National College of Ireland

Customer Churn Prediction using RAG-Based Sentiment Analysis with LLMs and CatBoost

MSc Research Project
Data Analytics

Sakshi Kacheshwar KalungePatil
Student ID: x23122366

School of Computing
National College of Ireland

Supervisor: Prof. Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sakshi KalungePatil
Student ID:	x23122366
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Prof. Rejwanul Haque
Submission Due Date:	11/08/2025
Project Title:	Customer Churn Prediction using RAG-Based Sentiment Analysis with LLMs and CatBoost
Word Count:	7431
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	*Sakshi KalungePatil*
Date:	12th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Customer Churn Prediction using RAG-Based Sentiment Analysis with LLMs and CatBoost

Sakshi Kacheshwar KalungePatil
x23122366

Abstract

In today's competitive e-commerce era with vast quantity of sentiment-rich textual and behavioral customer data available, predicting customer churn becomes crucial in understanding customer behavior for business sustainability, user engagement, and rapid market growth. This is achievable with the integration of advanced techniques required to extract meaningful insights for data-driven decision-making. This research investigates the use of a hybrid approach for predicting early signs of costumer churn by integrating a fine-tuned large language model LLM with an ensemble machine learning technique. The study explores the idea of combining the Retrieval-Augmented Generation (RAG) framework with instruction-following fine-tuned Large Language Model Meta AI (LLaMA) for sentiment analysis through customer review data. To boost predictive performance, sentiment-driven features achieved from the RAG module are combined with structural features such as verified purchase and review length, and are passed to the CatBoost model for final churn prediction. The research used a Kaggle dataset consisting of Amazon customer reviews 2023, containing the combination of textual and behavioral characteristics. This hybrid approach reveals that the model achieved an accuracy of 86.75% for the RAG-based fine-tuned LLaMA model and an accuracy of 75.9% for the CatBoost model. Adapting such a hybrid approach validates the effectiveness of combining sentiment-rich textual data with structural features for churn prediction in real-world applications.

1 Introduction

In today's immersive e-commerce and digital era, customer churn prediction becomes a crucial task from both business and customer's perspective to understand user needs which directly impacts the market growth. It is important to fully understand and learn customer behavior for business sustainability and maintaining customer engagement over the years. Traditionally, previously cited work on churn prediction focused solely on structural data such as demographic information, transaction history, account/order status, and purchasing details. However, with such limited structural data, churn prediction becomes less effective as these structural features alone do not leverage the business performance. Hence, with the touch of sentiment-rich textual data, including customer reviews, complaints, and feedback, businesses can enhance productivity, sustainability, and increase user engagement.

Recent studies by Onasanya et al. (n.d.) delivered the need to implement traditional machine learning models such as Random Forest, Support Vector Machine (SVM),

Logistic Regression, and others for churn prediction using structural features, whereas studies by Angelina et al. (2023), Hasan et al. (2024) discussed and explained the impact of ensemble gradient algorithms such as XGBoost, CatBoost in analyzing customer churn prediction using categorical and numerical features alongside the other structural features. In contrast, a study by Zhang et al. (2023) introduced an advanced technique for churn prediction by integrating a retrieval-augmented generation (RAG) module with a fine-tuned large language model LLM to understand and learn churn retention through sentiments captured from textual data. This approach was found to be more efficient than other methods, providing better accuracy in customer churn prediction by analyzing customer behavior through customer reviews, complaints, and feedback.

This research explores the idea of combining the hybrid approach - RAG-based LLM with CatBoost to predict whether sentiment analysis through sentiment-rich textual data from customer reviews can provide an early sign of churn retention by analyzing the hidden emotional and behavioral patterns alongside structural features.

Research Question: Can Sentiment Analysis (Customer reviews or complaints) help in early churn prediction from Customer Behavior Prediction?

Research Objectives:

1. To construct an instruction-following dataset before fine-tuning a TinyLLaMA model to simulate interactive Human-AI conversations.

2. To execute a Retrieval-Augmented Generation (RAG) module for extracting relevant context based on the input prompt provided.

3. To integrate RAG-based sentiment scores with structural features and further evaluate the model performance through CatBoost classifier for churn prediction.

4. To assess model performance of LLaMA + RAG and CatBoost models using standard performance metrics and visualizations.

1.1 Research Outline

This section provides the research workflow in a structured and organized way. Initially, Section 2 gives a detailed summary of previous work based on churn prediction using traditional and ensemble machine learning algorithms and sentiment analysis using RAG-based fine-tuned LLMs. In addition, a comparison table is displayed that defines the research gap. Section 3 describes the research methodology by dividing it into five different phases. Section 4 illustrates the design specification of the proposed research. Section 5 discusses the tools and libraries used and the implementation of the adapted models. Section 6 evaluates the model performance through a series of evaluation metrics, including classification report and confusion matrix with Tableau visualizations. Lastly, Section 7 concludes the research with the key findings, limitations, and discusses the possible enhancements for future work.

2 Related Work

This section provides an in-depth literature review by examining the strengths and limitations of each related work divided into subcategories focusing on different approaches. Sub-section 2.1 explores the use of traditional machine learning techniques to predict

consumer retention. Sub-section 2.2 focuses on the implementation of advanced ensemble machine learning algorithms and, lastly, Sub-section 2.3 incorporates the use of the Retrieval-Augmented Generation (RAG) module integrated with Large Language Models (LLMs) to enable early prediction of customer churn through sentiment analysis.

2.1 Customer churn prediction using traditional Supervised Machine Learning Models

Muradkhanli and Karimov (2023) combines the intelligence of big data analytics and machine learning to predict churn, additionally incorporating the use of sentiment analysis to understand the customer behavior by implementing supervised machine learning models including Random Forest and Support Vector Machine (SVM). Lastly, combines sentiment scores to detect customer churn and analyze models by evaluating performance metrics, exhibiting strong churn prediction performance. In contrast, the proposed research by Lin (2024) explores the idea of sentiments within the education industry focusing on improving curriculum excellence rather than churn by using classification models such as Support Vector Machines (SVM), Recurrent Neural Networks (RNN) and Naive Bayes but fails to provide numeric evaluation with no performance metrics such as accuracy, precision, recall and others.

In one of the studies, Onasanya et al. (n.d.) focuses on expanding small businesses to examine customer behavior through structured data consisting of transactional and demographic information and implements a series of four machine learning models such as Logistic Regression, Random Forest, XGBoost, and Gradient Boosting. The models were then measured through performance evaluation metrics - accuracy, precision, recall, F1 score, and Receiver Operating Characteristic - Area Under Cover (ROC-AUC) curve. Based on the comparative analysis, the findings hinted at Random Forest as the best suitable model to predict customer churn with an accuracy of 81.9%. Likewise, the research conducted by Manzoor et al. (2024) delivered a precise overview of several machine learning models including Decision Trees, Support Vector Machines (SVM), Logistic Regression and Naive Bayes with performance metrics - accuracy, precision, recall score and F1-score.

Another work demonstrated by Sam et al. (2024) utilized the transactional data from telecom services to analyze customer behavior for early churn retention. Similarly, Dhanushkodi et al. (2025) focuses on the adoption of historical sales data from the retail industry. Both strategies offer useful insights for predicting customer churn with the implementation of traditional supervised machine learning models, including Support Vector Machine (SVM) and Random Forest. However, neither of the strategies emphasizes the use of behavioral data to predict early churn detection. Therefore, this research focuses on the use of sentiment-rich data as an important key factor in early churn rate prediction.

The approaches Manzoor et al. (2024), Sam et al. (2024) and Dhanushkodi et al. (2025) highlight a few limitations that could have been overcome by inculcating the approach stated by Niu et al. (2017) which presents a distinctive way of using behavioral data rather than focusing on traditional modes to address the customer churn prediction. However, it still lacks potential limitations as the churn detection cannot rely solely on behavioral or demographic data. Sentiments, on the other hand, convey a stronger sense of understanding making it a vital early churn indicator.

2.2 Customer churn prediction using Ensemble Machine Learning Models

This sub-section of the literature reviews the performance of ensemble gradient boosting machine learning models such as CatBoost, XGBoost and LightGBM in the field of customer churn prediction known for their efficient and significant performance in handling numerical and categorical data.

A proposed work by Deng et al. (2021) predicted the bank user churn by addressing the use of advanced ensemble learning models such as Random Forest, AdaBoost, and Gradient Boosting on a real-world bank dataset consisting of 60,000 user data by achieving an accuracy of 86.7% and AUC score of 0.91. Similarly, Kimura (2022) in their research went a step ahead by tackling the class imbalance issue on telecom dataset by implementing the SMOTE resampling technique on various ensemble machine learning models such as Random Forest and XGBoost resulting in an optimal precision of 88.3% and recall score of 85.7%. Since, both studies showed exceptional performance in terms of accuracies using ensemble models, they fell short in analyzing emotional tone or sentiment cues from the user data by limiting their focus on structured and demographic data.

The study by Angelina et al. (2023), focused on implementing churn prediction using the CatBoost machine learning algorithm due to its native capability of processing categorical data without requiring heavy preprocessing, resulting in an accuracy of 95%. Although both studies performed exceptionally well on ensemble machine learning models with SMOTE technique, neither of the studies focused on textual sentiment-rich data ignoring the emotional context from customer behavior thus, resulting in limited churn detection.

Another work by Li et al. (2024) innovatively builds a marketing pipeline by blending approximately 40,000 marketing features and user demographic data with the help of CatBoost, Random Forest, and XGBoost ensemble machine learning models to detect customer churn by achieving an accuracy of (81.0%), (79%), and (78.2%), respectively. Comparatively, Hasan et al. (2024) stresses the idea of implementing the CatBoost machine learning model, beneficial for handling numerical and categorical data requiring minimal manual pre-processing for churn retention prediction on telecom dataset thus, achieving an accuracy of 94%. Even though both studies display high performance for various ensemble machine learning models, they still lack predicting customer sentiments, solely focusing on structural data neglecting the text-driven features that are essential for early churn prediction.

Lastly, the study conducted by Prasher et al. (2025) demonstrated the effectiveness of CatBoost model in analyzing sentiment-rich data with the help of Term Frequency–Inverse Document Frequency (TF-IDF) features on suicide-intended textual data by achieving an accuracy of 89.2%, outperforming other machine learning models such as Random Forest and SVM. This paper indicates the importance of CatBoost in handling rich textual data, but still omits the behavioral data, which provides more accurate results in early churn prediction.

2.3 Sentiment Analysis using RAG-Based models integrated with LLMs

In one of the approaches conducted by Zhang et al. (2023), the research evaluates the strengths of general-purpose LLMs like ChatGPT by implementing sentiment analysis tasks on 26 diverse datasets to critically assess the performance capabilities of LLMs. Even though the model performs well with the combination of zero-shot and few-shot techniques, when compared to the smaller pre-trained models, it struggles to analyze aspect-based analysis and sentiment detection, suggesting fine-tuning of the model to capture early signs of dissatisfaction leading to early churn prediction. In comparison, Touvron et al. (2023) in their study introduces the LLaMA model, a lightweight open source LLM model that provides unrestricted access to publicly available data, achieving comparatively better results as compared to GPT and ChatGPT hence prioritizing transparency. Although it surpasses models like GPT-3 by including NLP tasks, it still lacks in providing accurate results and indicates the need for fine-tuning of the trained models for sentiment-specific predictions.

The above approaches Zhang et al. (2023) and Touvron et al. (2023) demonstrate the importance of LLMs in sentiment analysis but highlights a gap in extensive sentiment prediction that requires fine-tuning of the models like GPT-3 and LLaMA to generate highly accurate results on customer feedbacks, reviews or complaints in order to detect sentiment signals more significantly.

Another research proposed by Kiesler et al. (2023) explores the idea of how ChatGPT responds to student tasks by generating formative programming feedback. The authors further analyze student-generated programming queries to evaluate the quality of feedback, as this methodology provides significant findings as to how the LLMs process and evaluate user-generated inputs. The research shows that LLMs like ChatGPT adapt and understand user responses effectively, but the clearance in the output may degrade if the input contains complex language which the model fails to understand. Hence, the study further explains the need for instruction-following engineering to carefully understand the sentiment-rich textual data or any informal feedback to extract sentiments from the user-driven inputs. Another work stated by Lou et al. (2024) explains a detailed summary of how various LLMs models such as FLAN-T5 and Alpaca yield improved responses when applied an instruction-following prompt to the models by following a user-specified prompt, resulting in a comparatively better output than traditional models. When an instruction-following prompt is applied to these models, they may produce prominent results in sentiment extraction, making them powerful tools for generating formative student feedback. However, the study highlights a limitation stating that without strict adherence to the prompt, the model may mislead and yield false outputs.

A study presented by Zhang et al. (2023) explores a new sentiment analysis framework by implementing a retrieval-augmented generation model on fine-tuned LLMs. This RAG-Based model is used to enhance financial texts through the creation of instruction-following dataset followed by a RAG module to extract the relevant financial context from external sources resulting in significant improvements in terms of accuracy compared to traditional models - FinBERT and base LLMs. Similarly, a proposed work by Mathebula et al. (2024) developed a novel structure, namely language feature extraction and adaptation (LFEAR) by integrating the RAG module with an auto-regressive fine-tuned model by achieving an excellent precision score of 98.45% but comes with a limitation of heavy computational resources.

To further bridge the gap, Zheng et al. (2024) introduces an advanced technique called Aspect-Based Sentiment Analysis (ABSA), providing high-quality performance on SemEval datasets by dynamically evaluating in-context examples rather than static examples instruction tuning as this approach optimizes the sentiment granularity by identifying specific product aspects by enhancing the detection of specific complaints directly linked to customer churn. Despite this, the model remains sensitive to the example choice and is unsuitable for longer textual data containing reviews or complaints. Nayinzira and Adda (2024) introduces a Retrieval-Augmented Generation RAG module, namely the SentimentCareBot chatbot designed to provide mental health support sentiment analysis. The chatbot combines real-time sentiments from users' emotional states to capture responses. The model was trained and evaluated on users' mental health data and user-driven inputs. This RAG-Based chat bot used Facebook AI Similarity Search (FAISS) transformer for generating faster document retrieval and responses. Despite the RAG-Based chatbot, the study exhibits a series of weaknesses including possible misclassification of sentiment and lacks in clinical validation.

2.4 Research Gap

Sub-section 2.1 of the related work covers the strengths and weaknesses of traditional supervised machine learning models focusing on structural and demographic data, thus failing in analyzing behavioral data containing sentiments through any form of textual data such as customer reviews, feedback, and complaints. As it fells short in analyzing the behavioral trends, it cannot provide early signs of churn prediction. Based on the limitations stated, this research addresses the gap by implementing customer churn prediction through sentiment analysis merging customer behavioral data.

Despite the merits listed in sub-section 2.2 for advanced ensemble machine learning models depicting high accuracy with other performance evaluation metrics, CatBoost classifier's ability in handling categorical data without extreme manual preprocessing, the studies still lack in capturing behavioral patterns, hence focusing only on the structured data. To address these shortcomings, this research proposes to bridge this niche by combining sentiment analysis with an ensemble machine learning model namely CatBoost.

Sub-section 2.3 of related work demonstrates the implementation of LLMs integrated with RAG-based sentiment analysis. Traditional general-purpose LLMs such as GPT3 provide flexibility, but fail to detect early signs of churn retention unless required to be fine-tuned. On the other hand, Instruction-tuned LLMs address this gap but still require prompt engineering to deliver accurate responses. This research intends to fill the gap by combining the fine-tuning of LLMs with a human-written instruction-following dataset to predict sentiments via customer feedback, complaints, and reviews, thus building a robust framework for early churn prediction. Hence, this research aims to bridge this gap by developing a retrieval-augmented framework (RAG) integrated with fine-tuned LLaMA LLM to detect early churn indicators through sentiment analysis. The table below summarizes the technique(s) used by the authors in their research studies.

Research Paper	Supervised ML	Ensemble ML	LLMs	RAG Module	Model(s) used
Manzoor et al. (2024)	✓	-	-	-	SVM, Logistic Regression, Naive Bayes
Dhanushkodi et al. (2025)	✓	-	-	-	SVM, Random Forest
Sam et al. (2024)	✓	✓	-	-	Random Forest and XBoost
Angelina et al. (2023)	✓	✓	-	-	CatBoost with fine tuning
Kimura (2023)	✓	✓	-	-	SMOTE + LightGBM, XGB
Shaikhsurab and Magadum (2024)	-	✓	-	-	CatBoost + LightGBM
Deng et al. (2024)	✓	✓	-	-	Random Forest, XGBoost, LightGBM
Onasanya et al. (2022)	✓	✓	-	-	Logistic Regression, Random Forest, XGBoost
Dorokhov et al. (2020)	✓	-	-	-	Logistic Regression, SVM, Random Forest
Hasan et al. (2024)	-	✓	-	-	CatBoost
Li et al. (2024)	-	✓	-	-	CatBoost, XGBoost
Zhang et al. (2023)	-	-	✓	-	ChatGPT
Touvron et al. (2023)	-	-	✓	-	LLaMA
Lou et al. (2024)	-	-	✓	-	FLANT-T5 and Alpaca
Zhang et al. (2023)	-	-	✓	✓	RAG + FinBERT
Mathebula et al. (2024)	-	-	✓	✓	RAG + LFEAR
Zheng et al. (2024)	-	-	-	✓	RAG + ABSA
Nayinzira and Adda (2024)	-	-	-	✓	RAG using FAISS
This Research	-	✓	✓	✓	LLaMA + RAG, CatBoost

Table 1: Comparison of Key Features in Related Works

3 Methodology

The following section outlines the methodology used for the research. The first step includes the selection of an appropriate dataset followed by effective data preprocessing techniques. The second step involves fine-tuning the dataset using LLMs by creating an

instruction-following dataset followed by integrating a Retrieval-Augmented Generation (RAG) module to enhance the results, and the last step includes model training using various performance metrics for model evaluation. Figure 1 demonstrates an overview of the methodology incorporated in this research study. The adapted methodology follows the same approach as the KDD methodology, as it consists of five different phases: data selection, data preprocessing, data mining, model training, and model evaluation.

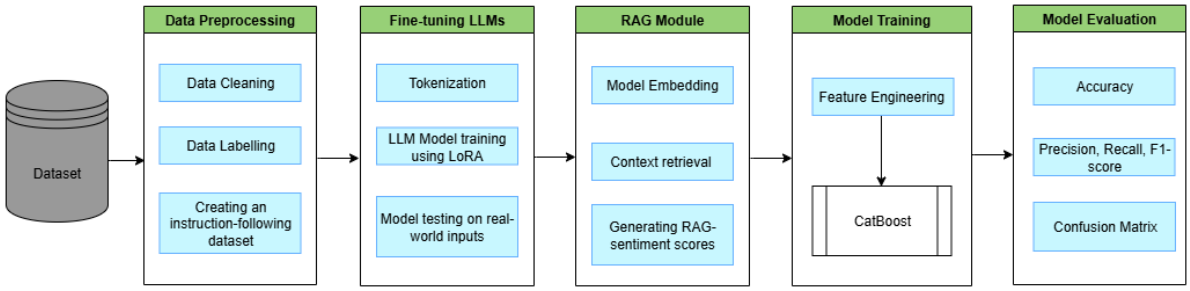


Figure 1: Research methodology

3.1 Data Selection

Data selection is one of the crucial tasks in order to address the research question put forward. This research study uses a publicly available dataset from Kaggle, namely - Amazon reviews 2023 Raviraj (2024). Since, the objective of this research is to predict early signs of customer churn by using not only the structured data but also through sentiment analysis, i.e. analyzing customers' emotional tone and behavior. Thus, making the selected dataset an appropriate choice as it consists of both structured metadata and customer feedback for churn prediction. This Amazon dataset provides a detailed collection of 14,798 customer reviews with respective ratings along with user metadata such as user IDs, timestamps, and verified purchase details indicating whether the customer has made a successful purchase. With the integration of structured data and sentiment-rich textual data, these features are essential for predicting customer churn by performing various tasks such as sentiment analysis and Natural Language Processing (NLP).

3.2 Data Preprocessing

After dataset selection, it is important to verify whether the raw data is ready to implement machine learning tasks by ensuring that the data is cleaned and refined. This step establishes a strong foundation for implementing machine learning tasks by ensuring that the data is refined into clean, transformed, and well-structured data.

3.2.1 Data Cleaning

In the initial phase, the dataset is cleaned to remove unnecessary columns with irrelevant information. In addition, sentiment-rich textual columns are combined. Finally, the combined text is refined into well-structured data by removing any special characters, symbols, punctuations, and converted to lower case suitable for performing the next required machine learning tasks.

3.2.2 Data Labeling

This subsection of the data preprocessing explains the need to convert the numerical values of the 'ratings' column into categorical data, each categorized with one of three labels: Positive, Neutral, or Negative. By inheriting this technique, it helps to analyze sentiments and classify user-satisfactory levels through discrete classes rather than predicting each sentiment-rich text by exact rating number.

3.2.3 Creating an Instruction-following dataset

Before fine-tuning the pre-trained LLMs, it is recommended to construct an instruction-following dataset, as pre-trained LLMs sometimes might not provide accurate results as they are not trained specifically to classify sentiments unless instructions provided explicitly. Lou et al. (2024) in their research demonstrate the importance of mapping an instruction-following dataset before fine-tuning an LLM, as this method significantly provides accurate results in sentiment-based churn prediction by recognizing sentiments through customers' emotional tone. This method of Instruction Following forces the systems to adapt a paradigm that focuses on following human-readable prompts rather than traditional training models. Another study by Zhang et al. (2023) focuses on constructing an instruction-following dataset, consisting of human-written instructions as input and expects responses derived through sentiment labeling as output. By generating such prompt-based dataset, it guides the LLMs to efficiently learn and understand the user's instructions. Following the same approach, this research creates 5 human-written instructions concentrating on the task of classifying the sentiment-rich textual data into categorized sentiment labels where the sample input from the dataset is combined with one randomly selected instruction to produce an output in the format as stated: "Human: [**instruction**] + [**input**], AI Assistant: [**output**]". This process is diagrammatically shown in Figure 2.

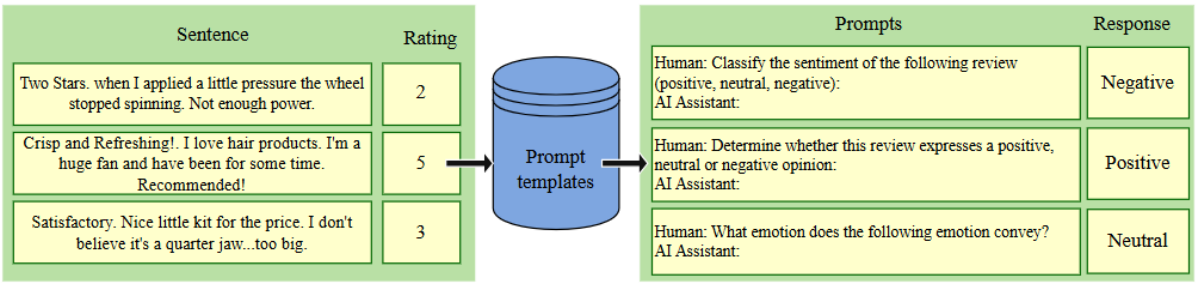


Figure 2: Formatting sentiment-rich data into instruction-following dataset

3.3 Fine-tuning LLMs

Once an instruction-tuning dataset is generated, the next step is fine-tuning the pre-trained LLMs using the constructed instruction-following dataset. With the help of this fine-tuning process, it becomes easier for the model to learn and understand user's instructions and generates the expected output efficiently and accurately when provided with additional human-written instructions to predict behavioral patterns through sentiment analysis.

Fine-tuning LLMs when integrated with an instruction-following dataset provides accurate results, as this approach effectively aligns the user instructions with the LLM system. The encouraging work of Kiesler et al. (2023) illustrates that when fine-tuning integrated with a limited and accurate instruction-following dataset, leads to generate remarkably well results as the fine-tuned LLM adheres as expected to user instructions. Similarly, Zhang et al. (2023) applies the instruction-tuning on an open-source pretrained model like LLaMA for predicting financial sentiments when provided with financial news, tweets or reviews to align the user's behavior with LLM model.

3.3.1 Tokenization

Before fine-tuning the LLaMA model, the raw data is tokenized, i.e. converted into tokens or smaller pieces of text so that the model can understand and process the raw data. The text is tokenized in batches to ensure speedy processing. This process of tokenization converts and returns tokens into numerical values which are then fed to the model for interpretability.

3.3.2 LLM model training using LoRA

Once the raw data is tokenized, the model is fine-tuned using causal language model (CLM), which predicts the next token in the sequence based on the previous context. The model injects Low-Rank Adaption (LoRA) technique to fine-tune smaller layers instead of training the entire model as it is computationally expensive and time consuming. With the adaptation of this technique, the model learns task-specific behavior, without the need to train the entire model, making LoRA highly efficient technique.

3.3.3 Model testing on real-world inputs

After fine-tuning the pre-trained LLaMA model, the model performance is evaluated on real-world input not included in the dataset. When fed a complex customer review as input, with the help of sentiment pipeline, namely text generation, the model accurately classifies it into one of the three labels: Positive, Neutral, or Negative based on the sentiment and emotional tone expressed in the review.

3.4 RAG Module

Retrieval-Augmented Generation (RAG) facilitates LLMs to generate accurate results with powerful prompt engineering by integrating external knowledge sources, hence producing context-aware responses based on human-written instructions. The RAG module proposed in this research is implemented in three phases: Model Embedding, Context retrieval and Generating RAG sentiment scores as follows:

3.4.1 Model Embedding

To implement the RAG pipeline, the raw textual data is first transformed into numerical vector representations capturing the semantic and emotional meaning of customer reviews. Next, a pre-trained sentence transformer model is used to transform each customer review from a set of raw strings to high-dimensional numerical vector embeddings to produce highly relevant responses by identifying emotions from sentiment-rich textual data.

3.4.2 Context Retrieval

RAG module operates in two steps - **Retrieval** and **Generation** where it combines the retrieval of relevant documents from a large dataset or an external data source and generates the expected responses by using an LLM to answer the questions using instruction-following fine-tuning. This subsection describes the first step of RAG pipeline i.e. **Context Retrieval**.

Once the raw data is transformed into embeddings, they are then stored in a FAISS index structure, allowing fast retrieval of similar searches. With the FAISS index, it enables the retrieval of the most relevant and nearby context for a given customer review or an input query resulting in finding the top k nearest neighbors similar to the passed input review.

3.4.3 Generating RAG sentiment scores

After retrieving the relevant documents, the final step is to generate expected responses with the integration of pre-trained fine-tuned model. In this phase, the retrieved context is merged with the prompt and passed through the sentiment classifier pipeline. This mechanism feeds each customer review into the sentiment pipeline in batches, called batch processing. When executed, classifies the input query into one of the binary sentiment labels: 0 for Positive sentiment (not likely churn) and 1 for Negative sentiment (likely churn) and are stored into a new column column RAG_sentiment_scores in the dataset.

3.5 Model Training

After preprocessing and fine-tuning the data using a RAG-based LLaMA model, the next step is to evaluate the performance by training the model. This subsection further discusses the model training process. For model training, the preprocessed data is split into training and testing data with an 80:20 ratio, where 80% of the data is utilized for model training and the rest of the 20% for model testing. The following section outlines the model structure of two discrete models - LLaMA LLM and CatBoost for churn prediction as follows:

1. **RAG-based LLaMA**: The dataset is trained by implementing **TinyLlama-1.1B-Chat**, a lightweight version of LLaMA model that uses causal language modeling (CausalLM) combined with Retrieval-Augmented Generation (RAG) methodology.

2. **CatBoost**: One of the ensemble gradient boosting models used for classification purposes and handling numerical and categorical data without heavy manual heavy preprocessing. This research uses CatBoost to predict the behavioral patterns of customers through the structural details of the customers.

3.6 Model Evaluation

Model evaluation is the final stage where it depicts the model performance based on various evaluation parameters. This section of the proposed methodology discusses the evaluation metrics used to assess the performance of the adapted models. The purpose of this research is to predict early signs of churn through sentiment analysis by analyzing structural and behavioral data. The model is evaluated with the most common evaluation metrics, such as accuracy, precision, recall, F1-score, and confusion matrix, which helps

to depict the confusion matrix elements - True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN). In addition to these metrics, the model also evaluated the validation loss or the cross-entropy loss on the testing data.

4 Design Specification

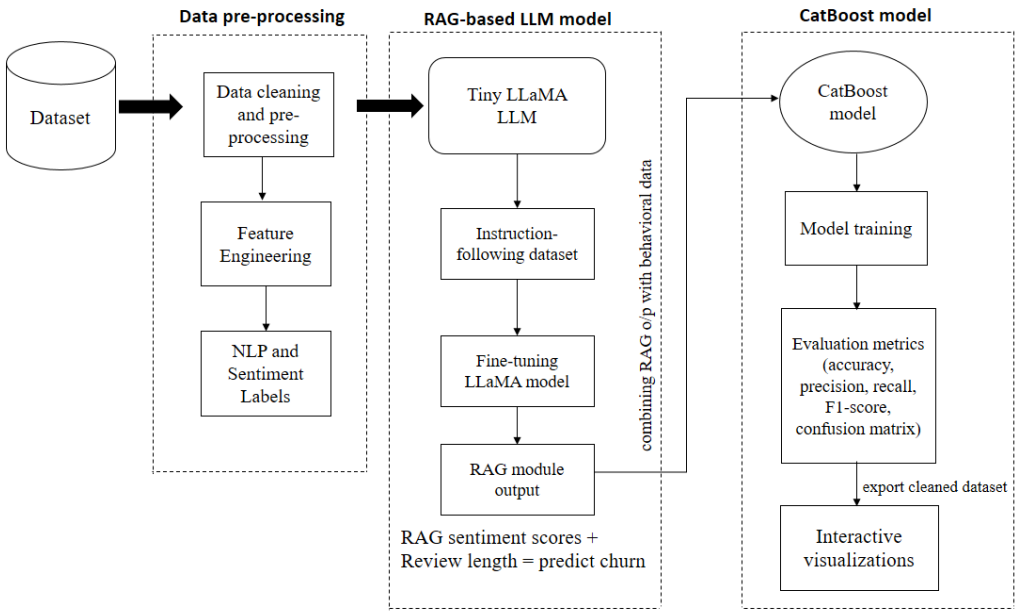


Figure 3: System architecture

This section outlines the system architecture followed to implement the churn prediction model used in the research. The proposed design specification describes the end-to-end workflow of the model carried out from pre-processing phase through to the evaluation phase. The system architecture is divided into the following three phases: Data pre-processing, Rag-based LLM and CatBoost as shown in Figure 3 above. The first phase, Data pre-processing, builds the foundation for transforming the raw data into cleaned data responsible for handling modeling tasks. In this phase, the raw data including structural and behavioral data is gone through a several data cleaning steps including removal of unnecessary fields, handling missing values, text normalization (lowercasing, punctuations, and symbols elimination) and generating sentiment labels from user-ratings which is then successfully passed to the next phase for fine-tuning of the model. The second phase, RAG-based LLM, deals with advanced sentiment analysis through creation of instruction-following dataset, fine-tuning of the LLaMA model and implementation of RAG Module which generates RAG sentiment scores and review length as output and are passed to the last phase. In the last phase, CatBoost model, the outputs of previous phase are combined with the structural data for final churn prediction; this combined dataset containing both the structural and behavioral traits is then trained using CatBoost to predict the churners and non-churners through customer reviews. Lastly, the model performance is evaluated through a set of standard performance metrics, including the

confusion matrix and the classification report, and the cleaned dataset is then exported to generate interactive visualizations using Tableau to visually identify the churn prediction.

5 Implementation

The implementation section of the research provides a detailed overview of tools and languages used to carry out the execution of two distinct models for sentiment analysis: LLaMA LLM integrated with the Retrieval-Augmented Generation (RAG) model based on textual data and a CatBoost model based on structural data. This section explains the tools and libraries used and the model implementation.

5.1 Tools, Libraries and Framework used

The following tools and languages were used to implement the proposed model:

1. **Google Colab**: The implementation was executed on T4 GPU runtime available in Google Colab.[1]
2. **Python**: All tasks including data preprocessing, model training, and model evaluation, were implemented using python as the programming language.
3. **Tableau**: To create interactive visualizations showing churn predictions.
4. **Libraries**:
   (a) **pandas** and **numpy**: For executing data analysis tasks such as data cleaning, handling, and manipulation tasks.
   (b) **transformers** and **peft**: For implementing LLaMA and parameter efficient fine-tuning of large pretrained LLMs using LoRA from Hugging Face. [2] [3]
   (c) **datasets**: For easy access of large datasets, training and evaluation splits for LLaMA model.
   (d) **faiss-cpu**: For fast retrieval of similar search of context within large dataset in RAG module.
   (e) **catboost**: For training the dataset using ensemble gradient boosting machine learning model for analyzing structural features by handling categorical data.
   (f) **scikit-learn**: For evaluating performance metrics such as accuracy, precision, recall, F1-score etc.
   (g) **matplotlib** and **seaborn**: For graphical plotting of the evaluation metrics and visualizations.

5.2 Model Implementation

The model implementation is executed by merging two distinct models - the LLaMA + RAG model and the CatBoost model for churn prediction through sentiment analysis.

[1] `https://colab.google/`
[2] `https://huggingface.co/docs/transformers/en/index`
[3] `https://huggingface.co/docs/peft/en/index`

5.2.1 RAG-based LLaMA Implementation

As the dataset comprises structural and textual data for analyzing sentiment prediction through customer reviews, the first pipeline of model implementation focused on fine-tuning a pretrained version of the LLaMA model - **TinyLLaMA (TinyLlama-1.1B-Chat-v1.0)** integrated with retrieval-augmented generation (RAG) methodology. This form of advanced sentiment analysis enables the model to efficiently classify sentiments based on customer reviews. The text was then tokenized using HuggingFace's tokenizer followed by LoRA, employing parameter-efficient fine-tuning. The training dataset was passed through **2 epochs** using **AutoModelForCausalLM** to predict the next tokens in the given input sequence. Lastly, the output logs and fine-tuned weights were stored in the directory **./llama_finetuned_sentiment**.

Later, the fine-tuned TinyLLaMA model was integrated with the RAG module using a **DistilBERT-based** sentiment analysis pipeline to retrieve similar context using **FAISS** and generate RAG sentiment scores which is then passed to the CatBoost model for final churn prediction using structural data.

5.2.2 CatBoost Implementation

The second pipeline of model implementation involved the analysis of structural features, including user ratings and verified purchase details of the customers integrated with RAG sentiment scores of the Tiny LLaMA + RAG model. CatBoost implementation involved **feature engineering** process in which a binary churn label was derived from the verified_purchase field of the dataset. Additionally, a numeric column was derived, namely review_length feature to capture the length of customer reviews by counting the total length of words in the cleaned text.

The vector representations generated from the RAG model were assigned to 5-dimensional vector embeddings. These vectors together were further combined with RAG sentiment scores and review_length for CatBoost model training. The dataset was divided into training and testing data using an 80:20 ratio. The class imbalance issue was also addressed by setting the class weights for the minority class i.e. churn (1) to weights of 3, while the majority class with not churn (2) remained unchanged with default class weight set to 1. After training, the model performance was assessed using standard performance evaluation metrics for each class.

6 Evaluation

In the Evaluation section of the proposed research, the performance of two distinct models i.e. TinyLLaMA + RAG module and CatBoost are evaluated using a set of standard matrices like accuracy, precision, recall, F1-score supported by a classification report and confusion matrix to briefly describe the predictive analytical capabilities throughout the churn and non-churn classes by combining structural and textual data through sentiment analysis for churn prediction. The model evaluation is categorized into three experiments as follows:

6.1 Experiment 1 (TinyLLaMA + RAG model)

The first experiment was evaluated on TinyLLaMA integrated with the RAG module to predict sentiments through customer reviews only containing textual data. As shown in Table 2, it demonstrates the model performance measured under various performance metrics by achieving an accuracy of **86.75%**, depicting strong overall performance in predicting customer churn. In addition to the accuracy, the model performance was also evaluated using precision, recall value and F1-score representing the model's sensitivity in capturing actual churn cases.

Metric	**Value**
Accuracy	0.8675
Precision	0.5773
Recall	0.8134
F1 Score	0.6753

Table 2: Evaluation Metrics for RAG-based LLaMA Model

Table 3, describes the confusion matrix providing a more detailed view of the prediction outcomes. For the Not Churn (0) class, the model was able to correctly identify 9,768 cases and misclassify 1,350. However, for the Churn (1) class, the model correctly classified 1,844 cases while producing false positives for 423 cases. From the results described, the model showed stronger capability in identifying and predicting churn rates with a slight trade-off in misclassifying non-churned users as churned.

Class	**Precision**	**Recall**	**F1-score**	**Predicted 0**	**Predicted 1**
Not Churn (0)	0.96	0.88	0.92	9768	1350
Churn (1)	0.58	0.81	0.68	423	1844

Table 3: Confusion Matrix for RAG-based LLaMA Model

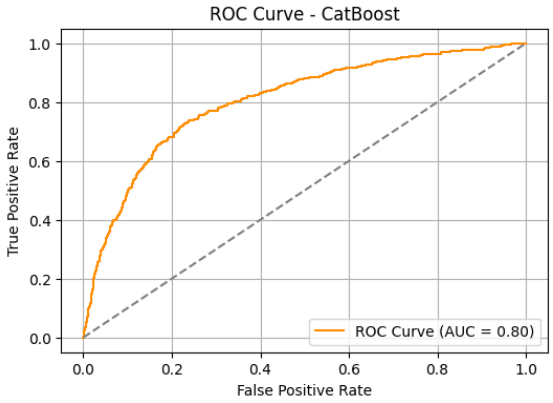
6.2 Experiment 2 (CatBoost model)

The second experiment includes evaluating the model performance through CatBoost. Table 4, summarizes the performance of the model through accuracy, precision, recall, F1-score, classification report, and confusion matrix. An accuracy of **75.9%** was achieved demonstrating the overall performance of the model.

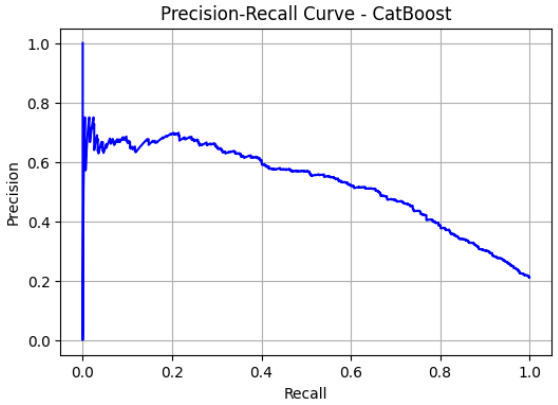
Class	Precision	Recall	F1-score	Support
Not Churn (0)	0.91	0.77	0.83	2336
Churn (1)	0.46	0.73	0.56	624
			Accuracy	0.759

Table 4: Classification Report for CatBoost Model

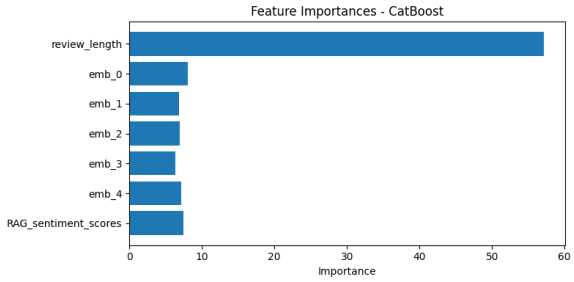
The Figure 4 shows the evaluation visualizations for the CatBoost model. Figure 4a shows the ROC curve plotting the true positives against the false positives. This helps in distinguishing the positive and negative classes in the model. Similarly, Figure 4b shows the relationship between precision and recall. This curve reflects how well the model performs in correctly identifying a positive class with fewer false positives. In the last figure 4c, the graph shows the features that played a major role in the model predictions, stating that the numerical feature: review_length contributed the most in churn prediction.



(a) ROC Curve



(b) Precision-Recall Curve



(c) Feature Importance

Figure 4: ROC Curve and Precision-Recall Curve for CatBoost

6.3 Experiment 3 (Tableau visualizations)

Figure 5 illustrates various visualizations created in Tableau. The figure shows the relationship between Ratings and Sentiment, plots a bar graph to display the sentiment distribution, displays the sentiment trend over timestamp, and categorizes the churn by sentiment labels in the form of a pie chart.

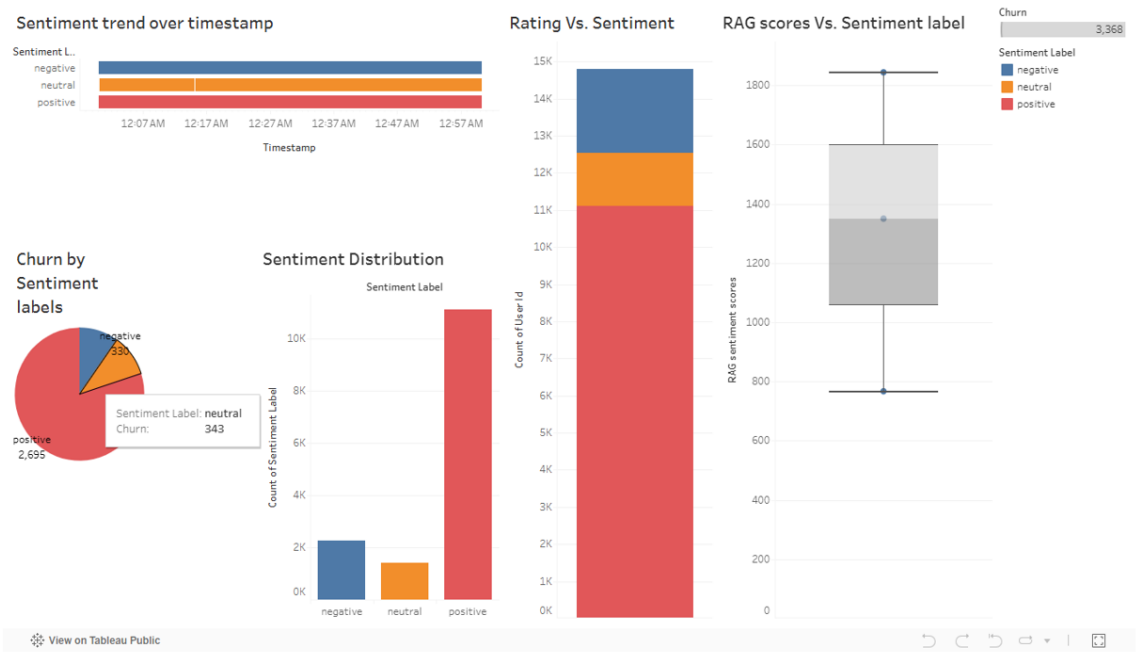


Figure 5: visualizations

6.4 Discussion

The above sections of the evaluation discussed the model performance based on two distinct experimental models such as the TinyLLaMA + RAG model and the CatBoost model in predicting customer churn using a combination of structural and textual data. From the final results, the experiments provided a brief understanding by stating the strengths and limitations of each modeling approach. This subsection aims to discuss the same as follows:

In **Experiment 1**, the TinyLLaMA integrated with the RAG model expressed how well the model performed in handling sentiment-rich textual data by achieving an accuracy of **86.75%** and a recall value of **81.34%** for the churn class, since the model successfully captured majority of the actual churners. However, the model evaluated the precision of **57.73%** suggesting that the model falsely predicted the non-churners being churners, highlighting as a limitation in identifying the actual churn prediction.

In contrast, **Experiment 2** showed the performance evaluation for the CatBoost model that focused primarily on structural features such as review length, RAG-based embeddings and RAG sentiment scores generated via RAG-based LLM model. With an accuracy of **75.9%** and precision of **91%** the model performed considerably well for the non-churn class whereas achieved a precision of just **46%** indicating a higher rate of false positives. The CatBoost performed well for some metrics and remained less effective in a few stating that even though it is efficient and interpretable, it still lacks in capturing the emotional patterns in textual data, limiting its overall efficiency in churn prediction.

To conclude, both models showed their respective strengths but also highlight a few shortcomings that could benefit from inculcating improved embedding strategies, better handling of class imbalance issue and with the use of extensive hyperparameter tuning.

7 Conclusion and Future Work

This proposed research investigated the use of a hybrid approach by integrating a fine-tuned TinyLLaMA + RAG model for sentiment-rich analysis with a CatBoost model for behavioral analysis for customer churn prediction. The aim of the research was to employ sentiment-rich textual data and structural data to elevate the predictive analysis and identification of early signs for churn prediction.

The RAG-based TinyLLaMA model for sentiment analysis through textual data achieved an accuracy of **86.75%** with a recall value of **81.34%** for the churn class, efficiently identifying the actual churners. However, the model evaluated a lower precision score, suggesting some false positives of non-churners. In contrast, CatBoost achieved a reasonable accuracy of **75.9%** with higher precision for the non-churn class and reduced performance for identification of potential churners. The results obtained highlight the strengths of both models and the importance of integrating a large language model with a machine learning model for advanced sentiment analysis.

The implications of the proposed research confirm the significance of a hybrid approach, i.e. integration of the LLM model with the ensemble machine learning model for churn prediction. In addition, it also validates the use of the RAG module to enhance the gap between static textual data and dynamic model interpretation. Although, despite the strengths, the model implementation faces few limitations, such as extensive computational resources, time-consuming fine-tuning, and slow RAG inference for huge data. Due to these limitations, future work may include the following enhancements:

1. **Aspect-Based Sentiment Analysis (ABSA)**: ABSA integration with LLM + RAG pipeline could improve churn prediction by identifying sentiment-rich textual data related to specific product features instead of general user data like product type, product price, thus, enabling accurate retention techniques.

2. **Use of actual retrieval-augmented external knowledge**: Instead of relying on RAG simulated embeddings, actual retrieval-augmented external knowledge resources could be used to yield high accuracy and model performance.

3. **Multi-class Sentiment Analysis**: Instead of using RAG-based sentiment scores, multi-class sentiment labels - positive, neutral, negative could be integrated with the CatBoost model to enable detailed behavioral customer intent.

4. **Real-time Churn Prediction System**: By combining both TinyLLaMA + RAG and CatBoost pipelines, integrated with cloud services such as AWS, a real-time churn prediction framework can be deployed.

References

Angelina, J., Subhashini, S. J., Harish Baba, S., Dheeraj Kumar Reddy, P., Sudheer Kumar Reddy, P. V. and Sameer Khan, K. (2023). A Machine Learning Model for Customer Churn Prediction using CatBoost Classifier, *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 166–172.
URL: *https://ieeexplore.ieee.org/abstract/document/10142823*

Deng, Y., Li, D., Yang, L., Tang, J. and Zhao, J. (2021). Analysis and prediction of bank user churn based on ensemble learning algorithm, *2021 IEEE International Conference*

on Power Electronics, Computer Applications (ICPECA), pp. 288–291.
URL: *https://ieeexplore.ieee.org/document/9362520*

Dhanushkodi, K., Bala, A., Kodipyaka, N. and Shreyas, V. (2025). Customer Behavior Analysis and Predictive Modeling in Supermarket Retail: A Comprehensive Data Mining Approach, *IEEE Access* **13**: 2945–2957.
URL: *https://ieeexplore.ieee.org/abstract/document/10542125*

Hasan, B., Zubair, Shaikh, S. A., Khaliq, A. and Nadeem, G. (2024). Data-Driven Decision-Making: Accurate Customer Churn Prediction with Cat-Boost, *The Asian Bulletin of Big Data Management* **4**(02).
URL: *https://abbdm.com/index.php/Journal/article/view/175*

Kiesler, N., Lohr, D. and Keuning, H. (2023). Exploring the Potential of Large Language Models to Generate Formative Programming Feedback, *2023 IEEE Frontiers in Education Conference (FIE)*, pp. 1–5.
URL: *https://ieeexplore.ieee.org/abstract/document/10343457*

Kimura, T. (2022). Customer Churn Prediction with Hybrid Resampling and Ensemble Learning.
URL: *https://www.researchgate.net/publication/360287935*

Li, A., Yang, T., Zhan, X., Shi, Y. and Li, H. (2024). Utilizing Data Science and AI for Customer Churn Prediction in Marketing, *Journal of Theory and Practice of Engineering Science* **4**(05): 72–79. Number: 05.
URL: *https://centuryscipub.com/index.php/jtpes/article/view/593*

Lin, F. (2024). Sentiment analysis in online education: An analytical approach and application, *Applied and Computational Engineering* **33**: 9–17.
URL: *https://www.ewadirect.com/proceedings/ace/article/view/9950*

Lou, R., Zhang, K. and Yin, W. (2024). Large Language Model Instruction Following: A Survey of Progresses and Challenges, *Computational Linguistics* **50**(3): 1053–1095.
URL: *https://direct.mit.edu/coli/article/50/3/1053/121669/Large-Language-Model-Instruction-Following-A*

Manzoor, A., Atif Qureshi, M., Kidney, E. and Longo, L. (2024). A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners, *IEEE Access* **12**: 70434–70463.
URL: *https://ieeexplore.ieee.org/document/10531735*

Mathebula, M., Modupe, A. and Marivate, V. (2024). Fine-Tuning Retrieval-Augmented Generation with an Auto-Regressive Language Model for Sentiment Analysis in Financial Reviews, *Applied Sciences* **14**(23): 10782. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute.
URL: *https://www.mdpi.com/2076-3417/14/23/10782*

Muradkhanli, L. G. and Karimov, Z. M. (2023). Customer Behavior Analysis Using Big Data Analytics And Machine Learning, *Problems of Information Society* pp. 61–67. Publisher: Problems of Information Society.
URL: *https://jpis.az/en/journals/306*

Nayinzira, J. P. and Adda, M. (2024). SentimentCareBot: Retrieval-Augmented Generation Chatbot for Mental Health Support with Sentiment Analysis, *Procedia Computer Science* **251**: 334–341.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050924033520*

Niu, X., Li, C. and Yu, X. (2017). Predictive Analytics of E-Commerce Search Behavior for Conversion, *AMCIS 2017 Proceedings* .
URL: *https://aisel.aisnet.org/amcis2017/DataScience/Presentations/7*

Onasanya, A. E., Aroyewun, O. and Okonkwo, R. (n.d.). Predictive Analytics For Customer Behaviour: Developing a Predicitve Model That Analyzes Customer Data To Forecast Future Buying Trends And Preferences, Enabling Small Businesses To Tailor Their Marketing And Product Strategies Effectively.
URL: *https://www.researchgate.net/publication/378176015*

Prasher, S., Nelson, L. and Pappa, K. (2025). Suicidal Risk Prediction Using Catboost Classifier: A Comparative Study of Six Machine Learning Classifiers, *ResearchGate*.
URL: *https://www.researchgate.net/publication/388767853*

Raviraj (2024). Amazon reviews 2023.
URL: *https://www.kaggle.com/datasets/ravirajbabasomane/amazon-reviews-2023*

Sam, G., Asuquo, P. and Stephen, B. (2024). Customer Churn Prediction using Machine Learning Models, *Journal of Engineering Research and Reports* **26**(2): 181–193.
URL: *https://journaljerr.com/index.php/JERR/article/view/1081*

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs].
URL: *http://arxiv.org/abs/2302.13971*

Zhang, B., Yang, H., Zhou, T., Ali Babar, M. and Liu, X.-Y. (2023). Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models, *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, Association for Computing Machinery, New York, NY, USA, pp. 349–356.
URL: *https://doi.org/10.1145/3604237.3626866*

Zheng, G., Wang, J., Yu, L.-C. and Zhang, X. (2024). Instruction Tuning with Retrieval-based Examples Ranking for Aspect-based Sentiment Analysis. arXiv:2405.18035 [cs].
URL: *http://arxiv.org/abs/2405.18035*