A Novel Text Mining Approach to Securities and Financial Fraud Detection of Case Suspects

Sundar Krishnan skrishnan@shsu.edu
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Narasimha Shashidhar karpoor@shsu.com
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Cihan Varol cxv007@shsu.com
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

ABM Rezbaul Islam ari014@shsu.com
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA

Abstract

Securities or stock fraud is a type of financial fraud involving securities or asset markets that can result in criminal charges and jail time. Detecting securities fraud from a large volume of electronic evidence without automation, statistical methods, and analytics can be a mammoth exercise for investigation teams due to the ever-increasing volumes of electronic data as case evidence. In this study, the authors propose a machine learning and neural network-based approach consisting of various analytical sub-approaches and automation that can assist financial forensic investigators, legal teams, paralegals, digital forensic investigators, and auditors in financial fraud case investigations such as “insider trading fraud” and “pump and dump fraud”. This comprehensive approach can help reduce investigation time, cost and rework when identifying internal trading fraud and pump and dump fraud indicators.

Keywords: Digital Forensic Analytics, Digital Forensics, Supervised Learning, Hybrid Learning, Unsupervised Learning, Insider Trading, Pump and Dump, Legal Analytics, Forensic Accounting, Financial Forensics, Legal Case Evidence, eDiscovery, Financial Fraud, Electronic Stored Information.

1. INTRODUCTION

Every nation’s economy is increasingly dependent on the success and integrity of the securities and commodities markets. Market manipulations can occur, leading to massive financial loss, erosion of public confidence, and investor fears. Such manipulations (fraud) are sometimes triggered by individuals and punishable under laws. Financial fraud or securities fraud, also known as stock fraud and investment fraud, is a deceptive practice in the stock or commodities markets that induces investors to make purchase or sale decisions based on false information (“Securities fraud - Wikipedia,” n.d.). Securities fraud can also include stock manipulation causing a loss for investors. An example of securities fraud is insider trading, in which trading of a public company’s stock or other securities (such as bonds or stock options) is based on nonpublic information about the company (“Insider trading - Wikipedia,” n.d.). The U.S. Securities and

Exchange Commission (SEC) defines insider trading as “buying or selling a security, in breach of a fiduciary duty or other relationship of trust and confidence, while in possession of material, nonpublic information about the security” (“Officers, Directors and 10 percent Shareholders | SEC.gov,” n.d.), (“Detecting Financial Statement Fraud,” n.d.). Usually, an insider privy to non-public information benefits from stock trading. Insider trading violations may also include “tipping” such information on securities trading by the person “tipped” [3]. Another example of securities fraud is the Pump & Dump (P&D) which involves artificially inflating the price of an owned stock through false and misleading positive statements, to later sell the cheaply purchased stock at a higher price (“Pump and dump - Wikipedia,” n.d.). This fraud need not be carried out by an insider of the company but can be detected from a pattern of statements and their subsequent trading on the stock. In both the fraud scenarios, these days, the Internet plays a major role coupled with advancing technology of smartphone devices and real-time information on stocks. Fraudsters leverage spam emails, investment research websites, social media, misinformation, advertisements, stenography, and various technology means to accomplish their goals. Detecting such fraud and forensically investigating them can involve many human resources often contracted to private firms specializing in this field.

In the U.S., many rules (“Selective Disclosure and Insider Trading,” n.d.), laws (“The Laws That Govern the Securities Industry,” n.d.), and regulations have been enforced by the Securities and Exchange Commission (SEC) and the Financial Industry Regulatory Authority (FINRA) accompanied by state-level securities laws. Usually, federal entities such as the SEC and FBI lead fraud investigations and prosecution. In the U.S., the government tries to prevent and detect insider trading by monitoring the trading activity in the market, especially around significant events such as earnings announcements, acquisitions, and other events material to company’s value. Such events that may move the company stock prices significantly and thus close monitoring of company insiders for illegal trading activity is the focus. Sometimes, whistle blowers and complaints from traders may also trigger investigations. Formally, the SEC defines an insider for insider trading as a “director, senior officer, or any person or entity of a company that beneficially owns more than 10% of a company’s voting shares” [3]. The SEC defines pump and dump schemes in two parts; 1. stock promotion through sharing misleading or false statements/information, 2.sale of stock during peak price (“‘Pump and dump’ Schemes,” n.d.), (“Pump & Dump Schemes - Securities Fraud Attorneys,” n.d.). Several activities may trigger a fraud investigation such as reverse mergers, wash trades, misleading press releases, whistleblowers, and complaints from traders [9]. Penalties of both frauds involve fines and federal jail time (“Insider Trading FAQ Part 1,” n.d.).

Financial forensics also known as forensic accounting is the art and science of investigating people and money (Dorrell & Gadawski, 2012).Financial forensics is a specialty area of financial accounting practice mainly focused on detecting financial crimes. Forensic accounting professionals are trained to identify red-flag discrepancies in accounting related documents and forensically follow clues to the source of fraudulent activity within the scope of the investigation. A trigger to financial forensics is a forensic audit of financial data. Forensic audits relate directly to an issue defined by the audit client and designed to focus on reconstructing past financial transactions for a specific purpose, such as concerns of fraud (“The Difference Between a Financial Statement Audit & a Forensic Audit,” n.d.), (“Forensic Audit vs. Internal Audit: Differences in Accounting,” n.d.).

Technology driven tools for automation of forensic investigations of fraud are few and often proprietary to the forensic investigation teams. The bulk of securities fraud investigations involves many painstaking hours of combing evidence. In the recent decades, evidence has transformed from traditional paper to electronic files. Technology leaps have led to ever increasing storage at lower costs leading to mammoth piles of electronic data to sift through. Technology has come to the aid of such investigations but is still considered as an assistant to human effort. This perception has now started to change with the leveraging of Artificial Intelligence and Machine Learning techniques. Financial data specific forensic tools may exist, but there is a lack of tools that mine non-finance data for clues of fraud. With fraudsters leveraging the Internet for social media platforms, finance related discussion forums, smartphones, etc., evidence of fraud has now moved away from traditional financial data. This also calls for exacting such evidence from networks, smartphone and computers thereby involving digital forensic professionals. Preparing clues for prosecution may later involve eDiscovery professionals and paralegals.

Natural language processing (NLP) is a subset of Artificial Intelligence involving programming computers and computational linguistics to process massive volumes of unstructured language data by breaking it down into a structured format("NLP vs. NLU vs. NLG: the differences between three natural language processing concepts," n.d.),("NLP vs. NLU: What's the Difference and Why Does it Matter?," n.d.). NLP focuses on processing the text in a literal sense, like what was said. Conversely, Natural Language Understanding (NLU) being a subset of NLP focuses on extracting the context and intent from the language. While humans naturally do this in conversation, the combination of NLP and NLU techniques is required for a machine to understand the intended meaning of different texts. In short, NLP looks at what was said, and NLU looks at what was meant ("NLP vs. NLU: What's the Difference and Why Does it Matter?," n.d.). The need to employ NLP and NLU in processing electronic forensic evidence for a case is required as evidence can largely be text based such as social media posts, SMS, emails, office documents, etc. Such evidence can be voluminous and searching for relevant content (also known as ediscovery) for winning legal arguments can take time, thus driving up investigation costs ("Three Trends Driving Up E-Discovery Costs," n.d.), (Fritz, n.d.) and risk. A financial fraud investigation these days involves skilled resources from many branches of forensics, analytics, legal, paralegals, law enforcement, etc. - all leveraging technology to speed up the process. With the recent popularity and growth of Artificial Intelligence (AI), deep learning, Neural Networks, and Machine Learning (ML) coupled with ever increasing automation power, large amounts of evidence can be quickly processed for clues. This saves time and reduces costs for forensic teams. In this research, the authors propose a methodology to harvest clues from non-finance evidentiary data for a suspect using multiple approaches. They also propose a tool that automates these approaches to assist the fraud investigation team.

2. LITERATURE REVIEW

Post investigation scoping, traditional financial forensics relies on tools and techniques such as Financial Status Audit Techniques (FAST), ICE/SCORE, Interviews, Interrogation, facial mappings, background research, forensic lexicology, valuation matrix, laboratory analysis, transaction analysis, etc.(Dorrell & Gadawski, 2012). However, in the recent decade, this has been supplemented with technology-driven tools such as block chain, deep learning, and machine learning, primarily to speed up the investigation process and address vast volumes of electronic case evidence. In academic literature, financial fraud areas of focus (evidentiary data) have been on credit card data, banking transactions, financial statements, and social media. The notion of financial fraud has been viewed differently in different academic disciplines. Legal scholars conceive of fraud as a legal concept that prescribes civil or criminal liability, scholarsin the fields of sociology and criminology think of it asa behavioral category that involves human actions, finance,and economics scholars, in turn, see fraud as a form of risk(Reurink, 2018). While approaches may vary, the common underlying need in investigations is speed and accuracy. Technology largely addresses this requirement by leveraging Computational Intelligence (CI) based techniques such as deep learning, data analytics, and data mining. An important use of technology in this field is in fraud detection. West et al. (West & Bhattacharya, 2016) analyze scientific literature for the association between fraud types, CI-based detection algorithms and their performance. David(Nam, 2020) employed classification techniques on internet discussion forum data to detect pump-and-dump (P&D) fraud schemes and found Convolutional Neural Network (CNN) to be the best performing model in classifying (P&D) fraud. A recent area of focus for P&D fraud is the Crypto currency market, wherein prices of Crypto currency increase in prices, volume, and volatility followed by quick reversals. Jiahua et al. (Xu & Livshits, n.d.) investigate multiple P&D activities organized in Telegram channels and built a model that predicts the pump likelihood of all coins listed in a crypto-exchange prior to a pump.

Tao et al.(Li, Shin, & Wang, 2021) show P&Ds are detrimental to the liquidity and price of Crypto currencies.

There are many warning (red flags) signs that precede the occurrence of fraud. A red flag is a set of circumstances that are unusual or vary from the normal activity, thereby warranting an investigation (Hancox & Dinapoli, n.d.). Insider fraud detection in academic research often focuses on banking transaction data, credit card data, Crypto currencies and point-of-sale systems(Samaneh Sorournejad, Zojaji, Atani, & Monadjemi, 2016). Srivastava et al. (Srivastava & Bhatnagar, 2021) employ process mining to examine the most common forms of insider fraud occurring in banks and attempt to categorize them. Internal fraud has spread to emerging nations as well. Roy et al. (Roy & Basu, 2021)identify the types and drivers of insider frauds in Indian banks that can be absorbed by policymakers in creating a more robust system for timely detection of frauds. With the popularity and benefits of applying statistical and analytical techniques coupled with ever increasing volume of financial data in investigations, machine learning algorithms, deep learning and transfer learning techniques have recently attracted the attention of academia in financial fraud detection. Lebichot et al. (Lebichot, Borgne, He-Guelton, Oblé, & Bontempi, 2019) use deep transfer learning approaches for credit card fraud detection in a deep neural network setting. They focus on transfer classification models learned on a specific category of transactions (e-commerce) to another (face-to-face). Jiang et al. (Jiang et al., 2019) design a GCN (Graph Convolutional Networks) based anomaly detection model to detect anomalous behaviors of users and malicious threat groups for insider threat and fraud detection. Liu et al. (Liu, Mai, Shan, & Wu, 2020) propose a textual analytic framework employing machine learning to predict potentially opportunistic insider trading from corporate textual disclosures. Islam et al. (Islam, Khaled Ghafoor, & Eberle, 2019) present a deep learning based approach that detects and predicts illegal insider trading proactively from large heterogeneous sources of structured and unstructured data. Lauar et al. (Lauar & Arbex Valle, 2020) employ machine learning techniques based on features from both spot and options markets to recognize suspicious negotiations before relevant events are disclosed to Brazilian authorities. Other sources of evidence such as social media communications and SMS texts can also offer key evidence towards financial fraud detection. To encompass such evidence into fraud detection, the authors in this research use previously created synthetic textual dataset from various possible electronic sources to closely mirror a typical financial/securities fraud investigation evidence stack. They then employ various analytical algorithms, build a labeled structured dataset of trading texts, correlate events against historical stock data while factoring in suspect's risk profile and sentiments to provide an approach for insider trading and P&D detection. The authors propose this approach as possible a path for fraud investigation teams when embarking on leveraging machine learning and Artificial Intelligence techniques to speed up their investigations.

3. METHODOLOGY

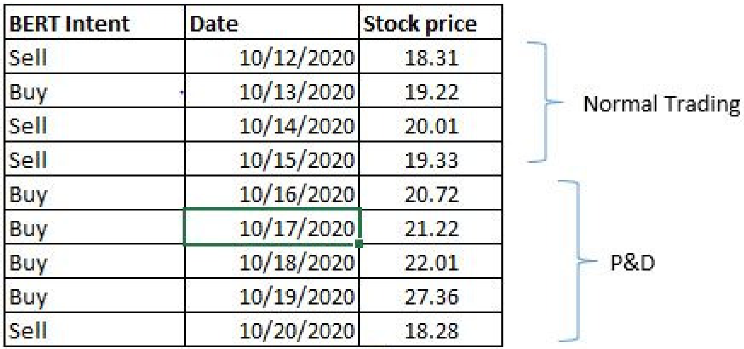
Evidence for a real-life forensic investigation of financial fraud was hard to find in public for academic research. Thus, the authors felt the need to customize and build random fictitious electronic evidence (ESI) for this experiment (Krishnan, Shashidhar, Varol, & Islam, 2022). The experiment was carried out using an Intel(R) Core (TM) i5-3470 CPU @ 3.20GHz 16 GB RAM PC and a 64-bit Windows 10 operating system. Software used was Python, SQL Server 2019, and Visual Studio 2019. This experiment is solely to showcase an umbrella approach to tackling securities/financial fraud investigations. To avoid bias, the experiment results are published as-is, and no attempt was made to withhold wayward results or showcase only high-fidelity results.

3.1 Insider Trading

An insider's motive (intent) to buy or sell stock using privileged information unknown to the public is key to indicators of insider trading fraud. The logic for detecting insider trading considers four different factors as shown in Figure 4. Provision for calculating the risk of an individual along with static and temporal anomalies coupled with machine learning techniques are the key areas of this logic. The crux of the logic is the intent exhibited by a suspect when trading stock. Intent can be gathered from the communications of the suspect and then correlated against stock prices of the same timestamp. The key to establishing insider trading is the suspect's knowledge of privileged information. This is marked by the suspect's access to this information by means of attending key meetings, access to IT systems, and access to colleagues who may have this information. A high-risk employee profile is defined in terms of Financial IT systems privileged access, action owner of past audits, etc. Sentiments of a suspect's communication can greatly assist with legal arguments and thus sentiments across case evidence were highlighted for the investigators. Figure 6 shows our proposed software implementing this logic and handling insider trading scenario across whole evidence.

3.2 Pump & Dump

Since we are determining intent from textual data, we can implement a simple logic for pump and dump (P&D). If we observe a pattern of intent to "buy" stock followed by an intent to "sell" and this pattern correlates to stock price increase followed by a drastic fall, then we can conclude there is an indication of P&D. For P&D the suspect need not be an employee of the company and thus any such metadata collected by the tool was ignored. Figure 1 describes this logic. This logic can be further tuned for parameters such as the amount of price increase (pump), the amount in price decrease (dump), the volume of stock sold between the timestamps of this pattern and communication time gaps (days, hours) to trigger a P&D indicator. For simplicity, the logic in our experiment required a minimum of three or more continuous text communication evidence with intent to buy stock followed by a single textual communication evidence of an intent to sell this stock. This should then correlate to an increase in stock price (due to buy/pump intent) and volume followed by an immediate decrease (due to sell/dump intent). Figure 7 shows our proposed software implementing this logic and handling P&D scenario across whole evidence.



BERT Intent	Date	Stock price
Sell	10/12/2020	18.31
Buy	10/13/2020	19.22
Sell	10/14/2020	20.01
Sell	10/15/2020	19.33
Buy	10/16/2020	20.72
Buy	10/17/2020	21.22
Buy	10/18/2020	22.01
Buy	10/19/2020	27.36
Sell	10/20/2020	18.28

Normal Trading
P&D

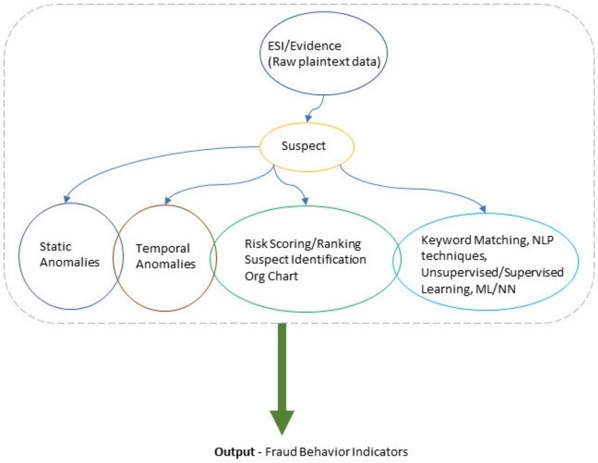
FIGURE 1: Example of Pump and Dump (P&D) logic using Intent

3.3 Experiment Design

The aim of this experiment is to propose an approach that helps the investigators to investigate financial fraud, especially cases of insider trading and Pump & Dump schemes. Case evidence was mined for human intent, unlabeled data was labeled using unsupervised learning, and various algorithms were implemented along with risk ranking, suspect profiling, and sentiment analysis to arrive at fraud indicators. To arrive at such conclusions, a mix of direct mining of evidence coupled with machine learned predictions from labeled data is employed. Figure 2 shows a high-level investigation approach with suspect profile, suspect intention, stock value, and risk as inputs producing a Boolean indicator of fraud as output along with the source of evidence. For ease of understanding, the proposed fraud detection approach utilizes the three buzzwords of any investigation, namely: means, opportunity and motive. While this detection approach can address means and opportunities to a certain degree, it is left to the investigators and prosecutors to establish a motive. However, for ease of understating, the motive is taken as profiting from stock prices. The overall methodology is inductive wherein the investigators look for patterns in the case data (ESI) and arrive at potential fraud conclusion. This is the preferred way in any criminal/civil investigation as otherwise it can be constructed as bias or prejudice by the investigators in hypothesizing crime/fraud and then looking for patterns in the case data. Figure 3 highlights the various machine learning and automation methods/techniques leveraged under the proposed financial fraud detection umbrella. The reason for proposing multiple analytical algorithms (sub-approaches) all wrapped into one main approach is that investigators are not bound by the results of one algorithm but instead have a mix to choose the best one. Also, each analytical algorithm has a built-in feedback feature that, when triggered by the investigator, will contribute back to labeled data that can be reused for supervised learning.

3.4 Dataset Preparation

The datasets used for this experiment was from prior research(Krishnan et al., 2022). The key types of data were from fictitious emails, Facebook posts, Tweets, WhatsApp/SMS messages, and random MSWord documents. Data was stored in SQL tables identified by their source/document identifier known as bates number/ID. Each email and MS Word documents were further broken down into sentences and stored in a separate SQL table. Data needs to be processed for analytics as there can be occurrences of emojis, hyperlinks, stop words, etc. that can inhibit the analytical process (Krishnan, Shashidhar, Varol, & Rezbaul Islam, 2021). All textual data was pre-processed using Natural Language Processing (NLP) techniques such as tokenization, stop words, stemming and lemmatization. All suspect names, key event dates; textual data and stock symbols used are solely for demonstration purposes and bear no resemblance in any shape or form in real life.



INPUT: Suspect, Investigation Timelines & Stock

MEANS:
Static Anomalies (via Inputs for Investigator) – Attendance at key company meetings, mail address, suspicious destination traffic, suspicious metadata, etc.

OPPOURTUNITY:
Temporal Anomalies (via Inputs for Investigator) – Colleagues network/access, organization roles, system access levels, communication frequency, etc.

MOTIVE :
Buy/Sell Intent detection and Sentiments (Use of available ML/NN models LSTM, NLP, CNN etc.)

OTHERS :
Scoring Engine – Risk/Score/Ranking/Flags. **(Weights)**
Subject Identification – Actors/Subjects in ESI
Investigator Input: Enterprise organization chart

OUTPUT: Fraud Behavior *Indicators* & *Cui Bono* (Who is to gain)

Motive is the reason for committing the **crime**, **means** are the tools or methods used to commit the **crime**; and **opportunity** is the occasion that presents itself to allow the **crime** to take place.

FIGURE 2: Financial fraud detection – High level approach.

1) Reddit Data: Stock trading and finance-related discussion data from Reddit forums was collected via allowed Reddit APIs ("reddit.com: api documentation," n.d.). For simplicity, sub reddits (community/channel/forum) considered were Wallstreet bets and Investing. Python scripts were written and executed between Nov/06/2021 and Nov/14/2021 to read each sub reddit data and write into .csv files that were later stored as SQL tables. A total of 155,651 rows of Reddit data was collected. This step can be altered if the investigation team has quality labeled stock-trading data for supervised learning.

2) Yahoo Finance Data: Historical market data from Yahoo Finance was obtained as needed using the python module yfinance (Aroussi, n.d.). The module yfinance is a python module that uses Yahoo! Finance's API and returns stock, crypto currency, forex, mutual fund, commodity futures, ETF, and U.S. Treasury financial data. Python scripts were written and executed via C#.NET on the prototype tool. These scripts also inserted data into SQL tables when executed.

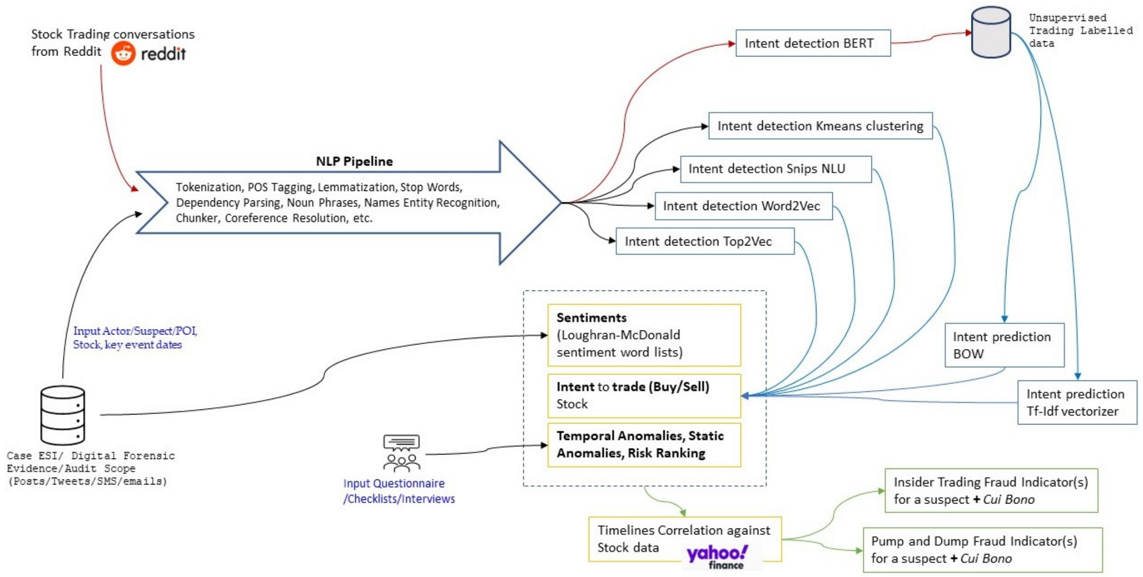

Stock Trading conversations from Reddit
reddit
NLP Pipeline
Tokenization, POS Tagging, Lemmatization, Stop Words, Dependency Parsing, Noun Phrases, Names Entity Recognition, Chunker, Coreference Resolution, etc.
Intent detection BERT
Unsupervised Trading Labelled data
Intent detection Kmeans clustering
Intent detection Snips NLU
Intent detection Word2Vec
Intent detection Top2Vec
Input Actor/Suspect/POI, Stock, key event dates
Case ESI/ Digital Forensic Evidence/Audit Scope (Posts/Tweets/SMS/emails)
Sentiments (Loughran-McDonald sentiment word lists)
Intent to trade (Buy/Sell) Stock
Temporal Anomalies, Static Anomalies, Risk Ranking
Input Questionnaire /Checklists/Interviews
Intent prediction BOW
Intent prediction Tf-Idf vectorizer
Timelines Correlation against Stock data
yahoo! finance
Insider Trading Fraud Indicator(s) for a suspect + Cui Bono
Pump and Dump Fraud Indicator(s) for a suspect + Cui Bono

FIGURE 3: Financial fraud detection process involving various approaches.

3) Ancillary Data: For various automation steps, ancillary data such as stock ticker/symbol data, emojis, emoticons, stop words, etc., were assembled from the Internet. Few data files were stored in SQL databases, while the rest were stored as flat files. All stock data was limited to NASDAQ, NYSE, and NYSE stock exchanges that can be further expanded to other exchanges.

3.5 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for NLP developed by Google (Devlin, Chang, Lee, & Toutanova, 2018). BERT can be used in a wide variety of NLP tasks such as question answering (SQuAD v1.1), Natural Language Inference (MNLI), and others. A python script was written to perform a sort of unsupervised classification of textual Reddit data into buy, sell, or other based on similarity. This approach helps us label the Reddit data in an unsupervised way. After text data preprocessing, creation of target clusters using Word2vec and gensim was performed, followed by word Embedding with transformers and BERT. The gensim package has a function that returns the most similar words for any given word. Lastly, we assigned observations to clusters by cosine similarity and evaluated the model's performance. Classification results were stored in SQL tables as labeled data using BERT.

1) TF-IDF: Term Frequency-Inverse Document Frequency (TF-IDF) statistical approach determines how important a word is by weighing its frequency of occurrence within the document. After data preprocessing (word cleaning, stop words removal, hyperlinks, stemming, lemmatization), the data from BERT was split into training & testing subsets. A Naive Bayes classifier was used to fit the training data, and predictions were obtained with the test dataset. Model's accuracy, precision, recall, confusion matrix, and ROC were obtained. This model was then applied against text from each document (identified by bates number) from the investigation caseload and prediction results of buy/sell/other were stored in a SQL table.

2) BOW: The Bag-of-Words (BOW) model builds a vocabulary from a corpus of documents and counts how many times the words appear in each document. A python script was created for implementing BOW. After data preprocessing (word cleaning, removal of stop words, hyperlinks, stemming/ lemmatization), the labeled dataset (using BERT technique) was split

into training & testing subsets. The Tf-Idf vectorizers and Naive Bayes classifier was applied to transform and predict test data. Model's accuracy, precision, recall, confusion matrix and ROC were obtained. This model was then applied against text from each document (identified by bates number) from the investigation caseload and prediction results of buy/sell/other were stored in a SQL table.

3.6 K-Means

This approach involves unsupervised text clustering using NLP and K-Means. Against the Reddit data, clustering was performed using K-Means to find top 10 clusters. After data preprocessing (word cleaning, removal of stop words, hyperlinks, stemming/ lemmatization), the clustering model was then applied directly against ESI case data for any synonyms (similar words). Results were stored in a SQL table.

3.7 Top2Vec

Top2Vec (Angelov, 2020) is an algorithm for topic modeling and semantic search in a large collection of documents. Top2Vec utilizes Doc2vec to first generate a semantic space that consists of word and document vectors in a continuous representation of topics. There was no need to remove stop words as such words will appear in almost all documents present in the corpus, therefore being equidistant from all topics. They will not appear as a nearest word to any specific topic. Stemming/lemmatization were not implemented, but text was cleaned for punctuation and made lowercase. A python script was created to implement Top2Vec against case evidence for keywords "buy" and "sell" and similar semantic words. Results were stored in a SQL table.

3.8 Word2Vec

Word2vec is a popular technique to learn word embeddings using deep learning and a two-layer neural network. Its input is a text corpus, and its output is a set of vectors wherein semantically similar words are placed close to each other. Word2Vec model comes in two flavors: Skip Gram Model and Continuous Bag of Words Model (CBOW). A python script was created using gensim library implementing both Skip Gram Model and CBOW approach of Word2vec directly against the case evidence. The script computed the similarity of words to "buy" and "sell" in each bates number of the investigation evidence caseload. Results were stored in a SQL table.

3.9 Snips

Snips NLU (Coucke et al., 2018) is an open-source Natural Language Understanding(NLU) python library that allows parsing sentences written in natural language and extract structured information("Snips Natural Language Understanding — Snips NLU 0.20.2 documentation," n.d.). The NLU engine first detects the from text the intention of the user, then extracts the parameters (called slots) of the query. As required by Snips, ajson/YAMLfile was fitted in the SnipsNLU Engine with custom utterances of buy/sell intent and stock symbols. Figure 4 shows a snapshot of this file contents. A good alternative to Snips NLU was Rasa NLU ("Open source conversational AI," n.d.). However, Snips NLU has been proven to be better than Rasa NLU ("NLU-benchmark/2017-06-custom-intent-engines at master · sonos/nlu-benchmark," n.d.), (GitHub, n.d.)and thus used in this experiment.

3.10 Sentiments

Suspects can display sentiments that can help in profiling. The authors decided to use the Loughran-McDonald sentiment word lists (Terblanche & Marivate, 2021) to perform sentiment analysis as this was specifically built and is maintained for textual analysis related to finance. This added information could help investigators better understand the behavioral aspect of the suspect at a particular timeline corresponding to the text origin. A python script was created to implement the Loughran-McDonald sentiment word lists against each bates number and suspect in the case evidence pile. Results were stored in a SQL table. Prior work by the authors on deducing suspect's sentiments from case data (Krishnan et al., 2022)can be used in addition to the Loughran-McDonald sentiment lexicon.

3.11 Calendar of Key Events

To correlate key dates of events against specific evidence (example tweet or SMS sent date) and lookup of historical stock prices, a C#.NET input screen/form was created to ingest multiple dates and event details. This information was manually input by the investigators and stored in a SQL table. For the sake of simplicity, only dates were considered although this can be extended to include the time of day.

3.12 Suspect Metadata

A C#.NET input screen/module was created to ingest suspect metadata such as their company designation, attendance at meetings (dates), and work hours. Such data can help support the investigation findings. This information was manually input by the investigators for suspects and stored in a SQLtable.

3.13 Risk Profile and Ranking

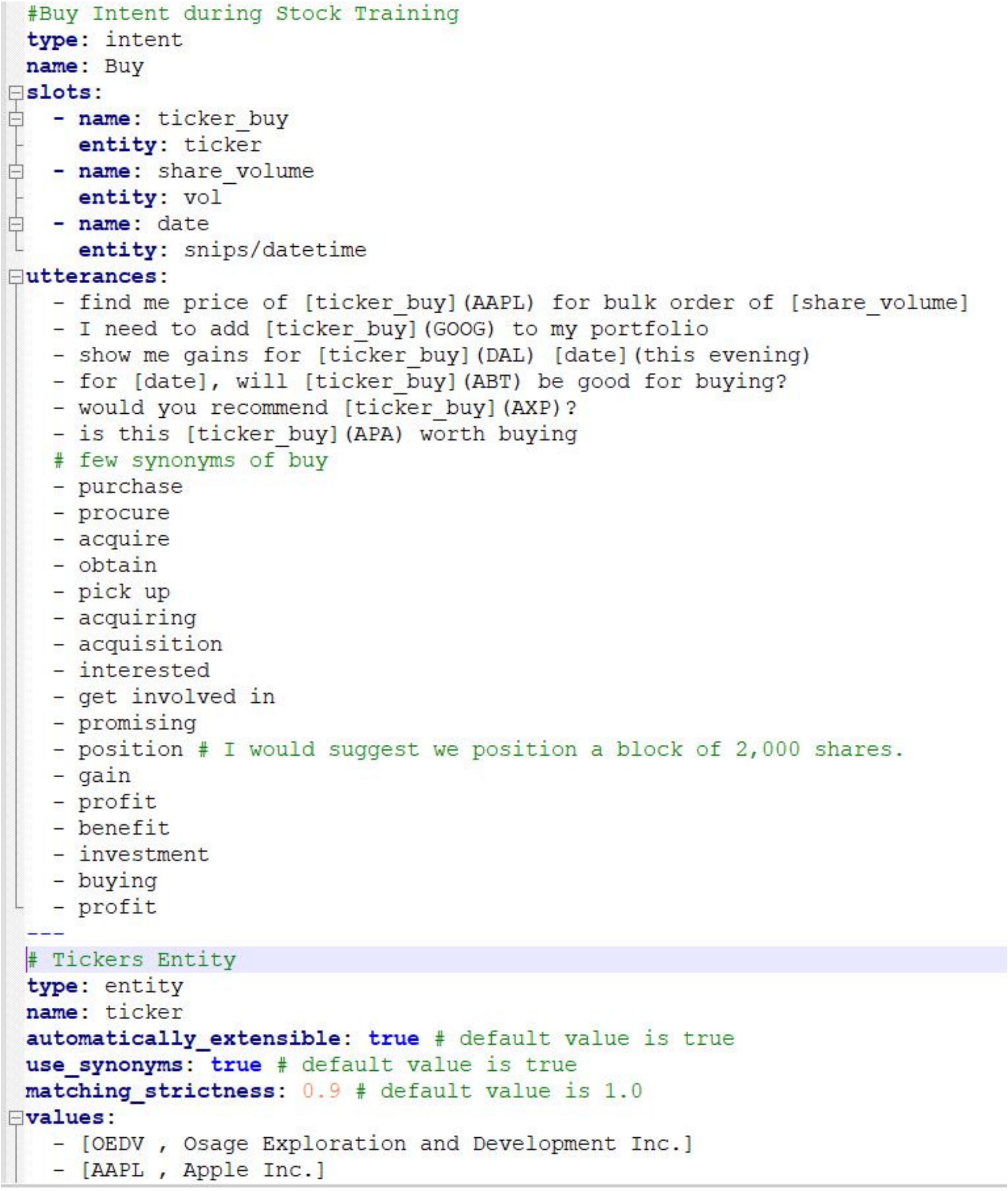
A C#.NET input screen/module was created to ingest suspect risk metadata such as involvement in financial audits, access to key colleagues, finance systems access levels, elevated privileges if any, prior red flags from Human Resources (HR)department of the company and a prior victim of phishing. Such data can help build a risk profile, provide circumstantial/observational evidence, and provide valuable insight of the suspect to the investigators. Risk ranking was based on weighted approach of this metadata that can be customized by the investigators. The investigators manually input this information for suspects and stored in a SQL table.

3.14 Presentation

A Windows Forms (client/server) module was created using C#.NET as a prototype software (Krishnan, n.d.) for use by the investigators. Figure 5and Figure 6 shows the main screen of this software when used for “insider trading fraud” and “pump and dump” detection. For an investigation to commence, the user first inputs suspect metadata, suspect risk details, and key event dates that lie within the scope of the investigation. The next step involves choosing the suspect, the company stock, and dates. The user can then choose the option to find evidence of insider trading or evidence of a pump and dump. The authors note that in the case of pump and dump scenario, certain data elements of the suspect collected earlier may not be relevant such as job designation, system access levels, or association to the company. On the prototype software, upon user action to find evidence of insider trading or pump & dump, data stored on various SQL tables is correlated against historical stock data obtained from Yahoo Finance API. Results of correlation is then displayed on the screen along with sentiment data and risk ranking for the suspect, pointing eventually to the bates number if there as any evidence found. If no evidence was found, a suitable message was displayed on the screen. The user screen allows for the download of a report and re-run any of the abovementioned algorithms. The software also allows for storing other investigation related details.

4. ANALYSIS

In this research, the authors combine supervised learning and unsupervised learning to help locate fraud indicators in a stack of electronic evidence. This approach is known as Hybrid supervised/unsupervised learning. The algorithms used in this research can vary and can be improved with user feedback (fine tuning). This approach is suitable for an investigation team that has no prior labeled data on trading intents. They can start with unlabeled data and over the period of many investigations, build a quality dataset. All the code files for this research are available on GitHub (Krishnan, n.d.).

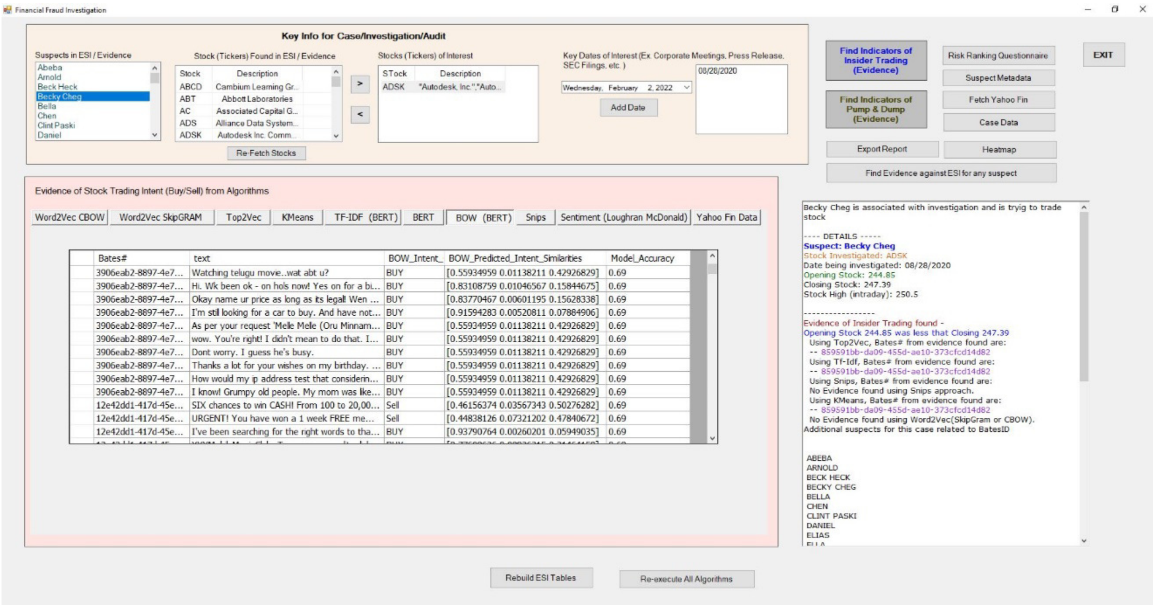


```
#Buy Intent during Stock Training
type: intent
name: Buy
slots:
  - name: ticker_buy
    entity: ticker
  - name: share_volume
    entity: vol
  - name: date
    entity: snips/datetime
utterances:
  - find me price of [ticker_buy](AAPL) for bulk order of [share_volume]
  - I need to add [ticker_buy](GOOG) to my portfolio
  - show me gains for [ticker_buy](DAL) [date](this evening)
  - for [date], will [ticker_buy](ABT) be good for buying?
  - would you recommend [ticker_buy](AXP)?
  - is this [ticker_buy](APA) worth buying
  # few synonyms of buy
  - purchase
  - procure
  - acquire
  - obtain
  - pick up
  - acquiring
  - acquisition
  - interested
  - get involved in
  - promising
  - position # I would suggest we position a block of 2,000 shares.
  - gain
  - profit
  - benefit
  - investment
  - buying
  - profit
---
# Tickers Entity
type: entity
name: ticker
automatically_extensible: true # default value is true
use_synonyms: true # default value is true
matching_strictness: 0.9 # default value is 1.0
values:
  - [OEDV , Osage Exploration and Development Inc.]
  - [AAPL , Apple Inc.]
```

FIGURE 4: Snips json/YAML logic.

1) Quality of Case Data: Typical case data can be a collection of files on electronic/computer systems housing any information related to the scope of investigations. Often the initial collection volume is more than required as the investigation scope may not be well defined and kept broad. The eDiscovery EDRM (EDRM, n.d.) model can be applied to the process of vetting and analysis to filter out the irrelevant data and retain data as critical evidence for legal arguments. The process of vetting native format involves a ton of data processing tasks such as data masking, redaction, culling, etc. A major problem that investigators may encounter is the language used in evidence. When Word2Vec was used against a good dataset like glove-wiki-gigaword-50, similar words were found to be more accurate. However, our case evidence data for Word2Vec was not as good of linguistic quality as that of glove-wiki-gigaword-50 and not encompassing English language. Thereby our results were not as expected. This is a common real-life scenario that investigators will encounter as communication data in case evidence these days is not the standard English language. This holds good to other

prominent world languages as well as the Internet has been bemoaned as the downfall of the written word, pronunciations, and grammar. For example, multiple languages can be found mixed with each other in a single SMS text. Thus, quality of case data can vary from case to case and care must be taken to first analyze linguistics within the textual evidence followed by management of non-textual data such as media, biological evidence, etc.


Financial Fraud Investigation
Key Info for Case/Investigation/Audit
Suspects in ESI / Evidence
Stock (Tickers) Found in ESI / Evidence
Stocks (Tickers) of Interest
Key Dates of Interest (Ex. Corporate Meetings, Press Release, SEC Filings, etc.)
>
<
Add Date
Re-Fetch Stocks
Find Indicators of Insider Trading (Evidence)
Find Indicators of Pump & Dump (Evidence)
Risk Ranking Questionnaire
Suspect Metadata
Fetch Yahoo Fin
Case Data
EXIT
Export Report
Heatmap
Find Evidence against ESI for any suspect
Evidence of Stock Trading Intent (Buy/Sell) from Algorithms
Word2Vec CBOW
Word2Vec SkipGRAM
Top2Vec
KMeans
TF-IDF (BERT)
BERT
BOW (BERT)
Snips
Sentiment (Loughran McDonald)
Yahoo Fin Data
Rebuild ESI Tables
Re-execute All Algorithms

Abeba
Arnold
Beck Heck
Becky Cheg
Bella
Chen
Clint Paski
Daniel

Stock	Description
ABCD	Cambium Learning Gr...
ABT	Abbott Laboratories
AC	Associated Capital G...
ADS	Alliance Data System...
ADSK	Autodesk Inc. Comm...

STock	Description
ADSK	"Autodesk, Inc.","Auto...

Wednesday, February 2, 2022

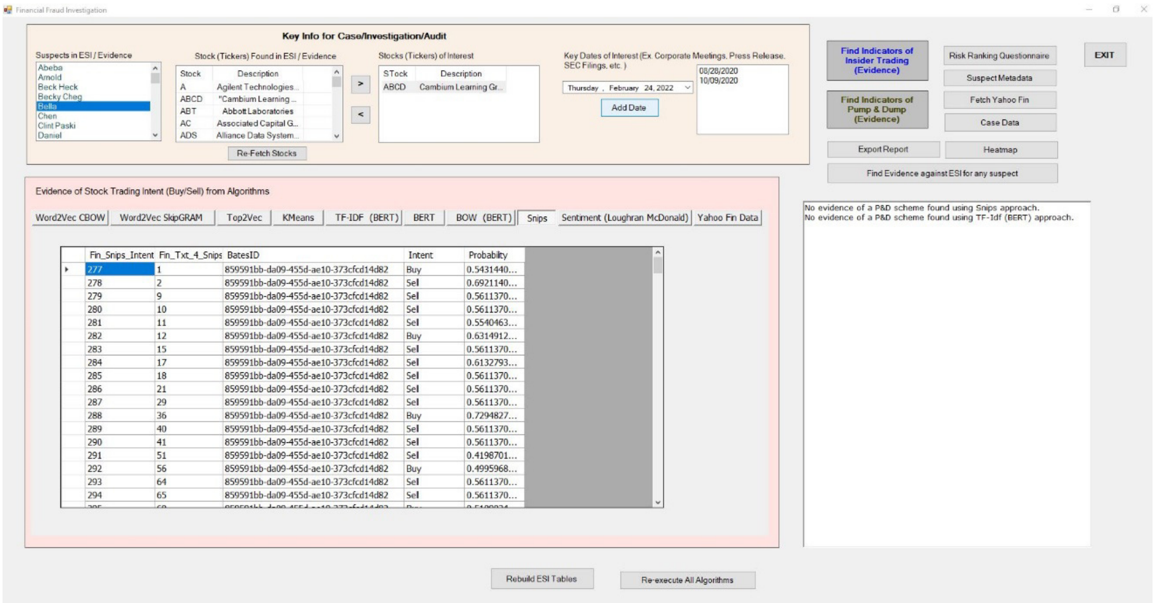
08/28/2020

Bates#	text	BOW_Intent_	BOW_Predicted_Intent_Similarities	Model_Accuracy
3906eab2-8897-4e7...	Watching telugu movie..wat abt u?	BUY	[0.55934959 0.01138211 0.42926829]	0.69
3906eab2-8897-4e7...	Hi. Wk been ok - on hols now! Yes on for a bi...	BUY	[0.83108759 0.01046567 0.15844675]	0.69
3906eab2-8897-4e7...	Okay name ur price as long as its legal! Wen ...	BUY	[0.83770467 0.00601195 0.15628338]	0.69
3906eab2-8897-4e7...	I'm still looking for a car to buy. And have not...	BUY	[0.91594283 0.00520811 0.07884906]	0.69
3906eab2-8897-4e7...	As per your request 'Melle Melle (Oru Minnam...	BUY	[0.55934959 0.01138211 0.42926829]	0.69
3906eab2-8897-4e7...	wow. You're right! I didn't mean to do that. I...	BUY	[0.55934959 0.01138211 0.42926829]	0.69
3906eab2-8897-4e7...	Dont worry. I guess he's busy.	BUY	[0.55934959 0.01138211 0.42926829]	0.69
3906eab2-8897-4e7...	Thanks a lot for your wishes on my birthday. ...	BUY	[0.55934959 0.01138211 0.42926829]	0.69
3906eab2-8897-4e7...	How would my ip address test that considerin...	BUY	[0.55934959 0.01138211 0.42926829]	0.69
3906eab2-8897-4e7...	I know! Grumpy old people. My mom was like...	BUY	[0.55934959 0.01138211 0.42926829]	0.69
12e42dd1-417d-45e...	SIX chances to win CASH! From 100 to 20,00...	Sell	[0.46156374 0.03567343 0.50276282]	0.69
12e42dd1-417d-45e...	URGENT! You have won a 1 week FREE me...	Sell	[0.44838126 0.07321202 0.47840672]	0.69
12e42dd1-417d-45e...	I've been searching for the right words to tha...	BUY	[0.93790764 0.00260201 0.05949035]	0.69

```
Becky Cheg is associated with investigation and is tryig to trade
stock

---- DETAILS -----
Suspect: Becky Cheg
Stock Investigated: ADSK
Date being investigated: 08/28/2020
Opening Stock: 244.85
Closing Stock: 247.39
Stock High (intraday): 250.5

-----------------
Evidence of Insider Trading found -
Opening Stock 244.85 was less that Closing 247.39
 Using Top2Vec, Bates# from evidence found are:
 -- 859591bb-da09-455d-ae10-373cfcd14d82
 Using Tf-Idf, Bates# from evidence found are:
 -- 859591bb-da09-455d-ae10-373cfcd14d82
 Using Snips, Bates# from evidence found are:
 No Evidence found using Snips approach.
 Using KMeans, Bates# from evidence found are:
 -- 859591bb-da09-455d-ae10-373cfcd14d82
 No Evidence found using Word2Vec(SkipGram or CBOW).
Additional suspects for this case related to BatesID


  ABEBA
  ARNOLD
  BECK HECK
  BECKY CHEG
  BELLA
  CHEN
  CLINT PASKI
  DANIEL
  ELIAS
```

FIGURE 5: Screen of custom software for use by case investigator for Insider trading. (Trade stocks and names shown are purely for academic study and have no bearing on an event/person/investigation).

2) Labeled/Unlabeled stock trading data: Due to the unavailability of a public labeled dataset on stock trading intents, case investigators may need to create a labeled dataset for supervised learning. The proposed approach can be executed against any historical case evidence data to arrive at a labeled dataset. Similarly, few public online sources of trading discussions such as news articles, discussion forums and financial market watch comments can be assembled to build an unlabeled dataset. The BERT approach discussed in this research can then be applied to label this dataset. A manual review of intents can then be completed to validate the quality of data upon which the labeled dataset on trading intents can then be used for analytical experiments. All text data must be carefully processed for hyperlinks, emojis, gifs, emoticons, smiles, abbreviations, etc. [32]. Such data should not be left behind but rather processed into their textual equivalent. This can be an uphill task for the investigation team if there is no such labeled dataset to begin with, but once created; can be reused with periodic updates for many investigations.


Financial Fraud Investigation
Key Info for Case/Investigation/Audit
>
<
Re-Fetch Stocks
Add Date
Find Indicators of Insider Trading (Evidence)
Find Indicators of Pump & Dump (Evidence)
Risk Ranking Questionnaire
Suspect Metadata
Fetch Yahoo Fin
Case Data
EXIT
Export Report
Heatmap
Find Evidence against ESI for any suspect
Evidence of Stock Trading Intent (Buy/Sell) from Algorithms
Word2Vec CBOW
Word2Vec SkipGRAM
Top2Vec
KMeans
TF-IDF (BERT)
BERT
BOW (BERT)
Snips
Sentiment (Loughran McDonald)
Yahoo Fin Data
Rebuild ESI Tables
Re-execute All Algorithms

Suspects in ESI / Evidence

Abeba
Arnold
Beck Heck
Becky Cheg
Bella
Chen
Clint Paski
Daniel

Stock (Tickers) Found in ESI / Evidence

Stock	Description
A	Agilent Technologies...
ABCD	"Cambium Learning ...
ABT	Abbott Laboratories
AC	Associated Capital G...
ADS	Alliance Data System...

Stocks (Tickers) of Interest

STock	Description
ABCD	Cambium Learning Gr...

Key Dates of Interest (Ex. Corporate Meetings, Press Release, SEC Filings, etc.)

Thursday , February 24, 2022

08/28/2020
10/09/2020

Fin_Snips_Intent	Fin_Txt_4_Snips	BatesID	Intent	Probability
277	1	859591bb-da09-455d-ae10-373cfcd14d82	Buy	0.5431440...
278	2	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.6921140...
279	9	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
280	10	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
281	11	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5540463...
282	12	859591bb-da09-455d-ae10-373cfcd14d82	Buy	0.6314912...
283	15	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
284	17	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.6132793...
285	18	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
286	21	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
287	29	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
288	36	859591bb-da09-455d-ae10-373cfcd14d82	Buy	0.7294827...
289	40	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
290	41	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
291	51	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.4198701...
292	56	859591bb-da09-455d-ae10-373cfcd14d82	Buy	0.4995968...
293	64	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...
294	65	859591bb-da09-455d-ae10-373cfcd14d82	Sell	0.5611370...

No evidence of a P&D scheme found using Snips approach.
No evidence of a P&D scheme found using TF-Idf (BERT) approach.

FIGURE 6: Screen of custom software for use by case investigator for Pump and Dump (P&D) scheme. (Trade stocks and names shown are purely for academic study and have no bearing on an event/person/investigation).

3) Supervised/Unsupervised learning: The datasets used in the experiment were randomly picked and assembled to mimic typical case evidence and investigation. Twitter data, WhatsApp data, SMS data, emails, random custom MSWord documents, and Facebook data constitute the case evidence. Thus, the accuracy of models and results of the experiments were solely for demonstration of the approach. The BERT model achieved a 41% accuracy when predicting financial fraud intent. The accuracy of the TF-IDF model was found to be at 65% and BOW model was 70%. The probability of snips in determining a "buy" intent was 79% and "sell" intent was 78%. The Top2vec algorithm had a 0.2 for similar word score. The similarity of words to "buy" or "sell" found using Word2Vec was between 0.97 to 0.99. Figure 6 displays the ROC curve of the BOW approach for "Buy", "Sell" and "Other" while Figure 7 displays the ROC curve of the TF-IDF approach for "Buy", "Sell", and "Other". BOW method was employed directly against the evidence while TF-IDF was employed against the labeled Reddit data (after using BERT to label this data). Thus, we cannot compare the BOW approach against TF-IDF as both are employed against different datasets. Case investigators can ignore or retain a model based on spot checks and manual analysis of evidence. This umbrella approach provides investigators with various approaches towards determining fraud indicators.

5. STUDY BENEFITS & LIMITATIONS

This study does have a few limitations. The methodology in this study is limited to U.S. English language. However, this can be scaled into supporting other languages. The use of emoticons in today's electronic communications can convey a ton of information that can be used by criminal minds. Due to time limitations, this study skipped emoticons but accounted for emojis. Electronic communications also involve sharing of media, GIFs, and images. They can be used as covert channels of communication. Due to time limitations, this study skipped such data. Risk profiling of suspect along with risk ranking techniques calculations by the authors are for demonstration purposes and can be further improved.

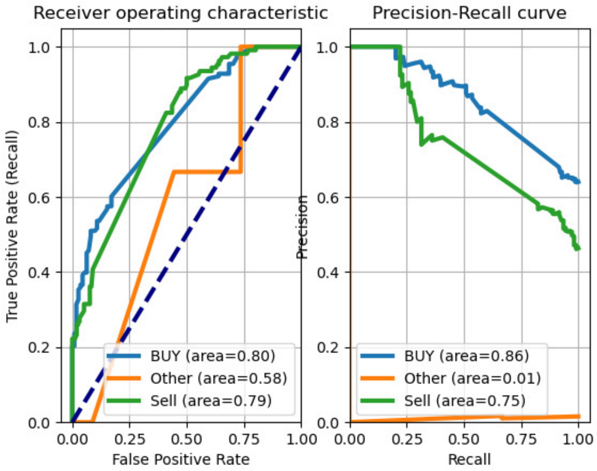


FIGURE 7: BOW approach – ROC, precision, and recall.

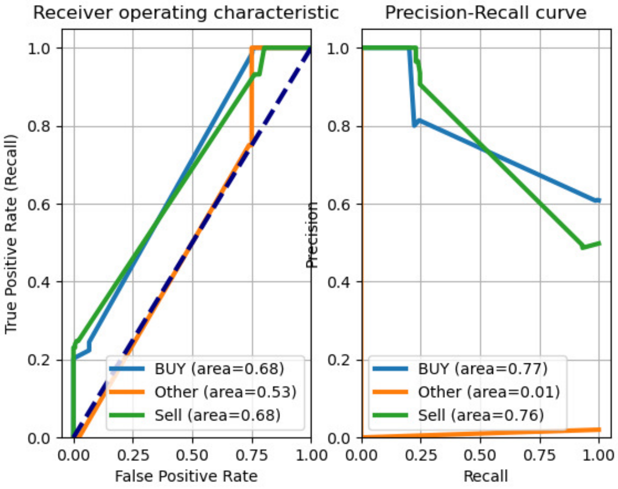


FIGURE 8: TF-idf approach – ROC, precision, and recall.

6. CONCLUSION

In this paper, the authors propose an approach that consists of multiple sub-approaches that together constitute a powerful tool for case investigators investigating financial frauds. The umbrella of sub-approaches leverages machine learning (supervised and unsupervised) algorithms, deep learning (transformers) techniques, risk profiling and sentiment analysis. Investigators can choose one sub-approach over another based on the results of each and the supporting indicators that they fetch. The authors apply the sub-approach against synthetic dataset that closely mimics real-world electronic case evidence such as Tweets, Facebook posts, emails, word documents, SMS texts and WhatsApp texts. These sources of data are notorious for deviating from traditional English language and thus the authors highlight the need to address the linguistic challenges in case evidence before applying analytical techniques. The sub-approaches work on the intent of a suspect towards internal trading and Pump and Dump (P&D) frauds. The authors propose pursuing the human intent during trading of stocks namely “buy” and “sell”. A suspect who is an employee of an organization (listed on the stock exchange) having privileged information from an event may trade stock (buy/sell). While insider trading is not always a cause for concern, misusing company privileged information can be investigated and is punishable. The proposed approach can narrow down the electronic document (bates number) that exhibits an intent to “buy” or “sell”. This intent when coupled with the job title of the suspect, risk profile, access to the key events, etc. can assist case investigators in building winning legal arguments for the case. Likewise, a suspect exhibiting a pattern of intent through a series of “buy” followed by an intent to “sell” can be deemed as a P&D. Thus, the approach of pursuing intent in both the fraud scenarios can assist investigators in pointing to the exact source of evidence (bates number)in the case evidence pile (ESI) thereby narrowing down the source and speeding-up the investigative process. In conclusion, this proposed analytical approach can help financial forensic fraud investigators, eDiscovery professionals, paralegals, and financial auditor’s process volumes of electronic data in short period of time thereby reducing investigation costs.

7. REFERENCES

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. Retrieved from https://arxiv.org/abs/2008.09470.

Aroussi, R. (n.d.). yfinance · PyPI. Retrieved February 4, 2022, from https://pypi.org/project/yfinance/.

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., … Dureau, J. (2018). Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. Retrieved from https://arxiv.org/abs/1805.10190v3.

Detecting Financial Statement Fraud. (n.d.). Retrieved January 23, 2022, from https://www.investopedia.com/articles/financial-theory/11/detecting-financial-fraud.asp.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. Retrieved from https://arxiv.org/abs/1810.04805v2.

Dorrell, D., & Gadawski, G. (2012). *Financial forensics body of knowledge*. John Wiley & Sons, Inc. Retrieved from https://books.google.com/books?hl=en&lr=&id=NFlP-vBIHXAC&oi=fnd&pg=PT10&dq=Financial+Forensics+Body+of+Knowledge+-+Darrell+D.+Dorrell,+Gregory+A.+Gadawski&ots=UyL_LdAsPu&sig=t6w1K9qIumEkzrDpDUqT2bv8Erc.

EDRM, D. (n.d.). Processing Guide. Retrieved from http://www.edrm.net/frameworks-and-standards/edrm-model/processing/.

Forensic Audit vs. Internal Audit: Differences in Accounting. (n.d.). Retrieved January 31, 2022, from https://www.eidebailly.com/insights/articles/2019/3/forensic-audit-vs-internal-audit.

Fritz, F. (n.d.). The Costs Of E-Discovery And What May be Recoverable Under 28 U.S.C. § 1920. Retrieved February 15, 2022, from https://www.jdsupra.com/legalnews/the-costs-of-e-discovery-and-what-may-36639/.

GitHub. (n.d.). nlu-benchmark. Retrieved February 5, 2022, from https://github.com/wenjingu/nlu-benchmark.

Hancox, S. J., & Dinapoli, T. P. (n.d.). *Red Flags for Fraud, State of New York Office of the State Comptroller*.

Insider trading - Wikipedia. (n.d.). Retrieved January 29, 2022, from https://en.wikipedia.org/wiki/Insider_trading.

Insider Trading FAQ Part 1. (n.d.). Retrieved January 31, 2022, from https://prisonprofessors.com/insider-trading-faq-part-1/.

Islam, S. R., Khaled Ghafoor, S., & Eberle, W. (2019). Mining Illegal Insider Trading of Stocks: A Proactive Approach. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 1397–1406. https://doi.org/10.1109/BIGDATA.2018.8622303.

Jiang, J., Chen, J., Gu, T., Choo, K. K. R., Liu, C., Yu, M., … Mohapatra, P. (2019). Anomaly Detection with Graph Convolutional Networks for Insider Threat and Fraud Detection. *Proceedings - IEEE Military Communications Conference MILCOM*, *2019-Novem*. https://doi.org/10.1109/MILCOM47813.2019.9020760.

Krishnan, S. (n.d.). Project · GitHub. Retrieved May 6, 2022, from https://github.com/kshsus.

Krishnan, S., Shashidhar, N., Varol, C., & Islam, A. R. (2022). Sentiment Analysis of Case Suspects in Digital Forensics and Legal Analytics. *International Journal of Security*, *13*(1). Retrieved from https://www.cscjournals.org/journals/IJS/issues-archive.php.

Krishnan, S., Shashidhar, N., Varol, C., & Rezbaul Islam, A. (2021). Evidence Data Preprocessing for Forensic and Legal Analytics. *International Journal of Computational Linguistics (IJCL)*, *12*(2), 24–34. Retrieved from https://www.cscjournals.org/library/manuscriptinfo.php?mc=IJCL-122.

Lauar, F., & Arbex Valle, C. (2020). Detecting and Predicting Evidences of Insider Trading in the Brazilian Market. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12461 LNAI*, 241–256. https://doi.org/10.1007/978-3-030-67670-4_15.

Lebichot, B., Borgne, Y.-A. Le, He-Guelton, L., Oblé, F., & Bontempi, G. (2019). Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection. *Springer*, 78–88. https://doi.org/10.1007/978-3-030-16841-4_8.

Li, T., Shin, D., & Wang, B. (2021). Cryptocurrency Pump-and-Dump Schemes. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3267041.

Liu, R., Mai, F., Shan, Z., & Wu, Y. (2020). Predicting shareholder litigation on insider trading from financial text: An interpretable deep learning approach. *Information & Management*, *57*(8), 103387. https://doi.org/10.1016/J.IM.2020.103387.

Nam, D. (2020). Modelling Stock Market Manipulation in Online Forums. Retrieved February 23, 2022, from https://qspace.library.queensu.ca/handle/1974/28239?show=full.

NLP vs. NLU: What's the Difference and Why Does it Matter? (n.d.). Retrieved February 15, 2022, from https://rasa.com/blog/nlp-vs-nlu-whats-the-difference/.

NLP vs. NLU vs. NLG: the differences between three natural language processing concepts. (n.d.). Retrieved February 15, 2022, from https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/.

NLU-benchmark/2017-06-custom-intent-engines at master · sonos/nlu-benchmark. (n.d.). Retrieved February 5, 2022, from https://github.com/sonos/nlu-benchmark/tree/master/2017-06-custom-intent-engines.

Officers, Directors and 10 percent Shareholders | SEC.gov. (n.d.). Retrieved January 31, 2022, from https://www.sec.gov/smallbusiness/goingpublic/officersanddirectors.

Open source conversational AI. (n.d.). Retrieved February 6, 2022, from https://rasa.com/.

Pump & Dump Schemes - Securities Fraud Attorneys. (n.d.). Retrieved January 31, 2022, from https://www.criminallawyergroup.com/practice-areas/securities-and-commodities-fraud/pump-dump-schemes/.

Pump and dump - Wikipedia. (n.d.). Retrieved January 29, 2022, from https://en.wikipedia.org/wiki/Pump_and_dump.

"Pump and dump" Schemes. (n.d.). Retrieved January 31, 2022, from https://www.sec.gov/rss/your_money/pump_and_dump.htm.

reddit.com: api documentation. (n.d.). Retrieved February 4, 2022, from https://www.reddit.com/dev/api/.

Reurink, A. (2018). FINANCIAL FRAUD: A LITERATURE REVIEW. *Journal of Economic Surveys*, *32*(5), 1292–1325. https://doi.org/10.1111/JOES.12294.

Roy, N. C., & Basu, S. (2021). Bank's battle against insider frauds ignitors and mitigators: an emerging nation experience. *Journal of Facilities Management*, *19*(4), 437–457. https://doi.org/10.1108/JFM-04-2020-0021/FULL/XML.

Samaneh Sorournejad, Zojaji, Z., Atani, R. E., & Monadjemi, A. H. (2016). A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective. Retrieved from https://arxiv.org/abs/1611.06439v1.

Securities fraud - Wikipedia. (n.d.). Retrieved January 29, 2022, from https://en.wikipedia.org/wiki/Securities_fraud.

Selective Disclosure and Insider Trading. (n.d.). Retrieved January 30, 2022, from https://www.sec.gov/rules/final/33-7881.htm.

Snips Natural Language Understanding — Snips NLU 0.20.2 documentation. (n.d.). Retrieved February 5, 2022, from https://snips-nlu.readthedocs.io/en/latest/.

Srivastava, S., & Bhatnagar, R. (2021). Process Mining Techniques for Detecting Fraud in Banks: A Study. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(12). Retrieved from https://turcomat.org/index.php/turkbilmat/article/view/8058.

Terblanche, M., & Marivate, V. (2021). Loughran McDonald-SA-2020 Sentiment Word List. Retrieved February 6, 2022, from https://researchdata.up.ac.za/articles/dataset/Loughran_McDonald-SA-2020_Sentiment_Word_List/14401178.

The Difference Between a Financial Statement Audit & a Forensic Audit. (n.d.). Retrieved January 31, 2022, from https://bizfluent.com/info-12085490-difference-between-financial-statement-audit-forensic-audit.html.

The Laws That Govern the Securities Industry. (n.d.). Retrieved January 30, 2022, from https://www.investor.gov/introduction-investing/investing-basics/role-sec/laws-govern-securities-industry.

Three Trends Driving Up E-Discovery Costs. (n.d.). Retrieved February 15, 2022, from https://www.forbes.com/sites/forbestechcouncil/2021/10/22/three-trends-driving-up-e-discovery-costs/?sh=2bb25be2724c.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, *57*, 47–66. https://doi.org/10.1016/J.COSE.2015.09.005.

Xu, J., & Livshits, B. (n.d.). The Anatomy of a Cryptocurrency Pump-and-Dump Scheme | USENIX. Retrieved February 23, 2022, from https://www.usenix.org/conference/usenixsecurity19/presentation/xu-jiahua.