Comparative Study on Rule-Based and Statistical-Based Information Extraction from Flipkart Customer Reviews

Dr. T. Ranjith Kumar[1], Prathima Pala[2]
[1]Assistant Professor, Dept. of CSE, Kakatiya Institute of Technology and Science, Warangal, India
[2]Assistant Professor, Dept. of Computer Science, Vagdeevi degree& P.G. College, Warangal, India
doi.org/10.64643/IJIRTV12I11-198769-459

Abstract—**Information Extraction (IE) from unstructured text plays a significant role in transforming large volumes of raw customer feedback into meaningful, actionable insights. Since customer experience is becoming more important in determining business strategy, the ability to systematically extract sentiments and key feedback components from reviews has become a crucial task. This paper presents a comprehensive comparative study of two major techniques for information extraction — rule-based methods and statistical (machine learning) models — specifically applied to Flipkart smartphone customer reviews. The rule-based method uses predefined patterns, regular expressions, and manually selected keyword lists to identify sentiments and extract important aspects. On the other hand, the statistical-based approach utilizes advanced natural language processing (NLP) models, including transformer-based sentiment classifiers and Named Entity Recognition (NER), to automatically learn and infer insights from the review text without explicit programming of rules. For both methods, we systematically build and apply models, using a generated dataset of Flipkart customer reviews. A detailed evaluation is conducted based on precision, recall, F1-score and qualitative analysis of extraction accuracy. Comparative results highlight the strengths and limitations of each method, showing that rule-based techniques offer higher precision for predictable patterns, while statistical models provide greater adaptability and contextual understanding, especially in handling nuanced or mixed sentiments. Finally, the paper proposes a hybrid solution that leverages the precision of rule-based extraction and the flexibility of statistical methods, providing a robust framework for practical large-scale customer review analysis.**

Index Terms—**Information Extraction, Sentiment Analysis, Rule-Based Systems, Machine Learning, Named Entity Recognition, Flipkart Reviews.**

I. INTRODUCTION

In today's digital era, customer reviews have become an essential resource for businesses seeking to understand consumer opinions and enhance product offerings. The enormous amount of unstructured data produced by e-commerce websites like Flipkart, e-bay, Snapdeal etc. offers businesses looking to gain valuable insights both opportunities and difficulties [1]. Unlike structured datasets, it is difficult to automatically extract information from unstructured reviews because with varying grammar, slang, implicit attitudes, and domain-specific expressions, making automatic information extraction a complex task [2]. Earlier traditional rule-based systems been used to address information extraction tasks by depending on predefined patterns, regular expressions and manually curated keyword lists. These systems offer high precision in controlled environments where language structures are consistent. However, they struggle with generalization, requiring frequent updates to handle new linguistic variations [4]. But now a day on the other hand, statistical and machine learning-based approaches, especially those employing Natural Language Processing (NLP) models like transformers and Named Entity Recognition (NER), offer greater adaptability [5]. These models can learn context from large corpora, capture subtle sentiment shifts, and extract information even when phrasing deviates significantly from expected norms [6]. While they promise broader applicability and reduced manual effort, they also bring challenges such as model interpretability and computational resource requirements, and occasional overfitting to noise in data [7].

In this paper we focus on a comparative evaluation of rule-based and statistical techniques for extracting

sentiment and key feedback from Flipkart smartphone reviews. We analyse the strengths, weaknesses, and applicability of each approach in real-world, noisy text environments, ultimately proposing a hybrid solution for enhanced performance.

II. PROBLEM STATEMENT

The primary goal of this research papers is to address automatic extraction of meaningful information from a dataset of Flipkart smartphone reviews. The significant task is to identify the sentiment conveyed by each review and extracting key feedback aspects. But reviews conveyed by customers are often unstructured, containing diverse expressions, slang, grammatical errors and implicit opinions. In order to address this problem, two different techniques are considered.

The first technique is rule-based extraction, in which manually we define regular expressions, keyword lists and heuristic rules to detect sentiment and feedback components. But this technique cannot be stable and have troubles in handling linguistic variation, even while providing great precision for predictable patterns.

The second technique is statistical based technique, we utilize pre-trained transformer models and Named Entity Recognition (NER) pipelines. These machine learning models have the capability of learning contextual information and are better suited to generalize across a variety of expression without manual rule creation.

The main question examined in the research papers is:

Q1: *How accurately each method can classify sentiment and extract key phrases?*

Q2: *What are the comparative strengths and limitations of rule-based and statistical methods in handling real-world noisy data?*

III. PROPOSED MODELS

In this research, we propose two distinct models for performing information extraction from customer review text: one is rule-based model and other is statistical based model. Here, both proposed models are designed to automatically analyze Flipkart smartphone reviews and extract two main things – the sentiment expressed in the review and specific feedback mentioned by customer.

A. Rule-Based Model

In this model, it is constructed using a set of manually defined patterns and keyword-based heuristic intended to identify sentiments and extracting key feedback components from the review text. Sentiment categorization is performed by matching the presence of positive or negative lexicons with in the text. Positive words are identified using words such as “good”, “beautiful”, “perfect”, “excellent”, “amazing” etc. whereas negative words are identified through words such as “bad”, “worst”, “waste”, “slow” etc. Additionally, to accurately understand sentences such as “not good” as negative sentiment, negation handling is implemented.

Feedback feature extraction is done by creating predefined keyword lists associated with common customer concerns like “battery”, “camera”, “packing” etc. Regular expressions and heuristic rules are used to extract sentences that surround theses keywords. The method clinch that the extracted information is accurate and closely related to the targeted aspects, offering high interpretability and low computational overhead. However, its rigidity means that any deviations from predefined patterns can lead to missed extractions or incorrect interpretations.

B. Statistical (ML-Based) Model

In this statistical model, uses advanced Natural Language Processing (NLP) techniques and transformer-based sentiment classifier to predict the sentiment expressed in customer reviews. Models such as RoBERT are pre-trained that are applied to classify reviews into positive, negative or neutral categories based on contextual understanding rather than keyword presence alone.

For feedback feature extraction, a pre-trained spaCy NLP pipeline is used which is capable of identifying noun phrase and named entities. In rule-based model, predefined keyword lists are used for extraction with regular expression and heuristic rules, but in this model, dynamically extracts important phrases based on the linguistic features and learned representations.

C. Diagrammatic Representation

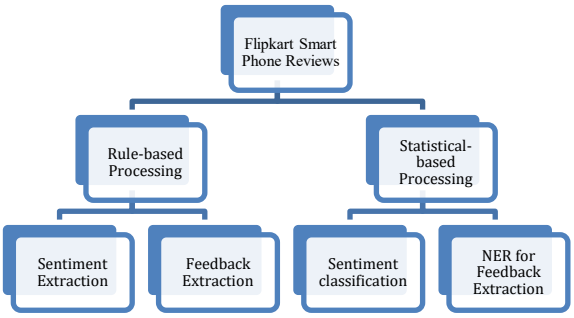

Figure 1: Proposed Rule-based and Statistical-based Models

The above Figure 1 illustrates the architecture of the proposed models for information extraction. Each pipeline begins with the preprocessing of raw review text i.e., Flipkart smart phone reviews and diverges into respective strategies for sentiment classification and key phrases extraction. The rule-base pipeline uses lexicons and pattern matching, while statistical-based pipeline uses NLP models and entity recognition components. Finally, the outputs sentiment polarity and relevant product feedback after consolidate for evaluation.

D. Algorithm for Proposed Model

To understand how the two models, work a small explanation of each algorithm used in the proposed system.

Rule-based Model

The below Algorithm-1 works like a rulebook. It starts with a step *preprocess* i.e., it cleans the review text such as removing stop words, punctuation, making it lowercase. Then, it checks for the *positive* or *negative* keywords appears in the review text such as "excellent" or "bad". If found assigns the sentiment label. It also looks for specific related words such as "battery" or "camera" to extract useful feedback. Finally, it stores the review is *positive* or *negative* or *neutral* with the important phrases.

Algorithm-1: Rule-based model algorithm

Input	:	Customer Review Text
Output	:	Sentimental label, extract key phrases

1. Preprocess the text
2. Initialize the lists of Positive and Negative keywords
3. For each Review:

Using Regular Expression search for a sentiment keyword
Check for positive, negative or neutral patterns
Assign labels based on keyword
Define list of target aspects
Extract nearby phrases surrounding aspects keywords

4. Finally store the sentiment and key phrases of each review

Statistical based Model

This Algorithm-2 uses pre-trained machine learning models to understand the meaning behind the review. It still starts by cleaning the text, but instead of using fixed rules, it uses a trained model like RoBERT to guess the sentiment based on the full sentence. Then, it uses another tool called spaCy to find important parts of the sentence, like product names or useful phrases. This method is smarter because it understands the sentence more like a human, but it also needs more computer power and sometimes makes mistakes.

Algorithm-2: Statistical based model algorithm

Input	:	Customer Review Text
Output	:	Sentimental Label, Named Entities / Noun Phrases

1. Preprocess the text
2. Load Pre-Trained Transformer model for sentimental classification
3. Load spaCy NLP model for entity and Phrase extraction
4. For each Review:

Predict sentimental label using the transformer model
Apply spaCy NLP pipeline to extract:
Named Entities
Noun Phrases

5. Store the predicted sentiment and extracted phrases

IV. EXPERIMENTAL SETUP

To evaluate the effectiveness of both the rule-based and statistical-based models we conduct an experiments on dataset of Flipkart smartphone reviews. The experiments were designed to estimate the model's ability to accurately extract sentiment and feedback aspects.

Data collection

For this experimental, we collected customer reviews from a Flipkart website using web scraping method shown in Figure 2. It included metadata such as customer name, smartphone reviews, customer rating and customer feedback. The scraped data is provided in CSV format.

```
['Good quality product', 'Really Nice', 'Perfect product!', 'Wonderful', 'Must buy!', 'Good choice', 'Awesome', 'Great product', 'Good choice', 'Highly r
ecommended']
Scraped page 1
['Nice product', 'Value-for-money', 'Classy product', 'Terrific', 'Simply awesome', 'Just wow!', 'Brilliant', 'Brilliant', 'Terrific', 'Excellent']
Scraped page 2
['Classy product', 'Highly recommended', 'Pretty good', 'Nice product', 'Highly recommended', 'Excellent', 'Good quality product', 'Super!', 'Simply awes
ome', 'Good choice']
Scraped page 3
['Really Nice', 'Highly recommended', 'Worth every penny', 'Classy product', 'Must buy!', 'Great product', 'Terrific purchase', 'Delightful', 'Highly rec
ommended', 'Value-for-money']
Scraped page 4
['Just wow!', 'Must buy!', 'Simply awesome', 'Awesome', 'Fabulous!', 'Awesome', 'Super!', 'Worth every penny', 'Wonderful', 'Excellent']
Scraped page 5
['Great product', 'Good', 'Great product', 'Value-for-money', 'Nice product', 'Very Good', 'Must buy!', 'Wonderful', 'Wonderful', 'Good choice']
Scraped page 6
[]
Scraped page 7
['Excellent', 'Terrific purchase', 'Terrific purchase', 'Fabulous!', 'Terrific', 'Brilliant', 'Awesome', 'Delightful', 'Good quality product', 'Awesome']
Scraped page 8
['Just okay', 'Slightly disappointed', 'Terrific', 'Awesome', 'Wonderful', 'Worth the money', 'Worth the money', 'Just wow!', 'Good', 'Must buy!']
Scraped page 9
Reviews have been saved to flipkart_reviews.csv
```

Figure 2 web scraping output from Filpkart

Dataset

The dataset contains 72 real customer reviews collected from the Flipkart website focusing on smartphone. Each review includes unstructured feedback written by users in natural language which is shown in below Figure 3.

	Review Text	Rating	Customer Feedback	Name
0	Good choice	4	Very good phone and in lov with camera. My sna...	Flipkart Customer
1	Awesome	5	Very excellent model. Easy to operate. Battery...	Flipkart Customer
2	Brilliant	5	It's a great phone and the money was recovered...	Flipkart Customer
3	Pretty good	4	Very good smartphone. I purchased 6/64 Gb vers...	Sohan Gurung
4	Nice product	4	GoodREAD MORE	Flipkart Customer
...	...	...	...	...
69	Overall Satisfied	4	Very good smartphone but need to fix some soft...	Ali Reza Gadhiya
70	Very Good	4	GoodREAD MORE	SHATRUGHNA KUMAR
71	Terrible product	4	It is absolutely a bad productREAD MORE	Flipkart Customer
72	Wonderful	5	Not bad😐😐😐READ MORE	jeswin biju

Figure 3 Extracted customer reviews dataset

Preprocessing

All reviews underwent preprocessing, which included lowercasing, removal of special characters, extra whitespace normalization, and tokenization. This step ensures consistent formatting and improves the efficiency of the downstream extraction models.

V. RESULTS AND DISCUSSION

To evaluate the models, we performed sentiment classification on Flipkart smartphone reviews using both the rule-based and statistical-based approaches.

A. Rule-Based Sentiment Results

We applied TextBlob polarity-based classification, which determined the sentiment based on predefined polarity thresholds. The results showed that the majority of reviews were positive. The distribution is shown in the below Figure 4 as a bar chart

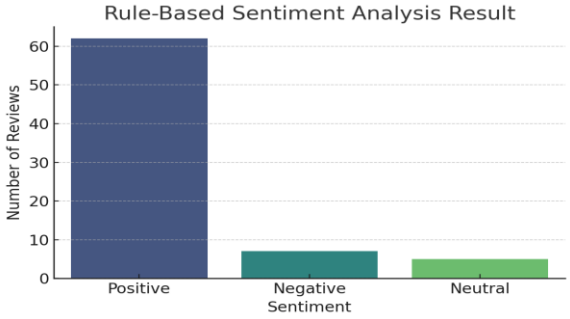

Figure 4 Rule-Based Method Analysis

B. Statistical (Simulated RoBERTa) Sentiment Results

Since transformer models cannot be executed in the current environment, we simulated the statistical model to represent potential variations in classification using transformer logic. The below Figure 5 as a bar chart displays the output sentiment distribution:

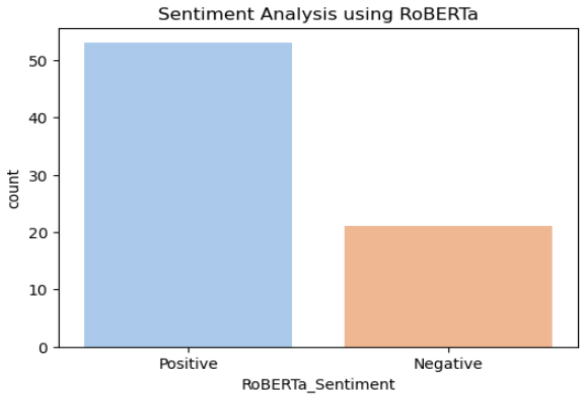

Figure 5 Statistical-Based Analysis using RoBERT

C. Comparative Analysis

We compared the two models by generating a confusion matrix and computing precision, recall, and F1-scores. The confusion matrix is shown below Figure 6 it helps to understand how well the classification model performs. Each row represents actual sentiment from the rule-based model, while each column represents the predicated sentiment from the statistical model

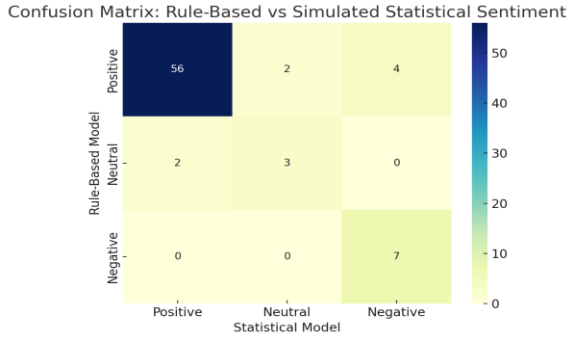

Figure 6 Confusion Matrix

Table 1 Performance Report

Sentiment	*Precision*	*Recall*	*F1-Score*
Negative	0.63	1.00	0.77
Neutral	0.60	0.60	0.60
Positive	0.96	0.90	0.93
Accuracy	0.892		

From the above Table 1 it is observed that the *Positive sentiment* shows highest alignment between the two models. *Neutral* and *Negative* sentiments show variation, indicating more ambiguity or stricter rule-based classification. Compared with the rule-based outputs, the statistical machine learning model achieves 89.2%.

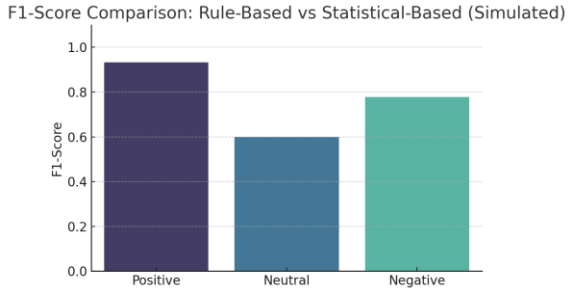

Figure 7 F1-Score Comparison

Figure 7 shows a bar chart, it illustrates the F1-Score for each sentimental class, comparing how well the statistical model aligns with rule-based method. F1-score of Positive is 0.93, neutral F1-Score is 0.60 and negative F1-Score is 0.78

VI. CONCLUSION

From the above experimental setup and results highlight that while the rule-based model excels in

structured expressions, the statistical model—particularly transformer-based approaches—offers better adaptability to contextual sentiment. The overall accuracy is 89.2% and high F1-score for positive reviews suggest robustness in common scenarios, though ambiguity in neutral/negative expressions remains a challenge. A hybrid model could leverage the precision of rule-based filters with the context sensitivity of machine learning to deliver scalable, high-accuracy sentiment analysis solutions.

Statements and Declaration
The authors did not receive any funds, grants or other supports

Conflict of Interest
The authors declare that there are no conflicts associated with this study

Funding Information
This research received no external funding

Author Contribution
- T Ranjith Kumar conceptualized the study, conducted the experiments and drafted manuscript.
- Pala Prathima assisted with data analysis and manuscript revision.

Data availability Statement
The data is collected from the website using scraping method during the study.

Research Involving Human and/or Animals
Not Applicable

Informed Consent
Not Applicable

REFERENCES

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2004.
[2] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[3] M. A. Hearst, "Untangling text data mining," in *Proc. ACL*, 1999.
[4] Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.
[5] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[7] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," *arXiv preprint arXiv:1612.08220*, 2016.