A COMPREHENSIVE APPROACH TO PREVENTING DATA LEAKAGE AND STRENGTHENING CYBERSECURITY

TMP-2023-24-082

PROPOSAL PROJECT REPORT

Fayas ACM – IT20637828

B.Sc. (Hons) in Information Technology Specializing in Cyber Security

Department of Computer Systems and Engineering

Sri Lanka Institute of Information Technology

August 2023

A COMPREHENSIVE APPROACH TO PREVENTING DATA LEAKAGE AND STRENGTHENING CYBERSECURITY

TMP-2023-24-082

PROPOSAL PROJECT REPORT

(Utilizing NLP Techniques for Enhanced Data Protection)

Fayas ACM – IT20637828

Supervisor – Mr. Amila Senarathne

Co-Supervisor –

B.Sc. (Hons) in Information Technology Specializing in Cyber Security

Department of Computer Systems and Engineering

Sri Lanka Institute of Information Technology

August 2023

DECLARATION

We declare that this is our own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature	Date
Fayas ACM	IT20637828		2023.08.25

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

Signature of the Supervisor Date

…………………………………… ……………………………………

Signature of the supervisor

Signature of the Co-Supervisor Date

…………………………………… ……………………………………

ABSTRACT

Nowadays, the escalating threat of data breaches has become a paramount concern for businesses. The safeguarding of sensitive data is a top priority, necessitating the attention of top management, IT administrators, and experts alike. Traditional security measures like firewalls are proving inadequate in the face of evolving cyber threats. Data Loss Prevention (DLP) systems are a new desire in the struggle for data security. This research initiative comprises 4 interconnected subcomponents aimed at providing a comprehensive strategy to minimize data loss and enhance cybersecurity.

The first subcomponent, "Safeguarding Systems by Identifying Unusual User Patterns " uses a machine learning algorithm to proactively identify unusual user behavior and access patterns, enabling timely responses to potential threats. The second, "Utilizing NLP Techniques for Enhanced Data Protection" explores natural language processing algorithms to automatically identify sensitive information in text, boosting data safety by detecting insider risks. The third, "Unveiling Patterns and Anomalies to Mitigate Data Breach Risks" employs innovative data analysis tools and machine learning to uncover hidden patterns and minimize breach risks. Lastly, "Malicious Image Detection and Classification Using Deep Learning Techniques" focuses on defending against image-based cyberattacks by utilizing convolutional neural networks to distinguish between legitimate and malicious images.

With a foundation in expertise and practicality, this research not only contributes to the theoretical advancements of cybersecurity but also equips businesses with effective tools to navigate the complex digital landscape and safeguard their critical data.

My research component delves into the domain of "utilizing NLP techniques for more desirable statistics protection." In the present-day digital milieu, where data breaches pose significant threats, novel techniques are imperative to improve the security of sensitive information. This aspect of the research focuses on harnessing the potential of natural language processing (NLP) algorithms to strengthen data security measures.

NLP involves training computer systems to understand and interpret human language. In a world where safeguarding confidential data is paramount, NLP emerges as a promising tool due to its ability to enable machines to comprehend and analyze written content. The objective of this research segment is to apply NLP techniques to textual data, specifically to identify and categorize sensitive information. By employing techniques like entity recognition and sentiment analysis, the goal is to enhance automated systems' capability to detect critical facts within text-based documents. This proactive identification process not only enhances data protection but also addresses potential insider threats, contributing to the broader spectrum of risk management. The integration of NLP techniques represents a commitment to innovative approaches in data security, yielding insights that contribute to the evolving landscape of cybersecurity in our increasingly digitized world.

TABLE OF CONTENTS

1.INTRODUCTION

In the modern days, protecting sensitive data has come to be greater vital than ever before. unfortunately, record breaches have become more and more common, with cybercriminals constantly finding new ways to infiltrate networks and steal valuable information. That's why it is essential to have a complete method to prevent statistics leakage and strengthen cybersecurity. By enforcing the proper security features and staying vigilant for potential threats, we will assist in defending ourselves and our companies from the devastating consequences of a data breach.

One commonplace example of information leakage is when an employee by accident sends private facts to the wrong man or woman. data leakage could have serious consequences for individuals, and businesses. economic loss is one of the maximum immediate and tangible consequences of statistics leakage. In some cases, the value of remedying the breach may be big, such as criminal costs, compensation to affected events, and damage to IT infrastructure. additionally, records leakage can damage an employer's reputation, main to lost business and diminished belief from clients and companions.

To effectively prevent data leakage and make stronger cybersecurity, a comprehensive approach is important. this indicates going beyond truly imposing safety features and alternatively taking a holistic approach that considers all aspects of an employer's operations. by way of doing so, capability vulnerabilities can be diagnosed and addressed before they may be exploited by cybercriminals. A comprehensive method additionally includes ongoing monitoring and assessment up-to-date make sure that security features stay effective in the face of evolving threats.

1.1 Research Background

Data protection is more essential than ever within the current global. Due to improvements in statistics series and sharing, the need for guaranteeing facts protection has elevated. Natural language processing (NLP) equipment has made it viable to improve records safety solutions and evaluate sensitive information. The idea for enhancing data security measures using NLP approaches is presented in this research. The aim of this study is to present a thorough understanding of the way NLP may be used to detect, categorize, and protect sensitive facts in textual data. The motive of this research is to combine advanced language processing skills with efficient information safety mechanisms to solve the problems associated with facts safety in a unique and substantial technique.

Natural language processing (NLP) techniques have caused significant upgrades in several categories, inclusive of sentiment evaluation, categorization of textual content, and named entity recognition. Those NLP methods allow us to routinely extract important details from textual content records, which may additionally enhance the effectiveness of identifying sensitive information. We can build a strong basis for growth facts security across a couple of sectors by combining those capabilities with information protection techniques. This effort aims to have an effective impact on the developing information protection environment by means of thorough research, thoughtful implementation, and rigorous evaluation. Similar to increasing the textual content of current expertise, we assume that by showing the efficacy of NLP-driven methodologies in information protection, we can also provide practical insights that can be used in a number of contexts.

This proposal elaborates on the project's objectives, methodology, expected outcomes, assessment plans, and ethical concerns. By initiating this journey to use the potential of NLP for increased data security, we want to bridge the gap between cutting-edge technology and the urgent need to safeguard sensitive information. This proposal elaborates on the task's goals, technique, predicted effects, evaluation plans, and ethical issues. By initiating this adventure to use the potential of NLP for increased information protection, we want to bridge the gap between cutting-edge technology and the pressing need to safeguard sensitive information.

1.2 Literature Survey

The aims, techniques, anticipated effects, evaluation strategies, and ethical issues of the study are all defined in this proposal. We desire to close the distance between advanced technology and the urgent need to protect sensitive information by way of beginning this journey to leverage the potential of NLP for improved information protection. A way well worth mentioning is the use of Named Entity Recognition (NER) algorithms. studies by means of Smith et al. (2020) confirmed the cost of NER in automatically recognizing components like names, locations, and dates. This method enables for the identification of capacity identities and sensitive statistics in unstructured textual input.

Sentiment analysis is a good method for evaluating the emotional context of textual data. Sentiment evaluation may use emotional cues to locate potentially sensitive material, in step with a study by means of Johnson and Lee (2021). The approach enhances data protection while also providing conclusions approximately the emotions associated with specific types of information.

This research highlights interesting approaches to mixing NLP techniques with data safety features, and there are still problems to be overcome. The practical use of NLP models in different scenarios, the trade-off between accuracy and calculation speed, and ethical questions near data privacy are all now being investigated. On top of those basics, this project seeks to further knowledge by using NLP techniques and comparing them inside a robust data security framework. By way of reading and combining the information found inside the literature, this research aims to provide helpful insights on how NLP and data protection might be combined for increased security.

Why are we utilizing NLP techniques?

1. **Identifying Sensitive Information in Documents**
   Sensitive information is any data that, if disclosed, may want to cause damage to an individual or organization. this may encompass personal information including social security numbers, credit card numbers, and clinical data, in addition to confidential business facts including financial statements and alternate secrets.

Figuring out sensitive information in files is crucial for protecting each individual and company from the consequences of records breaches. NLP techniques provide a powerful manner to automatically extract these statistics from files, taking into consideration more efficient and accurate identification of sensitive facts.

2. **Importance of Protecting Data from Unauthorized Access.**

data is a valuable asset that needs to be protected from unauthorized access. With the growing amount of private and sensitive information being stored online, it has emerged as more crucial than ever to make certain that this information is saved secure.

Unauthorized access to information could have extreme consequences, ranging from identification theft to financial loss. it's miles critical for businesses and individuals alike to take steps to protect their facts and prevent it from falling into the wrong hands.

3. **NLP techniques offer a powerful way to extract sensitive information from documents.**

Natural language Processing (NLP) techniques offer a powerful manner to extract sensitive information from files. With the help of machine learning algorithms, NLP can identify and categorize information which includes names, addresses, credit card numbers, and other personal records.

one of the benefits of the usage of NLP for sensitive data extraction is that it can analyze large volumes of text data speedy and accurately. This makes it an ideal tool for organizations that want to process massive quantities of facts on a everyday foundation, along with financial institutions, healthcare vendors, and government organizations.

4. **Analyzing text data uncovers insights and identifies personal information.**

Analyzing text data is a powerful tool which can discover valuable insights and identify sensitive personal data. With the assist of natural Language Processing (NLP) techniques, it's miles possible to extract meaningful information from large volumes of unstructured textual content information, which includes emails, social media posts, and patron feedback forms. by using analyzing this facts, organizations can gain a higher expertise of their customers, competitors, and market trends.

however, it is crucial to be aware that analyzing text information additionally comes with potential dangers. non-public information, which includes names, addresses, and credit score card numbers, may be inadvertently discovered via text analysis. This makes it critical for corporations to take measures to protect sensitive data from unauthorized access. through implementing appropriate security protocols and data anonymization techniques, organizations can ensure that they may be safeguarding their customers' privacy while still gaining valuable insights from text data.

1.3 Research Gap

Despite the fact that the existing literature on the combination of natural Language Processing (NLP) technologies and data security delivers useful information, there are a number of research shortages and possibilities for investigation.

First of all, while studies are looking at the use of NLP techniques for identifying sensitive information, which includes sentiment analysis and named entity recognition (NER), there is a need for more extensive methods that take context into account. sensitive information could be incorrectly identified as a consequence of the minute linguistic and contextual variations that existing methods can find challenging to identify.

Second, there are difficulties that call for extra research in the actual integration of NLP-driven data protection mechanisms into practical applications. effective data protection calls for the installation of security measures that safeguard data privacy throughout the analysis process in addition to the precise identification of sensitive data

A research hole in the area of NLP-driven data protection is ethical issues. The advantages of automated analysis must be carefully weighed against the requirement to respect user privacy and permission.

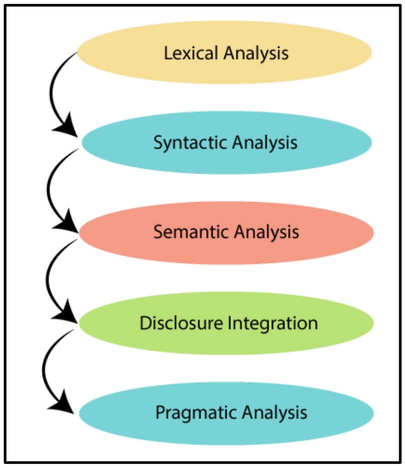
By offering a thorough framework that makes use of NLP approaches for improved data security, this project seeks to fill these research gaps. This project seeks to provide answers to the problems noted in the literature by improving current methodology, creating new techniques, and assessing the usefulness of combining NLP with information protection.

1. **Existing Systems using technology without NLP Techniques**
   This refers to the current state of data protection systems that do not include NLP techniques. These systems may rely on traditional security features but miss out on the benefits that NLP can provide in terms of expertise context, sentiment, and more textual data.

   - **Disadvantages of not using NLP in existing systems:**
     - **lack of Contextual Information:** Without NLP, the system may struggle to understand the context in which certain words or phrases are used, leading to potential false positives or negatives in data protection.
     - **limited Language guide:** non-NLP systems could have difficulties dealing with multilingual content effectively, leaving vulnerabilities in non-native language texts.

   - **Advantages of using the novel system:**
     - **Contextual awareness:** NLP enables understanding the context of language, reducing false alarms, and improving accurate threat detection.

- **Multilingual support:** NLP models can handle various languages, enhancing the system's ability to protect data across linguistic diversity.
- **Sentiment analysis:** NLP can help identify potential security threats based on sentiment, identifying malicious intent even in apparently innocuous text.
- **Advanced pattern recognition:** NLP enables the identity of complicated patterns that would indicate information leaks or unauthorized access.



2. **Need for Comprehensive Approaches**
   This shows that there is a requirement for more holistic and sophisticated techniques for data protection. NLP strategies should offer a more nuanced understanding of data, leading to better protection against various types of threats such as information leaks or unauthorized access.

   Existing system approaches would possibly consist of traditional security measures such as access controls, encryption, and firewalls. However, those might lack the depth of understanding that NLP brings. in our NLP-based system, we could use:

   - **Named Entity Recognition:** Identify specific entities within files, allowing for better categorization and protection of sensitive data.
   - **Keyword analysis:** detect specific keywords that could signify sensitive information or security breaches.
   - **Subject matter Modeling:** Understand the main topics in a record, assisting in classification and data protection.

3. **Limitations in Handling Document Formats**
   Some data protection systems may struggle with different document formats (e.g., PDFs, Word files, spreadsheets). By way of integrating NLP, those systems could potentially gain the ability to extract, analyze, and protect information regardless of the format it's offered in.

- **Existing system:** The existing system might only handle PDF and JPG formats, limiting its ability to process content material from Word documents or Excel spreadsheets.
- **Novel system (with NLP):** Your system improves this limitation by incorporating NLP techniques that can handle various formats like PDF, JPG, Word documents, and Excel spreadsheets enabling the extraction of text from different formats, taking into account consistent analysis and protection across formats.

4. **The system analyzes the document word by word.**
   This statement highlights the level of granularity at which the NLP techniques might operate. Analyzing documents word by word allows for a detailed exam of the content, that may uncover hidden patterns, anomalies, or sensitive information that might not be apparent at a higher level.
   - **Existing system:** The existing system might process documents at a better level, missing intricate details and potential threats present at the word level.
   - **Novel system (with NLP):** Your system's approach to analyzing documents word by word allows for a more detailed examination. NLP helps in understanding the nuances of each word and identifying hidden threats, sensitive information, or irregularities that could indicate data breaches.

1.4 Research Problem

There are concerns about records security and protection due to the growing amount of textual data that is created and exchanged in digital contexts. although natural Language Processing (NLP) tools have the potential to improve data security tactics, there are still difficulties in accurately locating and protecting sensitive content.

In the realm of data protection, the identification of sensitive data within uploaded documents provides a significant hurdle. conventional methods frequently fall short in precisely pinpointing such content, leaving data vulnerable to breaches and unauthorized access. The emergence of natural Language Processing (NLP) techniques holds the promise of substantially elevating the accuracy and efficiency of sensitive information extraction from various document types and formats. Moreover, the integration of an intelligent user awareness system can play a pivotal role in notifying users about the presence of sensitive information and fostering informed decision-making regarding document handling.

creating a reliable and accurate approach for leveraging NLP methods to improve data security is the main research challenge of this project.

The initiative specifically targets to solve the following difficulties:

- **Difficulty in identifying sensitive information from the uploaded document.**
  The core challenge revolves around the difficulty in identifying sensitive information within documents. This may include personally identifiable information (PII), financial facts, confidential business details, and more. existing methods often lack the nuance and contextual understanding required for accurate detection.

- **Advancements in NLP will drive improved accuracy and efficiency in sensitive information extraction.**
  Our proposed solution seeks to harness the advancements in NLP to address the identification challenge. NLP techniques consisting of named entity recognition, topic modeling, sentiment analysis, and advanced language models can be leveraged to extract sensitive information more accurately and efficiently.

- **Provide an awareness message or alert to users, notifying them about sensitive information.**
  The underlying premise is that incorporating NLP techniques can significantly enhance both the accuracy and efficiency of identifying sensitive information. By understanding the context, semantics, and relationships between words and phrases, NLP can overcome the limitations of rule-based or keyword-based techniques. Beyond technical enhancements, your research problem also highlights the importance of user awareness. Implementing an alert or notification system that informs users about the presence of sensitive information empowers them to make informed decisions regarding sharing, storing, or handling their documents.

- **Contextual Accuracy.**
  Pre-existing NLP methods, including Named Entity recognition (NER), can also have difficulty recognizing contextual alterations, which might result in inaccurate identification of sensitive information included in text.

- **Integration of security Mechanisms.**
  To guarantee data privacy throughout, NLP-driven data protection mechanisms like encryption or anonymization must be carefully integrated into the analytical pipeline.

- **Ethical Framework**

  To guarantee ethical and open practices, it is important to address the ethical implications of automated data analysis and protection, including questions of user permission and privacy. By tackling these problems, this project seeks to advance the area of data protection through the usage of NLP methods to improve the precision and efficacy of data security mechanisms inside textual data.

2. OBJECTIVE

The objective of this research component lies in harnessing the strength of natural Language Processing (NLP) techniques to strengthen data protection strategies. By integrating advanced NLP algorithms into the data analysis process, the aim is to enhance the system's ability to identify and flag sensitive information within textual data. This intelligent system endeavors to provide users with real-time alerts about potential risks, enabling them to take immediate action to prevent data breaches.

Additionally, the research seeks to explore the contextual intricacies of language, allowing the system to understand diffused versions in meaning and context, thereby contributing to the accuracy of sensitive information detection. Through the creation of user-friendly interfaces and continuous learning mechanisms, the research aspires to develop an adaptive and effective data protection solution that not only aligns with the evolving nature of cyber threats but also empowers users to make informed decisions in safeguarding critical information.

2.1 Main Objective

Develop an intelligent system that employs natural Language Processing (NLP) techniques to examine documents, thereby heightening user awareness regarding sensitive information present in these documents. The machine should promptly generate alerts to inform users about the identified sensitive information.

2.2 Sub Objectives

Implementing Natural Language Processing (NLP) techniques to analyze documents: The research aims to implement advanced Natural Language Processing techniques to analyze different types of documents. these techniques will help the system understand the text and identify patterns that might indicate sensitive information.

Ensuring the system's accuracy in identifying and flagging sensitive information: One of the key sub-objectives is to ensure the system's accuracy in recognizing and flagging sensitive information. The NLP algorithms should be fine-tuned to minimize false positives and fake negatives, ensuring reliable results.

Providing user-friendly interfaces for efficient interaction and understanding of alerts: Creating user-friendly interfaces is a crucial aspect of the research. The system must provide an intuitive and easy-to-understand interface that allows users to interact efficiently with the alerts generated via the system. This aspect is essential to ensure that users can take appropriate actions based on the identified sensitive information.

Enhancement of data classification: Develop methodologies to accurately classify and categorize different types of sensitive information, such as personal identifiers, financial data, and confidential keywords, using NLP techniques.

Real-time Alert generation: Design a mechanism to generate actual-time indicators and notifications for users on every occasion sensitive statistic is detected within the analyzed statistics, ensuring quick response to capacity dangers.

Contextual analysis: Explore advanced NLP techniques that consider the contextual meaning of words and phrases, enabling the system to identify subtle variations of sensitive information within varying contexts.

Evaluation of overall performance: assess the effectiveness of the built-in information protection mechanisms and the produced NLP models.

By fulfilling these goals, this research hopes to provide useful knowledge on how to use NLP approaches for improved data security. The initiative aims to close the gap between cutting-edge technology and the pressing need to protect sensitive information by means of integrating powerful language processing skills with strong security measures.

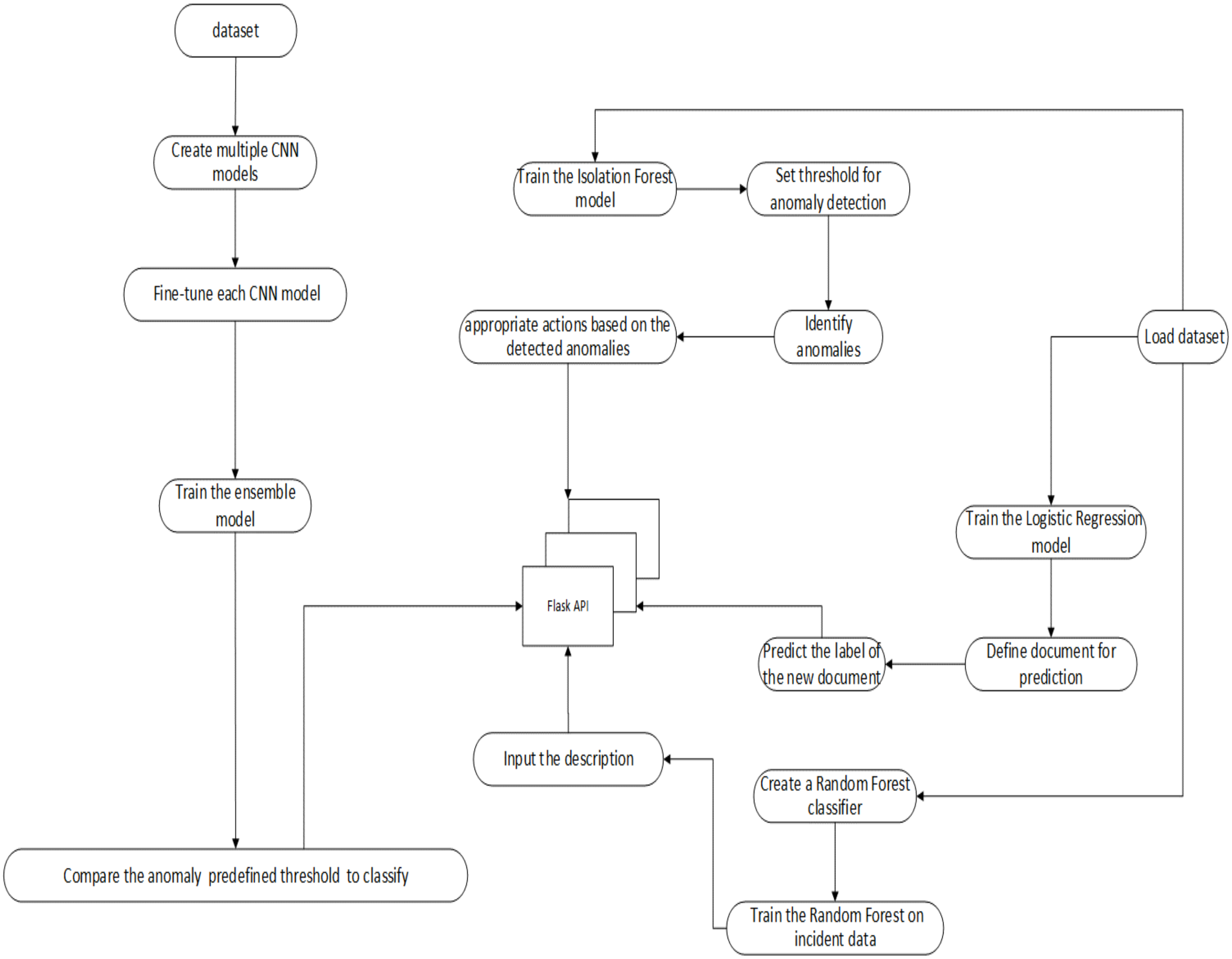
3. METHODOLOGY

The project's methodology includes a methodical approach that blends data security measures with natural Language Processing (NLP) tools. A thorough literature analysis will be conducted before the study gets started to identify pertinent NLP approaches, current data protection measures, and ethical issues. The next step entails data gathering and preparation, during which a wide range of textual datasets will be vetted and made ready for analysis. sensitive information will be accurately identified using NLP models, which will be created and put into use.

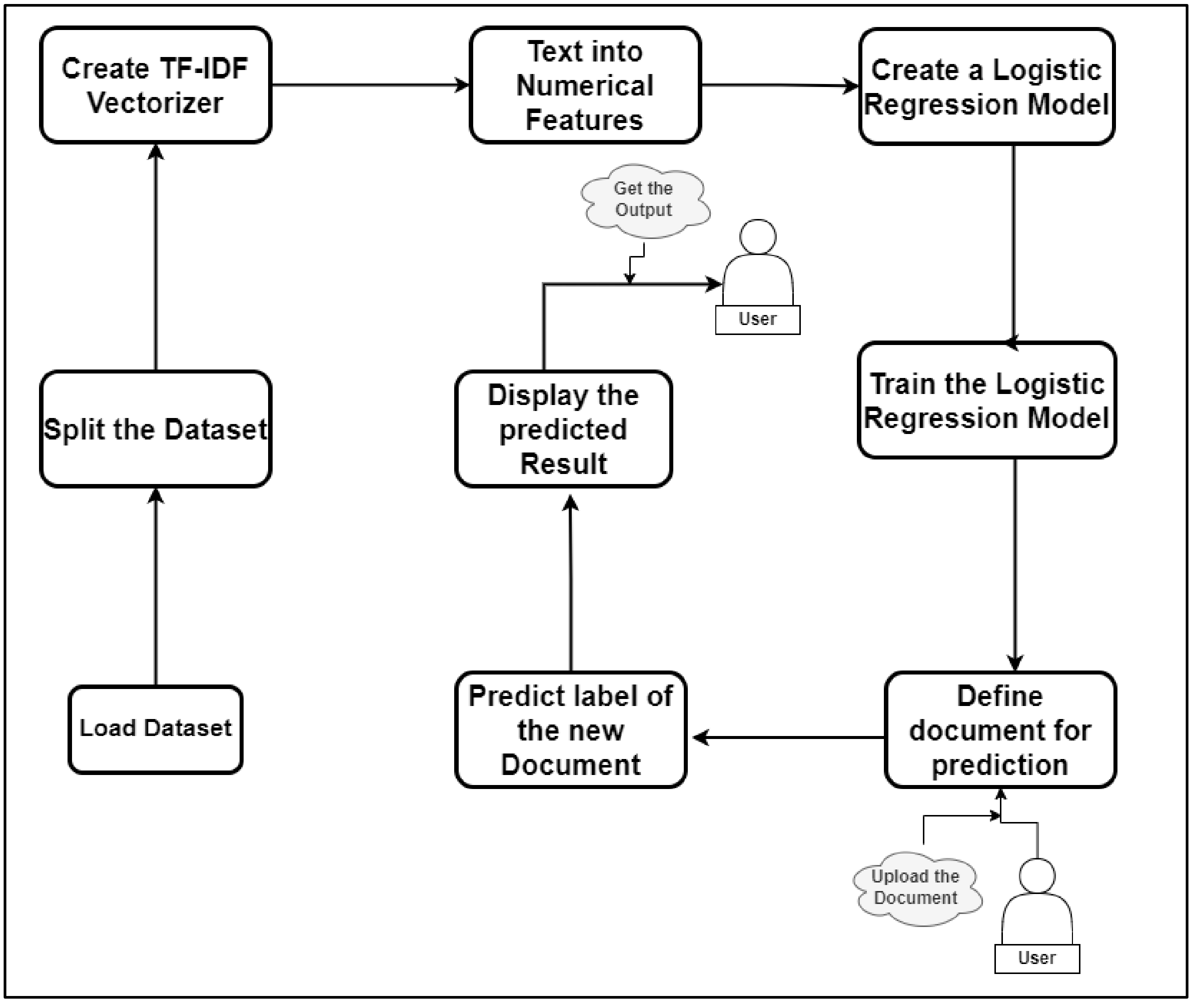
These models will include methods like sentiment analysis and Named Entity recognition (NER). Transformer-based models will also be investigated and improved to efficiently categorize sensitive information. To guarantee data security throughout NLP analysis, integration of data protection mechanisms like encryption and anonymization might be painstakingly carried out. Comprehensive testing utilizing benchmark datasets and comparison of outcomes to pre-established criteria are both included in the performance review process. Finally, an ethical framework stressing user permission, data protection, and responsible data processing will be designed and put into place. The approach of the project goals to integrate NLP developments with data protection requirements, providing useful insights into improved data security in textual data analysis.

3.1 System Diagram

Overall System Diagram


dataset
Create multiple CNN models
Fine-tune each CNN model
Train the ensemble model
Compare the anomaly predefined threshold to classify
Train the Isolation Forest model
Set threshold for anomaly detection
Identify anomalies
appropriate actions based on the detected anomalies
Load dataset
Train the Logistic Regression model
Define document for prediction
Predict the label of the new document
Flask API
Input the description
Create a Random Forest classifier
Train the Random Forest on incident data

System Diagram for Individual Research Component
(Utilizing NLP Techniques for Enhanced Data Protection)


Create TF-IDF Vectorizer
Text into Numerical Features
Create a Logistic Regression Model
Get the Output
User
Split the Dataset
Display the predicted Result
Train the Logistic Regression Model
Load Dataset
Predict label of the new Document
Define document for prediction
Upload the Document
User

The NLP Engine, which includes a couple of NLP fashions in rate of responsibilities like Named Entity Recognition (NER) and sentiment evaluation, is at the center of the device. For the cause of finding and categorizing sensitive data, these models have a look at incoming textual facts.

data safety Modules, which use encryption and anonymization techniques, are easily incorporated with the NLP Engine. those modules make sure that non-public statistics are stored private throughout the analysis method. To enhance transformer-based total models' accuracy in categorizing touchy data, a Transformer version Optimization Layer investigates and tweaks them. this accretion aids in the correct recognition of diffused patterns within the textual content. additionally, an ethical Framework Module carries user permission strategies, privateness rules, and transparency guidelines to guarantee the ethical remedy of statistics.

The machine gets input from the facts collection and preprocessing stage in the form of uncooked textual content information for analysis. The output that has been processed, complete with classifications and statistics security measures, is then made accessible for similar usage or take a look at. The gadget Diagram demonstrates how NLP abilities and facts security features may go collectively seamlessly to offer a complete framework that improves records security whilst allowing significant textual facts analysis.

- **Data Input Data Processing**

The gathering, curation, and compilation of the textual data for later analysis include the statistics consumption and preprocessing phase of the venture. relevant textual records may be gathered from a variety of assets, including net articles, social media postings, and documents, to guarantee a varied and consultant dataset. The selection-making manner may be ethically guided to make certain that the information is collected ethically and in regard to user privacy.

information will undergo sizeable preprocessing after series to standardize codecs, do away with noise, and manage any discrepancies. we can use strategies like tokenization and stemming to divide the textual content into digestible chunks and go back to inflected words to their authentic shape. text cleansing procedures will even remove useless characters, symbols, and formatting artifacts that could compromise the correctness of later NLP studies.

For accurate NLP analysis and version training, the data pretreatment level is a vital starting line. This stage seeks to minimize any biases and ensure that the venture's next levels are primarily based on a robust and representative textual facts basis through curating a high-quality, well-structured dataset.

- **Future Extraction**

The project's earliest stages, which center on gathering and getting the dataset geared up for analysis, are the information consumption and feature extraction segment. The dataset could be compiled from an expansion of applicable textual resources, including documents, blog posts, and internet articles. The technique of amassing statistics could be guided by way of moral concerns, assuring responsible acquisition at the same time as shielding a person's privateness.

The raw textual information may be transformed into established representations suitable for analysis after data series and the usage of feature extraction techniques. Tokenization will divide the text into understandable chunks, and extraordinary linguistic elements, consisting of element-of-speech tags and named entities, may be extracted to offer context.

- **Classification**

In the classification step, textual input is categorized and classified in keeping with predetermined standards using state-of-the-art devices gaining knowledge of algorithms. The herbal Language Processing (NLP) models created for this mission may be used to categorize text into one-of-a-kind organizations, which include sensitive and non-sensitive information. These models will employ information gleaned from the preprocessing stage and employ strategies that have evolved the use of labeled datasets. that allows you to discover touchy facts within the textual facts, dependable and effective class is wanted. The categorization process could be crucial in figuring out if the NLP-pushed statistics protection strategy is powerful.

- **Dataset**

I will be collecting data for datasets from some small and medium-level companies. To gather relevant textual data for analysis, business firms are worked with throughout the dataset acquisition procedure with the purpose of making sure that statistics gathering complies with felony requirements and respects the sensitivity of agency statistics, ethical issues, and facts privateness might be prioritized. The number of records collected, the sorts of textual fabric to be covered, and the security precautions in the vicinity will all be outlined in collaboration with interested commercial enterprise companions. After careful preprocessing to assure quality, consistency, and anonymity, the collected dataset will then be ready for future herbal Language Processing (NLP) analysis. The mission's NLP-pushed statistics safety techniques might be evaluated using a representative and critical dataset that becomes acquired collaboratively.

- **Simulation**

In the simulation phase, models and approaches are positioned to be used in a controlled setting to simulate actual-world instances for assessment. In this venture, simulated information situations reflecting the favored utility could be used to test the NLP fashions and facts protection mechanisms. those conditions will cover several language complexities, sensitivity, and types. The facts safety strategies will guarantee confidentiality and adherence to moral ideas while the NLP fashions will categorize and hold sensitive fabric. To assess the precision, efficacy, and performance of the incorporated NLP-pushed records protection answer, the simulation effects might be examined. prior to deployment in the actual international, this segment enables a thorough evaluation, permitting vital revision and optimization.

5. PROJECT REQUIREMENTS

4.1 Functional Requirements

The NLP-driven data safety machine's required capabilities are particular in the project's functional requirements. The NLP models must, first and main, efficiently categorize text into predetermined categories, isolating sensitive from non-sensitive material. To attain thorough identification, the models must also be able to control differences in language and context.

- Identify the sensitive Information.
- Data file handling.
- Analyze Data and Alerting.
- High processing rate

4.2 Non-Functional Requirements

The non-functional criteria for the mission include factors that go beyond specific talents to assure the gadget's widespread pleasant, overall performance, and value. First and predominant, the NLP models have to respond quickly, presenting real-time categorization without considerable latency. The device ought to be extra specific and correct than preset standards for identifying sensitive information.

- **Accurate:** The output given by the application is more accurate since we are using a highly accurate dataset with a lot of data.
- **Effective and efficient**
- **Availability:** The application is always available when the user is needed. Available 24/7.
- **Usability:** The application is easy to use and has a user-friendly interface, that allows any user without any technical knowledge to seamlessly interact with the product.
- **Security**

4.3 Technologies

The foundation is natural Language Processing (NLP), with version introduction and evaluation finished using NLP libraries. system getting to know frameworks with a view to making it simpler to teach and enhance NLP models. Encryption techniques that adhere to enterprise standards are used inside the integration of information safety techniques to ensure the secure management of touchy statistics. The venture intends to provide a powerful NLP-driven facts safety solution that clings to enterprise great practices by way of the use of this technology.

- Natural Language Processing (NLP)
- Feature Extraction Technology
- Supervised learning algorithm
-

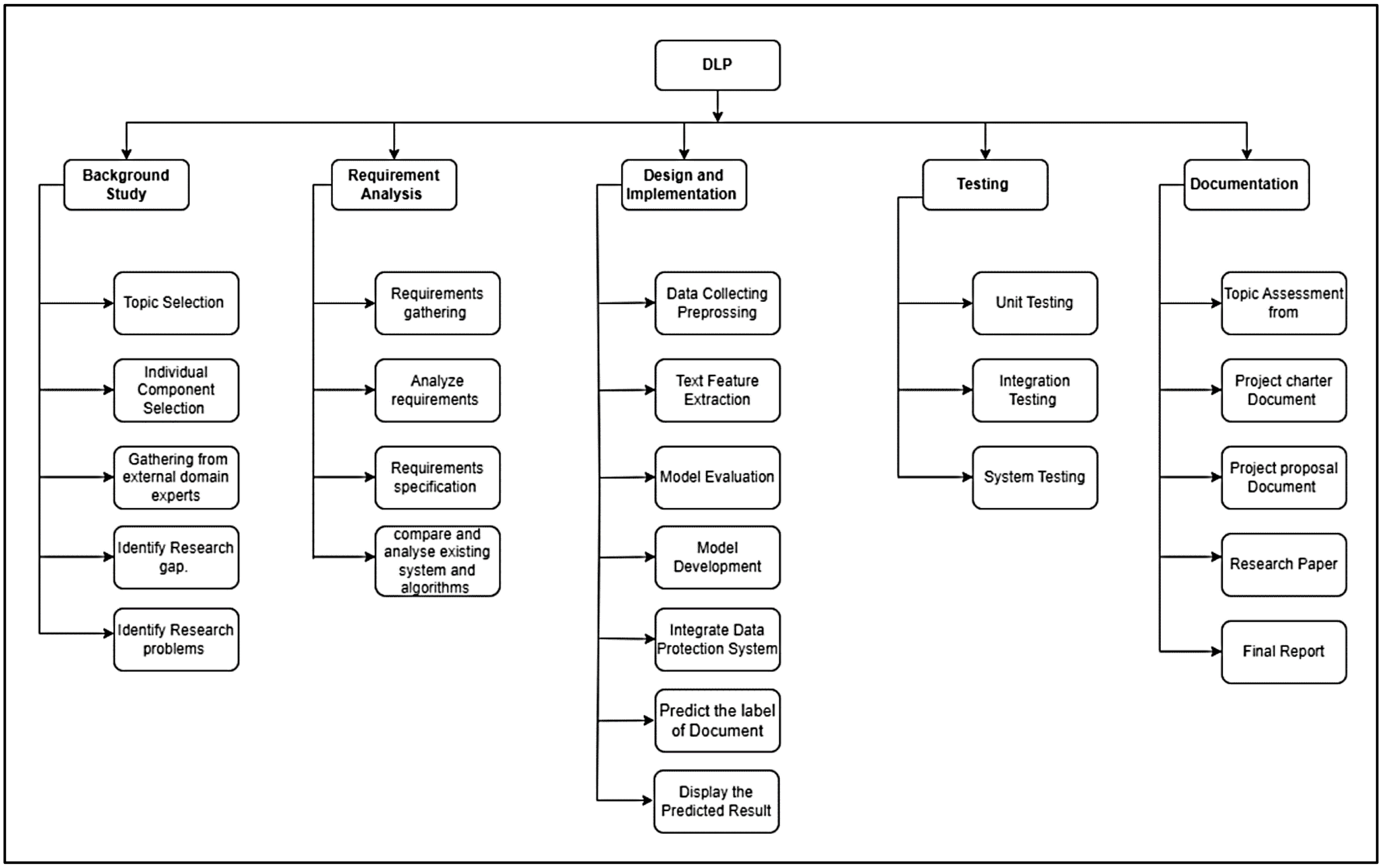
Language to use - Python.

- Python is one of the most extensively used programming languages.

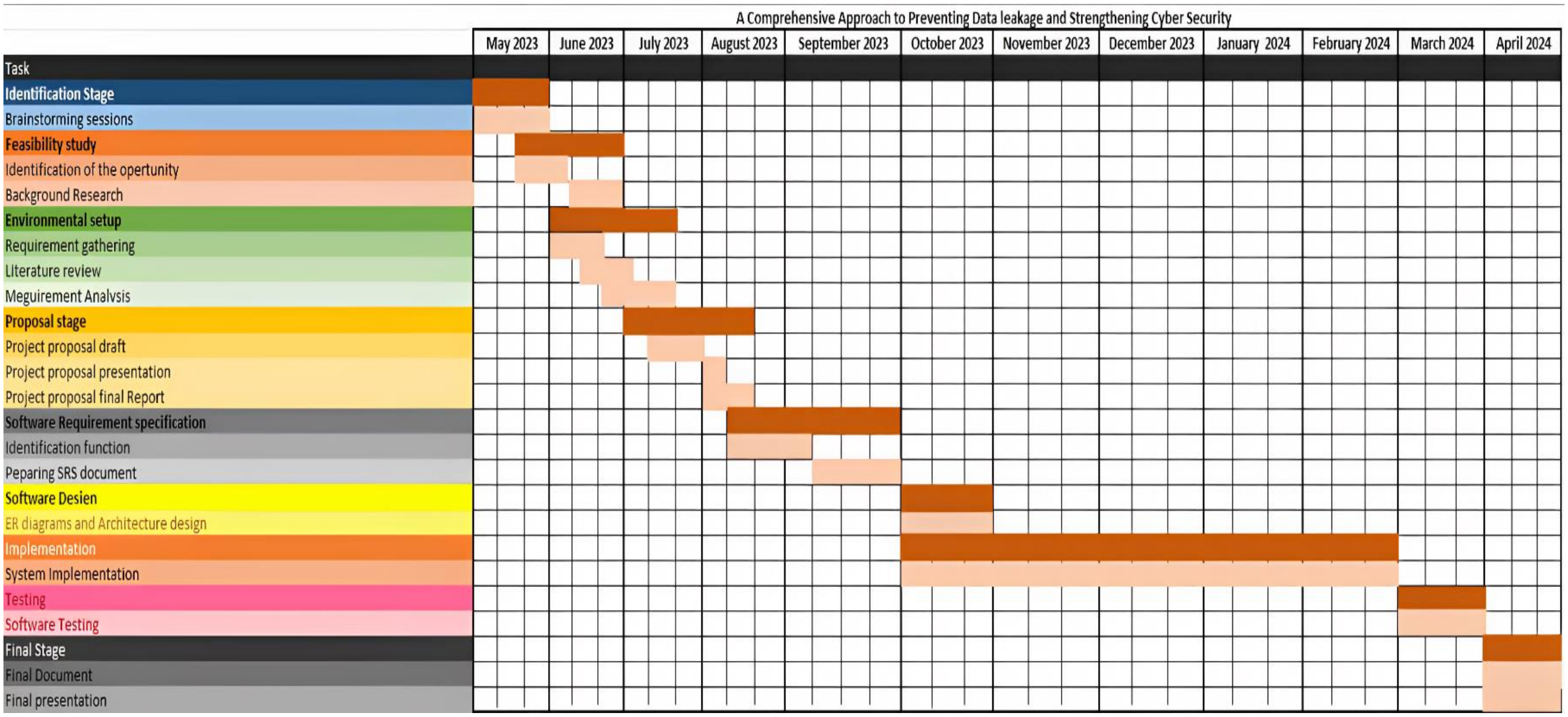
Tools

- PyCharm
- Visual Studio Code

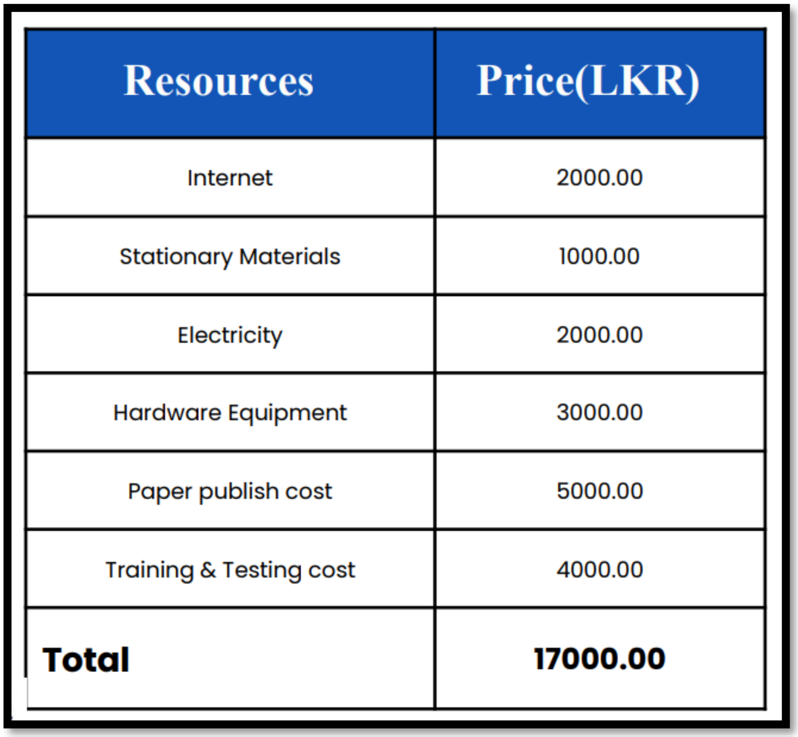
5. WORK BREAKDOWN STRUCTURE (WBS)


DLP
Background Study
Topic Selection
Individual Component Selection
Gathering from external domain experts
Identify Research gap.
Identify Research problems
Requirement Analysis
Requirements gathering
Analyze requirements
Requirements specification
compare and analyse existing system and algorithms
Design and Implementation
Data Collecting Preprossing
Text Feature Extraction
Model Evaluation
Model Development
Integrate Data Protection System
Predict the label of Document
Display the Predicted Result
Testing
Unit Testing
Integration Testing
System Testing
Documentation
Topic Assessment from
Project charter Document
Project proposal Document
Research Paper
Final Report

6. GANTT CHART



7. BUDGET



Resources	Price(LKR)
Internet	2000.00
Stationary Materials	1000.00
Electricity	2000.00
Hardware Equipment	3000.00
Paper publish cost	5000.00
Training & Testing cost	4000.00
Total	**17000.00**

8. COMMERCIALIZATION

- **Create a sales plan for the industry.**

Industry analysis: Conduct thorough market research to identify industries with the highest demand for comprehensive cybersecurity solutions. Understand their pain points, regulatory requirements, and specific challenges associated with data leakage and cybersecurity.

Tailored fee Proposition: Craft a compelling value proposition that directly addresses the unique needs of each targeted industry. Highlight how our system's capabilities align with its requirements, showcasing its effectiveness in preventing data leakage and strengthening overall cybersecurity.

Strategic Partnerships: Forge partnerships with influential industry associations, organizations, and thought leaders. Collaborate to co-host webinars, workshops, or events that position our system as an innovative solution, gaining credibility and expanding our reach within the industry.

- **Design customer subscription plans.**

  Tiered Plans: Develop various subscription levels, each catering to different business sizes and needs. Offer options like basic, standard, and premium plans, each with a distinct set of features and capabilities.

  Scalability: Ensure that your subscription plans are designed to accommodate the growth and changing requirements of businesses. Provide flexibility for clients to upgrade or adjust their plans as their needs evolve.

  Customization Flexibility: Integrate customization options within subscription plans, allowing clients to tailor features based on their specific data protection needs. This ensures that they only pay for the functionalities they require.

- **Provide excellent customer support.**

  Dedicated support team: Assign a team of knowledgeable support representatives to promptly address inquiries and issues.

  Multi-Channel assist: Provide support through various channels such as email, live chat, and phone.

  24/7 Availability: offer round-the-clock assistance for critical concerns and urgent inquiries.

  Knowledge Base: Develop an online resource with FAQs, tutorials, and troubleshooting guides.

  Continuous training: Offer training sessions to help clients maximize the benefits of the system.

9. REFERENCE

[1] Babarinde, L., & Ray, T. (n.d.). *Data loss prevention (DLP)*. Learning Center; Imperva Inc. Retrieved August 25, 2023, from https://www.imperva.com/learn/data-security/data-loss-prevention-dlp/

[2] Bhattacharyya, S. (n.d.). *The ethical considerations of natural language processing (NLP)*. Analyticssteps.com. Retrieved August 25, 2023, from https://analyticssteps.com/blogs/ethical-considerations-natural-language-processing-nlp

[3] Gokce, E. (2020, May 12). *Beginner's guide to data cleaning and feature extraction in NLP*. Towards Data Science. https://towardsdatascience.com/beginners-guide-for-data-cleaning-and-feature-extraction-in-nlp-756f311d8083

[4] Kumar, R. (n.d.). *Data Leakage Detection*. Globaljournals.org. Retrieved August 25, 2023, from https://globaljournals.org/GJCST_Volume17/3-Data-Leakage-Detection.pdf

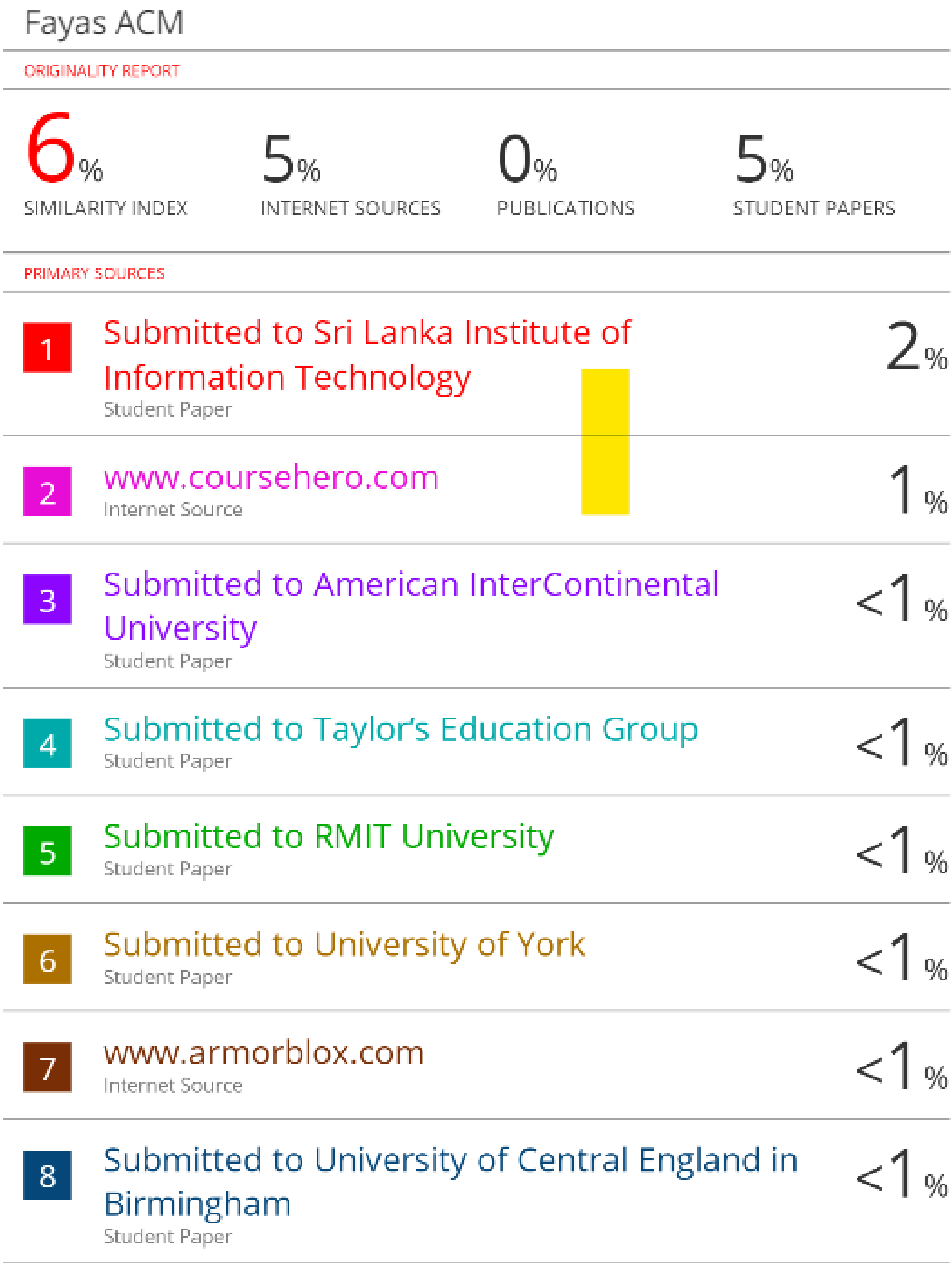
[5] *Major challenges of natural language processing (NLP)*. (2020, December 22). MonkeyLearn Blog. https://monkeylearn.com/blog/natural-language-processing-challenges/

[6] *Protecting personal information: A guide for business*. (2016, October 2). Federal Trade Commission. https://www.ftc.gov/business-guidance/resources/protecting-personal-information-guide-business

[7] Rashid, F. Y. (2022, March 16). *Enhancing DLP with natural language understanding for better email security*. Dark Reading. H ttps://www.darkreading.com/emerging-tech/enhancing-dlp-with-natural-language-understanding-for-better-email-security

[8] (N.d.-a). Researchgate.net. Retrieved August 25, 2023, from https://www.researchgate.net/publication/224146568_Data_Leakage_Detection

APPENDICES



Fayas ACM

ORIGINALITY REPORT

6%	5%	0%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Sri Lanka Institute of Information Technology Student Paper	2%
2	www.coursehero.com Internet Source	1%
3	Submitted to American InterContinental University Student Paper	<1%
4	Submitted to Taylor's Education Group Student Paper	<1%
5	Submitted to RMIT University Student Paper	<1%
6	Submitted to University of York Student Paper	<1%
7	www.armorblox.com Internet Source	<1%
8	Submitted to University of Central England in Birmingham Student Paper	<1%