ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY214

Sentim (IOCL)
Unlocking Sentiments: Enhancing IOCL Petrol Pump Experiences

[1]Megha Gupta; [2]Chirasha Jain; [3]Ishita Jain; [4]Shivam Bisht; [5]Deepanshu
[1]Assistant Professor; [2,3,4,5]U.G. Student,
Computer Science & Engineering
Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India

Abstract:- This study, titled "Sentim(IOCL):Unlocking Sentiments: Enhancing IOCL Petrol Pump Experiences," delves deeply into the rich tapestry of public comments surrounding petrol pumps, with focus on discerning the sentiments and opinions relevant to IOCL. By employing cutting-edge natural language processing techniques, we extract explicit aspects from these comments and to gain a nuanced understanding of the sentiments associated with different facets. Our goal is to develop a usability index for selected petrol pumps, offering invaluable insights into their strengths and areas for refinement as perceived by the general populace. We're moving away from the usual method where each sentence is looked at separately. Instead, we're taking a more detailed approach that considers how different parts of the comments relate to each other. This way, we can understand not just what people are saying but also the reasons behind it. Our goal is to make a big contribution to understanding people's opinions by creating a method that looks at the whole picture, not just individual parts. By doing this, we hope to give IOCL and other companies in the industry practical advice on how to make their customers happier and keep getting better.

Keywords:- *Sentim IOCL, IOCL, Petrol Pumps, Public Comments, Sentiment Analysis, Natural Language Processing, Usability Index, Opinion Mining, Customer Satisfaction, Improvement, Personalized Research.*

I. INTRODUCTION

Understanding customer sentiment is crucial in today's competitive business landscape, especially in industries like petroleum, where satisfaction significantly influences brand loyalty and repeat business. Traditional methods of assessing satisfaction, such as surveys or focus groups, often fall short in capturing the breadth of customer feedback and can be both costly and time-consuming. Sentim(IOCL) aims to revolutionize this process by harnessing the power of Natural Language Processing (NLP) to analyze public comments on petrol pumps, providing deeper insights into customer sentiment.

The proposed method utilizes NLP techniques to extract key aspects discussed in customer reviews and categorize sentiments towards these aspects. By analyzing vast amounts of feedback data, this approach can identify areas for improvement across various operational aspects of petrol pumps, including staff behavior, cleanliness, and product quality, which may not be easily accessible through traditional means.

To showcase the efficacy of this approach, Sentim(IOCL) conducted an experiment using customer reviews of IOCL petrol pumps. This experiment involved several stages, including text cleaning, aspect extraction, and sentiment orientation. The results of this experiment, alongside discussions on its limitations and potential for future applications within the petroleum industry, will be elaborated on in subsequent sections of this paper.

The objective of the research paper on the Sentim IOCL project is to investigate the efficacy of Natural Language Processing (NLP) techniques in analyzing customer sentiment towards Indian Oil Corporation Limited (IOCL) petrol pumps. Through a comprehensive exploration of customer feedback data, the paper aims to develop and evaluate a sentiment analysis framework that leverages NLP algorithms to extract insights into customer perceptions, preferences, and concerns. The research seeks to provide actionable recommendations for IOCL to enhance customer satisfaction, improve operational efficiency, and strengthen its competitive position in the petroleum industry.

II. RELATED WORK

Sentim(IOCL) represents a significant advancement in customer sentiment analysis within the petroleum industry, outperforming traditional methods plagued by manual processes and disconnected tools. By employing cutting-edge Natural Language Processing (NLP) techniques, Sentim(IOCL) delivers a more thorough understanding of customer feedback, tailored to diverse geographical regions and demographics. Its user-friendly interface and robust analytical capabilities empower Indian Oil Corporation Limited and stakeholders to make informed decisions, driving improvements in customer satisfaction and operational efficiency.

In related work, prior studies have extensively explored sentiment analysis and natural language processing (NLP) applications in understanding customer feedback. Hu and Liu (2004) and Liu (2012) laid foundational groundwork by mining and summarizing customer reviews, underlining the importance of sentiment analysis. Additionally, research by Chen et al. (2014), Pontiki et al. (2014), and Wang et al. (2016) advanced the field by investigating aspect extraction and sentiment categorization techniques. While these studies offer valuable insights, the proposed research focuses specifically on the petroleum industry, with a targeted examination of customer sentiment towards Indian Oil Corporation Limited (IOCL) petrol pumps. By leveraging state-of-the-art NLP algorithms, this study aims to contribute novel insights tailored to IOCL's challenges and opportunities in enhancing customer satisfaction and operational effectiveness.

III. METHODOLOGY

The main challenge is to analyze user reviews from our database to understand their emotions and determine overall satisfaction levels. We identify frequent nouns or phrases as aspects and analyze sentiment towards these aspects to create a sentiment orientation matrix. Each comment's sentiment orientation on a specific aspect is represented in a quadruple (ei, aij, sijk, hk), where ei is the topic, aij is an aspect of ei, sijk is the sentiment orientation on aij of ei, and hk is the opinion holder, with a focus on IOCL Petrol Pumps.This work tries to solve the problem by performing classification using the sentiment orientation matrix. The whole procedure is implemented as the following five steps.

A. *Text Cleaning*

The dataset used for analysis is a bagful of comments. For the following steps, all the comments are cleaned based on sentence-level. All the sentences are tokenized, and words in the sentences are lemmatized to get unigrams by Part-of-Speech Tagging. However, at the same time, stop words are not abandoned.

B. *Aspect Extraction*

Under the reviews, people are discussing different sub-topics, which are called aspects of this work, such as staff behavior. This project focuses only on explicit aspects based on the frequency, which are nouns and noun phrases appearing in the comment body. In this case, all the noun phrases are assumed to be composed of two words, i.e., all the noun phrases are bi-grams. Based on the uni-grams, bi-grams are built at the sentence level to find candidate noun phrases. This project assumes that the candidate noun phrases should be in the pattern of two nouns, or stop word plus noun or adjective plus noun. So, to efficiently get the bi-grams, bi-grams without nouns inside are abandoned. To obtain the valid noun phrases and rank them, the model calculates Pointwise Mutual Information (PMI) scores for all the bi-grams. As shown in equation 1, PMI values take into account the correlation between the two words inside the noun phrase, avoiding the case where the noun phrases, in fact, are the aspects of the aspects. 50 bi-grams with the highest PMI scores are selected, but the meaningless ones among them are abandoned.

➢ *Equation 1: PMI Score*

$$PMI(x,y) = \log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

Except for these selected noun phrases, all the other bi-grams are split into uni-grams again, and the corpus is recleaned by removing the stop words. With comments as documents, I calculate the term frequency-inverse document frequency (TF-IDF) for all the nouns and candidate noun phrases. Twenty features with the highest TF-IDF (refer to equation 2) values are selected as aspects.

➢ *Equation 2: TF-IDF*

$$TF-IDF-e = \sum_{j=1}^{n} \frac{TF-IDF_{j,d}}{e^{distance(i,j)}}$$

$$TF-IDF_{j,d} = TF * IDF$$

$$TF = \frac{n_{id}}{\Sigma\, kn_{id}}$$

$$IDF = log\frac{|D|}{N_i}$$

C. *Aspect Categorization*

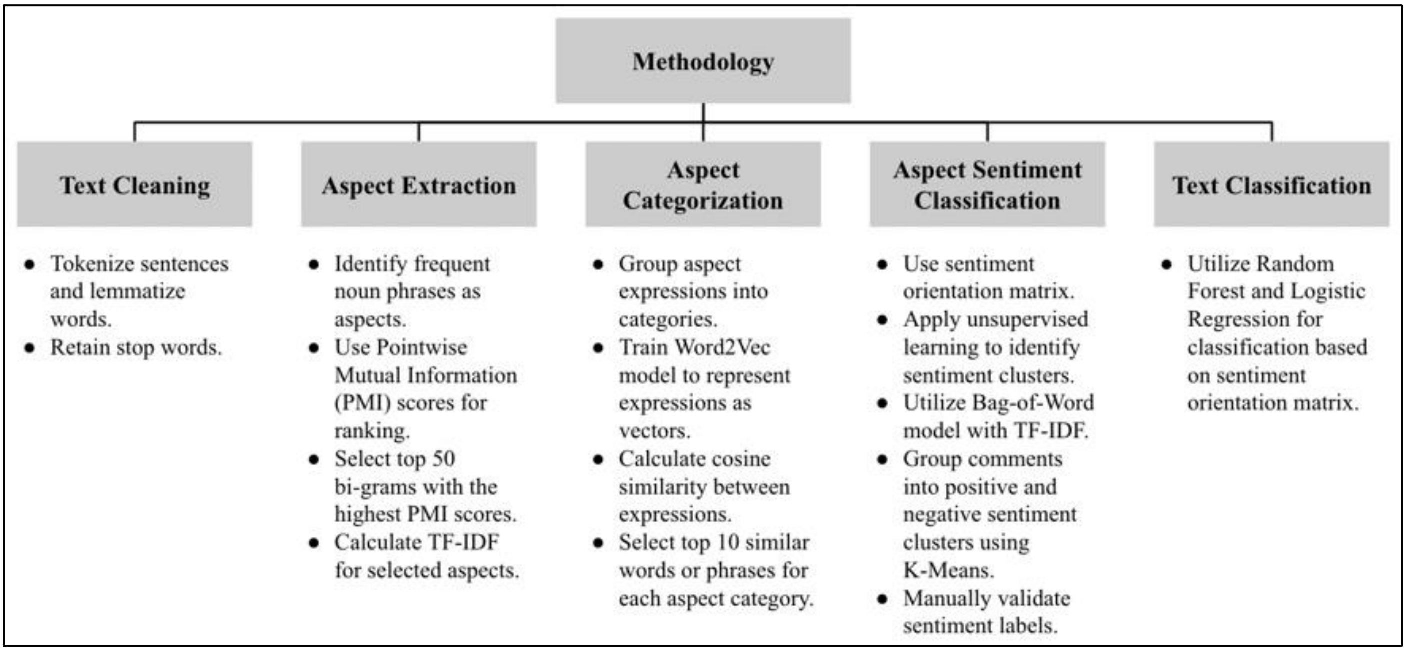
The aspects found from the last step are the explicit aspects of this analysis, but these aspects have other expressions because different people may have different describing habits. So the target of this step is to group aspect expressions into aspect categories.

Assuming that aspects of expressions belonging to the same category have the same context, it is necessary to compare the context of words and phrases. A Word2Vec model is trained to represent all the unigrams and bigrams as vectors; thus, it is possible to compare the context between them. These vectors can be seen as a description of the context of each element in the vocabulary. Skip-gram model is used to get the input for the word embedding representations. Through skip-gram, all the sentences in all the comments are organized into sequences. The length of the context window is 3, and each gram concerns the 2 grams surrounded as neighbors.

Since all the words and phrases are in a numeric form, the cosine similarity (refer to equation 3)distance between them can be calculated to see the difference in their contexts. With the word embedding model, for each aspect category extracted in the last step, I take the top 10 words or phrases closest to it, regarded as different aspect expressions. But some expressions may not be nouns or noun phrases, so these are removed from the category.

➢ *Equation 3: Cosine Similarity*

$$cos(\theta) = \frac{A\,.\,B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_I^2}}$$



Text Cleaning

- Tokenize sentences and lemmatize words.
- Retain stop words.

Aspect Extraction

- Identify frequent noun phrases as aspects.
- Use Pointwise Mutual Information (PMI) scores for ranking.
- Select top 50 bi-grams with the highest PMI scores.
- Calculate TF-IDF for selected aspects.

Aspect Categorization

- Group aspect expressions into categories.
- Train Word2Vec model to represent expressions as vectors.
- Calculate cosine similarity between expressions.
- Select top 10 similar words or phrases for each aspect category.

Aspect Sentiment Classification

- Use sentiment orientation matrix.
- Apply unsupervised learning to identify sentiment clusters.
- Utilize Bag-of-Word model with TF-IDF.
- Group comments into positive and negative sentiment clusters using K-Means.
- Manually validate sentiment labels.

Text Classification

- Utilize Random Forest and Logistic Regression for classification based on sentiment orientation matrix.

Fig 1: Methodology

D. Aspect Sentiment Classification

In this step, we use a sentiment orientation matrix to understand public opinions. Each row represents a comment, and each column represents an aspect. The values in the matrix range from -1 to 1, indicating whether the sentiment is negative, neutral, or positive, respectively. Since we don't have polarity labels, we use unsupervised learning. We identify opinion sentences in comments and assign them a sentiment score. If a comment doesn't have an opinion on an aspect, we label it as neutral. Then, we collect comments related to each aspect and use K-Means to group them into positive and negative sentiment clusters.

We use the Bag-of-Word model to represent comments and apply the TF-IDF approach. For each cluster, we find the two comments closest to its center, labeling them as positive and negative. Then, we manually check these comments to ensure accuracy. Finally, we assign sentiment labels to all comments in the cluster based on whether they align with the positive or negative center comments. This method helps us categorize comments as either positive or negative sentiment for each aspect.

➢ *Equation 4: Comparison Parameters*

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$Specificity = \frac{TN}{TN + FP}$$

E. Classification of Text

The end step is to create the connection between the sentiment orientation matrix and the editor's selection by classification. This step uses Random Forest and Logistic Regression as tools.

IV. EXPERIMENT RESULTS

After implementing the outlined procedures, we have successfully categorized aspects into three categories along with their respective expressions, detailed in Table 1. The sentiment orientation distribution across different aspects and the distribution of target variables are depicted in Figure 2.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY214

Table 1: Aspect Category & it's Aspect Expression

S.No.	Aspect Category	Aspect Expression
1	staff	staff be, all staff, bad, behavior, customer, behave, area
2	station	pump, road, place, petrol, fuel station, big, station with
3	card	part, cash, debit card, credit card, upi, accept, credit, payment, card accept
4	place	road, visit, location, this place, busy, all time
5	all facility	available, facility, toilet, water, drinking water, washroom, clean, free air, air
6	work	machine, fill, guy, put, person, cheat, ask, wait
7	quality	quality, end quality, good quality, accurate, petrol, product, petroleum, pure
8	oil	corporation, outlet, operate, engine, company, petroleum, leg, product, city
9	diesel	gas, petrol diesel, provide, premium, good, leg, pure, station with
10	fill	put, tank, bike, refill, person, puncture, wait, guy, clear

Comments lacking opinion sentences are excluded from analysis, resulting in a total of 26,717 comments for classification tasks. Among these, some are not chosen, leaving 25,907 comments selected for further evaluation. To address potential bias in labels, Adaptive Synthetic Sampling (ADASYN) is employed to achieve a more balanced dataset, thereby mitigating issues related to overfitting and suboptimal performance on True Positive. Following adjustment, the dataset comprises 25,907 unselected and 26,092 selected comments.

Table 2: Comparison of Results

Model	Description	CV Score	Test Score
A	Random Forest (Num)	0.750	0.747
B	Logistic Regression (Num)	0.562	0.561
C	Random Forest (Categorical)	0.676	0.672
D	Logistic Regression (Categorical)	0.592	0.600

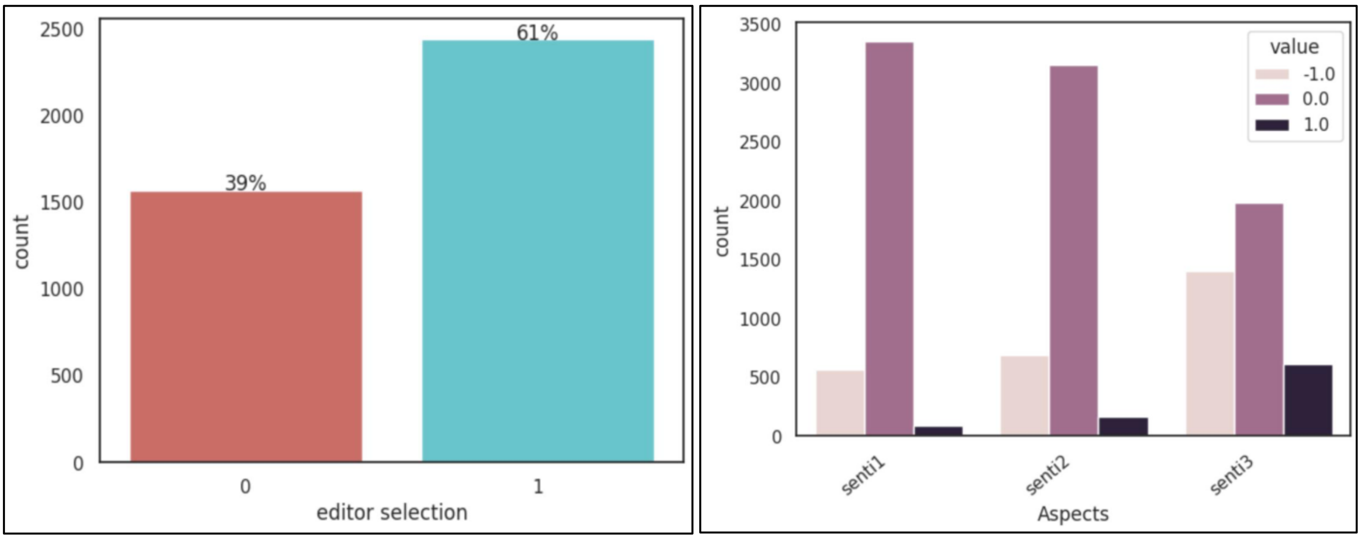


Fig 2: Dimension of Input & Target Variables

Figure 3 illustrates the output of each model, as outlined in Table 2, displaying a graphical representation of the confusion matrix for each model. Additionally, Figure 4 depicts the Receiver Operating Characteristic (ROC) curves for a visual assessment of each model's performance.

The classification process employs Random Forest and Logistic Regression on numeric and categorical inputs, respectively. For categorical input, sentiment orientation values are transformed into dummy variables, with 0 serving as the baseline. Evaluation utilizes a 10-fold Cross Validation strategy.

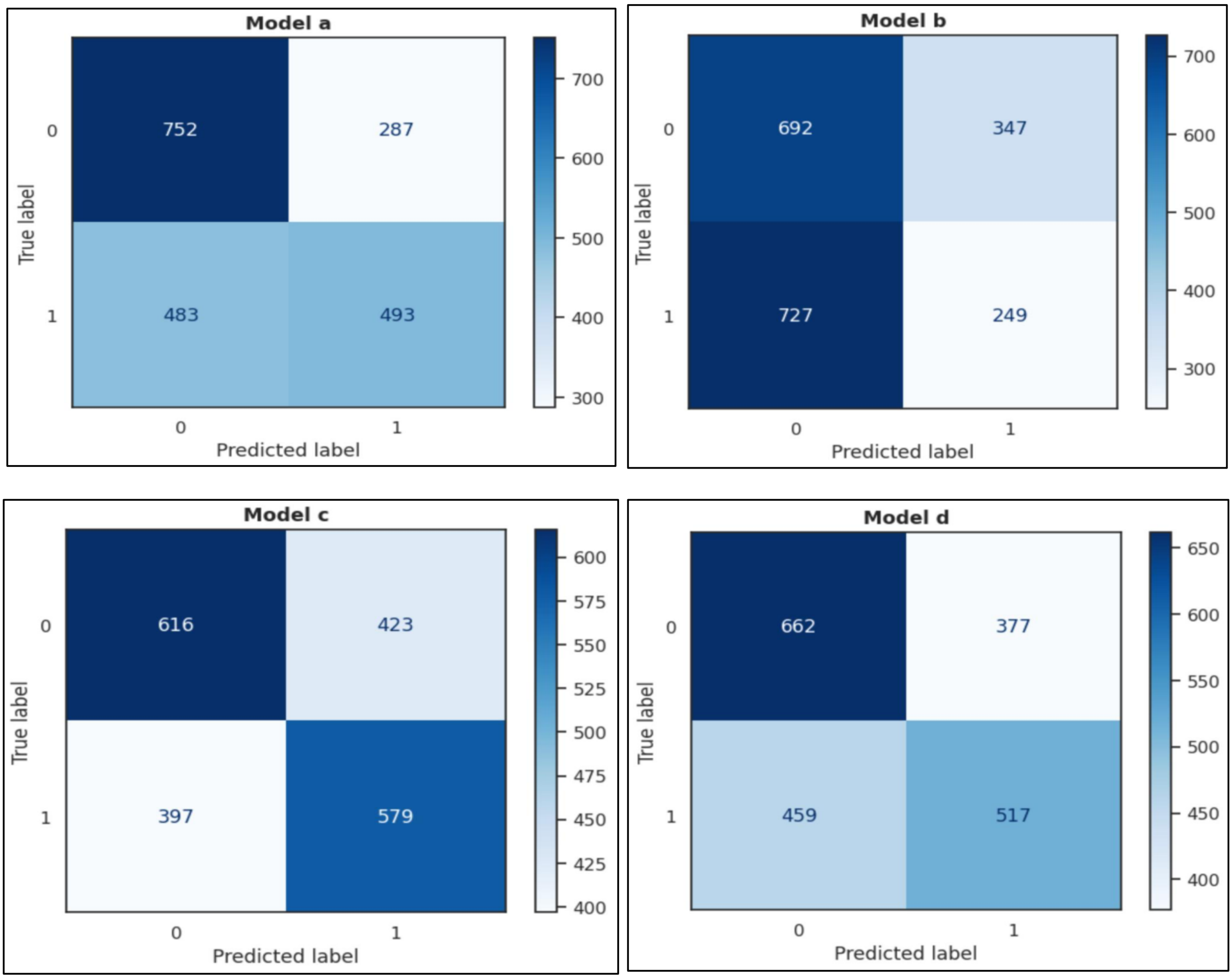


Fig 3: Conclusion Matrix

Table 3 presents the classification report of four classifier models. It's evident that Random Forest with numeric variables achieves the highest accuracy and demonstrates good precision and recall. Conversely, Logistic Regression's performance is subpar, marginally surpassing random classification. However, classifiers utilizing categorical input exhibit high true positives.

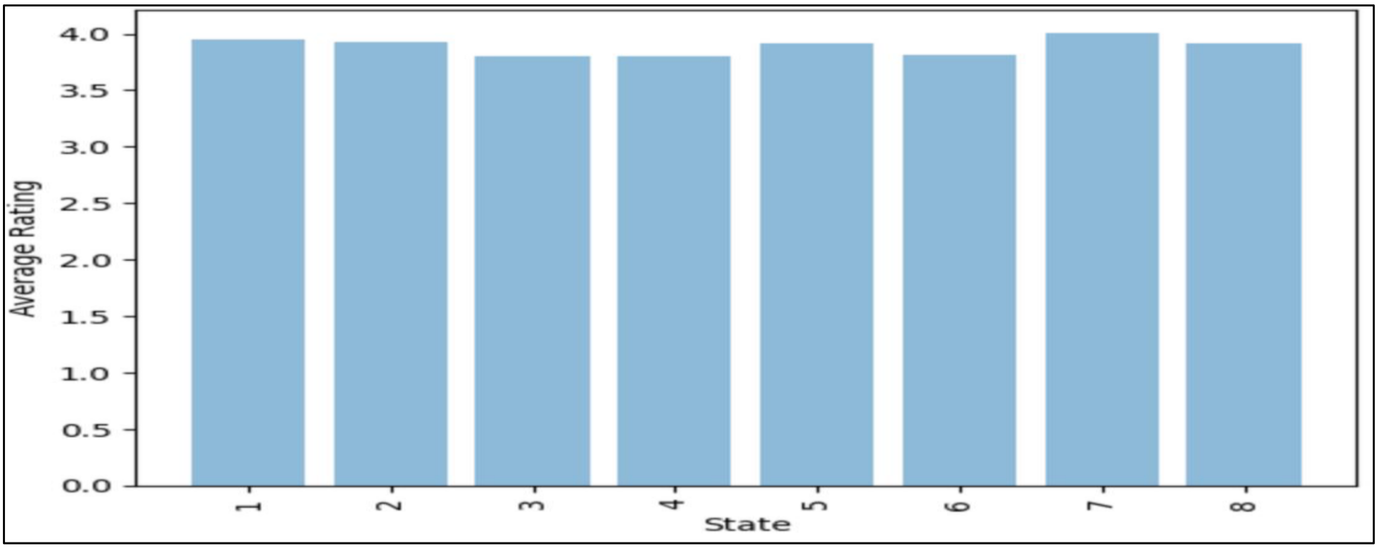


Fig 4: Average Rating by State

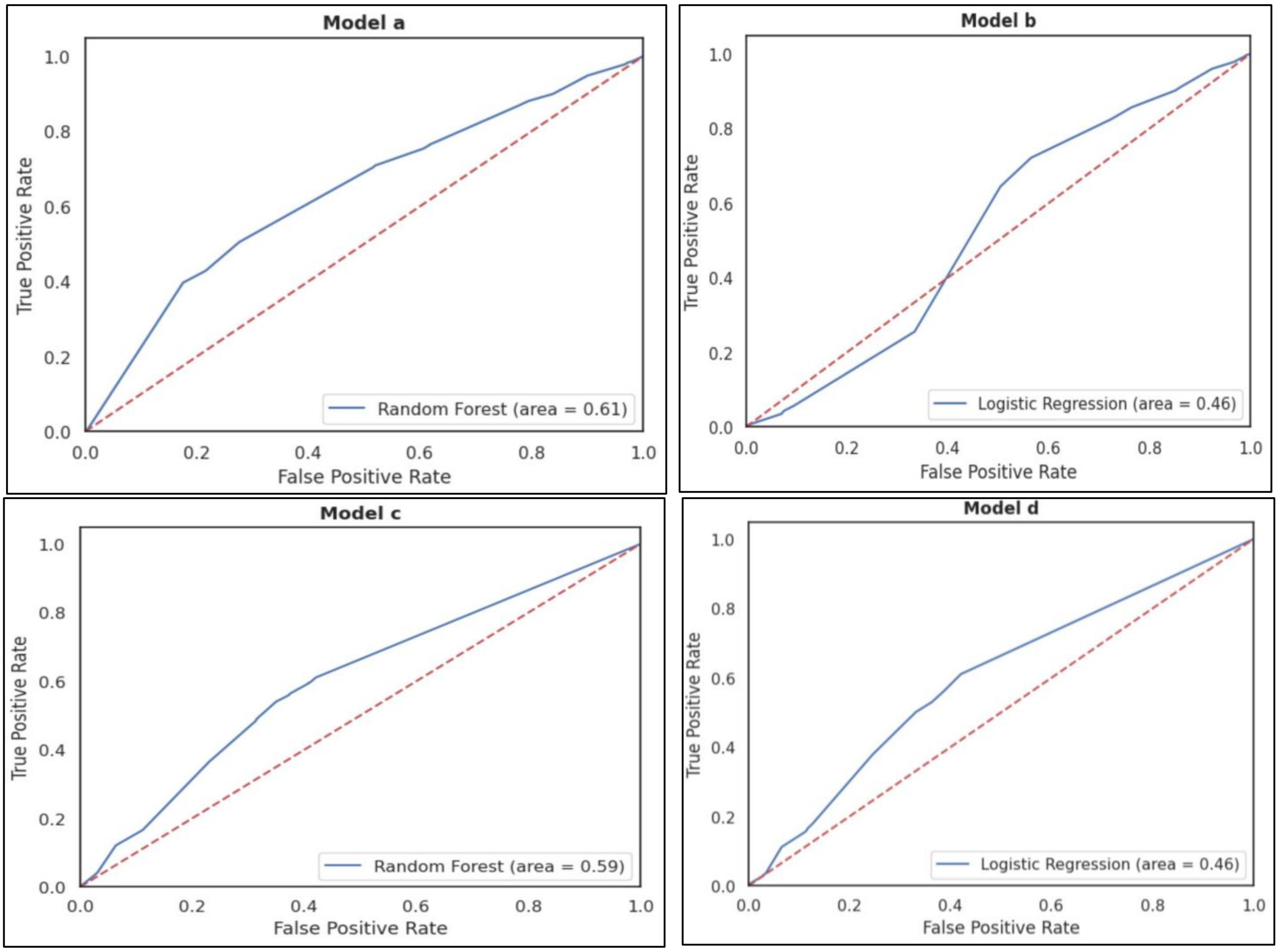


Fig 5: Receiver Operating Characteristic

Based on the results obtained from the Random Forest model with numeric variables, the Accuracy, Precision, Recall, and F1-Score (as per equation 4) are all relatively high, exceeding 0.6. This suggests a correlation between public sentiment orientation on various aspects and editors' selection decisions, indicating the influence of extracted aspects on editorial choices. Notably, the majority of errors stem from False Negatives, highlighting a potential area for improvement.

Conversely, classifiers utilizing categorical input demonstrate strong performance in minimizing False Negatives, indicating robust clustering despite occasional polarity misclassifications. This underscores the accuracy of the extracted aspects. However, it's plausible that certain significant aspects remain unextracted, contributing to the observed discrepancies.

In summary, the experimental findings clearly indicate that editorial selections reflect political opinions, affirming the existence of a relationship between public and editorial sentiments. To enhance results and explore this relationship further, several avenues for future research are suggested. Firstly, expanding the range of sentiment orientation labels and utilizing alternative information retrieval tools could enhance polarity precision, particularly in identifying nuances such as sarcasm and comparisons. Additionally, implementing multi-classification approaches could offer a more nuanced understanding of public opinions.

Secondly, considering implicit aspects not captured in the current analysis could leverage additional comments, possibly through topic modeling or sentiment word detection. Thirdly, exploring editors' sentiment orientation towards different aspects is crucial for understanding their selection criteria. Lastly, leveraging alternative classifiers or ensemble methods could improve the overall classification models for this complex problem.

V. CONCLUSION

In conclusion, Sentim IOCL represents a pivotal advancement in leveraging natural language processing (NLP) techniques to decode and harness customer sentiment within the petroleum industry, with a focal point on Indian Oil Corporation Limited (IOCL). Through rigorous sentiment analysis of varied customer feedback sources, ranging from social media interactions to surveys, the project has adeptly unveiled pivotal insights into customer preferences, concerns, and expectations. These findings furnish actionable intelligence for IOCL, empowering the company to refine service quality, streamline operations, and fortify brand resonance.

Looking ahead, the success of Sentim IOCL not only bolsters IOCL's strategic decision-making capabilities but also sets the stage for continued innovation and exploration in the realm of customer sentiment analysis. The scalable and adaptable nature of the sentiment analysis framework opens avenues for its application across diverse industries, promising to usher in a new era of data-driven customer-centricity. By embracing this data-driven paradigm, IOCL can cement its position as a market leader, fostering enduring customer loyalty and sustained business growth in an ever-evolving landscape.

Furthermore, Sentim IOCL represents a significant advancement in utilizing natural language processing (NLP) techniques to decode customer sentiment in the petroleum industry, focusing on Indian Oil Corporation Limited (IOCL). By analyzing diverse customer feedback sources, the project has revealed crucial insights into customer preferences and concerns, empowering IOCL to enhance service quality and brand resonance. Looking ahead, Sentim IOCL's success paves the way for further innovation in customer sentiment analysis, offering a scalable framework for data-driven decision-making across industries. Its impact underscores the transformative potential of data analytics in driving business growth and customer satisfaction.

APPLICATIONS

The Sentim IOCL holds promise for various applications within Indian Oil Corporation Limited (IOCL) and the broader petroleum industry:

A. Customer Satisfaction Enhancement:

By analyzing customer sentiment expressed in reviews and feedback, IOCL can identify areas for improvement in its services, facilities, and customer interactions. Insights gained from sentiment analysis can inform targeted strategies to enhance customer satisfaction and loyalty, ultimately improving brand perception and competitiveness.

B. Operational Efficiency Optimization:

Sentiment analysis can help IOCL optimize its operational processes by pinpointing inefficiencies or pain points highlighted by customers. For example, identifying recurring complaints about long wait times at petrol pumps could prompt IOCL to streamline operations, adjust staffing levels, or implement technology solutions to improve service efficiency.

C. Product and Service Innovation:

Understanding customer sentiments towards existing products and services can guide IOCL in developing new offerings or refining existing ones to better meet customer needs and preferences. Sentiment analysis can uncover emerging trends, customer preferences, and unmet needs, informing product development initiatives and marketing strategies.

D. Brand Reputation Management:

Monitoring sentiment towards the IOCL brand across various online platforms and media channels allows for proactive reputation management. Detecting negative sentiment early enables IOCL to address issues promptly, respond to customer concerns, and mitigate potential reputational damage. Conversely, identifying positive sentiment provides opportunities to amplify positive brand experiences and cultivate brand advocates.

E. Competitive Benchmarking:

Comparing sentiment analysis results with those of competitors provides valuable insights into IOCL's competitive positioning and areas of differentiation. Understanding how IOCL's customer sentiment compares to that of rival petroleum companies can inform strategic decision-making and help IOCL stay ahead in the market.

F. Employee Engagement and Satisfaction:

Sentiment analysis can extend beyond customer feedback to include internal communications and employee sentiment surveys. Analyzing employee feedback enables IOCL to gauge employee satisfaction, identify concerns, and address issues affecting morale and productivity. Engaged and satisfied employees are more likely to deliver excellent customer service, contributing to overall business success.

These applications demonstrate the diverse ways in which sentiment analysis can be leveraged to drive improvements and innovation across various aspects of IOCL's operations, ultimately leading to enhanced customer experiences, operational efficiency, and competitive advantage.

REFERENCES

[1]. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.

[2]. Liu, B. (2015). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 8(1), 1-167.

[3]. Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. Proceedings of the 21st International Conference on World Wide Web, 191-200.

[4]. Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. Proceedings of the 20th International Conference on Computational Linguistics, 1367-1373.

[5]. Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, 347-354.

[6]. Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine, 9(2), 48-57.

[7]. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches, and applications. Knowledge-Based Systems, 89, 14-46.

[8]. Lu, Y., & Zhai, C. (2009). Multi-faceted opinion analysis in text: Model and method. Proceedings of the 18th International Conference on World Wide Web, 911-920.

[9]. Mukherjee, A., Venkataraman, R., Liu, B., & Glance, N. (2013). What Yelp fake review filter might be doing? Proceedings of the 22nd International Conference on World Wide Web, 535-536.