Semi-Supervised Topic Discovery and Sentiment Extraction on Textual Feedback Systems

Aditya Gadepalli (19BM6JP08)
Srijan Gupta (19BM6JP19)
Anudeep Immidisetty (19BM6JP54)

Introduction

- Any consumer facing organization needs to constantly keep track of any **feedback**/**grievances** from the stakeholders.
- Any relevant feedback must be classified, and the **incumbent issues** need to be clustered into appropriate categories.
- In fact, any new categories need to be tracked in an **unsupervised** manner preferably.
- This **clustering** needs to be assessed over several **KPIs** that convey effect over business performance

Problem Statement

- We intend to prepare an end-to-end on-shot solution for **GST fraud** tracking, which **scraps text from online sources**, and categorize the given article into their fraudulent domain.
- Since there is no prepared dataset for it, we **prepared our own dataset** for the same using a **self-developed semi supervised feedback classification system**.
- We further develop the reason for the GST Fraud using **Topic Modelling, Classification and Clustering approaches**.

What is GST?

- GST (Goods and Services Tax) is a comprehensive value added tax on goods and services introduced on **1st July 2017** in India.
- It is collected on value added at each stage of sale/purchase in the supply chain and is hence a **seamless input tax credit system**.
- The Taxation is ultimately **borne by the final consumer**.
- The **total number of GST frauds** stand at **637**, till Feb 20th 2020 [ref].
- It is categorized under **four tax slabs** of 5%, 12%, 18% and 28%.
- Types include **Centre GST** (CGST), **State GST** (SGST), **Integrated GST** (IGST).

Project Outline

Data Preparation

1. Web Scrapping
from URLs of several **news providers** to get articles.

2. Preliminary EDA
is done over the data.

3. The Standard Corpus
is prepared by extensive **manual labelling** to help build a classifier to filter GST fraud articles from rest of the news in next step.

Data Filtering

1. Data Cleaning
using NLTK library for **stemming**, **lemmatization**, regex based filtering and custom stop words removal and data **contraction mapping (done first actually)**.

2. BERT Embedding
is used to embed each article into a finite size vector to capture **semantic meaning**.

3. Classification Model
is built using **cosine similarity** over the standard corpus to filter any **incoming articles** being published on news websites as GST fraud articles or not.

4. Hyperparameter Tuning
is performed on the above classifier and the **precision** and **recall** are greatly improved in this step. Also a **Latent Semantic Analysis (using PCA)** is performed to understand semantic relations between articles in a better way.

Data Exploration

1. Secondary EDA
is performed post the filtering. A **word web** is also generated to visualize the fraud categories.

2. Sentiment Analysis
is performed on the articles to understand the degree of **severity** each article poses in terms of the attention it can draw and the magnitude of the fraud. This helps us understand the **priority** of assignment of these topics to the respective departments for timely resolution/review of the issue.

3. Complex Network
is built to understand the **relation** between various aspects of GST fraud.

Categorization

1. Topic Modelling
is performed using **LDA (Latent Dirichlet Allocation)** and **iDF (inverse Document Frequency)** to identify the categories of GST fraud in a completely **unsupervised manner.**

2. Cluster Analysis
is performed on the articles using **Hierarchical clustering** determine any sub-categories or hierarchies.

3. Classification
is done over extracted RoBERTa features to allocate articles to respective **fraud categories** (concerned departments) and the results are manually reviewed, and the feedback is updated once model is in **deployment**.

Evaluation

1. Standard Metrics
such as **Top-3 accuracy** are used for model evaluation.

2. Indicators and Benchmarking
are thoroughly assessed. We propose **custom indicators** based on growth rate in number of misc. article allocations, **identification of new fraud categories**, weighted misclassifications (by fraud amount) etc. and also benchmark our model performance against human performance both in terms of **accuracy** and **scale**.

3. Going Further
is also presented wherein once the model is under deployment, **continuous feedback** is collected about what solutions have been used on which particular issue. This will help is developing **a fully scalable dashboard** that can **propose solutions** as soon as the fraud is classified.

Database Preparation

Web Scraping

Goal: Fetch **GST fraud articles** using the keyword *"GST" / "GST Fraud"* as the **primary search phrase**.

Libraries: BeautifulSoup4, newspaper3k

- We perform **web-scraping** to collect news articles from several sources* using python libraries.
- The URLs typically include GST articles, advertisements, social media links, and redirects to the same page.
- The links in the page were extracted using **regular expression** after loading the whole page's script using Beautiful Soup.

***Sources:** Economic Times, Financial Express, Livemint, Times of India, Hindustan Times, Business Standard, The Hindu, The Hindu-Business Line, Deccan Herald, Zee News, India Today, Indian Express, NDTV, Money Control, Business Today, Bloomberg, The Print, Google News, The Tribune, and GST Compendium.

Database Preparation

Preliminary EDA and the Standard Corpus

Preliminary EDA

- A preliminary EDA reveals that there is a **need for extensive cleaning**.
- We perform **secondary EDA** with extensive visualizations later.
- We note that the **headlines are not complete indicative** of the content or the sentiment of the articles.
- The **writing style largely varies** between authors and websites.
- The **sentiment** largely remain **same within a website.** This will be explained later.
- The **IPC Sections** applicable for certain kinds of GST frauds are also mentioned (eg. Section 67-A).

Standard and Reference corpus

- **A Standard Corpus** was prepared with total 50 articles, 25 for the favorable class (i.e. GST Fraud Case) and 25 non favorable articles (a mix of GST non fraud, and a few advertisement and social media links).
- **A Reference Corpus** is also prepared consisting of 75 articles on the desired topic

Data Filtering

Data Cleaning

Library: NLTK

- **Stemming and Lemmatization** are performed to obtain the root words of the textual data.
- **Regex based filtering** and **custom stop words removal** are performed to remove any unnecessary whitespaces, special characters, spam words etc.
- **Data contraction mapping** is performed to map spoken short word forms, mis-spelt words and abbreviations into regular textual English, singular spellings and full-forms respectively.

Example:

- Contractions like **he'd, she'll've,** and **they'll've** upon punctuation removal becomes **hed, shellve,** and **theyllve**, which is neither a logical word nor is specifying the context it serves.
- So using contraction mapping, they were developed into **he would, she will have and they will have** resp.
- Abbreviations like **GST, rs, cr, etc** were also converted to their respective full forms using the same mechanism.
- Inconsistent spelled words like **adhar, aadhar and aadhaar** were unified into a single entity, aadhaar.

Tax officials detected Rs 37,946 cr worth of tax fraud in 2019\n\nand Rs

→ Cleaning special characters and brackets using regular expression

tax officials detected rs 37946 cr worth of tax fraud in 2019 and rs

→ Mapping for the corpus

tax officials detected rupee 37946 crore worth of tax fraud in 2019 and rupee

→ Stop words and words of character lengths less than 3 were removed

tax officials detected rupee 37946 crore worth tax fraud 2019 rupee

Data Filtering

Embeddings, Classification and Hyperparameter Tuning

Embedding Generation

- Considered **TF.iDF, Word2Vec** and **RoBERTa** to obtain the best performing embedding.
- Used RoBERTa because it generated the best evaluation metric value on the standard corpus.

Reason for best performance: **attention ability of RoBERTa** best captures the necessary context and semantic content of the articles, which was essential for our similarity comparison.

Evaluation Parameter Selection

- Precision: selected articles should be relevant to the target class.
- Recall : essential to capture as many relevant articles from the corpus possible.

Latent Schematic Analysis

- Reduced feature space, using PCA, to better capture schematics of articles.
- Not used, since it did not improve evaluation scores further and would only act as a redundant process.

Similarity

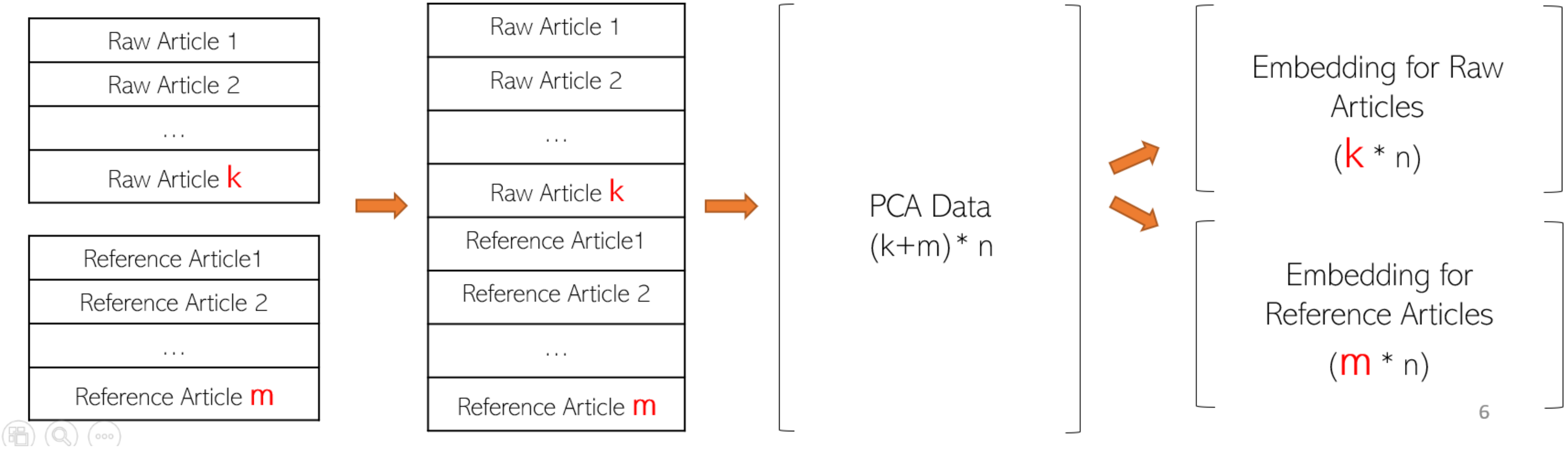
- Developed using cosine similarity and the Maximum similarity pairs are taken to compare results.
- Similarity threshold set for classification. Threshold value selected for value which gave the maximum f1 score .

Model Results (on test corpus)

- **Precision : 0.96**
- **Recall: 0.926**
- Able to even detect 2 articles which were incorrectly labelled (intentionally labelled incorrectly to replicate human performance), **thereby surpassing the human benchmark.**

Data Filtering

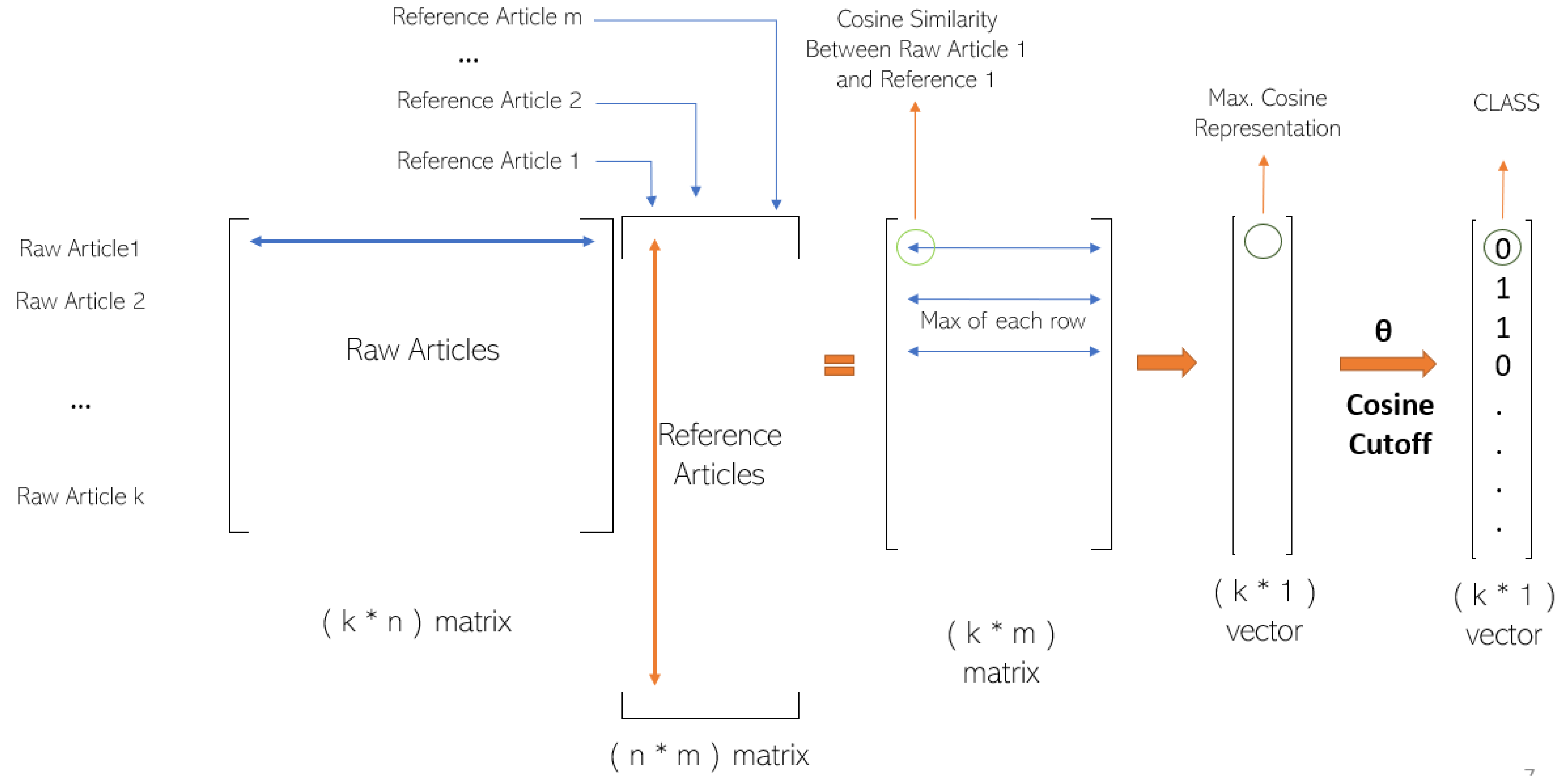
Embeddings, Classification and Hyperparameter Tuning



Data Filtering

Embeddings, Classification and Hyperparameter Tuning

For a system with **n** dimensions for each sentence
Where each sentence vector is a unit vector (achieved upon normalization)



Data Exploration

Secondary EDA

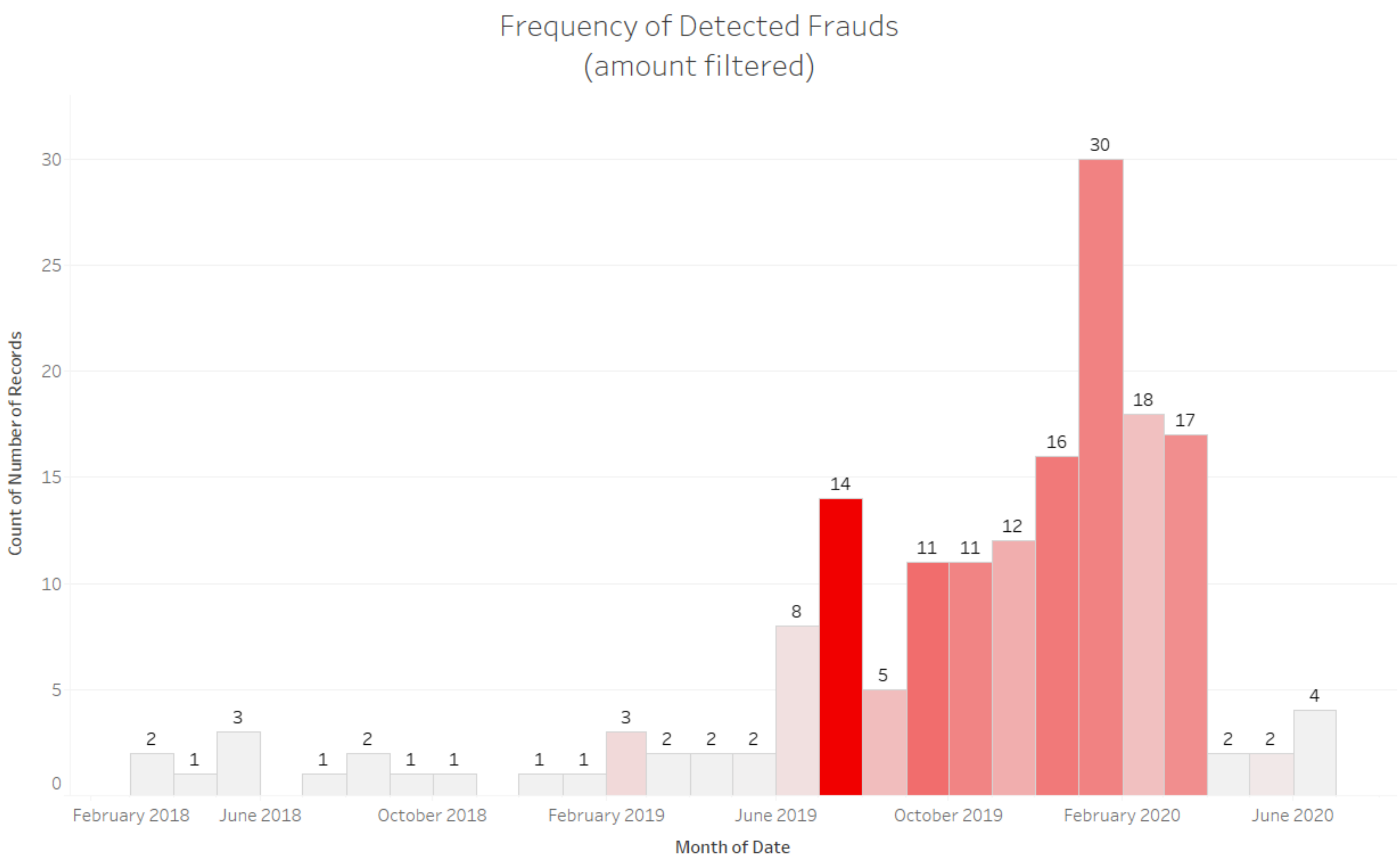
Word Cloud: Provides a visualization of keywords across the corpus.

A few are listed below:

- Tax credit
- Input tax
- Investigation
- Goods services
- Companies
- Fake invoice
- Accused
- GST
- Firm
- Fraud

Data Exploration

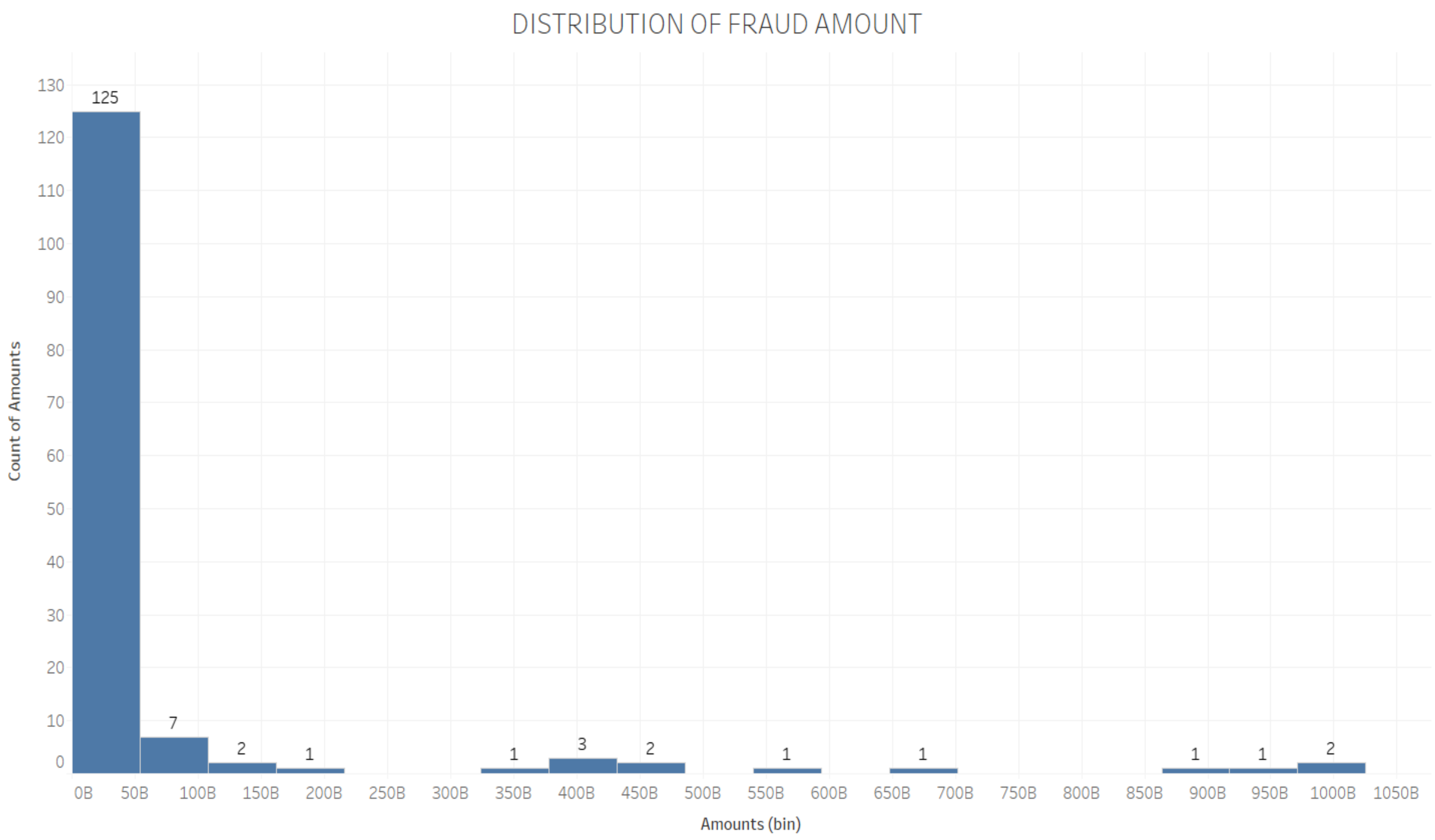
Secondary EDA



- The **biggest fraud** (in terms of monetary value) were reported in the period for July 2019
- **Maximum frequency of cases** was recorded for the period Jan 2020 to Feb 2020
- **Subsequent decline** in the number of cases for the following period due to introduction of **lockdowns** in India, due to the **COVID-19 pandemic.**
- Upon **lockdown relaxation**, we may begin seeing a rise in the number of GST fraud cases.

Data Exploration

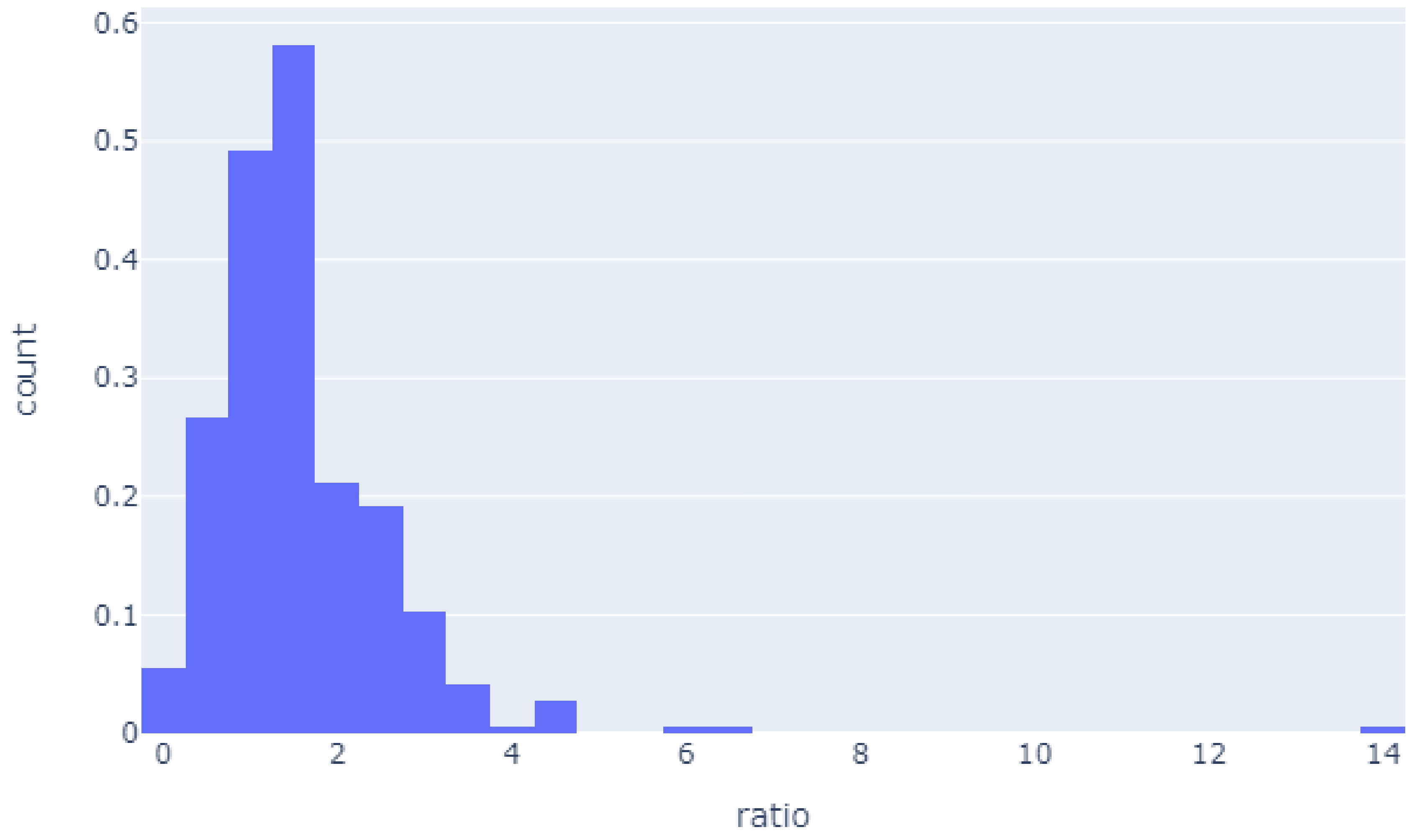
Secondary EDA



- Looking at the distribution, we see **majority of the fraudulent transactions are relatively on the lower end** of the amount involved (in the bin 0-50 Bn INR)
- Maximum fraud recorded for a single fraud event was recorded **at 97.1 thousand crore**, from the developed corpus.

Data Exploration

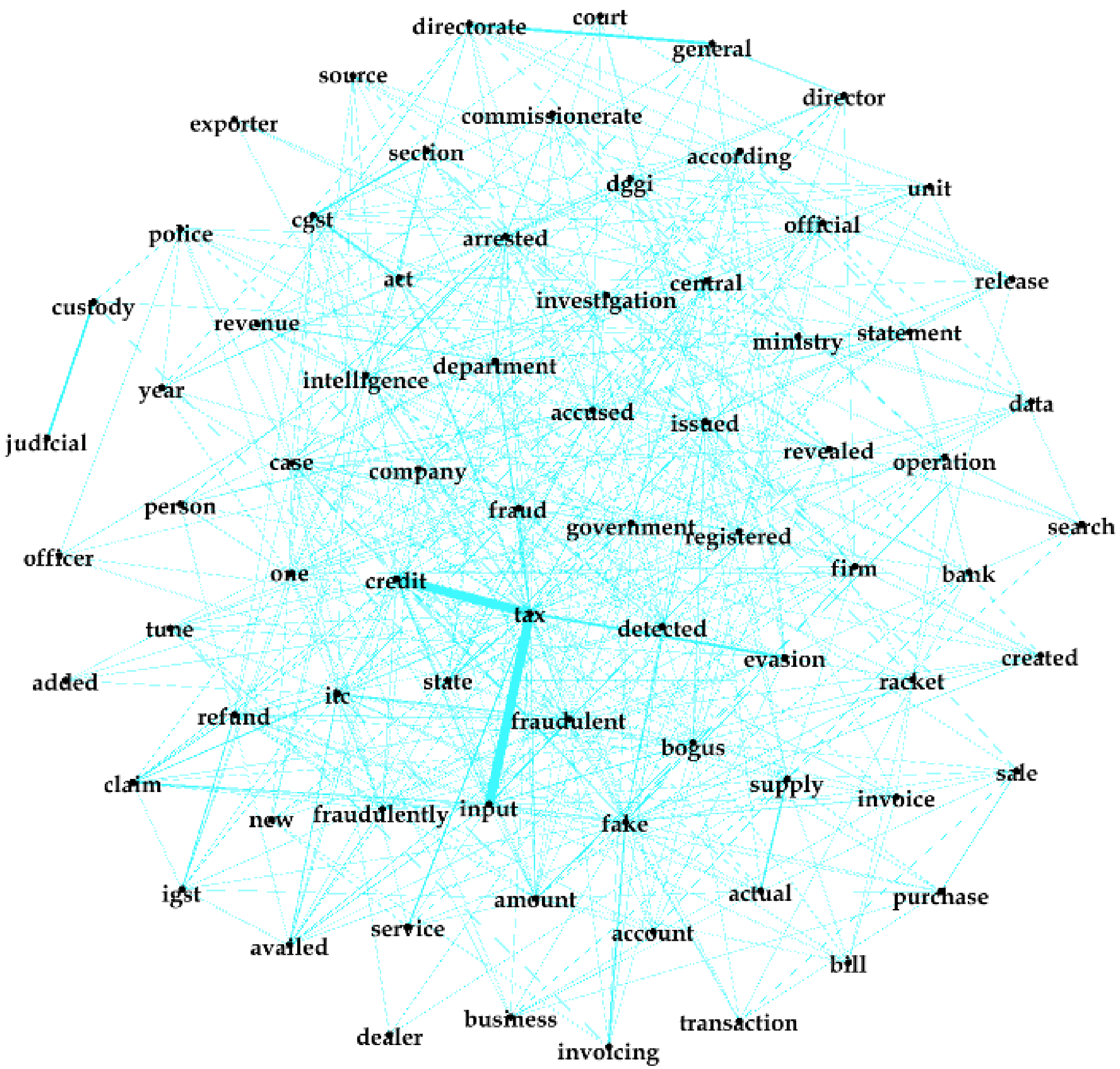
Sentiment Analysis



- We used ***NLTK's SID Polarity*** in this for sentiment analysis
- 70% articles have a **predominantly more negative sentiment than a positive sentiment.**
- This bias in sentiment scores highlights **media sentiment on the GST frauds.**
- **Turning point for us**
  - It showed that the semantic part of the articles is **quite important** to capture and predominantly categorize whether a given article is of favorable class or not (which is not always guaranteed in a supervised classifier)
  - But **this alone is not sufficient** to categorize an article a raw article as fraudulent or not.

Data Exploration

Complex Network



- **Network analysis** used to depict relations among factors and to analyze the social structures that emerge from the recurrence of these relations
- Gephi v0.9.2 is used to chart the co-occurence matrix of bigrams.
- A *"Fruchterman-Reingold"* rearrangement is used to depict nodes with higher node-centrality at the geometric centre of the graph using a gravity factor of 10.
- This graph helps assess **most common bigrams** in data
- Shows "input tax" and "tax credit" are most common word pairs
- Gave us an idea **on potential fraudulent areas**
  - Judicial custody
  - Tax evasion
  - Bogus claims
  - Fake amount
  - Fake invoicing
  - Fake transaction etc.

Categorization

Topic Modelling using Latent Dirichlet Allocation

To perform Topic Discovery over the refined corpus of articles, we now proceed with topic modelling using Latent Dirichlet Allocation (LDA). LDA is a text mining method based on “Bayes Hierarchy Model” first proposed in 2003.

The generative process of LDA:

1. Take a topic from a document；
2. Take a word from the chosen topic from 1；
3. Repeat 1 and 2 until every single word was matched with a topic in the document.

- The data is first filtered to retain only nouns and adjectives as they usually comprise the topics-words of our interest.
- The major topic of a document is inferred from the distributions of “document-topic” and “topic-word”.
- From the above distributions, we obtain a set of topics (comprised of relevant topic-words) in this **unsupervised** way.
- The number of topics needed is also obtained in an unsupervised manner using “**Hierarchical Dirichlet Process**” (HDP).

Libraries: Gensim, NLTK

Categorization

Topic Modelling using Latent Dirichlet Allocation

The obtained output is presented below, from which we intuitively assign topic labels as depicted through two examples below.

(1, '0.033*"tax" + 0.032*"gst" + 0.013*"credit" + 0.009*"input" + 0.008*"invoices" + 0.008*"crore" + 0.008*"gstr" + 0.006*"evasion" + 0.006*"companies" + 0.006*“bill"'), → Tax Evasion

(4, '0.027*"tax" + 0.026*"gst" + 0.024*“trading" + 0.023*"fake" + 0.019*"firms" + 0.019*"goods" + 0.017*"crore" + 0.016*"credit" + 0.013*"input" + 0.011*"companies"')] → Fake Firms

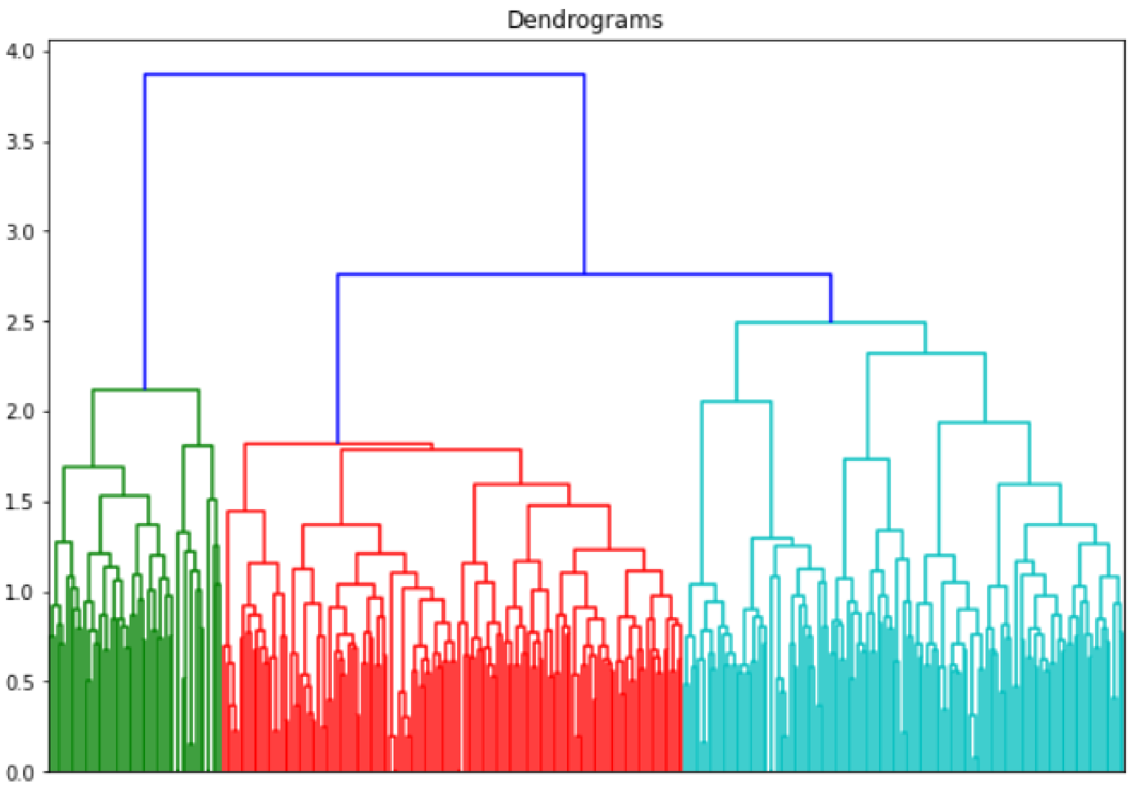
List of obtained topics

1. Tax Evasion
2. Fake Invoices
3. Fake Firms
4. Claims Without Receipts
5. Info
6. Others

Cluster Analysis

Hierarchical clustering

A clustering Analysis is performed to understand the relation between the obtained topic classes and their prevalence.



This represents “**Tax Evasion**” class and we observe that it’s the **third** most prominent type of fraud

This represents “**Fake Invoices**” class and we observe that it’s the **second** most prominent type of fraud

This represents “**Fake Firms**” class and we observe that it’s the **most prominent** type of fraud

Classification

Topic Allocation

- We perform a multi-class classification over RoBERTa (A Robustly Optimized BERT Pretraining Approach) features.
- The class imbalance is considered several classifiers are experimented with to obtain best classification results.
- The features were developed using Latent Semantic Analysis to yield 25 features and was able to account for 68.34% variance in the data.
- The number of features were decided to optimize the evaluation metric accuracy for the developed classifier.

Evaluation

Metrics

The classification results, for different models, are as follows:

Model	Train_Accuracy	Validation_Accuracy	f1
Random Forest	1.000	0.704918	0.686565
bagged LR	0.684	0.655738	0.650781
Naive-Bayes	0.656	0.639344	0.629129
XGB Classifier	1.000	0.622951	0.620330
LightGBM	1.000	0.622951	0.618613
Ada boost	0.760	0.606557	0.616051
Logistic Regression	0.688	0.590164	0.582434
Bagging	0.980	0.557377	0.561949
KNN	0.664	0.557377	0.534754
Decision Tree	1.000	0.442623	0.454486

A 73.77% accuracy upon hyperparameter tuning of Random Forest Classifier is noted as the best performing classifier.

Evaluation

Indicators and Benchmarking

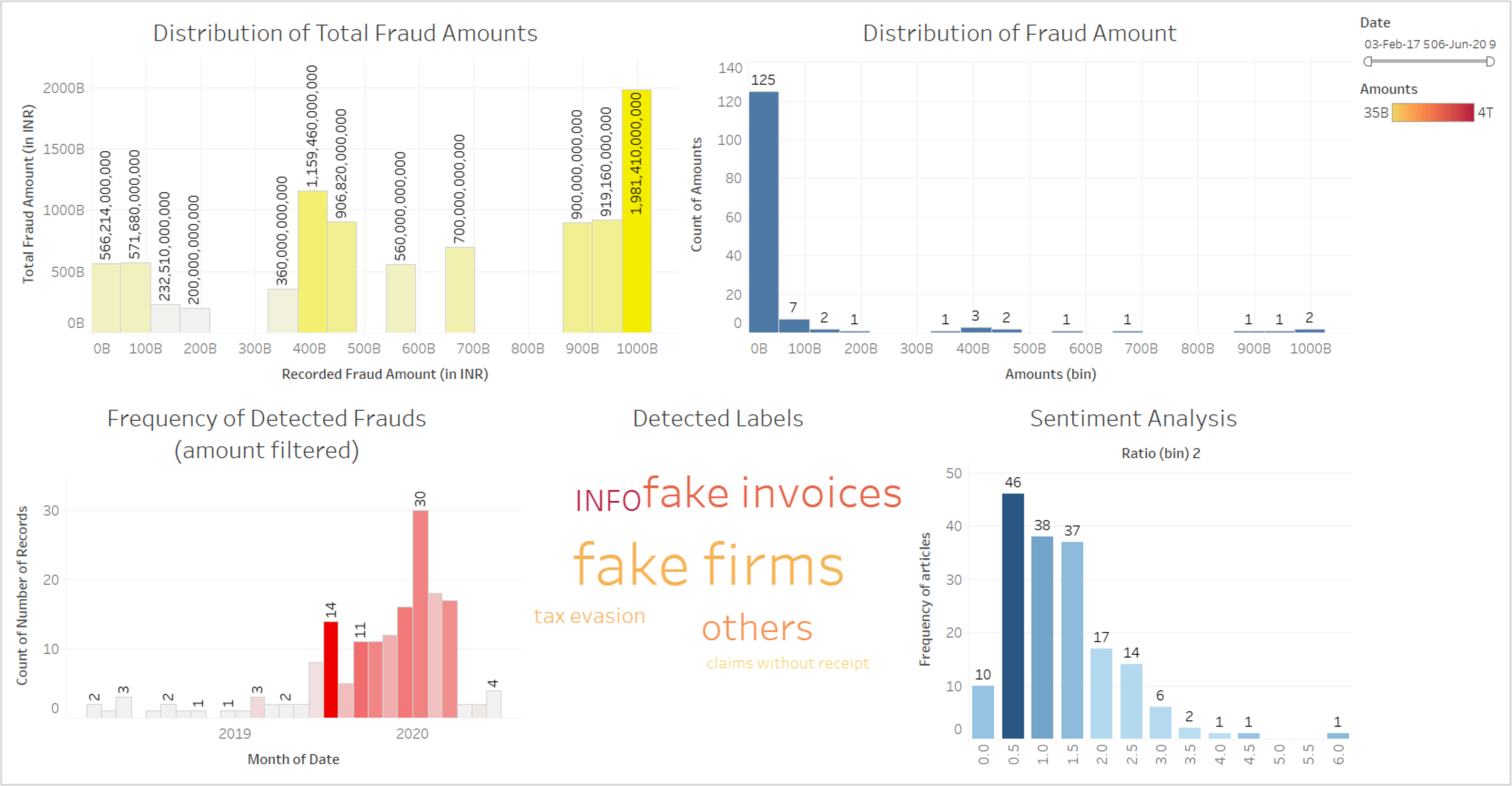
nltk_compound	nltk_neg	nltk_pos	nltk_neuteral	ratio	amount	labels
-0.9565	0.134	0.049	0.817	2.271186	400000000	fake invoices
-0.9524	0.126	0.079	0.795	1.41573	1E+09	fake firms
-0.693	0.056	0.056	0.889	0.848485	1.125E+11	INFO
-0.9918	0.128	0.071	0.801	1.580247	8E+09	fake invoices
0.9686	0.017	0.071	0.912	0.209877	3.5E+10	others
-0.9805	0.168	0.06	0.772	2.4	280000000	fake invoices
-0.9337	0.109	0.067	0.824	1.415584	1.2E+10	fake firms
-0.9027	0.109	0.065	0.826	1.453333	4.5E+09	fake invoices
0.872	0.04	0.06	0.9	0.571429	2E+10	others
-0.8591	0.075	0.053	0.872	1.190476	560000000	fake firms
-0.9819	0.132	0.047	0.821	2.315789	690000000	fake invoices
-0.9918	0.131	0.034	0.836	2.977273	1.2E+11	fake firms

- We see that the "*nltk_compound*" and the "*amount*" column are very useful indicators of the **severity** of the GST fraud case that has been reported.
- We propose **custom indicators** based on growth rate in number of misc. article allocations, **identification of new fraud categories**, weighted misclassifications (by fraud amount) etc. and also benchmark our model performance against human performance both in terms of **accuracy** and **scale.**
- The Topic modelling can be now applied within the classified label categories to discover **newly evolving sub-categories.**
- Under the "**others**" label we discover **new GST frauds** such as those seeking illegal benefits from foreign tourists' GST benefits.

Evaluation

Going Further

- Once the model is under deployment, **continuous feedback** is collected about what solutions have been used on which particular issue.
- This will help is developing **a fully scalable dashboard** that can **propose solutions** as soon as the fraud is classified.



Dashboard Preview and Features

Allows selective features on the following category (single or in combination of one another)

- Over a time period (on discrete and continuous scales)
- Over a fraud label category
- Over the class of fraud (based on amount of frauds)
- Over the sentiment of articles

Other Applications

Analyzing Reviews

- The current work discussed about the GST application in detail.
- We further provide an overview of how the same mechanism can prove useful for other use-cases like **an e-commerce** website like **Amazon.com** or a **hotel booking** website like **OYO** or even a **mobile application platform** like **Apple's AppStore**.
- The **reviews are analogous to articles**. One may include **tweets** and **blog posts** too; the cleaning process would be much simpler for these.
- The sentiment extraction can be done in the exact same manner.
- Additionally, the **star ratings** can also be accounted for and correlation can be established between top keywords and the average ratings.
- The topic discovery model would now give us a wide classification of user reviews say on the grounds of **duplicate products, broken items, delayed delivery, improper refunds** etc.
- This mechanism would speed up the process of **grievance addressal on platforms of such huge scale** such as Amazon.com

Questions?

Thank You!