ANOCCA: Analysis of Consumer Disputes & Impact of Social Media on Litigation

Bikram Mondal[*], Kritika Gupta[†], Abhijnan Chakraborty[‡]

Abstract

In this paper, we provide a deep analysis of the areas of consumer dissatisfaction. To achieve this we have taken two platforms into account; the consumer disputes redressal commission and the social media platform and then provided their separate analysis and the relation between them. As the consumer case website does not provide any statistical analysis, we start in this area by analyzing the available data for providing information about the people, advocates, and companies involved in most of the cases, details of the companies involved and their major sectors, overall most affected sectors, and summarized details of time and money involved in the consumer cases. To understand the influence of social media on consumer court, frequent areas of faults, sentiment analysis of tweets, and customer interaction rate for companies are calculated to reach the conclusion.

Keywords: Consumer, Cases, Analysis, Twitter, Social Media.

1 Introduction

Consumers play a vital role in the economy as their purchasing decisions can drive the demand for products, influencing businesses and market trends.

[*] Student, Department of Computer Science & Engineering, Indian Institute of Technology, Delhi, India Email: mcs212127@iitd.ac.in

[†] Student, Department of Computer Science & Engineering, Indian Institute of Technology, Delhi, India Email: mcs212136@iitd.ac.in

[‡] Assistant Professor, Department of Computer Science & Engineering, Indian Institute of Technology, Delhi, India Email: abhijnan@iitd.ac.in

Companies strive to understand consumer behavior and preferences to effectively provide products and services that meet their needs and preferences. With the rise in consumer awareness and the boost of products and services, the need to address disputes and ensure fair treatment has led to the establishment of consumer courts. Observing the rise in cases in consumer courts, India has also moved ahead in the direction of protecting consumer rights and introduced the Consumer Protection Act, 2019, which provides a 3-tier structure of the National Commission, the State Commissions, and the District Commissions for the speedy resolution of consumer disputes[1]. National e-Governance Plan (NeGP), 2006, is an initiative of the Government of India to make all government services available to the citizens of India via electronic media[2]. In spite of these efforts by the government, there are still a lot of consumers unaware of their rights, and the website[3] containing the dataset of consumer cases does not provide any analysis of the cases in India. So, even when a consumer wants to file a case in court, they are unclear about many things, such as how long the process will take and whether they will be sufficiently compensated for their grievances.

In this age of high-speed internet and smartphones, all the newly established companies use social media as a platform for direct reach to the customer to advertise their products and offers. Consumers also prefer to post on social media about their complaints with the hope of getting faster replies and solutions. Companies, too, tend to solve complaints fast and with high customer satisfaction on social media to achieve customer retention.

Therefore, this area requires a detailed connection between them that is available to all. Statistical analysis of consumer cases can help a new consumer know beforehand what is likely to happen by comparing it with other similar cases. Lawyers can get an idea of where to focus with the help of most litigated sectors. Companies can know about their area of fault if there is some insight giving the areas where usually consumers are complaining, either in consumer court or on social media.

[1] National Informatics Centre, 'National Consumer Disputes Redressal Commission' (NCDRC) <https://ncdrc.nic.in/history.html> accessed 10 June 2023.

[2] Wikimedia Foundation, 'National e-Governance Plan' (Wikipedia, 9 March 2023) <https://en.wikipedia.org/wiki/National_e-Governance_Plan> accessed 10 June 2023.

[3] National Informatics Centre, 'Computerization and Computer Networking of Consumer Commissions in Country' (CONFONET) <https://confonet.nic.in/> accessed 10 June 2023.

2 Background

In this section, we have discussed some background of the sources used for our analysis such as Web Scraping using Selenium, Fuzzy Matching and Levenshtein Distance, Bidirectional LSTM, Twitter requests API, Aspect Based Sentiment Analysis, Summarizing using TextRank, and Topic Modeling using Latent Dirichlet Allocation (LDA).

2.1 Web Scraping using Selenium

For extracting large datasets present on the websites, web scraping is required[4]. Selenium is an umbrella project for a range of tools and libraries that enable and support the automation of web browsers[5]. It provides like finding elements on a website, selecting one out of the options in the drop-down menu, filling in the text box automatically, automating clicks, and many more.

2.2 Fuzzy Matching and Levenshtein Distance

When an exact match cannot be found between the words, the fuzzy match is used, in which the user sets a threshold of the fuzzy match less than 100%, and then get all the matches that are giving higher value than the threshold[6]. Fuzzy matching can be done using Python module fuzzywuzzy[7] which uses the Levenshtein distance between the sequences for comparison. Levenshtein distance between two sequences is the minimum number of edits(insertions, deletions, or substitutions) required to convert one sequence to another[8].

[4] Wikimedia Foundation, 'Web scraping' (Wikipedia, 20 May 2023) <https://en.wikipedia.org/wiki/Web_scraping> accessed 12 June 2023.

[5] 'The Selenium Browser Automation Project' (Selenium) <https://www.selenium.dev/documentation/> accessed 12 June 2023.

[6] Wikimedia Foundation, 'Fuzzy matching (computer-assisted translation)' (Wikipedia, 17 March 2023) <https://en.wikipedia.org/wiki/Fuzzy_matching_(computer-assisted_translation)> accessed 12 June 2023.

[7] 'fuzzywuzzy 0.18.0' (PyPi) <https://pypi.org/project/fuzzywuzzy/#description> accessed 12 June 2023.

[8] Wikimedia Foundation, 'Levenshtein distance' (Wikipedia, 11 May 2023) <https://en.wikipedia.org/wiki/Levenshtein_distance> accessed 12 June 2023.

2.3 Bidirectional LSTM

Recurrent neural networks(RNN) is a class of artificial neural network where the output of some nodes are fed as an input to others. RNNs have a hidden layer that has memory to store information about the sequence. Bidirectional long short-term memory (BLSTM) is a type of recurrent neural network that incorporates information about the sequence in both directions (forward and backward) at every step[9,10]. It is helpful in cases where the word in the middle is required, so to understand the context in which it occurs we need to understand the pattern from both sides.

2.4 Twitter requests API

Twitter API provides access to the developers for their projects[11]. Twitter requests package can be used for cases when tweets have to be extracted in some unique manner. It gives the option of searching tweets in an advanced way, for example, tweets posted within a duration of time, tweets containing a particular tag, can set maximum tweets returned in a request, and many more.

2.5 Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis requires two steps; aspect extraction[12] and sentiment analysis of that respective aspect[13]. To achieve ABSA results, PyABSA can be used, it supports several ABSA subtasks, including aspect term extraction, aspect sentiment classification, and respective aspect-based sentiment analysis[14]. PyABSA is a framework built on PyTorch, making its

[9] M. Schuster and K. K. Paliwal, ‘Bidirectional recurrent neural networks’ in IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093.

[10] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[11] ‘Twitter API’ (Twitter) <https://developer.twitter.com/en/docs/twitter-api> accessed 12 June 2023.

[12] Shu, L., Xu, H., & Liu, B. (2017). Lifelong Learning CRF for Supervised Aspect Extraction. ArXiv. /abs/1705.00251

[13] Wang, Yequan & Huang, Minlie & Zhu, Xiaoyan & Zhao, Li. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. 606-615. 10.18653/v1/D16-1058.

[14] Yang, H., & Li, K. (2022). PyABSA: A Modularized Framework for Reproducible Aspect-based Sentiment Analysis. ArXiv. /abs/2208.01368.

results reproducible. It provides 29 models and 26 datasets to work on easily, and one can also test their dataset on PyABSA.

2.6 Summarizing using TextRank

TextRank is a graph-based ranking model for text processing and for extracting keywords and sentences[15]. The accuracy achieved by TextRank is competitive with that of previously proposed state-of-the-art algorithms. TextRank does not require deep linguistic knowledge, nor domain or language-specific annotated corpora, which makes it highly portable to other domains, genres, or languages.

2.7 Topic Modeling using Latent Dirichlet Allocation (LDA)

Topic modeling is a statistical technique for revealing the underlying semantic structure in a large collection of documents. It is a technique that comes with a group of algorithms that reveal, discover and annotate thematic structure in a collection of documents[16]. Latent Dirichlet Allocation is one of the ways to perform topic modeling. LDA is a generative probabilistic model for collections of discrete data such as text corpora[17]. pyLDAvis is designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization[18].

3 Related Work

There is growing attention from researchers and practitioners in recent years on the legal domain and the social media domain. This section focuses on some of the related works existing in the literature.

[15] Mihalcea, Rada, and Paul Tarau. ‘Textrank: Bringing order into text.’ Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[16] Kherwa, Pooja & Bansal, Poonam. (2018). Topic Modeling: A Comprehensive Review. ICST Transactions on Scalable Information Systems. 7. 159623. 10.4108/eai.13-7-2018.159623.

[17] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ‘Latent dirichlet allocation.” Journal of machine Learning research 3.Jan (2003): 993-1022.

[18] Mabey, Ben. ‘pyLDAvis documentation.’ (2021).

In the paper[19], the authors discuss consumer power in marketing theory and changing power dynamics with case studies. They concluded that companies need to increase their website awareness by efficiently using communication tools; enhancing their website's findability and credibility by working on their website's content; introducing convenient payment and on-time delivery options; and finally consulting with their consumers and e-communities for feedback at the post-purchase stages. So it's high time that companies wake up to the realization of the increased power of the online consumer and respond accordingly.

Konattu et al in their paper, have evaluated the performance of the Consumer Disputes Redressal Commission for the last 8 years by using mean, standard deviation, and correlation coefficient[20]. They have made a statistical analysis of the number of pending cases and the rate of disposal of the cases from the consumer court.

In the paper[21], they have focused on the analysis of the performance of consumer dispute redressal forum. They have analyzed various things like the satisfaction level of the respondents on the decisions made in the forum, how much of the claim was actually received, the time taken for a particular redressal case, and whether any pressure was put on the forum to take the decision in favour of a specific party in the context of Mathura, Uttar Pradesh.

For the retrieval of legal cases a knowledge representation model has been developed by taking 152 examples from the domain of accident compensation in the paper[22]. In their representation model, issues are decomposed into subissues, factors are categorized into pro-claimant and pro-respondent factors, and some contextual features like *Role of Claimant*, *Role of Respondent*, *Injury Person* and *Compensation Type* are considered and these features are identified manually. When a new case comes, after analyzing the case facts

[19] S. Umit Kucuk, Sandeep Krishnamurthy, An analysis of consumer power on the Internet, Technovation, Volume 27, Issues 1–2, 2007, Pages 47-56, ISSN 0166-4972, <https://doi.org/10.1016/j.technovation.2006.05.002>.

[20] Mrs. Elwin Paul Konattu and Dr.V.K.Sudhakaran, 'A Critical Evaluation on the Performance of Consumer Disputes Redressal commission in India', IOSR Journal of Business and Management, Volume 20, Issue 9. Ver. IV (September. 2018), PP 47-52.

[21] Sharma and Hari, 'An Analysis of Performance of Consumer Redressal Forum', Siddhant- A Journal of Decision Making, 17, 2017, 277-286.

[22] Yiming Zeng and others, 'A Knowledge Representation Model for the Intelligent Retrieval of Legal Cases', International Journal of Law and Information Technology, Volume 15, Issue 3, Autumn 2007, Pages 299–319, <https://doi.org/10.1093/ijlit/eal023>.

this retrieval scheme helps the users to effectively locate the most relevant precedents.

Michael Aikenhead et al have proposed distributed artificial intelligence-based simulation for the investigation of law[23]. They argued that both the existing simulations in the field of legal studies (*legal knowledge-based systems* and *knowledge-based micro simulatuon*) adopt the comparatively micro perspectives of law, whereas their approach encourages a more macro view in simulation.

In this paper, the author examined the use of machine learning technologies to suppress or block access to social media to al-Qaida and IS-inspired propaganda[24]. The concerns are not with this objective but with how social media purport to achieve this important goal. The author argues that, in their design and manner of addressing the legitimate goal of fighting terrorism, machine learning systems, if heavily used, will chill freedom of expression, potentially unfairly target the Muslim community, and at the same time not make us safer from valid terrorist threats.

4 Problem Statement

Currently, there is no detailed analysis of the consumer court cases in India like statistics of advocates, complainants, respondents, duration of cases, types of business implicated, and range of disputes. There is also a lack of in-depth understanding of how social media influences consumer court cases in India.

So, the main objectives of this paper are

- to analyze which are the most litigated sectors all over India. Most litigated sectors are the sectors for which there is a significant volume of cases filed against them in the consumer court[25].

[23] M Aikenhead and others, 'Exploring law through computer simulation', International Journal of Law and Information Technology, Volume 7, Issue 3, AUTUMN, Pages 191–217, <https://doi.org/10.1093/ijlit/7.3.191>.

[24] Jamil Ammar, Cyber Gremlin: social networking, machine learning and the global war on Al-Qaida-and IS-inspired terrorism, International Journal of Law and Information Technology, Volume 27, Issue 3, Autumn 2019, Pages 238–265, <https://doi.org/10.1093/ijlit/eaz006

[25] Sectors include areas like insurance, construction, finance, trading, and many more.

- to find the range of monetary value involved for the disputes addressed in the consumer court.
- to observe frequent parties and advocates involved in various cases in different regions of India.
- to make a comprehensive analysis of various types of cases[26] in a particular duration range and to make statistical analysis on the average time for disposing of the cases.
- to learn whether there is any relation between social media comments and consumer cases and to analyze how social media platforms like Twitter influence the reduction of consumer court cases in India.

5 Dataset

For our research, the datasets we used comprised three types of data. These are consumer court case records, company details datasets and social media data. The source, retrieval process, and description of these datasets are given below. We have used the Postgres SQL database for storing the data efficiently[27].

5.1 Consumer Court Case Records

The dataset for consumer court case records is collected from the consumer dispute redressal commissions website 'Confonet'[28] using web-scraping[29]. We have used Python selenium for web-scraping and collected the data efficiently[30].

The retrieved dataset consists of records of various cases filed in consumer disputes redressal forums for the duration of 5 years (2017 - 2021). Each of

[26] Various types of cases include First Appeal, Revised Petition, Transfer Application, Execution Application, and many more.

[27] 'PostgresSQL: Documentation' (Postgres, 25 May 2023) <https://www.postgresql.org/docs/> accessed 13 June 2023.

[28] National Informatics Centre, 'CONFONET JUDGEMENT SEARCH' (ConfoNet) <https://cms.nic.in/ncdrcusersWeb/search.do?method=loadSearchPub> accessed 14 March 2022.

[29] See n 4.

[30] See n 5.

these records includes case number, complainant, respondent, complainant advocate, respondent advocate, date of filing, date of disposal, date of upload, and case document link. There are a total of 38 288 records across 16 states in that duration.

5.2 Company Details Dataset

The whole company details dataset is collected from the Ministry of Corporate Affairs website[31].

The dataset contains records of all the registered companies in India. Each record consists of the corporate identification number, company name, company status, class of company, company category, authorized capital, paid-up capital, date of registration, registered state, register of companies, principal business activity registered office address, and sub-category. In total, there are 10 48 575 records.

This whole dataset is not required for our analysis. So, we have constructed a smaller dataset by using this dataset and collecting some additional data from various online forums and websites, we will discuss it more in the analysis section.

5.3 Social Media Data

We have collected this dataset from the popular social media 'Twitter' through API call[32].

The tweets are related to consumer grievances, opinions, and discussions about various products, services of a particular company, businesses, and many more. We have collected tweets from around 60 companies for which there is a large no of cases in the consumer court and tweets from 25 more new-age companies. Each tweet data comprises of tweet id, tweet text, author id, created at, conversation id and language.

[31] Ministry of Corporate Affairs, 'Data & Reports' (MCA, 12 June 2023) <https://www.mca.gov.in/MinistryV2/archiveofmasterdatadetails.html> accessed 31 August 2022

[32] See n 11.

Data Preprocessing

As the consumer court case dataset contains a lot of ambiguity, spelling mistakes, and non-uniform usage of punctuation, some sort of data-cleaning and data preprocessing is required. Following are the preprocessing steps we have followed.

- In some of the records, the same complainant or respondent name and the same advocate name are written in different forms with a difference of very few characters. So, to make them uniform we have used fuzzy-matching[33] and if two names are more than 80% similar, we have mapped them into one single name. We have used the *fuzzywuzzy* module of Python for this purpose[34].

- The date format in the date of filing and date of disposal field of the records is not uniform for all the states. We have converted all the date fields in DD-MM-YYYY format as this field is required for duration analysis.

6 Analysis

In this section, we examine all the objectives one by one by discussing the methods we have used and visualizing the results achieved.

6.1 Most litigated Sectors

To analyze the most litigated sectors, firstly we have retrieved the top 200 complainants and respondents all over India using SQL queries on consumer court case records. The stop-words and punctuation are removed from the company names and also the names are converted to lowercase for matching purposes. We have now matched each of those names with the company dataset using Levenshtein distance in Postgres SQL.

Algorithm 1 describes the algorithm of matching the complainant and respondent names with the company database and finally finding the sector(principal business activity) for that company. We have the company table and a table with the complainant and respondent names, we want

[33] See n 5.

[34] See n 6.

to find out the sector in which the company belongs. We have taken a *threshold* as 0.3. Now for each *data*1 and *data*2 where *data*1 belongs to the complainant and respondent name and *data*2 belongs to the company name in the company table, we have set *min_ratio* as 1 initially, the *res* is empty initially, we have calculated the *ratio* as Levenshtein distance of *data*1 and *data*2 divided by maximum length between *data*1 and *data*2. The lower the ratio, the better the accuracy. Now the *data*2 for which *ratio* is less than *threshold* and which gives minimum ratio, we have taken that *data*2 as mapping of *data*1. We have also stored the corresponding sector name.

Algorithm 1: Matching Company Names & Finding Sectors

```
Data: Company Table, Complainant and Respondent Names
Result: List with Complainant and Respondent Names, Original
        company name, sector name
$threshold \leftarrow 0.3$;
for data1 where data1 ∈ Complainant and Respondents name do
    for data2 where data2 ∈ Company name in Company table do
        $min\_ratio \leftarrow 1$;
        $res = empty$
        $ratio = \frac{levenstein\_dist(data1,data2)}{max(length(data1),length(data2))}$
        if $ratio < threshold$ then
            if $ratio < min\_ratio$ then
                $min\_ratio = ratio$;
                $res = (data2, data1, sector)$
            end
        end
    end
    output.append(res)
end
```

This process does not give matching information for all 200 companies. We required the help of online sources to get the sectors for those companies.

Using the above information we have found out the total number of cases for each sector and the companies involved in each sector using SQL queries.

Figure 1 shows the plot for different sectors vs the total number of cases all over India. From the plot, we can see that the Insurance sector is the most litigated sector whereas Mining & Quarrying is the least litigated sector. In the Insurance sector, there are 4295 cases whereas in Mining & Quarrying sector there are only 33 cases.

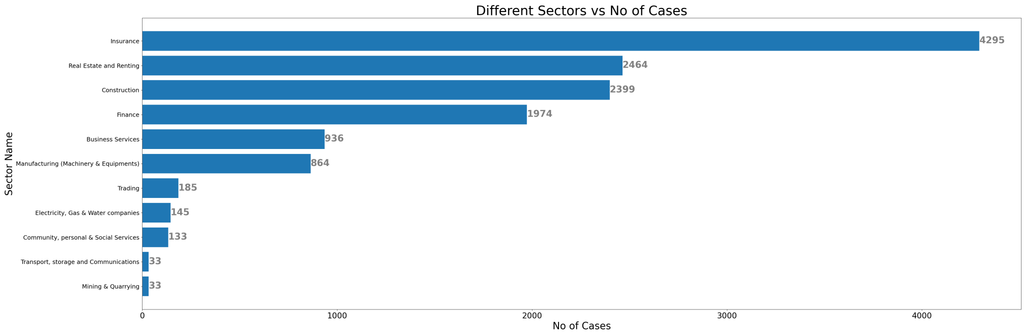

Figure 1: Sector-wise Analysis

6.2 Range of Disputed Monetary Value

The statistics of the range of amounts at stake for the disputes addressed in consumer court are very interesting. Figure 2 shows the algorithmic workflow for finding out these statistics.

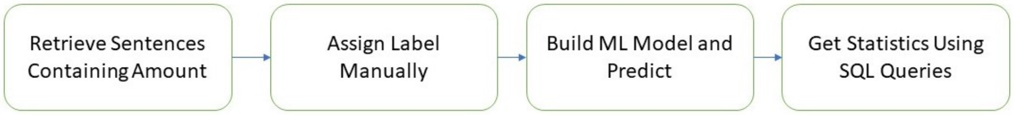
Retrieve Sentences Containing Amount
Assign Label Manually
Build ML Model and Predict
Get Statistics Using SQL Queries

Figure 2: Work-flow for finding the range of amount

The first step is to retrieve all the sentences that contain some amount part from each case document. The concept of regular expression is used for the matching purpose. We have used the *RegEx* module of python[35]. As there is a lot of ambiguity in the case documents for representing the amount, we are required to consider various regular expressions.

Then we have manually assigned labels to some of the sentences (around 410) for building a classification-based ML model. We need it because for one document there may exist more than one sentence with amount part among which all are not representing the amount at stake. Some sentences represent the cost of harassment, mental agony, litigation charges, and many

[35] The Python Software Foundation, 're — Regular expression operations' (Python, 12 June 2023) <https://docs.python.org/3/library/re.html> accessed 13 June 2023.

more. So if the sentence contains the amount at stake then the corresponding level has been assigned as 1, otherwise 0.

The next step is to fit a classification-based machine learning model on the built dataset. We have tried with Naive-Bayes classifier[36], Stochastic Gradient Descent classifier[37], Logistic Regression[38] and with Bi-directional LSTM[39] for classification with 20% of test data. We have noticed that BiLSTM is giving the best accuracy (77%) among all of these. So, we have trained a model using Bi-LSTM and predicted for all of the sentences retrieved in the first step.

Finally, we have created six groups of amount range[40] and using the results of the previous step, SQL queries are used to count the number of cases for each of the amount buckets.

Range of Amount	**Number of Cases**
Below Rs. 1000/-	260
Rs. 1000/- to Rs. 10 000/-	668
Rs. 10 000/- to Rs. 1 00 000/-	1299
Rs. 1 00 000/- to Rs. 10 00 000/-	1276
Rs. 10 00 000/- to Rs. 1 00 00 000/-	707
Above Rs. 1 00 00 000/-	78

Table 1: Statistics of range of amount

Table 1 depicts the statistics of the range of amount and the corresponding number of cases for the disputes. From the table we can say that the maximum number of cases lies in the range of Rs. 10 000/- to Rs. 10 00 000/-.

[36] scikit-learn developers, 'Naive Bayes' (scikit-learn) <https://scikit-learn.org/stable/modules/naive_bayes.html> accessed 13 June 2023.

[37] scikit-learn developers, 'SGDClassifier' (scikit-learn) <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html> accessed 13 June 2023.

[38] scikit-learn developers, 'LogisticRegression' (scikit-learn) <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html> accessed 13 June 2023.

[39] See n 9, 10.

[40] The ranges can be found in Table 1.

6.3 Frequent Parties and Advocates

In the data preprocessing section, we have discussed fuzzy matching and its importance in our context. Once that is done, to get the analysis of frequent parties, we need the information of people and companies mostly involved in cases. To achieve this, we merged the complainant and respondent columns and grouped them using the names to get the count of the number of cases for each company or person. The result is finally sorted in descending order based on the count of cases and the top 10 are retrieved. SQL queries are used for this purpose.

For the analysis of frequent advocates, the exact same process is followed. The only difference is instead of using the complainant and respondent column, we have used the complainant advocate and respondent advocate column in this case.

Parties Involved	**Number of Cases**
NIC Ltd.	416
HDFC ERGO General Insurance	376
Aminder Singh	273
Life Insurance Corporation	273
SP Developers	267
Sky Rock City Welfare Society	265
NGHI Developers India Ltd.	263
Bathinda Development Authority	261
Oriental Insurance Company Ltd.	255
Bajwa Developers Ltd.	240

Table 2: Frequent Parties vs Number of Cases

Table 2 and Table 3 represent the top 10 parties and advocates involved in the cases respectively with their corresponding number of cases all over India. National Insurance Company Limited (NIC Ltd.) has the maximum number of cases (Around 416) and Advocate Ashish Bhawnani represented the maximum number of cases (Around 435) in the consumer court of India during the interval of 2017-2021. We have also made a state-wise analysis and all the analysis is shown on the website developed by us[41,42].

Advocates Involved	Number of Cases
Ashish Bhawnani	435
P.K. Paul	341
Hari Babu	305
R.K. Pattnaik	246
Subrata Mondal	198
Sandeep Bhardwaj	183
K. Venkateswarlu	182
Shekhar Prakash Shrivastava	182
Prasanta Banerjee	170
Sunila Jain	153

Table 3: Frequent Advocates vs Number of Cases

6.4 Duration Analysis

This section describes the number of cases characterized based on the duration to solve them, the average time to solve the cases and the statistics of various types of cases for each duration range.

We have created six groups of duration range i.e. less than one week, one week to one month, one month to six months, six months to one year, one year to five years, and greater than five years. Using the date of filing and the date of disposal columns, all the cases are mapped to one of the six range buckets based on the difference between the two dates, and finally count of cases for each bucket is calculated using SQL queries.

Figure 3 shows the plot of the duration range against the number of cases. From the figure, we can conclude that around 64% of cases are resolved within 1 year. In fact, we have examined that the average time taken for all the cases is 1 year and 37 days.

[41] Link of the website (CONSUMER COURT ANALYTICS): <http://consumercourtcases-env-1.eba-pxp5mcew.ap-south-1.elasticbeanstalk.com/>

[42] Technologies Used: HTML, CSS, JS, JQuery, Python Flask, CanvasJS (Documentation: <https://canvasjs.com/docs/charts/basics-of-creating-html5-chart/>).

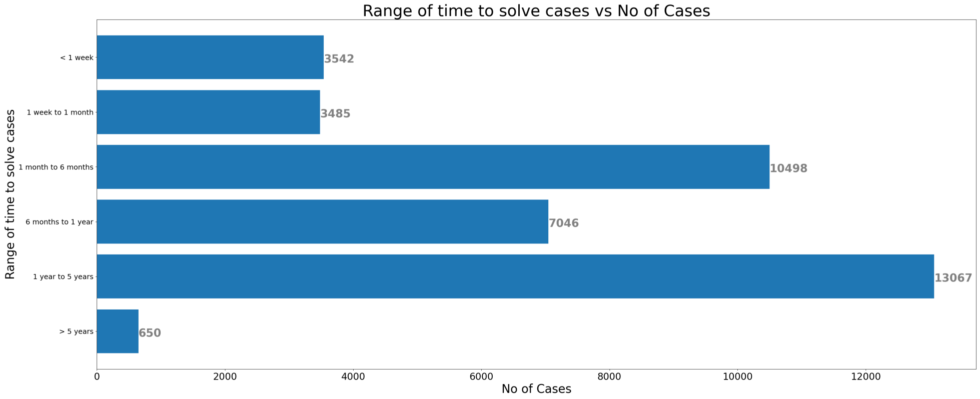

Figure 3: duration analysis

Generally, the prefix of the case number represents the type of the case[43] which we have used to create the categories of type of cases. We have retrieved the statistics of the number of cases for each type in each of the above-mentioned duration ranges. Table 4 represents the statistics of the cases which take more than 5 years. For the remaining range buckets, the analysis can be found on our website[44].

Type of Cases	**Number of cases**
Appeal/First Appeal	418
Consumer Case	202
Execution Application	17
Revised Petition	11
Miscellaneous Application	1

Table 4: Statistics of Cases solved in more than 5 years

We have also found out the average time to solve various types of cases and the result is depicted in Table 6.

[43] FA: First Appeal, CC: Consumer Case, RP: Revised Petition, TA: Transfer Application, IA: Interlocutory Application, RA: Review Application, EA: Execution Application, MA: Miscellaneous Application.

[44] See n 41.

Type of Cases	Average Time in Days
Consumer Case (CC)	504
Appeal/First Appeal (A/FA)	451
Execution Application (EA)	450
Revised Petition (RP)	202
Miscellaneous Application (MA)	102
Interlocutory Application (IA)	70
Review Application (RA)	63
Transfer Application (TA)	43

Table 5: Statistics of average time

6.5 Relation of Social Media with Consumer Cases

In this section, we discuss whether there is any common relation between the tweets posted by the consumers and the disputes filed in the consumer court. The procedure we followed is described in Figure 4.

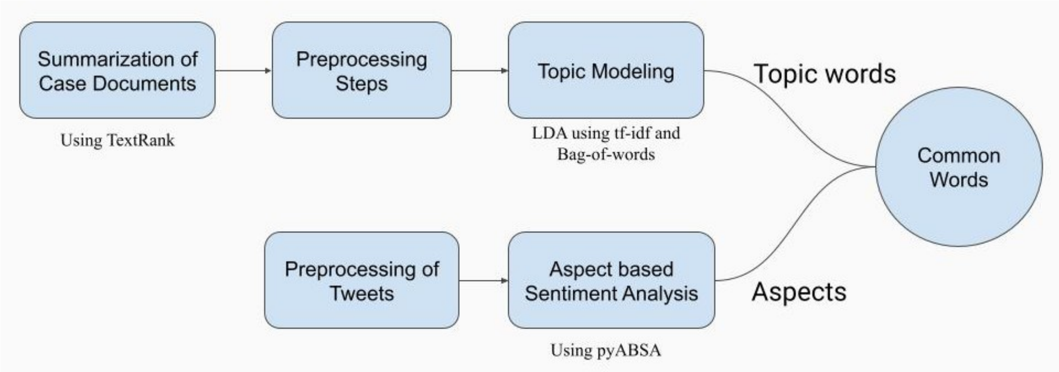

Figure 4: Work-flow for finding the relation between court cases and tweets

After reading the text of the case documents (in PDF form) using *PyPDF2*[45], we have obtained the summaries of those using *TextRank* summarizer[46].

[45] The Python Software Foundation, ‘PyPDF2’ (Python, 31 December 2022) <`https://pypi.org/project/PyPDF2/`> accessed 22 June 2023.

[46] See n 15.

These summaries contain stopwords and punctuation, which we have removed. The words are also lemmatized[47]. The summaries are then divided into clusters using LDA[48]. We have applied LDA using two different models; the Bag of Words Model[49] and the Term Frequency - Inverse Document Frequency (TF-IDF) model[50]. Table 6 gives the performance measure of these two models. Perplexity is the measure of how good a model is in predicting a given set of case documents, lower perplexity indicates better performance. Coherence is the measure of human interpretability of the topics, higher is better. As Table 6 indicates TF-IDF is better than the bag of words in our case, we have used LDA using TF-IDF for topic modeling. Along with the division of summaries, LDA also gives the topic words for each of the clusters.

	Coherence	**Perplexity**
Bag of Words	0.56	-7.26
TF-IDF	0.60	-9.34

Table 6: Performance measure of Bag-of-words and TF-IDF model

We also have the tweets of the litigated companies. After some preprocessing of the tweets like the removal of stop-words and punctuation, aspects from the tweets are extracted and for each aspect, the sentiment (positive, negative, or neutral) is checked. pyABSA is used for this purpose[51].

Common words between the topic words of case documents and the aspects with more negative sentiment for a particular company give a sense of the areas where the consumers are tweeting as well as filing cases against that company. Figure 5 describes this relationship for the company Unitech Limited for one topic cluster. A detailed analysis for all the companies is represented in our website[52]. The website also provides respective cases for each topic.

[47] Lemmatization is the process of grouping different forms of the same word.

[48] Seen n 17.

[49] Wikimedia Foundation, 'Bag-of-words model' (Wikipedia, 17 June 2023) <https://en.wikipedia.org/wiki/Bag-of-words_model> accessed 23 June 2023.

[50] Wikimedia Foundation, 'tf-idf' (Wikipedia, 27 May 2023) <https://en.wikipedia.org/wiki/Tf-idf> accessed 23 June 2023.

[51] See n 14.

[52] See n 41.

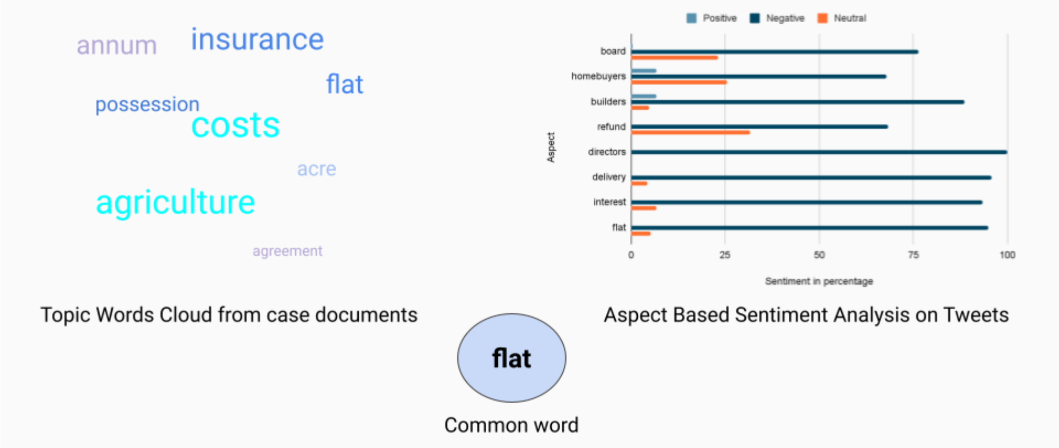

Figure 5: relationship of cases and tweets for Unitech Ltd.

This analysis helps the companies to find their broad area of faults like flat-related issues for Unitech Limited. It also helps the consumers to find similar cases by seeing the topic words.

6.6 Influence of Twitter in the reduction of the number of legal disputes

This section describes the influence of social media on consumer court cases. For this, we have divided the litigated companies into two types, Twitter inactive and Twitter active, and added a third type of companies that are active on Twitter but don't have any case against them. This gives us the set of three types: Twitter inactive and litigated[53], Twitter active and litigated[54], and Twitter active and non-litigated[55]. Now, we can compare the litigated and non-litigated companies that are active on Twitter using the Tweet dataset. Using the conversation id we have retrieved the whole conversation that happened after that tweet. In that conversation, if the respective company has replied, we consider it as the company's reply. Now, this dataset of tweets will be the Replies Dataset[56].

[53] Companies like SP Developers, Sky Rock City Welfare Society, and some others

[54] Companies like State Bank of India, National Insurance Company Ltd., and some others

[55] Companies like Swiggy, Zomato, Digit Insurance, and some others

[56] Contains fields: author_id, tweet_id, conversation_id, tweet_text

While retrieving this dataset, we can fetch the statistics of the reply rate of the company. This rate can be calculated by counting the number of times a company replies out of the number of times it is tagged in tweets. The reply percentage of 16 companies (8 for each type) is shown in the box plot 6. We can note that the customer interaction rate of litigated companies is much less than that of non-litigated companies with some exceptions.

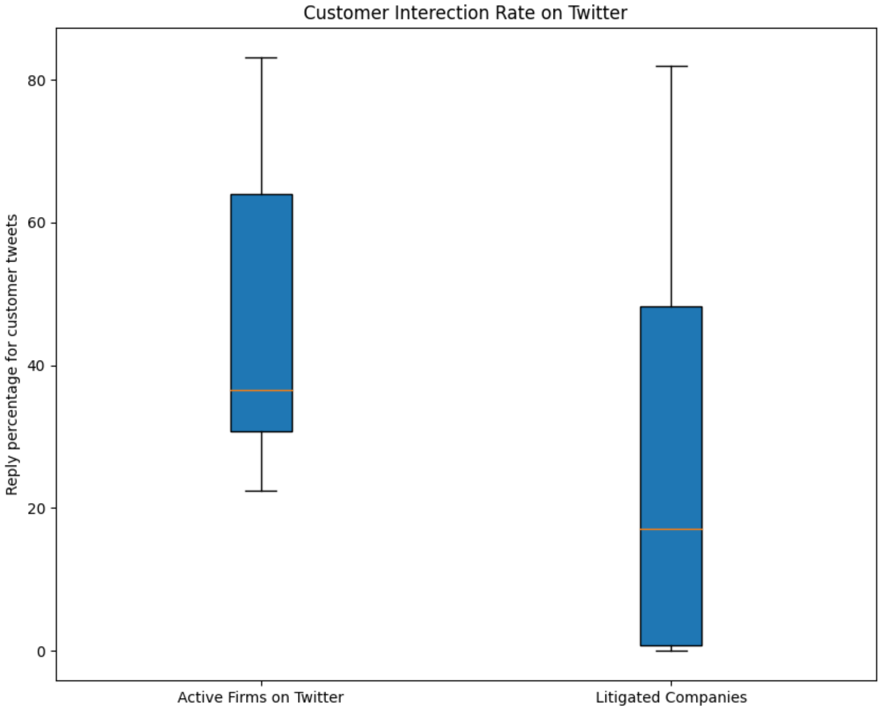

Figure 6: Customer Interaction Rate on Twitter

For two litigated companies whose customer interaction is higher, we have further researched to find their pattern of interaction growth over the years with comparison to progression in the number of cases against them. Table 7 shows that for companies like HDFC ERGO General Insurance Company and Central Bank of India, increasing customer interaction on Twitter has significantly reduced their respective number of cases.

Year	**Customer Interaction Rate**	**Number of Cases**
2017	0.00%	121
2021	71.12%	15

(a) Central Bank of India

Year	**Customer Interaction Rate**	**Number of Cases**
2019	70.58%	176
2021	81.99%	6

(b) HDFC ERGO General Insurance Company

Table 7: Year-wise customer interaction rate and number of cases

This analysis brings us to the following conclusion:

- Twitter active litigated companies: Increasing customer interaction rate can reduce the number of cases filed against them.
- Twitter inactive litigated companies: Start by creating a specific twitter handle for customer's problem resolution and try to solve most of the customer's complaints on such platform.

7 Conclusion

This paper gives a broad overview of the whole judicial system of India. On one hand, it provides the statistical analysis of the cases and on the other hand, it also draws a conclusion about the relationship between the tweets of the consumers for various companies and the cases filed against the same company in the consumer disputes redressal commission.

From the results of various analyses, some conclusions can be drawn. As the Insurance, Real Estate and Renting, Construction, and Finance sectors are most litigated so these companies have to analyze their faults and do the needful to make the consumers happy and reduce the pressure on the judicial system. The analysis of the range of disputed monetary value gives us the insight that the number of low-profile cases in terms of amount (Below Rs. 10 000/-) is less and high-profile cases in terms of amount (Above Rs. 10 00 000/-) is also less. The number of cases in the range of Rs. 10 000/- to Rs. 10 00 00/- is the maximum. Frequent companies like NIC Ltd, HDFC ERGO General Insurance, Life Insurance Corporation, and others have to meet customer expectations for the reduction of the number of cases against them. By seeing the analysis of the frequent advocates, people can easily find the popular advocates for approaching their cases. By observing the statistics of the average time, people can make an intelligent guess of how much time it may take to process their cases. The average time taken for all the cases in the duration of 2017-2021 is 1 year 37 days. The judicial system has to pay attention to this to reduce the processing time and should aim for the timely disposal of the cases.

Grouping case documents based on topic words helps customers to find similar types of cases while filing a case. Companies by looking at the common topics of tweets and cases can understand their broad area of faults as customers are complaining about it on Twitter and court. Twitter has a huge impact on the legal area because increasing customer interaction rate on social media can significantly reduce the number of consumer court cases against the companies, as we have seen in the case of HDFC ERGO General Insurance Company, and Central Bank of India. Hence, it is suggested that inactive companies take the initiative in interacting with their customers on social media platforms like Twitter, whereas already active and litigated companies still have a scope to improve their interaction rate by comparing them with the non-litigated active companies.