CHAPTER I

INTRODUCTION AND ITS BACKGROUND

Research plays a crucial role in the development of academic institutions, as it contributes to the creation of new knowledge and innovation. As a result, universities invest heavily in research initiatives to ensure that faculty members and students are actively involved in conducting research. However, managing the vast amount of research materials produced by academic institutions is becoming increasingly challenging with traditional research management practices such as physical storage and manual cataloging no longer being sufficient. This includes not just data, but also analytical processes, tools, and knowledge structures.

Research output management, as stated by Houweling et al., (2023), is the essential aspect of managing the vast amount of data and information generated by academic institutions, and it must be done carefully as part of the research process. By adhering to the FAIR principles and using research objects and RO-Crate solutions, researchers can effectively manage and enhance the value of research outputs. Expanding the concept of research data management to research output management fosters a more holistic approach to managing research outputs throughout the research process.

Laguna State Polytechnic University - Sta. Cruz Campus is an academic institution that promotes research initiatives supporting its vision and mission. The campus has numerous faculty members and students conducting research in various fields, which generates a vast amount of research materials, such as journal articles, conference papers, dissertations, theses, and other academic publications. However, the campus currently lacks a centralized research repository system, resulting in research materials being scattered and challenging to locate.

To address these challenges, this study proposes the development of an advanced web-based research repository system, called ResearchVault, for Laguna State Polytechnic University - Sta. Cruz Campus. The system will incorporate intelligent chatbot integration and collaborative features to enhance efficient research management. It will enable the seamless storage, organization, and management of research materials, making it easy for faculty members and students to locate research materials quickly. Advanced search capabilities will enable users to search for research materials based on keywords, authors, publication date, and other criteria.

An intelligent chatbot will also be included in the system to assist users in navigating the research repository and answering common queries. The chatbot will provide a conversational interface that enables users to access research materials and insights in a user-friendly and efficient manner.

In addition to the features mentioned above, ResearchVault will also include a special functionality that allows users to transform a manuscript into a succinct IMRAD format using Natural Language Processing (NLP) and text mining techniques. This particular feature will prove especially valuable for educators and students who are required to condense their research outcomes for presentation or publication purposes.

By implementing ResearchVault, Laguna State Polytechnic University - Sta. Cruz Campus will have a centralized research repository system that enables faculty members and students to quickly locate research materials, and collaborate efficiently, thereby enhancing research output and efficiency.

Research Problem

Research is essential to academic institutions, but managing the increasing volume of research materials produced is becoming more challenging. Google Drive, as the current system used for digital storage, has its limitations. One significant limitation is its provision of a limited amount of free storage space, and acquiring additional storage incurs additional costs. The current system makes it difficult for faculty members and students to access and share knowledge, which can limit collaboration and reduce research output.

The proposed study aims to develop an advanced web-based research repository system, called ResearchVault, for Laguna State Polytechnic University - Sta. Cruz Campus. The system will incorporate intelligent chatbot integration and collaborative features to enhance research management efficiency, enable the seamless storage, organization, and management of research materials, and increase collaboration and research output.

Specifically, the study sought to answer the following research problems:

1. How to design and develop an advanced web-based research repository system that includes an intelligent chatbot for RIU consultation and an IMRAD converter to condense thesis manuscripts into 10 pagers?
2. How to effectively preprocess research data, including tasks such as text extraction from PDFs, section classification, and content organization, to create a clean and structured dataset for machine learning model training?
3. How to train machine learning models, including LDA (Latent Dirichlet Allocation), Random Forest, and the deep learning model BERT (Bidirectional Encoder Representations from Transformers), using the preprocessed research dataset to ensure optimal performance and utility in the web-based research repository system?
4. How to conduct model evaluation of the trained machine learning model which utilized LDA, Random Forest and the deep learning model BERT (Bidirectional Encoder Representations from Transformers) with in terms of accuracy, precision, recall, and F1-score to identify which to integrate to the system for converting thesis manuscript into IMRAD and Chatbot?
5. How can the efficacy of integrated system-incorporated models be assessed through actual testing procedures?

Research Objectives

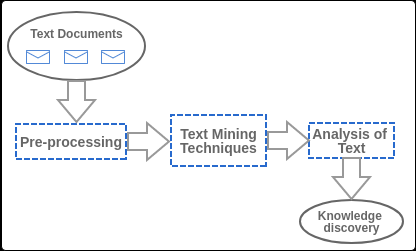
The general objective of this research is to develop an advanced web-based research repository for Laguna State Polytechnic University - Sta. Cruz Campus with intelligent chatbot integration and collaborative features for efficient research management.

Specifically, this research aims to:

1. To design and develop an advanced web-based research repository system with chatbot integration and IMRAD converter for Laguna State Polytechnic University - Sta. Cruz Campus that allows for the seamless storage, organization, and management of research materials and enhances efficient research management.
2. To perform data preprocessing tasks, including text extraction from PDFs, section classification, and content organization, to prepare the research dataset for machine learning model training.
3. To train machine learning models using the preprocessed research dataset, and leverage techniques such as LDA (Latent Dirichlet Allocation), Random Forest, and the deep learning model BERT (Bidirectional Encoder Representations from Transformers) with various sets of hyperparameters.
4. To conduct model evaluation based on the accuracy, precision, recall, and F1-score of trained machine learning model which utilized LDA (Latent Dirichlet Allocation), BERT (Bidirectional Encoder Representations from Transformers) and Random Forest with distinct numbers of parameters, to determine which machine learning model to integrate.
5. To assess the performance of the integrated models by conducting actual testing within the system.

Theoretical Framework

The theoretical framework establishes how the algorithms will be implemented on the system's development process. As highlighted by Jadhav et al.'s survey (2023), text mining, a facet of data mining, is delineated as the procedure for distilling implicit knowledge from textual data, proving essential for managing extensive volumes of unstructured text data. Thus, the framework of text mining serves as the guide structure for this study.



**Figure 1:** Text Mining Framework. Source: Saquicela et al., (2018).

The text mining process, drawing inspiration from Jadhav et al., (2023), is a comprehensive procedure that involves several crucial stages, including text preprocessing and the discovery of knowledge.

Text preprocessing is the initial stage, which deals with the transformation of unstructured text into a more structured format. This could be either document-based or concept-based, depending on the requirements of the analysis. This stage is critical as it prepares the raw text data, making it suitable for further processing and analysis.

Following preprocessing, the next stage is knowledge discovery. In a document-based intermediate form, knowledge discovery involves uncovering patterns or trends across a collection of documents or texts. This is achieved through the application of various text mining techniques. Techniques such as topic modeling, clustering, and algorithms play a pivotal role in this stage. These techniques are primarily used for grouping or categorizing related textual data, thereby simplifying the process of analysis.

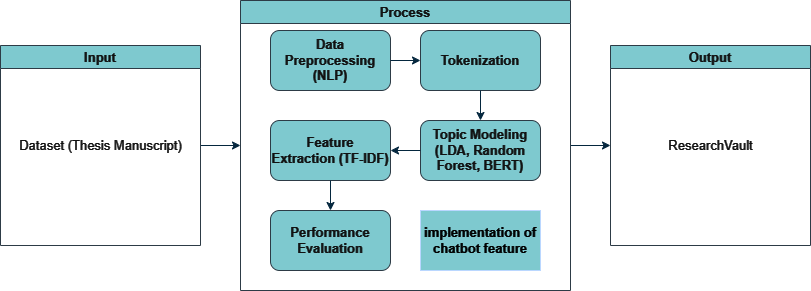
On the other hand, when dealing with a concept-based intermediate form, knowledge discovery takes on a slightly different approach. It involves identifying patterns or relationships between objects or concepts through rigorous analysis. This process allows for a deeper understanding of the data, revealing insights that might not be immediately apparent.

Conceptual Framework

The proposed research repository system for Laguna State Polytechnic University - Sta. Cruz Campus is designed to provide an efficient and effective platform for research management and knowledge sharing among faculty members, students, and other external stakeholders.

The ResearchVault will serve as a digital repository for research materials, accessible to registered users. The repository will offer several key features to support research activities, such as chatbot, citation generator, chatbot and IMRAD converter. The ResearchVault repository and intelligent chatbot will work together to provide an enhanced research experience for users.

The conceptual framework of the proposed research repository system is summarized in the following diagram:



**Figure 2:** Conceptual framework of the study.

The conceptual framework draws on existing theories and literature to explain a phenomenon. It outlines the steps needed for the study based on insights from other researchers' viewpoints and observations related to the study subject. This framework guides the research process with a clear and informed approach. The study integrated three essential text mining techniques: LDA topic modeling, TF-IDF feature conversion, and Random Forest classification along with Bidirectional Encoder Representations (BERT). The process commenced by amalgamating various CSV data sources into a unified dataset, laying the groundwork for streamlined data handling.

In the analytical phase, by simplifying text into clusters of words and employing LDA, the study revealed underlying themes within the data. This technique illuminated prominent topics and their prevalence across different documents. The process then goes towards numerical conversion, transforming preprocessed text into weighted TF-IDF vectors. This conversion allocated values based on word importance within documents and the larger dataset. Culminating in a pivotal juncture, a classifier underwent training. This classifier effectively grouped text data into predefined categories. This training process involved a subset of the dataset, with subsequent assessment occurring using an independent testing set. To analyze the effectiveness of the classifier, a comprehensive evaluation was undertaken. This examination provided insights into precision, recall, F1-score, and support metrics for each category, offering a clear understanding of classification performance.

Scope and Limitations of the Study

The proposed research repository system for Laguna State Polytechnic University - Sta. Cruz Campus aims to provide an efficient platform for research management and collaboration. The scope of the project is limited to the development of the web-based research repository, as well as the integration of an intelligent chatbot feature and IMRAD converter using NLP techniques. The following are the detailed scope and limitations of the project:

* Development of a web-based research repository that incorporates key features such as Sort & Filter when browsing research materials, Comments, Citation Generation, PDF Viewer for manuscript, and IMRAD converter.
* Integration of an intelligent chatbot feature that provides assistance and answer queries related RIU consultations.
* The proposed research repository system is designed specifically for Laguna State Polytechnic University - Sta. Cruz Campus, and may not be applicable to other institutions.
* The system will rely on the availability and quality of the research materials that are uploaded by the users, and cannot guarantee the accuracy and completeness of the research materials.

Significance of the Study

The proposed research repository system for Laguna State Polytechnic University - Sta. Cruz Campus aims to provide an efficient and user-friendly platform for managing and sharing research materials. The significance of this study lies in its potential to address the challenges and limitations associated with traditional research management approaches and to enhance the research productivity and collaboration among faculty members and students.

Specifically, the system's output would be of great benefit to the following individuals or groups:

**Research Implementing Unit Head**. The system would enhance RIUH’s efficiency by providing an overview of student research papers, enabling effective management and tracking of undergraduate theses, and facilitating the creation of an IMRAD journal.

**Dean of the College.** The system would provide the dean with an overview of the progress of undergraduate theses, allowing them to monitor the research productivity of faculty members and students more efficiently

**Students.** The system would provide undergraduate students with unlimited access to the online research paper repository and enable them to gather more relevant research papers for their studies

**Technical Panel Members.** The system would provide technical panel members with ease of checking the works of the student researchers they are working with, allowing them to provide more timely and relevant feedback.

**Course Instructors.** The system would provide course instructors with ease of checking the works of the students they are handling, allowing them to monitor and assess their students' research productivity more efficiently.

**Researchers.** The study would be beneficial to present and future researchers who would be conducting a study related to text mining, which involves topic modeling and clustering.

**Community.** The system would encourage a particular group of researchers to avoid searching for research papers outside the university for their own convenience. Moreover, other institutions could replicate this study in their own communities to be part of the growing popularity of digital repositories and to enhance research productivity and collaboration.

CHAPTER II

REVIEW OF RELATED LITERATURE

This chapter presents a review of relevant literature and studies, both local and foreign. The aim is to synthesize and assess the various related sources to provide a comprehensive understanding of the topic at hand. The review will cover current research, theories, and approaches related to research repository systems, chatbot integration, and collaborative features. It will also explore the potential benefits and challenges of these technologies in the context of efficient research management.

Natural Language Processing (NLP)

In the words of Long (2022) entitled “A Grammatical Error Correction Model for English Essay Words in Colleges Using Natural Language Processing” it was discussed that Natural language processing technology is a theory and approach for exploring and developing successful human-computer communication. With the rapid growth of computer science and technology, statistical learning methods have become an important research area in artificial intelligence and semantic search. If there are errors in the semantic units (words and sentences), it will affect future text analysis and semantic understanding, eventually affecting the whole application system performance. As a result, intelligent word and grammatical error detection and correction in English text are a significant and difficult aspect of natural language processing. Therefore, this paper examines the phenomena of word spelling and grammatical errors in undergraduate English essays and balances the mathematical-statistical models and technology solutions involved in intelligent error correction. The research findings of this study are represented in two aspects. (1) In nonword mistakes, four sorts of errors are studied: insertion, loss, replacement, and exchange between letters. It focuses on nonword mistakes and varied word forms (such as English abbreviations, hyphenated compound terms, and proper nouns) produced by word pronunciation difficulties. This paper utilizes the nonword check information to recommend an optimum combination prediction method based on the suggested candidate list for actual word errors, and the genuine word repair model is trained. This approach is 83.78% accurate when used with actual words with spelling errors in the context. (2) It verifies and corrects sentence grammar using context information from the text training set, as well as grammatical rules and statistical models. In addition, it has investigated singular and plural inconsistency, word confusion, subject, and predicate inconsistency, and modal (auxiliary) verb errors. It includes sentence boundary disambiguation, word part-of-speech tagging, named entity identification, and context information extraction. The software for checking and fixing sentence grammatical mistakes presented in this article works on English texts with difficulty levels 4 and 6. Furthermore, this work obtains a clause correctness rate of 99.70%, and the system’s average corrective accuracy rate for four-level and six-level essays is more than 80%.

Furthermore, Guler & Akgul (2022) present a study titled “A Review on The Science of Natural Language Processing” highlighting that in today’s world, it is not possible for a normal human brain to understand and interpret the rapidly growing data piles at the same time. The use of artificial intelligence techniques is a great requirement for the perception and interpretation of big data with the help of machines. One of the most popular sub-branches of artificial intelligence is natural language processing. The spoken language used by each country has its own characteristics. Natural language processing is a technique that analyzes the features of world languages and makes sentiment analysis. Natural language processing is the most important building block in the creation of language between humans and machines. In this way, it enables the textual data to be converted into the desired language in the digital environment, the logical answers to the questions asked, the synthesis and summary of digital texts, and the guidance of the machines with voice commands. The natural language processing process begins by moving words, which are the smallest building blocks that make sense in the text, into the digital world. It brings the model to the creation process after preprocessing, rooting and feature extraction operations on the words transferred to the digital world. After these processes, a connection is established between machine language and human. With natural language processing technology, communication between human and machine can now be provided with only spoken language without the need for a port input.

In the study conducted by Chaudhary et al., (2022) titled “Intelligent Virtual Research Environment for Natural Language Processing (IvrE-NLP)” In the 21st +century, natural language processing (NLP) has obtained much prominence for human–machine interaction (HMI). With this interest in natural language processing (NLP) has grown significantly, numerous NLP tools (e.g., morphology, the tagger, and a parser, etc.) have been developed all over the world. Despite having huge importance and requirements, we have noticed gaps for having a comprehensive single framework or platform, which encompass all NLP-related tools and technologies for promoting the research in NLP and sharing the knowledge and resources among NLP researchers required for understanding and building the solution for HMI. Our objective is to apply Software engineering in natural language processing with the concept of an object-oriented model by using a collection of reusable objects by defining the communication protocol, consisting of a set of rules that must be applied to exchange data between two NLP modules. We proposed state of art ivrE—A virtual environment for creating, modifying, executing, and analyzing various NLP solutions and technology. The proposed idea is broadly based on to define own ivrE-NLP object framework model that permits the developer to create, modify, and execute the application and analyze their outcomes by operations on visual representations of the modules. A variety of NLP-based applications (tools, modules, and plugins) already exist, they can publish into store available with environment so it can be used by research community at large. To develop complete NLP framework or platform, we require much more than just assembling or collecting these tools or modules at one place, no matter how good any tool or module is working individually. It requires not only the standards and a set of protocols, but also requires a compliant composition than a pre-defined algorithms and their implementation. In brief, we require a comprehensive open framework to bundle, manage, and integrate set of NLP tools, modules, components, applications, algorithms, and define their associated rules, comprehensive data structures, and knowledge.

In a recent study by Ali (2022), titled “AI-Natural Language Processing (NLP)” Natural Language Processing (NLP) could be a branch of Artificial Intelligence (AI) that allows machines to know the human language. Its goal is to form systems that can make sense of text and automatically perform tasks like translation, spell check, or topic classification. Natural language processing (NLP) has recently gained much attention for representing and analysing human language computationally. It's spread its applications in various fields like computational linguistics, email spam detection, information extraction, summarization, medical, and question answering etc. The goal of the Natural Language Processing is to style and build software system which will analyze, understand, and generate languages that humans use naturally, so as that you just could also be ready to address your computer as if you were addressing another person. Because it’s one amongst the oldest area of research in machine learning it’s employed in major fields like artificial intelligence speech recognition and text processing. Natural language processing has brought major breakthrough within the sector of COMPUTATION AND AI.

As previously discussed by Sanadi et al., (2022) titled "A Review Paper on Natural Language Processing (NLP)," the authors explore the technology of Natural Language Processing (NLP), which serves as a bridge between humans and machines, making machines more human-like and facilitating easier communication between the two. NLP has witnessed significant advancements and a wide range of applications over the past few decades. These applications have proven highly beneficial in everyday life, such as voice-command-activated machines. Numerous research organizations are actively working on enhancing NLP to develop more practical and user-friendly solutions. The potential of NLP lies in its ability to create computer interfaces that are more intuitive for humans, enabling individuals to communicate with computers in their natural language instead of requiring them to learn a specific computer language.

To sum up the studies of Long (2022), Guler & Akgul (2022), Chaudhary et al. (2022), Ali (2022), and Sanadi et al. (2022) collectively provide insights into the multifaceted field of Natural Language Processing (NLP). Long (2022) focuses on the practical applications of NLP, particularly in the context of grammatical error correction in English essays, showcasing the significance of accurate language understanding for effective human-computer communication. Guler & Akgul (2022) emphasize the pivotal role of AI techniques, including NLP, in processing and interpreting the growing volumes of data, underlining NLP's ability to bridge the gap between human language and machine understanding. Chaudhary et al. (2022) contributes by proposing an integrated framework, ivrE-NLP, that addresses the need for a comprehensive platform to unify various NLP tools and technologies, fostering research collaboration and knowledge sharing in the NLP community. Ali (2022) highlights NLP's broad applicability, from computational linguistics to spam detection, illustrating its enduring significance in the domain of machine learning and artificial intelligence. Finally, Sanadi et al. (2022) explore the evolving landscape of NLP, emphasizing its role in making human-computer interactions more natural and intuitive, offering a vision of NLP as a bridge that brings machines closer to human understanding. These studies collectively underscore the diverse applications and growing importance of NLP in enhancing language comprehension, data analysis, and human-machine interactions.

Bidirectional Encoder Representations

In their research, Mohan et al., (2023). Presented The Internet has become a crucial space for customer feedback and the budding of various ideologies across different cultures. But some people provide their opinions whose sole meaning is different from what their figurative meanings imply. This sophisticated form of sentiment expression through mockery or irony is called sarcasm. Sarcastic form of texts has taken a prominent place in social media and their growth has become exponential in recent years through messages and posts. Sarcastic comments, tweets, or feedback can be misleading in data mining activities and may result in wrong predictions since the boundaries of the sarcastic context are not well defined. There is a need to model new systems that can precisely define the boundaries of the sentence and detect sarcasm in them. Different methods of classification have been used for sarcasm detection in various works including deep learning, neural networks, and other architectures. This paper focuses on combining the capabilities of both Bidirectional Encoder Representations from Transformers (BERT) and Graph Convolutional Networks (GCN) for detecting sarcastic content from text. A BERT-GCN architecture is proposed which takes as input, the graph, and text representations and learns the complex structural and semantic patterns in the text, thereby detecting sarcastic content. The efficiency of BERT-GCN is compared with various baseline methods.

As noted by Long (2023) entitled “Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model” Sentiment is currently one of the most emerging areas of research due to the large amount of web content coming from social networking websites. Sentiment analysis is a crucial process for recommending systems for most people. Generally, the purpose of sentiment analysis is to determine an author’s attitude toward a subject or the overall tone of a document. There is a huge collection of studies that make an effort to predict how useful online reviews will be and have produced conflicting results on the efficacy of different methodologies. Furthermore, many of the current solutions employ manual feature generation and conventional shallow learning methods, which restrict generalization. As a result, the goal of this research is to develop a general approach using transfer learning by applying the “BERT (Bidirectional Encoder Representations from Transformers)”-based model. The efficiency of BERT classification is then evaluated by comparing it with similar machine learning techniques. In the experimental evaluation, the proposed model demonstrated superior performance in terms of outstanding prediction and high accuracy compared to earlier research. Comparative tests conducted on positive and negative Yelp reviews reveal that fine-tuned BERT classification performs better than other approaches. In addition, it is observed that BERT classifiers using batch size and sequence length significantly affect classification performance.

As explored in previous literature by Kalaivani et al., (2020) entitled “Sarcasm identification and detection in conversion context using BERT.” Sarcasm analysis in user conversion text is automatic detection of any irony, insult, hurting, painful, caustic, humour, vulgarity that degrades an individual. It is helpful in the field of sentimental analysis and cyberbullying. As an immense growth of social media, sarcasm analysis helps to avoid insult, hurts and humour to affect someone. In this paper, we present traditional machine learning approaches, deep learning approach (LSTM -RNN) and BERT (Bidirectional Encoder Representations from Transformers) for identifying sarcasm. We have used the approaches to build the model, to identify and categorize how much conversion context or response is needed for sarcasm detection and evaluated on the two social media forums that is twitter conversation dataset and reddit conversion dataset.

In a cross-cultural study by Bhardwaj et al., (2022) entitled Every youngster around us uses sarcasm as an indirect way to say a negative statement. With the growth of artificial intelligence and machine programming in the field of natural language programming (NLP), the detection of sarcasm efficiently and accurately has become a challenge. To contribute as a solution to this ever-growing field of interest, this paper proposes a novel approach for sarcasm detection with the use of machine learning and deep learning. This approach uses bidirectional encoder representations from transformers (BERT) to pre-process the sentence and feed it to a hybrid deep learning model for training and classification. This hybrid model uses convolutional neural networks (CNN) and long short-term memory (LSTM).

In the investigation of Deepa (2021). entitled “Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task.” The latest trend in the direction of sentiment analysis has brought up new demand for understanding the contextual representation of the language. Among the various conventional machine learning and deep learning models, learning the context is the promising candidate for the sentiment classification task. BERT is a new pre-trained language model for context embedding and attracted more attention due to its deep analyzing capability, valuable linguistic knowledge in the intermediate layer, trained with larger corpus, and fine-tuned for any NLP task. Many researchers adapted the BERT model for sentiment analysis tasks by influencing the original architecture to get better classification accuracy. This article summarizes and reviews BERT architecture and its performance observed from fine-tuning different layers and attention heads.

As explored in previous literature, various aspects of Natural Language Processing (NLP) and sentiment analysis. Mohan et al. (2023) focus on sarcasm detection, highlighting the challenges posed by sarcastic expressions in social media texts and proposing a BERT-GCN architecture to detect sarcasm accurately. Long (2023) delves into sentiment analysis, emphasizing the importance of sentiment classification in web content and introducing the BERT-based model as a powerful approach for this purpose. Kalaivani et al. (2020) extend the discussion on sarcasm identification, using traditional machine learning approaches, LSTM-RNN, and BERT to detect sarcasm in user conversation texts across social media platforms. Bhardwaj et al. (2022) address the challenge of sarcasm detection in the context of artificial intelligence and machine programming, proposing a hybrid model that combines BERT, CNN, and LSTM for efficient sarcasm detection. Finally, Deepa (2021) explores the use of BERT as a pre-trained language model for sentiment analysis, discussing its contextual embedding capabilities and its application in improving classification accuracy. Collectively, these studies showcase the versatility of NLP techniques, particularly BERT, in addressing various language-related tasks such as sarcasm detection and sentiment analysis, with a focus on improving accuracy and efficiency in understanding and interpreting textual data.

**Term Weighting Schemes**

Alshehri et al., (2023) presented a novel supervised term weighting scheme called TF-TDA (Term Frequency-Term Discrimination Ability) for sentiment analysis tasks. Traditional unsupervised term weighting schemes, such as TF-IDF, may not be sufficient for sentiment analysis. In text classification tasks, such as sentiment analysis (SA), feature representation and weighting schemes play a crucial role in classification performance. Traditional term weighting schemes depend on the term frequency within the entire document collection; therefore, they are called unsupervised term weighting (UTW) schemes. One of the most popular UTW schemes is term frequency–inverse document frequency (TF-IDF); however, this is not sufficient for SA tasks. Newer weighting schemes have been developed to take advantage of the membership of documents in their categories. These are called supervised term weighting (STW) schemes; however, most of them weigh the extracted features without considering the characteristics of some noisy features and data imbalances. Therefore, in this study, a novel STW approach was proposed, known as term frequency–term discrimination ability (TF-TDA). TF-TDA mainly presents the extracted features with different degrees of discrimination by categorizing them into several groups. Subsequently, each group is weighted based on its contribution. The proposed method was examined over four SA datasets using naive Bayes (NB) and support vector machine (SVM) models. The experimental results proved the superiority of TF-TDA over two baseline term weighting approaches, with improvements ranging from 0.52% to 3.99% in the F1 score. The statistical test results verified the significant improvement obtained by TF-TDA in most cases, where the p-value ranged from 0.0000597 to 0.0455.

Nugroho et al., (2022) conducted a study on detecting emotion in Indonesian tweets. tecting Emotion in Indonesian Tweets: A Term-Weighting Scheme Study. Journal of Information Systems Engineering and Business Intelligence. 8. 61-70. 10.20473/jisebi.8.1.61-70. Term-weighting plays a key role in detecting emotion in texts. Studies in term-weighting schemes aim to improve short text classification by distinguishing terms accurately. This study aims to formulate the best term-weighting schemes and discover the relationship between n-gram combinations and different classification algorithms in detecting emotion in Twitter texts. The data used was the Indonesian Twitter Emotion Dataset, with features generated through different n-gram combinations. Two approaches assign weights to the features. Tests were carried out using ten-fold cross-validation on three classification algorithms. The performance of the model was measured using accuracy and F1 score. The term-weighting schemes with the highest performance are Term Frequency-Inverse Category Frequency (TF-ICF) and Term Frequency-Relevance Frequency (TF-RF). The scheme with a supervised approach performed better than the unsupervised one. However, we did not find a consistent advantage as some of the experiments found that Term Frequency-Inverse Document Frequency (TF-IDF) also performed exceptionally well. The traditional TF-IDF method remains worth considering as a term-weighting scheme. This study provides recommendations for emotion detection in texts. Future studies can benefit from dealing with imbalances in the dataset to provide better performance.

Wang et al., (2021) explored entropy-based term weighting schemes for text categorization. In text categorization, Vector Space Model (VSM) has been widely used for representing documents, in which a document is represented by a vector of terms. Since different terms contribute to a document’s semantics in various degrees, a number of term weighting schemes have been proposed for VSM to improve text categorization performance. Much evidence shows that the performance of a term weighting scheme often varies across different text categorization tasks, while the mechanism underlying variability in a scheme’s performance remains unclear. Moreover, existing schemes often weight a term with respect to a category locally, without considering the global distribution of a term’s occurrences across all categories in a corpus. In this paper, we first systematically examine pros and cons of existing term weighting schemes in text categorization and explore the reasons why some schemes with sound theoretical bases, such as chi-square test and information gain, perform poorly in empirical evaluations. By measuring the concentration that a term distributes across all categories in a corpus, we then propose a series of entropy-based term weighting schemes to measure the distinguishing power of a term in text categorization. Through extensive experiments on five different datasets, the proposed term weighting schemes consistently outperform the state-of-the-art schemes. Moreover, our findings shed new light on how to choose and develop an effective term weighting scheme for a specific text categorization task.

Doğan et al., (2020) proposed a novel term weighting strategy called TF-MONO for text classification. The effective representation of the relationship between the documents and their contents is crucial to increase classification performance of text documents in the text classification. Term weighting is a preprocess aiming to represent text documents better in Vector Space by assigning proper weights to terms. Since the calculation of the appropriate weight values directly affects performance of the text classification, in the literature, term weighting is still one of the important sub-research areas of text classification. In this study, we propose a novel term weighting (MONO) strategy which can use the non-occurrence information of terms more effectively than existing term weighting approaches in the literature. The proposed weighting strategy also performs intra-class document scaling to supply better representations of distinguishing capabilities of terms occurring in the different quantity of documents in the same quantity of class. Based on the MONO weighting strategy, two novel supervised term weighting schemes called TF-MONO and SRTF-MONO were proposed for text classification. The proposed schemes were tested with two different classifiers such as SVM and KNN on 3 different datasets named Reuters-21578, 20-Newsgroups, and WebKB. The classification performances of the proposed schemes were compared with 5 different existing term weighting schemes in the literature named TF-IDF, TF-IDF-ICF, TF-RF, TF-IDF-ICSDF, and TF-IGM. The results obtained from 7 different schemes show that SRTF-MONO generally outperformed other schemes for all three datasets. Moreover, TF-MONO has promised both Micro-F1 and Macro-F1 results compared to other five benchmark term weighting methods especially on the Reuters-21578 and 20-Newsgroups datasets.

Tang et al., (2019) addressed the deficiencies of the widely used term weighting scheme, TF-IDF, and proposed an improved scheme called term frequency-inverse exponential frequency (TF-IEF). Text representation is a necessary and primary procedure in performing text classification (TC), which first needs to be obtained through an information‐rich term weighting scheme to achieve higher TC performance. So far, term frequency‐inverse document frequency (TF‐IDF) is the most widely used term weighting scheme, but it suffers from two deficiencies. First, the global weighting factors IDF in TF‐IDF approaches infinity if a certain term does not occur in a text. Second, the IDF is equal to zero if a certain term appears in any text. To offset these drawbacks, we first conduct an in‐depth analysis of the current term weighting schemes, and subsequently, an improved term weighting scheme called term frequency‐inverse exponential frequency (TF‐IEF) and its various variants are proposed. The proposed method replaces IDF with the new global weighting factor IEF to characterize the global weighting factor log‐like IDF in the corpus, which can greatly reduce the effect of feature (term) with high local weighting factor TF in term weighting. As a result, a more representative feature can be generated. We carried out a series of experiments on two commonly used data sets (corpora) utilizing Naïve Bayes and support vector machine classifiers to validate the performance of our proposed schemes. Experimental results explicitly reveal that the proposed term weighting schemes come with better performance than the compared schemes.

To sum it all, these studies tackled the challenges of sentiment analysis, emotion detection, and text categorization by introducing innovative term-weighting approaches designed to improve classification performance. Alshehri et al. (2023) introduce TF-TDA, a supervised scheme that outperforms traditional unsupervised methods. Nugroho et al. (2022) explore various term-weighting schemes for emotion detection, emphasizing the strengths of supervised approaches while acknowledging the continued relevance of TF-IDF. Wang et al. (2021) propose entropy-based schemes that consistently surpass existing methods, shedding light on global term distribution's importance. Doğan et al. (2020) introduce TF-MONO and SRTF-MONO, delivering superior results, particularly on Reuters-21578 and 20-Newsgroups datasets. Tang et al. (2019) address TF-IDF's limitations with TF-IEF, showcasing better performance in text classification. Together, these papers highlight the critical role of tailored term weighting in enhancing text analysis and classification outcomes across diverse domains.

Text Mining

In a comprehensive review of Abdusalomovna & T. D. (2023). argued that text mining is a science about the generalized language of informatics, which appeared on the basis of the methods of machine learning and the rules of statistics. Text mining (also called text analytics) is the use of natural language processing (NLP) to analyze free (unstructured) text in documents and databases or transform it into normalized, structured data suitable for machine control. artificial intelligence (AI) technology. learning (ML) algorithms.In our independent work, we provide an introduction to these technologies and highlight some of the features that contribute to an efficient solution.

In the study titled “A review on authorship attribution in text mining” by Zheng et al., (2023), The issue of authorship attribution has long been considered and continues to be a popular topic. Because of advances in digital computers, this field has experienced rapid developments in the last decade. In this article, a survey of recent advances in authorship attribution in text mining is presented. This survey focuses on authorship attribution methods that are statistically or computationally supported as opposed to traditional literary approaches. The main aspects covered include the changes in research topics over time, basic feature metrics, machine learning techniques, and the advantages and disadvantages of each approach. Moreover, the corpus size, number of candidates, data imbalance, and result description, all of which pose challenges in authorship attribution, are discussed to inform future work.

In a recent article Tavana et al., (2022). A Review of Digital Transformation on Supply Chain Process Management Using Text Mining. Industry 4.0 technologies are causing a paradigm shift in supply chain process management. The digital transformation of the supply chains provides enormous benefits to organizations by empowering collaboration among multiple internal and external organizations and systems. This study presents a narrative review explaining the existing knowledge on digital transformation in supply chain process management using text mining. It summarizes the existing literature to explain the current state of the art in supply chain digitalization. This comprehensive review identifies the most important topics and technologies and determines the future trends in this emerging field. We investigate the articles published in Web of Science and Scopus databases and use text mining techniques (clustering and topic modeling) on the article contents. Using VOS viewer, a bibliometric analysis of 395 articles with 12,700 references is analyzed. The contents of the articles are explored using text mining approaches. The synthesized results reveal that the most important topics in digital transformation are “sustainable supply chain management” and “circular economy and industry 4.0 technologies”. The study further discovers big data, data analytics, blockchain, artificial intelligence, machine learning, and the Internet of Things as the most critical technologies for facilitating supply chain digital transformation. Finally, an overlay heatmap analysis of the research articles found that digital transformation, supply chain management, industry 4.0, decision-making, and sustainability are emerging trends in supply chain digitalization.

According to a report by Antons et al., (2020) The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. Unstructured data in the form of digitized text is rapidly increasing in volume, accessibility, and relevance for research on innovation and beyond. While traditional attempts to analyze text (i.e., qualitative analysis) are limited in processing large amounts of data, text mining presents a set of approaches that allow researchers to explore large-scale collections of texts in an efficient manner. Given the potential of text mining as a method of inquiry, the primary purpose of this manuscript is to enable both novice and more experienced innovation researchers to select, specify, document, and interpret text mining techniques in a way that generates valid and reliable knowledge for the innovation management community. This involved taking stock of text mining applications in the field of innovation research to date by means of a systematic review of 124 journal articles employing text mining techniques and are published in a basket of the 10 premier innovation management and 8 top general management journals. The results of the systematic manual and computational analysis of these articles do not only illustrate the state and evolution of text mining applications in our field, but also allow for evidence-based recommendations regarding their future use. Here, our paper presents methodological, conceptual, and contextual development priorities that will contribute to establishing higher methodological standards in text mining and enhance the methodological richness in our field.

As mentioned in the literature review Kumar et al., (2015). “Text mining and similarity search using extended tri-gram algorithm in the reference based local repository dataset.” In the emerging technological scenario world is becoming a digital hub where data is easily accessible on the internet. When we take into consideration the academic and research & development fields, the digital resources become more important. Academic research and their publications in current environment can be easily accessed through internet. Easy availability of such research work attracts academic literature dishonesty or plagiarism. Many of the research papers published in the several Conference-proceedings and Journals may have some percentage of plagiarized contents. At the same time, the author(s) may cite irrelevant references in the research paper. In the present research paper, a reference based extended trigram approach has been reported to check the textual plagiarism of the text written in the research papers. References form the pivotal part of any research paper or dissertation which defines the area of research and the state of the art of the research based on which originality of the contribution is adjudged. The present article also discusses the behavior of the referencing in three major research categories i.e., Research papers, Master Dissertations, and the Doctoral Dissertations.

In the analysis of these case studies, it collectively shed light on various aspects of text mining and its applications in different domains. Abdusalomovna & T. D. (2023) introduce text mining as a science rooted in machine learning and statistical rules, emphasizing its role in transforming unstructured text into structured data, suitable for artificial intelligence and machine learning algorithms. Zheng et al. (2023) delve into authorship attribution in text mining, focusing on statistically or computationally supported methods and addressing challenges like data imbalance and result description in this context. Tavana et al. (2022) explore the digital transformation of supply chain management through text mining, identifying key topics and technologies shaping the field, including sustainability and industry 4.0 technologies. Antons et al. (2020) conducts a systematic review of text mining applications in innovation research, offering insights into the state and evolution of text mining in this domain and providing recommendations for future use. Kumar et al. (2015) tackles the issue of textual plagiarism in academic research by proposing an extended trigram approach for reference-based similarity checking, highlighting the importance of maintaining originality and proper referencing in research papers, dissertations, and theses. Together, these papers showcase the versatility and growing relevance of text mining in various research domains, from authorship attribution to supply chain management and plagiarism detection, highlighting its potential to uncover valuable insights from unstructured textual data.

Deep Learning

This study of “Deep Learning” by LeCun et al., (2015). Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech, and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Sharifani et al., (2023). present the study “Machine Learning and Deep Learning: A Review of Methods and Applications”. Machine learning and deep learning have rapidly emerged as powerful tools in many fields, including image and speech recognition, natural language processing, and even medicine. In this article, we provide a review of the methods and applications of machine learning and deep learning, including their strengths and weaknesses, as well as their potential future directions. We also discuss the challenges associated with these technologies, including data privacy, ethical considerations, and the need for transparency in the decision-making process. Machine learning and deep learning are two of the most revolutionary technologies in the field of artificial intelligence. They have become increasingly popular in recent years due to their ability to make predictions, analyze large datasets, and provide insights that were previously impossible to obtain. This article will explore the basics of machine learning and deep learning, their differences, applications, and their impact on various industries. Machine learning and deep learning are transforming the way we interact with technology and unlocking new possibilities for innovation. These technologies have already made significant impacts in various industries and have the potential to continue to revolutionize the world. This article provides a comprehensive overview of the basics of machine learning and deep learning, their differences, applications, and their impact on society. With a focus on current literature and research, this article aims to provide a better understanding of the potential of machine learning and deep learning and their implications for the future.

Mathew et al., (2021). present the study “Deep learning techniques: an overview.” Deep learning is a class of machine learning which performs much better on unstructured data. Deep learning techniques are outperforming current machine learning techniques. It enables computational models to learn features progressively from data at multiple levels. The popularity of deep learning amplified as the amount of data available increased as well as the advancement of hardware that provides powerful computers. This article comprises the evolution of deep learning, various approaches to deep learning, architectures of deep learning, methods, and applications.

Shinde et al., (2018, August). present the study “A review of machine learning and deep learning applications.” machine learning is one of the fields in the modern computing world. A plenty of research has been undertaken to make machines intelligent. Learning is a natural human behavior which has been made an essential aspect of the machines as well. There are various techniques devised for the same. Traditional machine learning algorithms have been applied in many application areas. Researchers have put many efforts to improve the accuracy of that machinelearning algorithms. Another dimension was given thought which leads to deep learning concept. Deep learning is a subset of machine learning. So far few applications of deep learning have been explored. This is definitely going to cater to solving issues in several new application domains, sub-domains using deep learning. A review of these past and future application domains, sub-domains, and applications of machine learning and deep learning are illustrated in this paper.

In a study on workplace dynamics Gopinath et al., (2023). entitle “A comprehensive survey on deep learning-based malware detection techniques.” Recent theoretical and practical studies have revealed that malware is one of the most harmful threats to the digital world. Malware mitigation techniques have evolved over the years to ensure security. Earlier, several classical methods were used for detecting malware embedded with various features like the signature, heuristic, and others. Traditional malware detection techniques were unable to defeat new generations of malware and their sophisticated obfuscation tactics. Deep Learning is increasingly used in malware detection as DL-based systems outperform conventional malware detection approaches at finding new malware variants. Furthermore, DL-based techniques provide rapid malware prediction with excellent detection rates and analysis of different malware types. Investigating recently proposed Deep Learning-based malware detection systems and their evolution is hence of interest to this work. It offers a thorough analysis of the recently developed DL-based malware detection techniques.

In their study, LeCun et al. (2015) delve into the realm of deep learning, showcasing its capability to enable computational models to learn complex data representations through multiple layers of abstraction. They emphasize its transformative impact on various domains, from speech recognition to genomics, and elucidate how backpropagation facilitates the adjustment of internal parameters for representation computation in each layer. Meanwhile, Sharifani et al. (2023) provide a comprehensive review of machine learning and deep learning, elucidating their strengths, weaknesses, applications, and future prospects. They emphasize the revolutionary potential of these technologies, highlighting their impact on diverse industries and addressing associated challenges like data privacy and ethical considerations. Mathew et al. (2021) offers an overview of deep learning, emphasizing its superiority in handling unstructured data and its evolution, approaches, architectures, methods, and applications. Shinde et al. (2018) explore the broad landscape of machine learning and deep learning applications, highlighting their evolution and their potential to address various challenges across domains. Lastly, Gopinath et al. (2023) focus on deep learning-based malware detection techniques, highlighting their effectiveness in combating evolving threats and providing rapid malware prediction with excellent detection rates, thus offering insights into the crucial domain of cybersecurity.

Machine Learning (ML)

In the course of study of Möller (2023) entitled “Machine Learning and Deep Learning” states that Machine Learning is a sub-category of Artificial Intelligence enabling computers with the ability of pattern recognition, or to continuously learn from, making predictions based on data, and carry out decisions without being specifically programmed for doing so. In this context, Machine Learning is a broader category of algorithms being able to use datasets to identify patterns, discover insights, and enhance understanding and make decisions or predictions. Compared with Machine Learning, Deep Learning is a particular branch of Machine Learning that makes use of Machine Learning functionality, and moves beyond its capabilities. Deep Learning Algorithm is interpreted as a layered structure that tries to replicate the structure of the human brain. These capabilities enable Machine Learning and Deep Learning Algorithms usage in applications to identify and respond to cybercriminals manifold cyberattacks. This is achieved by analyzing Big Datasets of cybersecurity incidents to identify patterns of malicious activities. For this purpose, Machine Learning and Deep Learning compare known threat event attacks with detected threat event attacks to identify similarities they automatically dealt with trained Machine Learning or Deep Learning model for response.

As explored in the research of Chopra & Khurana (2023) entitled “Introduction to Machine Learning”, Machine learning is a subfield of artificial intelligence, broadly defined as a machine's capability to imitate intelligent human behavior. Like humans, machines become capable of making intelligent decisions by learning from their past experiences. Machine learning is being employed in many applications, including fraud detection and prevention, self-driving cars, recommendation systems, facial recognition technology, and intelligent computing. This book helps beginners learn the art and science of machine learning. It presents real-world examples that leverage the popular Python machine learning ecosystem, The topics covered in this book include machine learning basics: supervised and unsupervised learning, linear regression and logistic regression, Support Vector Machines (SVMs). It also delves into special topics such as neural networks, theory of generalisation, and bias and fairness in machine learning. After reading this book, computer science and engineering students - at college and university levels - will receive a complete understanding of machine learning fundamentals and will be able to implement neural network solutions in information systems, and also extend them to their advantage.

Bell (2022) states in the study entitled “What is Machine learning”, Over the last six decades, several pioneers of the industry have worked to steer us in the right direction. There are a number of different algorithms that one can employ in machine learning (ML). The required output is what decides which to use. ML algorithms characteristically fall into one of two learning types: supervised or unsupervised learning. ML is widely used in software to enable an improved experience with the user. Using ML, robots can acquire skills or learn to adapt to the environment in which they are working. Robots can acquire skills such as object placement, grasping objects, and locomotion skills through either automated learning or learning via human intervention. The race is on for ML to be used in healthcare analytics. A number of start-ups are looking at the advantages of using ML with big data to provide healthcare professionals with better-informed data to enable them to make better decisions.

The study of Gupta et al., (2022) entitled “Machine Learning Techniques for Digital Library Services” discusses the digital library as an organised system of digital storage, digital computing and communications mechanism with the data and software needed to reproduce, emulate and execute the services. Point out on digitization, need of digitization, digitization process, challenges of digitization, hardware and software for digitization, challenges of library professional, role of library professionals, digital library initiatives and future of digital libraries in India. What an Artificial Intelligence and machine learning technique are help to enhance the quality of digital library services the end users.

In the context of the study of Jordan & Mitchell (2015) entitled “Machine Learning: Trends, perspectives, and prospects” states that Machine learning addresses the question of how to build computers that improve automatically through experience. It is one of today’s most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, and marketing.

The reviewed literature on machine learning and deep learning underscores their pivotal roles in the realm of artificial intelligence. Möller (2023) elucidates that machine learning encompasses a broader category of algorithms designed to identify patterns and make predictions based on data, while deep learning, as a specialized branch of machine learning, mimics the human brain's layered structure for enhanced capabilities. In a similar vein, Chopra & Khurana (2023) emphasize machine learning's ability to replicate human intelligence through learning from experiences and its applications in various domains. Bell (2022) delves into machine learning algorithms, distinguishing between supervised and unsupervised learning, and highlighting its wide-ranging applications, including robotics and healthcare analytics. Gupta et al. (2022) discusses the role of machine learning in enhancing digital library services, emphasizing its potential for improving user experiences and managing digital resources. Lastly, Jordan & Mitchell (2015) provide a broader perspective, noting the rapid growth of machine learning driven by evolving algorithms, increasing data availability, and its widespread adoption across diverse fields, emphasizing the shift toward evidence-based decision-making.

Random Forest

In the context of the study of Jaiswal, J. K., & Samikannu, R. (2017, February) entitled “Application of random forest algorithm on feature subset selection and classification and regression.”. It discards insignificant variables and produces efficient and improved prediction performance on the class variables that is more cost effective and more reliable understanding of the data. Random forest has been emerged as a quite efficient and robust algorithm that can handle feature selection problem even with the higher number of variables. It is also very much efficient while dealing with Missing data imputation, classification, and regression problems. It can also handle outliers and noisy data very well. In this paper we applied the concept of random forest algorithm on the feature subset selection and classification and regression to perform the comparative study of the random forest algorithm in different perspectives.

As indicated by the research of Parmar et al., (2019). entitled “A review on random forest: An ensemble classifier.” Ensemble classification is an information mining approach which utilizes various classifiers that cooperate for distinguishing the class label for new unlabeled thing from accumulation. Arbitrary Forest approach joins a few randomized choice trees and totals their forecasts by averaging. It has grabbed well-known attention from the community of research because of its high accuracy and superiority which additionally increase the performance. Now in this paper, we take a gander at improvements of Random Forest from history to till date. Our approach is to take a recorded view on the improvement of this prominently effective classification procedure. To begin with history of Random Forest to main technique proposed by Breiman then successful applications that utilized Random Forest and finally some comparison with other classifiers. This paper is proposed to give non specialists simple access to the principle thoughts of random forest.

As noted in the abstract of Mohapatra et al., (2020) entitled “Mohapatra, N., Shreya, K., & Chinmay, A. (2020)” Optimization algorithms are implemented for making the field of machine learning more efficient by comparing various solutions until an optimum or a satisfactory answer is found to yield a better accuracy score than the earlier existing one. In this paper, optimization of the Random Forest is performed which is a supervised learning model for classification and regression. A detailed analysis of the optimization technique of this model is done, which follows the unequal weight voting strategy, where weight is assigned based on how well an individual tree performs.

According to a report of Ren et al., (2017). “Research on machine learning framework based on random forest algorithm.” With the continuous development of machine learning, industry and academia have released a lot of machine learning frameworks based on distributed computing platform, and have been widely used. However, the existing framework of machine learning is limited by the limitations of machine learning algorithm itself, such as the choice of parameters and the interference of noises, the high using threshold and so on. This paper introduces the research background of machine learning framework, and combined with the commonly used random forest algorithm in machine learning classification algorithm, puts forward the research objectives and content, proposes an improved adaptive random forest algorithm (referred to as ARF), and on the basis of ARF, designs and implements the machine learning framework.

In the examination of Disha et al., (2022). Entitled “Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique.” To protect the network, resources, and sensitive data, the intrusion detection system (IDS) has become a fundamental component of organizations that prevents cybercriminal activities. Several approaches have been introduced and implemented to thwart malicious activities so far. Due to the effectiveness of machine learning (ML) methods, the proposed approach applied several ML models for the intrusion detection system. In order to evaluate the performance of models, UNSW-NB 15 and Network TON\_IoT datasets were used for offline analysis. Both datasets are comparatively newer than the NSL-KDD dataset to represent modern-day attacks. However, the performance analysis was carried out by training and testing the Decision Tree (DT), Gradient Boosting Tree (GBT), Multilayer Perceptron (MLP), AdaBoost, Long-Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) for the binary classification task. As the performance of IDS deteriorates with a high dimensional feature vector, an optimum set of features was selected through a Gini Impurity-based Weighted Random Forest (GIWRF) model as the embedded feature selection technique. This technique employed Gini impurity as the splitting criterion of trees and adjusted the weights for two different classes of the imbalanced data to make the learning algorithm understand the class distribution. Based upon the importance score, 20 features were selected from UNSW-NB 15 and 10 features from the Network TON\_IoT dataset. The experimental result revealed that DT performed well with the feature selection technique than other trained models of this experiment. Moreover, the proposed GIWRF-DT outperformed other existing methods surveyed in the literature in terms of the F1 score.

To sum it all, Jaiswal and Samikannu (2017) demonstrate how Random Forest aids in feature subset selection, classification, and regression, improving prediction performance and handling issues like missing data, outliers, and noisy data effectively. Parmar et al. (2019) provide an overview of Random Forest's evolution, highlighting its history, the main technique proposed by Breiman, successful applications, and comparisons with other classifiers, emphasizing its high accuracy and effectiveness. Mohapatra et al. (2020) delve into the optimization of Random Forest, focusing on its supervised learning model for classification and regression, particularly exploring unequal weight voting strategies based on individual tree performance. Ren et al. (2017) contribute to the field by introducing an improved adaptive Random Forest algorithm and designing a machine learning framework, addressing limitations in existing machine learning frameworks and emphasizing the importance of addressing algorithmic constraints. Lastly, Disha et al. (2022) explore the application of Random Forest in intrusion detection systems, employing a Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique to enhance performance. They compare various machine learning models for intrusion detection, highlighting the effectiveness of the GIWRF-DT model in achieving superior results, especially in terms of the F1 score, compared to existing methods.

Topic Modeling

Churchill & Singh (2022) conducted a comprehensive analysis of the evolution of topic modeling in their study titled "The Evolution of Topic Modeling" Topic models have been applied to everything from books to newspapers to social media posts in an effort to identify the most prevalent themes of a text corpus. We provide an in-depth analysis of unsupervised topic models from their inception to today. We trace the origins of different types of contemporary topic models, beginning in the 1990s, and we compare their proposed algorithms, as well as their different evaluation approaches. Throughout, we also describe settings in which topic models have worked well and areas where new research is needed, setting the stage for the next generation of topic models.

In the study titled "A Review of Topic Modeling Methods" by Vayansky & Kumar (2020) Topic modeling is a popular analytical tool for evaluating data. Numerous methods of topic modeling have been developed which consider many kinds of relationships and restrictions within datasets; however, these methods are not frequently employed. Instead many researchers gravitate to Latent Dirichlet Analysis, which although flexible and adaptive, is not always suited for modeling more complex data relationships. We present different topic modeling approaches capable of dealing with correlation between topics, the changes of topics over time, as well as the ability to handle short texts such as encountered in social media or sparse text data. We also briefly review the algorithms which are used to optimize and infer parameters in topic modeling, which is essential to producing meaningful results regardless of method. We believe this review will encourage more diversity when performing topic modeling and help determine what topic modeling method best suits the user needs.

Dieng et al., (2020) present the study "Topic Modeling in Embedding Spaces" Topic modeling analyzes documents to learn meaningful patterns of words. However, existing topic models fail to learn interpretable topics when working with large and heavy-tailed vocabularies. To this end, we develop the embedded topic model (etm), a generative model of documents that marries traditional topic models with word embeddings. More specifically, the etm models each word with a categorical distribution whose natural parameter is the inner product between the word’s embedding and an embedding of its assigned topic. To fit the etm, we develop an efficient amortized variational inference algorithm. The etm discovers interpretable topics even with large vocabularies that include rare words and stop words. It outperforms existing document models, such as latent Dirichlet allocation, in terms of both topic quality and predictive performance.

Barde & Bainwad (2017) conducted an overview of topic modeling methods and tools in their study titled "An Overview of Topic Modeling Methods and Tools" Topic modeling is a powerful technique for analysis of a huge collection of a document. Topic modeling is used for discovering hidden structure from the collection of a document. The topic is viewed as a recurring pattern of co-occurring words. A topic includes a group of words that often occurs together. Topic modeling can link words with the same context and differentiate across the uses of words with different meanings. In this paper, we discuss methods of Topic Modeling which includes Vector Space Model (VSM), Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) with their features and limitations. After that, we will discuss tools available for topic modeling such as Gensim, Standford topic modeling toolbox, MALLET, BigARTM. Then some of the applications of Topic Modeling covered. Topic models have a wide range of applications like tag recommendation, text categorization, keyword extraction, information filtering and similarity search in the fields of text mining, information retrieval.

Hu et al., (2014) conducted a study titled "Interactive Topic Modeling" Topic models are a useful and ubiquitous tool for understanding large corpora. However, topic models are not perfect, and for many users in computational social science, digital humanities, and information studies—who are not machine learning experts—existing models and frameworks are often a “take it or leave it” proposition. This paper presents a mechanism for giving users a voice by encoding users’ feedback to topic models as correlations between words into a topic model. This framework, interactive topic modeling (ITM), allows untrained users to encode their feedback easily and iteratively into the topic models. Because latency in interactive systems is crucial, we develop more efficient inference algorithms for tree-based topic models. We validate the framework both with simulated and real users.

The discussed research papers collectively provide an extensive view of the evolution, diversity, and applications of topic modeling in the field of natural language processing and text analysis. Churchill & Singh (2022) offer a comprehensive analysis of the historical development of topic models, examining their origins, algorithmic variations, and evaluation approaches, setting the stage for the future of topic modeling. Vayansky & Kumar (2020) emphasize the need for a broader selection of topic modeling methods beyond Latent Dirichlet Analysis, introducing various approaches capable of handling complex data relationships, temporal changes in topics, and short text data. Dieng et al. (2020) introduces the embedded topic model (ETM), which enhances traditional topic modeling with word embeddings, improving interpretability even with large vocabularies and outperforming existing models. Barde & Bainwad (2017) provide an overview of popular topic modeling methods and tools, including Vector Space Model, Latent Semantic Indexing, Probabilistic Latent Semantic Analysis, and Latent Dirichlet Allocation, along with practical applications in text mining and information retrieval. Lastly, Hu et al. (2014) introduce interactive topic modeling (ITM), enabling non-expert users to incorporate their feedback into topic models, enhancing user-friendliness and efficiency in topic modeling applications.

Chatbot

The study of Luo et al., (2022) titled “A critical review of state‐of‐the‐art chatbot designs and applications” Chatbots are intelligent conversational agents that can interact with users through natural languages. As chatbots can perform a variety of tasks, many companies have committed numerous resources to develop and deploy chatbots to enhance various business processes. However, we lack an up-to-date critical review that thoroughly examines both state-of-the-art technologies and innovative applications of chatbots. In this review, we not only critically analyze the various computational approaches used to develop state-of-the-art chatbots, but also thoroughly review the usability and applications of chatbots for various business sectors. We also identify gaps in chatbot-related studies and propose new research directions to address the shortcomings of existing studies and applications. Our review advances both academic research and practical business applications of state-of-the-art chatbots. We provide guidance for practitioners to fully realize the business value of chatbots and assist in making sensible decisions related to the development and deployment of chatbots in various business contexts. Researchers interested in the design and development of chatbots can also gain useful insights from our critical review and identify fruitful research topics and future research directions based on the research gaps discussed herein.

In the study of Long (2021) entitled “Development of Intelligent Telegram Chatbot Using Natural Language Processing.” Intelligent chatbots have been gaining interest in the past years due to the advance of artificial intelligence algorithms. As a result, many studies have been conducted on emotional transition and dialog structures. One of the benefits is on the medical applications, in which psychological assessment, clinical counseling, autism diagnostics, and advanced cognitive models could be provided. On the other hand, chatbot knowledge generally comes from a web-based information repository, in which the information is reliable, but it is rather not versatile as it does not contain emotions. The objective of this paper is to develop an intelligent chatbot using natural language processing and Telegram API. First, text processing using Telegram API on Python was developed. Next, emotion recognition was performed on the recorded chats. The appropriate response is then sent to the user. Results showed that our Telegram chatbot could interact smoothly with the users and identify the user's emotions.

This study of “Intelligent Chatbot for Lab Security and Automation” by Prasad et al., (2020) Artificial intelligence based communicative artefacts are called chatbots. The purpose of a chatterbot or chatbot is to render an interaction between a human and a robot in the form of speech or text. They offer the best services in a variety of areas, such as education, healthcare, transportation, etc. As per the research, nearly 85 % of product offerings will be automated by 2020. The proposed system is a voice-based chatbot that helps to improve the security and automation of a lab. The system makes use of Automatic Speaker Recognition (ASR) algorithm in order to recognize a person and allow him/her inside the lab. This allows only authorized person to access the lab facilities. The same algorithm is used for generating the list of available components in the lab based on the person's keyword input thus automating the lab components dispatch.

As outlined by Poongodi et al., (2019) Chat-bot is a computer code that interacts with humans by itself without taking any human assistance. The chat-bot uses natural language processing (NLP) for understanding a human request or query. Blog is an information network where people (group) post information in their area of expertise and people with similar interests read and learn from them. The problem with traditional blog websites is that the user can only get the data which is linked to the home page and also the user have to manually go to every page to get the data. The proposed research work aims in designing and building a chat-bot which, with which the administrators and any common man can interact in English and ask for questions like 'show all the posts by abc', which hence provides a natural language (NL) interface between user's language (English) and the blog's (an information network) database.

In study of Satu et al., (2015) entitled “Review of integrated applications with aiml based chatbot.” Artificial Intelligence Markup Language (AIML) is derived from Extensible Markup Language (XML) which is used to build up conversational agent (chatbot) artificially. There are developed a lot of works to make conversational agent. But low cost, configuration and availability make possible to use it in various applications. In this paper, we give a brief review of some applications which are used AIML chatbot for their conversational service. These applications are related to cultural heritage, e-learning, e-government, web base model, dialog model, semantic analysis framework, interaction framework, humorist expert, network management, adaptive modular architecture as well. In this case, they are not only providing useful services but also interact with customers and give solution of their problems through AIML chatbot instead of human beings. So, this is popular day by day with entrepreneur and users to provide efficient service.

In summary. Luo et al. (2022) offer a comprehensive review of chatbot technologies and their applications across different business sectors, emphasizing the need for bridging research gaps and proposing future directions for both academia and industry. Long (2021) focuses on the development of an intelligent chatbot for emotional recognition, particularly in medical applications, showcasing the potential of chatbots in psychological assessment and counseling. Prasad et al. (2020) present an AI-based chatbot designed for lab security and automation, demonstrating the versatility of chatbots in enhancing security and streamlining lab processes. Poongodi et al. (2019) discuss the use of chatbots in the context of blog networks, enabling users to interact naturally in English and retrieve specific information from blog databases. Satu et al. (2015) provide insights into AIML-based chatbots and their integration into various applications, highlighting the cost-effectiveness and configurability of chatbot technology.

Software Development Model

The study by Umeugo, et al., (2023) investigates the effect of different Software Development Life Cycle (SDLC) models on the perception of Secure SDLC (SSDLC) innovation characteristics and the intention to adopt SSDLC. Software security remains an important issue. Security must be prioritized as a functional requirement to build secure software. Security must also be incorporated in every stage of the SDLC by practicing a secure SDLC (SSDLC). There are various SDLC models, each with emphasized priorities, strengths, and weaknesses. Increasing the security of more published software requires that SMEs, the majority of software publishers, adopt and practice the SSDLC. In promoting the SSDLC, there is a need to know if efforts should be adapted to the various SDLC models. This study empirically examined the effect of SDLC models on the innovation characteristics of the SSDLC derived from the Diffusion of innovation theory and the intention to adopt the SSDLC. A sample of software security managers of software SMEs in the United States was surveyed for the SDLC model used, their perception of the relative advantage, trialability, observability, complexity, and compatibility of the SSDLC, and intention to adopt the SSDLC. A Kruskal-Wallis test performed on the data showed no statistically significant differences between SDLC model groups for relative advantage, compatibility, trialability, observability, complexity, and intention to adopt the SSDLC. Results also indicated that SME Software security managers, on average, would be inclined to adopt the SSDLC if given the impetus. SSDLC adoption efforts can be mostly uniformly applied across the SDLC models. Software security policymakers may find the results of this study useful for SSDLC adoption policy formulation.

Patel (2023) provides an insight into Software Development Life Cycle (SDLC) process models. The paper highlights the software development industry has rapidly grown in recent past years. Let's look at sectors like Healthcare, Tech, Education, E-commerce Transport, Food, beverages, etc. They depend on software developed for them because software is the main part that has grown these industries and businesses, whether in the form of websites, AI-based machines, robots, or apps. The computer software saves time regarding bits, helps solve complex problems, and prolonged and repeated processes fast and accurately without any error. To handle these problems and features, industries need software programs to facilitate the employees' administration work, offices, banks, departments, etc. Developing suitable high-quality software according to the requirements given by the client is the primary goal of software engineering. The software requires a plan with the whole team, and team members cover different parts of the software. The software project manager and his team follow a specific SDLC (Software development Life Cycle) model throughout the project to achieve their goals and complete the project in time according to the given requirements. These SDLC models will be covered in depth in this paper

Kara (2023) focuses on using secure SDLC models to develop secure web applications. Software takes part in center of the digital transformation. Also Covid-19 has been accelerated the digital transformation. Every organization tries to transfer their processes from manuel to electronic environment and this is provided by software. This transformation also enforces corporates to use a secure Software Development Lifecycle (SDLC) methodology or framework. Web applications which are special internetwork application have a great importance in digital transformation due to reasons such as being accessible from anywhere in the world and not requiring a special client application. In this study we investigated that how web application security requirements can be met with SDLC.

Alzaid & Khalfan (2022) examine the involvement of top management in different phases of the System Development Life Cycle (SDLC). One of the most essential factors in the success of system implementation has been recognized as top management support and involvement. Few research, however, have addressed the question of what sort of engagement is necessary through the various stages of the system development life cycle (SDLS). Given the many challenges to top management involvement and support in the various SDLC phases. The objective of this research was twofold. First, to examine the relationship between top management support and the phases of SDLC in order to give guidance for top management practices to ensure the success of information system projects. Second, this study sought to investigate approaches of motivating top management to participate in the SDLC as well as the barriers that hinder them from doing so. This study investigates the function of top management in various phases of system implementation, which will help us in understanding the support mechanism from top management in various SDLC stages. To achieve this goal, the author performed a qualitative study in five different firms in Kuwait, interviewing top management, project management, system analysts, and IT managers. The research established criteria for top management participation and indicated that top management should be involved primarily in the planning and implementation phases, as well as other phases as needed.

Gupta (2021) presents a comparative study of different SDLC models. The Software Development Life Cycle (SDLC) refers to a methodology with clearly defined processes for creating highquality software which are cost effective and reliable. This method of software developing process is quite systematic and structural. SDLC defines the framework that has different activities and tasks to be administered during the software development process. Software development process is quite complex, and to do it without any proper planning would be inefficient. So, we use these SDLC models to make the Software development process simple and systematic. There are various software development life cycle models that are used in the software development process, all having their own advantages and limitations. In this paper, we have included six of these SDLC models - Waterfall Model, Spiral Model, V Model, Agile Model, Iterative Model and Rapid Application Development (RAD) Model.

Umeugo et al. (2023) examine the effect of different SDLC models on the adoption of Secure SDLC (SSDLC) and innovation characteristics, highlighting the uniformity of SSDLC adoption efforts across various SDLC models. Patel (2023) provides an overview of SDLC process models, emphasizing their importance in the rapid growth of the software development industry and their role in facilitating the development of high-quality software. Kara (2023) focuses on the use of secure SDLC models in developing secure web applications, emphasizing the significance of web applications in digital transformation. Alzaid and Khalfan (2022) investigate top management's involvement in different phases of SDLC, underlining the importance of top management support and involvement in ensuring the success of information system projects. Gupta (2021) offers a comparative study of different SDLC models, outlining the advantages and limitations of models such as Waterfall, Spiral, Agile, Iterative, and Rapid Application Development (RAD).

Related Systems

**Table 1**. Features comparison of existing systems.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Researchvault | Research Gate | Google Scholar | Academia Edu |
| Collaborative features | ✓ | ✓ | ✓ | ✓ |
| Research library | ✓ | ✓ | ✓ | ✓ |
| Accessible online | ✓ |  |  |  |
| Download/upload documents | ✓ | ✓ | ✓ | ✓ |
| Semantic search engine | ✓ | ✕ | ✓ | ✕ |
| IMRAD Journal | ✓ | ✕ | ✕ | ✕ |
| Chatbot assistance | ✓ | ✕ | ✕ | ✕ |
| IMRAD converter | ✓ | ✕ | ✕ | ✕ |

ResearchGate is a social networking platform founded in 2008, designed specifically for scientists, researchers, and academics to collaborate and share scientific research. It provides a digital space where researchers can create profiles, share publications, engage in discussions, and connect with peers. Users can upload full-text versions of their papers, join research groups, and follow other researchers to stay updated on the latest developments in their field. ResearchGate has become a popular platform for scientists to disseminate their work, increase their visibility, and foster collaborations, aiming to facilitate knowledge exchange and communication within the scientific community and beyond. However, it has also faced legal challenges and criticism concerning copyright infringement and the sharing of unauthorized full-text articles.

Google Scholar is a web-based search engine provided by Google that focuses specifically on scholarly literature, including articles, theses, books, conference papers, and preprints. Launched in 2004, Google Scholar aims to help researchers and academics find relevant scholarly information from a wide range of disciplines. It indexes content from various sources, including academic publishers, universities, and other websites, making it a comprehensive resource for academic research. Google Scholar provides citation metrics, such as the number of times an article has been cited, and allows users to set up alerts to receive notifications about new publications in their areas of interest. It has become a widely used tool for researchers, students, and professionals seeking scholarly information and conducting literature reviews.

Academia.edu is an online platform that allows researchers, scholars, and academics to share and access scholarly articles, papers, and research materials. Founded in 2008, Academia.edu provides a space for individuals to create profiles, upload their research work, and connect with others in their respective fields. Users can follow specific researchers or topics of interest, enabling them to receive updates on new publications and research activities. The platform also offers features for networking, discussion forums, and the ability to track citations and analytics of uploaded papers. Academia.edu aims to foster collaboration and facilitate the dissemination of knowledge within the academic community, providing a platform for researchers to showcase their work and connect with a wider audience. However, it's important to note that some content on Academia.edu may require a subscription or be subject to copyright restrictions.

Synthesis

As discussed by Long (2022) and Mohan et al. (2023), both delve into natural language processing (NLP) but focus on different aspects. Long (2022) discusses grammatical error correction, emphasizing the importance of accurate text analysis for NLP. It addresses nonword mistakes, spelling errors, and the training of word repair models. In contrast, Mohan et al. (2023) concentrates on sarcasm detection using advanced techniques like Bidirectional Encoder Representations from Transformers (BERT) and Graph Convolutional Networks (GCN). While both studies involve NLP, Long's work is centered on error correction, while Mohan's paper explores the complexities of identifying sarcasm in text, showcasing the versatility of NLP.

Similarly, Abdusalomovna & T. D. (2023) and Jaiswal & Samikannu (2017) focus on data analysis but with different applications and methods. Abdusalomovna & T. D. (2023) discuss text mining as a broad approach that utilizes natural language processing and statistics to convert unstructured text into structured data. They emphasize its applications in various domains, including missing data imputation. In contrast, Jaiswal & Samikannu (2017) concentrate on the random forest algorithm, highlighting its role in feature selection and data analysis. They discuss its robustness, especially in handling noisy data. While both papers address data analysis, Abdusalomovna & T. D. explore the broader field of text mining, while Jaiswal & Samikannu delve deeply into a specific algorithm's capabilities.

In a related vein, Churchill & Singh (2022) and Luo et al. (2022) explore the applications of computational techniques but in different contexts. Churchill & Singh (2022) provide an analysis of the evolution of topic modeling, highlighting its applications in various domains such as business and academia. They emphasize the importance of topic modeling for understanding prevalent themes in text corpora. On the other hand, Luo et al. (2022) conduct a critical review of chatbot designs and their practical applications in business sectors. While both papers involve computational techniques, Churchill & Singh focus on topic modeling's historical development, whereas Luo et al. delve into chatbot technology and its usability in business contexts.

Furthermore, LeCun et al. (2015) and Möller (2023) introduce fundamental concepts in artificial intelligence and its subsets. LeCun et al. (2015) discuss deep learning's contributions to various applications, including speech and image recognition. They emphasize deep neural networks' ability to capture complex patterns and features. In contrast, Möller (2023) introduces machine learning as a subset of artificial intelligence, emphasizing its role in pattern recognition and decision-making. While both papers provide foundational knowledge, LeCun et al. focus on deep learning's advancements, while Möller introduces broader concepts in machine learning.

Additionally, Mohan et al. (2023) and Churchill & Singh (2022) explore different areas within the field of natural language processing (NLP). Mohan et al. (2023) focus on sarcasm detection, utilizing advanced techniques like Bidirectional Encoder Representations from Transformers (BERT) and Graph Convolutional Networks (GCN). They discuss the challenges of identifying sarcasm in text and compare their method to baseline approaches. In contrast, Churchill & Singh (2022) provide an overview of the evolution of topic modeling, highlighting its applications in diverse domains. They emphasize the importance of topic modeling for understanding prevalent themes in text corpora. While both papers involve NLP, Mohan et al. delve into sentiment analysis, while Churchill & Singh focus on topic modeling's historical development.

Moreover, Abdusalomovna & T. D. (2023) and LeCun et al. (2015) touch on different aspects of artificial intelligence and its applications. Abdusalomovna & T. D. (2023) discuss text mining and its versatility in handling unstructured text data. They emphasize the conversion of unstructured text into structured data and its applications in various domains, including missing data imputation. In contrast, LeCun et al. (2015) delve into deep learning's contributions to applications such as speech and image recognition. They discuss the advancements in deep neural networks and their ability to capture complex patterns. While both papers relate to artificial intelligence, Abdusalomovna & T. D. focus on text mining, whereas LeCun et al. emphasize deep learning's applications.

Lastly, Jaiswal & Samikannu (2017) and Luo et al. (2022) explore computational techniques and their applications in data analysis, but within different contexts. Jaiswal & Samikannu (2017) discuss the random forest algorithm's applications, particularly in feature selection and data analysis. They emphasize its robustness, especially in handling noisy data. In contrast, Luo et al. (2022) conduct a critical review of chatbot designs and their practical applications in business sectors. While both papers involve data analysis, Jaiswal & Samikannu focus on a specific algorithm's capabilities, whereas Luo et al. delve into chatbot technology and its usability in business contexts.

With all these studies in consideration, a comprehensive overview of the evolving landscape of computational techniques in various domains emerges. Long (2022) and Mohan et al. (2023) delve into natural language processing (NLP), where Long emphasizes grammatical error correction and Mohan focuses on sarcasm detection. Abdusalomovna & T. D. (2023) and Jaiswal & Samikannu (2017) tackle data analysis, with the former encompassing text mining and the latter focusing on the robustness of the random forest algorithm. Churchill & Singh (2022) and Luo et al. (2022) explore computational techniques, with Churchill & Singh tracing the evolution of topic modeling and Luo et al. critically reviewing chatbot designs. LeCun et al. (2015) and Möller (2023) introduce fundamental concepts in artificial intelligence, with LeCun emphasizing deep learning's applications and Möller providing a broader understanding of machine learning. Additionally, Mohan et al. (2023) and Churchill & Singh (2022) delve into different aspects of NLP, with Mohan focusing on sentiment analysis and Churchill & Singh highlighting topic modeling's historical development. Abdusalomovna & T. D. (2023) and LeCun et al. (2015) touch on artificial intelligence, with Abdusalomovna & T. D. discussing text mining's versatility and LeCun et al. emphasizing deep learning's applications. Jaiswal & Samikannu (2017) and Luo et al. (2022) explore computational techniques in data analysis, with Jaiswal & Samikannu focusing on the random forest algorithm's capabilities and Luo et al. examining chatbot technology's usability in business contexts.

The reviewed studies collectively span a wide range of topics in the fields of Natural Language Processing (NLP), sentiment analysis, text mining, deep learning, machine learning, Random Forest, topic modeling, chatbot technology, and Software Development Life Cycle (SDLC) models. These studies showcase the multifaceted nature of NLP, with a focus on practical applications, data interpretation, and collaborative platforms, while also addressing sentiment analysis challenges and emphasizing the role of tailored term weighting. Deep learning and machine learning are explored extensively, demonstrating their transformative impact on various domains, including cybersecurity. The studies on Random Forest highlight its evolution, optimization, and application in intrusion detection. In the context of topic modeling, various approaches and tools are discussed to handle complex data relationships. Additionally, chatbot technologies and their applications across sectors are examined, and SDLC models are evaluated for their impact on secure software development.

CHAPTER III

RESEARCH METHODOLOGY

This chapter outlines the methodology employed to achieve the objectives of the study. It presents the research design, the local of the studies, as well as the data collection methods, data model generation, development methodology, performance evaluation of algorithms, software used, system architecture and software testing.

Research Design

The research design of "ResearchVault: An Advanced Web-based Research Repository for Laguna State Polytechnic University - Sta. Cruz Campus with Intelligent Chatbot Integration and Collaborative Features for Efficient Research Management" will utilize both experimental and developmental research design methods. As stated by Aynyemi (2023) experimental research is a scientific approach to research, where one or more independent variables are manipulated and applied to one or more dependent variables to measure their effect on the latter. The effect of the independent variables on the dependent variables is usually observed and recorded over some time, to aid researchers in drawing a reasonable conclusion regarding the relationship between these 2 variable types. This method was utilized in order to determine the best-performing trained natural language processing (NLP) algorithm model in terms of extracting information with respect to the datasets collected.

On the other hand, the developmental research, as defined by Richey (1994) as the systematic study of designing, developing, and evaluating instructional programs, processes, and products that must meet criteria of internal consistency and effectiveness. It involves a situation in which the product-development process is analyzed and described, and the final product is evaluated. Therefore, the researchers used this method to examine the consistency and efficacy of the development of the system.

**Locale of the Study**

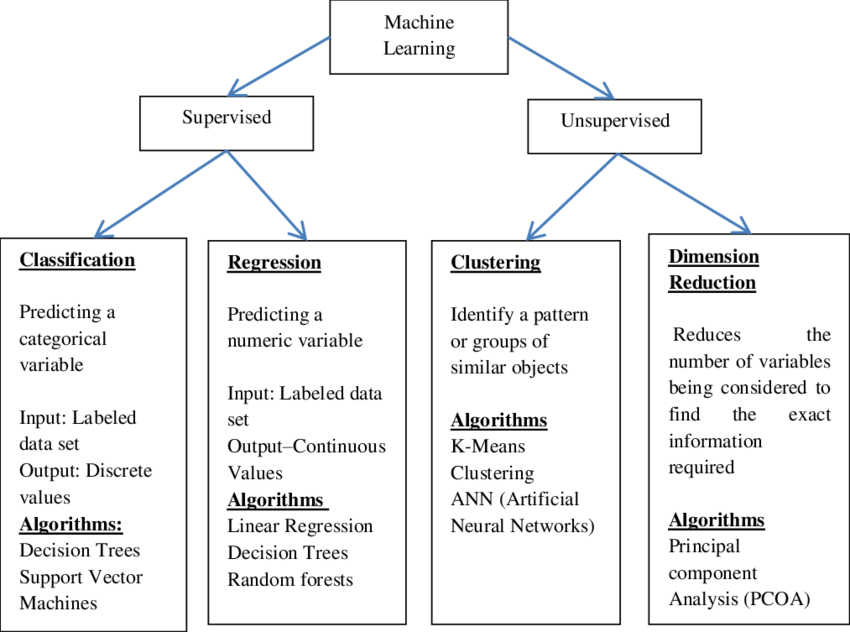
This study was conducted at the Laguna State Polytechnic University - Sta. Cruz Campus (LSPU-SCC), situated in Santa Cruz, Laguna. The research primarily involves the RIUH and students of LSPU-SCC, who are the intended users of the web-based research repository system being developed. Conducting the research within this academic setting allows for a unique opportunity to integrate the system into the institution's research workflow and engage with the university community for real-world testing and feedback.

**Applied Concepts and Techniques**

This section states the applied concepts and techniques to be used in solving the research problem and achieving the research objectives. This indicates theoretical disciplines computer science such as follow:

***Machine Learning Techniques***

Chopra & Khurana (2023) Machine learning is a subfield of artificial intelligence, broadly defined as a machine's capability to imitate intelligent human behavior. Like humans, machines become capable of making intelligent decisions by learning from their past experiences. Machine learning is being employed in many applications, including fraud detection and prevention, self-driving cars, recommendation systems, facial recognition technology, and intelligent computing.

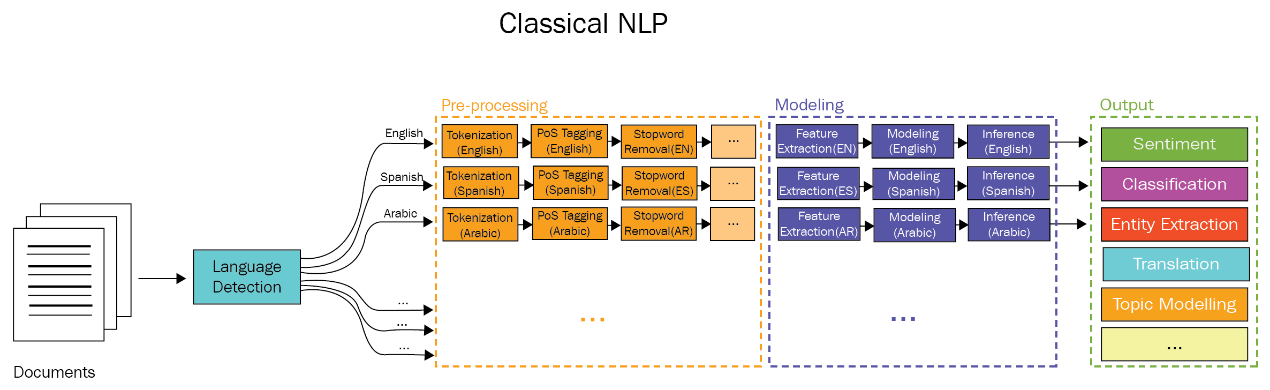


**Figure 3:** Classification of machine learning techniques. Source: Suryakanthi (2020)

Machine Learning enable computers with the ability of pattern recognition, or to continuously learn from, making predictions based on data, and carry out decisions without being specifically programmed for doing so, as stated by Möller (2023). The researchers in this study would use ML to automatically classify sections within a PDF document and extract their content. The predictions determine how the text is segmented into sections, facilitating the conversion of the PDF into IMRAD format.

***Natural Language Processing Techniques***

Ali (2022) Natural Language Processing (NLP) could be a branch of Artificial Intelligence (AI) that allows machines to know the human language. Its goal is to form systems that can make sense of text and automatically perform tasks like translation, spell check, or topic classification. Natural language processing (NLP) has recently gained much attention for representing and analysing human language computationally. It's spread its applications in various fields like computational linguistics, email spam detection, information extraction, summarization, medical, and question answering etc.

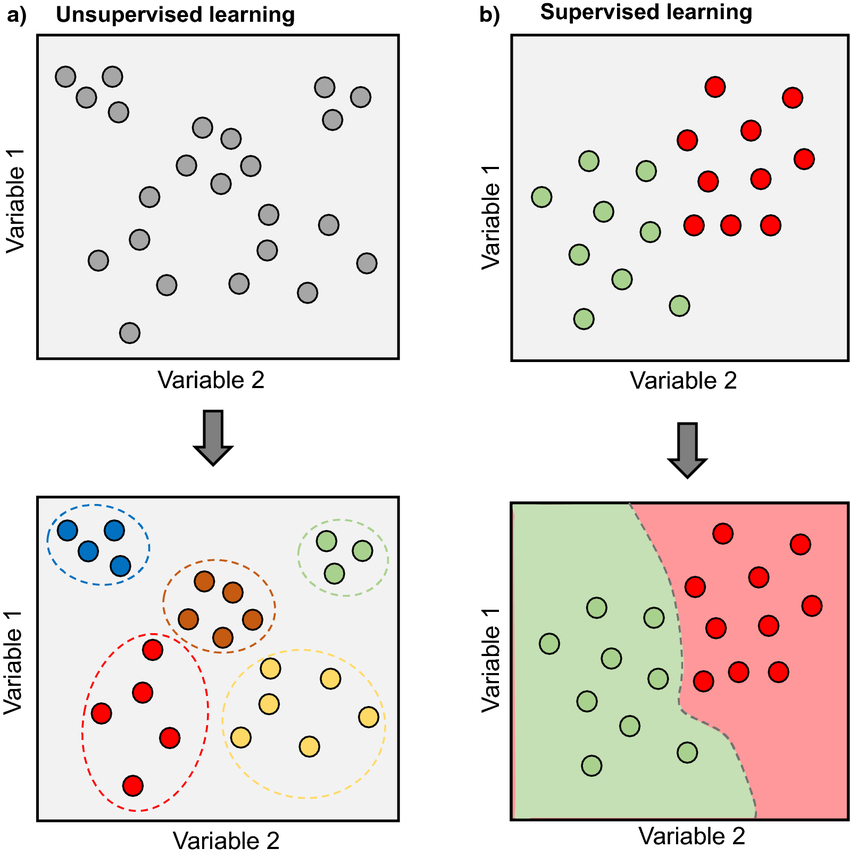


**Figure 4:** Classical NLP approach. Source: Oreily (2023)

Natural Language Processing enables the textual data to be converted into the desired language in the digital environment, the logical answers to the questions asked, the synthesis and summary of digital texts, and the guidance of the machines with voice commands. The natural language processing process begins by moving words, which are the smallest building blocks that make sense in the text, into the digital world, as explained by Guler & Akgul (2022). NLP was used by the researchers to handle various aspects of text processing and classification, making it possible to automatically categorize and structure the content of a PDF document based on its sections.

***Unsupervised Learning***

Akanksha\_Rai et al., (2023) Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically, supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

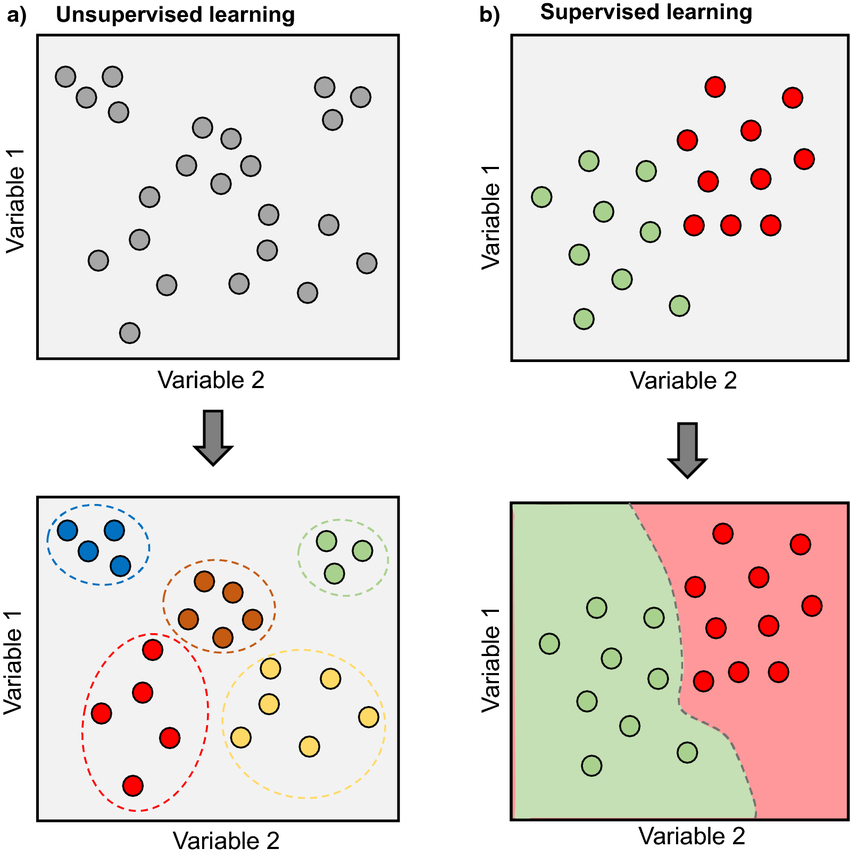


**Figure 5:** Unsupervised learning. Source: Morimoto & Ponton (2021)

Unsupervised Learning Algorithms allow users to perform more complex processing tasks compared to supervised learning. Although, unsupervised learning can be more unpredictable compared with other natural learning methods. Unsupervised learning algorithms include clustering, anomaly detection, and neural networks as discussed by Johnson (2023) The researchers employed unsupervised learning, specifically Latent Dirichlet Allocation (LDA), to discover hidden topics within a collection of text documents during the data modeling process

***Supervised Learning***

Cunningham et al., (2008) Supervised learning accounts for a lot of research activity in machine learning and many supervised learning techniques have found application in the processing of multimedia content. The defining characteristic of supervised learning is the availability of annotated training data. The name invokes the idea of a ‘supervisor’ that instructs the learning system on the labels to associate with training examples. Typically, these labels are class labels in classification problems. Supervised learning algorithms induce models from these training data and these models can be used to classify other unlabeled data.

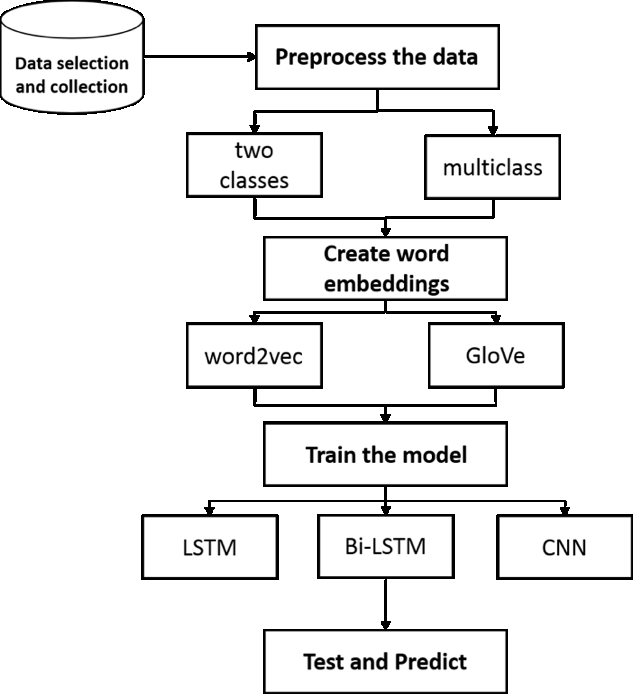


**Figure 6:** Supervised learning. Source: Morimoto & Ponton (2021)

Supervised learning is also called classification or inductive learning in machine learning. This type of learning is analogous to human learning from past experiences to gain new knowledge in order to improve our ability to perform real-world tasks as stated by Liu et al., (2011) The researchers employed Supervised learning, for training a machine learning model to predict labels based on input text data.

***Text Classification***

Cai et al., (2018) Text classification is one of the most widely used natural language processing technologies. Common text classification applications include spam identification, news text classification, information retrieval, emotion analysis, and intention judgment, etc. Traditional text classifiers based on machine learning methods have defects such as data sparsity, dimension explosion and poor generalization ability, while classifiers based on deep learning network greatly improve these defects, avoid cumbersome feature extraction process, and have strong learning ability and higher prediction accuracy.



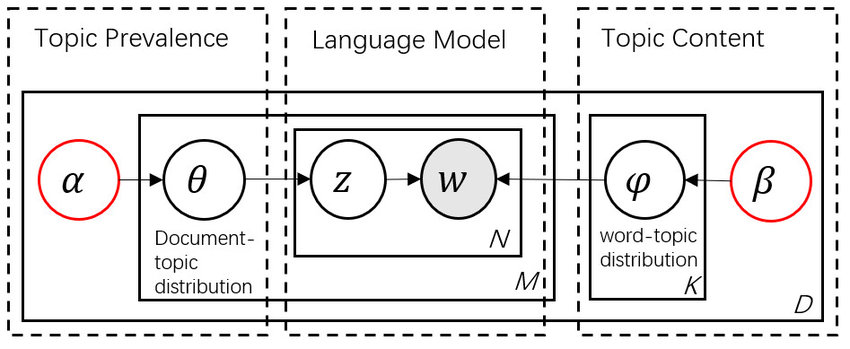
**Figure 7:** Text Classification. Source: Alsurayyi et al., (2019)

Text classification is used to organize documents in a predefined set of classes. It is very useful in Web content management, search engines; email filtering, etc. Text classification is a difficult task due to high- dimensional feature vector comprising noisy and irrelevant features. Various feature reduction methods have been proposed for eliminating irrelevant features as well as for reducing the dimension of feature vector. Relevant and reduced feature vector is used by machine learning model for better classification results as explored by Agarwal & Mittal (2014) The researchers used text classification to categorize text documents into four IMRAD sections: Introduction, Method, Result, and Discussion.

Algorithm Analysis

This section on algorithm analysis presents the methodologies and techniques used by the researchers in applying machine learning to the development of the system. It enumerates the LDA topic modeling, the TF-IDF term weighting scheme, the random forest classifier, and the Bidirectional Encoder Representation (BERT), which were all used to achieve the objectives of the study. It also includes an overview of the different machine learning algorithms, their architectures, advantages, and disadvantages, to provide a better understanding of their general behavior and how they were used in the development of the system.

*LDA Topic Modeling*

Introduction, Methodology, Results, and Discussion. LDA will be a vital component of our data preprocessing pipeline, aiding in the extraction of relevant topics from the text. By employing LDA, we can identify and categorize the thematic content of each section, helping us distinguish between introductory content, methodological descriptions, result discussions, and analytical interpretations

**Figure 8:** LDA plate diagram. Source: He et al., (2020).

The advantages of LDA include being a probabilistic model with interpretable topics and not needing to know the topics in advance. However, its disadvantages include the inability to represent the relationship among topics and the requirement to specify the number of topics in advance. Therefore, it is important to consider the specific needs of the research when selecting a topic modeling approach.

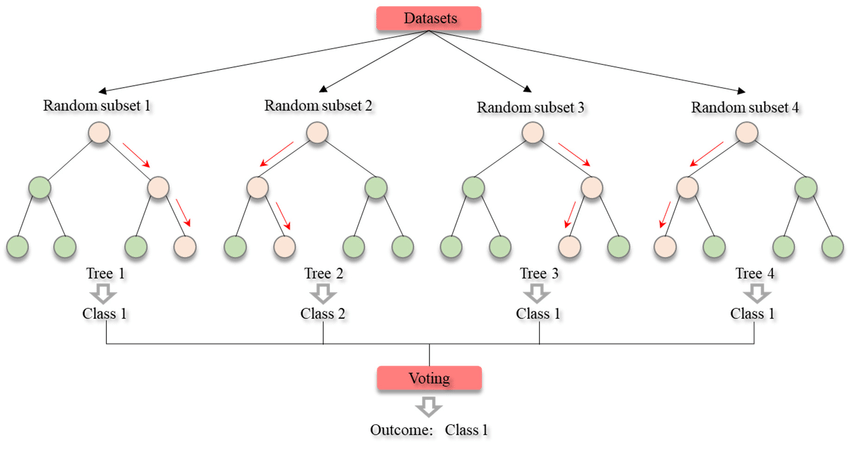
*TF-IDF Term Weighting Scheme*

**Equation 1:** Calculation process of the TF-IDF term weighting scheme. Source: Sabbah et al., (2015).

TF-IDF (term frequency-inverse document frequency) is a widely recognized term weighting scheme employed for assessing term relevance within a document corpus. This method computes a weight for each term that signifies its significance within a document relative to the entire corpus. The TF aspect is derived from a term's frequency within a document divided by the document's total term count, while the IDF component considers the inverse of a term's frequency across all documents in the corpus. These two components are then multiplied to yield the term's final weight. Researchers have explored alternative term weighting schemes to enhance information retrieval and text classification systems. For example, Paik's (2013) proposed TF-IDF variant excels at capturing diverse aspects of term importance, making it effective for both short and long queries. Additionally, Chen et al.'s (2016) TF-IGM scheme, incorporating a novel statistical model, has demonstrated superior performance compared to traditional TF-IDF and other supervised term weighting schemes in text classification. These studies provide valuable insights for enhancing the design and development of information retrieval and text classification systems.

IMRAD section classification often involves distinguishing between these distinct sections to enable efficient information extraction and analysis. When applied to this task, TF-IDF's ability to weigh the importance of terms within a document corpus becomes particularly valuable. For instance, in the "Introduction" section, certain terms and phrases may be more indicative of background information, while the "Methods" section might contain specialized technical terminology. By utilizing TF-IDF, we can assign appropriate weights to terms, allowing the model to discern the unique linguistic characteristics of each IMRAD section. This not only improves the accuracy of classification but also enhances the model's ability to extract relevant information for subsequent analysis, thereby advancing the automation of document processing in scientific research contexts.

*Random Forest*

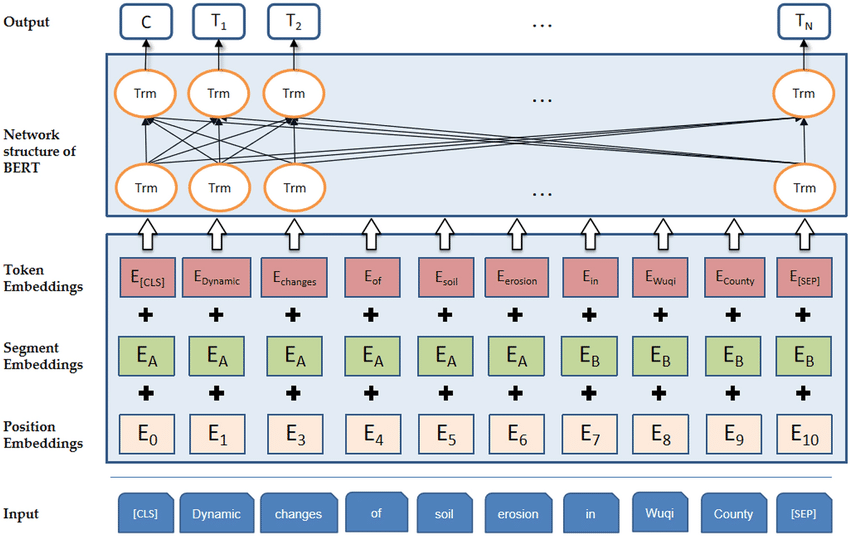


**Figure 9:** Random Forest algorithm. Source: Shih et al., (2019).

Random Forest is a widely employed ensemble learning method for classification and regression, building numerous decision trees during training and deriving the final prediction through class mode (classification) or mean prediction (regression) of these individual trees. It extends the decision tree algorithm by randomly selecting feature subsets at each split, then determining the best split among those features for each tree in the forest. This approach not only accommodates large, high-dimensional datasets but also handles missing data while exhibiting robustness to noise and outliers. Nevertheless, Random Forest can be computationally demanding, especially with sizable datasets. In the realm of information retrieval and text classification, alternative term weighting schemes have been explored. For instance, Paik (2013) introduced a TF-IDF variant adept at capturing diverse aspects of term saliency, catering to both short and long queries. Additionally, Chen et al. (2016) developed a TF-IGM scheme incorporating a novel statistical model, which has demonstrated superior performance in text classification compared to traditional TF-IDF and other supervised term weighting schemes.

IMRAD section classification is a complex task that demands the ability to discern distinct linguistic patterns and content structures within each section. Random Forest's ability to construct multiple decision trees and combine their outputs makes it particularly well-suited for this task. Each IMRAD section can exhibit distinct linguistic patterns and vocabulary, and the Random Forest algorithm excels at capturing these nuances. By randomly selecting a subset of features (such as specific terms or phrases) at each decision point within the trees, the algorithm can effectively discern the unique characteristics of each section. For example, the "Methods" section might frequently include technical terms or experimental details that distinguish it from the more general language found in the "Introduction" or the analytical results presented in the "Results" section. By leveraging Random Forest, the model can efficiently learn and differentiate between these linguistic cues, leading to accurate and reliable classification results for each IMRAD section, ultimately enhancing the automation and efficiency of scientific document analysis and organization.

*Bidirectional Encoder Representation*



**Figure 10:** The overall structure of the BERT model. Source: Sun et al., (2022).

BERT (Bidirectional Encoder Representations from Transformers) is a powerful model that has greatly advanced Natural Language Processing (NLP) and text classification. Its distinctive feature is its capability to understand word context from both the left and right sides during training. Text classification using BERT involves two key phases: pre-training, where the model learns language representations from extensive text corpora, and fine-tuning, where BERT is adapted to specific tasks with an additional output layer trained on labeled data. In a study published in IEEE Transactions on Geoscience and Remote Sensing by He et al. (2020), HSI-BERT (Hyperspectral Image-Bidirectional Encoder Representations from Transformers) was introduced for hyperspectral image classification. This model, leveraging multi head self-attention in an MHSA layer, possesses a global receptive field for capturing pixel dependencies regardless of spatial distance, accommodating dynamic input regions, and demonstrating excellent generalization across regions with different shapes without the need for retraining. While BERT excels at capturing intricate language nuances, it demands substantial computational resources and extensive data.

The use of BERT (Bidirectional Encoder Representations from Transformers) represents a significant advancement in the field of Natural Language Processing (NLP) that can greatly enhance the accuracy and precision of classification tasks. BERT's distinguishing feature is its unique bidirectional context understanding during training, enabling it to capture complex linguistic nuances and the intricate relationships between words. For IMRAD section classification, this means that the model can effectively discern not only specific keywords but also the broader context in which they appear, leading to a more robust and context-aware classification process. By leveraging BERT during pre-tuning, the model can acquire a deep understanding of the linguistic patterns and structural characteristics that define each IMRAD section, making it highly adaptable and capable of handling a wide range of scientific documents with varying writing styles and content structures. This enhanced contextual understanding empowers the model to make more accurate and nuanced predictions, ultimately improving the automation and efficiency of scientific document analysis and organization.

Data Collection Methods

The researchers used data collection methods, such as interviews, and online research for the collection of data and information which were critical for the study. The researchers conducted short interviews with the Library Services Head Director of the College of Computer Studies (CCS) to gain access to research materials. The collected manuscripts served as the primary source for training the machine learning models used in the system.

**Table 2:** Datasets of manuscript and IMRAD from interview.

|  |  |  |  |
| --- | --- | --- | --- |
| **Program** | **Manuscripts** | **10 Pagers** | **Years** |
| **Bachelor of Science in Information Technology** | 13 | 13 | 2016 |
| 43 | 43 | 2017 |
| 33 | 33 | 2018 |
| 23 | 23 | 2021 |
| **Bachelor of Science in Computer Science** | 5 | 5 | 2017 |
| 2 | 2 | 2022 |
| **TOTAL** | 119 | 119 |  |

**Table 3:** Document titles and document pages of datasets.

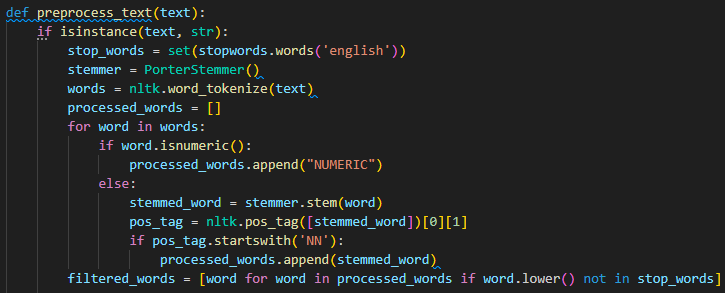
|  |  |
| --- | --- |
| **Document Titles** | **Number of Pages** |
| **E-Health: Victoria Health Center Online Information System Using SMS Notification** | 90 |
| **Thermoelectric BBQ Grill** | 56 |
| **CCS InfoCast: An Information Dissemination Management System using SMS Blast** | 55 |
| **CCS Online Grading System** | 51 |
| **CCS Online Thesis Management System** | 54 |
| **eFort: An Electronic Faculty Portfolio Management System** | 42 |
| **FEMS: A Lan-Based Students Evaluation on Faculty Performance** | 58 |
| **Document Titles** | **Number of Pages** |
| **College of Computer Studies Web-based Announcement System Via SMS** | 67 |
| **Jesus the Saviour Hospital Management Information System** | 55 |
| **LSPU SCC Payment Transaction Management System** | 113 |
| **OMS: Office Management Solutions for College of Computer Studies** | 41 |
| **Premiumbikes Lending Management System** | 43 |
| **Syllabus Generating System A Syllabus Generator for the College of Computer Studies** | 37 |
| **SMS: A Web-Based CCS Syllabus Management System for the College of Computer Studies** | 43 |
| **Virus Attack: The Development of a Mobile Game** | 43 |
| **AFScan: Android File Scanner and Translator with Optical Character Recognition** | 77 |
| **Anonymous Restriction Application Using AES Algorithm** | 80 |
| **Crime Analysis in 4th District of Laguna using JRip Algorithm** | 96 |
| **Smart Exam Checker Using Speed Up Robust Feature (SURF) Algorithm** | 63 |
| **Executor: Massive Open World Role-Playing Game** | 101 |
| **Computer Aided Instruction for Hotel Reservation Management System** | 98 |
| **Tiwi Food Product Company Web-based Human Resource Integrated System** | 89 |
| **A Salon Proprietor’s Smart Assistant System** | 56 |
| **All-in-one Kiosk** | 64 |
| **ASADO: Automated Student File and Document Organizer** | 00 |
| **Document Titles** | **Number of Pages** |
| **Augmented Element: An Interactive Augmented Reality Game of Elements** | 84 |
| **Automated Canteen Transaction and Gate Pass Monitoring Using RFID of St. Mary’s Montessori** | 59 |
| **CCS Research LAB** | 95 |
| **CCS-eBASS “College of Computer Studies – Electronic Bulleting Announcement System Using SMS”** | 79 |
| **Checkmate: A Mobile Scanning and Score Generating Application** | 77 |
| **College of Computer Studies File Submission Kiosk** | 77 |
| **COP: CCS OJT Portal** | 76 |
| **CS Online: An Online Clerical Service System for the Guidance Office** | 72 |
| **EatSmart: A Smart Dining System** | 74 |
| **E-Bins Online Barangay Nutrition Scholar** | 74 |
| **Faculty Availabilit and Monitoring Sstem Using SMS Technology** | 48 |
| **Feliciano Dental Clinic Management System with SMS Notification** | 67 |
| **HR Online: An Online Human Resource Management Information System of Laguna State Polytechnic University Santa Cruz Campus** | 115 |
| **IKNo'SS: Integrated Knowledge Supervision System** | 120 |
| **Information Assimilation Learning Tool for Lumban National High School** | 102 |
| **ITAid: A LAN Based Courseware for College of Computer Studies** | 69 |
| **LABVIEW: LAN-Based LSPU-SCC CCS Laboratory Attendance Monitoring System using Finger-Print Biometric Technology** | 127 |
| **Laguna Shop: An E-commerce Platform** | 66 |
| **Document Titles** | **Number of Pages** |
| **LDH-InfoKiosk: Laguna Doctors Hospital Information Kiosk** | 104 |
| **LSPU-ARCSS: LSPU Social Media with Thesis Archive** | 94 |
| **LSPU SCC Asset Management System** | 106 |
| **LSPU Sports Equipment Monitoring System Using Biometrics and SMS Notification** | 63 |
| **LSPU Web-Based ISO Record Management System** | 58 |
| **Laguna State Polytechnic University Sta. Cruz, Campus Safety and Security Management System Using Barcode and SMS Technology** | 69 |
| **Mathrix Adventure: A Mathematical Adventure Game Using Virtual Reality Application** | 66 |
| **OSYAID: An Online Learning Management System for Out of School Youth** | 81 |
| **A Coin-Operated Photo Booth** | 54 |
| **Santa Cruz Water District Customer Relationship Management System** | 87 |
| **School Event Attendance Monitoring System Using Fingerprint and SMS Notification** | 86 |
| **SMART Office: Office Automation and Security Monitoring System** | 85 |
| **SMP (Service Management Program) E-Learning System** | 72 |
| **Students Attendance and Violation Monitoring System Using RFID with SMS Notification** | 156 |
| **TESDA School Division Payroll and Management System** | 63 |
| **THEREPO: A Thesis Repository Android Based Application** | 90 |
| **Document Titles** | **Number of Pages** |
| **Class Scheduler: An Expert System Timetabling for Faculty Members Using Genetic Algorithm** | 89 |
| **Eye Guide: An Obstacle Detection and Avoidance for Blind Navigation** | 72 |
| **The Chemistry of Solution: Chemistry Courseware with Virtual Reality Laboratory Simulation** | 77 |
| **Leafy: An Android Application for Leaves Identification Using Tensor Flow** | 146 |
| **Fate Flames of Revolution** | 51 |
| **CALEMS: CCS Advance Laboratory Electric Management System** | 116 |
| **3D INFODELO: LSPU Augmented Reality 3D Information and Locator** | 64 |
| **Advance Learning for CCS SCC Using Augmented Reality and Moral Support** | 52 |
| **Disasterville - Android Game** | 88 |
| **E-DOC: LSPU Document Tracking and Scheduling System with SMS Notification** | 93 |
| **E-Sked: Online Scheduling Management System of Provincial Population Office with SMS Notification** | 117 |
| **Laguna State Polytechnic University Extension and Training Services Document Management System** | 123 |
| **The Bible Warrior** | 51 |
| **PDRRMO Information Management System** | 91 |
| **Nexus point: A College of Computer Studies Web Portal** | 69 |
| **ASADO: Automated Student File and Document Organizer** | 90 |
| **UpMART: UPLB Museum of Natural History Tour using Augmented Reality** | 68 |
| **The Patrol Methods in Pila, Laguna: Input to Crime Prevention** | 65 |
| **Document Titles** | **Number of Pages** |
| **RFVMS: Radio Frequency Vehicle Monitoring System; A Modernization Project of SMO LSPU-SCC** | 102 |
| **Cross Platform Unified President Report** | 78 |
| **e-GYNE: Mobile Application for Women's Health** | 137 |
| **e-Steno: A Mobile Stenography Courseware with Scanning Capability** | 72 |
| **FAMS:"Web-Based Farm Monitoring System with 3D Mapping for Brgy. Sta Clara Sur, Pila, Laguna** | 73 |
| **Human AR: Augmented Human Internal Organs** | 84 |
| **TILA: T.L.E. Interactive Learning Aid** | 67 |
| **SaveLIFE: Connecting Blood Donors Mobile Application** | 85 |
| **PINHS Interactive Online Courseware for Araling Panlipunan with Virtual Assitance** | 97 |
| **Online Courseware of DMRMNHS in Filipino Subject** | 66 |
| **LSPU Web Based Research and Development Management System** | 99 |
| **LSPU System Scholarship and Financial Assistance Record Management System with SMS Announcement and Notification** | 79 |
| **L-SMS: LSPU-SCC Syllabus Management System** | 65 |
| **Laguna GENSERV Cross-Platform Web and Mobile Application** | 74 |
| **Jobs Management and Dissemination System Using SMS Technology for Public Employment Service Office (PESO) At Santa Cruz Laguna** | 96 |
| **BAT (Budget Office, Accounting Office, Treasury Office) Monitoring Record Management System of Municipality of Sta. Cruz, Laguna** | 86 |
| **Document Titles** | **Number of Pages** |
| **Crime and Incident Mapping of Bay, Municipal Police Station** | 78 |
| **E-Mayor: Scheduling Appointment System** | 83 |
| **E-Report Mo: An Online Crime Reporting System for Santa Cruz Municipal Police Station** | 74 |
| **Instructional Materials (IMs) Submission and Monitoring System** | 72 |
| **Integration of Cooperatives Under PCDO Loaning Management System** | 84 |
| **iSeekLawyer: Online Meets Law; Justice on the Go!** | 102 |
| **LSPU – SCC Physical Plant and Site Development Scheduling Management System** | 76 |
| **LSPU Students: OJT Web Portal** | 63 |
| **Modernization Program for Kingdom Plantae: Identification and Proper Care with the Use of Mobile Application (South East Asian Plant)** | 79 |
| **Municipal Tricycle Franchise Regulatory Board and Business Permit Record Management System (MTFRBBPRMS)** | 73 |
| **Municipal of Magdalena Billing and Collection with Meter Reading Application and SMS Notification** | 82 |
| **Municipality of Santa Cruz Public Cemetery Record Management System with SMS Notification** | 78 |
| **Municipality of Sta. Cruz Scholarship Program Validation with SMS Notification** | 87 |
| **NSTP Web Portal Student Information Management System with Auto Generated QR Coded** | 58 |
| **Online Management Information System for Girl Scouts of the Philippines –Laguna Council** | 81 |
| **Peso Laguna Integrated Website with Data Mapping** | 72 |
| **Document Titles** | **Number of Pages** |
| **Plastic in Exchange for Rice Using Kiosk: A Technological Way to Addressed Waste Management and Hunger** | 106 |
| **Rainfall Advisory System with SMS Notification for Municipal Disaster Risk Reduction and Management Office Santa Cruz, Laguna** | 113 |
| **Sustainable Livelihood Program Management Information System with SMS Notification** | 96 |
| **The Development of MDRRMO Report Management Information System** | 80 |
| **The Development of Municipal Planning and Development Coordinator Office (MPDCO)** | 94 |
| **Web Portal: Municipal Youth and Sports Development Office of Sta. Cruz Laguna** | 87 |

Data Model Generation

This section would enumerate the various methodologies employed in developing the algorithm models. It would include data preprocessing, tokenization, feature extraction, topic modeling and training the transfer learning-based models.

*D****ata Preprocessing***

Data preprocessing is a crucial step in text mining tasks that involves transforming raw text data into a more manageable form for machine learning algorithms to process. This is achieved through various techniques such as converting all characters to lowercase, removing stop words and punctuations, tokenization, stemming, lemmatization, and vectorization. These techniques help to eliminate noise or unhelpful parts of the data, making it easier for machine learning algorithms to extract meaningful insights from the text.



**Figure 11:** Data preprocessing code in LDA model generation.

The figure above illustrates a Python function used for text data preprocessing. It begins by verifying whether the input is a valid text string. Upon confirmation, the function proceeds with several fundamental preprocessing steps. It initializes a set of common English stopwords, which are words typically devoid of substantial meaning in textual analysis. The function then employs the Porter stemming algorithm to reduce words to their base form, effectively eliminating suffixes and converting them to their root form. While performing this operation, it identifies and retains words classified as nouns (NN) using part-of-speech tagging. Numeric values are replaced with a standard 'NUMERIC' token to maintain consistency across the text. Subsequently, the function filters out words found in the stopwords set, ensuring that the resulting text comprises only meaningful, stemmed words.

*Tokenization*

Tokenization is the process of breaking down text into individual tokens or words. This is an important step in natural language processing and is used to prepare text for further analysis. Tokenization can be done using various techniques such as whitespace tokenization, rule-based tokenization, and statistical tokenization.

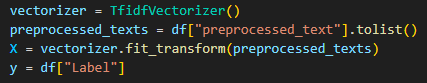


**Figure 12:** Tokenizing text chunk

The figure above represents a critical stage in natural language processing known as tokenization. Within this process, the input text is divided into smaller units, referred to as tokens, to facilitate subsequent analysis and understanding. For each 'text\_chunk' in a collection of such chunks, the code employs a tokenizer to segment the text into tokens. It utilizes the 'return\_tensors' parameter to return these tokens in a format compatible with PyTorch, a popular deep learning library. Additionally, the 'padding' and 'truncation' parameters are used to ensure that the tokens are uniformly sized, enabling efficient processing by machine learning models. Tokenization is a fundamental preprocessing step in text analysis, enabling the transformation of unstructured text data into a structured format that can be readily consumed by various natural language processing tasks such as text classification, sentiment analysis, and machine translation.

*Feature Extraction*

Feature extraction is the process of extracting relevant features from raw data. In natural language processing, this can involve techniques such as bag-of-words, term frequency-inverse document frequency (TF-IDF), and word embeddings. Feature extraction is an important step in preparing text data for machine learning algorithms.

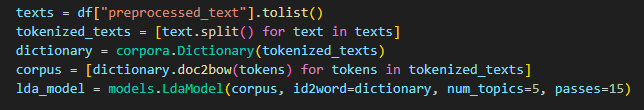


**Figure 13:** Feature extraction using TF-IDF.

The figure above demonstrates the essential step of text vectorization using Term Frequency-Inverse Document Frequency (TF-IDF) encoding. In this process, a TF-IDF vectorizer is employed to convert preprocessed text data into numerical features suitable for machine learning. The 'preprocessed\_texts' variable contains a list of preprocessed text strings, where each string has been cleaned and processed for analysis. The TF-IDF vectorizer, represented by 'vectorizer,' transforms these text strings into a numerical matrix 'X.' Each row of 'X' corresponds to a preprocessed text document, and each column represents a unique term found in the entire corpus. The TF-IDF values in 'X' quantify the importance of each term within each document relative to the entire corpus. Additionally, 'y' contains the corresponding labels or categories for each text document. This vectorization process is crucial for training machine learning models on text data, as it allows algorithms to work with structured numerical data rather than raw text."

*Topic Modeling*

Topic modeling is a type of statistical modeling used in natural language processing to discover abstract topics that occur in a collection of documents. It involves using algorithms to analyze and cluster text data into groups based on their content. This can help to uncover hidden patterns and relationships within the data, making it easier to understand and interpret.

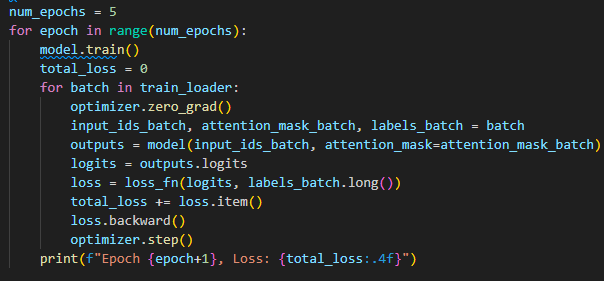


**Figure 14:** Topic modeling using LDA.

The code snippet depicted in Figure demonstrates the process of topic modeling using Latent Dirichlet Allocation (LDA), a widely used technique in natural language processing. Initially, the 'texts' variable gathers the preprocessed text data from the DataFrame and converts it into a list. Subsequently, the text is tokenized into individual words or tokens, resulting in 'tokenized\_texts,' which is a list of lists, with each inner list containing tokens from a respective document. Next, a 'dictionary' is constructed from the tokenized texts, creating a mapping of terms to numerical IDs. The 'corpus' is then generated by converting the tokenized texts into bag-of-words format, where each document is represented as a list of term-frequency pairs. Finally, the LDA model ('lda\_model') is trained on this 'corpus' to discover underlying topics in the text data. The 'num\_topics' parameter specifies the number of topics to be extracted, and 'passes' controls the number of iterations during training. This process enables the extraction of latent topics from the preprocessed text, making it valuable for tasks such as document categorization and content understanding

*Training*

Training refers to the process of teaching a machine learning algorithm to make predictions or decisions based on data. In natural language processing, this can involve training algorithms such as decision trees, support vector machines, and neural networks on text data. Training involves adjusting the parameters of the algorithm to minimize error on a training dataset.



**Figure 15:** Pre-tuning the BERT model.

In the above figure, a training loop is implemented for a deep learning model, typically used for tasks such as text classification or natural language processing. The loop iterates over a specified number of epochs, which represent the complete pass through the training dataset. During each epoch, the model is put into training mode with 'model.train().' Within the epoch loop, the training data is divided into batches using a 'train\_loader.' For each batch, the optimizer's gradients are zeroed to prepare for gradient descent. The batch consists of 'input\_ids\_batch,' 'attention\_mask\_batch,' and 'labels\_batch.' The model is then applied to the input data, generating 'logits' (raw model predictions). A loss function ('loss\_fn') is employed to compute the difference between the model's predictions and the ground truth labels. The cumulative loss for the epoch is updated, and backpropagation is performed to adjust the model's parameters using gradient descent ('optimizer.step()'). After processing all batches within the epoch, the total loss for that epoch is printed. This code snippet exemplifies the training process for deep learning models, with the goal of minimizing the loss and improving the model's predictive capabilities.

*Data Model Evaluation*

In order to calculate the performance of the transfer learning-based Bidirectional Encoder Representation models, they would need to be run through a test set. The test set used for evaluation was from the IMRAD dataset only. This was due to the fact that testing on the other developed transfer learning-based Bidirectional Encoder Representation models as well would just return the same performance levels.

Since the models were used on the prediction of section content, the metrics for classification algorithms that were based on the confusion matrix were used.



**Figure 16:** Confusion matrix.

the confusion matrix is a crucial tool for evaluating the performance of the BERT-based text classification model. The confusion matrix is used to quantify the model's classification results by breaking down its predictions into four categories: true positives, true negatives, false positives, and false negatives. This breakdown allows for a detailed analysis of the model's behavior, helping to identify specific areas where it excels and where it may be making errors. For example, it can reveal whether the model tends to confuse certain classes or if it performs well across all classes. This information is valuable for fine-tuning the model, making decisions about threshold values, and gaining insights into its strengths and weaknesses, ultimately aiding in the refinement of the classification task. The metrics, derived from the confusion matrix, used to measure the performance of the models were discussed in the following:

*Accuracy*

**Equation 2:** Accuracy Formula. Source: Hasty (2023)

The accuracy of a model would measure how often the classifier correctly predicts. It can be defined as the ratio of the number of correct predictions and the total number of predictions.

*Precision*

**Equation 3:** Precision Formula. Source: Hasty (2023)

The precision of a model would explain how many of the correctly predicted cases actually turned out to be positive. It can be defined as the number of true positives divided by the number of predicted values.

*Recall*

**Equation 4:** Recall Formula. Source: Hasty (2023)

The recall of a model would explain how many of the actual positive cases were able to be correctly predicted. It can be defined as the number of true positives divided by the total number of actual positives.

*F1-Score*

**Equation 5:** F1-score Formula. Source: Hasty (2023)

The F1-score of a model would give a combined idea about the precision and recall metrics. It would be maximum when precision would be equal to recall. It can be defined as the harmonic mean of precision and recall.

*Average Precision (AP)*

**Equation 6:** Average precision formula.

It was the sum of precisions divided by the number of relevant documents in the ranked list.

*Mean Average Precision (MAP)*

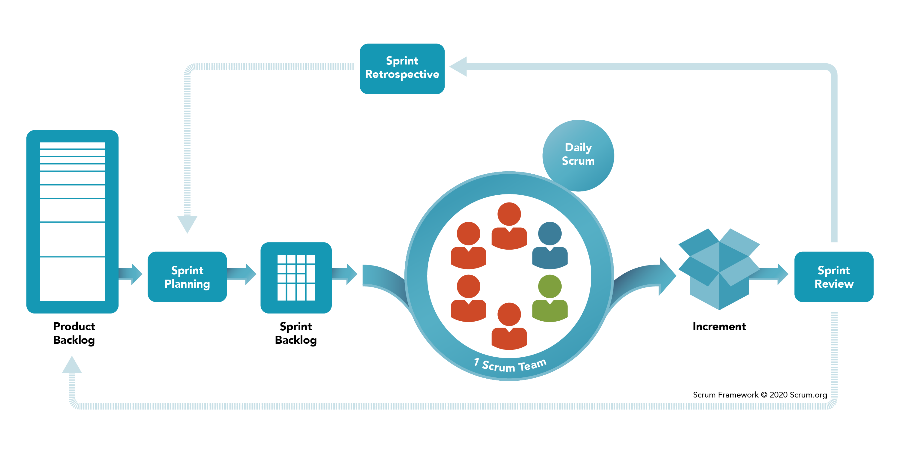
**Equation 7:** Map Average precision formula.

It was the mean of the average precision scores for each query where, Q was the number of queries in the set and *AveP(q)* was the average precision (AP) for a given query, *q*.

System Development Methodology

The development of the “Researchvault: An Advanced Web-Based Research Repository for Laguna State Polytechnic University - Sta. Cruz Campus with Intelligent Chatbot Integration and Collaborative Features for Efficient Research Management” system followed a Software Development Life Cycle (SDLC) model, specifically the SCRUM software development methodology.

SCRUM is an agile framework for managing and completing complex projects. It is based on the principles of transparency, inspection, and adaptation. The SCRUM methodology involves several key stages, including Product Backlog, Sprint Planning, Sprint Backlog, Sprint, Increment, Sprint Review, and Sprint Retrospective.



**Figure 17:** SCRUM software development methodology. Source: Scrum 2023

*Product Backlog*

The SCRUM workflow starts with the construction of a product backlog. This is simply a list of features or functionalities for the system or product. The items on the list are chosen based on suggestions from stakeholders and ideas from the development team. The items are also prioritized, with high-priority items at the top of the list and non-priority items at the bottom. The product backlog is regularly maintained to ensure prioritization is correct and new items are added or removed as needed.

**Table 4:** Product backlog of the researchers.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PRODUCT BACKLOG | | | | | | |
| ID | As a… | I want to be able to… | So that… | Priority | Sprint | Status |
| 1 | User | Search research papers | I can find relevant research materials | High | 1 | done |
| 2 | User | View related research papers | I can see other research materials that may be relevant to my search | High | 1 | done |
| 3 | User | Upload research materials | I can share my research with others | High | 1 | done |
| 4 | User | Browse research materials by category or topic | I can easily find research materials that are relevant to my interests | High | 1 | done |
| 5 | User | Create an account | I can have full access to the system’s features | High | 1 | done |
| 6 | User | Login to my account | I can upload research materials and use system’s features | High | 1 | done |
| 7 | User | Convert manuscript into IMRAD format | I can have a standardized format for my research materials | Medium | 2 | In progress |
| 8 | User | Use the chatbot for inquiries related to the manuscript | I can get instant responses and save time | Medium | 2 | In progress |
| 9 | RIUH | Comment on unapproved research materials | I can provide feedback to improve the quality of the research materials | Low | 3 | To be Started |
| 10 | RIUH | Monitor research materials through the IMRAD journal | I can track progress, review content, and provide feedback | Low | 3 | To be Started |

*Sprint Planning*

In sprint planning, the development team decides which product backlog items will be worked on in the upcoming sprint. A meeting is held to determine what can be delivered and how it will be achieved. The duration of the sprint is also decided during this meeting.

*Sprint Backlog*

The output from sprint planning is a sprint backlog. This is a list of product backlog items that will be worked on during the sprint.

**Table 5:** Sprint backlog for the first sprint.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SPRINT BACKLOG | | | | | | |
| ID | As a… | I want to be able to… | So that… | Priority | Sprint | Status |
| 1 | User | Search research papers | I can find relevant research materials | High | 1 | done |
| 2 | User | View related research papers | I can see other research materials that may be relevant to my search | High | 1 | done |
| 3 | User | Upload research materials | I can share my research with others | High | 1 | done |
| 4 | User | Browse research materials by category or topic | I can easily find research materials that are relevant to my interests | High | 1 | done |
| 5 | User | Create an account | I can have full access to the system’s features | High | 1 | done |
| 6 | User | Login to my account | I can upload research materials and use system’s features | High | 1 | done |

*Sprint*

During the sprint, the development team works on completing the items in the sprint backlog. Daily scrum meetings are held to discuss progress and ensure everything is on track.

*Increment*

The outcome of a sprint is a usable product called an increment. This is the integration of all completed items from the sprint backlog.

*Sprint Review*

At the end of a sprint, a sprint review is held where the development team presents the increment to stakeholders for feedback.

*Sprint Retrospective*

After the sprint review, a sprint retrospective is held where the development team reflects on their performance during the sprint and identifies areas for improvement.

Performance Evaluation of Algorithms

In the performance evaluation phase, the researchers assessed the effectiveness of the algorithms integrated into the web application, including the LDA topic model, Random Forest, Bidirectional Encoder Representation (BERT). This was done through actual testing of data, allowing for an objective evaluation of the models' performance in terms of accuracy, efficiency, and other relevant metrics.

*LDA Topic Modeling*

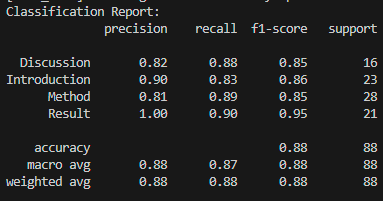
The evaluation of the LDA topic model was primarily based on quantitative metrics designed to assess the quality of the topics generated by the model. These metrics provide objective measures of the model's performance without the need for human interpretation.

The process begins with data preprocessing, where text data from multiple CSV files undergoes essential tasks such as tokenization, stemming, and named entity recognition. Missing values are also handled appropriately.

Following preprocessing, the code employs Latent Dirichlet Allocation (LDA) for topic modeling, uncovering latent themes within the text corpus. The text data is then transformed into numerical features using TF-IDF vectorization, preparing it for machine learning.

The performance is evaluated using a dedicated test dataset. It generates a classification report, offering essential metrics like accuracy, precision, recall, and F1-score for assessing the model's performance.

To ensure the model's availability for future use, it is saved as a file using the joblib library, allowing it to make predictions on new, unseen text data when needed.



**Figure 18:** Classification report on LDA model training

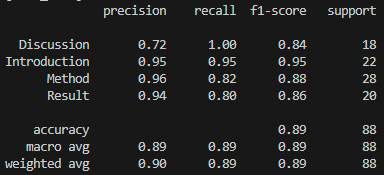
The classification report provided in the above figure is based on the performance of an LDA (Latent Dirichlet Allocation) model in classifying IMRAD sections. It offers insights into the model's ability to correctly classify different section types. Precision measures how often the model's positive predictions are correct, while recall assesses its ability to capture all relevant instances. The F1-score is a balanced metric that combines both precision and recall. In this context, the LDA model demonstrates moderate to high performance in classifying IMRAD sections, with an overall accuracy of 88%. The precision and recall values vary across sections, with Method and Discussion sections exhibiting higher precision and recall compared to Introduction and Result sections. These metrics collectively indicate the model's effectiveness in categorizing IMRAD sections, with room for improvement in some areas.

*Random Forest*

A Random Forest algorithm is utilized to perform text classification. The model's performance is evaluated using common metrics such as accuracy, precision, recall, and F1-score. The evaluation is conducted on a test dataset, which is split into a training set and a testing set.

The training set is used to train the Random Forest model. During this training phase, the model learns patterns and relationships within the preprocessed text data to make predictions.

The testing set is employed to assess the model's performance. It serves as an independent dataset to evaluate how well the trained model can generalize to new, unseen text documents. The evaluation results provide insights into the model's accuracy and effectiveness in classifying documents into predefined categories.



**Figure 19:** Classification report of random forest model training.

The classification report presented in the figure above summarizes the performance of a Random Forest model in classifying IMRAD sections. It provides essential evaluation metrics for each class, including precision, recall, and F1-score, which assess the model's ability to make correct predictions for each section type. The high precision values indicate that the model generally makes accurate positive predictions for each section. Meanwhile, the recall values show the model's effectiveness in identifying relevant instances for each class. The F1-scores strike a balance between precision and recall, offering an overall measure of classification performance. In this context, the model achieves a high level of accuracy, correctly classifying IMRAD sections with an accuracy of 89%. These metrics collectively demonstrate the model's capability to classify IMRAD sections effectively.

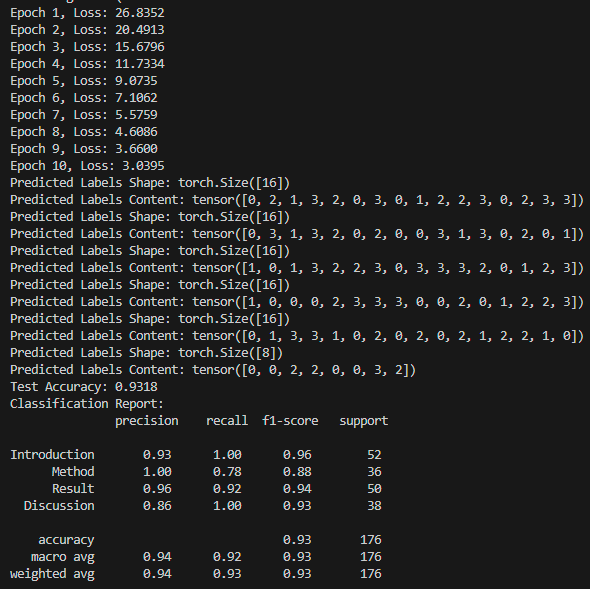
***Bidirectional Encoder Representation (BERT)***

The BERT algorithm underwent a comprehensive evaluation, employing a range of performance metrics including accuracy, precision, recall, and F1-score. This evaluation process involved the utilization of a dedicated test dataset that was thoughtfully divided into a training set and a validation set.

The training set was designated for the primary purpose of training the BERT model, allowing it to learn from the available data. Meanwhile, the validation set was reserved for assessing the model's performance, ensuring a rigorous evaluation.

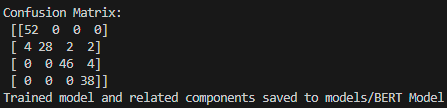
The results of this evaluation exhibited the BERT model's commendable performance across multiple dimensions, most notably in terms of accuracy, underlining its efficacy in classifying text data according to the specified IMRAD labels.

The researchers conducted an in-depth analysis of the attention weights within the BERT model. This scrutiny of attention weights provided valuable insights into the inner workings of the model, shedding light on how it makes predictions. These insights not only enhanced the researchers' understanding of the model but also paved the way for potential improvements, further refining its overall performance. This holistic evaluation approach underscores the thoroughness and rigor of the research effort.



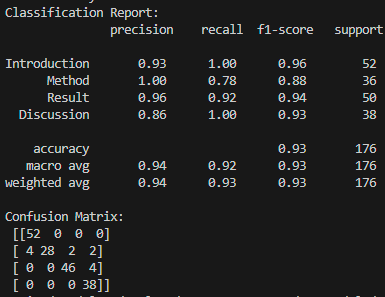
**Figure 20:** Epoch Loss from training BERT model.

The output provided in the figure above represents the training progress of a BERT model for classifying IMRAD sections over five epochs. The loss values steadily decrease from 26.8352 in the first epoch to 9.0735 in the fifth epoch. This trend indicates that the model is progressively improving its predictive accuracy as it learns from the training data. Lower loss values signify better alignment between the model's predictions and the actual IMRAD section labels, demonstrating that the model is effectively adapting its parameters to the task.

****

**Figure 21:** Confusion matrix from training BERT model.

The confusion matrix presented here summarizes the performance of a BERT model for IMRAD section classification. It consists of a 4x4 matrix, where each row corresponds to the true labels, and each column represents the predicted labels made by the model. The values in the matrix denote the number of instances falling into specific combinations of true and predicted labels. For instance, the top-left cell (52) signifies the count of accurate predictions where the true label was "Introduction," and the model correctly classified it as such. Conversely, the matrix also highlights misclassifications, such as the bottom-right cell (38), indicating the count of instances where the true label was "Discussion," but the model predicted it as "Discussion."



**Figure 22:** Classification report of BERT model training.

The classification report in the above figure offers a comprehensive evaluation of a BERT model's performance in IMRAD section classification. It provides precision, recall, and F1-score metrics for each IMRAD section class (Introduction, Method, Result, Discussion) and an overall summary. Precision measures the accuracy of positive predictions, reflecting how many instances classified as a particular section are actually that section. Recall gauges the model's ability to correctly identify all instances of a particular section within the dataset. F1-score is the harmonic mean of precision and recall, offering a balanced measure of a model's accuracy. For instance, in the "Introduction" section, the model achieved a precision of 0.93, indicating that 93% of the instances itclassified as "Introduction" were accurate. The recall of 1.00 suggests that the model correctly identified all instances of "Introduction" in the dataset. The F1-score of 0.96 is a balanced measure of precision and recall. The macro average and weighted average metrics provide overall summaries of the model's performance across all classes, showing that the model has an average F1-score of 0.93, indicating strong overall classification performance. It achieved an accuracy of 0.93, reflecting the proportion of correctly classified instances in the dataset. These metrics collectively indicate that the BERT model has performed well in IMRAD section classification.

Software Used

A variety of software was employed in the development and testing of algorithms for this research. The following sections provide a discussion on the software used in this study:

*Visual Studio Code*

Visual Studio Code is a source code editor developed by Microsoft for Windows, Linux, and macOS. It includes support for debugging, embedded Git control and GitHub, syntax highlighting, intelligent code completion, snippets, and code refactoring. It is highly customizable and has a large and active community that contributes to its development and provides support to its users.

*Python Libraries*

Several Python libraries were used in this research. These include:

*OS*

It provides a portable way of using operating system dependent functionality. It allows for interaction with the underlying operating system, including file and directory management, process management, and environment variable manipulation.

*Secrets*

It is used for generating cryptographically strong random numbers suitable for managing data such as passwords, account authentication, security tokens, and related secrets. It provides a secure way of generating random data that is suitable for use in cryptographic applications.

*Spacy*

It is an open-source software library for advanced natural language processing. It provides tools for tasks such as tokenization, part-of-speech tagging, named entity recognition, dependency parsing, and text classification.

*IO*

It provides the Python interfaces to stream handling. It allows for the creation and manipulation of streams of data, including reading from and writing to files.

*Torch*

It is an open-source machine learning library for Python that provides tensor computation with strong GPU acceleration. It is widely used in the field of deep learning and provides tools for building and training neural networks.

*PyPDF2*

It is a pure-Python library built as a PDF toolkit. It is capable of extracting document information (title, author, etc.), splitting documents page by page, merging documents page by page, cropping pages, merging multiple pages into a single page, encrypting and decrypting PDF files. It provides a wide range of functionality for working with PDF files.

*Reportlab*

It is an open-source engine for creating complex, data-driven PDF documents and custom vector graphics. It provides tools for generating high-quality PDF documents with rich formatting and layout options.

*Transformers*

It is a state-of-the-art Natural Language Processing library for TensorFlow 2.0 and PyTorch. It includes pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers) that can be fine-tuned for specific tasks such as text classification or named entity recognition.

*Pandas*

It is an open-source data analysis and manipulation tool, built on top of the Python programming language. It provides data structures such as DataFrame and Series that allow for easy manipulation of tabular data.

*Sklearn.metrics*

It provides several functions to measure the quality of predictions from classification algorithms, including the classification\_report function which generates a text report showing the main classification metrics, and the confusion\_matrix function which computes the confusion matrix to evaluate the accuracy of a classification.

*Flask*

Is a micro web framework written in Python that allows you to build web applications quickly and with minimal code. It provides tools such as routing and request handling that make it easy to develop web applications.

*Database Connection*

A database connection was established using the DBConnection class from the db\_connection module. This allowed for efficient storage and retrieval of data used in this research. The class provides methods for connecting to a database server, executing SQL queries, and managing transactions.

System Architecture

This section presents the system architecture of the developed system. Enumerate the functions and features highlight the integration of computing science solutions. Discuss the structure, behavior, and more views of a system.

Users can interact with the ResearchVault system through a web application using its frontend or client, which is developed using HTML, CSS, and JavaScript. The frontend provides a user-friendly interface for browsing research materials, leaving comments, generating citations, and utilizing other platform features. It communicates with the backend server via HTTP requests and APIs to fetch and display research data.

The backend of the ResearchVault system is responsible for processing user requests, managing data, and implementing the core functionality of the platform. It is built using the Flask framework in Python to handle HTTP requests and provide routing for different features. The backend server interacts with the SQLite database to retrieve and store research materials, user data, comments, and other relevant information.

The system's database uses SQLite to store all research materials, user profiles, comments, and metadata related to research papers. This lightweight database system is suitable for smaller-scale applications and can efficiently handle data storage and retrieval for ResearchVault.

Additionally, ResearchVault includes an IMRAD Converter that was created using BERT (Bidirectional Encoder Representations from Transformers), NLP (Natural Language Processing) techniques, and a Random Forest model. This converter analyzes and structures research papers according to the IMRAD (Introduction, Methods, Results, and Discussion) format. The converted documents are then stored in the SQLite database for easy retrieval and access.

Software Testing

Software testing, as defined by Sawant et al. (2012), is a critical process aimed at evaluating software to identify and rectify errors. In this research context, the testing focused on assessing the accuracy and performance of integrated models designed for classifying IMRAD sections. The researchers conducted this testing after the system's development, with a particular emphasis on evaluating the functionality of the web application.

To assess the performance of the IMRAD converter, the researchers made use of a comprehensive dataset containing diverse research manuscript materials. This dataset served as the input for the system during the testing phase, encompassing a wide range of research papers, reflecting real-world variations in academic writing styles, disciplines, and content.

These datasets were then fed to the system for testing. The system utilized various machine learning. Specifically, the researchers harnessed the power of text classification techniques to categorize sections of research manuscripts into the distinct IMRAD structure components, such as Introduction, Methods, Results, and Discussion. The researchers also employed Bidirectional Encoder Representation from Transformer (BERT), a state-of-the-art transformer-based model, to gain a deep understanding of contextual nuances and relationships within the text. This advanced technique significantly improved the accuracy of IMRAD section classification.

The researchers also utilized the Random Forest algorithm to enhance overall classification accuracy. This involved aggregating predictions from multiple decision trees, contributing to the system's proficiency in IMRAD classification. Latent Dirichlet Allocation (LDA) Topic Modeling played a vital role in uncovering underlying themes and topics within research manuscripts. This added layer of analysis further refined the classification process.

To assess the system's performance comprehensively, the researchers employed several well-established evaluation metrics. The confusion matrix allowed them to visualize the models' performance, helping identify true positives, true negatives, false positives, and false negatives. Monitoring the training process over multiple epochs provided insights into how quickly the models converged and whether further training was necessary. Classification reports offered detailed statistics on precision, recall, F1-score, and support for each IMRAD section, providing a thorough understanding of the models' proficiency in classifying research manuscripts