ISSN 2723-6471

Unveiling Hybrid Model with Naive Bayes, Deep Learning, Logistic Regression for Predicting Customer Churn and Boost Retention

Devibala Subramanian[1,*], Ajitha[2], Siti Sarah Maidin[3]

[1,2]*Dept of Computer Science, KG College of Arts and Science, Coimbatore, India.*

[3]*Faculty of Data Science and Information Technology (FDSIT), INTI International University, Nilai, Malaysia*

(Received: January 9, 2025; Revised: February 10, 2025; Accepted: April 2, 2025; Available online: April 30, 2025)

Abstract

The telecommunications sector is rapidly evolving but is increasingly challenged by customer churn, where subscribers switch to competing service providers. This study introduces a hybrid model for churn prediction and customer retention by combining machine learning methods—Naive Bayes, Deep Learning, and Logistic Regression—with sentiment analysis on user-generated content (UGC). Data was gathered through two primary sources: survey responses and 352 social media comments from users aged 20–35. The survey data was enriched with features such as gender, age, subscription period, complaints, and retention efforts. The preprocessing steps included handling missing values, scaling features, and encoding categorical variables to ensure model robustness. Experimental results demonstrated that Logistic Regression achieved the highest accuracy (88.45%) and sensitivity (91.33%) in detecting potential churners. The PCA-based approach followed closely with an accuracy of 86.77% and a balanced sensitivity-specificity profile (89.95% and 83.58%, respectively), effectively capturing key churn indicators. Random Forest and Decision Tree classifiers yielded lower sensitivity but remained strong in specificity, indicating their suitability for identifying loyal customers. Attribute weight analysis across models revealed that subscription plan, age, and retention effort were consistently influential in churn prediction. Furthermore, the integration of sentiment analysis provided emotional context to churn behavior, with negative comments triggering alerts for proactive engagement. The study highlights the predictive strength of combining structured survey data and unstructured UGC through machine learning and sentiment analytics. It underscores the importance of personalized retention strategies based on model interpretability and correlation weight findings. This hybrid approach equips telecom companies with actionable insights to minimize churn and sustain customer loyalty in a competitive market.

Keywords: Customer Churn, Churn Prediction, Sentiment Analysis, Machine Learning Methods, Process Innovation, Product Innovation

1. Introduction

The telecommunications industry stands out as a reliable and competitive sector with numerous players vying for market share. [1] Telecom companies generate vast amounts of data annually, given the breadth of services they offer. Customers are presented with a range of options based on quality and cost, resulting in a competitive landscape where each company strives to maximize profits and retain customers. Moreover, customer dissatisfaction often causes churn, where customers move to other providers. This turnover creates a significant challenge for telecom companies, as it has a direct impact on their revenue [2].

Recent advancements in technology, such as the Internet of Things (IoT), Internet Protocol (IP), and big data analytics, are playing a pivotal role in accelerating the growth of the business-to-business (B2B) telecommunication market. B2B telecommunications involve communication solutions tailored to meet the needs of businesses, enabling seamless exchange of information and services between organizations [3]. These solutions facilitate critical activities such as voice communication, video conferencing, secure data transfers, messaging, and customized communication services designed to address specific business requirements. B2B telecommunication services empower companies to establish robust communication frameworks that enhance operational efficiency. They enable organizations to coordinate across supply chains, streamline collaboration with partners, and deliver superior customer support. Additionally, these services are instrumental in supporting day-to-day operations, driving innovation, and ensuring businesses remain

***Corresponding author: Devibala Subramanian (shiiv30@gmail.com)**
DOI: https://doi.org/10.47738/jads.v6i2.675

adaptable in rapidly evolving markets. The ability to exchange information effortlessly through these advanced systems supports informed decision-making, crisis response strategies, and efficient service delivery across industries [4].

Moreover, B2B communication systems are integral to customer relationship management (CRM), remote teamwork, and managing disruptions in operations. The reliance on such services highlights their importance in fostering collaboration, driving innovation, and maintaining competitiveness in the global marketplace. However, the telecommunication industry faces challenges such as customer churn, which has a direct and significant impact on revenue streams. When customers terminate their subscriptions or contracts, telecom providers lose recurring income, which is critical for sustaining business operations. This loss not only affects profitability but also increases customer acquisition costs as providers invest more in marketing and promotional activities to replace churned customers. Consequently, addressing churn is essential for maintaining revenue stability and ensuring long-term growth in the B2B telecommunications sector. This can lead to a decrease in overall revenue and profitability. To replace lost customers, providers must invest in marketing, sales, and promotional activities, increasing customer acquisition costs and potentially hindering profitability. Moreover, losing long-term, loyal customers means missing out on potential revenue, further impacting profitability and limiting the provider's ability to maximize each customer's value. If a substantial number of customers defect to competitors, the provider's market position weakens, resulting in a loss of market share and reduced ability to attract new customers [5].

Fluctuations in the customer base necessitate adjustments in resource allocation, capacity planning, and network infrastructure management. Sudden changes in demand can lead to inefficiencies, increased costs, it also presents challenges in sustaining service quality. In summary, churn affects B2B telecommunication in multiple ways, including customer loss, revenue impact, customer acquisition costs, reputation, customer lifetime value, service quality, competitive landscape, customer relationships, and operational efficiency. Support Vector Machines (SVM) are also used for binary classification, focusing on identifying the best hyperplane that separates churn from non-churn cases [6]. These models can manage both linear and non-linear separations through kernel functions [7], [8], which map input data into higher-dimensional spaces.

2. Literature Review

This review explores existing literature on customer retention and predictive methodologies, focusing on various aspects such as churn prediction, managerial strategies, and the use of social media to gather user feedback for improving retention efforts. Several studies have utilized advanced data mining techniques to extract insights from telecommunications call data, successfully building robust models for predicting customer churn. These models demonstrate promising accuracy and showcase the effective integration of data mining into churn prediction frameworks [9], [10], [11].

Other research has introduced novel predictive approaches that emphasize the role of data certainty in enhancing the accuracy of churn prediction models, particularly in telecommunications. This highlights the importance of considering data reliability when developing decision-support systems [12]. Comprehensive evaluations of customer retention strategies in mobile telecommunications reveal the significance of managerial decision-making. These studies offer strategic recommendations tailored to the dynamic nature of the industry and provide frameworks for optimizing retention practices [13].

Further investigations into B2B customer dynamics explore the management of churn, retention, and profitability. The findings present a strategic model that balances theoretical insights with practical applications, offering actionable solutions for sustaining long-term business relationships [14]. An innovative perspective on churn prediction leverages social network analysis to identify key influencers within customer networks. By analyzing network dynamics, these studies uncover new avenues for understanding and preventing customer attrition [15], [16], [17].

Additional research within the telecommunications industry employs predictive modeling techniques to forecast churn. These efforts emphasize practical applications and present insights relevant to real-world implementation of retention strategies in the telecom sector [18], [19], [20]. Studies focusing on ensemble methods explore the effectiveness of combining multiple classifiers for churn prediction. These works highlight the advantages of classifier integration and offer practical guidance for improving model performance in predictive analytics [21].

 ISSN 2723-6471

Finally, sentiment analysis in social media platforms, particularly Twitter, has been advanced by incorporating deep neural networks and behavioral data. This approach enhances the accuracy of sentiment classification, underscoring the potential of integrating user behavior into predictive models on dynamic digital platforms.

3. Methodology

In this research paper, we employed various machine learning algorithms, with a focus on three specific ones, along with sentiment analysis.

3.1. Machine Learning Approaches - Classification Methods

3.3.1. Deep Learning Algorithm

Deep learning algorithms, known for their ability to automatically learn hierarchical representations from data, are increasingly utilized for churn prediction tasks. Neural networks, particularly feedforward neural networks, as well as more complex architectures like recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), are popular choices. Deep learning requires ample labeled data and careful hyperparameter tuning. While these models can capture complex patterns traditional methods may miss, they demand cautious handling to prevent overfitting and ensure interpretability. The choice of a specific deep learning model depends on data characteristics and churn prediction task requirements [19].

3.3.2. Naïve Bayes

Naïve Bayes is a probabilistic model based on Bayes' theorem. Despite its simplicity, it can effectively predict churn, particularly with a large number of features. Although it assumes independence between features, which may not always hold true, it often yields good results in practice [20], [21].

3.3.3. Logistic Regression

Logistic Regression is a foundational model widely used for binary classification tasks such as churn prediction, as it estimates the probability that a customer will churn based on input features. The model applies a logistic function to model the relationship between a binary dependent variable (e.g., churn or not) and one or more independent variables, producing an output between 0 and 1 that represents the probability of churn. This probability can then be converted into a binary classification using a predefined threshold, typically 0.5. One of the strengths of logistic regression lies in its interpretability—each coefficient reveals the direction and magnitude of the influence of a feature on the likelihood of churn. For example, a positive coefficient indicates that an increase in the corresponding feature value is associated with a higher chance of churn. Although it assumes a linear relationship between the input variables and the log-odds of the outcome, logistic regression remains a robust and interpretable baseline model. In the context of telecommunications, it is particularly useful for identifying key factors that influence customer behavior, thereby enabling companies to implement more targeted customer retention strategies [21], [22].

3.2. Sentimental Analysis

Monitoring sentiment on social media platforms helps companies understand public perceptions of their services. A surge in negative sentiment or consistent customer dissatisfaction can signal potential churn. Analyzing customer feedback offers insights into their sentiments, with positive ones indicating satisfaction and negative ones suggesting dissatisfaction or frustration. Integrating sentiment scores from customer interactions can bolster churn models' predictive accuracy [14].

However, it's essential to use sentiment analysis alongside other relevant features in a comprehensive churn prediction model. Additionally, sentiment analysis models must be tailored and refined for the specific industry domain and customer language to ensure precise results. The analysis categorizes content into positive, neutral, or negative sentiments.

Tracking sentiment on social media platforms helps businesses understand how the public perceives their products or services. A rise in negative sentiment or ongoing customer dissatisfaction may indicate the possibility of churn, as customers unhappy with the service might consider leaving. Analyzing feedback from customers provides valuable

ISSN 2723-6471

insight into their emotions, with positive sentiment reflecting contentment and negative sentiment pointing to frustration or discontent.

Incorporating sentiment analysis into churn prediction models can improve their accuracy by adding an emotional dimension to the data. However, sentiment analysis should not be used on its own. For a more reliable churn prediction, it should be combined with other relevant factors, such as customer behavior, usage patterns, and demographic data, to create a comprehensive model.

3.3. Dataset Description

This dataset is designed for churn prediction in the telecommunications sector. It contains a variety of features that reflect customer demographics, behavior, usage patterns, and interactions with telecom services. The data was gathered from two primary sources: social media comments and customer surveys.

The initial phase involved collecting 352 user comments from Vodafone's social media platforms over a six-month period. These comments came from both male and female users, primarily aged between 20 and 35 years. This qualitative data provides insight into customer sentiment and experiences.

In addition, a structured survey was conducted across various telecom companies to support machine learning analysis. The survey collected quantitative data on several attributes crucial for understanding customer churn behavior. These attributes are summarized in the table 1.

Table 1. Customer Churn Dataset Attributes

Attribute	Description
Gender	Indicates whether the customer is male or female.
Age	Customer's age.
Telecom Company	Name of the telecom company the customer is associated with.
Period	Duration the customer stayed with the company.
Line Type	Type of line used (personal, business, or corporate).
Complaints	Details on how the customer lodged complaints.
Problems	Problems encountered by the customer with the telecom company.
Retention Efforts	Describes the company's efforts to retain the customer.
Churn	Indicates whether the customer switched to another provider.

These variables form a comprehensive basis for analyzing churn patterns and building predictive models. Next, we employed three distinct machine learning algorithms to analyze the survey data and identify the key factors that contribute to predicting customer churn. To ensure a robust analysis, the dataset included not only current customer feedback but also historical data from former customers. This comprehensive dataset allowed us to develop a deeper understanding of churn behaviour over time, offering valuable insights into the patterns and triggers that influence customer retention or departure.

Prior to applying the algorithms, we performed several essential preprocessing steps on the dataset to ensure the quality and accuracy of our analysis. These tasks involved identifying and removing any duplicate records or missing entries that could skew the results. We also addressed any outliers in the data that could introduce noise, ensuring that the model would focus on relevant trends rather than anomalous data points. Additionally, we standardized the features by scaling them to a consistent range, which is crucial for improving the performance of certain algorithms, particularly those sensitive to feature magnitudes, such as logistic regression and decision trees. Furthermore, we encoded categorical variables to convert them into a numerical format that the machine learning models could interpret. This step was critical for ensuring that the data was properly structured for analysis, as most machine learning algorithms require numerical inputs. By performing these preprocessing tasks, we ensured that the dataset was clean, well-structured, and ready for analysis, enabling the algorithms to make accurate predictions and identify the most significant factors influencing churn.

 ISSN 2723-6471

3.4. Exploratory Data Analysis

As part of the Exploratory Data Analysis (EDA), Figure 1 displays the output of checking for missing or null values in the dataset after preprocessing steps. The dataset was initially stripped of the customerID column, as it is not relevant for modeling or analysis purposes. Then, the TotalCharges column was explicitly converted to a numeric data type using coercion to handle non-numeric entries, which may result in missing values.

```
dtype: object

In [9]:
df = df.drop(['customerID'], axis = 1)
df.head()

Out[9]:
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSuppo
0	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	N
1	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	N
2	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	N
3	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Y
4	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	N

```
In [10]:
df['TotalCharges'] = pd.to_numeric(df.TotalCharges, errors='coerce')
df.isnull().sum()

Out[10]:
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges       11
Churn               0
dtype: int64
```

Figure 1. Evaluating null values in dataset

The output shows that all columns except TotalCharges contain zero missing values. The TotalCharges column has 11 null values, likely due to invalid entries that could not be converted to numeric values. This assessment is crucial to ensure data completeness and determine the need for further cleaning or imputation before building predictive models. Prior to commencing modeling, Exploratory Data Analysis (EDA) helps uncover initial patterns and insights within the data, guiding both preprocessing and feature selection decisions. Figure 2 presents a dual-ring donut chart visualizing customer churn distribution across gender groups.

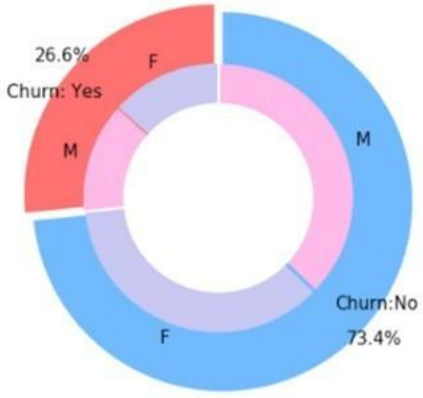

Figure 2. Converting Field Type

The outer ring represents customers who did not churn, accounting for 73.4% of the total population, while the inner ring illustrates customers who did churn, comprising 26.6%. Both segments are further categorized by gender: male (M) and female (F). From the visualization, it can be observed that male customers dominate both the churned and non-churned groups, although the difference is more pronounced among those who stayed. This suggests that gender may have some influence on churn behavior and should be considered in the modeling process.

3.4.1. KNN Classification

This approach can be used to solve both classification and regression problems. It works by classifying a data point based on the characteristics of nearby data points. When a new data point is added, the algorithm looks at the stored

ISSN 2723-6471

data points and compares their similarities. One of the most common ways to measure similarity is by using Euclidean distance, which calculates how far apart the points are from each other in space.

3.4.2. Random Forest Classifier

Random forest is a classification method that uses many decision trees to make predictions. Each tree is built by randomly selecting features and using a technique called bagging to create different variations. This results in a group of trees that are independent of each other. When it's time to make a prediction, the random forest combines the results from all the trees, leading to a more accurate prediction than any single tree could provide on its own. Random Forest relies on decision trees as base learners. A decision tree recursively splits the data based on features, creating a tree-like structure. The tree makes predictions at its leaves [8].

4. Research Contribution

4.1. Analyzing User Generated Content Using Sentiment Analysis

The customer churn prediction model that incorporates user-generated content (UGC) involves a structured sequence of analytical steps. It begins with the collection of comments, reviews, or opinions submitted by users across various platforms. These textual data points serve as the primary input for analysis. Following data collection, preprocessing techniques are applied to clean and standardize the text, preparing it for subsequent analysis as illustrated in Figure 4. This preprocessing includes typical text normalization steps such as tokenization, removal of stop words, and lowercasing.

Once the data is prepared, sentiment analysis is conducted using methods like stemming and lemmatization. Each word in the user-generated content is assigned a sentiment polarity—positive, negative, or neutral—and the overall sentiment of the comment is determined based on these classifications. This sentiment classification becomes a crucial input for the churn prediction process. Based on the sentiment outcomes, comments are categorized as either positive or negative. Negative sentiments are treated as potential indicators of customer dissatisfaction and thus, possible churn. This is visually demonstrated in Figure 5, which shows the classification output based on sentiment scores. When a negative comment is identified, the system triggers an alert that includes the user's ID. This alert is directed to the customer retention team, enabling them to take timely and proactive measures to engage the at-risk customer and potentially prevent churn.

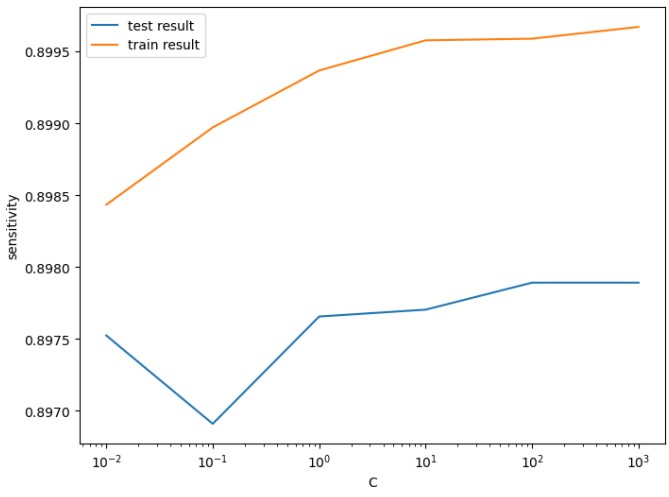

Figure 4. KNN Model Building 1

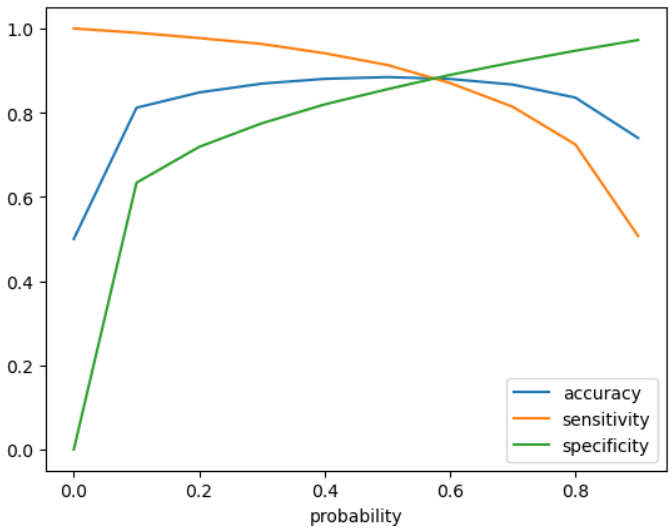

Figure 5. KNN Model Building 2

Through this integrated approach, sentiment analysis becomes a valuable component of the churn prediction model, offering real-time insights and supporting effective retention strategies. As an example, the model can analyze a sentence like "I hate this company," breaking it down word by word to classify each term and determine the overall negative sentiment.

4.2. Churn Prediction

The mathematical representation involves computing the co- variance matrix, eigenvalues, and eigenvectors. Given a dataset X, the covariance matrix is computed as

$$\Sigma = 1 \cdot (XT \cdot X) \quad (1)$$

In the context of Principal Component Analysis (PCA), the number of samples in the dataset is denoted as n. PCA works by transforming the original data into a new set of variables called principal components. These components are derived from the eigenvectors of the covariance matrix, denoted by σ. Each eigenvector represents a direction in the feature space, and the corresponding eigenvalue signifies the variance or importance of that direction. The higher the eigenvalue, the more significant the principal component is in explaining the variance of the data. To assess the effectiveness of PCA, its performance is compared to that of models trained without dimensionality reduction. Key performance metrics such as accuracy, precision, and recall are used to evaluate how well the model performs both before and after applying PCA. This comparison helps determine whether reducing the dimensionality of the data leads to better or comparable results in terms of predictive performance. The PCA process involves several steps, starting with the selection of the optimal number of principal components to retain. This is typically done by examining the explained variance ratio of the components, ensuring that the most important features are preserved while reducing the dimensionality of the dataset. Once the relevant components are selected, a classifier, such as logistic regression or another suitable model, is trained on this reduced feature space. The goal is to achieve a balance between maintaining sufficient information from the original data and reducing computational complexity. This experimentation process helps to identify the most effective number of components that improve model performance while minimizing overfitting and noise.

5. Experimental Results and Discussions

This section provides an overview of the experiments and results of the proposed study, which consists of two parts.

5.1. Identifying The Main Factors

Figure 6 illustrates the proposed sentiment analysis model that integrates user-generated content (UGC) to detect customer churn. The model begins with the extraction and preprocessing of textual data from web-based user comments, including tokenization, extraction, and lemmatization. This processed text is then analyzed using sentiment lexicons such as the Minqing Hu and Bing Liu opinion lexicon to determine word polarity. Comments classified as negative trigger an alert indicating a high probability of churn, enabling timely customer retention interventions.

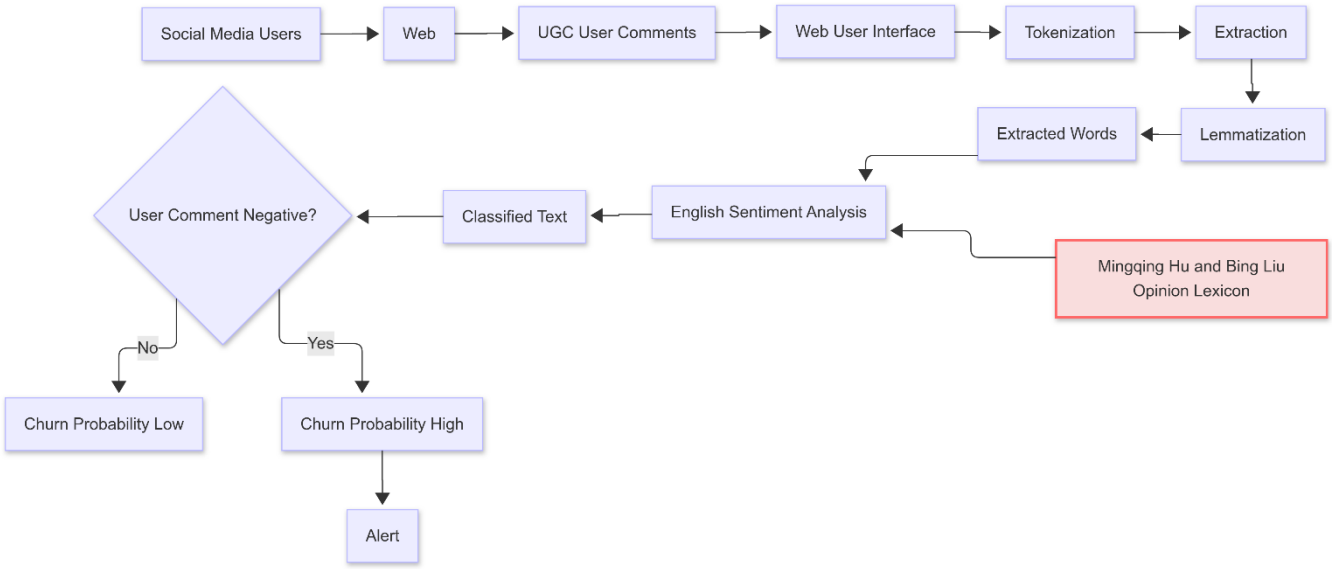

Figure 6. Proposed Sentiment Analysis Model.

 ISSN 2723-6471

To assess the relative importance of features influencing churn prediction, three machine learning models were compared: Naive Bayes, Logistic Regression, and Deep Learning. Each model provides distinct insights based on how they assign weight to different attributes within the dataset. Naive Bayes, which relies on probabilistic reasoning and assumes feature independence, identifies the user's subscription plan as the most significant predictor (weight = 0.1721), followed by company retention efforts (0.0905) and age (0.0756). These findings are detailed in table 2, highlighting the importance of customer engagement and demographic information in churn risk assessment.

Table 2. Naïve Bayes

Attribute	Weight
Your subscription plan	0.172103
The company made efforts to keep you	0.090457
Age	0.075618
An issue occurred on your end with the company.	0.063299
Gender	0.005086
Utilized for lodging a complaint	0.039478
How long have you been a customer of the company?	0.038443

Logistic Regression, a model suited for binary classification, emphasizes gender as the most influential attribute (0.1015), along with lodging complaints (0.0924) and retention efforts (0.0845). The role of tenure and customer-reported issues also ranks highly. These insights are summarized in table 3, showcasing the model's ability to capture relationships between features and churn probability.

Table 3. Logistic Regression

Attribute	Weight
Gender	0.101549
Utilized for lodging a complaint	0.092412
The company made efforts to keep you	0.084525
How long have you been a customer of the company?	0.082095
An issue occurred on your end with the company.	0.064137
Age	0.046731
Your subscription plan	0.006412

In contrast, Deep Learning—leveraging neural network architectures—places the highest importance on age (0.1499) and subscription plan (0.1303), as shown in table 4. This suggests that the model is particularly effective at detecting nuanced patterns in demographic and behavioral data.

Table 4. Deep Learning

Attribute	Weight
age	0.149923
Your subscription plan	0.130300

Several patterns emerge across models. Age, gender, and line type are consistently influential in predicting churn, although their ranking and significance vary. Gender proves especially critical in Logistic Regression and Naive Bayes, while line type holds considerable weight in Naive Bayes and Deep Learning models. On the other hand, retention efforts, though important in Naive Bayes and Logistic Regression, appear less impactful in Deep Learning.

These differences underscore the importance of selecting the appropriate model based on the nature of the data and the business context. Understanding how each algorithm interprets and prioritizes input features allows organizations to refine customer retention strategies with model-specific insights, ultimately supporting more accurate and actionable churn predictions.

ISSN 2723-6471

5.2. Detecting Customers with High Probability of Churning

The PCA-based churn prediction model achieves strong performance with an accuracy of 86.77%. It excels in identifying customers likely to churn, with a high sensitivity of 89.95%. This indicates the effectiveness of the PCA approach in capturing relevant churn patterns. Additionally, the model demonstrates commendable specificity of 83.58%, accurately recognizing customers less prone to churn. The balanced sensitivity and specificity suggest the PCA-driven model achieves an optimal balance in predicting both positive and negative churn instances, making it robust for identifying and mitigating customer churn. The decision tree classifier for churn prediction attains an accuracy of 86.03%, effectively classifying instances. However, it exhibits a relatively lower sensitivity of 69.95%, suggesting it may struggle in identifying potential churners. Nonetheless, with a specificity of 86.61%, the model excels in recognizing customers less likely to churn. The trade-off between sensitivity and specificity indicates potential room for improvement through further tuning or ensemble methods to enhance true positive churn case capture. Despite this, the decision tree classifier remains promising for identifying non-churning customers, reducing false positives and ensuring accurate retention predictions.

The random forest classifier achieves an accuracy of 80.18%, competently making correct classifications. With a sensitivity of 75.65%, it moderately identifies customers at risk of churning, while a specificity of 80.35% effectively distinguishes less likely churners. Despite the robust overall accuracy, there's room for improvement in balancing sensitivity and specificity to capture more true positive churn cases. Nevertheless, the random forest classifier proves reliable in discerning non-churning customers, reducing false positives and ensuring accurate retention predictions. The logistic regression churn prediction model demonstrates impressive overall accuracy at 88.45%, effectively classifying instances. With a high sensitivity of 91.33%, it excels in identifying potential churners, capturing true positive instances effectively.

Additionally, a specificity of 85.56% showcases its proficiency in recognizing less likely churners. The balanced performance between sensitivity and specificity indicates the logistic regression model's reliability in accurate churn predictions, minimizing false positives while maintaining high overall accuracy.

5.3. Result Summary

5.3.1. User Generated Content (UGC) Analysis:

The Principal Component Analysis (PCA)-based churn prediction model delivers highly reliable performance, achieving an impressive accuracy rate of 86.77%. Its strength lies in effectively identifying customers who are likely to discontinue their subscriptions, as evidenced by its high sensitivity score of 89.95%. This high sensitivity underscores the model's ability to detect potential churners accurately by capturing critical patterns and trends within the data. Additionally, the model exhibits a notable specificity of 83.58%, which highlights its capability to correctly classify customers who are less likely to churn. This dual effectiveness ensures that the model not only identifies potential risks of customer loss but also avoids false positives, thereby minimizing unnecessary retention efforts targeted at customers who are not at risk.

The balance between sensitivity and specificity reflects the robustness of the PCA-based approach, demonstrating its capacity to handle both positive and negative churn scenarios effectively. By reducing the dimensionality of the dataset and focusing on the most significant features, PCA simplifies complex data patterns without compromising the integrity of the predictions. This streamlined analysis enables the model to deliver accurate and actionable insights, making it a valuable tool for proactive customer retention strategies. In summary, the PCA-driven model combines precision with reliability, offering a comprehensive solution for identifying at-risk customers while maintaining a balanced approach to minimize errors. Its strong performance underscores its potential to support telecom companies in mitigating churn and enhancing customer retention efforts.

5.3.2. User Generated Content (UGC) Analysis:

In evaluating four different models for predicting customer churn, each demonstrates distinct advantages and challenges, making them suitable for varying contexts depending on business needs. The PCA-based model stands out for its strong overall performance, achieving an accuracy of 86.77%. Its high sensitivity of 89.95% highlights its ability to accurately detect customers at risk of churning, while its specificity of 83.58% ensures it correctly identifies

customers unlikely to churn. This balance between sensitivity and specificity makes the PCA-based model particularly effective for capturing both positive churn cases and those who will remain loyal. The decision tree classifier also shows promising results, with an accuracy of 86.03%. It performs particularly well in terms of specificity, achieving 86.61%, which indicates strong reliability in identifying customers who are not likely to churn. However, its sensitivity is lower at 69.95%, suggesting it may miss a significant number of potential churners. This limitation could be addressed by fine-tuning the model or exploring ensemble methods to enhance its ability to identify positive churn cases without sacrificing accuracy.

The random forest classifier delivers balanced results, achieving an accuracy of 80.18%. Its sensitivity of 75.65% and specificity of 80.35% indicate that it is a reliable model for differentiating between churners and non-churners. While its performance is not as high as some of the other models, its ability to manage complex datasets and reduce overfitting makes it a dependable option for practical applications, particularly in scenarios where balanced predictions are critical. Among the four models, the logistic regression model emerges as the top performer, with an accuracy of 88.45%. Its high sensitivity of 91.33% ensures exceptional detection of potential churners, while its specificity of 85.56% reflects its strength in correctly identifying loyal customers. This combination of high accuracy, sensitivity, and specificity makes logistic regression a robust and interpretable choice for churn prediction, especially in cases where understanding the relationship between predictors and churn likelihood is crucial.

Overall, each model presents unique strengths that cater to different predictive needs. While the PCA-based and logistic regression models excel in accuracy and sensitivity, the decision tree and random forest classifiers offer advantages in specificity and balanced performance.

Selecting the appropriate model depends on the specific objectives, data characteristics, and operational requirements of the churn prediction task. This makes it a robust and interpretable choice for churn prediction, with well-calibrated predictions across both positive and negative cases. Each model presents trade-offs, underscoring the importance of selecting the most suitable model based on specific objectives and requirements.

5.3.3. User Generated Content (UGC) Analysis

In this section, we explore the relationship between user-generated content (UGC) and customer churn by combining sentiment analysis with correlation analysis. This dual approach allows for a more nuanced understanding of how user sentiment relates to key behavioral indicators and customer retention decisions. In addition to evaluating sentiment polarity, we conducted a correlation analysis to examine how sentiment scores align with specific key performance indicators (KPIs) such as engagement patterns, complaint types, and company retention efforts. This analysis revealed how shifts in sentiment may precede changes in customer behavior, offering valuable insights for churn prediction and strategic intervention.

The correlation analysis was applied across different machine learning models, including Deep Learning and Logistic Regression. Table 5 presents the top-weighted attributes identified by the Deep Learning-based correlation model (COR Deep Learning). The most influential factor was "opting out of credit without utilization," which had the highest weight of 0.1883, indicating a strong association with churn behavior. Other significant attributes included efforts by the company to retain users—such as offering complimentary minutes, additional internet megabytes, or invoice reductions—all of which consistently showed high correlation with churn likelihood.

Table. 5. COR Deep Learning

Attribute	Weight
An issue arose for you with the company = Opting out of the credit without utilization	0.188313
The company attempted to keep you = Provided complimentary minutes	0.175153
An issue occurred for you with the company = Internet service was disconnected	0.148716
The company made efforts to keep you = Provided additional internet megabits	0.148716
An issue occurred for you with the company = Payment plan complication	0.123320
An issue occurred for you with the company = The internet was exceptionally slow	0.123320
The company endeavored to keep you = Provided complimentary gigabits	0.123320

 ISSN 2723-6471

The company sought to retain you = Reduced the invoice amount	0.123320
The company attempted to keep you = Through special offers = By offers	0.122237
Age = 36-45	0.115637

Similarly, table 6 displays the attribute weights as determined by the COR Logistic Regression model. The pattern remains consistent, with "opting out of credit without utilization" again emerging as the most critical indicator of churn (weight = 0.1883), followed closely by "provided complimentary minutes" (0.1764) and "internet service was disconnected" (0.1487). These findings reinforce the importance of both service issues and retention strategies in influencing customer decisions.

Table 6. COR Logistic Regression

Attribute	Weight
An issue arose for you with the company = Opting out of the credit without utilization	0.188313
The company attempted to keep you = Provided complimentary minutes	0.176361
An issue occurred for you with the company = Internet service was disconnected	0.148713
The company made efforts to keep you = Provided additional internet megabits	0.148713
An issue occurred for you with the company = Payment plan complication	0.123322
An issue occurred for you with the company = The internet was exceptionally slow	0.123322
The company endeavored to keep you = Provided complimentary gigabits	0.123322
The company sought to retain you = Reduced the invoice amount	0.123322
The company attempted to keep you = Through special offers	0.120172
Age = 36-45	0.115638

By comparing the results across models, we observe a recurring trend: the attribute "withdrawal from credit without usage" holds the highest weight across all algorithms, signifying its dominant role in predicting customer churn. Additionally, offering free minutes ranks among the most effective strategies for customer retention. This consistency in attribute rankings across models strengthens the validity of the correlation analysis and enhances its applicability in real-world churn management scenarios.

Moreover, the analysis highlights some variations in attribute significance across models. In Naive Bayes, the duration of the customer-company relationship holds the least weight. In Logistic Regression, line type contributes minimally, while in Deep Learning, retention efforts show the lowest relative influence. These differences underscore the diverse interpretative strengths of each model and the importance of model selection in predictive analytics. Overall, the integration of sentiment and correlation analysis offers a comprehensive framework for interpreting UGC in churn prediction. It not only improves the understanding of customer behavior but also provides a data-driven basis for developing targeted and effective retention strategies.

6. Conclusion

This paper holds significant importance in the context of customer churn within the telecommunications sector, a crucial factor for companies aiming to enhance profitability. Customer churn significantly impacts revenue in the telecom industry, prompting the development of a model to analyze customer behavior and predict potential churn. Deep Learning, Naive Bayes, and Logistic Regression algorithms were utilized in this study to identify key factors influencing the churn process. The paper discusses the percentage and weights of comments utilized in classifying each algorithm. Additionally, correlation among these algorithms was explored to determine the attribute with the highest weight in decision-making. Furthermore, sentiment analysis was conducted on User-Generated Content (UGC) to evaluate and categorize customer opinions. By leveraging survey responses and comments, churned customers were successfully identified and directed to the customer retention department for personalized retention efforts, complete with their customer ID.

7. Declarations

7.1. Author Contributions

Conceptualization: D.S., A., and S.S.M.; Methodology: S.S.M.; Software: D.S.; Validation: D.S., S.S.M., and A.; Formal Analysis: D.S., S.S.M., and A.; Investigation: D.S.; Resources: S.S.M.; Data Curation: S.S.M.; Writing Original Draft Preparation: D.S., S.S.M., and A.; Writing Review and Editing: S.S.M., D.S., and A.; Visualization: D.S. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Bharambe, P. Deshmukh, P. Karanjawane, D. Chaudhari, and N. M. Ranjan, "Churn prediction in telecommunication industry," *in 2023 International Conference for Advancement in Technology (ICONAT)*, Goa, India, vol. 2023, no. Jan., pp. 1–5, 2023.

[2] N. I. A. Razak and M. H. Wahid, "Telecommunication customers churn prediction using machine learning," *in 2021 IEEE 15th Malaysia International Conference on Communication (MICC)*, Malaysia, vol. 2021, no. Dec., pp. 81–85, 2021.

[3] P. A. Salgueiro and H. S. Mamede, "Which factors influence the adoption of online self-service technologies by B2B customers of a telecom?," *in 2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, Chaves, Portugal, vol. 2021, no. Jun., pp. 1–6, 2021.

[4] X. Hu, Y. Yang, L. Chen, and S. Zhu, "Research on a customer churn combination prediction model based on decision tree and neural network," in *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, Chengdu, China, vol. 2020, no. Apr., pp. 129–132, 2020.

[5] W. N. Wassouf, R. Alkhatib, K. Salloum, and S. Balloul, "Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study*," J. Big Data,* vol. 7, no. Apr., pp. 1-20, 2020.

[6] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *in IEEE Access*, vol. 7, no. May, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[7] A. S. M. Alharbi and E. De Doncker, "Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information," *Cogn. Syst. Res.*, vol. 54, no. May, pp. 50–61, 2019.

[8] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *J. Bus. Res.*, vol. 94, no. Jan., pp. 290–301, 2019.

[9] F. Abdi and S. Abolmakarem, "Customer behavior mining framework (CBMF) using clustering and classification techniques," *J. Ind. Eng. Int.*, vol. 15, no. Suppl. 1, pp. 1–18, 2019.

[10] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, pp. 1-21, 2019.

[11] E. M. M. van der Heide, R. F. Veerkamp, M. L. van Pelt, C. Kamphuis, I. Athanasiadis, and B. J. Ducro, “Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle,” *J. Dairy Sci.*, vol. 102, no. 10, pp. 9409–9421, 2019.

[12] J. Vijaya and E. Sivasankar, “Computing efficient features using rough set theory combined with ensemble classification techniques to improve customer churn prediction in telecommunication sector,” *Computing*, vol. 100, no. 8, pp. 839–860, 2018.

[13] A. Ankit and N. Saleena, “An ensemble classification system for Twitter sentiment analysis,” *Procedia Comput. Sci.*, vol. 132, no. 1, pp. 937–946, 2018.

[14] R. Abdalla and M. Esmail, *WebGIS for Disaster Management and Emergency Response*. Cham, Switzerland: Springer, 2019. doi: 10.1007/978-3-030-03828-1.

[15] A. Gaur and R. Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques," *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, Bhopal, India, vol. 2018, no. Dec., pp. 1-5, 2018, doi: 10.1109/ICACAT.2018.8933783.

[16] M. A. Hossain, M. R. Chowdhury, N. Jahan, “Customer retention and telecommunications services in Bangladesh,” *Int. J. Asian Soc. Sci.*, vol. 7, no. 11, pp. 921–930, 2017.

[17] A. Idris and A. Khan, “Churn prediction system for telecom using filter-wrapper and ensemble classification,” *Comput. J.*, vol. 60, no. 3, pp. 410–430, 2017.

[18] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University - Engineering Sciences*, vol. 30, no. 4, pp. 330–338, Oct. 2018.

[19] J. Bughin, “Reaping the benefits of big data in telecom,” *J. Big Data*, vol. 3, no. 14, pp. 1-12, 2016.

[20] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, no. Jun., pp. 1–9, 2015.

[21] A. T. Jahromi, S. Stakhovych, and M. Ewing, “Managing B2B customer churn, retention, and profitability,” *Ind. Mark. Manag.*, vol. 43, no. 7, pp. 1258–1268, 2014.

[22] Y. Wei, H. B. Hashim, S. H. Lai, K. L. Chong, Y. F. Huang, A. N. Ahmed “Comparative analysis of artificial intelligence methods for streamflow forecasting,” *IEEE Access*, vol. 12, no. 1, pp. 10865–10885, 2024, doi: 10.1109/ACCESS.2024.3351754.