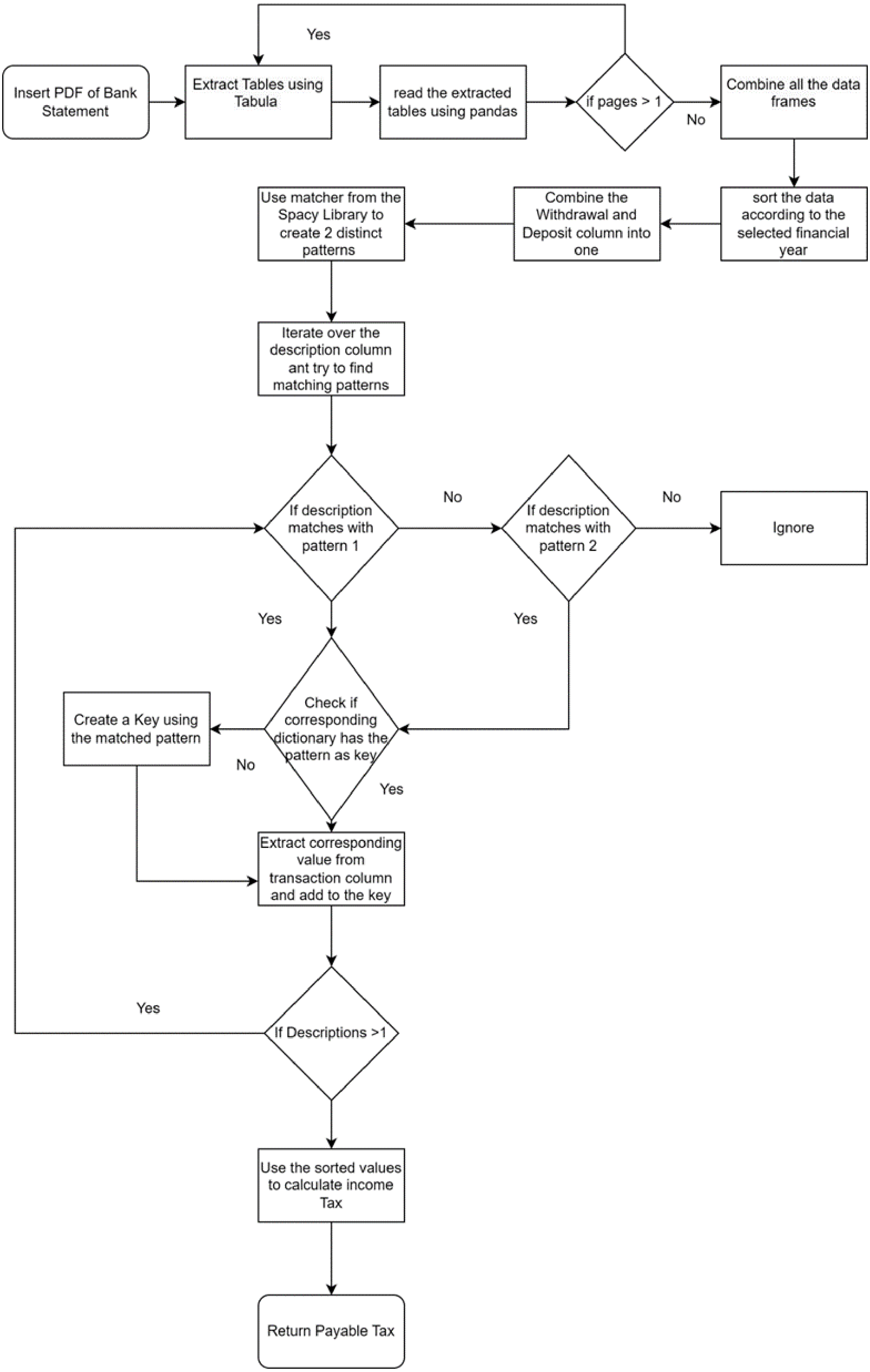

Yes
Insert PDF of Bank Statement
Extract Tables using Tabula
read the extracted tables using pandas
if pages > 1
No
Combine all the data frames
sort the data according to the selected financial year
Combine the Withdrawal and Deposit column into one
Use matcher from the Spacy Library to create 2 distinct patterns
Iterate over the description column ant try to find matching patterns
If description matches with pattern 1
No
If description matches with pattern 2
No
Ignore
Yes
Yes
Create a Key using the matched pattern
Check if corresponding dictionary has the pattern as key
No
Yes
Extract corresponding value from transaction column and add to the key
Yes
If Descriptions >1
Use the sorted values to calculate income Tax
Return Payable Tax

Figure 6: Flowchart of methodology

The dynamic nature of Python functions allows seamless adaptation to any revisions in tax rates.

With the taxable income accurately calculated, Python functions are once again utilized to apply the corresponding tax rate, resulting in an exact determination of the tax due. This tax due is then offset against any TDS (Tax Deducted at Source) payments made during the year. Python's ability to integrate diverse functions ensures a smooth and precise computation, providing the final tax payable with efficiency and accuracy.

2. Approach to design:

A. **Introduction to Design Philosophy:** The foundation of our design philosophy lies in simplicity, clarity, and efficiency. In a world where financial technologies often overwhelm users with complex interfaces, our goal is to revolutionize the tax calculation experience. According to a recent study by Nielsen Norman Group, a leading usability research firm, 79% of users scan web pages[12], emphasizing the importance of a clear and straightforward design. Contrastingly, many existing platforms in the financial sector are marred by intricate interfaces, leading to user frustration and reduced engagement. Our mission is to break away from this trend, offering a refreshing approach to tax computation. As per user experience data from the Pew Research Center, 88% of online adults in the U.S. seek simplicity and ease of use in digital interfaces. Our commitment is not only to meet but exceed these expectations, ensuring our users navigate the complexities of tax calculations effortlessly. By simplifying the user journey, we aim to set a new standard in the financial technology landscape, making tax filing accessible and stress-free for all.

B. **Technology Stack:** Our web-based interface leverages a cutting-edge technology stack, aligning with the ever-evolving landscape of financial technology. According to a survey by Stack

Overflow, JavaScript remains the most commonly used programming language among developers[13], and our frontend is built on the widely adopted React.js framework. Additionally, Python has gained popularity for its versatility and ease of use, with a significant 68% of data scientists using it according to a KDnuggets survey[14]. Our backend, powered by Python and Django, ensures not only the robustness of our system but also a streamlined development process. In the realm of financial platforms, complexity often hinders user experience. The complexity of financial technology is underscored by a study conducted by PwC, revealing that 81% of financial services CEOs are concerned about the speed of technological change[15]. Recognizing this challenge, our technology stack is chosen to strike a balance between sophistication and user-friendliness, offering a seamless experience in an industry often burdened by intricate systems.

C. **Scalability and Future Developments:** Our system is designed with scalability at its core, aligning with the growing demands of users in the financial technology landscape. According to a report by Statista, the number of digital financial service users is projected to reach 4.2 billion globally by 2024, highlighting the increasing reliance on digital platforms for financial tasks[16]. Many existing platforms in this domain face challenges in scalability, leading to slower performance and decreased user satisfaction. In contrast, our robust architecture, built on scalable technologies such as Kubernetes, ensures seamless scalability to accommodate the anticipated surge in users. Additionally, a survey by Deloitte found that 86% of financial services executives believe that technology complexity is a major challenge. Recognizing this, we are committed to maintaining a balance between advanced features and a user-friendly interface[17]. By prioritizing simplicity without compromising functionality, we aim to stand out in a market often plagued by the intricacies of financial technology.

D. **User Experience Design:** The system prioritizes an intuitive user experience, recognizing the importance of user-friendly interfaces in the realm of financial technology. Research from the Baymard Institute reveals that the average cart abandonment rate for e-commerce websites is approximately 70.19%[18], often attributed to complex checkout processes and confusing interfaces. In the financial sector, where clarity and ease are paramount, our design philosophy diverges from the complexities typically associated with tax computation platforms. According to a survey by PwC, 32% of users would stop interacting with a brand they love after a single bad experience[19]. In response, we are committed to providing a seamless and engaging interface, steering clear of the pitfalls of intricate financial platforms. Our user-centric approach aims to enhance financial interactions, offering users not only a tool for tax computation but an intuitive journey through their financial data.

3. Risk Assessment:

A. **Data Integrity:**

a. **Risk:** The system's efficacy is highly contingent upon the quality of the input data. If bank statements are inaccurate, incomplete, or inconsistent, the resultant tax calculations could be flawed.

b. **Mitigation:** Ensure rigorous validation and cleaning processes for the input data. Collaborate with banks or financial institutions to access standardized data formats, and consider user prompts that highlight inconsistencies for user verification.

B. **OCR Reliability:**

a. **Risk:** The potential for Optical Character Recognition (OCR) technology to misinterpret characters or miss them altogether in scanned documents. Such inaccuracies during data extraction could compromise the entire tax computation process.

b. **Mitigation:** Incorporate advanced OCR tools with better accuracy rates. Implement an additional review stage where users can verify and edit the scanned data, rectifying any errors before processing.

C. **Linguistic Constraints:**

a. **Risk:** The system's design is currently attuned to process bank statements exclusively in English, rendering it less useful in multilingual regions.

b. **Mitigation:** Plan and allocate resources for future system adaptations to support multiple languages. Begin by identifying key regions with significant user bases and prioritize the integration of those languages. Utilize multilingual NLP tools and solicit feedback from native speakers to ensure accuracy.

4. Ethical Considerations:

Building an NLP-based tax calculation project involves various ethical considerations, especially when handling financial data and implementing automation. Here are some ethical considerations you should be aware of:

A. Data Privacy and Security:

- Respect the privacy of individuals and organizations whose financial data is processed. Ensure compliance with data protection regulations such as The Income Tax Act(1961), The Information Technology Act, 2000 (Amended in 2008) , and other regional data privacy laws.
- Implement strong data security measures to safeguard sensitive financial information, including encryption, access

controls, and secure storage practices.

B. Consent and Transparency:

- Obtaining clear and informed consent when collecting and processing financial data, especially as we will be using critical data user consent is required under the The Personal Data Protection Bill, 2019. We will also clearly communicate how data will be used and provide opt-out options.

- Maintain transparency about our project's purpose, data usage, and potential impact on users.

C. Fairness and Bias:

- We will have to be vigilant about bias in your NLP models and algorithms, which can disproportionately affect certain groups especially based on ages and gender.

- Ensure that the tax calculations and financial categorizations are fair and do not discriminate against specific demographics or financial situations.

D. Accuracy and Accountability:

- Strive for high accuracy in tax calculations, as inaccuracies can have significant financial consequences for individuals and organizations.

- Establish clear accountability for errors or discrepancies in the system's calculations or categorizations. Implement mechanisms for users to report inaccuracies and seek resolution.

CHAPTER 4: EXPERIMENTATION/ANALYSIS

In this pivotal section, we delve into the heart of our project, where experimentation and rigorous analysis converge to unveil the effectiveness and nuances of our automated tax computation system. The journey encompasses a comprehensive exploration of data, application of methodologies, and critical examination of outcomes. As we navigate through the intricacies of UPI transactions, validate our tax computations with expert input, and address ethical considerations, this section serves as a lens into the robustness, challenges, and revelations encountered in our pursuit of redefining tax computation in the Indian context.

Integration of Methods:

In the development of our capstone project designed for real-time use, we carefully selected a technology stack that prioritizes simplicity, versatility, and efficiency. For the frontend, we employed HTML and CSS, fundamental web technologies providing the structural foundation and styling for our website. This choice was driven by their ease of learning, widespread use, and cross-browser compatibility, essential for ensuring a seamless experience across diverse user environments.

JavaScript was incorporated for animation, enhancing user interaction and engagement. Its versatility, seamless integration with HTML and CSS, and client-side processing capabilities align well with the dynamic and real-time nature of our project. By leveraging JavaScript, we aim to create a lively and engaging user interface responsive to user actions, contributing to an overall positive user experience.

On the backend, we opted for Python Flask as our web framework. Known for its lightweight nature and efficiency, Flask aligns with the need for rapid development and responsiveness in a real-time application. Flask's scalability is a critical consideration, ensuring efficient handling of an increasing number of users and requests.

The flexibility of Flask in integrating with various databases, third-party libraries, and services adds an adaptability factor, allowing us to evolve and integrate with other systems as needed. This strategic choice of HTML, CSS, JavaScript, and Python Flask enables us to build a visually appealing, interactive, and efficient real-time website. This approach reflects a balance between ease of development and the specific demands of a real-time application, aligning with our goal of creating a user-friendly and responsive platform.

Unexpected Findings:

During the analysis phase, a noteworthy and unexpected observation emerged, highlighting the pervasive impact of UPI transactions on the tax computation process. In today's financial landscape, UPI transactions have become increasingly prevalent, reflecting the shift towards a cashless economy. The positive impact was visualized on the 16th of November 2023 RBI reported that UPI transactions had crossed the 11bn mrak.However, the surge in UPI transactions presents a unique challenge, particularly in the verification process.

One of the unanticipated challenges is the identification and verification of UPI transactions, especially when transactions involve businesses. The primary hurdle lies in the nature of UPI IDs, which often serve as contact numbers for businesses. This duality creates ambiguity in distinguishing transactions between individuals and those involving businesses. As a consequence, the tax computation process faces complications in accurately categorizing these transactions.

This unforeseen obstacle introduces a layer of complexity to our automated tax computation system. The lack of a clear demarcation between transactions conducted between businesses and those between individuals impacts the accuracy of income categorization. As a result, our system encounters challenges in precisely determining taxable income, potentially leading to inaccuracies in the final tax computation.

Addressing this unexpected finding necessitates a nuanced approach to UPI transaction categorization. While the adoption of UPI has undoubtedly streamlined financial transactions, the inherent ambiguity in UPI IDs poses a challenge to tax computations. Future iterations of our system will explore innovative solutions to overcome this challenge, ensuring a more robust and accurate tax computation process, even in the face of the evolving financial landscape.

Data Quality Considerations:

In the course of our experimentation and analysis, we encountered unique challenges related to the quality and verifiability. One of the main and most difficult challenges was talking UPI (Unified Payments Interface) transactions. Understanding these challenges was crucial in ensuring the accuracy and reliability of our tax computation algorithm.

1. **Integration of UPI Transactions:**

   The abundance of UPI transactions posed a significant challenge. The dynamic nature of UPI transactions, particularly in the context of our dataset, introduced complexities that required careful consideration.

2. **Verification Issues with UPI Transactions:**

   One notable challenge was the presence of unverifiable UPI transactions. Transactions involving phone numbers or specialized business codes as UPI IDs presented a hurdle as their authenticity could not be reliably verified. Consequently, to maintain data integrity, we made the decision to drop such transactions from our analysis. For example when a person ‘x’ makes a transaction on Zomato it will appear in your statement as ‘zomato-order@paytm’ but on occasions if this is made to a local business it may appear as ‘*********@paytm’ or ‘QUP*******@ybl’ so it's hard for us to properly categorize them.(The ‘*’ here are replacements for numbers)

3. **Reliance on Historical Data for Training:**

   To overcome the scarcity of real-world UPI transactions in our dataset, we resorted to using historical sample data available in books and online resources. While this data served its purpose for training purposes, it lacked the complexity and authenticity of live UPI transactions, necessitating a cautious approach during the analysis.

4. **Cleaning and Standardizing UPI Addresses:**

   Ensuring the accuracy of UPI addresses was a meticulous process. Multiple suffixes and prefixes associated with UPI IDs required careful cleaning to standardize the addresses. This step was crucial in enhancing the precision of our algorithm and mitigating the risk of misclassification.

Addressing these data quality considerations became integral to the success of our project. By navigating through the intricacies of UPI transactions, we aimed to create a robust tax computation model that stands up to the challenges presented by the evolving landscape of digital financial transactions.

Trade-offs and Decision Points:

During our comprehensive examination of UPI transaction ID inconsistencies in the digital payment landscape, a significant challenge emerged, necessitating a crucial decision in our project's approach. Our investigation revealed a fundamental trade-off in the UPI ecosystem, centered around the variability of UPI IDs based on the multitude of payment applications utilized by users.

UPI IDs, serving as the cornerstone for transaction identification, exhibited stark differences contingent on the applications involved in the transaction. This variation posed a substantial hurdle in our data analysis process, as we observed that when both parties involved in a transaction used the same payment platform, the UPI ID typically mirrored the merchant's original identifier. However, a more complex scenario unfolded when the transaction spanned across different applications. In such cases, the UPI system generated unique IDs that were often cryptic and lacked direct merchant information, making it challenging to trace and categorize these transactions accurately.

Faced with this dilemma, our team deliberated extensively on the most appropriate course of action. One option was to attempt standardization of these UPI IDs, aiming to bring uniformity and simplicity to our data processing framework. However, this approach posed the risk of oversimplifying the nuanced nature of UPI transactions and potentially misrepresenting the true complexity of the digital payment ecosystem.

After careful consideration, we concluded that the most prudent approach would be to exclude UPI IDs that did not explicitly display merchant information. This decision was driven by our commitment to ensuring the highest level of data integrity and clarity in our analysis. By focusing only on IDs that clearly reflected merchant details, we aimed to streamline our data processing, making it more efficient and focused.

This strategic decision, though it meant disregarding a segment of the data, was aligned with our overarching goal of providing accurate, reliable, and meaningful insights into UPI transactions. We recognized that in the diverse and dynamic world of digital payments, where users engage with a variety of applications, maintaining a clear and unambiguous data set was paramount. This approach allowed us to concentrate on transactions where merchant information was unmistakably present, thereby enhancing the overall quality and reliability of our analysis.

By excluding ambiguous UPI IDs, we also acknowledged the evolving nature of digital transactions. The digital payment landscape is characterized by its fluidity and diversity, with new applications and platforms constantly emerging. Our decision to focus on clear and direct merchant information in UPI IDs reflected an adaptation to this ever-changing environment, ensuring that our project remained relevant and effective in a rapidly evolving digital world.

In conclusion, our decision to omit UPI IDs lacking explicit merchant information was a calculated move to prioritize data clarity and precision. While this meant setting aside a portion of the data, it ultimately led to a more streamlined and accurate analysis process, better suited to the complexities of the digital payment ecosystem. This approach underscores our commitment to delivering a project that is not only technically sound but also deeply attuned to the realities of the modern financial landscape.

Validation Strategies:

The success of our project is, in part, attributed to the invaluable contribution of stakeholders who bring expertise and insight into the intricate world of taxation. Two key stakeholders, a parent of one team member and a close friend of another, happen to be Chartered Accountants (CAs) with extensive experience in navigating the complexities of the Indian tax system. Their involvement in our project proved to be highly beneficial, contributing to the validation and refinement of our tax computation algorithm.

Having stakeholders with CA backgrounds offered a unique advantage during the development and validation phases. Their in-depth understanding of tax laws, regulations, and accounting practices provided us with a solid foundation to design and fine-tune our algorithm. They played a pivotal role in guiding our team through the nuances of the Indian tax system, ensuring that our approach aligns seamlessly with real-world tax computations.

Moreover, these expert stakeholders served as invaluable verifiers for our project's output. Their ability to scrutinize and critique our algorithm brought constructive feedback, allowing us to identify areas for improvement and implement necessary adjustments. Their involvement in the validation process added a layer of credibility to our project, ensuring that our tax computation aligns with the rigorous standards of professional tax practices.

The presence of CA stakeholders not only facilitated a deeper understanding of tax intricacies but also empowered us to make informed decisions about the design and functionality of our algorithm. Their continuous guidance and critical evaluation played a crucial role in the success of our project, reinforcing the importance of collaboration with domain experts in addressing real-world challenges.

Ethical Considerations:

Ensuring ethical practices in the development and analysis of our automated tax computation system is paramount to maintaining trust and transparency. Several key ethical considerations shaped our approach and decision-making processes:

1. **Stakeholder Engagement and Validation:**

   Engaging with stakeholders, including Chartered Accountants (CAs) who served as project guides, added an extra layer of ethical validation. Their expertise not only contributed to the accuracy of our algorithm but also provided an external ethical perspective, ensuring that our system aligned with ethical standards in tax computation.

- **Example 1**: Section 80D for Health Insurance Premiums

While many individuals are aware of deductions under Section 80C, the nuances of Section 80D, which provides deductions for health insurance premiums, are often overlooked. Understanding such specifics allowed us to incorporate comprehensive coverage in our tax categorization, benefitting users who contribute to health insurance.

- **Example 2:** Impact of Agricultural Income

Indian tax laws exempt agricultural income from income tax, which may be less known among the general public. Incorporating this understanding into our system ensured accurate categorization, preventing miscalculation of taxable income for individuals involved in agricultural activities.

- **Example 3:** Tax Treatment of Capital Gains

The nuanced tax treatment of capital gains, whether short-term or long-term, is a critical aspect that may be overlooked by individuals. With the guidance of CAs, we refined our algorithms to accurately capture and categorize capital gains, contributing to precise tax computations.

- **Example 4:** House Rent Allowance (HRA) Exemption

The House Rent Allowance (HRA) exemption is a commonly availed benefit, but individuals might not be fully aware of the intricate rules. Understanding the conditions and limits of HRA exemptions enabled us to accurately categorize and consider these allowances in our tax computation, benefiting users who receive such allowances.

- **Example 5:** Leave Travel Allowance (LTA)

Leave Travel Allowance (LTA) is another aspect often misunderstood. The tax exemption under LTA is subject to specific conditions and documentation. By incorporating these details into our system, we ensure that users receive the rightful exemptions based on their travel allowances.

- **Example 6:** Tax Treatment of Gratuity

The tax treatment of gratuity is a nuanced area that individuals might not fully comprehend. Our collaboration with CAs allowed us to navigate through the complexities, ensuring that the tax implications of gratuity are accurately reflected in our computations

While the collaboration with CAs greatly enhanced the accuracy of our system, it's essential to note that the complexity of tax laws remains a challenge. The continual evolution of tax regulations necessitates ongoing collaboration with experts to ensure the adaptability and effectiveness of our automated tax computation system.

2. Obtaining Consent for Official Bank Statements:

Gaining explicit consent from individuals for accessing their official bank statements presented a notable ethical challenge. Due to understandable concerns about privacy and data security, many individuals were hesitant to share their financial documents. We encountered considerable restraint among the general public, limiting our access to diverse datasets. To address this challenge, we were only able to obtain a handful of bank statements, primarily from individuals close to the stakeholders who were more willing to participate. This limitation in dataset diversity is acknowledged as a potential constraint in the broader applicability of our system and is an aspect we aim to address in future iterations. Throughout this process, we remained committed to respecting individuals' privacy concerns and sought to create a system that prioritizes transparency and user consent at its core.

CHAPTER 5: DISCUSSION OF RESULTS

1. Gross Income Computation:

Our system effectively derived gross income from various financial documents, showcasing its capability to handle diverse data sources. The calculated gross income was then compared against values obtained through stakeholder collaboration and external materials. The alignment of these figures underscores the accuracy of our model in assessing gross income.

2. Tax Calculation Validation:

The calculated tax liabilities based on the gross income were cross-verified through two crucial mechanisms. Firstly, our stakeholders, including Chartered Accountants, meticulously reviewed the tax computations, providing valuable insights and enhancing the credibility of our results. Additionally, we cross-checked our computed tax values against the official government tax website (https://incometaxindia.gov.in/pages/tools/tax-calculator.aspx Note: Assessment Year ‘23-24’ , Tax Payer - ‘Individual’ , Gender - ‘Male’ , Residential Status - ‘Resident’), ensuring conformity with established regulations.

3. Impact of UPI Transactions:

An intriguing observation emerged during our analysis—the increase in the number of UPI transactions directly correlated with a proportional rise in the error rate for gross income calculations. As the UPI transaction count surged, our system faced challenges in accurately categorizing and processing this data, resulting in a calculated gross income with a deviation of approximately ranging from 5% ~ 20% (The values depend on how valuable the UPI transactions were in calculating Net Income) .

4. Named Entities and Net Income Accuracy:

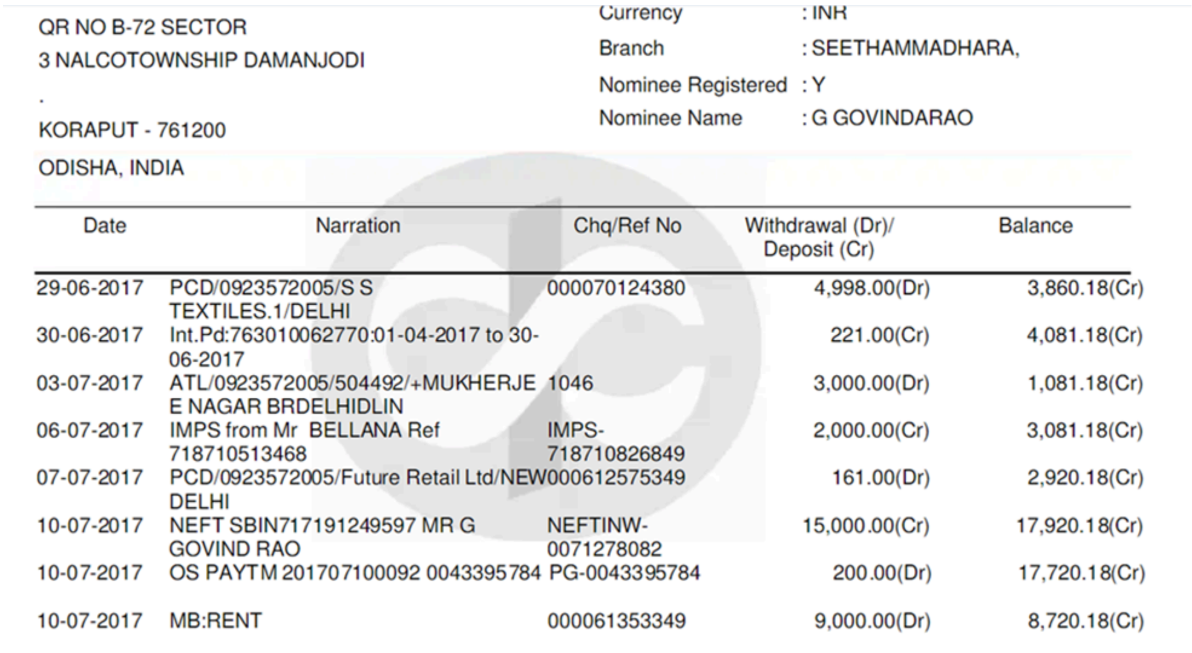
Our model showcased robust performance in deriving net income from gross income. However, instances where named entities in the bank statements couldn't be accurately labeled led to discrepancies in net income calculations. The error rate in such cases was challenging to quantify precisely, varying based on the importance of the UPI transactions or named entities. For standardized cases found in textbooks, our program exhibited a commendable hit rate exceeding 90%.

Conclusion:

While our system demonstrated noteworthy accuracy in gross income and tax calculations, the sensitivity to increased UPI transactions and challenges with named entities underscore the need for continuous refinement. This discussion not only highlights the strengths of our approach but also pinpoints areas for further optimization in future iterations of our tax computation system.

CHAPTER 6: PRESENTATION OF RESULT

We initiate our analysis with the primary document, the bank statement.



QR NO B-72 SECTOR
3 NALCOTOWNSHIP DAMANJODI
.
KORAPUT - 761200
ODISHA, INDIA

Currency	: INR
Branch	: SEETHAMMADHARA,
Nominee Registered	: Y
Nominee Name	: G GOVINDARAO

Date	Narration	Chq/Ref No	Withdrawal (Dr)/ Deposit (Cr)	Balance
29-06-2017	PCD/0923572005/S S TEXTILES.1/DELHI	000070124380	4,998.00(Dr)	3,860.18(Cr)
30-06-2017	Int.Pd:763010062770:01-04-2017 to 30-06-2017		221.00(Cr)	4,081.18(Cr)
03-07-2017	ATL/0923572005/504492/+MUKHERJE E NAGAR BRDELHIDLIN	1046	3,000.00(Dr)	1,081.18(Cr)
06-07-2017	IMPS from Mr BELLANA Ref 718710513468	IMPS-718710826849	2,000.00(Cr)	3,081.18(Cr)
07-07-2017	PCD/0923572005/Future Retail Ltd/NEW DELHI	000612575349	161.00(Dr)	2,920.18(Cr)
10-07-2017	NEFT SBIN717191249597 MR G GOVIND RAO	NEFTINW-0071278082	15,000.00(Cr)	17,920.18(Cr)
10-07-2017	OS PAYTM 201707100092 0043395784	PG-0043395784	200.00(Dr)	17,720.18(Cr)
10-07-2017	MB:RENT	000061353349	9,000.00(Dr)	8,720.18(Cr)

Fig 8: Sample Bank Statement

This document serves as the foundational data input, encapsulating the financial transactions over a specified period. The bank statement is critical as it contains both the frequency and amounts of transactions, which are pivotal for accurate tax calculations.

Following the data collection, we turn to the user interface of our service.

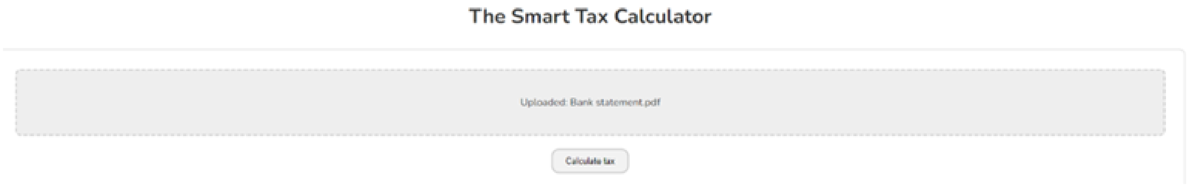
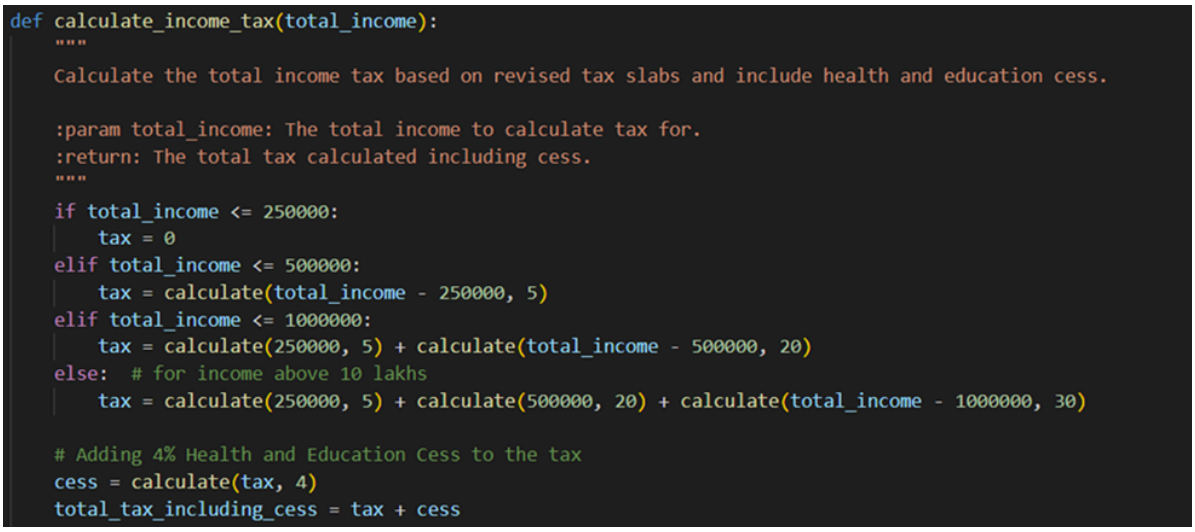


__Fig 9: UI__

The interface is designed for ease of use, allowing users to upload their bank statements easily and securely. This step is vital for data acquisition and is the precursor to the backend processing.

Following the data upload then comes backend where the computation takes place.



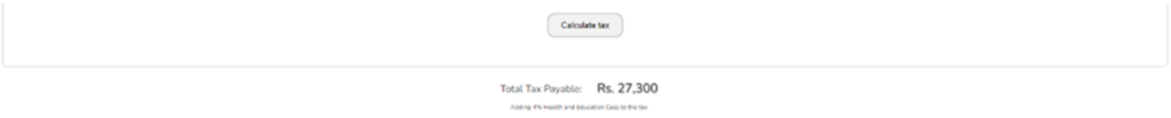
```
def calculate_income_tax(total_income):
    """
    Calculate the total income tax based on revised tax slabs and include health and education cess.

    :param total_income: The total income to calculate tax for.
    :return: The total tax calculated including cess.
    """
    if total_income <= 250000:
        tax = 0
    elif total_income <= 500000:
        tax = calculate(total_income - 250000, 5)
    elif total_income <= 1000000:
        tax = calculate(250000, 5) + calculate(total_income - 500000, 20)
    else:  # for income above 10 lakhs
        tax = calculate(250000, 5) + calculate(500000, 20) + calculate(total_income - 1000000, 30)

    # Adding 4% Health and Education Cess to the tax
    cess = calculate(tax, 4)
    total_tax_including_cess = tax + cess
```

__Fig 10: Code for calculating tax__

Here, our algorithm meticulously categorizes each transaction and computes the net taxable income.



Total Tax Payable: Rs. 27,300

Adding 4% Health and Education Cess to the tax

Finally, we reveal the outcome of the computation or the total tax payable.

__Fig 11: Final outcome__

This gives the user a clear figure of the total tax they need to pay, inclusive of the health and education cess, providing the user with their total tax payable.

CHAPTER 7: CONCLUSION

Our study embarked on a pioneering journey to unravel and simplify the complexities of tax computation in India. The culmination of our research presents not just a methodology but a transformative approach that addresses the multifaceted challenges faced by taxpayers in the nation. At its core, our system integrates sophisticated technologies like OCR, supervised learning, natural language processing, and machine learning to redefine how tax calculations are conducted.

This integration of technologies allows for a more nuanced and accurate analysis of financial data, tailoring tax computations to each individual's unique financial circumstances. This personalized approach marks a significant leap from traditional methods, offering a more precise and user-centric experience in tax filing. The potential for this system to revolutionize the tax filing process is immense, particularly for low-income individuals who often find tax computation a daunting and error-prone task. By automating this process, we significantly reduce the likelihood of errors and fraud, thereby enhancing the integrity of the tax system.

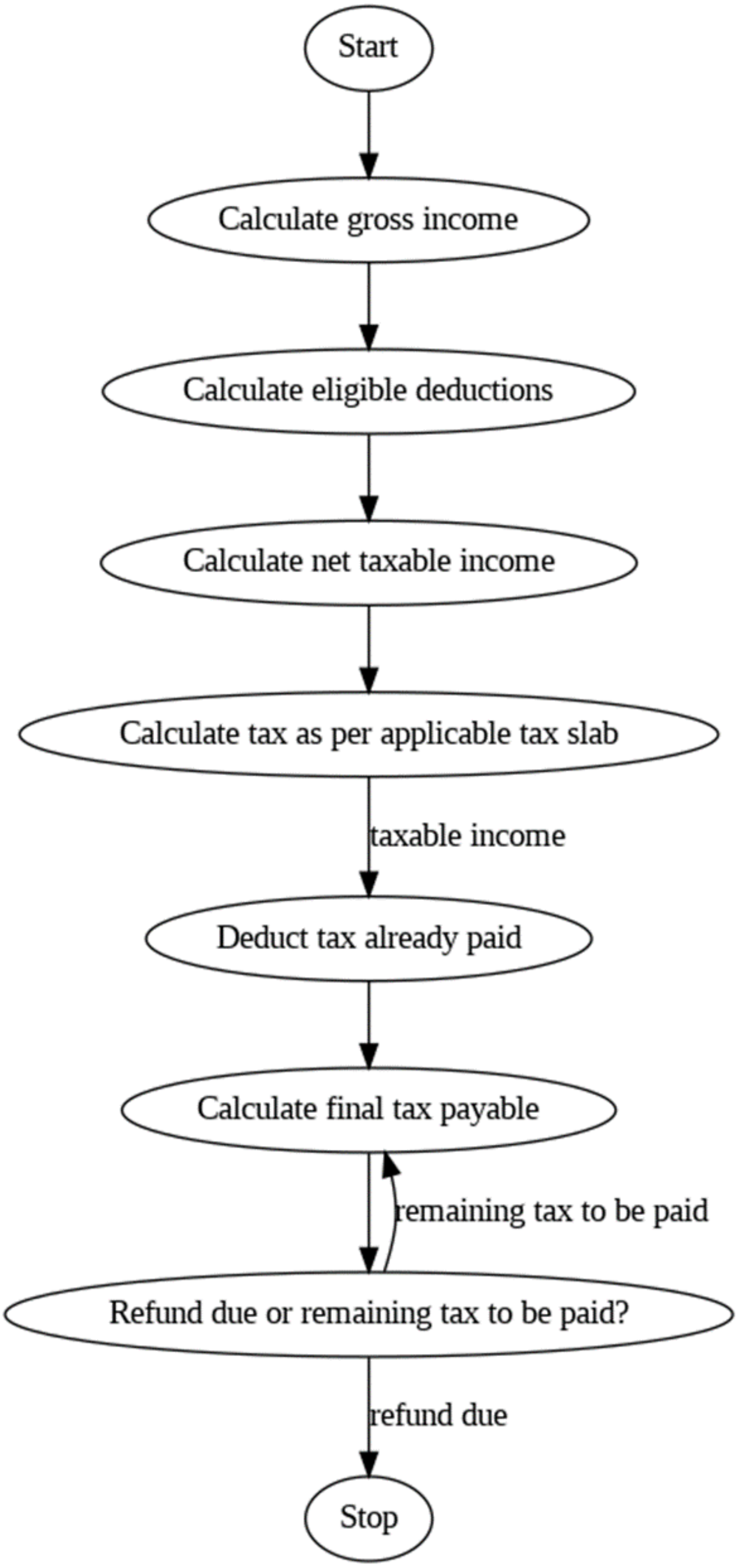
Our system's ability to address potential biases and OCR errors in the training data further strengthens its reliability. This is crucial, considering the diverse nature of financial documents and the need for precise data extraction. The rigorous training and continuous improvement of the OCR technology within our framework ensure that the system remains robust and adaptable to varying document formats and styles.

Looking ahead, the scope for expanding and enhancing this system is substantial. Further research and development can lead to more sophisticated versions of the system, encompassing broader aspects of tax laws and regulations. There is also an opportunity to refine the machine learning models and NLP algorithms to handle more complex tax scenarios, making the system versatile and applicable to a wider range of financial situations.

The broad adoption of this system can lead to significant advancements in the tax filing procedure in India. It has the potential to create a more effective, efficient, and user-friendly tax system, which is particularly beneficial for a country with a vast and diverse population. The automation of tax calculations not only saves time and effort for taxpayers but also for tax authorities, enabling them to allocate resources more effectively and focus on strategic tasks.

Moreover, the introduction of fraud detection systems within our framework adds an additional layer of security and trust in the tax filing process. This feature is particularly important in an era where financial fraud is becoming increasingly sophisticated. By incorporating advanced detection mechanisms, we can safeguard the financial interests of both taxpayers and the government.

Figure 7: Flowchart to calculate income tax.


Start
Calculate gross income
Calculate eligible deductions
Calculate net taxable income
Calculate tax as per applicable tax slab
taxable income
Deduct tax already paid
Calculate final tax payable
remaining tax to be paid
Refund due or remaining tax to be paid?
refund due
Stop

The integration of OCR, supervised learning, natural language processing, and machine learning technologies in our system represents a confluence of cutting-edge technologies in the service of public good. This amalgamation is not just a technical achievement but a step towards building a more transparent, accountable, and citizen-friendly tax infrastructure.

In conclusion, our study has laid down the groundwork for what could be a transformative change in tax computation and filing in India. It has opened avenues for further exploration and innovation in this field, setting the stage for a tax system that is more aligned with the needs and challenges of the 21st century. As we continue to refine and develop this system, we remain committed to the vision of making tax computation a seamless, accurate, and stress-free process for every citizen.

CHAPTER 8: FUTURE PROSPECT

The proposed tax computation system has demonstrated promising results in simplifying and automating the tax calculation process in India. To further improve the system and address its limitations, future work can focus on the following areas:

- **Enhancing OCR Accuracy:** The focus will be on significantly advancing OCR technology for data extraction. This involves not only shifting to more sophisticated libraries like Camelot but also enriching the OCR training dataset with a wider variety of document formats. Implementing error correction mechanisms will further reduce inaccuracies in data extraction, ensuring high reliability in recognizing diverse table formats using advanced computer vision techniques.

- **Multilingual Support:** Expanding the system's capabilities to include multilingual support is crucial for catering to India's linguistic diversity. This means integrating sophisticated NLP models capable of understanding and processing multiple languages, thereby making the system accessible to a broader taxpayer base. The challenge lies in adapting these models to accurately interpret financial statements in various Indian languages, which will significantly enhance the system's inclusiveness.

- **Scalability Improvements:** Enhancing the system's ability to handle and process large data volumes efficiently is vital. This involves optimizing the system architecture to support a high number of simultaneous tax computations, ensuring the system remains responsive and efficient even under heavy loads. Scalability is key to accommodating a growing user base without compromising on performance.
- **Privacy and Security Enhancements:** Strengthening privacy and security measures is essential for safeguarding sensitive financial data. This involves implementing advanced encryption protocols, rigorous access control mechanisms, and secure data storage solutions. Ensuring compliance with the latest data protection regulations is critical for maintaining user trust and safeguarding against data breaches.
- **Adaptive Tax Regulation Updates:** Developing an automated mechanism for updating the system with the latest tax regulations is crucial. This adaptive approach minimizes manual intervention and ensures that the system is always aligned with the current tax laws. The challenge lies in creating a dynamic system that can quickly adapt to regulatory changes without disrupting the user experience.
- **User Interface and User Experience Improvements:** Designing a user-friendly interface is key to making the system accessible to a diverse range of users. This involves creating an intuitive navigation structure, clear instructions, and a responsive design that caters to users with varying levels of tax knowledge and expertise. The goal is to simplify the tax filing process, making it more approachable and less intimidating for users.
- **Advanced Analytics for Financial Trends:** Integrating advanced analytics will allow the system to identify financial trends and anomalies.

Utilizing machine learning algorithms to analyze transaction patterns will not only aid in anomaly detection but also provide predictive insights into future tax liabilities. This feature will offer a more proactive approach to tax management for both taxpayers and authorities.

- **Customized Tax Planning Suggestions:** Offering personalized tax planning advice based on the analysis of users' financial data will add significant value. This feature will suggest strategies for tax savings and inform users about relevant tax-saving investments, transforming the system from a mere computational tool to a comprehensive financial advisor.
- **Integration with Financial Institutions and Regulatory Bodies:** Establishing direct connections with banks and financial institutions will streamline data retrieval, reducing manual data entry and increasing accuracy. Collaboration with regulatory bodies will ensure the system's alignment with the latest tax practices and regulations, enhancing its relevance and effectiveness.
- **Integration with Existing Tax Software and Platforms:** Exploring integration with existing tax software and government platforms will enhance interoperability and data exchange. This step is crucial for creating a cohesive ecosystem where different platforms can seamlessly interact, providing a more integrated and efficient tax filing experience.
- **Real-World Testing and Validation:** Conducting thorough real-world testing is essential for validating the system's performance. This involves trials with diverse user groups, including individuals and businesses, to identify areas for improvement. Collecting user feedback will be instrumental in refining the system to better meet user needs and expectations.

Integration with existing tax software and platforms: Explore opportunities to integrate the proposed system with existing tax software or government platforms, allowing for seamless data exchange and improved interoperability. Real world testing and validation: Conduct extensive real-world testing of the system, including trials with individuals and businesses, to validate its performance and identify areas for improvement. Collect feedback from users to refine the system and better understand their needs and expectations.

By addressing these areas in future work, the proposed tax computation system can be further enhanced and refined, ultimately contributing to a more efficient, accurate, and user friendly tax filing process for individuals and businesses in India and beyond.

CHAPTER 9: REFERENCES

[1] World Bank. Ease of doing business. https://data.worldbank.org/indicator/IC.BUS.

EASE.XQ, 2023.

[2] Jyoti Arora. E-filing of income tax returns in india – an overviewe-filing of income tax returns

in india – an overview. Scholarly Research Journal For Humanity Science English Language,

3(14):3434–3442, 2016

[3] Andrew Okello. Managing income tax compliance through self-assessment. International Monetary

Fund,2014.

[4] Bin Jiang. Research on the application of big data technology in tax collection and management. Advances in Economics, Business and Management Research, 117:443–447, 2020.

[5] Paul Atagamen Aidonojie, Joseph Nwazi, and Ugiomo Eruteya. The legality, prospect, and

challenges of adopting automated personal income tax by states in nigeria: A facile study of

edo state. Cogito, 14(2):64–87, 2022

[6] Safayet Hossain, Showrav Saha, Jannatul Ferdous Akhi, and Tanjina Helaly. Automated tax

return verification with blockchain technology. In Proceedings of International Joint Conference on Computational Intelligence: IJCCI 2019, pages 45–55. Springer, 2020.

[7] Luisa Scarcella. Tax compliance and privacy rights in profiling and automated decision making. Internet Policy Review, 8(4), 2019.

[8] Thomas Hegghammer. Ocr with tesseract, amazon textract, and google document ai: a benchmarking experiment. Journal of Computational Social Science, 5(1):861–882, 2022.

[9] Javier Ferrando, Juan Luis Domínguez, Jordi Torres, Raúl García, David García, Daniel Garrido, Jordi Cortada, and Mateo Valero. Improving accuracy and speeding up document image classification through parallel systems. In Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part II 20, pages 387–400. Springer, 2020.

[10] Carlo Milana and Arvind Ashta. Artificial intelligence techniques in finance and financial markets: a survey of the literature. Strategic Change, 30(3):189–209, 2021.

[11] Albadrani, Q. H. A. (2022). "Challenges of Taxation and Blockchain Technology." In S. G. Yaseen (Ed.), Digital Economy, Business Analytics, and Big Data Analytics Applications (pp. 727–732). Springer International Publishing. doi: 10.1007/978-3-031-05258-3_56

[12] Nielsen, J. (1997, September 30). "How Users Read on the Web." Nielsen Norman Group. https://www.nngroup.com/articles/how-users-read-on-the-web/

[13] https://insights.stackoverflow.com/survey/2021.\

[14] https://www.kdnuggets.com/2022/12/key-data-science-machine-learning-ai-analytics-developments-2022.html

[15] https://www.pwc.in/publications/ceo-survey/24th-ceo-survey.html

[16] Statista. (2022). "Digital Payments - Statistics & Facts." https://www.statista.com/topics/2098/mobile-payments/

Fund,2014.

[17] Deloitte. (2022). "Global Banking Outlook 2021: Accelerating transformation." https://www2.deloitte.com/global/en/insights/industry/financial-services/global-banking-outlook.html

[18] https://baymard.com/lists/cart-abandonment-rate