Synthetic Data Generation from Amazon Reviews for Supplements and Vitamins

1 Introduction

This report outlines the methodology for generating synthetic data from a dataset of Amazon reviews for supplements and vitamins products. The objective was to create a statistically similar dataset suitable for analytical purposes while ensuring privacy and data utility. The report details the steps followed, including data cleansing, the models and architectures used, and insights gained throughout the process.

2 Methodology

2.1 Data Cleansing

Before generating synthetic data, the original dataset underwent thorough cleaning using Google Colab. The cleaning process involved several key steps:

- **Loading the Dataset:** The dataset was imported from a CSV file into a pandas DataFrame.
- **Text Cleaning:** A custom function was created to remove non-alphanumeric characters from the title and text columns while retaining non-string values like NaN (not a number) to preserve data integrity.
- **Handling Numeric Columns:** The `helpful_vote` column was converted to a numeric format, filling NaN values with zero.
- **Boolean Conversion:** The `verified_purchase` column was transformed into boolean values (true/false).
- **Dropping Unnecessary Columns:** The `date`, `time`, and `timestamp` columns were removed to simplify the dataset.
- **Data Type Specification:** Each column was assigned the correct data type to ensure smooth processing in subsequent steps.
- **Saving the Cleaned Data:** The cleaned dataset was saved as a new CSV file for future use.

2.2 Generating Synthetic Data with Gretel API

Following the data cleaning, the next steps involved generating synthetic data using the Gretel API:

- **Repository Setup:** Necessary libraries were cloned and installed from the Gretel repository to prepare the working environment.
- **Anonymization Configuration:** An Anonymizer object was created, defining project parameters, run modes, and configuration files for data transformation and synthetic data generation.
- **NER Report Modification:** A modified function for the Named Entity Recognition (NER) report was established to manage potential errors and ensure accurate entity detection.
- **Anonymization Execution:** The anonymization process commenced by iterating through the cleaned dataset, utilizing the configured Anonymizer to create synthetic data.

2.3 Model/Architecture Used

The primary architecture employed for generating synthetic data was the Anyway Conditional Tabular Generative Adversarial Network (ACTGAN).

2.3.1 Explanation of ACTGAN

ACTGAN consists of two main components: the Generator and the Discriminator, which are trained to compete against each other.

- **Generator:** This component learns to produce synthetic data that resembles the original dataset. Starting with random weights, it learns to create realistic samples by understanding patterns in the training data.
- **Discriminator:** This part evaluates the authenticity of the data, distinguishing between real samples from the original dataset and synthetic samples produced by the Generator. Feedback from the Discriminator helps improve the Generator's output.

ACTGAN is an advanced version of the CTGAN model, incorporating enhancements that improve speed, accuracy, and memory efficiency. It adeptly handles both categorical and continuous data types, making it well-suited for generating tabular datasets.

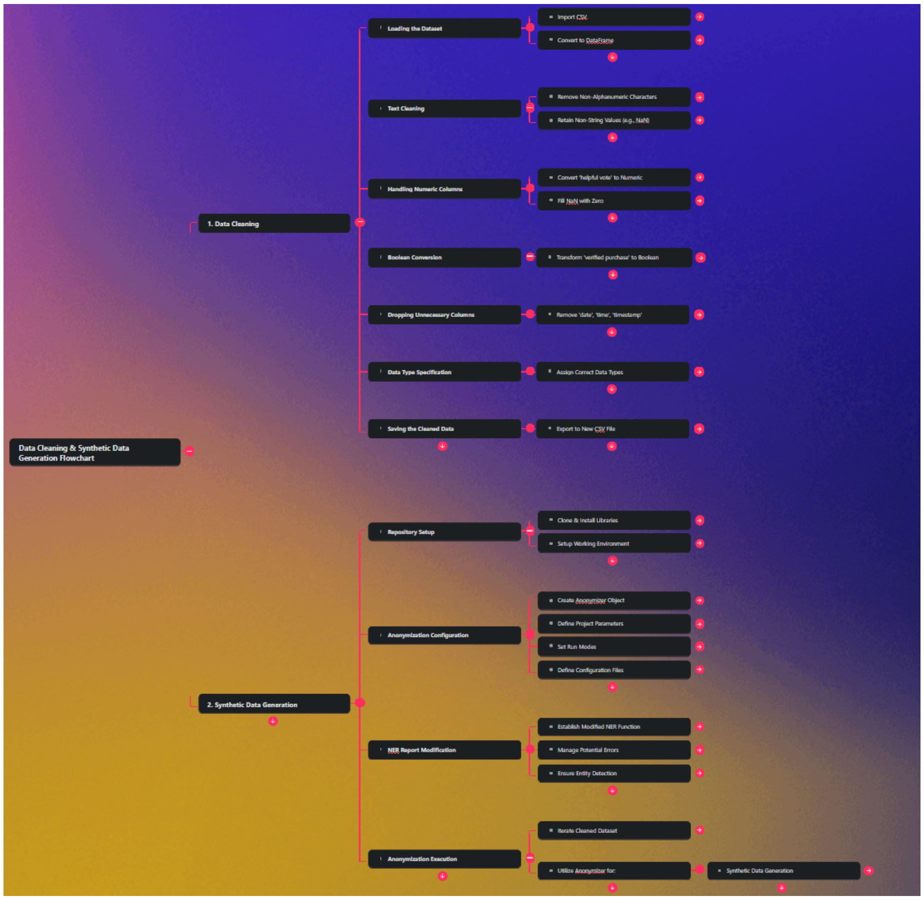
Data Cleaning & Synthetic Data Generation Flowchart
1. Data Cleaning
Loading the Dataset
Import CSV
Convert to DataFrame
Text Cleaning
Remove Non-Alphanumeric Characters
Retain Non-String Values (e.g., NaN)
Handling Numeric Columns
Convert 'helpful vote' to Numeric
Fill NaN with Zero
Boolean Conversion
Transform 'verified purchase' to Boolean
Dropping Unnecessary Columns
Remove 'date', 'time', 'timestamp'
Data Type Specification
Assign Correct Data Types
Saving the Cleaned Data
Export to New CSV File
2. Synthetic Data Generation
Repository Setup
Clone & Install Libraries
Setup Working Environment
Anonymization Configuration
Create Anonymizer Object
Define Project Parameters
Set Run Modes
Define Configuration Files
NER Report Modification
Establish Modified NER Function
Manage Potential Errors
Ensure Entity Detection
Anonymization Execution
Iterate Cleaned Dataset
Utilize Anonymizer for:
Synthetic Data Generation

Figure 1: Methodology Mind-map

2.3.2 Alternative Models Considered

While ACTGAN was the chosen model, several alternatives were evaluated:

- **CTGAN:** A predecessor to ACTGAN, offering a solid foundation for synthetic data generation but lacking some enhancements present in ACTGAN.
- **Variational Autoencoders (VAEs):** Effective for generating diverse data but less optimal for capturing the complexities of tabular data compared to GANs.

Research on various architectures aimed to assess their suitability in retaining the original dataset's statistical properties while minimizing privacy risks.

2.4 Factors Considered for Generating the Dataset

Several critical factors were considered during the synthetic dataset generation process:

- **Length of Reviews:** The character count of review texts was analyzed to ensure diversity in length, reflecting the variety found in user feedback.
- **Topic Diversity:** A range of topics was included in the reviews to represent different user experiences, enhancing the realism of the synthetic dataset.
- **Rating Distribution:** The distribution of ratings was examined to accurately represent user sentiment, balancing positive, neutral, and negative reviews.
- **Categorical Balance:** Discrete columns, such as verified purchase status, were analyzed to ensure proportional representation, further improving the dataset's realism.
- **Data Cleansing:** Text columns were cleaned to eliminate irrelevant characters, and columns were cast to their respective data types for consistency.

2.5 Measuring the Efficacy of a Synthetic Dataset

The efficacy of the synthetic dataset is evaluated using key metrics to ensure it mirrors the original data's characteristics.

- **Field Correlation Stability:** Measures the average absolute difference in correlations between fields in the training and synthetic datasets, with lower values indicating higher stability. Heatmaps illustrate these correlations, crucial for statistical analysis and machine learning applications.
- **Principal Component Analysis (PCA):** Compares the distributional distances of principal components from both datasets. Closer components result in higher synthetic quality scores, providing insights into the data's utility for machine learning.
- **Field Distribution Stability:** Uses Jensen-Shannon Distance to assess how well the distributions of numeric and categorical fields match. Lower average scores indicate higher stability, with visual aids like bar charts to enhance comparison. Maintaining distribution integrity is vital for the synthetic data's intended purpose.

Reports generated by Gretel provided insights into how well the synthetic dataset captured essential characteristics of the original data.

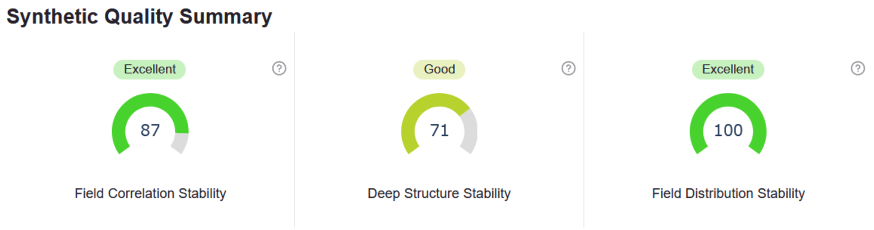
Synthetic Quality Summary
Excellent
87
Field Correlation Stability
Good
71
Deep Structure Stability
Excellent
100
Field Distribution Stability

Figure 2: Synthetic Data Quality Summary

2.6 Ensuring Synthetic Dataset is Inspired by the Source Dataset

To ensure that the synthetic dataset was inspired by the source dataset without being an exact copy, specific parameters in the ACTGAN model were fine-tuned:

- Mode-specific normalization was applied to manage continuous columns with complex distributions.
- Conditional generator mechanisms were employed to effectively balance categorical columns during training, ensuring diverse representation in the generated data.

During the review of the synthetic data, it was noted that some reviews shared similar structures. However, measures were taken to ensure that if one column's value was repeated, the values in other columns varied, maintaining overall sentiment while avoiding exact replicas.

2.7 Top Challenges in Solving the Problem Statement

Several challenges emerged during the synthetic data generation process:

- **Availability of LLMs and Resources:** Many large language models (LLMs) required paid subscriptions, while free models often struggled to allocate sufficient resources for effective training.
- **Generative Model Limitations:** Despite employing a GAN-based approach, the model required careful configuration to avoid overfitting and ensure that the generated data was sufficiently diverse.
- **Quality Control:** Maintaining high quality while achieving diversity in the synthetic data presented a continuous challenge, necessitating ongoing evaluation and adjustments during training.

2.8 Experiments Conducted

Key experiments included:

- **Model Training:** The ACTGAN model was trained on the cleaned dataset, with ongoing evaluations of its outputs.
- **Parameter Tuning:** Various hyperparameters were adjusted to optimize model performance, including batch size, learning rate, and architectural depth.
- **Evaluation Metrics:** Metrics were utilized to assess the quality of the synthetic data, including statistical tests for distribution similarity and evaluations of model performance.

2.9 Process Improvement

To enhance the synthetic data generation process further, the implementation of Large Language Model (LLM) APIs is recommended. This would involve prompting LLMs in a loop, where each iteration generates a new row of synthetic data based on the fine-tuned model. Utilizing various Langchain models allows for a more nuanced generation of synthetic data, potentially leading to greater variability and richness in the reviews. Each iteration can produce a complete review by sequentially generating different components, such as the review text, title, rating, and user ID, contributing to a more authentic representation of the original dataset.

3 Conclusion

The process of generating synthetic data from Amazon reviews for supplements and vitamins products was successfully executed, employing a structured approach that encompassed data cleansing, model training, and evaluation. The use of ACTGAN demonstrated significant promise in generating a dataset that retains essential statistical properties of the original data while ensuring user privacy. This synthetic dataset serves as a valuable resource for analytics and machine learning applications, enabling further insights into user experiences with supplements and vitamins.