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

Aditya Gadepalli[a], Srijan Gupta[a], Anudeep Immidisetty[a]

[a]*Post Graduate Diploma in Business Analytics, Jointly offered by IIT Kharagpur, ISI kolkata and IIM Calcutta*
Guide: Prof. Sujoy Bhattacharya, Associate Professor, Vinod Gupta School of Management
Project Repository: https://github.com/AdityaGadepalli/DSL2020

Abstract

Consumer facing organizations need to constantly keep track of feedback/grievances from their stakeholders. Relevant feedback must be classified and incumbent issues need to be clustered into appropriate categories. This clustering needs to be assessed over several indicators and benchmarks that convey effect over improvements in functional and operational utility. We propose a semi-supervised approach for performing relevant news article selection using extensive text cleaning, pre-trained RoBERTa embedding and Latent Semantic Analysis, followed by extensive exploratory data analysis, complex network analysis and sentiment extraction on the developed corpus. The processed articles are subjected to Latent Dirichlet Allocation, Hierarchical Dirichlet Process and Clustering to discover potential topics in an unsupervised manner. We then label the articles accordingly to devise a training set for building a multi-class classifier for topic categorization and assignment. The Top-3 accuracy metric is chosen along with several other indicators to understand the performance of the classifier and also ascertain the severity and other aspects of the incoming news articles. New topic discovery is also performed to improve the classifier classes and the model is re-calibrated at appropriate intervals to sufficiently stay above human benchmark in scale and accuracy.

Keywords:
RoBERTa Embedding, Sentiment Extraction, Latent Dirichlet Allocation, Hierarchical Dirichlet Process, Hierarchical Clustering

1. Introduction

The feedback/grievances of a consumer facing organization is ever evolving, with continuously arising newer categories with time. These newly evolving categories are tracked, usually in an unsupervised manner. Among a major way to express public opinion, news and perceptions about the same is using textual feedback systems, in the form of user reviews, news articles, tweets, LinkedIn feeds, Facebook posts and blog posts. With this mechanism, we intend to explore these developed insights in detail and in a well-structured manner. To highlight the working of our mechanism, we made use of GST Fraud as a use case scenario.

India, previously, had a cascading system of taxation, where taxes were extracted both by the central and the state government independently. Also, there were a variety of Value Added Tax (VAT) laws, Entry Taxes, Octroi etc. in the country with disparate tax rates and dissimilar tax practices, which divided the country into separate economic spheres. This greatly hindered the free flow of trade, and related foreign investors to invest in the country. (2)

GST (Goods and Services Tax) is a comprehensive value added tax on goods and services which was introduced on 1st July 2017 in India. All the taxes mentioned earlier got subsumed into a single taxation system called the Goods and Services Tax (GST) which were now being levied on supply of goods or services or both at each stage of the supply chain, starting from the manufacturing stage or imports, till the end consumer level. In summary, any tax levied by the Central or State Government on the supply of goods or services converged into GST, where the final taxation is ultimately borne by the final consumer. (2) The stated intent of these reforms was to formalize the Indian economy, giving a boost to foreign investments, harmonization of laws, procedures and rates, help improve tax buoyancy and improve environment for compliance.
The broad domains of GST fraud available are as follows:
Tax evasion: illegal activities in which the person or entity deliberately avoids paying its liable tax. The ones caught evading taxes are subjected to criminal charges and substantial penalties.
Fake invoices: if any of the vital information required in a typical GST invoice is missing, it can be said to be a fake invoice. Every registered business under GST must issue an invoice which contains a valid GSTIN. If a business has not been registered under GST but uses a fake GSTIN on the invoice and charges GST, it will also be considered a fake invoice. (6)
Fake firms: Creation of firms that are not registered with governing authorities, or falsely registered
Claims without receipt: refers to those claims which does not have any proof of legal transaction acceptable under the GST act, and are non-valid claims (1)

Contact: adityagadepalli@iitkgp.ac.in (Aditya Gadepalli), srijangupta@iitkgp.ac.in (Srijan Gupta), ianudeep@iitkgp.ac.in (Anudeep Immidisetty)
ID: 19BM6JP08 (Aditya Gadepalli), 19BM6JP19 (Srijan Gupta), 19BM6JP54 (Anudeep Immidisetty)

Others: those domains of frauds which do not categorize in any of the above categories.

General Mechanism of Fraud Detection: To understand the process, it is necessary to understand how frauds are detected. The GST filing cases are first reviewed by different agencies of the Central Board of Indirect Taxes and Customs (CBIC), majorly by the two premier intelligence agencies Directorate General of GST Intelligence (DGGI) and the Directorate General of Revenue Intelligence (DRI). The use of data analytics is quite intensive in the detection of fraud cases, to scrutinize all past and pending refund claims filed all over the country for inverted duty structure. The potential cases are analyzed in detail by experts and appropriate actions are taken, on the magnitude and severity of the crime. (5) The total number of GST frauds stand at 637, till Feb 20th 2020.

2. Data Preparation

Since there is no pre-existing dataset on this topic, we developed our own corpus. This step involves the mechanism used to filter the necessary text from the corpus, relevant to our application.

The idea behind this mechanism is similar to how a human being would work. To classify a given article is of a favorable class, a human being already has a reference, idea and/or concept in his mind. He/she upon seeing a new article would compare this article with the reference that is already preconceived with him/her, and make a judgement. Similar is our overall approach. We provided a select reference corpus which the system uses and compares the new corpus with, to classify accordingly. We also control the extent to which the corpus should be similar to be classified.

Favorable class of articles for us, here, are those articles that talk about a particular GST fraud occurrence, where a fraudulent person or entity is involved, and the victim here is the tax collection agency (and/or GST Council) of India.

2.1. Data Extraction

The primary aim for us was to get GST fraud articles that talked about a fraudulent GST event. The available GST fraud articles are to be selected out from all the available list of links. So, all the http links, under the topic "GST" were extracted from a varied range of sources. These links typically included GST articles, advertisements, social media links, and redirects to the other pages of the same website. The links in the page were extracted using regular expression after loading the whole page's script using Beautiful Soup.

2.2. Basic Exploratory Analysis

We explored the GST articles in different online news reporting, namely 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, Business Today, Bloomberg, The Print, Google News, The Tribune, and GST Compendium.

There are different variants of GST fraud articles available. For example, this article (3) talks about whether Aadhar services can help mitigate GST fraud, which falls under the topic of GST Fraud but does not fall in our favorable class. Similarly this article (4) talks about GST fraud details in aggregate, and not a particular use case. Such cases are unfavorable for us. But these articles do not refer to the occurrence of a GST fraud, which is the desired result. So, the headlines were not the complete indication of the content or the sentiment expressed in the articles. there was a need for a mechanism to select the desired class of articles from the corpus of entire articles. The writing style largely varied between authors and websites, but the sentiment largely remained the same within a website.

GST is a relatively new domain, introduced only in 2017. The reported GST fraud cases are also limited in number (in a few hundreds). For a good feature representation of the given articles in feature space ideally requires a large number of feature representation (sometimes possibly in a few thousands). Hence, developing a supervised classifier u sing s o m any features and so fewer data points would result in imperfect modelling, and ineffective classification and bias in the model would be prevalent. This motivated us to consider a semi-supervised approach based on similarity of articles to select favorable class articles from a corpus of mixed classes of articles.

A preliminary EDA reveals that there is a need for extensive cleaning, due to presence of special characters, abbreviations, contractions and improper formatting of text upon extraction, since the system would not be able to understand them properly. A secondary EDA was also performed upon the extracted relevant articles, which will be discussed later in the paper.

2.3. Selection Mechanism

Data Cleaning: We began the operation with text cleaning. Special characters (brackets, slashes and symbols) were removed from the raw text to make the text more structured and meaningful. Regular expression based filtering and custom stop words removal were performed to remove any unnecessary whitespaces, special characters, spam words etc. Data contraction mapping was performed, using a manual amp mechanism, to map spoken short word forms, mis-spelt words and abbreviations into regular textual English, singular spellings and full forms respectively. For 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 represented. Abbreviations like GST, rs, cr, etc were also converted to their respective full forms using the same mechanism, and inconsistent spelled words like adhar, aadhar and aadhaar were unified into a single entity (here aadhaar). Then stop words were removed, since they act as a noise for the given text, and kept only meaningful tokens. Words of character lengths less than 3 were them removed from the text corpus. For instance, there were mentions of words like "Mr. X" that can mislead the embedder, and give false and significant weights to non-meaningful words. A sample of the raw text and the processed outputs at each step are as follows:

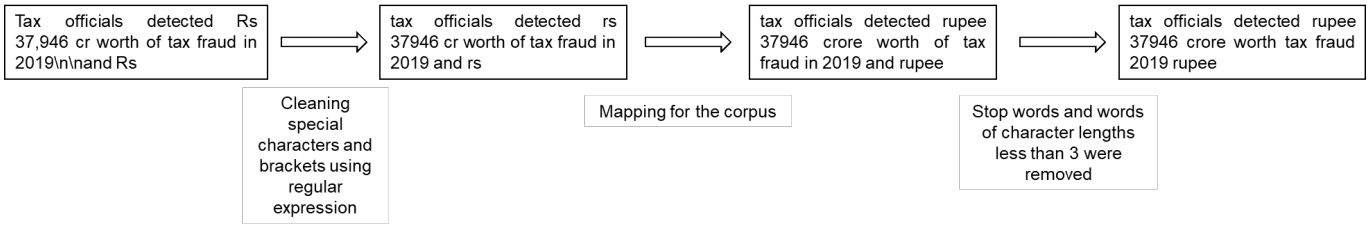

Figure 1: Step-wise cleaning mechanism

Standard Corpus: To develop the most optimal model and check model's effectiveness on an unknown, a standard corpus was prepared with 50 articles, 25 for the favorable class (i.e. GST Fraud Case) and 25 non favorable articles (mix of different kinds of articles, including few advertisement and social media links). We developed a reference corpus that consists of 75 articles on the desired class of articles. This corpus was used to compare the articles in the raw data with, and make the appropriate selection.

Evaluation parameter: Both precision and recall were independently important for this operation; high precision value because the selected articles should be relevant to the target class. Good quality data is essential for subsequent processes. High recall value is essential to capture as many relevant articles from the corpus possible.

Embedding Generation: The embeddings we explored were TF.iDF embeddings(using TF.iDF vectorizer of Sci-Kit Learn's Package), Word-to-vector embeddings(using Gensim Package and averaging over all words in a text to yield article embeddings) and RoBERTa embeddings (developed using PyTorch's Tokenizers). Ultimately, RoBERTa embedding was selected as it was able to generate the best precision and recall scores for the standard corpus. RoBERTa-base with mean-tokens pooling served the purpose quite precisely. A reason for the effectiveness of the embedding mechanism could be attributed to its ability to focus attention to the contextual and semantic parts quite accurately in the articles, and hence resulting the comparison between the articles along the similar aspects.

Latent Semantic Analysis: Latent semantic analysis (LSA) is a technique used for analyzing relationships between a set of documents and the terms they contain by producing a combination of the available set of concepts related to the documents and terms. Here, it was achieved using dimensionality reduction. These approaches do not center the data before computing the dimensionality reduction operation. Hence, data was first centered, then the features were then generated using Principal Component Analysis (PCA) transformation on the existing data, and explored for reduced dimensions 20, 50 and 100. The maximum dimension was restricted till hundred as the available corpus had only 124 articles with 1024 RoBERTa features. Exploring features more than 124 would be incorrect. Ultimately, since the process failed to improve evaluation metric values, this approach was not further explored for selection of the desired class of articles.

Developing Similarities: For a given set of raw corpus, we compute the cosine similarity between all possible pairs of raw

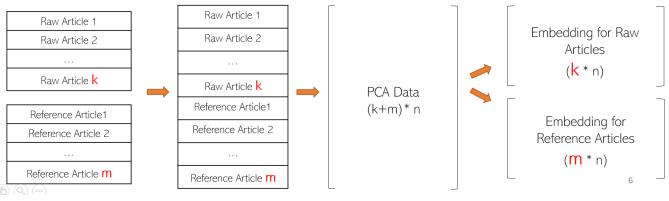

Figure 2: Depicting the mechanism for LSA

articles to a reference article, using the developed semantic embeddings of the pair of sentences. The data of each embedding array was first L2 normalized to make each sentence vector a unit vector in magnitude. For each document, the following operation was done to develop all possible pairs of summary-full text sentence pair's cosine similarity.

$$C_i = A_i^T B_i \tag{1}$$

A_i : Matrix containing normalized raw article embeddings for document i (dim: m^*1024)

B_i : Matrix containing normalized reference article embeddings for document i (dim: n^*1024)

C_i : Matrix with each element ($< j, k >$) representing cosine similarity if j^{th} raw article and k^{th} reference article for document i (dim : m^*n)

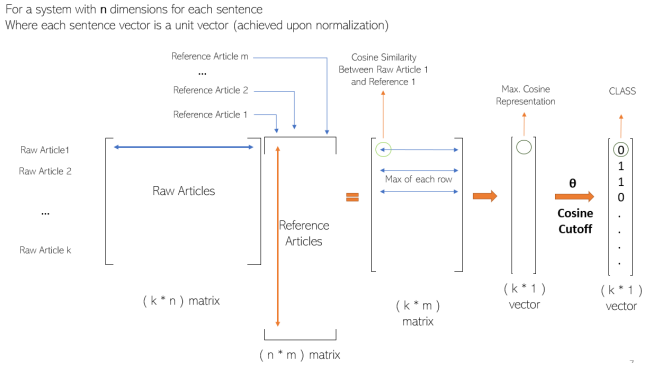

Figure 3: Describes the mechanism for developing similarity

Then, we find the maximum cosine similarity of a given raw article to any reference article, and develop this for all the raw articles. The similarity threshold was decided to give the best value of our evaluation metric.

Model Results: The precision and recall achieved on the test corpus was 0.96 and 0.926 respectively. This classifier was even able to detect 2 articles which were incorrectly labelled (intentionally labelled incorrectly) to replicate human performance, thereby surpassing the human benchmark. Testing the model's extraction on one day's archives of Times of India were also at par, further validating our mechanism for deployment. Using this, 2800+ GST articles were compared to ultimately yield a corpus of 294 relevant class articles.

3. Understanding Data

The developed corpus now consists of 294 favorable class articles. The exploratory data analysis results for the corpus are as follows:

3.1. Word Visual Representation

Figure 4: Describing the important words in the corpus using word cloud

With word cloud (Figure 4), we sought to understand the keywords that were present in the process. The word cloud was constructed using wordcloud package, under bilinear arrangement of words for the whole developed corpus. It provided a weighted visualization of keywords in the corpus. A few are Tax credit, Input tax, Investigation, Goods services, Companies, Fake invoice, Accused, GST, Firm and Fraud. These were later essential in reviewing and validating topic modelling results and potential fraud domains.

3.2. Time Frequency of Fraud Articles

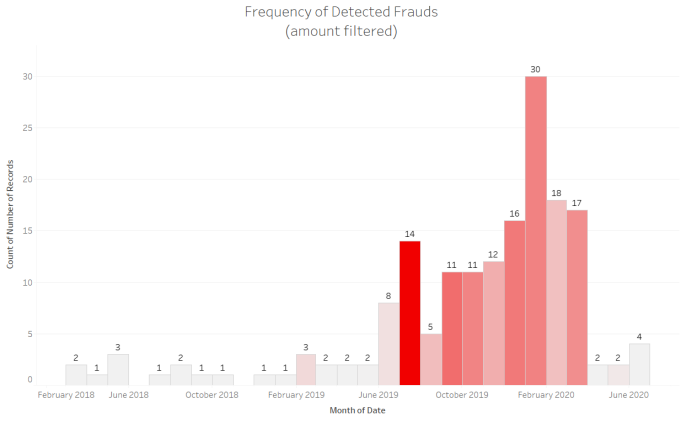

Figure 5: Number of Articles on the time scale

The curve (Figure 5) shows the number of GST fraud related articles published in the news, along a time scale. GST was introduced in July 2017, and first case observed in Feb 2018, and only started full scale proper operation by mid of 2019. This is reflected by the number of articles recorded on the matter. The color scale in the curve represents the average amount of fraud that has been recorded for that period, on a relative scale. The darker the color implies higher average fraud amounts. The highest average fraud (in terms of monetary value) were reported in the period of July 2019. The plot also shows the maximum frequency of cases was recorded for the period Jan 2020 to Feb 2020, and there was subsequently a decline in the number of cases for the subsequent period due to introduction of lockdowns in India, due to COVID-19 pandemic. Upon lockdown relaxation, we may begin seeing a rise in the number of reported GST fraud cases.

3.3. Distribution of Fraud Amounts

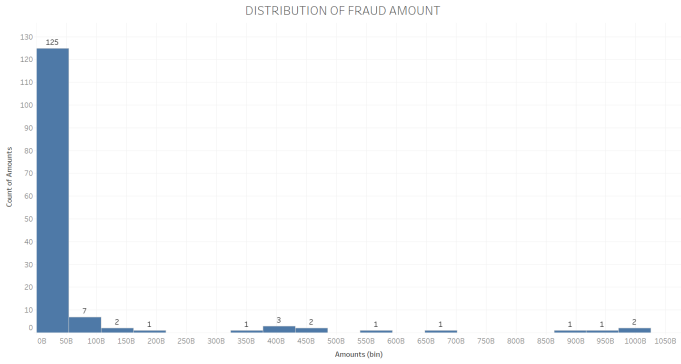

Figure 6: Frequency plot for Fraud Amounts

The plot (Figure 6) shows the frequency distribution of the amounts of fraud that have been reported in the extracted articles. Not all articles talked about the fraud amounts, as some articles talked about arrests due to GST related activities, talked about an ongoing investigation or some were just aggregate representation about the scenario. Based on the extracted amounts, we observe that majority of the fraudulent transactions are relatively on the lower end of the amount involved in the bin Rs 0 to Rs 5,000 cr. Maximum fraud recorded for a single fraud event was recorded at Rs 97,100 thousand crore from the developed corpus.

3.4. Sentiment analysis

The fraction of a given article that had a positive sentiment, a negative sentiment and a neutral sentiment were extracted using a pretrained model SID polarity, available in the NLTK package. The distribution of the negative and positive sentiments (as shown in figure 7) did not provide any visible distinction between the two classes. The blue trace indicates the negative sentiment and the red trace the positive sentiments. But the ratio of these two values gave an interesting insight.

The fraction of article that describes a negative sentiment to the fraction of article that describes a positive sentiment gave an understanding of how sentiment for GST frauds are depicted in

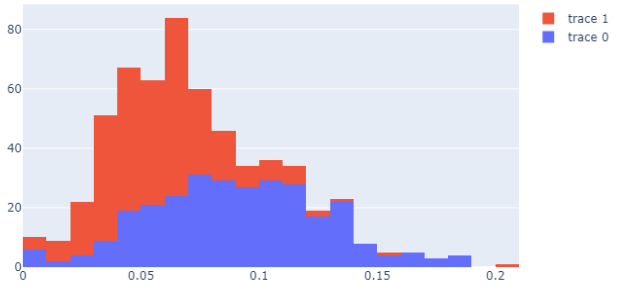

Figure 7: Positive and Negative Sentiment Analysis

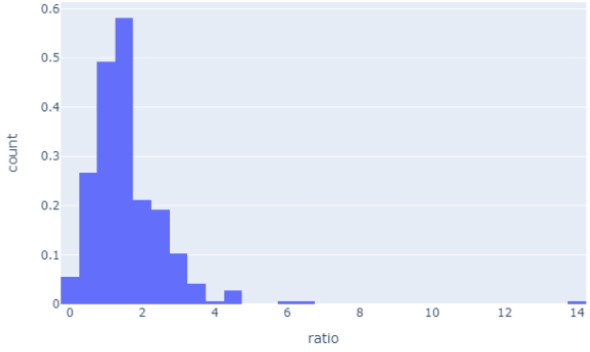

Figure 8: Negative-to-Positive Ratio Distribution

the news. The above figure shows the distribution of this ratio value, which was developed using SID polarity. 70% articles have a predominantly more negative sentiment than a positive sentiment. (Anti-evasion wing detects Rs 241-crore GST fraud, one held: Article by CNBC TV 18 had very negative in sentiment, with ratio value close to 14).

This was a turning point for us, and 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). This further motivated us to work in lines of a Semi Supervised Approach that ensures this quality is considered to classify.

3.5. Complex network curve

Network analysis was 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-occurrence matrix of bigrams. A Fruchterman-Reingold rearrangement is used to depict nodes with higher node-centrality at the geometric center of the graph using a gravity factor of 10. This graph helps assess most common bigrams in data and showed input tax and tax credit were most common bi-gram in the system. The developed network gave us an idea on potential

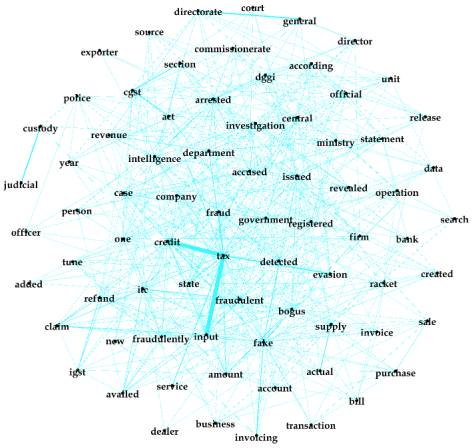

Figure 9: Complex Network Graph developed for keywords

fraudulent areas, namely Judicial custody, Tax evasion, Bogus claims, Fake amount, Fake invoicing, and Fake transaction (to name a few). These detected classes were not exclusive in nature, and usually the frauds that occurred were a combination of one or more of the above categories.

4. Topic Modelling

4.1. Topic Modelling

Large text corpuses are usually abundant with numerous documents/articles that can be categorised under appropriate topics. Such topics serve as the means to label the documents (here news articles) sharing similarities on semantic grounds. This process usually resembles a clustering methodology which is largely unsupervised. Topic modelling is hence, a statistical approach to mine potential topic words from the corpus. We have implemented Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP) and a Hierarchical Clustering (HC) methodology to achieve such "unsupervised topic discovery" purely based on the semantic context of the articles agnostic to the length or the authoring news provider.

4.2. Latent Dirichlet Allocation

LDA is a generative probabilistic model that is used to model abstract topics over collections of discrete data. It is a mutli-level Bayesian Model in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document and can hence be used for clustering and categorization tasks by means of unsupervised labelling. It is interesting to note that though introduced in 2000 by J.K. Pritchard et al in the context of population genetics, it was in 2003 when David M.Blei et al published their work on LDA in machine learning context. This work is considered seminal to the field of unsupervised topic modelling and discovery.

We utilised the Gensim library which is an open-source library for unsupervised topic modeling and natural language processing, using modern statistical machine learning. Firstly, only nouns and adjectives have been filtered from the initial corpus since they are most suitable as topic words. The topic modelling approach has provided us with six different topic labels composed of collections of "topic-words" which can be intuitively interpreted to assign "document-topics" accordingly for the sake of categorization (multi-class classification).

An example is presented 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""),

(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"")]
```

Figure 10: Sample LDA Analysis Results

The first collection of topic words represent "Tax Evasion" category and the second represent "Fake firms" category. Similarly we obtain four other categories namely, "Claims without receipts", "Fake invoices", "Info" and "others". Next we present how we decided on choosing six topics.

4.3. Hierarchical Dirichlet Process

HDP is a non-parametric Bayesian approach to clustering grouped data. Unlike its finite counterpart, latent Dirichlet allocation, the HDP topic model infers the number of topics from the data. HDP is hence is a powerful mixed-membership model for the unsupervised analysis of grouped data. We extract the number of topics using this approach and obtain six definitive clusters. This forms the basis of our initial choice of six topics while modelling topic discovery using LDA.

4.4. Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It helps us decide on the number of clusters using a dendrogram visualisation and outputs the cluster membership of each example in the data. The major challenge in the current study was that the classes are not fixed and continuously evolving. In such scenarios, unsupervised methods are preferable. However the most widely used clustering approach K-means would not prove effective in our case because the number of clusters needs to be fixed at the beginning itself. Thus, we state that hierarchical clustering would bring scalability and dynamicity to the solution.
For the current data on news articles related to GST Fraud, the process followed to implement hierarchical clustering is illustrated below.

1. Clean the data using the data cleansing procedure as explained earlier.
2. Generate RoBERTa embedding features from the cleaned text obtained from step 1.
3. Obtain a dendrogram visualisation to understand the number of clusters in the data.

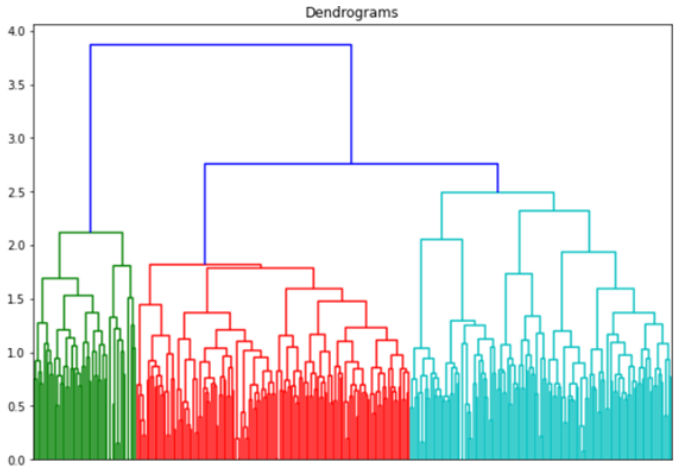

Figure 11: Dendrogram Representation of probable clusters for the corpus

4. Perform hierarchical clustering in python using the above obtained features.

The output obtained from the above process for the given data is as follows:

- Green represents the *'tax evasion'* class as it is the third most prominent type of fraud. This class, upon keyword analysis, predominantly talks about tax, crore and/or evasion related activities.
- Red represents the *'fake invoices'* class as it is the second most prominent type of fraud. This class, upon keyword analysis, predominantly talks about illegal, bogus and/or firms related activities.
- Sky Blue represents the *'fake firms'* class as it is the first most prominent type of fraud. This class, upon keyword analysis, predominantly talks about firms and/or bogus related activities.

4.5. Classifier Development

Based on the prepared dataset, we seek to develop a model approach that predicts the class of fraud a given article falls in based on its article text. There are 5 target classes, and the number of training data points was 294. The class distribution for the target class are as follows:

The text was cleaned using the same mechanism as we developed for the process before. The text content embedding was experimented using TF.iDF, and RoBERTa. TF.iDF yielded a feature space of 80,000+ features, and RoBERTa generated 1024 features. Were all these features used to develop a model with so little training dataset would make the model highly biased, and greatly overfit on the training dataset, hence leading to poor generalizability of the developed model. Hence, the feature space was reduced using latent semantic analysis (LSA) to a lower dimensional space, by using Principal Component Analysis (PCA).

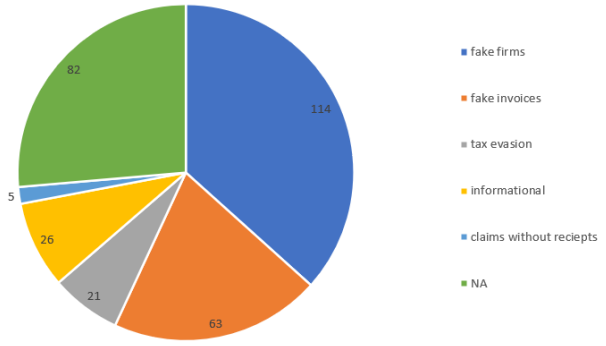

Figure 12: Distribution of target classes in the corpus

Model	Train Accuracy	Validation Accuracy	F1 score
Random Forest	1.000	0.705	0.687
bagged LR	0.684	0.656	0.651
Naive-Bayes	0.656	0.639	0.629
XGB Classifier	1.000	0.623	0.620
LightGBM	1.000	0.623	0.619
Ada boost	0.760	0.607	0.616
Logistic Regression	0.688	0.590	0.582
Bagging	0.980	0.557	0.562
KNN	0.664	0.557	0.535
Decision Tree	1.000	0.443	0.454

Figure 13: Model Results for different Classifiers

The dimensions of LSA were selected so as to yield maximum validation accuracy. The values were observed in an incremental form, and the most optimal performance was achieved using 25 PCA dimensions. The model results achieved using these 25 features are as follows:

Best model performance (validation accuracy) was achieved using Random Forest Classifier. Upon hyperparameter tuning, the best validation accuracy achieved was 73.77%, using 25 LSA features. Top-n-accuracy is an important evaluation metric especially in NLP problems. We calculated the top-3 validation accuracy in our case. The best validation top-3 accuracy was obtained as 90% by using an *rbf-kernel-SVM* model.

4.6. Indicators and Bench-marking

It is important to assess the incoming articles quantitatively. In this regard, we develop custom indicators, like here, the "nltk compound" and the "amount" column are very useful indicators of the severity of the GST fraud case that has been reported. We also propose custom indicators based on growth rate in the 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.

While the initial binary classifier (GST fraud vs Non-Fraud) had a precision of 0.96 and recall of 0.926, we were surprised to find that it was so accurate to even detect 2 articles which were incorrectly labelled (intentionally labelled incorrectly to replicate human performance), thereby surpassing the human benchmark.

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" is one example of such novel topic discovery.

5. Further scope

5.1. Dashboard

Once the model is under deployment, continuous feedback is collected about what solutions have been used on which particular issue. This will help in developing a fully scalable dashboard that can propose solutions as soon as the fraud is classified. The proposed is a dashboard with aggregated analysis of all available features and information in a visually structured manner, that discusses in detail about the amount distributions, reported fraud time distributions, and the negative-to-positive sentiment associated with such sub sections of articles.

The proposed dashboard was developed in Tableau. It allows multi-layer filtering mechanism, i.e. in single or in combination with one another, where filtering can be done easily through visual depictions. The dashboard allows selective features over a time period (on discrete and continuous scales), over a fraud label category, over the class of fraud, or over the sentiment of articles. Features are auto integrated with each other, hence allowing multi-layer and personal customization and according visualizations.A sample representation of the dashboard is depicted in Figure 14.

5.2. The Complete Model

Once the model is under deployment, continuous feedback is collected about what solutions have been used on which particular issue. This will help in developing a fully scalable dashboard that can propose solutions as soon as the fraud is classified. The User Interface (UI) will have a text box wherein one can paste the article content or alternatively choose to provide the URL to the news article. This article is then cleaned as per the procedures mentioned earlier. If it's classified as a GST Fraud article, we proceed with computing its sentiment score and also try to compare its co-occurrence against the standard corpus. We further proceed with the multi-class topic categorization and label it accordingly. We display the results of "Top-3" topic categories for the user to choose from. Going further, we can improve our classifier by taking feedback from the users on the classification and thereby provide a single label once we achieve sufficient data and confidence. Also, for topics labelled as "others", we probe further using topic modelling to discover any newly evolving topics and even those results can be displayed on application of filters. If there is a significant improvement in a new topic's article count, we can include that category for our future re-calibration of the Multi-class classifier. Hence, the current system can then move towards being a completely automated unsupervised topic discovery and sentiment extraction mechanism on textual feedback systems.

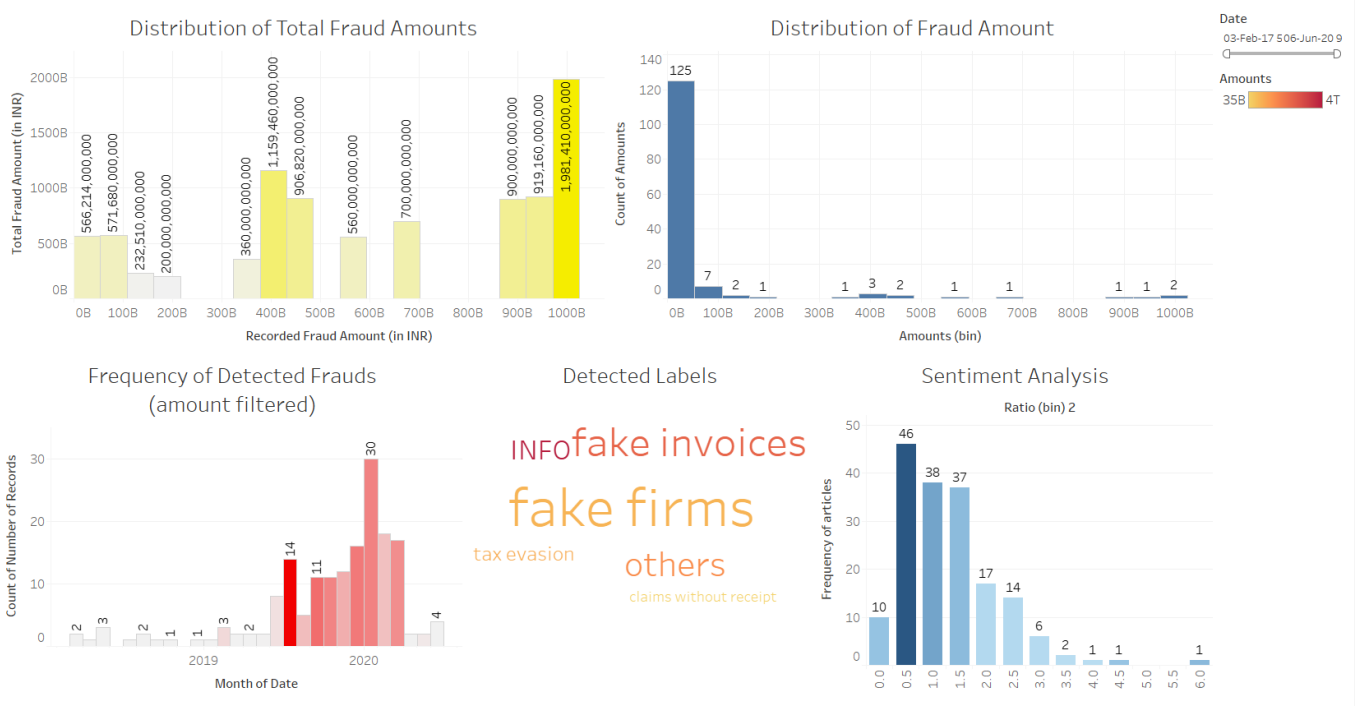

Figure 14: Dashboard representation

5.3. Reviews use-case

The current work discussed GST's application in detail. We further provide an overview of how the same mechanism can be proven 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 redressal on platforms like Amazon.com.

6. Conclusion

The present study aims at analysing the textual data coming from feedback systems to extract the class membership information from the unlabelled training corpus. We have utilised the data of news articles on GST fraud for a practical implementation of our proposition. The news articles are extracted using a novel and efficient mechanism, and compared different embedding mechanisms like TF.iDF, Word2Vec, and RoBERTa. In the context of GST fraud, the proposed method finds the reasons for the GST fraud from the unlabelled news articles and bucket them into those labels. We have generated a multitude of visualisations and dashboards to perform EDA. Then topic modelling using Latent Dirichlet Allocation(LDA) is used to extract labels. The number of topics which is an input for this activity can be obtained using Hierarchical Dirichlet Process(HDP). We state that the clusters can also be obtained using hierarchical clustering but the cluster labels are often hazy and unclear in most of the instances. We also modelled a RoBERTa feature based classifier which gave a Top-3 validation accuracy of 90% on the GST fraud article dataset. Although the current study implements the proposed algorithm on GST fraud dataset, it can be scaled and applied to cases with topics evolving with time.

7. References

[1] *Article on Claiming GST*.

[2] Council, G. *About GST*. http://gstcouncil.gov.in/about-gst.

[3] Financial Express, *GST Fraud Reference Article 1*. https://www.financialexpress.com/economy/could-aadhaar-linking-stop-gst-frauds/1727527/.

[4] Times of India, *GST Fraud Reference Article II*. https://timesofindia.indiatimes.com/business/india-business/government-blocks-rs-40000-crore-gst-claims-on-returns-mismatch/articleshow/73682209.cms.

[5] Business Std. , *Article on DRI DGGI Joint Operation*. https://www.business-standard.com/article/pti-stories/dri-dggi-carry-out-biggest-ever-joint-operation-against-gst-violators-at-336-locations-119091201164$_1$.*html*.

[6] Clear Tax, *Article on Fake GST Invoice Verification*. https://cleartax.in/s/fake-gst-invoice-verify-gstin-gst-tax-rates.

Appendix A. Labelling of GST Fraud Articles

The major challenge to build a supervised classifier was lack of labelled data. Hence, we have labelled the news articles related to GST fraud into five categories namely fake firms, fake invoices, tax evasion, others. These were obtained from topic modelling exercise on the entire data.

- The articles where the fraud was committed by establishing shell companies are labelled as *'fake firms'*.
- The articles where the fraud was committed using fake invoices generated from existing going concern business entities are labelled as *'fake invoices'*.
- The articles where tax was evaded by understating invoice or voluntarily defaulting taxes are labelled as *'tax evasion'*.
- All the other articles which do not fall into the above framework and need to be further explored for new topic discovery are labelled as *'others'*.
- Some articles which were not related to GST but not GST fraud have crept into our dataset even though the number is very few. They have been labelled as *'info'* if they summarise about GST Fraud. On the other hand if they are not related to GST, they are labelled as *'NA'*.

Article	Label
New Delhi: The Ministry of Finance announced yesterday that the Anti Evasion wing of CGST Delhi South Commissionerate has detected a case of Input Tax Credit (ITC) fraud through fake invoices issued by bogus firms. Investigations led to discovering the accused had also been generating bogus e-way bills to back the fake invoices. Over 35 entities are involved in the bogus transactions, involving fake invoicing to the tune of Rs. 214.74 crores and tax evasion of Rs. 38.05 crores. One person has been arrested on 19 February and remanded to judicial custody of 14 days. Further investigations in the matter are in progress.	Fake Firms
Moradabad, Nov 21 (KNN) A Moradabad based exporter has been arrested for Central Goods and Services Tax (GST) fraud of as much as Rs 28 crore and for fraudulently availing Input Tax Credit (ITC) of around Rs 8 crore. The Customs and GST authority have been cracking down on fake GST invoices and Input Tax Credit (ITC) cases. The officials in Meerut arrested a Moradabad-based exporter of clutch plates and disc pads for fake invoices of approximately Rs 28 crore and fraudulently availing ITC of around Rs 8 crore. According to a media report, the exporter has been arrested and Rs 7.80 crore recovered from him. This is considered as one of the biggest individual cases of fake invoices.	Fake Invoices
Ludhiana: The directorate general of GST intelligence (DGGI), Ludhiana zonal unit, has registered a case of GST evasion against prominent passenger bus services companies — Libra Bus Services Limited, Ludhiana, and Maharaja Travels, Amritsar — for non-payment of GST. Sources said the case has been registered as both the firms were neither filing their GST returns regularly nor were discharging their due liabilities on taxable services (transportation of passengers on AC buses). On preliminary scrutiny of various documents and evidences gathered from the business premises of both the firms, the tax liability worth more than Rs 1 crore.	Tax Evasion
Noida: According to officials, the accused firms, RoundPay Techno Media and RoundPay Voice Tech, both registered at the same address in Lucknow, were involved in e-payment for utility services.They were involved in circular trading and had shown transactions worth Rs 72.7 crore between them, on which they claimed an input tax credit of Rs 13.09 crore at 18 per cent GST, a senior official said. "The STF and the sales tax department were jointly entrusted with the responsibility to probe the case, a first of its kind operation in Uttar Pradesh. This trading module was studied for 45 days. It was found that their activities were being carried out in Lucknow, Pratapgarh, Noida and Khiri also," Deputy Superintendent of Police, STF Noida, Rajkumar Mishra said."The firms were identified, their locations physically verified. Over 100 officers were involved in the operation," he said. "When the investigating teams asked the companies for an invoice of the trade done between them, they failed to produce any. The system which was used by these companies for generating the invoice/bills also had no record of this trade," the officer said."The director of the parent company has admitted in writing to the probe agencies about the Rs 13 crore fraud," the DSP said, adding 6 TB (terrabyte) data has been recovered from the firms' servers in Noida and a detailed investigation of it was underway."The fraud amount is likely to go up, we are awaiting the details," he said, adding Mishra said no FIR has been lodged as of now neither any arrest made.	Others