

PROJECT REPORT ON

“Detection of Fraud using Topic Modelling techniques on text data”

A Project Work Submitted in Partial Fulfilment of the requirements
for

The Course

ALGORITHMS FOR INTELLIGENCE WEB AND INFORMATION RETRIEVAL

by

Aronya Baksy (PES1201800002)

Ansh Sarkar (PES1201800275)

Vishesh P (PES1201800314)

Under the supervision of

Prof. Nagegowda K S

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

PES UNIVERSITY

RR CAMPUS

TABLE OF CONTENTS:

1. INTRODUCTION	1
2. OVERVIEW OF THE PROJECT	1
3. ALGORITHM PSEUDO CODE	3
4. RESULTS AND DISCUSSION	7
5. CONCLUSION AND FUTURE SCOPE	11
6. REFERENCE & SOURCES	12

1 Introduction

1.1 Abstract of the Project

The Securities and Exchange Commission (SEC) of the United States government is a regulatory body that was established in 1933 in order to protect financial markets against market manipulation and enforce laws regarding the same.

The SEC publishes **press releases** that highlight recent actions taken by the SEC as well as other newsworthy information regarding financial markets in the USA. These press releases are publicly available on the SEC's official website [here](#). These can be accessed and retrieved in a year-wise fashion.

In our current scenario, where so much of financial activity worldwide takes place online, the press releases put out by the SEC at regular intervals are an important source of data regarding financial fraud. Text mining techniques have been used for a long time to detect instances of financial fraud in text data, most notably during the Enron financial scandal and the release of internal office communications on the same [1].

Text mining techniques such as TF-IDF based cosine-similarity models, and topic modelling techniques such as Latent Semantic Indexing (LSA) and its advanced cousin, the Latent Dirichlet Allocation (**LDA**) are used as techniques to extract semantic relations between topics. All of these models are unsupervised techniques that require human interpretation in order to produce useful results.

1.2 Problem Statement

The problem statement of this project is to be able to detect instances of financial fraud using text mining techniques, from the corpus of press releases output by the SEC on its website.

Using LDA, TF-IDF and cosine similarity as the selected approaches, we aim to be able to classify documents in the corpus as containing information about financial fraud or not, as well as allow a user to search for more specific instances of financial fraud using free-text queries.

2 Project Overview

2.1 Proposed Approach

2.1.1 Data Gathering

The first step in this project is the gathering of the data from the SEC's website into a usable format. This is achieved using a custom web crawler designed in Python using the `requests` library, and the `BeautifulSoup` HTML Parser.

The `requests` module is used to retrieve the HTML response from the server, and the `BeautifulSoup` module is used to parse the HTML, and extract the article text, as well as the metadata regarding the article, such as the place and date of publishing, the title and the unique identifier for the article.

Once gathered, the data is placed into a hierarchical directory structure. The root Data directory contains sub-folders organized by the year of the press release that are named in the format Year_20xx where 20xx is the 4-digit year.

The articles are stored in .txt format which is readable across platforms and carries minimal overhead, and hence is most efficient in storing in terms of disk space utilization.

2.1.2 Data Cleaning

The gathered text data must be cleaned, i.e. pre-processed into an appropriate and usable form so that useful information can be extracted from it.

The first step of cleaning is sentence tokenization, which is an algorithmic procedure for converting a body of text into a list of *sentences*. Each sentence is further tokenized into *words*.

Now that the corpus is tokenized into words, case folding is carried out, wherein the entire corpus is converted to lowercase.

The tokenized and case-folded corpus is put through a Parts-of-Speech Tagger (PoS Tagger) that outputs the token as well as the corresponding part of speech (verb, noun, adverb, adjective, etc.) that it belongs to. This extra information is useful for the next cleaning step.

The PoS-tagged text is then put through a *lemmatizer*. Lemmatization is a process by which various forms (i.e. verb forms, adjective forms etc.) are collected into a single base form. Lemmatization is different from stemming in its use of morphological and dictionary information in comparison to stemming which is a purely algorithmic approach.

2.1.3 Text Mining Approach 1: TF-IDF

This approaches uses a TF-IDF based approach to rank documents based on their similarity to a given query, and output the most relevant documents based on this score. The similarity metric used in this case is the **cosine similarity** which measures the semantic similarity of two entities as the cosine of the angle between their vector representations.

2.1.4 Text Mining Approach 2: LDA

First, each document in the corpus of documents is labelled manually to indicate whether it contains an instance of financial fraud (the label is a binary value, with a value of 1 indicating an instance of financial fraud and a value of 0 indicating otherwise).

The basic intuition behind an LDA topic model is that each word in each document comes from a topic and the topic is selected from a per-document distribution over topics. The topic that covers the most number of terms is considered to be the dominant topic, and for each document that is related to this dominant topic, the predicted label is said to be 1, and 0 for all other documents.

3 Algorithms and Pseudocode

3.1 Search Engine based on TF-IDF

After cleaning all the documents in our dataset, a matrix consisting of numerical values is constructed to represent all the documents. This matrix is used to determine the similarity between any pair of documents and between a user provided query and any document. This matrix is called the Term-Document Matrix. The rows of this matrix represent the unique terms collected from the documents and the columns represent all the documents in the form of numerical identifiers. The TF-IDF method is used to calculate the value inside every cell in the matrix. TF-IDF consists of two parts namely the Term-Frequency(TF) and the Inverse Document Frequency(IDF).

Term-Frequency(TF) of a term t is a function of the frequency of the term on a document d .

$$tf(t, d, D) = \frac{f_{t,d}}{\sum_{t' \in D} f_{t',d}} \quad (1)$$

The Inverse Document Frequency of a term acts as a weight for the TF value calculated earlier. The IDF value of a term is inversely proportional to the frequency of the word across the set of documents.

$$idf(t, D) = \log_{10} \left(\frac{|D|}{|\{d \in D : t \in d\}|} \right) \quad (2)$$

The two values for each term are multiplied to get the cell value of that particular term in the corresponding document. This process is performed on all the cells to get the Term-Document Matrix. Once the Term Document Matrix has been constructed, the next step is to determine the similarity between a search query provided by the user and every single document. To achieve this, the method of Cosine Similarity is implemented. The Cosine Similarity between two vectors is calculated by taking the dot product of the two vectors and dividing it by the product of the lengths of both the vectors. The value of the Cosine function ranges from -1 to 1 usually. However, the values in our query vector and Term-Document Matrix are always greater than or equal to 0. Hence the value of the similarity measure in this case always ranges from 0 to 1.

$$\text{cosineSimilarity}(q, d) = \frac{\vec{q} \cdot \vec{d}}{\|\vec{q}\| \|\vec{d}\|} \quad (3)$$

The query q is first transformed as a vector on the Term-Document Matrix. Then the pair-wise similarity between the query vector and every document's vector d is calculated and sorted in descending order. Finally, we retrieve the required number of documents all having similarity greater than 0 with the query vector.

3.1.1 Pseudocode

The pseudocode below details the calculation of the top N most similar documents that are returned by the system given an user query q and a corpus of documents D .

Algorithm 1 Tf-IDF Similarity Algorithm

```
procedure GETSIMILARDOCUMENTS(D: corpus, query, N)
    vocab ← {}
    for document d ∈ D do
        for term t ∈ d do
            if t ∉ vocab then
                vocab.add(t)
    TfIDFModel ← TFIDFVectorizer(vocabulary)
    TfIDFModel.fit(D)
    q ← TfIDFModel.fit(query)
    documentScores ← ⟨⟩
    for v ∈ TfIDFModel do
        similarity = cosineSimilarity(v, q)
        if similarity > 0 then
            documentScores.add(docID, similarity)
    topNdocs ← sort(documentScores, on=similarity)[:N]
    return topNdocs
```

3.2 LDA-based classification

Latent Dirichlet Allocation (LDA) is a “generative probabilistic model” of a collection of composites made up of parts. It is a form of unsupervised learning that views documents as bags of words i.e order does not matter. This algorithm was used for topic modelling on our cleaned data. Topic modelling is the process of identifying topics in a set of documents. In terms of topic modelling, the composites are documents and the parts are words and/or phrases. This is essentially a clustering problem where both words and documents can be thought as being clustered.

LDA model has a few assumptions behind it:

- Documents exhibit multiple topics (but typically not many)
- LDA is a probabilistic model with a corresponding generative process – each document is assumed to be generated by this (simple) process
- A *topic* is a distribution over a fixed vocabulary – these topics are assumed to be generated first, before the documents
- Only the number of topics is specified in advance.

LDA works by first making a key assumption: the way a document was generated was by picking a set of topics and then for each topic picking a set of words. To find topics LDA simply reverse engineers this assumption.

The process can be defined for each document as follows:

- Assume there are k topics across all of the documents
- Distribute these k topics across document m (this distribution is known as α and can be symmetric or asymmetric) by assigning each word a topic.
- For each word w in document m , assume its topic is wrong but every other word is assigned the correct topic.
- Probabilistically assign word w a topic based on two things:
 - what topics are in document m .
 - how many times word w has been assigned a particular topic across all of the documents (this distribution is called β).
- Repeat this process a number of times for each document.

3.2.1 Dirichlet Distribution

The distribution described in the above process is called the Dirichlet distribution. [Need to add some info pls help].

The α and β values are defined as following:

- α is the per-document topic distributions
- β is the per-topic word distribution,

The distribution can be defined as follows:

$$p(x|\alpha) = \frac{\Gamma(\sum_{i=1}^d \alpha_i)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_{i=1}^d x_i^{\alpha_i - 1} \quad (4)$$

for observations

$$\sum_{i=1}^d x_i = 1, \quad x_i \geq 0$$

where $\Gamma()$ is the Gamma distribution

3.2.2 Formal look at LDA algorithm

The following definitions are put in place for the definition of the LDA algorithm.

- $\beta_{1:K}$ are the K topics where each β_k is a distribution over the vocabulary
- θ_d is the topic proportions for each of the K topics in document d .
- $\theta_{d,k}$ is the topic proportion for topic k in document d
- z_d are the topic assignments for document d .
- $z_{d,n}$ is the topic assignment for word n in document d

- w_d are the observed words for document d

Then the joint distribution of the hidden and observed variables is given as below:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D \theta_d \prod_{n=1}^N p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \quad (5)$$

3.2.3 Pseudocode

Algorithm 2 LDA Fraud Classification

```

procedure CLASSIFYFRAUD(D: corpus, query, N)
    vocab ← {}
    for document d ∈ D do
        for sentence s ∈ d do
            vocab.add(s)
    LDAModel ← LDAModel(vocabulary, numtopics = 5)
    documentTopics ← ⟨⟩
    for document d ∈ D do
        t ← getTopicwithHighestProp(LDAModel, d)
        documentTopics.add((d, t))
    fraud_dominant_topic ← LDAModel.getLargestTopic()
    unseenDocument ← testSet.getDocument()
    if unseenDocument.dominantTopic == fraud_dominant_topic then
        Label = 1
    else
        Label = 0

```

4 Results and Discussion

4.1 Output Screenshots 1

```
Aronya@ARONYA ~\..\..\sec-analysis> python .\search_engine.py
Enter search query:fraud
Query: fraud
RESULTS:
Similarity: 0.1749226922370378
Article ID: 2026-323
Similarity: 0.07232195881822201
Article ID: 2026-41
Similarity: 0.03923593816446632
Article ID: 2026-18
Similarity: 0.03920884914623244
Article ID: 2026-181
Length of raw_text_list is 4
WRITING 2026-239 to file
The Securities and Exchange Commission today charged a Swedish national living in Thailand with conducting a multi-million dollar online offering fraud that victimized thousands of retail investors worldwide, including hundreds of investors from the Deaf, Hard of Hearing, and Hearing Loss communities. The SEC's complaint alleges that from November 2012 to June 2019, Roger Nils-Jonas Karlsson, through his entity, Eastern Metal Securities, defrauded over 2,000 retail investors in nearly every state in the United States, as well as in over 45 countries around the world. According to the complaint, Karlsson solicited investors for what he described as a "Pre Funded Reversed Pension Plan," falsely claiming that the investment platform was run by award-winning economists and promising returns based on the value of gold. Karlsson allegedly claimed that the investments had no risk of loss. At least some of the investors were members of the community for the last three years, invested more than $100,000 in Eastern Metal Securities, and claimed to be deaf, retired, or disabled. "We are all about what Karlsson raised," said the SEC's New York Regional Office, "and we are committed to fighting securities fraud that targets our country's most vulnerable communities," said Richard R. Best, Director of the SEC's New York Regional Office. "As alleged in the complaint, Karlsson's scheme jeopardized the hard-earned savings of thousands of retail investors." The SEC alleges that Karlsson violated the registration provisions of Sections 5(a) and 5(c) of the Securities Act of 1933 and the antifraud provisions of 17(a)(1) and 17(a)(3) of the Securities Act and Section 10(b) of the Securities Exchange Act of 1934 and Rules 10b-5(a) and 10b-5(c) thereunder, and seeks permanent injunctions, disgorgement with prejudgment interest, and a civil penalty. The case was investigated by Karen Lee Schleman, Director of the SEC's New York Regional Office, and the case is being supervised by Sanjay Wadhwani, Head of the SEC's New York Regional Office. The Retail Strategy Task Force's litigation will be led by Richard Hong, and the case will be overseen by Sanjay Wadhwani and Karen L. Chaudhury, Chair of the Division of Enforcement's Retail Strategy Task Force. The litigation will be led by Richard Hong, and the case is being supervised by Sanjay Wadhwani, Head of the SEC's New York Regional Office. The Retail Strategy Task Force encourages investors in the Deaf, Hard of Hearing, and Hearing Loss communities to learn more about how to spot frauds in their communities and how to protect themselves and others from investment fraud through the Task Force's investor outreach video "For the Deaf, Hard of Hearing, and Hearing Loss communities. The SEC appreciates the assistance of the U.S. Attorney's Office for the Northern District of California, the Internal Revenue Service, the securities and financial markets regulatory authorities in Austria, Finland, France, Hong Kong, Malaysia, Singapore and Thailand, and the National Bureau of Investigation of the United States.
WRITING 2026-41 to file
The Securities and Exchange Commission today announced that G. Jeffrey Boujoukos, the Director of its Philadelphia Regional Office, will leave the agency at the end of the month after nearly 11 years of service. Mr. Boujoukos has served at the helm of the SEC's Philadelphia office since December 2016, overseeing the agency's enforcement and examinations in the region. Before his appointment to Regional Director, Mr. Boujoukos supervised the enforcement program in the Philadelphia office as an Associate Director, and managed the Philadelphia trial unit as Regional Trial Counsel. Since late 2014, Mr. Boujoukos has led a staff of approximately 30 enforcement attorneys, accountants, investigators, and compliance examiners who investigate and enforcement financial institutions, funds, and broker-dealers. The Philadelphia office has overseen investigations involving more than 200 broker-dealers, mutual fund complexes, and other financial companies with over $10 trillion in assets under management, over 65 mutual fund complexes, and over 240 broker-dealers with over 14,500 branch offices. "Jeff has been an exemplary leader in the Commission's commitment to investors. Jeff has been by example in fighting fraud and educating our Main Street investors," said SEC Chairman Jay Clayton. "His dedication to our mission and to the Philadelphia Regional Office will have a lasting impact in Philadelphia and beyond, and I thank him for his counsel to me and his many years of service to the Commission." "Jeff has made great contributions to the SEC's mission and leaves behind a legacy of accomplishments," said Steven Peikin, Co-Director of the SEC's Division of Enforcement. "Under Jeff's leadership, the Philadelphia Regional Office has continued to bring innovative ideas and focus on the challenges facing Main Street investors. Jeff is an exceptional attorney and an even better colleague," said Stephanie Avakian, Co-Director of the SEC's Division of Enforcement. "Jeff's substantial experience and enthusiasm greatly benefit the agency and the Philadelphia office. We will miss him." "Jeff has been a tremendous advocate for investors during his time at the Philadelphia Regional Office," said Peter B. Driscoll, Director of the SEC's Office of Compliance Inspections and Examinations (OCIE). "Jeff's thoughtfulness and leadership on examination issues and investor outreach greatly advanced OCIE's mission of protecting investors. We thank him for all of his efforts." Mr. Boujoukos said, "It has been an honor to come to work each day with the staff of the Philadelphia Regional Office and the SEC. Their dedication, collaboration and unwavering commitment to the SEC's mission is extraordinary. I am proud of the work we have done together and grateful for the opportunity to be part of the SEC's culture of excellence." Under Mr. Boujoukos' supervision and leadership, the Philadelphia Regional Office has brought numerous groundbreaking enforcement actions that have protected Main Street investors and involved a variety of securities law violations, including: Mr. Boujoukos has also spearheaded the SEC's outreach efforts to retail investors i
WRITING 2026-219 to file
```

Figure 1: Output for the query **fraud**

```
Aronya@ARONYA ~\..\..\sec-analysis> python .\search_engine.py
Enter search query:financial fraud whistleblower
Query: financial fraud whistleblower
RESULTS:
Similarity: 0.37522967503475313
Article ID: 2026-219
Similarity: 0.06068592192238791
Article ID: 2026-323
Similarity: 0.04055229638984
Article ID: 2026-41
Similarity: 0.029863659199232
Article ID: 2026-18
Similarity: 0.02986365912636111616
Article ID: 2026-181
Similarity: 0.015742705796513805
Article ID: 2026-86
Similarity: 0.0066857198397992523
Article ID: 2026-264
Similarity: 0.00668571772703595
Article ID: 2026-191
Length of raw_text_list is 8
WRITING 2026-219 to file
The Securities and Exchange Commission today voted to adopt amendments to the rules governing its whistleblower program that are designed to provide greater clarity to whistleblowers and increase the program's efficiency and transparency. Concurrently, to protect additional investors, the SEC is adopting rules to enhance the clarity and transparency in the internal review process. The SEC's Office of the Whistleblower has provided guidance regarding the process for internalizing and accounting for eligible whistleblower claims. The SEC's whistleblower program was created to incentivize individuals to report high-quality tips to the Commission and assist the agency in its efforts to combat wrongdoing and, as a result, better protect investors and the marketplace. Since the program's inception ten years ago, whistleblowers have made a significant impact on the Commission's enforcement efforts and protection of investors. Original information provided by whistleblowers has led to enforcement actions in which the Commission has obtained over $2.5 billion in financial remedies, most of which has been, or is scheduled to be, returned to harmed investors. For the whistleblower, in the last three and a half years, the agency has made the first two largest awards in the program's history, two in excess of $1 million, and one each in the amount of $1.5 million, and $1.3 million. It is also across the program at which it is both processing claims and awarding amounts. This year, so far, even with the challenges posed by COVID-19, the Commission has processed more claims than in any previous year. "The Commission's enforcement efforts, and most importantly, American investors and markets, have greatly benefitted from the credible information and assistance that whistleblowers have provided," said SEC Chairman Jay Clayton. "Whistleblowers often take professional and reputational risks in reporting the information to the SEC and we are committed to rewarding them for taking those risks and contributing to our enforcement efforts. Today's rule amendments will help us get more money in to the hands of whistleblowers, and at a faster pace. Experience demonstrates this added clarity, efficiency and transparency will further incentivize whistleblowers, enhance the whistleblower culture, and reduce the cost of enforcement. The amendments will also facilitate the internal review process, which is critical to the success of the program and will help to expand and bolster the program in several ways. The rule amendments increase efficiencies around the review and processing of whistleblower award claims, and provide the Commission with additional tools to appropriately reward meritorious whistleblowers for their efforts and contributions to a successful matter. Among other enhancements, the amendments provide a mechanism for whistleblowers with potential awards of less than $5 million (which historically have represented nearly 75% of all whistleblower awards), subject to certain criteria, to qualify for a presumption that they will receive the maximum statutory award amount. Other awards will continue to be evaluated consistent with past practice. The amendments also affirm that award amounts and the percentage of discretion can determine the percentage tier, dollar amount, and complexity of the award. The awards will be determined by applying a formula based on the application of the award factors set forth in the Commission's whistleblower rules. In other words, it is not separate (post-application of the award factors) assessment of whether award amounts are too large. The amendments further clarify that the Commission may waive compliance with the TCR filing requirements if a whistleblower complies with the requirements within 30 days of first providing the information of first obtaining actual or constructive notice of the TCR filing requirements. The whistleblower rule amendments will become effective 30 days after publication in the Federal Register. *** FACT SHEET SEC Open Meeting September 23, 2020 Background Section 922 of the Dodd-Frank Wall Street Reform and Consumer Protection Act added Section 21 F to the Securities Exchange Act of 1934 (the "Exchange Act"), establishing the Commission's whistleblower program. Among other things, Section 21F authorizes the SEC to make monetary awards to eligible individuals who voluntarily provide original information that leads to successful SEC enforcement actions resulting in monetary sanctions over $1 million. Awards must be made in an amount equal to not less than 10 percent, and not more than 30 percent, of the monetary sanctions collected in the covered SEC action and certain related actions. The amendment clarifies that the form of an action—e.g., settlement agreements, deferred prosecution agreements (DPAs) and non-prosecution agreements (NPAs)—will not affect whether the action is a cover
```

Figure 2: Output for the query **financial fraud whistleblower**

4.2 Description of results of Tf-IDF based retrieval system

From the above results, it is clear that the system is able to retrieve relevant documents with reasonable accuracy when queried. The vocabulary of the document set is a limi-

```

C:\_VCS332 (IR)\sec-analysis >..\..\sec-analysis> main # +0 -1 -0 !
python .\search_engine.py
Enter search query:litigation
['litigation']
RESULTS:
Similarity: 0.0719191481654177
Article ID: 2020-10
Similarity: 0.04134358934985446
Article ID: 2020-41
Similarity: 0.040366384955635146
Article ID: 2020-108
Similarity: 0.0341870952025822
Article ID: 2020-41
Length of raw_text list is 4
WRITING 2020-10 to file
The Securities and Exchange Commission today announced that it filed an emergency enforcement action and obtained a temporary restraining order and asset freeze against Illinois resident Kenneth D. Courtright, III and his company, Todays Growth Consultant Inc., in connection with an alleged Ponzi-like scheme that raised at least $75 million from more than 500 investors throughout the United States and abroad. According to the SEC's complaint, from at least 2017 through at least October 2019, TGC, which also operated under the name "The Income Store," and Courtright, the company's founder and current chairman, promised investors an endless minimum guaranteed rate of return on revenues generated by websites. In exchange for an investor's "upfront fee," TGC claimed that it would either buy or build a website for the investor, and develop market, and maintain the website. As alleged, TGC falsely promised that it would use investors' funds exclusively for expenses related to the investor's website. In reality, as alleged, the sales were conducted through unregistered securities offerings, and TGC used new investors' funds to pay Courtright's personal expenses, including his mortgage and private school tuitions for his family. "TGC and Courtright's alleged fraud promised a guaranteed return when the company's business model and financial condition could not possibly support it," said Antonia Chion, Associate Director in the SEC's Division of Enforcement. "To avoid further harm to investors and preserve the misused assets that have not already been dissipated, we have sought and obtained emergency relief." The SEC's complaint, filed in federal court in Chicago on Dec. 27, 2019, and unsealed on Jan. 14, 2020, charges Courtright and TGC with violations of the antifraud and registration provisions of the federal securities laws, and seeks certain emergency relief as well as permanent injunctions, return of ill-gotten gains with prejudgment interest, and civil penalties. On Dec. 30, 2019, the Court issued a temporary restraining order, ordered an asset freeze and other emergency relief, and appointed a receiver for TGC. Investors can learn more about Ponzi scheme red flags by using the SEC's Investor.gov website. Investors should be cautious any time there are promises of a guaranteed rate of return in perpetuity. The SEC's investigation was conducted by Michael Brennan, Patrick L. Feeney, Jeffrey Anderson, Donato Furlano, and Michi Harthcock, with assistance from Suzanne J. Romajas. The investigation was supervised by Kevin Guerrero, Peter Rosario, and Antonia Chion. The litigation will be led by Ms. Romajas and supervised by Stephan Schlegelmilch; Robert M. Moye will assist with the litigation. The SEC's investigation is continuing.
WRITING 2020-41 to file
The Securities and Exchange Commission today charged a Swedish national living in Thailand with conducting a multi-million dollar online offering fraud that victimized thousands of retail investors worldwide, including hundreds of investors from the Deaf, Hard of Hearing, and Hearing Loss communities. The SEC's complaint alleges that from November 2012 to June 2019, Roger Nils-Jonas Karlsson, through his entity, Eastern Metal Securities, defrauded over 2,000 retail investors in nearly every state in the United States, as well as in over 45 countries around the world. According to the complaint, Karlsson solicited investors for what he described as a "Pre Funded Reversed Pension Plan," falsely claiming that the investment platform was run by award-winning economists and promising a payout based on the value of gold. Karlsson allegedly claimed that the investment had no risk of loss. At least 847 of the investors were members of a community for the Deaf that invested more than $2 million in Eastern Metal Securities since 2015 as their retirement investment. The SEC alleges that Karlsson raised $3.5 million from December 2017 through June 2019, and misappropriated at least $1.5 million to purchase real estate in Thailand and for other personal expenses. "We are committed to fighting securities fraud that targets our country's most vulnerable communities," said Richard R. Best, Director of the SEC's New York Regional Office. "As alleged in the complaint, Karlsson's scheme jeopardized the hard-earned savings of thousands of retail investors." The SEC alleges that Karlsson violated the registration provisions of Sections 9(a) and 9(c) of the Securities Act of 1933 and the

```

Figure 3: Output for the query litigation

```

C:\_VCS332 (IR)\sec-analysis >..\..\sec-analysis> main # +0 -1 -0 !
python .\search_engine.py
Enter search query:investigation
['investigation']
RESULTS:
Similarity: 0.0712360312836769
Article ID: 2020-10
Similarity: 0.0284840853306311735
Article ID: 2020-108
Similarity: 0.022021406819921557
Article ID: 2020-181
Length of raw_text list is 3
WRITING 2020-10 to file
The Securities and Exchange Commission today announced that it filed an emergency enforcement action and obtained a temporary restraining order and asset freeze against Illinois resident Kenneth D. Courtright, III and his company, Todays Growth Consultant Inc., in connection with an alleged Ponzi-like scheme that raised at least $75 million from more than 500 investors throughout the United States and abroad. According to the SEC's complaint, from at least 2017 through at least October 2019, TGC, which also operated under the name "The Income Store," and Courtright, the company's founder and current chairman, promised investors an endless minimum guaranteed rate of return on revenues generated by websites. In exchange for an investor's "upfront fee," TGC claimed that it would either buy or build a website for the investor, and develop market, and maintain the website. As alleged, TGC falsely promised that it would use investors' funds exclusively for expenses related to the investor's website. In reality, as alleged, the sales were conducted through unregistered securities offerings, and TGC used new investors' funds to pay investor returns, in Ponzi-like fashion, and to pay Courtright's personal expenses, including his mortgage and private school tuitions for his family. "TGC and Courtright's alleged fraud promised a guaranteed return when the company's business model and financial condition could not possibly support it," said Antonia Chion, Associate Director in the SEC's Division of Enforcement. "To avoid further harm to investors and preserve the misused assets that have not already been dissipated, we have sought and obtained emergency relief." The SEC's complaint, filed in federal court in Chicago on Dec. 27, 2019, and unsealed on Jan. 14, 2020, charges Courtright and TGC with violations of the antifraud and registration provisions of the federal securities laws, and seeks certain emergency relief as well as permanent injunctions, return of ill-gotten gains with prejudgment interest, and civil penalties. On Dec. 30, 2019, the Court issued a temporary restraining order, ordered an asset freeze and other emergency relief, and appointed a receiver for TGC. Investors can learn more about Ponzi scheme red flags by using the SEC's Investor.gov website. Investors should be cautious any time there are promises of a guaranteed rate of return in perpetuity. The SEC's investigation was conducted by Michael Brennan, Patrick L. Feeney, Jeffrey Anderson, Donato Furlano, and Michi Harthcock, with assistance from Suzanne J. Romajas. The investigation was supervised by Kevin Guerrero, Peter Rosario, and Antonia Chion. The litigation will be led by Ms. Romajas and supervised by Stephan Schlegelmilch; Robert M. Moye will assist with the litigation. The SEC's investigation is continuing.
WRITING 2020-108 to file
The Securities and Exchange Commission today released an updated roster of the executive staff of Chairman Jay Clayton, including several individuals who have recently joined the office. Chairman Clayton's executive staff is responsible for advising the Chairman on all matters before the Commission, working closely with agency staff, and helping the Chairman perform all day-to-day operations needed to fulfill the SEC's mission. "The women and men in the Chairman's office have demonstrated an unwavering commitment to advancing the SEC's mission, particularly during these challenging times," said Chairman Jay Clayton. "Their diverse backgrounds, wide ranges of experience, a no dedication, together with that of my fellow Commissioners and our 4,500 talented colleagues, have allowed the agency to continue our important work on behalf of Main Street investors." Below is a full list of Chairman Jay Clayton's executive staff as of May 2020. Sean Memon
Chief of Staff
Bio Bryan Wood
Deputy Chief of Staff
Bio Kristene Blake
Director of Administration
Bio Aleah Borgward

```

Figure 4: Output for the query investigation

Query	Similarity Value
"litigation"	0.072
"investigation"	0.076
"fraud"	0.174
"financial fraud whistleblower"	0.3752

Figure 5: Maximum cosine similarity value obtained for a set of representative queries

tation to the overall query performance, as being from an official source, the vocabulary

is very restricted in size (size is only around the order of 10^4 tokens for a set of 2600 documents). However on in-vocabulary words, the system manages to retrieve relevant documents.

From the table shown above, it is clear that the more specific a query becomes, the better is the performance of the system at retrieving relevant documents with regards to the submitted query.

4.3 Output Screenshots 2

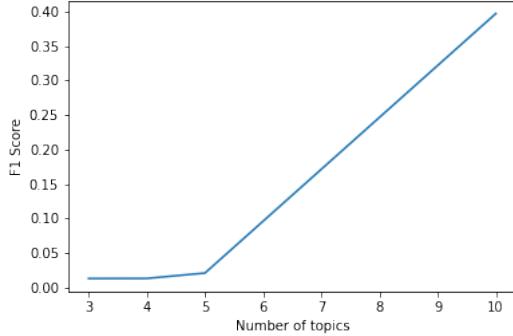


Figure 6: F1 Score for Classification

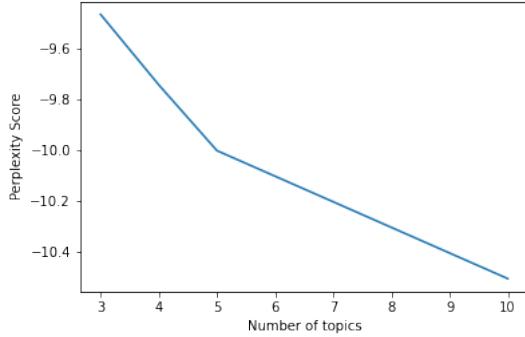


Figure 7: Model Perplexity

4.4 Description of results for LDA-based fraud classification

From the above plots, it is clear that up to a point, increasing the number of topics results in an improvement not only in the quality of the model (as evidenced by the perplexity score) but also the performance of the model on the classification task (as evidenced by the improvement in F1 score)

However, beyond a point, it is clear that the law of diminishing marginal returns affects the model's performance with respect to increasing number of topics. Hence we have chosen 5 topics as a good compromise between the model quality and classification

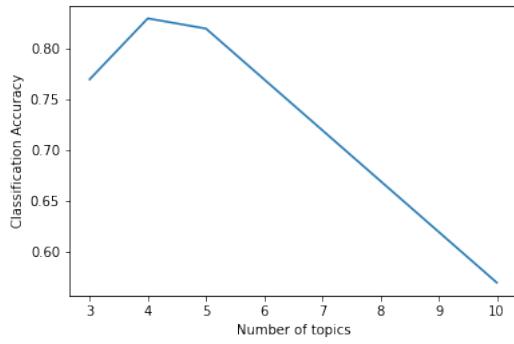


Figure 8: Model Accuracy

performance.

The images below show the term-topic distribution for a varying number of topics

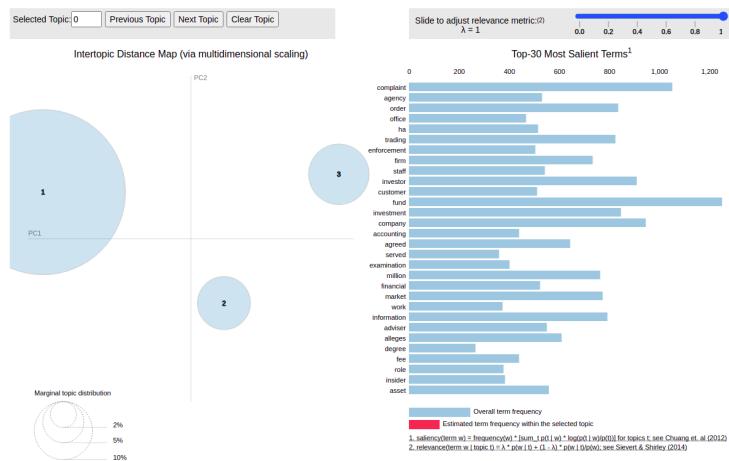


Figure 9: Visualization for 3 topics

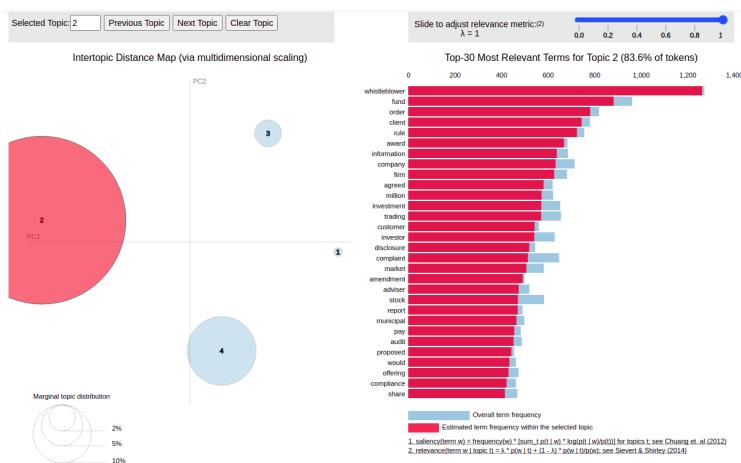


Figure 10: Visualization for 4 topics

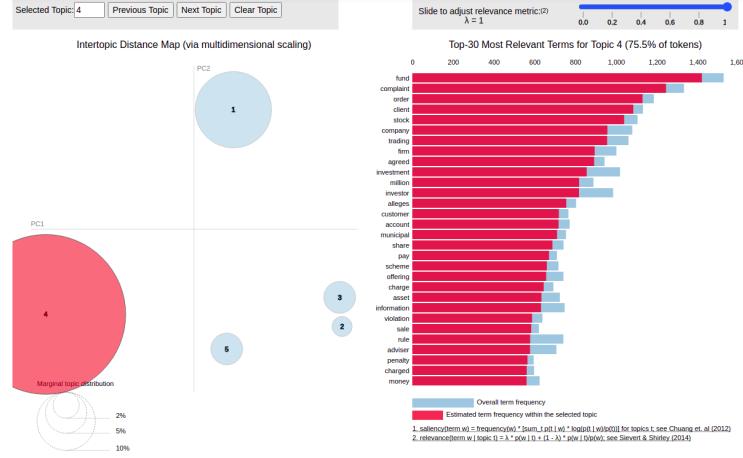


Figure 11: Visualization for 5 topics

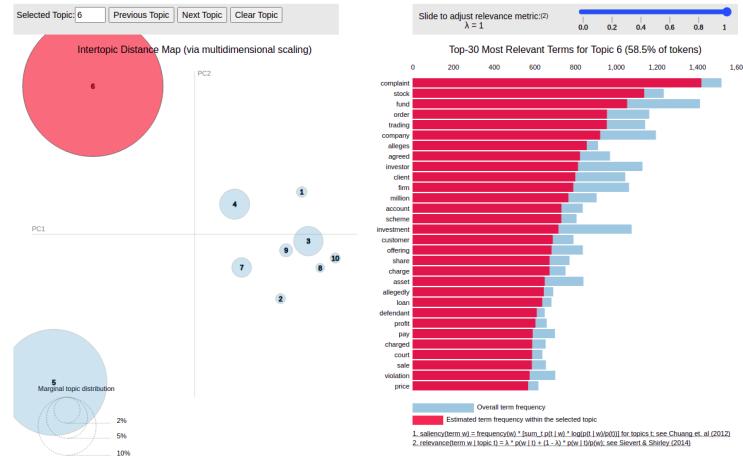


Figure 12: Visualization for 10 topics

5 Conclusion and Future Scope

5.1 Future Scope

The following areas for improvement can be suggested based on our experience of working with the dataset and modelling:

- Explore alternate corpora for financial fraud detection, and use them to either replace or augment the existing corpus of SEC Press Releases.
- Exploring other methods of data cleaning such as stemming, and their impact on the performance of both selected models
- Explore alternate formulae for calculating the TF and IDF scores and their impact on the relevance of the retrieved results
- Explore alternate methods of topic modelling such as latent semantic analysis (LSA) and their impact on the classification accuracy and the term-topic and document-topic distributions obtained.

5.2 Conclusion

This work implements text mining techniques and topic modelling techniques in order to detect instances of financial fraud. The steps involving gathering the data, cleaning the data, modelling the data and analyzing the results were successfully carried out in the course of this project. While the results leave some room for improvement, the improvement strategies have been outlined in the future scope section. The project for the course Algorithms for Intelligence Web and Information Retrieval gave us many ideas as to the real world implementation of the concepts taught in the course, as well as the constraints regarding said implementation, and we have come out of the project experience as more aware students.

We wish to thank Dr. Nagegowda KS, Asst. Professor at the Department of Computer Science and Engineering, PES University, for his unwavering support and insightful guidance in the course of this project.

References

- [1] "Analysis of communication patterns with scammers in Enron corpus", Dinesh Balaji Sashikanth, arXiv:1509.00705
- [2] "Financial Statement Fraud Detection using Text Mining", Nasib Singh Gill, International Journal of Advanced Computer Science and Applications 2012, DOI: 10.14569
- [3] "Building a Semantic document search engine using TF-IDF", Zayed Rais, 2020 <https://medium.com>
- [4] "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents", Shahzad Qaiser, Ramsha Ali, International Journal of Computer Applications 2019. volume 181(1), DOI: 10.5120
- [5] "Latent dirichlet allocation in web spam filtering", Istvan Biro, Jacint Szabo, Andras A Benczur, AIRWeb '08
- [6] "Financial Statement fraud detection using text mining: a systemic functional linguistics theory perspective", Wei Dong, Shaoyi Liao, Liang Liang, PAICS 2016
- [7] "Topic Modeling in Financial Documents", Patrick Grafe, Dept. of Computer Science, Stanford University
- [8] "Text mining techniques to detect corporate fraud", silfratech.com
- [9] GenSim module for topic modelling applications, [PyPI](https://pypi.org/project/gensim/)
- [10] PyLDAVis for interactive topic model visualizations, [PyPI](https://pypi.org/project/pyldavis/)