

The School of Mathematics



THE UNIVERSITY  
*of* EDINBURGH

# Financial Ombudsman Service decision analysis: using public data to understand customer complaints

by

Aravindh Sankar Ravisankar, S2596860

Dissertation Presented for the Degree of  
MSc in Statistics with Data Science

August 2024

Supervised by

Amy Wilson, Jordan Richards, Sergio Gomez Anaya (University of  
Edinburgh) and  
Ekaterina Zaytseva, Antonio Feregrino (Simply Business)

## Executive Summary

In earlier times, consumers frequently faced difficulties in resolving financial disputes with the companies[Gilad et al.(2008)[1]]. The creation of the Financial Ombudsman Service (**FOS**) by the government addressed this issue by providing an impartial review of complaints. Thus, a comprehensive analysis of the FOS review process was crucial for the financial institutions to thrive. Following initial data fetching from the decision documents, preprocessing is performed on the text data which thereon led to some exploratory data analysis for generating key insights. Two traditional machine learning models (**Logistic Regression and Random Forest**) and a large language model (**BERT**) were developed for a binary classification task to predict FOS decisions. While Random Forest performed well across three embedding techniques, the deficit between precision and recall was large(**0.17**). In contrast, BERT demonstrated balanced performance across all metrics with the optimal model having a precision of 0.686 and a recall of 0.615. Therefore, BERT is highly recommended for this classification task with adequate training time and an efficient set of hyperparameters.

## Acknowledgments

I would like to extend my heartfelt thanks to everyone who supported me during my second dissertation project. First and foremost, I would like to express my deep gratitude to my project supervisors: Amy Wilson, Jordan Richards, and Sergio Gomez Anaya from the School of Mathematics at the University of Edinburgh, as well as Ekaterina Zaytseva, Antonio Feregrino from Simply Business for their invaluable guidance, insightful suggestions, and unwavering support. I would like to extend a special thanks to Simply Business for providing the problem statement and to the School of Mathematics for facilitating this project. I am also grateful to my friends and family for their constant support and encouragement.

Thank you all for your contributions and support.

Yours sincerely,  
Aravindh Sankar Ravisankar

## University of Edinburgh – Own Work Declaration

This sheet must be filled in, signed and dated - your work will not be marked unless this is done.

Name: Aravindh Sankar Ravisankar

Matriculation Number: S2596860

Title of work: Financial Ombudsman Service decision analysis: using public data to understand customer complaints.

I confirm that all this work is my own except where indicated, and that I have:

- Clearly referenced/listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Not sought or used the help of any external professional academic agencies for the work
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

I understand that any false claim for this work will be penalised in accordance with the University regulations (<https://teaching.maths.ed.ac.uk/main/msc-students/msc-programmes/statistics/data-science/assessment/academic-misconduct>).

Signature: Aravindh Sankar Ravisankar

Date: 9th August 2024

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background and Motivation . . . . .	1
1.2	Problem Statement . . . . .	1
1.3	Data Overview . . . . .	1
<b>2</b>	<b>Literature Review</b>	<b>2</b>
<b>3</b>	<b>Exploratory Data Analysis (EDA)</b>	<b>3</b>
3.1	Organization Analysis . . . . .	3
3.2	Timeline Analysis . . . . .	4
3.2.1	Yearly Analysis . . . . .	4
3.2.2	Monthly Analysis . . . . .	6
3.3	Text Data Analysis . . . . .	7
3.3.1	Word Length Analysis . . . . .	7
3.3.2	Top 10 common words . . . . .	7
3.3.3	Wordcloud Analysis . . . . .	8
3.3.4	Top 15 words influencing the decision making . . . . .	9
3.4	Category Evaluation . . . . .	10
3.5	Insights on Ombudsman Findings . . . . .	11
<b>4</b>	<b>Feature Extraction</b>	<b>13</b>
4.1	Word2Vec Method . . . . .	13
4.2	TF-IDF Method . . . . .	13
4.3	GloVe Method . . . . .	14
<b>5</b>	<b>Modelling</b>	<b>14</b>
5.1	Traditional machine learning models . . . . .	14
5.1.1	Logistic Regression . . . . .	14
5.1.2	Random Forrest . . . . .	14
5.2	Large Language Model . . . . .	15
5.2.1	BERT . . . . .	15
<b>6</b>	<b>Results and Inferences</b>	<b>16</b>
6.1	Model Diagnostics . . . . .	16
6.2	Model Evaluation . . . . .	16
<b>7</b>	<b>Conclusions</b>	<b>18</b>
	<b>Appendix</b>	<b>21</b>
<b>A</b>	<b>Data Fetching</b>	<b>21</b>
<b>B</b>	<b>Data Preprocessing</b>	<b>21</b>
<b>C</b>	<b>Topic Modelling - LDA Implementation</b>	<b>22</b>
<b>D</b>	<b>Word Count Check</b>	<b>22</b>

## List of Tables

1	Model Evaluation Table - Conventional ML approach . . . . .	16
2	Model Evaluation Table - Large Language Model . . . . .	17

## List of Figures

1	Frequency Plot of complaints for the Top 10 companies . . . . .	3
2	Distribution of Upheld ratio aggregated by companies . . . . .	3
3	Number of complaints aggregated over the years . . . . .	4
4	Upheld complaints aggregated over the years . . . . .	5
5	Not upheld complaints aggregated over the years . . . . .	5
6	Upheld and Not upheld ratio distribution over the years present . . . . .	5
7	Frequency of complaints aggregated over the months . . . . .	6
8	Upheld and not upheld ratio aggregated over the months present . . . . .	6
9	Distribution of word length on the incoming complaint . . . . .	7
10	Barplot indicating the Top 10 words in the corpus . . . . .	8
11	Wordcloud generated on the complaint data corpus . . . . .	8
12	Top 15 words contributing to Upheld decisions . . . . .	9
13	Top 15 words contributing to Not upheld decisions . . . . .	10
14	Bar plot on the Frequency of complaints for every category . . . . .	10
15	Upheld and Not upheld ratio distribution across categories . . . . .	11
16	Wordcloud generated on Ombudsman Findings . . . . .	11
17	Word Count Image . . . . .	22

# 1 Introduction

## 1.1 Background and Motivation

Consumers faced significant challenges in resolving disputes with financial institutions before the establishment of the Financial Ombudsman Service (**FOS**). **Gilad et al.(2008)**[1] describes the conventional litigation process as complex and time-consuming which often discourages consumers from pursuing their complaints. The high costs and formal nature of court proceedings contributed to a substantial imbalance of power between consumers and financial organizations leading many people to abandon their complaints and accept unsatisfactory resolutions. **Benohr et al. (2012)**[2] indicates that the lack of availability of effective redress mechanisms left consumers frustrated and distraught.

The introduction of FOS significantly alleviated these challenges by providing a more accessible, efficient, and consumer-friendly platform for dispute resolution. The FOS offers an informal process where complaints can be resolved quickly and at no cost to consumers. **Hodges et al.(2012)** [3] highlights that the ombudsman service emphasizes mediation and non-binding recommendations which are typically accepted by both parties. **Merrick et al.(2007)** [4] suggests that the FOS model is a viable alternative to traditional court systems for resolving financial disputes as it reduces the time and financial burden on consumers. Furthermore, **Beqiraj et al.(2018)** [5] indicates that the FOS also provides useful feedback to financial institutions to help improve industry standards and customer satisfaction.

Therefore, it is essential for financial institutions to understand the nature of financial disputes and the methods employed by the ombudsman service which involves their approach towards a complaint, subsequent steps involved in the review process, and the rationale behind their decision-making.

## 1.2 Problem Statement

The Financial Ombudsman Service (**FOS**) plays a crucial role in resolving disputes between consumers and financial institutions. Established by the government, the FOS addresses complaints that could not be resolved directly between the parties involved. This project aims to harness public FOS decision files to gain actionable insights into dispute patterns and outcomes. These actionable insights would help them guide product development, identify and mitigate risks, and improve complaint-handling processes which tend to refrain from minimal escalation of complaints to the ombudsman service ensuring customer satisfaction. These key aspects were attained through building a pipeline that encompasses various intermediate phases like cleaning semi-structured text documents, identifying patterns in the preprocessed data using statistical techniques, visualizing data distributions in key feature columns, and creating word embeddings using various natural language processing techniques. Binary classification models were built using traditional machine learning algorithms and a large language model. The results were compared and contrasted in proposing an optimal model for future deployment.

## 1.3 Data Overview

The dataset encompasses 160,797 records (**check appendix on the data fetching**) collected from 2013 to 2024 (**until 15th of July**) with 12 feature columns covering over 2,653 different financial institutions across insurance and PPI sectors. The eight features fetched from the metadata offer some basic insights on the complaint whereas the last four features elucidate significant information about the various sections of the complaint document. The response variable 'decision' has two values to offer, 'Upheld' implying that the ombudsman has agreed with the complainant (**consumer**) whereas 'Not upheld' pertains to taking a stand against the complainant. Other notable variables of interest include 'decision\_id' (**unique identifier**), 'date' (**the date of complaint submission**), 'company' (**the financial institution in question**), 'Complaint\_info' (**overview of the complaint**) and 'Complaint\_explanation' (**detailed explanation of the complaint**). After preprocessing (**see appendix**), we were left with 160348 records.

## 2 Literature Review

In his research paper, **Kowsari**[6] provides a comprehensive overview of various text classification algorithms and their applications. The survey breaks down text classification systems into four primary phases: feature extraction, dimensionality reduction, classification algorithms, and model evaluation. Towards the end, he compares and contrasts the classification algorithms, presenting their evaluation metric scores along with their limitations. The paper concludes that the choice of model depends on the specific requirements of the problem statement.

**Santiago**[7] provides a detailed comparison between the BERT model and traditional machine learning approaches for text classification. The study emphasizes the ability of BERT to handle various natural language processing (**NLP**) tasks which delivers impeccable results compared to the traditional machine learning approaches. Experiments conducted across different datasets such as IMDB reviews, disaster-related tweets, and Portuguese news articles have consistently shown BERT outperforming the traditional models in accuracy making it a default choice for the NLP tasks due to its robust performance.

Delving deeper into the financial domain, **Forster**[8] has presented an advanced method for handling unstructured customer complaint letters in the insurance sector. The author proposes a cognitive computing approach that encapsulates classical natural language processing techniques, conventional machine learning algorithms, and sentiment analysis to develop a robust multi-label text classification model. The approach utilizing a MaxEnt machine learning algorithm along with TF-IDF and sentiment analysis achieved an impressive F1-score of 0.9 highlighting its significance in accurately classifying and managing customer complaints.

**Ashtiani**[9] offers a comprehensive review of the state-of-the-art techniques for detecting financial fraud in corporate statements. He systematically selected and analyzed 47 different study groups to identify the most effective machine learning and data mining methods used for fraud detection. While traditional machine learning approaches are considered for benchmark evaluation, he advocates for increased use of ensemble and deep learning methods with a slight motivation towards developing bio-inspired algorithms.



### 3 Exploratory Data Analysis (EDA)

#### 3.1 Organization Analysis

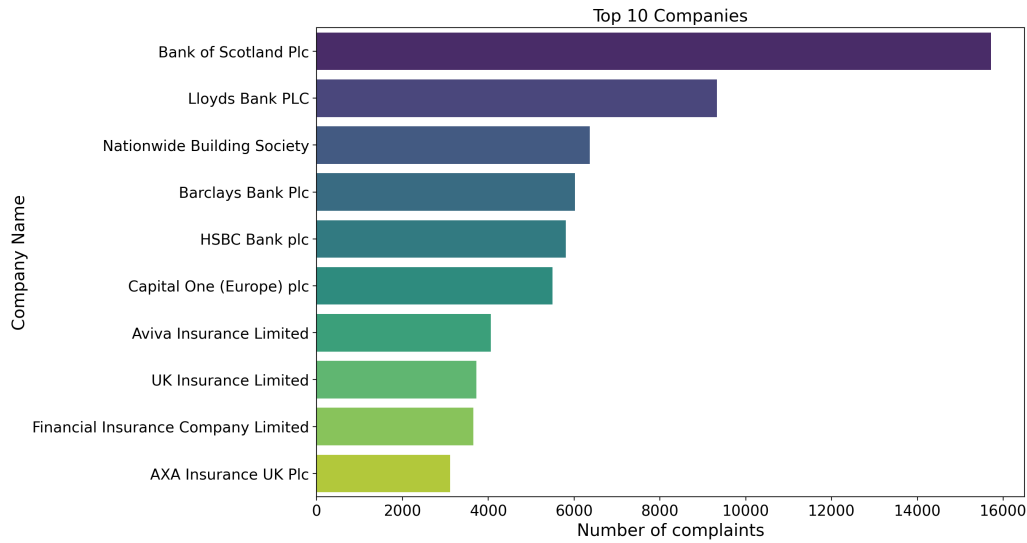


Figure 1: Frequency Plot of complaints for the Top 10 companies

Figure 1 displays the **frequency of complaints for the top 10 financial institutions** which are ranked in descending order based on the number of incoming complaints. Bank of Scotland leads the chart with approximately 16,000 complaints, followed by Lloyds Bank and Nationwide Building Society. These numbers suggest that these three institutions either face significant customer service issues in resolving financial claims or have larger customer bases, resulting in a higher number of complaints. This plot helps us understand the landscape of financial institutions suggesting that a more proactive approach could be undertaken by these companies to enhance their service quality and customer relations.

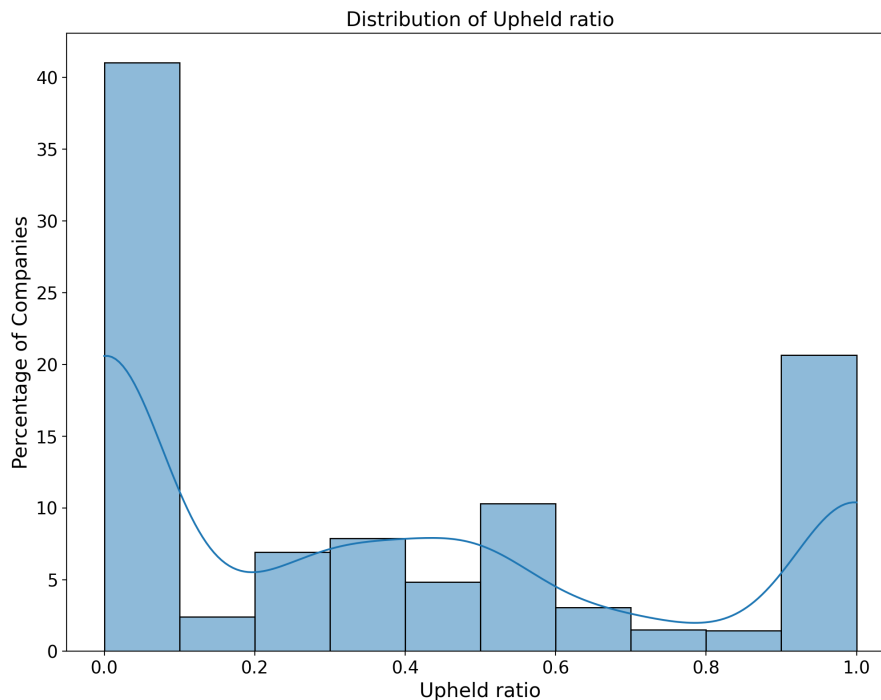


Figure 2: Distribution of Upheld ratio aggregated by companies

The histogram plot in Figure 2 illustrates the **distribution of the upheld claim ratio across companies**. The X-axis represents the continuous ratio measures ranging from 0 to 1, while the Y-axis indicates the percentage of companies. The plot shows that about 20% of companies have an upheld ratio close to 1 meaning all incoming complaints for these companies result in an 'upheld' decision. On the other hand, approximately 40% of the companies have an upheld ratio close to zero (**not upheld ratio is 1**) indicating that they have not had a single complaint adjudged as 'upheld' which is a significant achievement. This trend suggests a balanced distribution around the mid-ratio measure of 0.5 with a fairly equal spread of companies on either side. Companies with lower ratio measures may have less cause for concern while those with higher ratios especially those with a ratio of 1.0 are advised to take proactive measures in their complaint handling.

## 3.2 Timeline Analysis

### 3.2.1 Yearly Analysis

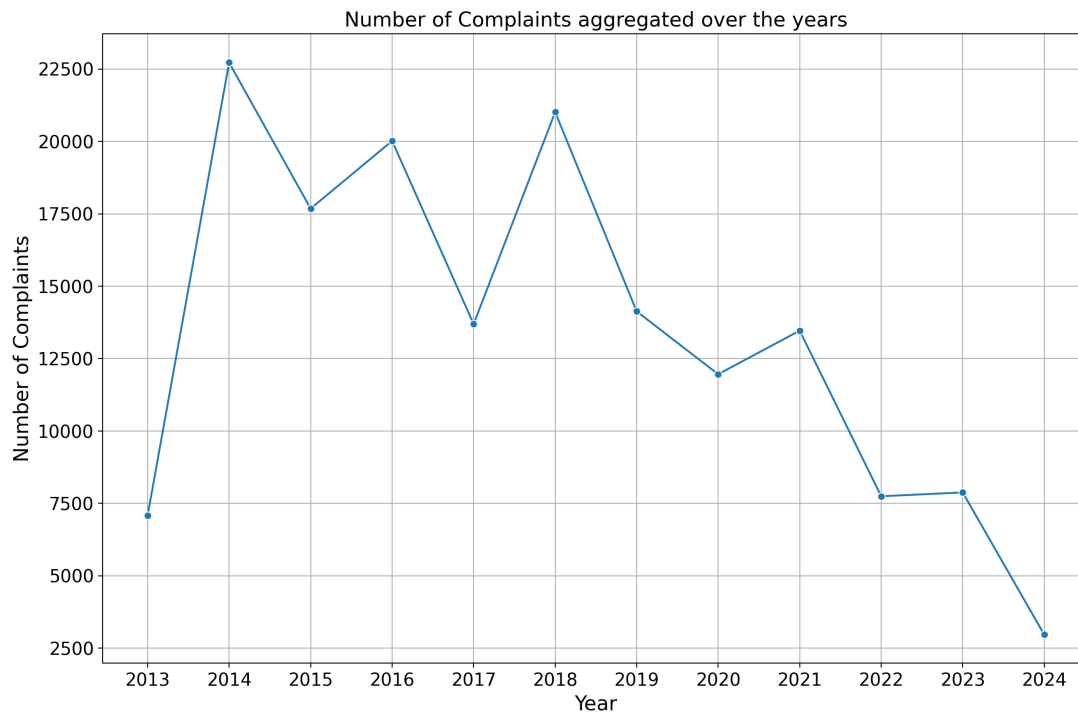


Figure 3: Number of complaints aggregated over the years

The line plot depicted above in Figure 3 illustrates the trend on the **number of complaints** received by the **ombudsman service** from 2013 until mid-July 2024. The data shows an inconsistent upward trend with irregular fluctuations during the first half of the timeline. After 2018, the number of complaints gradually decreased reaching a new low of about 2,500 complaints in 2024. This decline in complaints towards the later part of the timeline suggests that companies have implemented strong, proactive measures to address consumer grievances more effectively over time. However, further improvements could help ensure that the inflow of complaints continues to decrease in the years to come.

The bar plot depicted in Figure 4 shows the **yearly distribution of upheld claims** from 2013 to 2024. There was an upward trend in upheld complaints starting in 2013, peaking in 2018 with around 8,500 complaints. After 2018, there is a steep decline in the number of complaints averaging around 3,000 complaints each year until the end of the timeline hitting a new low in 2024. This trend suggests significant improvements in complaint handling or a reduction in consumer grievances in recent years as very few incoming complaints are being upheld by the ombudsman service.

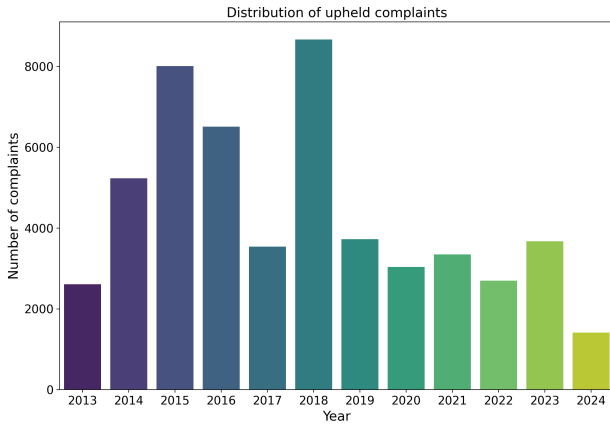


Figure 4: Upheld complaints aggregated over the years

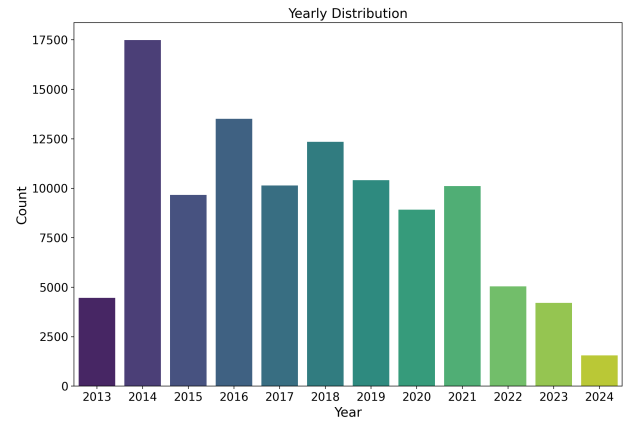


Figure 5: Not upheld complaints aggregated over the years

Figure 5 shows the **distribution of the number of not upheld complaints** from 2013 to 2024. The maximum number was in 2014 with about 17,500 complaints whereas 2024 had the lowest number with about 1,500 complaints. In the middle period, the number hovered around 10,000 to 12,500 complaints per year until a significant dip after 2021. This trend also reflects improvements in complaint-handling processes, better customer service, or changes in consumer behavior that led to fewer disputes being resolved against consumers(as indicated earlier). However, it would be wise to observe the upheld and not upheld ratios over the timeline to provide conclusive insights.

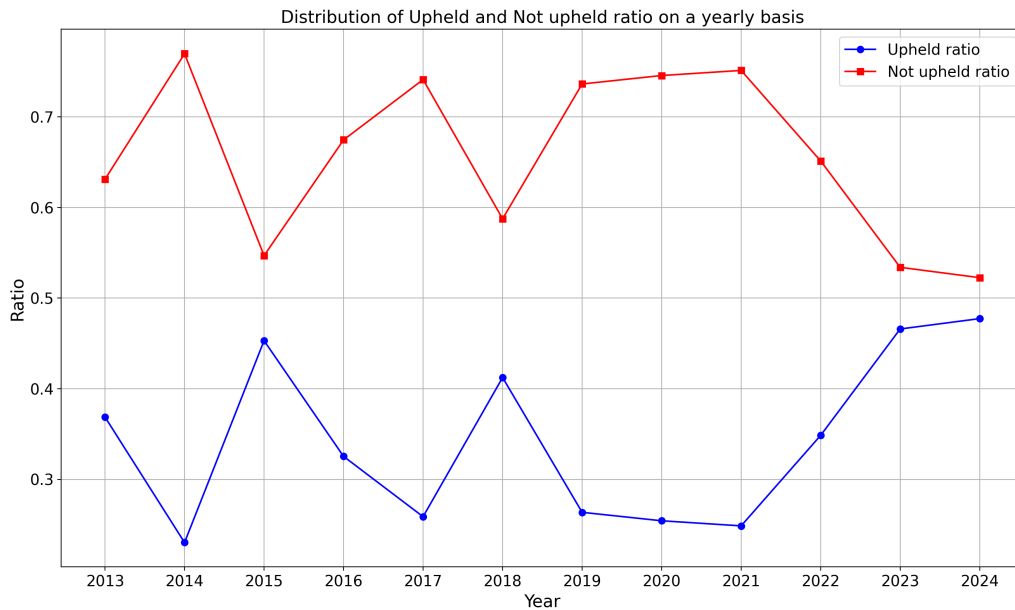


Figure 6: Upheld and Not upheld ratio distribution over the years present

To gain more conclusive evidence on the trends of upheld and not upheld complaints, we observe Figure 6 which provides insights into the **proportion of complaints resolved** in favor of consumers every year. Since there are only two possible decisions, the two line plots would be mirror images of one another (**if the upheld ratio is higher, the not upheld ratio will be lower for that particular year**). The upheld claim ratio exhibits considerable fluctuations over the timeline with consistent ups and downs. Notably, there is a significant rise in 2024, with the ratio reaching nearly 0.50 indicating that almost half of the incoming complaints were upheld (**ruled in favor of the complainant**). In contrast, years like 2014 and 2017 showed a lower trend in the upheld ratio.

The overall trend showcases periods of high and low upheld ratios reflecting varying effectiveness in complaint resolution processes and the potential impact of regulatory changes or internal improvements within companies. However, the sudden spike in the upheld ratio in recent years raises questions about the redressal mechanisms offered by financial institutions which calls for some serious proactive measures.

### 3.2.2 Monthly Analysis

The line plot depicted in Figure 7 illustrates the **number of complaints** received by the ombudsman service **aggregated over every month** of the year. March stands out with incoming complaints exceeding 16000 indicating a period of heightened customer dissatisfaction whereas a substantial dip is observed in the next month with the number going down to 11,000. The trend is erratic with no regular pattern observed over any phase and on an overall basis, there is a consistent inflow of complaints across every calendar month.

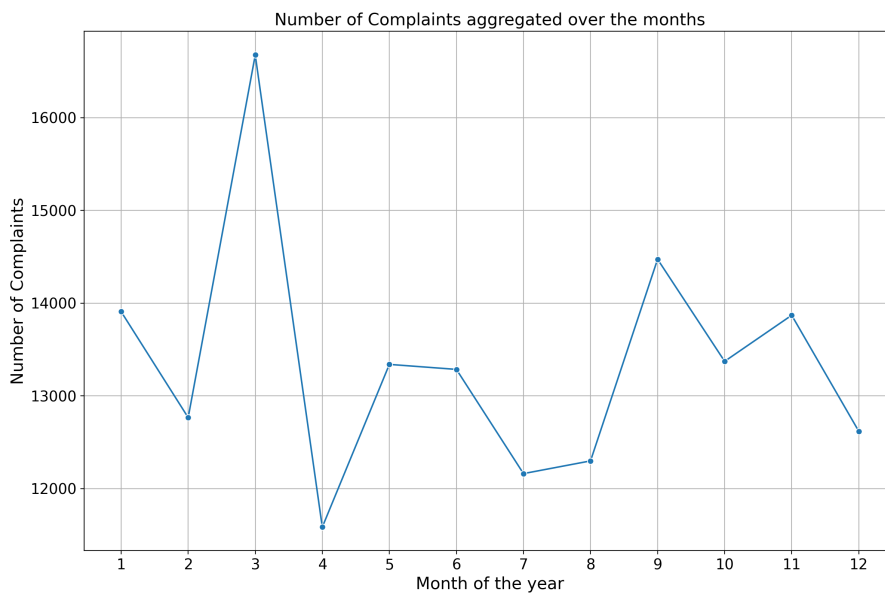


Figure 7: Frequency of complaints aggregated over the months

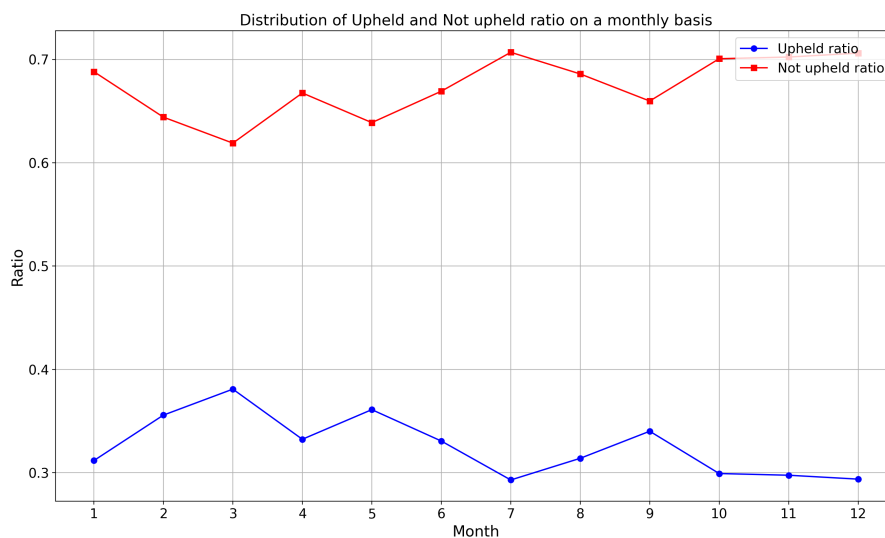


Figure 8: Upheld and not upheld ratio aggregated over the months present

Figure 8 presents two line graphs representing the **upheld and not upheld ratios for every month of the year**. The highest upheld claim ratio occurs in March peaking at approximately 0.38

indicating that a significant proportion of complaints during this month are resolved in favor of the complainant. In contrast, December shows the lowest upheld ratio of around 0.29 indicating fewer complaints are upheld during this month. The ratio fluctuates throughout the year with an average upheld ratio of approximately 0.30 maintained across the majority of months. One conclusive takeaway is the need for **careful monitoring** of complaints every month as there is **no clear seasonal pattern** in the data.

### 3.3 Text Data Analysis

#### 3.3.1 Word Length Analysis

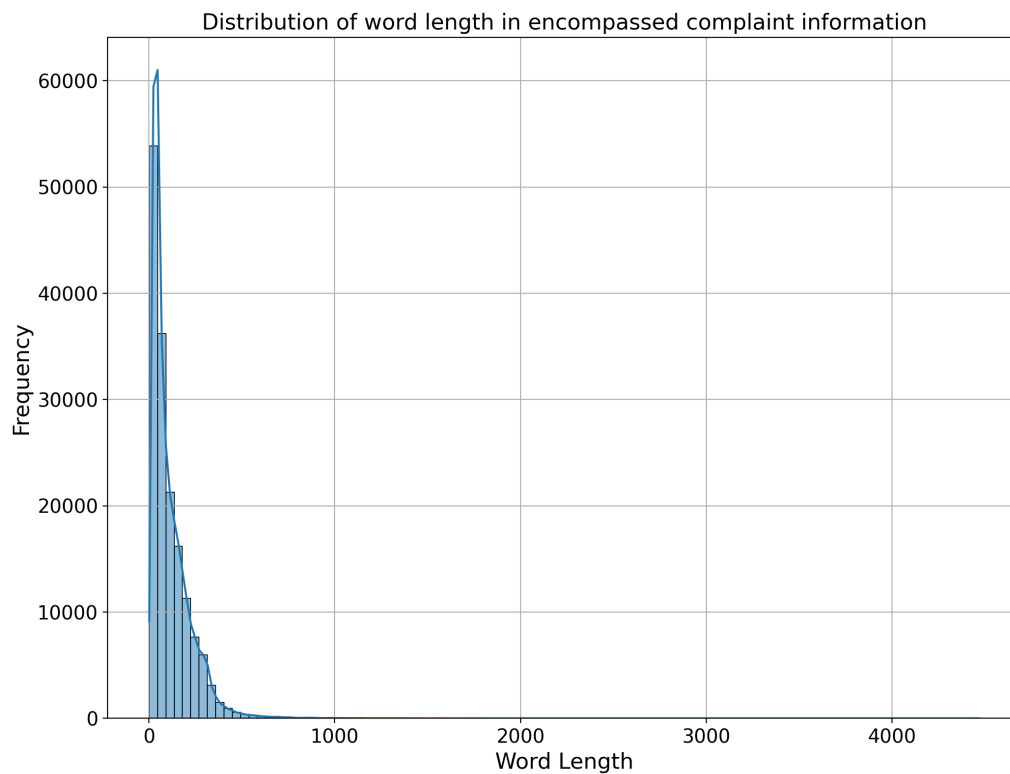
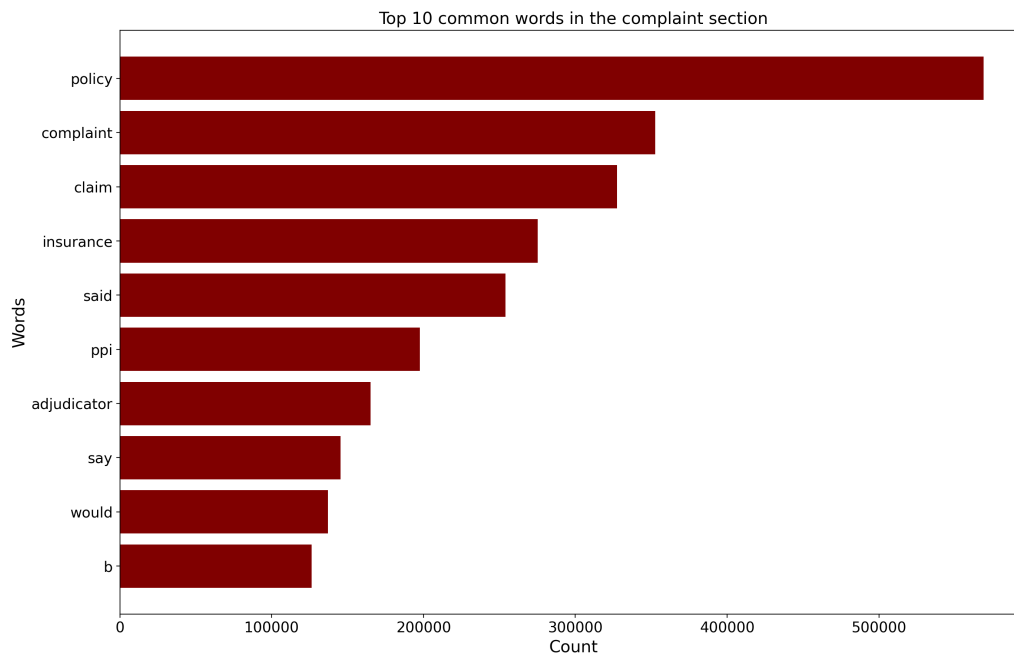


Figure 9: Distribution of word length on the incoming complaint

The histogram plot in Figure 9 shows the distribution of word length across every incoming complaint present (**considering only the complaint information**). The peak of the distribution shows that the majority of complaints contain fewer than 100 words with the frequency reaching over 60,000 complaints at the lower end of the spectrum. As the word count increases, the frequency of complaints diminishes sharply. The tail end of the distribution shows sparse instances of complaints with word counts up to and beyond 4,000 words which are extremely rare. This skewed distribution highlights that most complaints are brief and less detailed.

#### 3.3.2 Top 10 common words

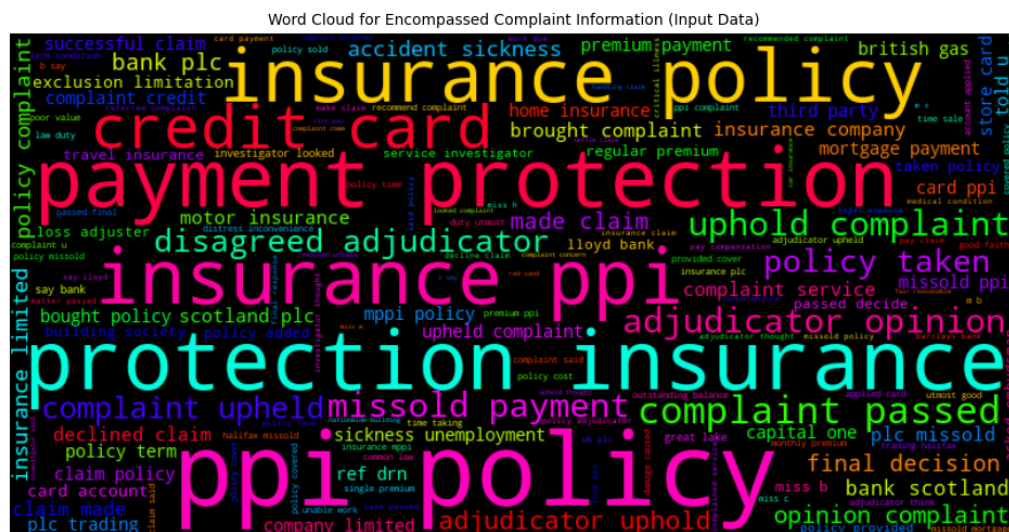
Figure 10 depicts a bar plot showing the top 10 common words that are used more frequently in describing the complaints alongside their frequency count.



The bar plot shows the top 10 words with 'policy,' 'complaint,' and 'claim' occupying the top three positions.

### 3.3.3 Wordcloud Analysis

Figure 11 shows the word cloud generated from the corpus of complaints which contains all the necessary information about the incoming complaints and provides valuable insights into the most frequently occurring terms in the complaint narratives. Prominent words such as "policy," "insurance," "ppi," "protection," "insurance," and "credit" are significantly highlighted indicating their central role in the complaints. Terms like "adjudicator," "decision," and "upheld" also appear frequently reflecting the common themes and processes involved in resolving these disputes.



### 3.3.4 Top 15 words influencing the decision making

The bar plot as observed in Figure 12 illustrates the top 15 words contributing to upheld decisions measured by their Term-Frequency Inverse Document Frequency (**TF-IDF**) scores. The term "ppi" (**Payment Protection Insurance**) stands out with the highest score followed closely by "policy" and "complaint," indicating these terms are highly influential in cases where complaints are upheld. Other significant words include "card," "insurance," "claim," and "missold," reflecting common elements involved in disputes that result in upheld decisions. This analysis underscores the importance of these specific terms in understanding the nature and reasons behind upheld decisions. It also provides valuable insights for improving complaint handling by alerting relevant personnel to closely monitor the presence of these words, as there is a higher likelihood that such complaints will transition to an upheld verdict by the ombudsman.

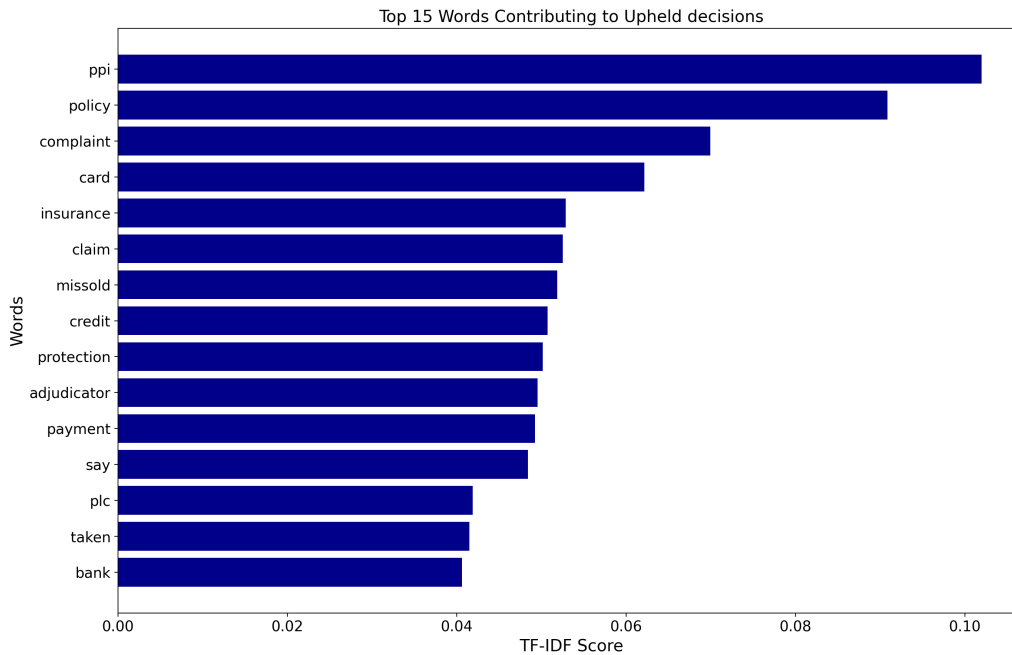


Figure 12: Top 15 words contributing to Upheld decisions

The bar plot as shown in Figure 13 illustrates the top 15 words contributing to not upheld decisions, measured by their Term-Frequency Inverse Document Frequency (**TF-IDF**) scores. The term "point" stands out with the highest score followed closely by "position" and "company," indicating that these terms are highly influential in cases where complaints are not upheld. Other significant words include "adjudicator," "cancel," and "installed," reflecting some common factors involved in disputes that could result in not upheld decisions. This analysis provides valuable insights into the decision-making process for financial institutions suggesting that any future complaint containing these words could be flagged for closer review as their presence indicates a higher likelihood that the complaint may not be upheld at the ombudsman level.

However, the significant point to be noted from both analyses is that the presence of these terms should not be used as the sole criteria for classification but rather as one of several factors to be wary of in future assessments of incoming complaints.

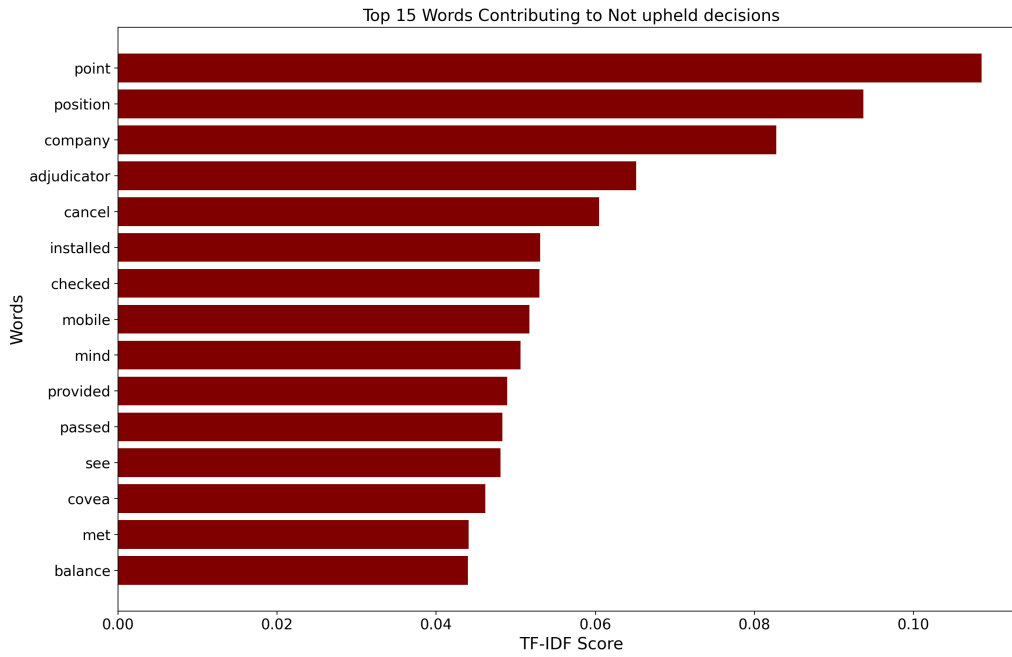


Figure 13: Top 15 words contributing to Not upheld decisions

### 3.4 Category Evaluation

Topic Modelling is performed through which categories of the complaints are identified (**see Appendix**).

Figure 14 is a bar plot that shows the distribution of complaints across various finalized subcategories. Mortgage PPI tops the chart with over 37,000 complaints closely followed by Vehicle Insurance and Credit Card PPI with 34,000 and 33,000 complaints respectively. Loan Repayment PPI and Legal/Business Insurance have the fewest complaints with around 5,000 each. The bar chart highlights clear consumer dissatisfaction with the top three categories (**as mentioned above**) indicating a need for proactive measures as the magnitude of incoming complaints is significantly high.

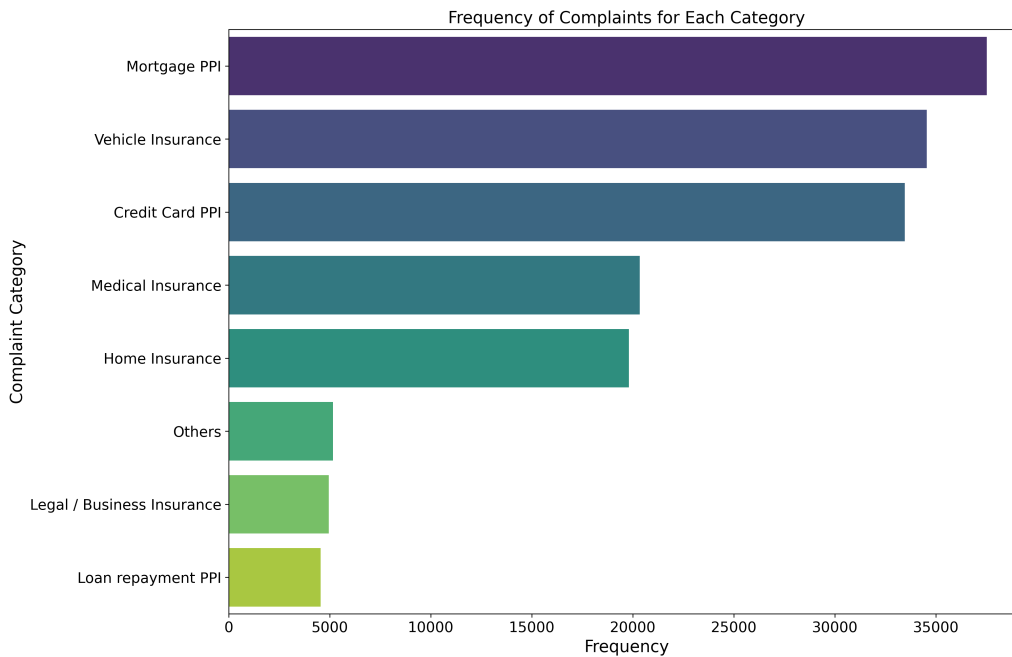


Figure 14: Bar plot on the Frequency of complaints for every category



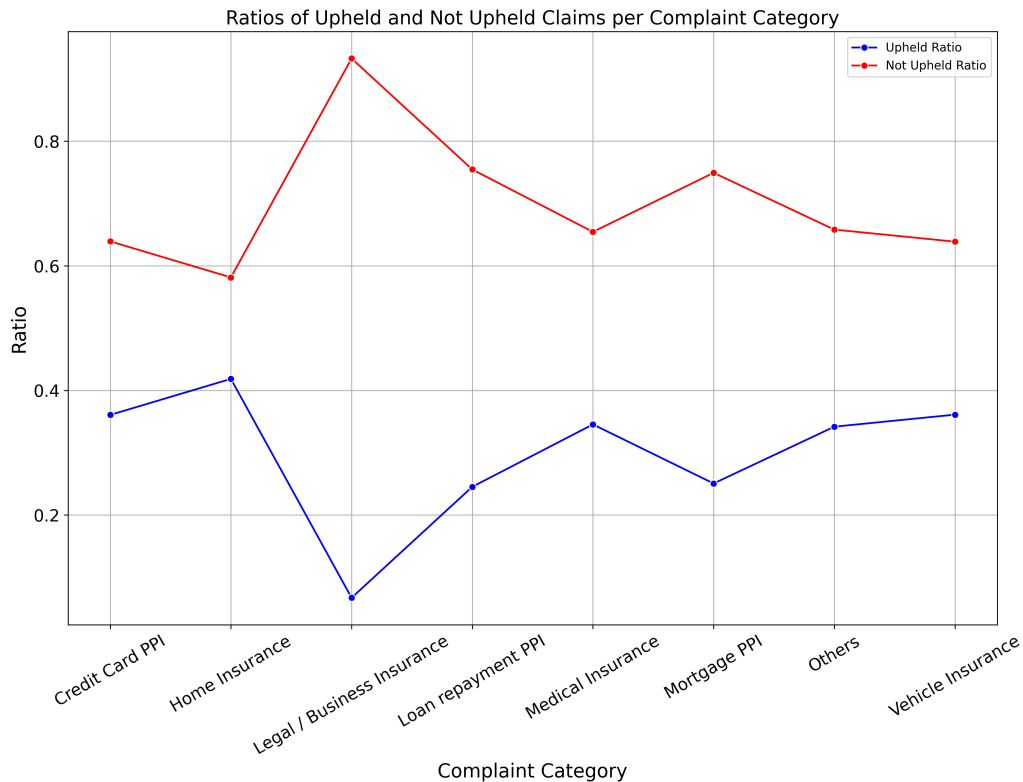


Figure 15: Upheld and Not upheld ratio distribution across categories

Figure 15 includes two line plots depicting the upheld and not upheld ratios aggregated across evaluated categories. Visual inspection reveals that the upheld ratio peaks in the Legal/Business Insurance category, with over 90% of complaints resulting in upheld decisions. For other domains, the upheld ratio ranges between 0.6 and 0.8 indicating that 60 to 80 percent of complaints are upheld. This trend highlights that a significant proportion of complaints escalated to the ombudsman are ruled in favor of the complainants (**regardless of the category**) with more focus required on the Legal/Business Insurance category.

### 3.5 Insights on Ombudsman Findings



Figure 16: Wordcloud generated on Ombudsman Findings

The word cloud generated (Figure 16) from the ombudsman analysis reveals key terms frequently mentioned in their findings which could lead to some plausible explanations. Words such as "complaint," "considered," and "argument" indicate that the ombudsman carefully reviews the arguments presented by both parties involved (**the company and the consumer**) before beginning their analysis. The frequent appearance of words like "circumstance," "decide," and "available" suggests that the ombudsman requires evidence submitted by the parties to understand the rationale behind the financial institution's verdict. Additionally, the presence of words like "fair" and "reasonable" reflects the ombudsman's commitment to delivering impartial and balanced decisions without favoring either side.

## 4 Feature Extraction

### 4.1 Word2Vec Method

**Word2Vec** introduced by Mikolov[10] at Google is a neural network model designed to create word embeddings by transforming words into continuous vector representations. It encompasses two primary model architectures: Continuous Bag of Words (**CBOW**) and Skip-gram. The CBOW model predicts a target word given its surrounding context, whereas the Skip-gram model predicts the context words given a target word.

The main objective of the Skip-gram model[11] for a given sequence of words  $(w_1, w_2, \dots, w_T)$  is to maximize the average log probability which is given as follows:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \quad (4.1)$$

where the size of the context is denoted by  $c$ .

The basic formulation of the Skip-gram model defines  $p(w_{t+j} | w_t)$  using the softmax function, which is:

$$p(w_O | w_I) = \frac{\exp(\mathbf{v}'_{w_O} \mathbf{v}_{w_I})}{\sum_{w=1}^W \exp(\mathbf{v}'_w \mathbf{v}_{w_I})} \quad (4.2)$$

where  $\mathbf{v}_w$  and  $\mathbf{v}'_w$  are the input and output vector representations of the word  $w$  and  $W$  is the number of words in the vocabulary.

Its enhanced approach to effectively capture semantic relationships between words is achieved by first building a vocabulary from the training corpus and then learning the vector representations. Additionally, the algorithm positions words in a high-dimensional vector space where similar words are placed together making it an efficient word embedding technique for various natural language tasks.

### 4.2 TF-IDF Method

Another important statistical word embedding technique that is widely used across various natural language tasks is **Term Frequency - Inverse Document Frequency (TF-IDF)**. This approach [12] helps to evaluate the significance of words in the corpus which encompasses two different metrics - Term Frequency (**TF**) which measures the frequency of a word 'w' in a document 'd' whereas the inverse document frequency (**IDF**) evaluates the rarity of the word across all the document present in the corpus. Therefore, common words have a low TFIDF score whereas rare words have a higher TFIDF score.

Given a document collection  $D$ , a word  $w$ , and an individual document  $d \in D$ , we calculate:

$$w_d = f_{w,d} \cdot \log \left( \frac{|D|}{f_{w,D}} \right) \quad (4.3)$$

where  $f_{w,d}$  equals the number of times  $w$  appears in  $d$ ,  $|D|$  is the size of the corpus, and  $f_{w,D}$  equals the number of documents in which  $w$  appears in  $D$ .

As emphasized earlier, this method effectively captures more important terms while reducing the weight of more common ones that are less informative.

### 4.3 GloVe Method

**Global Vectors (GloVe) for Word Representation** is an unsupervised learning algorithm developed by **Pennington**[13] at Stanford University. The approach generates word embeddings by aggregating global word-to-word co-occurrence statistics from the corpus. Unlike other word embedding techniques, GloVe constructs word vectors based on the statistical information of word co-occurrences, ensuring that the vector spaces capture both the global and local context of words.

Mathematically, GloVe leverages the probability that a word  $w_i$  will appear in the context of another word  $w_j$  represented by the ratio of co-occurrence probabilities. The primary equation concerning GloVe is as follows :

$$J = \sum_{i,j=1}^V f(X_{ij}) \left( w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2 \quad (4.4)$$

where  $X_{ij}$  represents the co-occurrence count between word  $i$  and word  $j$ ,  $w_i$  and  $\tilde{w}_j$  are the word vectors, and  $b_i$  and  $\tilde{b}_j$  are the bias terms. As a result of the enforcement of both global matrix factorization and local context window techniques, GloVe appears to be a standout approach across various natural language tasks.

In recent times, **Shapal** [14] demonstrated that the GloVe technique outperformed traditional methods such as TF-IDF and k-means clustering in various tasks while also being computationally efficient.

## 5 Modelling

### 5.1 Traditional machine learning models

#### 5.1.1 Logistic Regression

Logistic regression is a **supervised machine learning** algorithm commonly used for binary classification problems[15]. It predicts the probability of a binary outcome based on one or more predictor variables. This discriminative classifier uses the sigmoid function (**activation function**) to map the predicted values to probabilities that operate on a range from 0 to 1. The equation concerning the sigmoid function is given as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (5.1)$$

where  $z$  is the logit of the probability.

The log-odds or logit of the probability as a linear combination of the linear features could be formulated as follows:

$$\text{logit}(S) = b_0 + b_1 M_1 + b_2 M_2 + b_3 M_3 + \dots + b_k M_k \quad (5.2)$$

where  $S$  is the probability of the presence of features of interest,  $M_1, M_2, \dots, M_k$  are the predictor values, and  $b_0, b_1, \dots, b_k$  are the coefficients of the model.

Logistic Regression estimates the coefficients  $b_0, b_1, \dots, b_k$  by maximizing the likelihood of the observed data using optimization methods like gradient descent. The model must adhere to some assumptions such as having no linear relationship between the dependent and independent variables, the dependent variable must be dichotomous, independent variables to be linearly related within the group and groups must be mutually exclusive.

#### 5.1.2 Random Forrest

Random Forest is an ensemble learning method [16] that builds multiple decision trees using a process called **bootstrap aggregating (bagging)** where several subsets of the original training data are

created by sampling with replacement. Each subset thereon is used to train a separate decision tree which reduces overfitting since every tree is involved in training on a slightly altered data sample. In addition to it, Random Forest introduces **feature randomness** by selecting a random subset of features for splitting at each node of the decision trees present ensuring that the trees offer minimal overlap and improved generalization ability.

The algorithm calculates the entropy for each possible split to determine the best feature and threshold to split the data, with entropy defined as:

$$\text{Entropy}(D) = - \sum_{k=1}^c P_k \log_2 P_k \quad (5.3)$$

where  $D$  is the dataset,  $c$  is the number of classes, and  $P_k$  is the proportion of class  $k$  in the dataset. The information gain for each split is then calculated as:

$$\text{Information Gain}(D, t) = \text{Entropy}(D) - \frac{|D_t|}{|D|} \text{Entropy}(D_t) \quad (5.4)$$

where  $D_t$  is the subset of  $D$  after applying the split based on feature  $t$ .

The algorithm iteratively identifies the splits that maximize information gain to grow each decision tree. The final prediction is made by majority voting on the outputs of the individual decision trees present for classification tasks resulting in a more robust and accurate model.

## 5.2 Large Language Model

### 5.2.1 BERT

BERT (**Bidirectional Encoder Representations from Transformers**) represents a significant advancement in language representation models by pre-training deep bidirectional representations from unlabeled text data [17]. The BERT architecture encompasses several stacked encoders that perform exceptionally well on language modeling and other text analysis tasks because of its ability to capture the context in both directions (**left to right and vice versa**) which was absent in the earlier prototypes.

The working paradigm of BERT involves two main phases: **pre-training and fine-tuning**. Pre-training allows BERT to understand the language of the input text data whereas fine-tuning enables the model to learn and perform specific tasks efficiently. Pre-training in turn consists of two steps: the Masked Language Model (**MLM**) and Next Sentence Prediction (**NSP**). The MLM randomly masks some input tokens and the model learns to predict these masked tokens based on their context, thus enabling bidirectional understanding. The NSP helps the model understand the relationship between pairs of sentences, determining if the second sentence is a follow-up to the first. This ensemble approach allows BERT to be fine-tuned for various downstream tasks with minimal task-specific modifications, achieving state-of-the-art performance across different natural language processing tasks present.

## 6 Results and Inferences

### 6.1 Model Diagnostics

Model diagnostics are a set of techniques and metrics used to evaluate and understand the performance of the model. These diagnostics help interpret the model's predictions and identify areas for improvement. By analyzing diagnostic metrics, we can determine how well the model is performing and what adjustments are necessary to enhance its predictive capabilities.

The common metrics that are evaluated for conventional classification models are as follows:

- **Confusion Matrix:** A table (**2 x 2 dimensions**) that emphasizes the performance of the classification model built. It is based on four metrics which are True Positives(**TP**), True Negatives(**TN**), False Positives(**FP**), and False Negatives(**FN**).

- **Accuracy:** Help us understand the overall correctness of the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.1)$$

- **Precision:** Indicates the proportion of positive predictions that are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6.2)$$

- **Recall:** Indicates the proportion of positive predictions out of actual positives present.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6.3)$$

- **F1 Score:** Harmonic mean of Precision and Recall.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6.4)$$

- **AUC:** Area Under the Curve measure which signifies the distinguishing ability of the classification model.

### 6.2 Model Evaluation

In accordance with the normal conventional rule, the dataset has been split with 80% of data (**128278 records**) for the training dataset and the remaining 20% (**32070 records**) on the test dataset. Logistic Regression and Random Forest are the two machine learning models deployed for the binary classification task with three different feature extraction techniques tried out on every phase of training. The resulting scores which are evaluated on the test dataset for the earlier defined metrics are tabulated in the below Table 1.

Feature Extraction	Model Name/Metric	Accuracy	Precision	Recall	F1 Score	AUC
TFIDF	Logistic Regression	0.795	0.725	0.607	0.661	0.747
	Random Forest	<b>0.829</b>	<b>0.805</b>	<b>0.632</b>	<b>0.708</b>	<b>0.779</b>
Word2Vec	Logistic Regression	0.728	0.631	0.421	0.505	0.650
	Random Forest	0.766	0.748	0.436	0.551	0.682
GloVe	Logistic Regression	0.700	0.580	0.313	0.407	0.600
	Random Forest	0.740	0.770	0.300	0.432	0.628

Table 1: Model Evaluation Table - Conventional ML approach

In the context of model comparison, Table 1 clearly demonstrates that the Random Forest model outperforms Logistic Regression across all evaluation metrics on every feature extraction technique used. A significant difference in scores is observed across all metrics except for recall where the scores are nearly identical.

On the other hand, in the context of feature extraction techniques, TFIDF consistently delivers superior performance for both models across all evaluated metrics. There is a significant difference observed with recall and F1 scores over the other techniques. Therefore on considering every aspect involved, **Random Forest coupled with the TFIDF feature extraction** technique is the preferred model since it strikes a good balance among the scores evaluated.

In addition to two traditional machine learning models deployed, the research extended to implement a BERT large language model for the same binary classification model. The model is trained over three sets of epochs (**iterations**) and the corresponding scores on the test dataset for the evaluating metrics have been tabulated below in Table 2.

Model Name	Number of Epochs	Accuracy	Precision	Recall	F1 Score	AUC
BERT	<b>3</b>	<b>0.787</b>	<b>0.714</b>	0.590	0.646	0.737
	<b>5</b>	0.781	0.686	<b>0.615</b>	0.649	<b>0.739</b>
	<b>8</b>	0.776	0.702	0.557	0.621	0.720

Table 2: Model Evaluation Table - Large Language Model

The table shows the performance of the BERT model across three different sets of epochs (**3, 5, and 8**) in terms of model diagnostics present. When trained over **three epochs**, the BERT model achieves a **strong balance** between **precision and recall**, effectively identifying both upheld and not upheld cases. However, when the training is increased to **five epochs**, the model **slightly improves in recall**, becoming better at detecting upheld cases at the cost of a minor dip in precision leading to more false positives while other metrics remain largely unchanged. Finally, when the training is extended to **eight epochs**, the model's performance begins to decline with **recall** taking a **significant dip** accounting for the fact that the model is starting to **overfit**. Therefore, **training over five epochs** is considered **optimal** as it offers a reasonable trade-off between precision and recall (**with a narrower gap**) ensuring that the model consistently detects true positives while minimizing false positives.

On comparing the above two tables, it is observed that Random Forest with TFIDF outperforms the optimized BERT model (**trained over five epochs**) across all metrics. However, a notable difference of 0.17 between precision and recall is evident in the best-performing Random Forest model. In contrast, the scores of the BERT model remain within a closer range with the best-performing model (**trained over five epochs**) showing only a slight 0.07 difference between precision and recall. This highlights that the BERT model strikes an ideal balance by effectively identifying true positives while maintaining a low rate of false positives. This balance suggests that BERT is particularly adept at accurately distinguishing between upheld and not upheld decisions, making it a highly reliable model for the task at hand. In addition to that, since BERT is considered as one of the state-of-the-art models for natural language processing tasks, **it is highly recommended to implement the BERT model with further exploration on hyperparameter tuning.**

## 7 Conclusions

The study demonstrated the effectiveness of financial ombudsman service in handling consumer complaints. The research work started with scraping the data from the financial complaints extending it to a tabular format followed by preprocessing the text data which led to a comprehensive exploratory data analysis emulating key insights about the data. Three different feature extraction techniques are deployed where a compare and contrast evaluation is performed on the two machine learning models built for the binary classification task present. Logistic Regression proved to be a good benchmark model over which the ensemble Random Forest model was built which provided an impeccable performance across the model diagnostics. It was further extended to deploy the state-of-the-art BERT large language model which was trained and fine-tuned to implement the specific task over different training sets. Although the model had almost equal results with Random Forest exceeding on some metrics, the BERT model showed a good balance across all metrics evaluated. Therefore, the BERT model is the recommended architecture to be implemented for the binary classification task at hand. Future work could involve implementing other large language models like DistilBERT and RoBERTa, performing hyperparameter tuning, and exploring boosting techniques.



## Generative AI Acknowledgment

- I acknowledge the use of ChatGPT-3.5 to get the template code for various EDA and modeling techniques.
- I acknowledge the use of ChatGPT-3.5 to get a concise version of my written content at times.
- I acknowledge the use of ChatGPT-3.5 to get the latex code for formulae and figures used in the report.

## References

- [1] S. Gilad, “Accountability or expectations management? the role of the ombudsman in financial regulation,” *Law & Policy*, vol. 30, no. 2, pp. 227–253, 2008.
- [2] C. Hodges, I. Benöhr, and N. Creutzfeldt-Banda, “Consumer-to-business dispute resolution: the power of cadr,” in *era Forum*, vol. 13, pp. 199–225, Springer, 2012.
- [3] C. Hodges, “Consumer adr in europe,” *Zeitschrift für Konfliktmanagement*, vol. 15, no. 6, pp. 195–197, 2012.
- [4] W. Merricks, “The financial ombudsman service: not just an alternative to court,” *Journal of Financial Regulation and Compliance*, vol. 15, no. 2, pp. 135–142, 2007.
- [5] J. Beqiraj, S. Garahan, and K. Shuttleworth, “Ombudsman schemes and effective access to justice: A study of international practices and trends,” 2018.
- [6] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” *Information*, vol. 10, no. 4, p. 150, 2019.
- [7] S. González-Carvajal and E. C. Garrido-Merchán, “Comparing bert against traditional machine learning text classification,” *arXiv preprint arXiv:2005.13012*, 2020.
- [8] J. Forster and B. Entrup, “A cognitive computing approach for classification of complaints in the insurance industry,” in *IOP Conference Series: Materials Science and Engineering*, vol. 261, p. 012016, IOP Publishing, 2017.
- [9] M. N. Ashtiani and B. Raahemi, “Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review,” *Ieee Access*, vol. 10, pp. 72504–72525, 2021.
- [10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, vol. 26, 2013.
- [12] J. Ramos *et al.*, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 29–48, Citeseer, 2003.
- [13] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.
- [14] S. M. Mohammed, K. Jacksi, and S. R. Zeebaree, “Glove word embedding and dbscan algorithms for semantic document clustering,” in *2020 international conference on advanced science and engineering (ICOASE)*, pp. 1–6, IEEE, 2020.
- [15] A. Prabhat and V. Khullar, “Sentiment classification on big data using naïve bayes and logistic regression,” in *2017 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–5, IEEE, 2017.
- [16] Y. Sun, Y. Li, Q. Zeng, and Y. Bian, “Application research of text classification based on random forest algorithm,” in *2020 3rd international conference on advanced electronic materials, computers and software engineering (aemcse)*, pp. 370–374, IEEE, 2020.
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

# Appendix

## A Data Fetching

The dataset involved is extracted from the Financial Ombudsman Service (**FOS**) portal, which offers a comprehensive collection of records across various financial sectors since 2013. Each record follows a generic template comprising four sections as follows:

- **First Section:** Introduces the issue brought forth by the complainant, clearly presenting their perspective and the basis of their dissatisfaction.
- **Second Section:** Provides a detailed understanding of the situation, including all relevant historical information.
- **Third Section:** A thorough examination of the arguments and evidence submitted by both the complainant and the financial institution, followed by a detailed analysis and the provisional decision of the ombudsman.
- **Fourth Section:** The final section outlines the decision taken (**upheld or not upheld**) and the remedial actions to be taken, such as the type and amount of compensation and the deadline for resolution.

Given the nature of Simply Business operations, we restrict our research to analyzing complaints related to **Insurance** and **Payment Protection Insurance (PPI)**. The key insights and predictive modeling results attained from this analysis would be extremely specific and valuable to them.

To manage the time and space complexities, decision files are downloaded on a yearly basis. Each file is then processed using a user-defined function that parses the decision file, captures the content from each section, and stores it under appropriate headers in a tabular format. Data scraping was challenging because the generic template's section names changed over time, even though the content remained the same. This issue was addressed by manually reviewing and correcting the discrepancies. The resulting dataframe was then merged with the metadata (**initially downloaded**) to produce a new dataframe containing all the information for that particular year's insurance complaints.

This cyclic extraction process is repeated annually, with the additional step of appending each year's dataframe to a cumulative dataframe. The complete dataframe at the end encompasses information from all the years, which will be used for exploratory data analysis and predictive modeling.

## B Data Preprocessing

Each feature column had some level of significance in providing insights about the data, except for the 'extras' column, which was removed as it only contained NULL values. During data extraction, the code used for downloading decision files for a year's duration inadvertently included duplicate records on the threshold date, as it was processed twice in subsequent cycles. These duplicate records were identified and removed, retaining only the first occurrence. Additionally, records with NULL values in both the 'Complaint\_Info' and 'Complaint\_Explanation' columns were discarded, as they provided no information on the complaint. However, records where either of these two sections had content were retained for further analysis and modeling. Upon manual review, it was found that some records had content in the 'Complaint\_Explanation' section that included information on the ombudsman's provisional decision which could influence the decision-making of the model, thereby the information is scraped from the 'Complaint\_Explanation' column and provided to the newly created 'provisional\_decision' column.

We then cleaned and standardized the text data using Python's Natural Language Toolkit (**NLTK**). This process involved converting text to lowercase, removing punctuation and numbers, tokenizing the

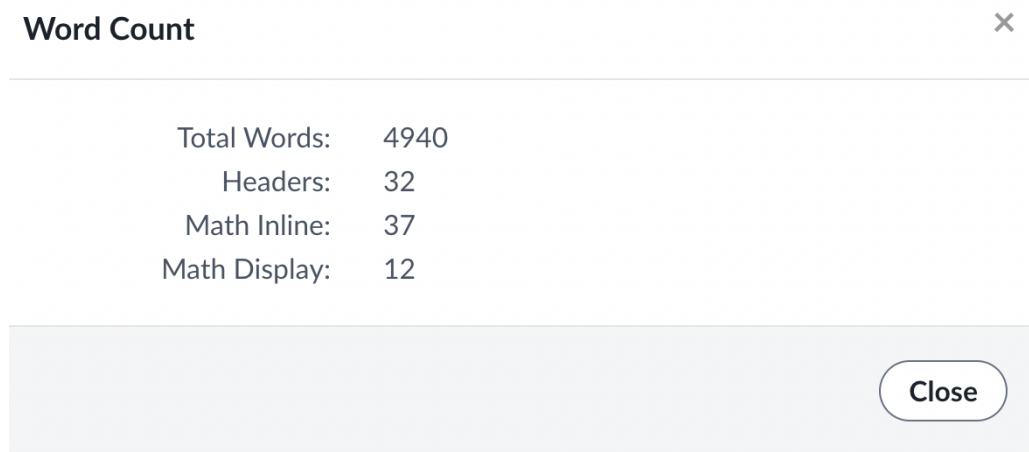
text into individual words, and removing stopwords using NLTK's stopwords list along with additionally specified stopwords like 'mr' and 'mrs'. Each word was then lemmatized to its base form, and special characters and extra whitespace were eliminated. This preprocessing was applied to each text column in the dataset, resulting in a new dataframe (**new\_df**) with cleaned text ready for further analysis and modeling. Additionally, we merged the two text columns (**'Complaint\_Info'** and **'Complaint\_Explanation'**) into a single feature column (**'complaint\_data'**) containing preprocessed text information about the complaint. This consolidated column will be later used as input for our predictive modeling.

## C Topic Modelling - LDA Implementation

The dataset encompasses complaints raised against various financial companies, specifically in the domains of insurance and payment protection insurance. To explore potential subcategories within these domains, topic modeling using Latent Dirichlet Allocation (**LDA**) is performed. The text data is preprocessed by converting it to lowercase and splitting it into words. A dictionary of words is created, with extreme words filtered out, and the corpus is prepared for LDA modeling. Multiple LDA models are trained with different numbers of topics (ranging from 7 to 11), based on visual inspection and data familiarity. Coherence scores are computed for each model to determine the best fit. The model with the highest coherence score is selected, and the top 10 words from each topic cluster are printed. By manually reviewing these words and applying domain knowledge, coherent topics are identified, and a corresponding dictionary is created. The dominant topics for each record in the data frame are then extracted and mapped to the identified topic names.

## D Word Count Check

Hereby attaching the word count image for reference. Total words were **4940** which includes just the main content (**excluding appendix, executive summary, and appendix**).

A screenshot of a 'Word Count' dialog box. The title bar says 'Word Count' with a close button (X) on the right. The dialog contains a table with four rows: 'Total Words: 4940', 'Headers: 32', 'Math Inline: 37', and 'Math Display: 12'. At the bottom right, there is a 'Close' button.

Word Count		X
Total Words:	4940	
Headers:	32	
Math Inline:	37	
Math Display:	12	
		Close

Figure 17: Word Count Image