‘Financial Fraud Detection System Using NLP AND ML’

Project Report submitted to Shri Ramdeobaba College of Engineering & Management,Nagpur in partial fulfillment of requirement for the award of degree of

Bachelor of Technology

in

COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

by

1.Divya Chandak

2.Anurag Thakre

3.Amitesh Jaiswal

Guide

Prof. A. V. Zadgaonkar

RCOEM

Shri Ramdeobaba College of
Engineering and Management, Nagpur

Computer Science and Engineering (Data Science)

Shri Ramdeobaba College of Engineering & Management, Nagpur
(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur)

December 2023

SHRI RAMDEOBABA COLLEGE OF ENGINEERING & MANAGEMENT, NAGPUR

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur)

Department of Computer Science and Engineering (Data Science)

CERTIFICATE

This is to certify that the project on "Financial Fraud Detection System Using NLP AND ML" is a bonafide work of

1.Divya Chandak
2.Anurag Thakre
3.Amitesh Jaiswal

submitted to the Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur in partial fulfillment of the award of a Degree of Bachelor of Engineering, in Computer Science and Engineering (Data Science). It has been carried out at the Department Computer Science and Engineering (Data Science), Shri Ramdeobaba College of Engineering and Management, Nagpur during the academic year 2023-24.

Date: 30.12.2023

Place: Nagpur

Prof. Prof. A. V. Zadgaonkar
Project guide

Prof. A. M. Karandikar
H.O. D
Department of Computer Science and Engineering (Data Science)

Dr. R. S. Pande
Principal (RCOEM)

DECLARATION

I, hereby declare that the project titled **“Financial Fraud Detection System Using NLP AND ML”** submitted herein, has been carried out in the Department of Computer Science and Engineering (Data Science) of Shri Ramdeobaba College of Engineering & Management, Nagpur. The work is original and has not been submitted earlier as a whole or part for the award of any degree / diploma at this or any other institution / University

Date: 30.12.2023

Place: Nagpur

Divya Chandak **(Roll no.: 05)**	Anurag Thakre **(Roll no.: 30)**
Amitesh Jaiswal **(Roll no.: 27)**	

ACKNOWLEDGEMENT

We would like to express our deep and sincere gratitude to our guide **Prof. A. V. Zadgaonkar**, Prof. Ashwini Vinayak Zadgaonkar Assistant Professor, Department of Computer Science & Engineering (Data Science), RCOEM, for giving us the opportunity to work on this project and providing valuable guidance throughout the project. It was a great privilege and honor to work under his guidance. We are extremely grateful for the experience we had in this project with him.

We express our sincere gratitude to **Prof. A. M. Karandikar** , Head of the Department of Computer Science and Engineering (Data Science), RCOEM for his guidance. Talent wins games, but teamwork and intelligence win championships. We would like to take this opportunity to express our deep gratitude to all those who extended their support and guided us to complete this project.

Divya Chandak
(Roll no.: 05)

Anurag Thakre
(Roll no.: 30)

Amitesh Jaiswal
(Roll no.: 27)

(Specimen `D') Approval Sheet

This report entitled (Title) by (Author Name) is approved for the degree of Bachelor of Technology, in Computer Science and Engineering (Data Science)

Name & signature of Supervisor(s) Name & signature of External.
Examiner(s)

------------------ ------------------

------------------ ------------------

Prof. A. M. Karandikar

H.O. D

Department of Computer Science and Engineering

(Data Science)

Date: 30.12.2023

Place: Nagpur

ABSTRACT

This comprehensive report presents an advanced software solution strategically developed to counter the widespread threat of financial fraud within the global financial system. The innovative system uses state-of-the-art technologies, namely **natural language processing (NLP)** and **machine learning (ML)**, to create a robust defense against fraudulent activities.

At the heart of our pioneering approach is the application of Latent Dirichlet Allocation (LDA), a technique used to identify 30 distinct topics from a carefully curated set of relevant documents. These documents cover a diverse range, including balance sheets, contractual agreements, credit agreements and court cases, all specifically related to financial fraud. This broad scope ensures a comprehensive understanding of the various aspects of fraudulent activities in the financial sector.

The proposed system carefully processes unstructured data, uses NLP for efficient feature extraction and ML techniques for accurate document identification, fraud detection and risk prioritization. This pioneering methodology equips financial institutions, regulators and law enforcement agencies with proactive tools to address and mitigate various forms of financial fraud. By significantly reducing risk and preserving the integrity of the global financial system, our software represents a major leap forward in document-based fraud detection.

In addition, the system's ability to analyze different types of documents, including balance sheets and legal agreements, ensures a holistic approach to fraud detection that covers different dimensions of financial transactions. This multifaceted strategy increases the effectiveness of fraud prevention and facilitates a more thorough understanding of potential risks in the financial environment. Essentially, our software serves as a key ally in protecting the financial sector from evolving and sophisticated fraud schemes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In the ever-evolving environment of global finance, the looming threat of financial fraud poses a significant challenge that can undermine the very foundations of the global economic system. Traditional fraud detection methods struggle to keep pace with the increasingly sophisticated and dynamic nature of fraudulent activity. Recognizing the urgent need for a more robust and proactive approach, we present a revolutionary software solution that harnesses the power of natural language processing (NLP) and machine learning (ML) to identify potential fraudulent activity in a complex network of financial records.

Financial fraud, with its potential to disrupt the global financial system, has emerged as a critical problem. The lack of current fraud detection methods requires a paradigm shift towards innovative and advanced technologies. The complex nature of financial transactions and the variety of document types make it difficult to effectively detect fraudulent activity. The need of the hour is a solution that can not only sift through vast amounts of unstructured data, but also recognize patterns and anomalies indicative of fraudulent behavior.

Our proposed software solution serves as a technological barrier against financial fraud by leveraging NLP and ML capabilities. Tailored to identify suspicious activity in financial records, the software follows a systematic process. It starts by collecting and pre-processing unstructured data and then uses NLP techniques to extract features. The extracted features then become input to ML algorithms that are trained to detect fraud and prioritize risks. This synergistic approach ensures comprehensive and intelligent analysis of financial documents and increases the ability to identify potential fraud before it can cause significant damage.

The variety of document types plays a vital role in detecting financial fraud. Our software is adept at categorizing documents such as balance sheets, contracts, credit agreements and court cases related to financial fraud. By tailoring the analysis to the specifics of each document type, the software increases its accuracy in identifying subtle patterns of fraudulent behavior. This adaptability is a key strength that ensures the system remains effective across the spectrum of financial instruments and legal documents.

Example Scenario:
Suppose an income tax investigator is tasked with reviewing a set of documents related to a potential financial fraud case. Among the documents, there's a mix of balance sheets, loan agreements, contractual agreements, business reports, and court-related financial fraud cases.

-Balance Sheet Identification:
Our software, equipped with advanced NLP algorithms, can swiftly identify patterns indicative of a balance sheet. It recognizes financial terms, account names, and the overall structure typical of balance sheets.For instance, if the document contains terms like "assets," "liabilities," and "equity," along with numerical values organized in a structured format, the system would confidently categorize it as a balance sheet.

-Loan Agreement Recognition:
In documents pertaining to loan agreements, the software leverages ML techniques to detect specific language patterns, legal terms, and clauses associated with loans. If phrases such as "borrower," "lender," "interest rate," and "repayment terms" are identified, the system categorizes the document as a loan agreement, aiding investigators in focusing on potential areas of financial scrutiny.

-Business Report and Court Case Analysis:
The software uses ML techniques for nuanced document types like business reports and court-related financial fraud cases. It recognizes contextual language and specific legal terminology. For example, if a document contains references to financial performance indicators, market analysis, or legal proceedings related to financial fraud, the system appropriately categorizes it as a business report or court case.

CHAPTER 2

PROBLEM DEFINITION

Definition: Financial Frauds have the potential to threaten the global financial system, and current fraud detection methods are insufficient.

Analysis: Financial frauds, with their evolving complexity, demand a proactive and sophisticated approach to safeguard the integrity of the financial ecosystem. The acknowledgment of inadequacies in existing detection methods emphasizes the need for innovation in combating these threats. This project, by addressing the limitations of current practices, aims to introduce a cutting-edge software solution equipped with Natural Language Processing and Machine Learning capabilities. The emphasis on the potential global impact of financial frauds underscores the project's significance in fortifying the financial sector against emerging risks and ensuring the stability of the international financial system. Our software collects and preprocesses the unstructured data, applies NLP for feature extraction, and employs ML for fraud detection and risk prioritization. The software’s output is skillfully represented through a visual narrative employing charts(for comparisons) and word clouds(for summarized view) of the dataset.

CHAPTER-3

MODELS

3. Models for Topic Modelling

3.1 LSA

Latent Semantic Analysis (LSA) plays a crucial role in Natural Language Processing (NLP) when it comes to uncovering topics and analyzing text. Often referred to as Latent Semantic Indexing (LSI), LSA is a method that transforms complex text data into a simpler form, revealing hidden patterns and connections. In the context of figuring out the main themes within a group of documents, LSA identifies these underlying topics by looking at how words co-occur. The idea behind LSA is pretty cool: if words show up in similar contexts, they probably share some meaning. LSA makes this happen by creating a term-document matrix that shows how words and documents relate. Then, it uses a technique called Singular Value Decomposition (SVD) to simplify this matrix while keeping the important semantic connections intact.

In the world of topic modeling, LSA's magic lies in turning documents into a space where topics are revealed as combinations of words with different weights. This makes it handy for tasks like grouping documents, finding information, and summarizing content. But, like all superheroes, LSA has its quirks. It might struggle with words that have multiple meanings or lose some of the finer details of how words fit together. That's where fancier models like Latent Dirichlet Allocation (LDA) step in. Nevertheless, LSA is still a go-to method, especially when you want things to be straightforward, easy to understand, and fast in the world of NLP topic modeling.

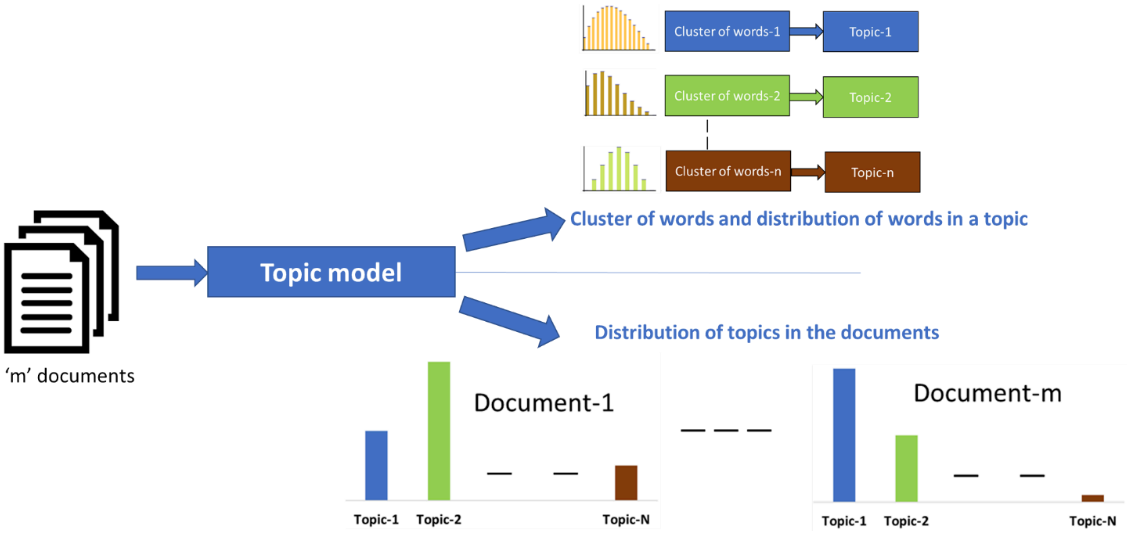


Figure 3.1 LSA Model

3.1.1 Problems with LSA

1. **Loss of Local Context:** One significant issue with LSA is that it tends to lose local context, as it relies heavily on the global structure of the term-document matrix. This means that the model might struggle to capture the subtle nuances of meaning that arise from the specific context of words within a document. As a result, LSA may not be as effective when dealing with short documents or when the precise arrangement of words is crucial for understanding the topic.

2. **Difficulty Handling Polysemy:** LSA may face challenges in handling polysemy, where a single word has multiple meanings. Since LSA primarily relies on co-occurrence patterns, it might struggle to disambiguate the different meanings of a polysemous word. This can lead to less precise topic modeling, where the model may assign a word to a topic based on its overall co-occurrence patterns rather than its specific contextual meaning in a given instance.

3.2 NMF

Non-Negative Matrix Factorization (NMF) stands out as a valuable tool in Natural Language Processing (NLP), especially when it comes to uncovering hidden topics within a collection of documents. Unlike other matrix factorization methods, NMF keeps things positive, limiting its factorized matrices to non-negative values—a perfect fit for scenarios where features like word frequencies can't go negative. In the realm of topic modeling, NMF works its magic by breaking down a term-document matrix into two smaller matrices. One of these represents the discovered topics, while the other signifies how much of each topic is present in each document. The underlying assumption is that the original data can be roughly approximated as the product of these non-negative matrices. This results in documents being expressed as combinations of topics with non-negative weights, making the representation both clear and easy to understand.

What makes NMF shine is its interpretability. The topics it uncovers are essentially positive blends of words, making them intuitive and user-friendly. This aspect proves invaluable in situations where a straightforward grasp of topics is crucial, such as in exploratory text analysis or when sharing findings with non-tech-savvy audiences. Nevertheless, NMF does have its quirks. It's a bit picky about initial values, and sometimes reaching a solution isn't a guaranteed smooth ride. Dealing with noisy data and outliers can also be a bit tricky for NMF. Despite these quirks, it remains a go-to choice for many in the NLP community, offering a nice balance between interpretability and computational efficiency. When deciding between NMF and other topic modeling approaches, practitioners often weigh the specific traits of their data against the need for clear topic interpretation in the world of NLP.

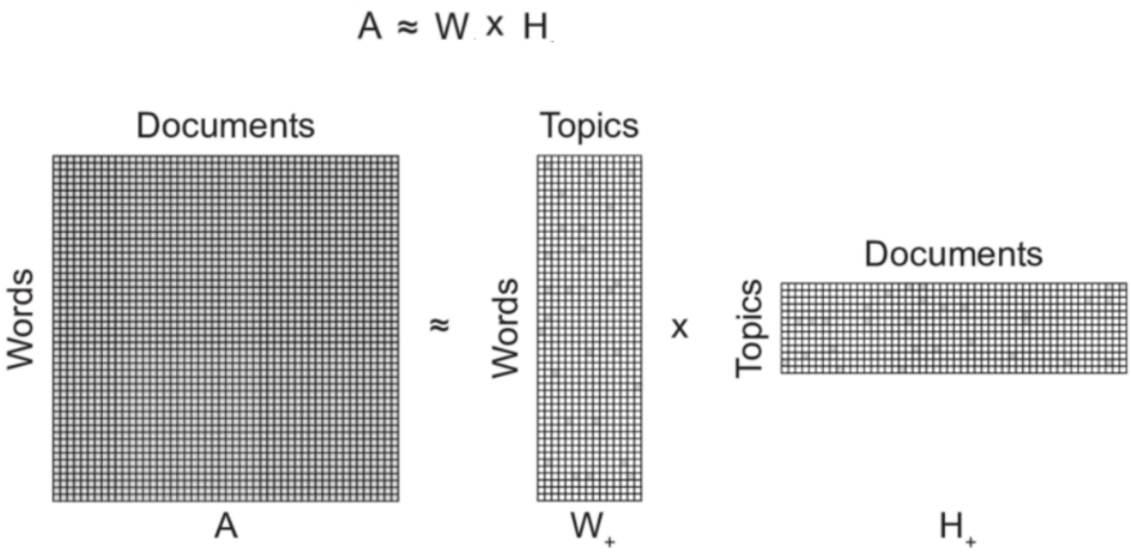
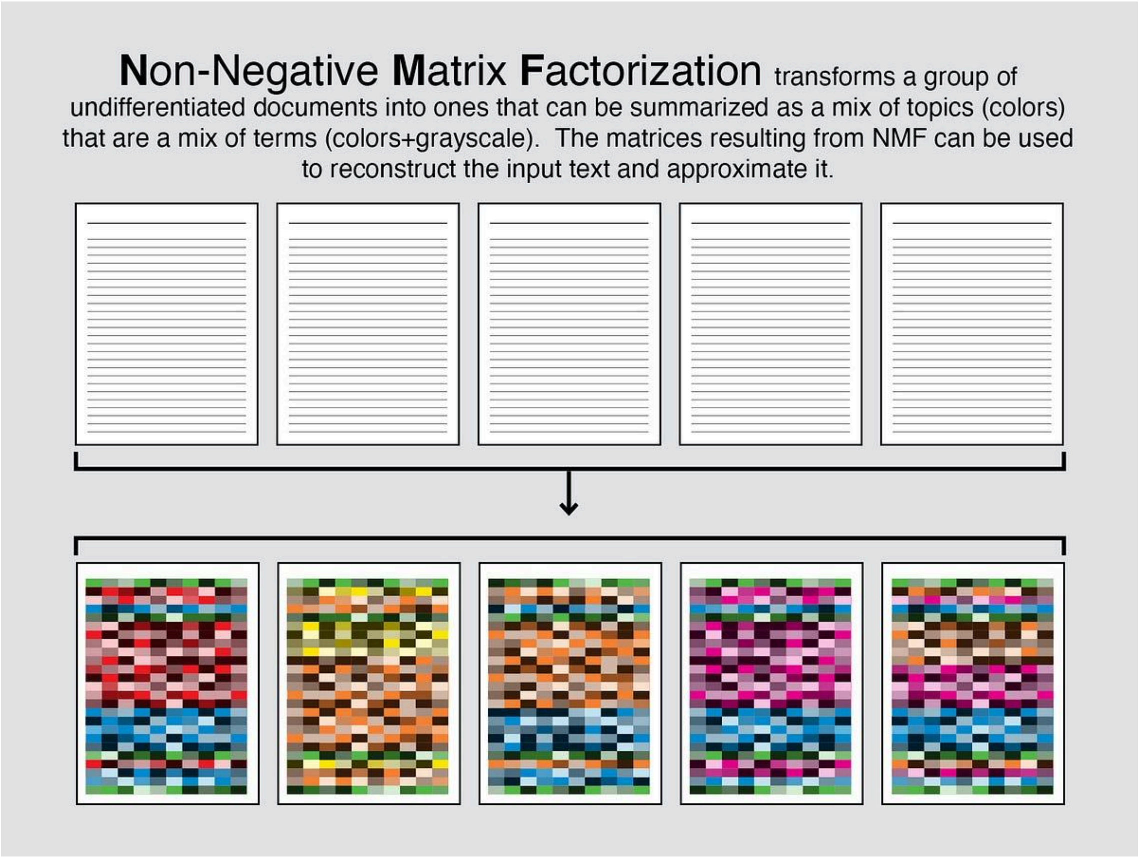


Figure 3.2 Topic modeling using NFM



Non-Negative **M**atrix **F**actorization transforms a group of undifferentiated documents into ones that can be summarized as a mix of topics (colors) that are a mix of terms (colors+grayscale). The matrices resulting from NMF can be used to reconstruct the input text and approximate it.

Figure 3.3 Non-Negative Matrix Factorization

3.2.1 Problems with NMF

1. **Sensitivity to Initialization:** NMF is sensitive to the choice of initial values during the factorization process, which can impact the convergence of the algorithm and the quality of the obtained topics. Different initializations may lead to different local minima, affecting the stability and reproducibility of the results. This sensitivity poses a challenge in ensuring the consistency and reliability of topic modeling outcomes.

2. **Limited Handling of Noise and Outliers:** NMF may struggle with noisy data or outliers, as it seeks to represent the input data as a combination of non-negative components. Noisy or outlier-laden documents can distort the factorization process, potentially leading to the extraction of topics that are influenced by irrelevant or anomalous information. Robustness to noise is crucial in real-world applications, where datasets may contain varying degrees of noise and inconsistencies.

3.3 LDA

In the fascinating realm of Natural Language Processing (NLP), the Latent Dirichlet Allocation (LDA) model emerges as a powerful force in the art of topic modeling. Picture this: a vast sea of text documents, each holding a unique story. LDA, akin to a skilled detective, steps in to unveil the hidden themes and patterns interwoven within this textual tapestry. In the world of topic modeling, LDA paints a vivid picture where documents are like intricate blends of various topics, and each topic, in turn, is a mosaic of carefully chosen words. It's a symphony of probabilities, where every word in a document dances to the tune of a particular topic, revealed through a probabilistic lens.

LDA's magic lies in its iterative journey, refining its understanding of document-topic and topic-word relationships until it captures the essence of the corpus. The outcome? A set of topics, each with its own unique word composition, and each document adorned with a distinctive blend of these topics. In practical terms, LDA finds its application in categorizing documents, suggesting relevant content, and aiding in the quest for information. It's like a guide through the labyrinth of words, helping us make sense of large volumes of text by uncovering the prevalent topics and their significance in the dataset. LDA, in the hands of researchers and industry experts, becomes a beacon, illuminating the path to extracting meaningful insights from the seemingly chaotic world of unstructured textual data.

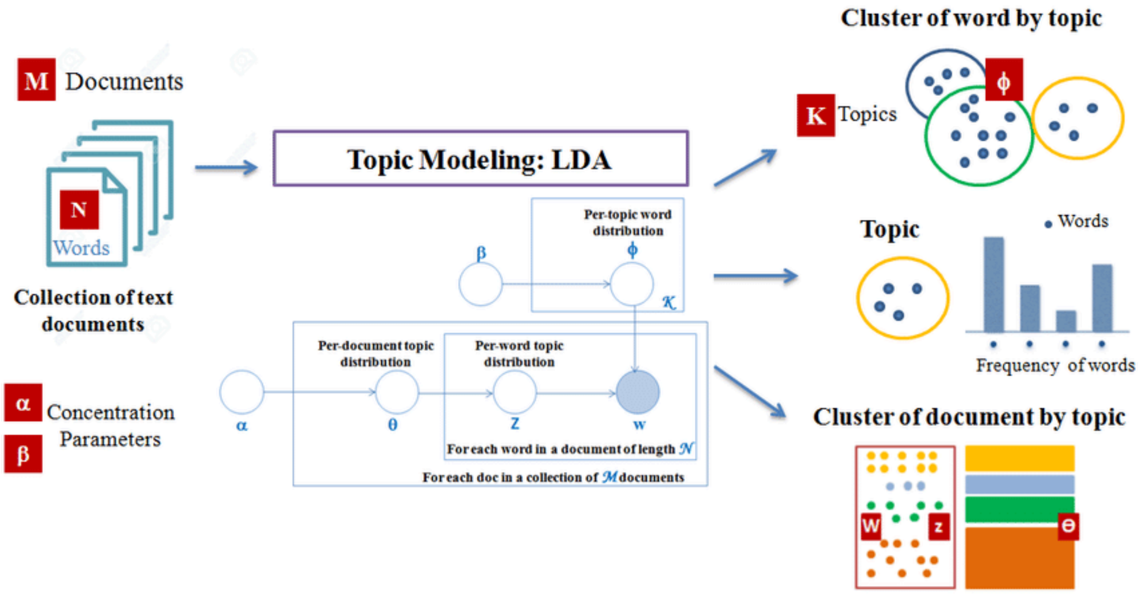

M Documents
N Words
Collection of text documents
α Concentration Parameters
β
Topic Modeling: LDA
Per-topic word distribution
β
ϕ
K
Per-document topic distribution
Per-word topic distribution
α
θ
z
w
For each word in a document of length N
For each doc in a collection of M documents
Cluster of word by topic
K Topics
ϕ
Topic
Words
Frequency of words
Cluster of document by topic
W
z
θ

Figure 3.4 LDA Model

3.4 Comparison of models

3.4.1 LSA vs. LDA

1. **Loss of Local Context in LSA:** LDA mitigates the issue of losing local context by employing a probabilistic generative model. Unlike LSA, which relies on a global structure, LDA assumes that each document is a mixture of topics, and each word in the document is attributed to a particular topic. This probabilistic approach allows LDA to capture the local context more effectively. By considering the distribution of topics within a document, LDA provides a more nuanced understanding of how words relate to each other within a specific context.

2. **Difficulty Handling Polysemy in LSA:** LDA addresses the challenge of polysemy by modeling words based on their likelihood of co-occurring in topics. In LDA, a word can belong to different topics with varying probabilities, reflecting its polysemous nature. This allows LDA to better capture the nuanced meanings of words in different contexts. By considering the probability distribution of words across topics, LDA provides a more nuanced representation, distinguishing between the different senses of a polysemous word.

3.4.2 NMF vs LDA

1. **Sensitivity to Initialization in NMF:** NMF's sensitivity to initialization can lead to different local minima and, consequently, inconsistent results. LDA, on the other hand, relies on a generative probabilistic model, which makes it less sensitive to initialization variations. The inherent randomness in the topic assignment process of LDA helps achieve more stable and reproducible results compared to NMF. LDA's probabilistic nature provides a more robust foundation for inferring topics, reducing the impact of initialization choices.

2. **Limited Handling of Noise and Outliers in NMF:** NMF may struggle with noisy data or outliers, as it aims to represent the input data as a combination of non-negative components. LDA, being a probabilistic model, accounts for uncertainty in topic assignments and is more resilient to noise. By probabilistically modeling the generation of documents, LDA can better distinguish between relevant patterns and outliers, resulting in more accurate topic modeling in the presence of noisy or outlier-laden data.

3.4.2 LDA vs. Other models (F1 score, Recall, Precision) LDA vs. Other models (F1 score, Recall, Precision)

Latent Dirichlet Allocation (LDA) stands out as a superior performer in topic modeling and text similarity tasks, showcasing higher F1 scores, precision, and recall compared to competing methods such as NMF, PCA, LSA and RP.

LDA's probabilistic approach, modeling documents as mixtures of topics and words as distributions across topics, allows for more nuanced representations. This flexibility results in improved precision and recall by capturing the underlying thematic structures in text data more accurately.

Unlike deterministic methods like NMF, PCA, LSA, and RP, LDA's probabilistic nature aligns better with the inherently uncertain nature of language, leading to more robust performance in topic modeling and text similarity tasks.

The proof is in the metrics: LDA consistently outshines its counterparts when it comes to extracting meaningful topics and gauging textual similarity.

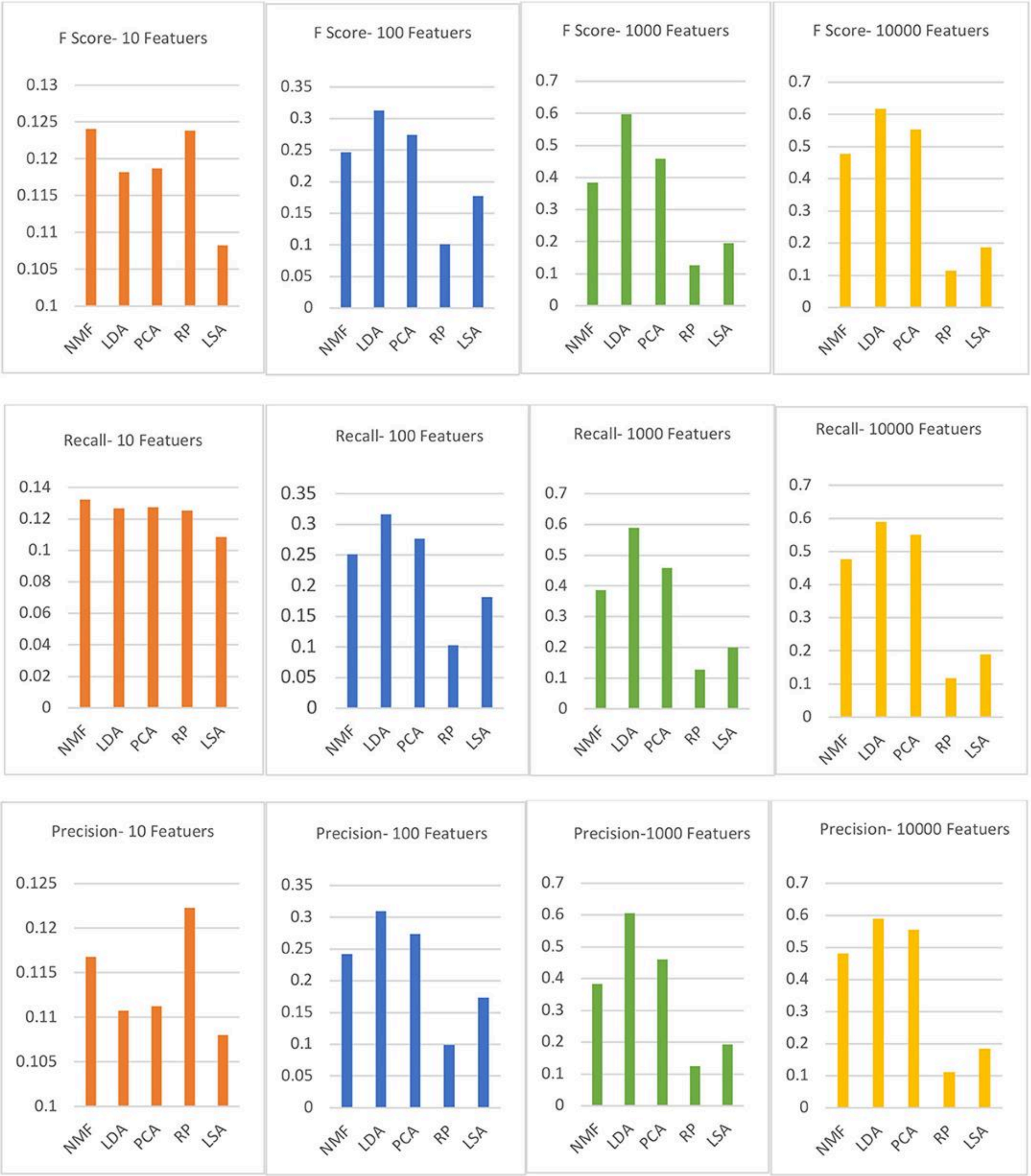


Figure 3.5 Comparison of LDA vs. Other Models

CHAPTER 4

Fraud Detection and Recognition

Stage 1: Data Collection and Preprocessing

Objective: Gather raw, unstructured data relevant to financial transactions or documents, preprocess the data, and form a corpus of words.

Activities: Collect a dataset of 50 financial documents. Preprocess the data by cleaning, tokenizing, converting to lowercase, removing stop words, and stemming or lemmatizing. Form a corpus of words representing the vocabulary in the financial documents.

Stage 2: Gensim and Bigram Application

Objective: Apply Gensim library and identify bigrams (two-word combinations) within the corpus.

Activities: Use Gensim to identify bigrams in the corpus, considering common two-word combinations. Create a bigram model and a trigram model to enhance the meaning captured by word combinations.

Stage 3: Stopword Removal

Objective: Remove stopwords from the corpus.

Activities: Define a function (remove_stopwords) to remove stopwords from the tokenized text.

Stage 4: Creating Document Term Matrix

Objective: Convert the list of documents (corpus) into a Document Term Matrix using the dictionary created from the trigrams.

Activities: Use Gensim to create a dictionary mapping words to numerical IDs.

Create the Document Term Matrix using the dictionary and the trigram-transformed corpus.

Stage 5: Finding Coherence Score

Objective: Assess the quality of the topic model by finding the coherence score.

Activities: Calculate the coherence score to evaluate the consistency and interpretability of the topics.

Stage 6: LDA Model Implementation

Objective: Apply Latent Dirichlet Allocation (LDA) to identify latent topics within the textual data.

Activities: Implement LDA with parameters like the number of topics, random state, chunk size, and passes. Calculate perplexity and coherence score to evaluate the model's performance.

Topic	Words
16	"share," "quarterly," "ordinary," "standalone," "income," "result," "items," "bank," "crores," "total"
26	"contract," "agreement," "company," "service," "product," "term," "party," "time," "period," "court"
17	"court," "amount," "section," "employee," "scheme," "word," "date," "application," "high," "pension"
25	"fraud," "united," "states," "enforcement," "recovery," "group," "victim," "attorneys," "agency," "mortgage_fraud"
21	"contract," "statement," "agreement," "account," "time," "business," "party," "bank," "service," "asset"
2	"borrower," "bank," "loan," "lender," "agreement," "time," "interest," "security," "amount," "date"

Table 1.1 Topic and Words

Document	Words
1	'share', 'HDFC', 'Bank', 'Consolidated', 'Quarterly', 'Share', 'total', 'result', 'Extra', 'Ordinary'
16	'Agency', 'shall', 'Agreement', 'Authority', 'Party', 'obligation', 'Documentation', 'mean', 'Research', 'Contract'
21	'section', 'proceeding', 'Section', 'corporate', 'debtor', 'criminal', 'would', 'provision', 'word', 'case'
28	'shall', 'Registrar', 'financial', 'Court', 'account', 'rule', 'States', 'Parties', 'period', 'Assembly'
35	'year', 'include', 'Company', 'sale', 'increase', 'cash', 'impact', 'share', 'value', 'rate'
47	'Borrower', 'Bank', 'shall', 'borrower', 'agreement', 'time', 'Asset', 'security', 'loan', 'interest'

Table 1.2 Document and Words

Stage 7: Document Topic Distribution

Objective: Explore the distribution of topics for each document.

Activities: Print the topics assigned to each document along with their corresponding probabilities.

Stage 8: Probability Visualization

Objective: Visualize the probability of each document having topics related to balance sheets, contract agreements, loan agreements, financial reports, and court cases related to financial fraud

.

Activities: For each document, print the probability distribution of topics in a bar chart, representing the likelihood of each document being associated with different topics.

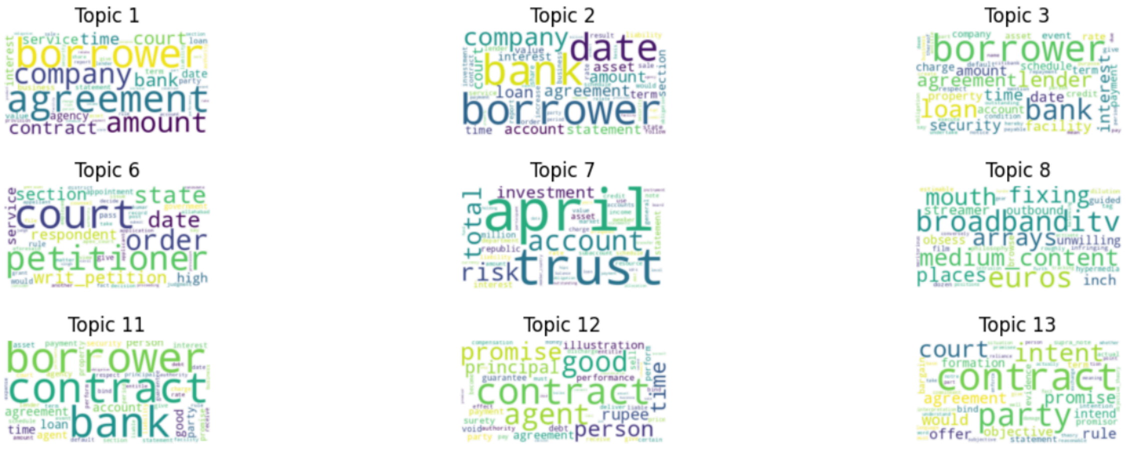


Figure 4.1 Topic-wise Word Cloud

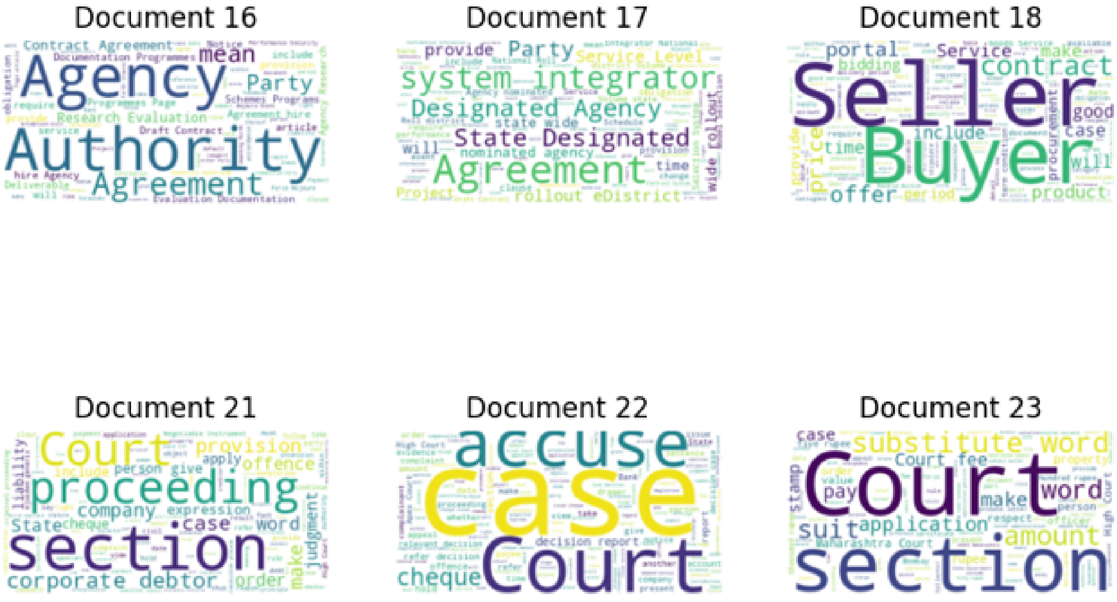


Figure 4.2 Document-wise Word Cloud

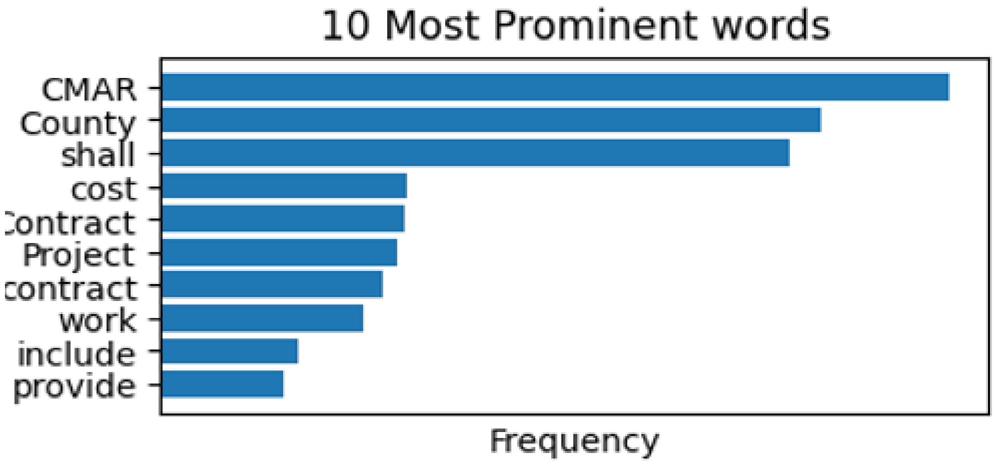


Figure 4.3 Top 10 Most Prominent words - Document 15

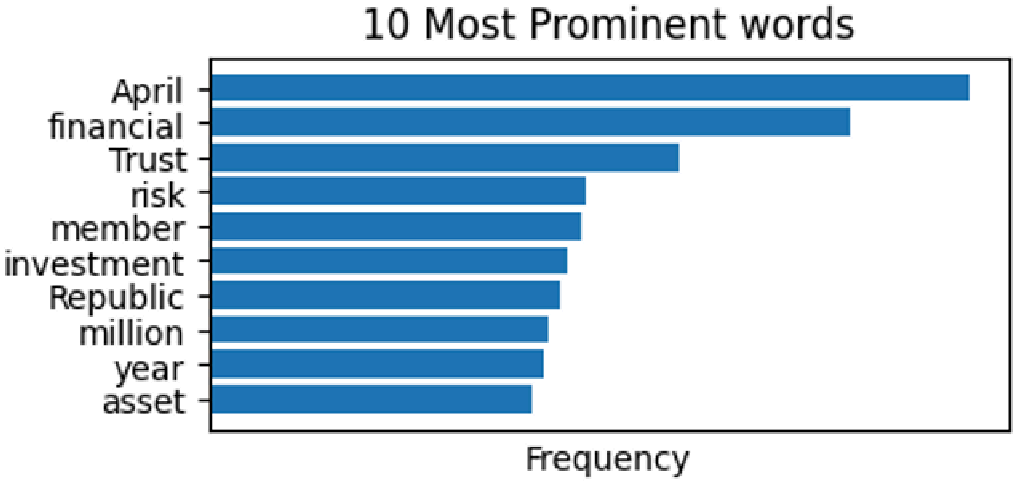


Figure 4.3 Top 10 Most Prominent words - Document 31

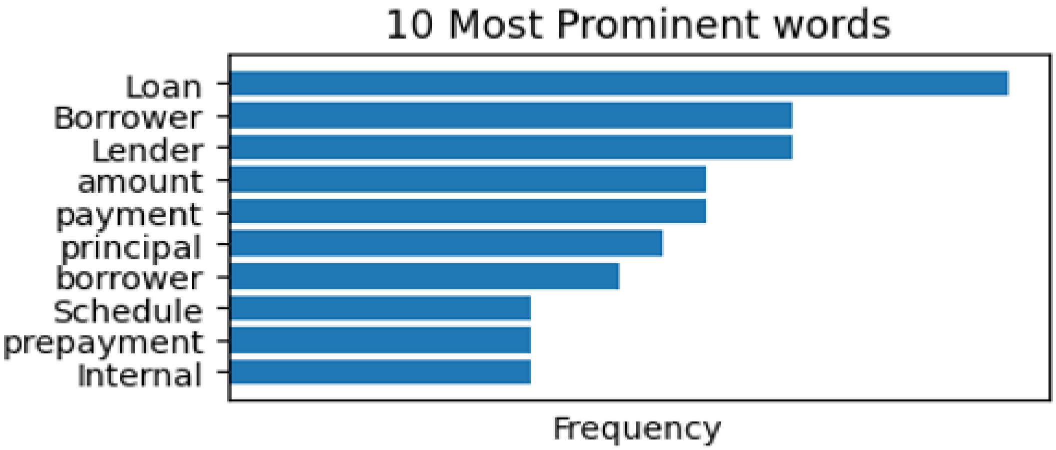
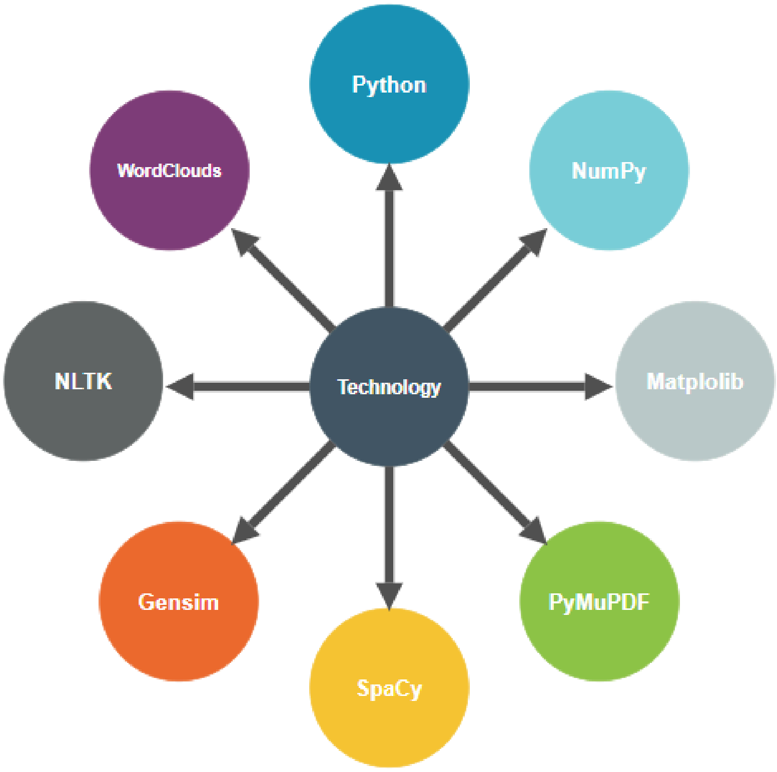


Figure 4.3 Top 10 Most Prominent words - Document 49

CHAPTER 5
TECHNOLOGY USED

5.1 Technology Stack



5.1.1 Python

Python is a high-level, dynamically-typed, and interpreted programming language renowned for its simplicity, readability, and versatility. Created by Guido van Rossum and first released in 1991, Python has evolved into a popular language for various applications, including web development, data analysis, artificial intelligence, and machine learning.

In the realm of Natural Language Processing (NLP), Python plays a pivotal role due to its rich ecosystem of libraries and frameworks. The language's simplicity and readability make it an ideal choice for processing and manipulating textual data. Libraries like NLTK (Natural Language Toolkit) and SpaCy provide powerful tools for tasks such as tokenization, part-of-speech tagging, and sentiment analysis. Additionally, Python is widely used in topic modeling techniques, where algorithms like Latent Dirichlet Allocation (LDA) are implemented using packages like Gensim.

In Machine Learning (ML), Python's popularity is unmatched. Libraries like Scikit-learn offer a comprehensive set of tools for tasks ranging from classification and regression to clustering and dimensionality reduction. TensorFlow and PyTorch, two prominent deep learning frameworks, are extensively used for building and training neural networks.

5.1.2 NumPy

NumPy, short for Numerical Python, is a fundamental library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to operate on these arrays. NumPy serves as the foundation for various scientific and data-related libraries, making it an essential tool in the Python ecosystem.

In the realm of Natural Language Processing (NLP) and Machine Learning (ML), NumPy plays a crucial role. NLP often involves working with matrices representing word embeddings, document-term matrices, or other numerical representations of language data. NumPy's capabilities in handling these numerical operations efficiently contribute to the performance and scalability of NLP algorithms.

5.1.3 Matplotlib

Matplotlib is a powerful and versatile Python library for creating static, animated, and interactive visualizations in a variety of formats. Developed by John D. Hunter in 2003, Matplotlib has become a cornerstone in the Python data visualization ecosystem. It provides an extensive array of plotting functions, enabling users to generate high-quality charts, plots, histograms, and more.

In the realm of Natural Language Processing (NLP) and topic modeling, Matplotlib plays a crucial role in visualizing and interpreting complex textual data. Specifically, when using techniques like Latent Dirichlet Allocation (LDA) for topic modeling, Matplotlib allows researchers and practitioners to present the results in a comprehensible and visually appealing manner.

Researchers often leverage Matplotlib to create bar charts, word clouds, and other visualizations that help convey the distribution of topics, the prevalence of certain terms, or the relationships between different elements in a textual corpus. These visualizations aid in understanding the underlying patterns and trends within the data, making it easier to communicate findings and insights.

5.1.4 PyMuPDF

PyMuPDF, also known as Fitz, is a Python library for interacting with PDF documents. It provides a range of functionalities for reading, accessing, and manipulating PDF files. PyMuPDF is particularly notable for its efficiency in handling large PDF documents and its support for various features like text extraction, image extraction, and annotations.

In the context of Natural Language Processing (NLP) and topic modeling, PyMuPDF becomes a valuable tool for extracting text content from PDF documents. Many legal, academic, and business documents are stored in PDF format, and topic modeling often involves analyzing and categorizing information from such files. PyMuPDF enables NLP practitioners to preprocess and extract text data from PDFs, making it accessible for subsequent analysis using topic modeling techniques.

NLP tasks, such as topic modeling, benefit from PyMuPDF's capabilities in converting PDF documents into a format suitable for text analysis. By leveraging PyMuPDF in the preprocessing phase, researchers and analysts can seamlessly integrate PDF content into their NLP pipelines, facilitating a more comprehensive understanding of topics present in a diverse range of documents.

5.1.5 SpaCy

SpaCy stands as a cornerstone in the realm of Natural Language Processing (NLP), contributing significantly to the efficiency and accuracy of text processing tasks. Renowned for its speed and precision, SpaCy offers a comprehensive suite of linguistic tools and pre-trained models, making it a preferred choice for developers and researchers alike.

In the context of topic modeling within NLP, SpaCy's capabilities shine through in its ability to seamlessly handle various linguistic nuances. Its tokenization, part-of-speech tagging, and named entity recognition features provide a solid foundation for extracting meaningful information from text. SpaCy's user-friendly interface and extensive language support further enhance its utility, enabling practitioners to focus on the intricacies of topic modeling without getting bogged down by the complexities of language processing.

As an integral component in the NLP toolkit, SpaCy empowers users to build robust and accurate topic models by efficiently preprocessing and analyzing textual data. Its versatility, combined with its emphasis on performance, positions SpaCy as a valuable asset in the pursuit of uncovering patterns and insights within large text corpora.

5.1.6 Gensim

Gensim stands as a pivotal player in the realm of Natural Language Processing (NLP), particularly renowned for its prowess in topic modeling. As a robust and user-friendly Python library, Gensim facilitates the implementation of unsupervised machine learning algorithms for extracting semantic structures from large text corpora. Its versatility shines prominently in the context of topic modeling, where it plays a crucial role in discerning latent thematic patterns within documents.

Gensim employs various algorithms, including Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), providing practitioners with a suite of tools to explore and analyze textual data. Its intuitive interface and seamless integration make it a preferred choice for researchers, data scientists, and developers working on tasks such as document similarity, keyword extraction, and topic identification. Gensim's impact in the field is underscored by its ability to transform unstructured text into meaningful insights, contributing significantly to advancements in natural language understanding and document clustering.

5.1.7 NLTK

Natural Language Toolkit (NLTK) is a comprehensive library for Python that serves as a powerhouse in the domain of Natural Language Processing (NLP). Developed to facilitate the exploration and manipulation of human language data, NLTK offers a broad spectrum of tools and resources, making it a go-to solution for linguistic research, education, and application development.

In the context of Natural Language Processing, NLTK plays a pivotal role in various tasks, including topic modeling. Its rich suite of modules encompasses functionalities for tokenization, stemming, lemmatization, and part-of-speech tagging, essential components in extracting meaningful insights from text data. NLTK contributes to the preprocessing phase, aiding in the transformation of raw textual data into a format suitable for algorithms like Latent Dirichlet Allocation (LDA). Its versatility and user-friendly design make NLTK an indispensable asset, bridging the gap between complex linguistic analysis and practical applications in natural language understanding and information retrieval.

5.1.8 WordClouds

WordClouds serve as captivating visual representations of textual data, contributing both aesthetic appeal and analytical insights to the field of Natural Language Processing (NLP). In the context of NLP, particularly in the realm of topic modeling, WordClouds offer a succinct and visually engaging summary of the most frequent words within a given corpus. As a powerful tool in topic modeling, WordClouds facilitate a quick understanding of the dominant subjects, aiding researchers, analysts, and enthusiasts alike in comprehending the essence of large volumes of text. The artful arrangement of words not only enhances interpretability but also makes the process of identifying and exploring topics within a corpus a visually intuitive experience. In essence, WordClouds seamlessly blend the artistic with the analytical, making them an invaluable asset in unraveling the intricacies of language and content in the vast landscape of NLP.

CHAPTER 6

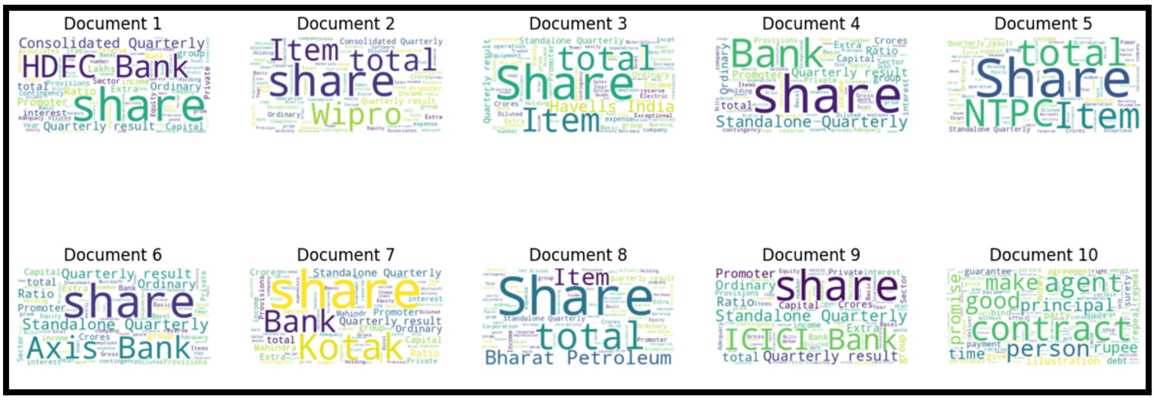
METHODOLOGY

CHAPTER 7
RESULT AND DISCUSSION

The results and discussions stemming from our project showcase the impactful utilization of Latent Dirichlet Allocation (LDA) and topic modeling in handling a substantial volume of files. Employing LDA, we transformed a large number of files into a coherent corpus, allowing for a more systematic and structured analysis of unstructured data. The application of topic modeling facilitated the extraction of key themes and topics within the documents. The generated graphs and charts, illustrating the most occurring words both topic-wise and document-wise, provide valuable insights into the underlying patterns present in the dataset. This visual representation proves instrumental in document classification, offering a comprehensive view of the prominent themes associated with each file.

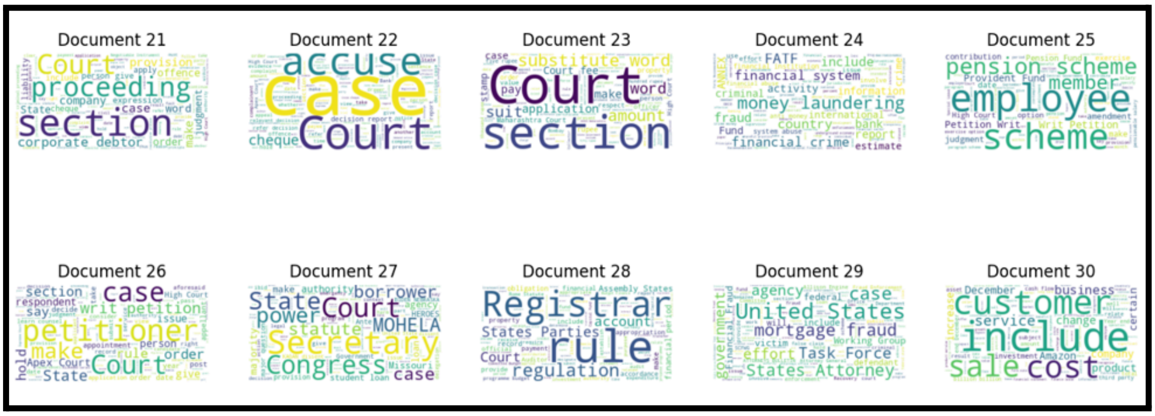
The software's ability to categorize documents based on the most prevalent topics not only streamlines the data but also enhances its utility for file tagging. This functionality is particularly advantageous for efficiently organizing and managing large datasets. Moreover, the software's proficiency in identifying suspicious files within the corpus adds a layer of security, making it a valuable tool for fraud detection and risk assessment. The visualizations generated through topic modeling serve as an intuitive means of understanding the content distribution, aiding users in making informed decisions regarding document prioritization and review. In essence, the combination of LDA and topic modeling proves to be a powerful approach in enhancing document analysis, classification, and detection of suspicious files, thereby contributing significantly to the overall efficacy of our software solution.

7.1 Outputs


Document 1 Document 2 Document 3 Document 4 Document 5

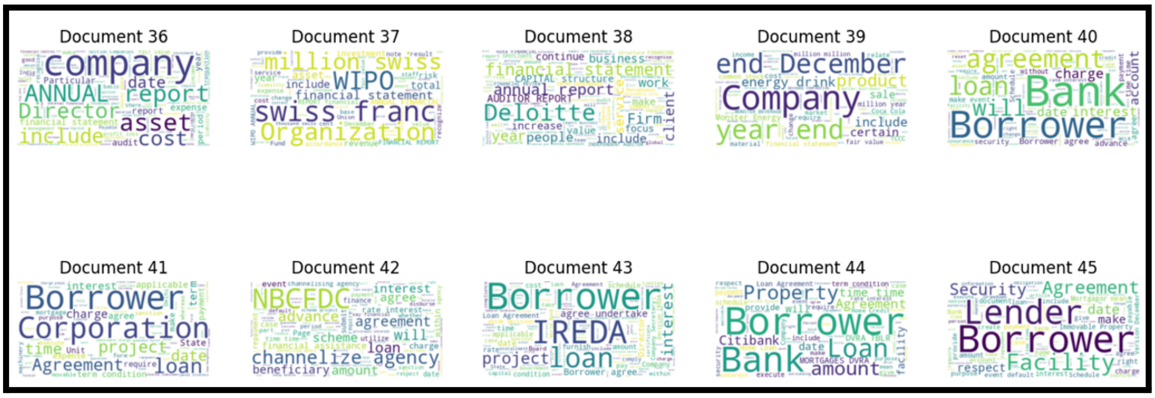
Document 6 Document 7 Document 8 Document 9 Document 10

Figure 7.1 Document-wise Word Cloud


Document 21 Document 22 Document 23 Document 24 Document 25

Document 26 Document 27 Document 28 Document 29 Document 30

Figure 7.2 Document-wise Word Cloud


Document 36 Document 37 Document 38 Document 39 Document 40

Document 41 Document 42 Document 43 Document 44 Document 45

Figure 7.3 Document-wise Word Cloud

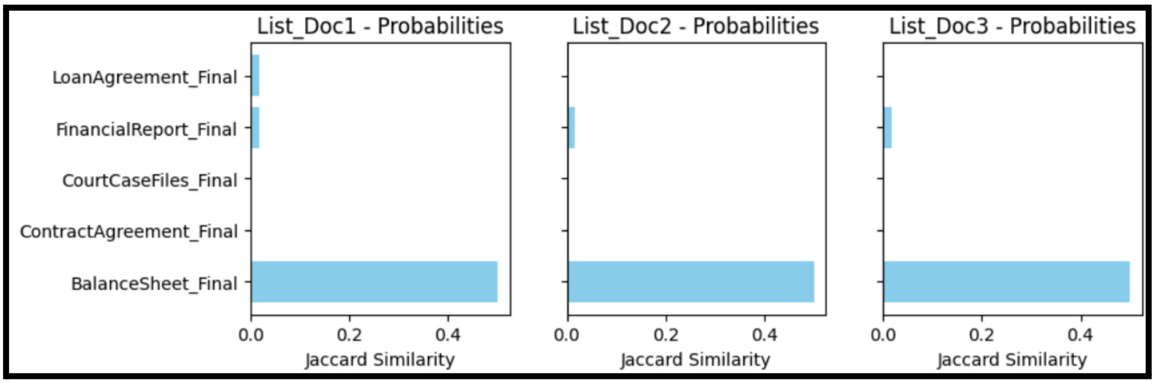


Figure 7.4 Identification of document type

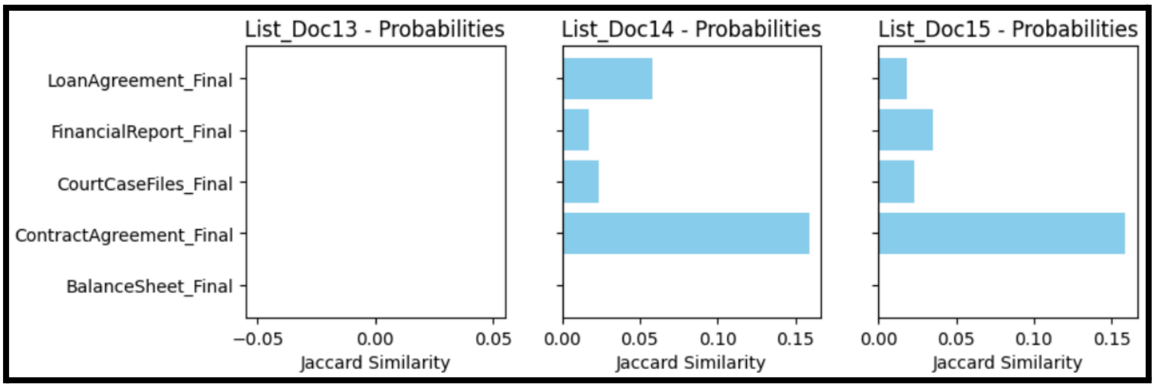


Figure 7.5 Identification of document type

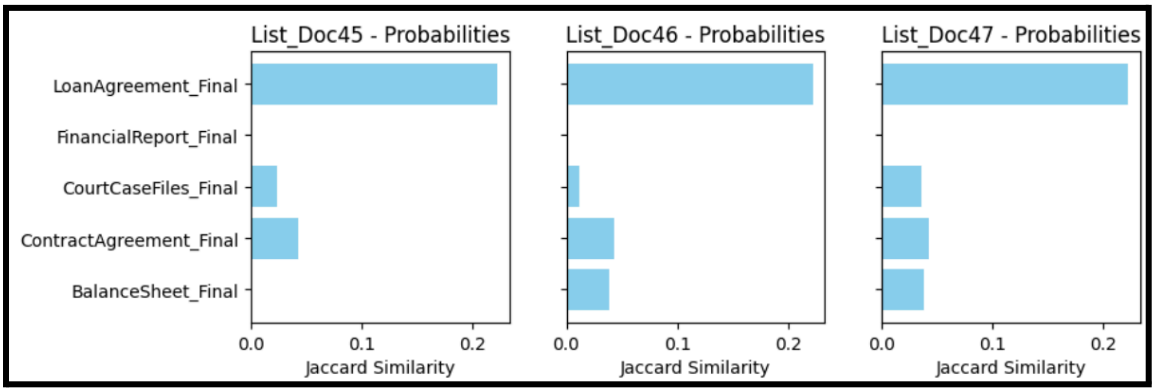


Figure 7.6 Identification of document type

CHAPTER 8

CONCLUSION

In conclusion, the presented project addresses a critical issue in the financial landscape—the inadequate methods for detecting potential financial frauds that pose a threat to the global financial system. The proposed solution, a sophisticated software application integrating **Natural Language Processing (NLP)** and **Machine Learning (ML)** techniques, aims to revolutionize fraud detection by systematically analyzing extensive unstructured data. Through a well-defined process involving data collection, preprocessing, NLP-driven feature extraction, and ML-based fraud detection and risk prioritization, the software demonstrates a comprehensive approach to identifying suspicious financial documents.

The primary objective is to proactively combat financial frauds, particularly those related to money laundering and tax evasion, by leveraging the capabilities of NLP and ML. The software's output includes a user-friendly interface for initiating scans, reviewing results, and presenting suspicious files with detailed explanations. The real-time scanning and alert mechanisms enhance the responsiveness to potential threats, allowing for immediate action upon the detection of suspicious files. Notably, the project places emphasis on optimizing the software's efficiency and scalability, ensuring its effectiveness in handling large corpora of files within a system. With its innovative approach and robust functionalities, this software represents a significant stride forward in bolstering the financial sector's defenses against evolving fraudulent schemes.

Chapter 9

References

1. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4317320

2.https://www.researchgate.net/publication/352145844_Deep_Learning_and_Explainable_Artificial_Intelligence_Techniques_Applied_for_Detecting_Money_Laundering-A_Critical_Review

3.https://www.researchgate.net/publication/322992807_Analysis_on_User_Interface_Aspects_of_Software_Used_by_Commercial_Banks_in_India

4.https://jfin-swufe.springeropen.com/articles/10.1186/s40854-020-00205-1

5. https://en.wikipedia.org/wiki/Natural_language_processing

6. G. V. Cormack and M. R. Grossman. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In SIGIR 2014.

7. Scott Deerwester, Susan T. Dumais, George W. Firmas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391-407, September 1990.

8. Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR ’99). Association for Computing Machinery, New York, NY, USA, 50–57.
DOI:https://doi.org/10.1145/312624.312649

9. Dimo Angelov. Top2Vec: Distributed Representations of Topics. arXiv:2008.09470v1, (2020).

10. A. Kanapala, S. Pal, R. Pamula. Text summarization from legal documents: a survey. Artificial Intelligence Review 51(3), 371–402 (2019).

11. Ylja Remmits. Finding the Topics of Case Law: Latent Dirichlet Allocation on Supreme Court Decisions. Thesis. Radboad Universiteit, (2017).

12. Pedro Henrique Luz de Araújo, and Teófilo de Campos. Topic Modelling Brazilian Supreme Court Lawsuits. JURI SAYS, 113, (2020).

13. James O’ Neill, Cécile Robin, Leona O’ Brien, Paul Buitelaar. An Analysis of Topic Modelling for Legislative Texts. ASAIL 2017, London, UK, June 16, (2017).

14. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003.

15. Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the

22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999.

16. Jordan L Boyd-Graber and David M Blei. Syntactic topic models. In Advances in neural information processing systems, pages 185–192, 2009.

17. S. Syed and M. Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 165–174, 2017.

18. Kai Yang, Yi Cai, Zhenhong Chen, Ho-fung Leung, and Raymond Lau. Exploring topic discriminating power of words in latent dirichlet allocation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2238–2247, 2016.

19. Geoffrey E Hinton et al. Learning distributed representations of concepts. In Proceedings of the eighth annual conference of the cognitive science society, volume 1, page 12. Amherst, MA, 1986.