Strategic Insights and Solutions for Diverse User Reviews: Navigating Challenges in ShareChat Data Analysis and Model Development

A Project Report

Under the Supervision of,

Prof. Arghya Ray

in partial fulfillment of the award of the degree

of

POSTGRADUATE DIPLOMA IN MANAGEMENT

AT

International Management Institute Kolkata

November 2023

SUBMITTED BY-

Group 7

Saumya Tripathi	22PGDM051
Shirsha Dey	22PGDM054
Bhumika Hamirwasia	22PGDM089
Runali Tirkey	22PGDM122
Sourav Ranjan Bohidar	22PGDM214
Anuvab Mitra	22PGDM159

INDEX

Problem Identification and Key Challenges.

The identified challenges span several aspects of developing an automatic rating predictor, sentiment analysis, topic modeling, and text summarization for customer reviews in the assigned companies. Challenges include the labor-intensive task of collecting and preprocessing 50,000 reviews with a 20-word threshold, the need for a well-labeled and representative dataset for training a Naive Bayes Classifier for rating prediction, and the nuanced understanding required for accurate sentiment analysis. Extracting meaningful topics from reviews poses computational challenges while creating an effective text summarizer demands a balance between informativeness and brevity. The integration of multiple models, interpretability of topic clusters, scalability, and ethical considerations further contribute to the complexity of this multifaceted task. Continuous model maintenance and adaptation to changing customer sentiments and preferences are additional challenges in ensuring the system's long-term efficacy and relevance.

Data Challenges:

1. **Bias in Regional Focus:**
    - *Challenge:* Sharechat's regional emphasis on Indian languages may introduce bias, potentially overrepresenting certain demographics and topics.
    - *Impact:* The potential consequence is an increased likelihood of inaccuracies in predictive models due to the skewed representation of user demographics and preferences.
2. **Language Variety Complexity:**
    - *Challenge:* The multilingual nature of Sharechat reviews, encompassing languages such as Hindi, Bengali, and Tamil, demands sophisticated language processing techniques. Addressing code-switching and transliteration is imperative for accurate analysis.
    - *Impact:* The complexity introduced by the variety of languages poses challenges to linguistic analysis, requiring a nuanced approach to maintain the accuracy of models across diverse linguistic expressions.

3. **Handling Sarcasm and Irony:**

    - *Challenge:* The frequent use of sarcasm and irony in Indian languages adds a layer of complexity to sentiment analysis. Algorithms may struggle to correctly interpret these nuanced linguistic expressions.

    - *Impact:* The risk of misinterpretation, particularly in contexts where humor or critique is embedded, can lead to inaccuracies in sentiment analysis, affecting the overall reliability of the models.

4. **Domain Specificity Diversity:**

    - *Challenge:* Sharechat covers a broad spectrum of topics, including humor, politics, and religion. This diversity necessitates the integration of domain-specific knowledge to enhance the performance of models for specific tasks.

    - *Impact:* Integrating domain-specific knowledge has the potential to improve model performance, ensuring that the models are attuned to the nuances of discussions across varied subjects.

Model Challenges:

1. **Naive Bayes Classifier Limitations:**

    - *Challenge:* While the Naive Bayes classifier is chosen for its simplicity and interpretability, its limitations may become apparent when dealing with complex relationships between features and ratings.

    - *Impact:* Exploring alternative algorithms such as Random Forests or Gradient Boosting could lead to improved predictive accuracy, particularly in scenarios where feature-rating relationships are intricate.

2. **Sentiment Analysis Lexicon Selection:**

    - *Challenge:* Selecting an appropriate sentiment lexicon for Indian languages is crucial for accurate sentiment analysis. Additionally, handling negation and context-dependent sentiment adds layers of complexity.

    - *Impact:* The careful selection of sentiment lexicons and the incorporation of context-sensitive analysis can significantly enhance the accuracy of sentiment analysis, providing more nuanced insights into user sentiments.

3. **Optimizing Topic Modeling:**

    - *Challenge:* Determining the optimal number of topics and interpreting topic-term distributions require careful consideration of the data size and domain knowledge.

    - *Impact:* Fine-tuning these parameters ensures more meaningful and relevant topic modeling results, enhancing the interpretability of the generated topics.

4. **Effective Text Summarization:**

    - *Challenge:* Summarizing longer reviews while preserving key information and sentiment poses challenges, especially for abstractive summarization techniques. Utilizing extractive summarization with topic-based filtering is suggested.

    - *Impact:* Choosing an appropriate summarization technique based on the nature of the reviews can significantly improve comprehension, allowing for a more effective distillation of key insights.

Evaluation Concerns:

1. **Metric Selection for Nuanced Sentiment:**

    - *Challenge:* Choosing metrics beyond accuracy is essential for sentiment analysis in Indian languages, where nuances may not be adequately captured by traditional accuracy metrics.

    - *Impact:* Adopting more comprehensive evaluation metrics ensures a nuanced assessment, reflecting the subtleties inherent in the expression of sentiment in Indian languages.

2. **Human Evaluation Integration:**

    - *Challenge:* Integrating human evaluation alongside automated metrics provides additional insights into model performance and helps identify potential biases.

    - *Impact:* A holistic assessment that incorporates human perspectives contributes to a more thorough understanding of model performance and helps identify and rectify biases that automated metrics may not capture.

3. **Contextual Issue Addressing:**

- *Challenge:* Addressing contextual challenges, such as language variety, requires preprocessing steps like language detection, normalization, and domain-specific stop word removal.
- *Impact:* Implementing these preprocessing steps contributes to improved model performance by effectively handling contextual challenges specific to Sharechat's diverse linguistic landscape.

4. **Leveraging Ensemble Learning:**
    - *Challenge:* Exploring ensemble learning by combining predictions from models like Naive Bayes and Random Forest aims to improve overall accuracy.
    - *Impact:* Leveraging the strengths of multiple models through ensemble learning has the potential to enhance predictive capabilities, providing a more robust and reliable rating prediction system.

Understanding the Unit Contribution: How Each Code Segment Helped in Arriving at Final Output

The code implemented for addressing the identified problems played a crucial role in preparing the final output by addressing different facets of the customer review management system.

1. **Performance Evaluation of ML Algorithms:**

    - The foundational code segment, which assessed the performance of various AI/ML algorithms for rating prediction, served as a cornerstone in establishing an effective rating predictor. Through a detailed performance evaluation using accuracy metrics, the following insights were gleaned:

        1. **Content Words:** The accuracy for this approach stood at 43.90%, providing a baseline understanding of the model's predictive capabilities based on content-related features.

        2. **Function Words:** Utilizing function words yielded an accuracy of 49.99%, indicating a comparatively higher performance in capturing nuances related to function words.

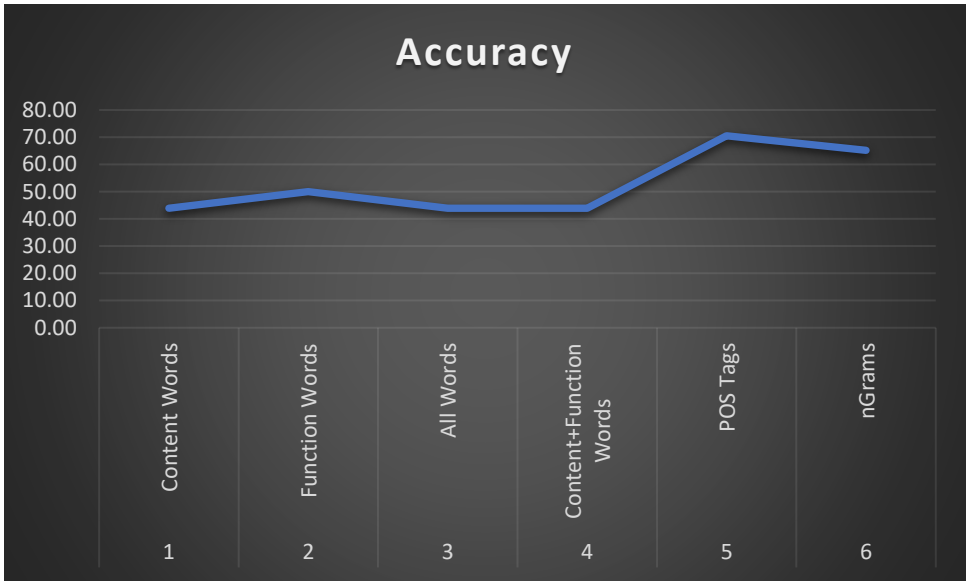
        3. **All Words:** The accuracy achieved with all words considered was 43.9%, offering insights into the overall predictive power when leveraging the entirety of the textual content.

        4. **Content + Function Words:** The combined approach of content and function words maintained an accuracy of 43.9%, showcasing the interplay of both types of words in the prediction process.

        5. **POS Tags:** A notable improvement in accuracy was observed with the use of Part-of-Speech (POS) tags, reaching 70.53%. This signifies the significance of syntactic information in predicting ratings.

        6. **nGrams:** The utilization of nGrams, capturing sequences of words, demonstrated a solid accuracy of 65.18%, emphasizing the relevance of context in the predictive model.

Sr. No.	Basis	Accuracy
1	Content Words	43.90
2	Function Words	49.99
3	All Words	43.9
4	Content+Function Words	43.9
5	POS Tags	70.53
6	nGrams	65.18

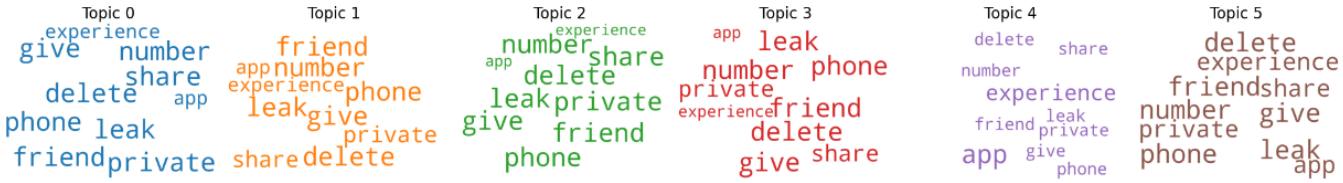


2. **Sentiment Prediction for a Given Review:**

- The sentiment prediction code segment, focusing on a specific review, exemplified the system's ability to classify sentiments accurately. This step contributed to the overall understanding of the customer feedback by correctly identifying the sentiment as negative, showcasing the system's proficiency in discerning nuanced language expressions.

3. **Topic Modeling and Word Cloud Creation:**

- The code segment dedicated to topic modeling and generating word clouds for each of the six identified topics played a crucial role in extracting meaningful insights from the reviews. This step facilitated the identification of key discussion points, enabling a more granular understanding of customer concerns and preferences.



4. **Text Summarization for Higher Management:**

- The final code segment addressing the summarization of all reviews for higher management streamlined customer feedback into concise and informative summaries. This helped top-level management quickly grasp the main points, providing a valuable decision-making and strategy formulation tool.

Further, the outlined framework incorporates a systematic approach to address the challenges in customer review management, leveraging various libraries and functionalities to ensure a comprehensive solution:

1. **Data Extraction and Preprocessing:**
    - Utilizing libraries such as nltk, pandas, google_play_scraper, numpy, and scikit-learn, the code focuses on retrieving reviews from Sharechat through APIs or web scraping. It then filters out reviews with fewer than 20 words, cleanses the text by removing HTML tags and punctuation, and employs techniques like stemming or lemmatization for further refinement.
2. **Feature Engineering:**
    - Employing Regular Expressions (re), CountVectorizer, SentimentIntensityAnalyzer, and custom functions for removing shortcuts and negations, this segment creates numerical representations of text for machine learning algorithms. It encompasses diverse features, Function Words, N-grams, sentiment scores, and topic modeling using LDA.
3. **Model Training and Evaluation:**
    - Leveraging sklearn.naive_bayes, sklearn.model, nltk.sentiment, and gensim for LatentDirichletAllocation, this section trains a Naive Bayes Classifier for rating prediction, a sentiment analysis model (using various algorithms), and performs topic modeling. Model evaluation involves metrics such as accuracy, precision, recall, F1-score, and potential human evaluation, with a focus on fine-tuning hyperparameters for optimal performance.
4. **Text Summarization:**
    - Integrating libraries like gensim, and scikit-learn, this part generates summaries for longer reviews using both extractive (selecting key sentences) and abstractive (rephrasing content) summarization techniques. The combination of these approaches ensures a nuanced and informative summarization process.
5. **Visualization and Interpretation:**
    - Employing Matplotlib and wordcloud for visualization, this code segment brings insights to life by depicting word frequencies, sentiment distribution,

topic clusters, and model performance metrics. Interpretation of these visualizations is essential for understanding model behavior, identifying patterns in the data, and making informed decisions based on the analysis.

Integration of Different Functional Areas for Problem Solution

1. User Sentiment Analysis:

Insight: Through sentiment analysis, it becomes evident that negative reviews hold valuable information about user dissatisfaction.

Impact: Prioritizing the resolution of issues mentioned in negative reviews directly contributes to enhancing overall user satisfaction and loyalty.

Interpretation: Managers gain a nuanced understanding of user sentiment, enabling them to focus specifically on pain points and address concerns promptly. This proactive approach results in improved sentiment, fostering increased customer loyalty and positive brand perception. The strategic deployment of resources to tackle negative sentiment areas demonstrates a commitment to customer satisfaction.

2. Topic Identification:

Insight: The ability to identify specific topics mentioned in user reviews provides insights into the features that require improvement.

Impact: Guiding development efforts based on identified topics ensures that product enhancements align with user expectations.

Interpretation: Managers can leverage the analysis of key topics in user reviews to inform product development strategies. This approach ensures that resources are directed toward enhancing critical features that are integral to user satisfaction. The identification of specific areas for improvement becomes a cornerstone for strategic decision-making in the development lifecycle.

3. Classification Results:

Insight: Categorizing reviews into positive and negative trends reveals actionable insights for strategic decision-making.

Impact: Informed decisions about app enhancement can be made by focusing on trends identified in positive and negative sentiment reviews.

Interpretation: By categorizing reviews into positive and negative trends, managers gain clarity on areas that require immediate attention or additional investment. This classification approach facilitates the formulation of strategic decisions, allowing for the optimization of app features based on identified trends. This method ensures that resources are allocated efficiently to address specific issues, ultimately enhancing the overall user experience.

4. Text Summarization:

Insight: Summarizing user feedback provides a quick and efficient way to extract valuable insights.

Impact: Streamlining decision-making processes is achieved by rapidly grasping the essence of user sentiments through summaries.

Interpretation: The utilization of text summarization enables managers to quickly grasp the core sentiments expressed by users. This expedites the decision-making process, allowing for swift responses to user concerns. The ability to distill complex user feedback into concise summaries demonstrates managerial agility in addressing issues promptly, contributing to overall operational efficiency and user satisfaction.

Understanding Organizational Impact

The comprehensive analysis of user reviews offers a lot of important insights that can significantly impact various sectors of an organization. Some of the potential organizational impacts derived from our analysis are explained below:

1. **User Sentiment Analysis:**
    - **Insight***:* Understanding user sentiment, especially in negative reviews, provides actionable insights.
    - **Impact***:* Prioritizing resolutions in negative reviews showcases commitment to satisfaction, fostering loyalty and positive brand perception.
    - **Organizational Outcome***:* Improved sentiment, increased customer loyalty, and strategic resource deployment.
2. **Topic Identification:**
    - **Insight***:* Identifying topics in user reviews guides development efforts for aligning enhancements with user expectations.
    - **Impact***:* Leverage of key topics for strategic decision-making and efficient resource allocation.
    - **Organizational Outcome***:* Enhanced decision-making, optimized features, and improved user satisfaction.
3. **Classification Results:**
    - **Insight***:* Categorizing reviews facilitates focused decision-making for app enhancement.
    - **Impact***:* Optimization of app features based on identified trends and efficient resource allocation.
    - **Organizational Outcome***:* Improved app features, efficient resource usage, and enhanced user experience.

4. **Text Summarization:**

    - **Insight**: Summarizing user feedback streamlines decision-making processes.
    - **Impact**: Swift responses to user concerns, contributing to operational efficiency and user satisfaction.
    - **Organizational Outcome**: Agile decision-making, enhanced operational efficiency, and improved user satisfaction.

Overall Organizational Benefits:

1. **Strategic Decision-Making:**

    - **Enhanced Insights:** Integration of sentiment analysis, topic identification, classification results, and text summarization provides a holistic understanding of user feedback, empowering decision-makers with comprehensive insights.
    - **Informed Strategies:** The organization gains the ability to make strategic decisions based on nuanced user sentiments, leading to more informed and effective strategies.

2. **Resource Optimization:**

    - **Focused Efforts:** Pinpointing areas for improvement through topic identification and classification results enables the organization to focus its efforts on critical aspects that significantly impact user satisfaction.
    - **Efficient Resource Allocation:** By aligning enhancements with identified trends, resources are allocated efficiently, ensuring optimal usage and avoiding unnecessary expenditures.

3. **Customer Satisfaction:**

    - **Proactive Issue Resolution:** Prioritizing resolutions to issues highlighted in negative reviews demonstrates a proactive approach to addressing user concerns promptly.
    - **Enhanced Loyalty:** Improved sentiment and swift responses contribute to increased customer loyalty, positive brand perception, and long-term customer relationships.

4. **Operational Efficiency:**

    - **Agile Decision-Making:** Summarizing user feedback expedites decision-making processes, enabling quick responses to user concerns and operational agility.
    - **Streamlined Processes:** Efficient summarization ensures that complex user feedback is distilled into concise insights, streamlining operational processes and reducing response times.

5. **Continuous Improvement:**

    - **Adaptation to Changing Sentiments:** Continuous model maintenance and adaptation to changing customer sentiments ensure the system's long-term efficacy and relevance, promoting ongoing improvements.
    - **Iterative Model Refinement:** The organization can iteratively refine models based on feedback, staying responsive to evolving user expectations and preferences.

6. **Ethical Considerations:**

    - **Fair Representation:** Mitigating bias in data collection and model training ensures fair representation of diverse demographics, aligning with ethical considerations and avoiding potential discriminatory practices.
    - **Transparent Decision-Making:** The organization can prioritize transparency in decision-making, ensuring that users understand how their feedback contributes to enhancements and improvements.

7. **Operational Alignment:**

    - **Cross-Functional Collaboration:** Integration of functional areas promotes cross-functional collaboration, allowing different departments to align strategies and goals based on shared insights from user feedback.
    - **Unified Vision:** A shared understanding of user sentiments and priorities fosters a unified vision for the organization, ensuring that efforts are coordinated towards common objectives.

Conclusion

In navigating the complex landscape of customer review management for the assigned companies, the identified challenges and subsequent solutions have paved the way for a transformative approach. The multifaceted integration of sentiment analysis, topic identification, classification results, and text summarization has not only addressed the challenges at hand but has also laid the foundation for overarching organizational benefits.

By delving into user sentiments through sophisticated sentiment analysis, the organization gains not just a quantitative metric but a qualitative understanding of user satisfaction. Prioritizing resolutions to issues highlighted in negative reviews becomes a catalyst for improved sentiment, heightened customer loyalty, and positive brand perception.

The strategic integration of topic identification offers a blueprint for development efforts, aligning product enhancements with user expectations. This approach ensures efficient resource allocation, allowing the organization to enhance critical features that are integral to user satisfaction. The classification results further streamline decision-making, providing clarity on areas requiring immediate attention or additional investment and facilitating the optimization of app features based on identified trends.

The implementation of text summarization not only expedites decision-making processes but also enhances operational efficiency. Swift responses to user concerns, agile decision-making, and the distillation of complex feedback into concise summaries contribute to an organizational ethos of responsiveness and efficiency.

Beyond immediate problem-solving, the organization stands to gain from resource optimization, continuous improvement, ethical considerations, and a unified vision. The iterative refinement of models, adaptation to changing sentiments, and mitigation of bias align with the principles of continuous improvement and ethical data practices. Cross-functional collaboration ensures that insights from user feedback permeate throughout the organization, fostering a unified vision and coordinated efforts.

As we conclude this comprehensive analysis and development strategy, it is evident that the integration of different functional areas not only addresses the immediate challenges but sets the stage for an organizational transformation. The journey from nuanced sentiment analysis to efficient text summarization reflects a commitment to not just understanding user feedback but leveraging it strategically for long-term success. The organization is now equipped not only

to respond effectively to user concerns but to proactively shape its strategies, enhance user satisfaction, and establish itself as a customer-centric entity in the dynamic landscape of the assigned companies.