National College of Ireland

# Decoding Online Pharmacy Trends: Clustering, Prediction and Business Insights

MSc Research Project
Data Analytics

Sneha Bhatgaonkar
Student ID: x22228136

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | | | |
|---|---|---|---|
| **Student Name:** | Sneha Bhatgaonkar | | |
| **Student ID:** | x22228136 | | |
| **Programme:** | Data Analytics | **Year:** | 2025 |
| **Module:** | Research Project | | |
| **Lecturer:** | Mohammed Hasanuzzaman | | |
| **Submission Due Date:** | 29/01/2025 | | |
| **Project Title:** | Decoding Online Pharmacy Trends: Clustering, Prediction and Business Insights | | |
| **Word Count:** | 6532 | | |
| **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:** | Sneha Bhatgaonkar |
| **Date:** | 29/01/2025 |

**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): | |

# Decoding Online Pharmacy Trends: Clustering, Prediction and Business Insights

Sneha Bhatgaonkar
x22228136

**Abstract**

The online pharmacy market in India is experiencing significant growth, driven by increasing consumer demand for convenience and accessibility in healthcare services. In this competitive market, it is essential to extract insights from available data, identify areas for improvement, and enhance business operations and services. This study aims to uncover patterns within online pharmacy data sourced from the Kaggle platform. Analysing biomedical text is challenging but domain-specific transformer-based models are effective in extracting relevant entities in such scenarios. To achieve this, named entity recognition (NER) models such as Med7 and Clinical-AI-Apollo (Medical-NER) are used to extract features. The data is analysed using K-Means clustering, and a classification model is built to predict product reviews using supervised machine learning models Random Forest, XGBoost, and Easy Ensemble classifiers. For vectorization, transformer-based models such as BioBERT, BioFormer-16L, and Clinical-AI-Apollo are used. Evaluation results show that the Easy Ensemble classifier with XGBoost estimator effectively handles class imbalance, and when combined with the Clinical-AI-Apollo model for vectorization, it outperforms other models in terms of ROC AUC performance. This work contributes to a deeper understanding of the data, providing business insights into identifying disease profiles, side-effects, medicine forms, composition and manufacturer etc. along with the review-based performance probability of products. These insights can inform better strategies to enhance underperforming products.

**Keywords**: clustering, NER, vectorization, classification

# 1 Introduction

The rapid growth of e-commerce, amid COVID-19 pandemic, has significantly reshaped industries worldwide, including the pharmacy sector. According to a recent Statista report, the online pharmacy market in India is projected to see user penetration increase to 6.74% by 2024, with expectations of reaching 10.04% within the next five years. Additionally, the annual revenue growth rate is forecasted to be 13.33% (Statista Research Department, 2024). These projections highlight the need to study pharmacy-related data in order to better understand user behavior and current market demands. Effective decision-making in this sector is crucial for business success, and such decisions rely heavily on trustworthy, contextually rich data. Unlike generic data, pharmacy data requires specialized knowledge to derive meaningful insights. The use of domain specific pre-trained models to analyse biomedical text can identify the patterns and trends within this data. These insights can be leveraged by businesses, healthcare professionals, and pharmacy teams to shape product offerings, improve services, and effectively respond to evolving market needs.

Prior to the development of domain-specific or context-aware models, some studies used generic BERT models to capture context from input data and perform text summarization for biomedical texts (Moradi *et al.* 2020). A large amount of data is produced

on daily basis, making it difficult to understand and extract new knowledge or trends without assistance. The knowledge graph technique used by Harnoune *et al*. (2021) was employed to extract structured information from biomedical clinical notes. This approach was tested on real-world clinical notes and performed well in Named-Entity Recognition (NER) tasks and relation extraction. Lee *et al*. (2020) introduced BioBERT, a model pre-trained on large biomedical corpora, which significantly outperformed previous state-of-the-art models in data mining tasks.

This study focuses on using domain-specific BERT models—BioBERT, Clinical-AI-Apollo, and BioFormer-16L, to generate embeddings for the biomedical text in the dataset. Cluster analysis using k-means is performed to group the data and facilitate a classification task that predicts reviews as either positive or negative. Random Forest, XGBoost, and Easy Ensemble classifiers are employed for this classification. Due to the imbalanced nature of the data, the performance of these classifiers is evaluated using ROC AUC, F1-score, precision, and recall. Integrating in-depth data analysis involving disease trends, side-effect profiles, product clustering, and reviews can provide a powerful framework for continuous product and brand improvement. The outcome of this approach will help businesses focus on high-demand areas by shifting resources toward improving or promoting better-performing products and considering the discontinuation or reformulation of underperforming products. This analysis can also help pharmaceutical companies better align research and development efforts to target high-impact areas that are in demand.

## 1.1 Research Questions

- How can unsupervised machine learning algorithms, combined with Named Entity Recognition (NER), extract meaningful patterns from online pharmacy data, and what actionable insights can these patterns provide for healthcare businesses?

- How can transformer-based vectorization models improve the performance of supervised machine learning classifiers in predicting binary medicine reviews from clustered pharmacy data?

## 1.2 Proposed Research Objectives

- Apply NER models (Clinical-AI-Apollo and Med7) to extract key entities such as diseases and side effects, followed by KMeans clustering to uncover meaningful patterns in online pharmacy data.

- Utilize transformer-based embeddings (BioBERT, Clinical-AI-Apollo, and BioFormer) with supervised classifiers (Random Forest, XGBoost, EasyEnsemble) to predict medicine review sentiment, aiming for high ROC AUC and F1-score performance.

The paper is organized as follows: Section 2 discusses related work on clustering, domain-specific NER models for entity extraction, techniques to handle imbalanced data, and the role of supervised models in classification tasks. Section 3 details the methodology used in this study. The design diagram for the implementation is presented in Section 4, followed by a discussion of model implementation in Section 5. Section 6 compares the performance of various models, and finally, Section 7 concludes the paper and suggests possible directions for future research.

# 2 Related Work

Understanding data and its attributes is important when performing classification or prediction tasks. Unsupervised techniques, such as clustering, can uncover the inherent groups or patterns in the data. This approach helps in learning more about structure of the data and how it can be used to generate insights (Naeem *et al*. 2023). Additionally, utilizing pre-trained models to extract information from domain specific text can improve the results of classification models. Studies carried out in the field of biomedical domain for text summarization, clustering using pre-trained models are discussed along with techniques to handle imbalanced data while performing classification tasks. Following section is divided into data mining using clustering, Named Entity Recognition, Vectorization, Evaluation metrics for imbalanced data and classification models.

## 2.1 Data Mining using clustering

Ashabi *et al.* (2020) conducted a systematic review on K-means and highlighted that K-means algorithm is essential to get overview of the data. Objective of K-means is grouping of data based on similarity and form cluster of coherent and homogeneous data points. After analysing 835 studies, K means is one of the most popular data mining techniques, used widely over a half century. Studies also mentioned about the limitations of K-means, this involves identifying k value, initial centroid, size of data and noise present in the data. Identifying K value for the data in dynamic environment is challenging. Choosing optimum K-value is important to get appropriate and meaningful clusters. This information plays important role for current study as K-means clustering will be used for data mining. Now, to identify cluster quality, the elbow method and silhouette are used in the study done by Saputra *et al.* (2020). This study showed distance metrics has less influence on identifying k value and we can use Elbow and Silhouette method to get optimum value of k. All the three distances Eucledian, Manhattan and Minkowaski were used in the analysis. When clusters are formed using Kmeans, Cluster cohesion and cluster separation values help to validate the quality of the cluster. 'Within cluster sum of square' (WSS) measures cluster cohesion and 'between cluster sum square'(BSS) measures cluster separation.

Bandyopadhyay *et al.* (2021) implemented the K-means clustering algorithm combined with Principal Component Analysis (PCA) for customer segmentation. This approach proved to be an effective solution for segmenting customers. For businesses aiming to maintain long-term sustainability and profitability, customer segmentation is essential. The results from the clusters formed through K-means and PCA aligned with the identified customer needs and preferences. By reducing the dimensionality of various features related to products and customers with PCA, meaningful patterns in customer behavior were observed. This approach successfully provided valuable insights that aligned with customer expectations, ultimately enhancing the business's ability to meet diverse customer needs. These studies are relevant to this research, as it highlights the effectiveness of using K-means with PCA to gain an overview of the dataset.

## 2.2 Named Entity Recognition

In this digital era, data is growing ubiquitously, especially in the healthcare domain. Data is available in various forms, including electronic health records, patient data, prescriptions, medicines purchased through online platforms, and more. Extracting meaningful information from these sources is a challenging task, as it requires domain-specific knowledge in medicine. Kormilitzin *et al*. (2021) developed and validated a named-entity recognition (NER) model called Med7 for clinical natural language processing. This model is useful for

performing analytical tasks and is trained to identify seven categories: drug names, route of administration, frequency, dosage, strength, form, and duration. As part of the current research, Med7 will be beneficial for identifying drug names, frequency, strength, and form information from the dataset. The model is freely available as a Python package for SpaCy.

The Hugging Face platform provides Clinical-AI-Apollo/Medical-NER, a model developed by training the DeBERTa V3 base model on the PubMed dataset. This model consists of 12 layers and a hidden size of 768. DeBERTa is a transformer-based neural language model built on top of BERT/RoBERTa. It uses a disentangled attention mechanism and an enhanced mask decoder. In DeBERTa, each word is represented using two vectors: a content vector and a positional vector. As a result, the attention weights provide importance to both content and position.

Clinical-AI-Apollo/Medical-NER is capable of identifying 41 medical entities, as listed in the model's config.json. These entities include ACTIVITY, ADMINISTRATION, AGE, BIOLOGICAL_ATTRIBUTE, BIOLOGICAL_STRUCTURE, CLINICAL_EVENT, COLOR, COREFERENCE, DATE, DETAILED_DESCRIPTION, DIAGNOSTIC_PROCEDURE, DISEASE_DISORDER, DOSAGE, DISTANCE, DURATION, FAMILY_HISTORY, FREQUENCY, HEIGHT, HISTORY, LAB_VALUE, MASS, MEDICATION, NONBIOLOGICAL_LOCATION, OCCUPATION, OTHER_ENTITY, PERSONAL_BACKGROUND, QUALITATIVE_CONCEPT, SEVERITY, SEX, SIGN_SYMPTOM, SUBJECT, TEXTURE, THERAPEUTIC_PROCEDURE, TIME, VOLUME, and WEIGHT.

This model is well-suited for the current research to extract relevant entities. Specifically, the DISEASE_DISORDER and SIGN_SYMPTOM entities will be especially useful for gaining insights from the current dataset.

## 2.3 Vectorization

In field of machine learning, while performing NLP related tasks irrespective of domain requires conversion of text data to numeric representation. To achieve this transformation, embeddings are required as model can understand the data in numeric form. Selva Berunda and Kanniga Devi (2021) have provided a review on Word embeddings which specifies it is a technique to represent the text in form of a vector in n-dimension space. Traditional word embedding, Static word embedding and Contextualized word embedding are dominant word embedding techniques. Traditional word embedding approach includes use of TF-IDF, Count Vector, Co-Occurrence Matrix. Most common approach is to use Bag of Words technique, Word2Vec, TFIDF. These are context independent word representation models. In advanced approach, transformer based pre-trained models such as ALBERT, BERT, DistilBERT, ELECTRA, RoBERTa are used which are encoder only models and provide contextualized word representation. Encoder models are well suited for Named Entity recognition, extractive question answering and sentence classification.

A study conducted to identify morbidity in clinical notes by Dessi *et al.* (2021) shows that the traditional machine learning model, Support Vector Machine (SVM) with TF-IDF, outperformed the deep learning models that used Word2Vec and GloVe embeddings. The performance was consistently strong across all 16 classes in the dataset. However, the authors note that the results may be biased and suggest that future studies should focus on improving data preprocessing to achieve more accurate and reliable outcomes.

Another study on the biomedical document labelling task, followed a three-step process - preprocessing, generating vector embeddings from pre-trained models, and applying Gaussian Mixture Modeling (GMM). The results indicated that the model using BioBERT embeddings achieved the best performance, followed by models with BioWordVec, GloVe, and Word2Vec embeddings. In contrast, TF-IDF with Autoencoder (AE) and TF-IDF with

Principal Component Analysis (PCA) produced the lowest scores. Text data transformation was performed using a pre-trained BioBERT model. To evaluate and compare model performance, several metrics were used, including the Fowlkes-Mallows score, silhouette coefficient, adjusted Rand index, and Davies-Bouldin score (Davagdorj *et al.* 2022). Another similar study carried out for biomedical document analysis showed BioBERT based K-means model achieved better FM score followed by AE-based TF-IDF and PCA based TF-IDF. TF-IDF based approach, cannot formulate relationship between words hence it is less informative when compared with BioBERT which is trained on large scale biomedical data.

BioBERT is the first domain-specific BERT based model developed by DMIS lab (Lee *et al*. 2020). This model is pretrained on biomedical corpora which includes PubMed abstracts and PMC full-text articles. Study mentions model outperformed BERT and other state-of-the-art models in text mining tasks, these tasks include biomedical named entity recognition, biomedical relation extraction and biomedical question answering. It was trained using eight NVIDIA V100 GPUs for 23 days. Analysis presented in the study shows model can efficiently understand complex biomedical text. Model can identify disease, Drug/Chem., Gene/Protein, Species entity types as part of named-entity recognition and Gene–disease, Protein–chemical as part of relation extraction.

BioBERT embeddings were successfully used by Das *et al*. (2020) to retrieve and extract information from scientific articles related to COVID-19. As per the original paper, BioBERT was trained on 29 million PubMed articles as of January 2019. However, there is a high probability that the pre-trained model lacks contextual information related to COVID-19. A study by Hebbar and Xie (2021) demonstrated that fine-tuning the base BioBERT model on COVID-19-related literature improved its performance. Sharaf and Anoop (2023) suggest that the responsible use of BioBERT in healthcare applications can improve clinical decision support, patient care, and healthcare processes.

Fang *et al.* (2023), implemented a compact BERT based model named BioFormer, that can perform four NLP tasks in biomedical domain - named entity recognition, relation extraction, question answering and document classification. The model is pre-trained using PubMed abstracts (as of Jan 2021) and PMC full-text articles. Model vocabulary formation was done with WordPiece technique to avoid out-of-vocabulary issue. Utilizing this model for the current study can help to identify the data and retrieve embeddings. These embeddings can be used with chosen supervised model to predict the reviews. Also, this model is 2X as fast as PubMedBERT.

Fang and Wang (2022) applied BioFormer to COVID-19 literature data for a multilabel classification task, comparing its performance with BioBERT and PubMedBERT models. The results demonstrated that BioFormer outperformed the other two models. The authors suggest that BioFormer's superior performance may be attributed to its release in 2021 and its inclusion of a COVID-19-specific corpus, whereas BioBERT and PubMedBERT were released prior to the pandemic and lacked this specialized data. From the perspective of current research, both BioBERT and BioFormer are suitable models for use.

## 2.4 Evaluation metrics for imbalanced dataset

In real-world medical applications, disease diagnosis and biological disorder data are often imbalanced. This data imbalance can significantly affect the performance of machine learning algorithms, as some models may become biased towards the majority class due to the skewed data distribution. Kumar *et al*. (2021) discussed various standard techniques to address the issue of highly imbalanced datasets. These techniques can be applied at both the data level and the algorithmic level.

At the data level, random oversampling and undersampling methods are commonly used to balance class distribution. Undersampling removes instances from the majority class

randomly to achieve balance, while oversampling replicates minority class instances to create a more uniform distribution. Oversampling can be achieved through methods like focused oversampling, synthetic sampling, random oversampling, and advanced techniques such as synthetic minority oversampling (SMOTE).

At the algorithmic level, hybrid methods like ensemble learning and classical methods such as thresholding, one-class classification, and cost-sensitive learning can be employed. These classifier-level techniques aim to improve learning from the minority class. For instance, in the threshold-moving technique, the decision boundary is adjusted based on a specified threshold, influencing how the model classifies the data. In ensemble learning, methods like Easy Ensemble and BalanceCascade are used to train and group classifiers on under-sampled subgroups. A survey paper presented by Susan and Kumar (2021) discussed applying balancing trick while working on an imbalanced dataset. In large datasets under sampling majority class can be considered as it requires low computation but, in the process, there is a loss of useful training data. On the other hand, oversampling minority class can balance the dataset at the cost of duplicated samples and might lead to overfitting. Choosing appropriate performance evaluation metrics plays a significant role in case of skewed class distributions. Area Under the Receiver Operating Characteristics Curve (AUC), G-mean, and the F1-score provide more information on the scores of the minority class. Also, Brownlee (2021) provides a comprehensive guide to choose right evaluation for the models in case of imbalance data. The author has suggested that when model is used to predict the probability and both classes are important, ROC AUC can be used to measure classifier performance.

A similar study conducted by Szeghalmy and Fazekas (2023) emphasized the importance of evaluating model performance using AUC and F1 values when handling imbalanced data. The study also highlighted that incorrect validation techniques can lead to misleading outcomes. To avoid this, the use of stratified k-fold cross-validation (SCV) is recommended. SCV splits the dataset into k folds while ensuring that each fold contains the same percentage of samples from both the minority and majority classes. These insights are valuable in building a solution for this study and guide the selection of appropriate data sampling techniques.

## 2.5 Classification Models

In real world data, class imbalance and high dimensionality is observed. Machine learning model cannot always interpret outcomes correctly in such scenarios and are majorly biased towards majority class. Pes (2021) conducted an experimental study on three datasets - image classification, text classification and text categorization to analyse hybrid learning strategies to overcome class imbalance and high dimensionality using Random Forest model. Hybrid learning with Random Under sampling and MinCost method obtained robust results. Another study for Alzheimer's disease Diagnostic classification by Song *et al*. (2021) mentioned RF performs well with high dimensional and highly correlated data. These conclusions are important for current study to use Random Forest as a classification model considering the high dimensionality.

The XGBoost model can handle large scale tasks efficiently and applied in various fields for classification and regression. Mixed views are observed on XGBoost performance when used with imbalanced dataset. Wang *et al*. (2020) implemented Imbalance-XGBoost by adding two components - weighted and focal losses to improve model performance in binary imbalanced classification. An experiment by Zhang *et al*. (2022) showed employing appropriate sampling at data level, and then using XGBoost at algorithm level improves the performance of the model in classification task in case of imbalanced data scenario.

A study conducted by Wang *et al*. (2022) aimed at classifying heartbeats using the Easy Ensemble technique suggested that this model could be effective for classification tasks

involving imbalanced datasets. The study highlighted that both sampling and class-weight techniques can be used to address data imbalance. Using class-weight adjustments along with the Easy Ensemble Classifier yielded the best performance in heartbeat classification, particularly improving the performance of the minority class. This approach could provide meaningful results in the current study as well.

Given the high dimensionality and class imbalance in current study, Random Forest, XGBoost, and the Easy Ensemble classifier are considered while building a solution for predicting product reviews using machine learning algorithms.

# 3 Research Methodology

This study follows the core stages of the CRISP-DM methodology to perform data mining and binary review classification for pharmaceutical products in the dataset. The 7-stage approach is presented in Figure 1, followed by a detailed description of each stage.

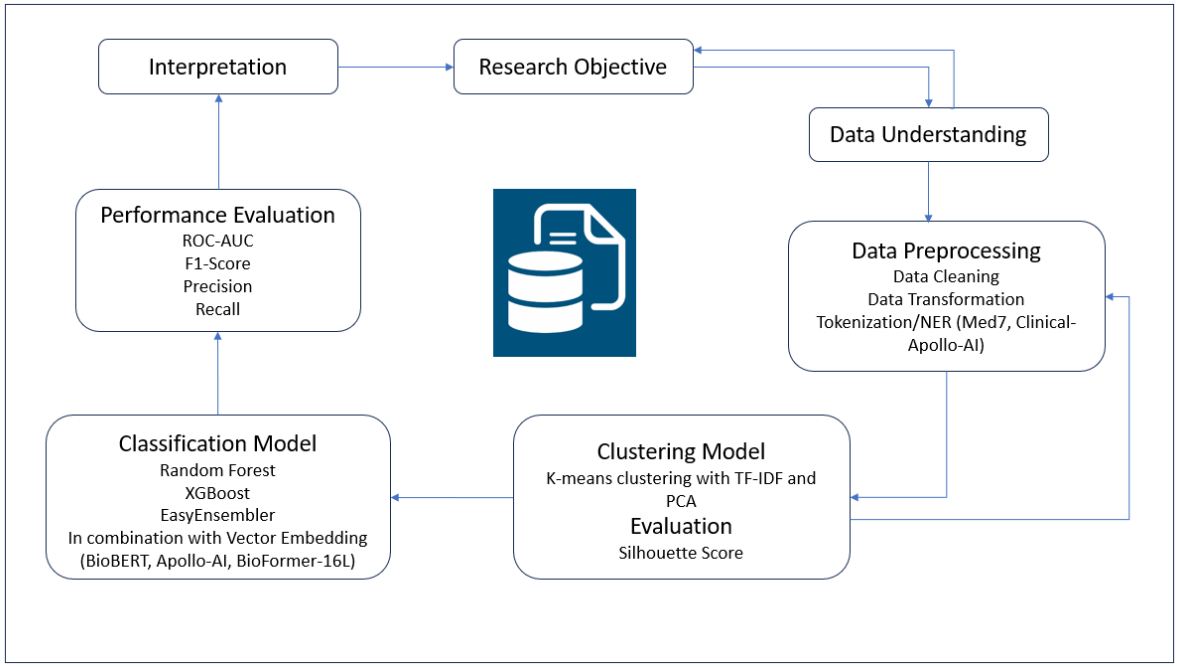



Figure 1: CRISP -DM Methodology

## 3.1 Research Understanding

The rapid expansion of the e-commerce sector across various domains has increased the need for data-driven analysis to enable informed decision-making. This study focuses on leveraging machine learning algorithms to analyse online pharmacy data. The primary objective of this research is to extract valuable insights from e-commerce pharmacy data by applying clustering techniques, named-entity recognition and supervised machine learning algorithms to classify product reviews. By analysing consumer feedback on medicinal products sold by online pharmacies, this study aims to provide businesses with a deeper understanding of product performance and consumer preferences, ultimately supporting strategic decision-making.

## 3.2 Dataset Understanding

The dataset used in this study is taken from Kaggle named 11000 Medicine details authored by Singh (2024), it is web scrapped from 1mg India's leading service provider in online

pharmacy domain. This dataset has 11825 records and carries information on medicine name, composition, uses, manufacturing information, side-effect and reviews.

## 3.3 Data Pre-processing

Data preprocessing involves cleaning, separating, and transforming data to make it more meaningful. Following key steps are taken during pre-processing:

**3.3.1 Data Cleaning:** As part of the data cleaning process, standard procedures such as checking for null and NaN values, identifying duplicate records, and removing them are performed. Post this activity, dataset has 11741 records.

**3.3.2 Data Processing:** The dataset primarily contains text data, the stopwords library from nltk.corpus is used to remove commonly used words in the English language.

**3.3.3 Feature Extraction:** The data includes information on medicinal components, which are presented as a combination of the component name and its amount (in mg, ml, etc.). This data was separated using appropriate regular expression techniques to identify the different components within the dataset. It was found that there were 1,065 unique component names.

**3.3.4 Review Label:** For the reviews, a weighted average technique was applied to the review columns to generate an overall review score. This score was then converted into positive and negative labels based on a predefined threshold. If the score was greater than or equal to three, it was labelled as positive; otherwise, it was labelled as negative.

## 3.4 Named Entity Recognition

Since the current dataset is from the healthcare domain, general domain pretrained models cannot be directly applied. Therefore, as mentioned in Section 2, various transformations were performed to effectively understand the underlying data, using named-entity recognition (NER) techniques. The Med7 library was employed to extract and separate information from the "Medicine Name" column. The DRUG and FORM entities returned by Med7 were used to create new columns. It was found that there are approximately 111 unique forms in the dataset (e.g., tablet, capsule, cream, injection, eyedrop, etc.). In some instances, where Med7 did not correctly identify entities, the data was manually cleaned by removing unrelated words.

Clinical-AI-Apollo/Medical-NER model from Hugging face is used as part of Named-Entity Recognition task. Various entities return by the model after inputting Uses and Side-Effects data. Considering the amount of data, processing time required to retrieve the entities was comparatively high nearly two to three hrs. Model returned tokenized data in json format which was analysed by creating a reusable function, which stores data in separate columns as per entities. This approach improved readability of data and cross evaluation. Data tagged with entities (B-DISEASE_DISORDER, I-DISEASE_DISORDER, B-SIGN_SYMPTOM, I-SIGN_SYMPTOM, B-DETAILED_DESCRIPTION, I-DETAILED_DESCRIPTION, B-DIAGNOSTIC_PROCEDURE, I-DIAGNOSTIC_PROCEDURE, B-OUTCOME) helped to generate more meaningful and clutter-free data.

A word cloud generated from the cleaned data highlights the presence of diseases found in the dataset, with high-frequency terms such as diabetes, blood pressure, and hypertension. Nausea, vomiting, and headache are seen as common side effects in present data. The medicine compositions show a strong presence of Glimepiride, Metformin, and

Telmisartan, which are primarily used to treat high blood sugar levels in type 2 diabetes and high blood pressure, respectively. Paracetamol, commonly used as a painkiller, is also frequently mentioned. The dataset includes information on the manufacturers for each medicine, and the word cloud reveals that there are many market players involved in pharma industry, including Sun Pharma, Intas, and Cipla etc.

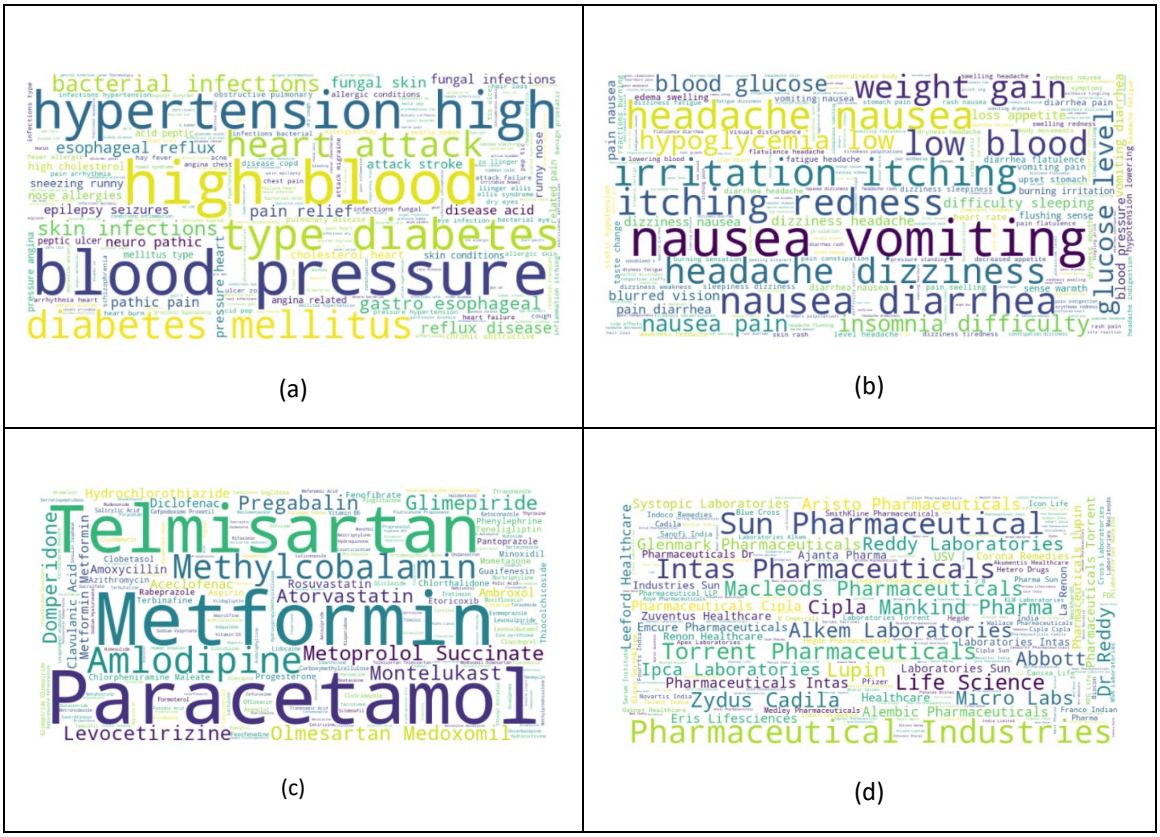



Figure 2: Word cloud representation
(a) Disease Profile (b) Side-Effects (c) Medicine components (d) Various Manufacturers

## 3.5 Clustering and Classification Models

To analyse data in an efficient way, related studies have shown K-means clustering can be performed, and it is an effective approach to get insights. To perform the classification task, Random Forest, XGBoost, and Easy Ensemble machine learning algorithms are used. Previous studies show that Random Forest is a widely adopted and powerful algorithm, capable of handling complex classification tasks with high accuracy. As mentioned in Section 2, XGBoost and Easy Ensemble classifiers are excellent options when dealing with data imbalance.

### 3.5.1 K-Means Clustering with PCA

Studies have shown that clustering has played an outstanding role in the field of marketing, medicine, biology and bioinformatics. As mentioned in section 2, PCA-based K-means clustering is effective at uncovering patterns within data and performs well on data with reduced dimensions, minimizing information loss.

### 3.5.2 Random Forest

The Random Forest algorithm is an ensemble of multiple decision trees. Each tree operates independently, with data sampled using the bootstrap technique. Randomness is introduced through feature bagging to ensure diversity among the trees. For classification tasks, the majority voting technique is applied. The ensemble learning nature of Random Forest is one of the key factors that boosts its performance across various tasks. It also includes the class_weight parameter, which automatically adjusts the weights of different classes.

### 3.5.3 XGBoost

XGBoost is considered a powerful and popular algorithm, especially in Kaggle competitions. It is a sparsity-aware algorithm and includes built-in regularization to handle overfitting (Chen and Guestrin, 2016). The model can be fine-tuned using hyperparameters such as learning rate, n_estimators, gamma, and max_depth. As mentioned in Section 2, XGBoost can deliver good performance in the case of imbalanced data, provided appropriate sampling is done at the data level.

### 3.5.4 Easy Ensemble Classifier (EEC):

The Easy Ensemble Classifier (EEC) is considered an advanced ensemble learning technique, used for classification tasks to address class imbalance. It reduces the bias towards the majority class by utilizing ensemble outcomes generated from balanced data samples. The EEC employs the AdaBoost algorithm as a boosting technique to improve classification results.

## 3.6 Evaluation Metrics

Model performance is assessed using confusion matrix and various other appropriate performance parameters. The confusion matrix provides information on the number of samples classified by the model into true negatives (TN), false negatives (FN), true positives (TP), and false positives (FP), giving an overall view of the model's performance. As discussed in Section 2, previous studies have shown that when evaluating the performance of imbalanced data, metrics such as ROC_AUC, F1-Score, Precision, and Recall are crucial. The ROC_AUC score indicates the model's ability to distinguish between positive and negative classes; a higher score reflects better performance. The F1-Score, which is the harmonic mean of precision and recall, provides an overall measure of how well the model handles both positive and negative samples. Recall, or the true positive rate, measures the model's ability to correctly identify actual positives. In this study, minimizing the false negative count is particularly important to effectively identify underperforming products. Therefore, the ROC_AUC, F1-Score, and Recall scores will be used to determine the best-performing models.

# 4 Design Specification

The designed architecture aims to perform efficient data analysis on biomedical texts and predict the outcome of medicine reviews. The end-to-end process involves data pre-processing, vectorization, clustering, classification, and evaluation stages. The solution is built using Jupyter Notebook, Python programming language and various libraries to achieve optimal performance in analysing biomedical data. The process begins with data pre-processing, where the dataset, stored in CSV format, is loaded using Pandas. Text cleaning using NLTK, tokenization and Named entity recognition (NER) are performed using

ApolloAI Medical NER and Med7 libraries, which are specialized for biomedical texts. The cleaned data is then vectorized using TF-IDF to convert text into a numerical representation, followed by PCA (Principal Component Analysis) for dimensionality reduction, ensuring computational efficiency. The next step involves K-Means clustering to group data, which helps in structuring the data for classification. Various machine learning models, including Random Forest, XGBoost and Easy Ensemble Classifier are trained by using various vectorizers namely Bio-BERT, Apollo-AI, BioFormer-16L. Each cluster is then used to predict the review labels. Model performance is evaluated using metrics such as ROC-AUC, F1-Score, Precision, and Recall, ensuring a balanced approach to classification. The best model is selected based on its ability to minimize misclassification, particularly in the negative class, ensuring robust and reliable predictions.

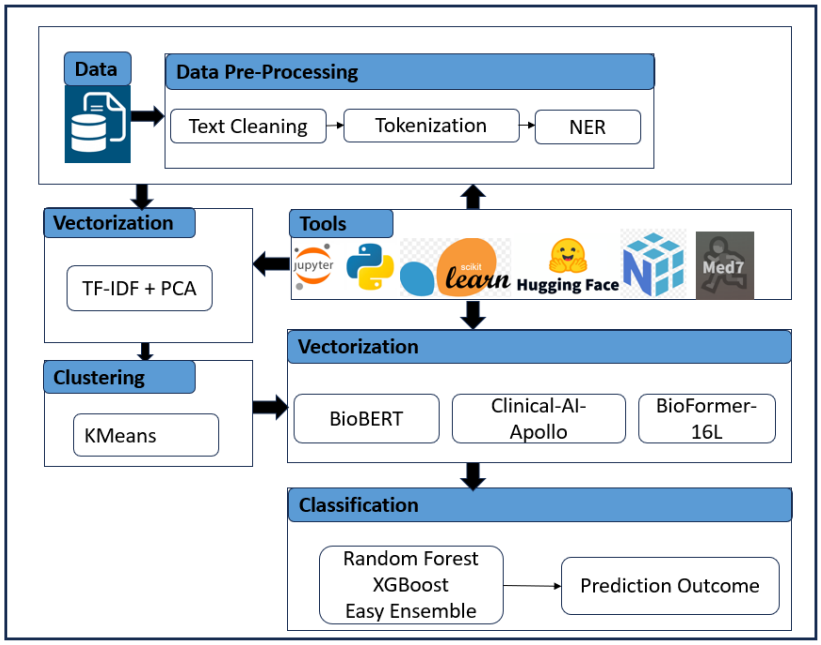



Figure 3: Design Architecture

# 5 Implementation

This section provides a detailed description of the architecture implementation. It outlines each step taken to achieve the business objective, explaining how various techniques were used to transform the data into a structured format, predict reviews for pharmaceutical products, and gain insights into their performance.

## 5.1 Data Pre-processing

1mg online pharmacy data from Kaggle used in this analysis, after data pre-processing activities, the dataset now contains 11741 records. Detailed data pre-processing activities are covered in Section 3.3

## 5.2 Data Analysis with K-Means Clustering

The initial step in understanding the data and its relationship with other variables involves the use of an unsupervised learning algorithm, K-means clustering. On pre-processed data, word embeddings are generated using TF-IDF, and after vectorization, the data has dimensions of 11,741 by 2,243. To reduce the dimensions while retaining maximum information, Principal Component Analysis (PCA) is applied. The Silhouette score, along with the Elbow plot, was

used to assess the quality of the clusters. The highest Silhouette score of 0.79 was observed for a cluster value of 3, but the resulting clusters were not very informative. For a cluster value of 7, the Silhouette score was 0.65. In this, the data separation was clearer, and the grouping of diseases appeared more structured. When the clusters were plotted, the distinction and grouping of data points were more informative and visually evident. Therefore, a k value of 7 was chosen for further analysis, as it provided a well-structured clustering result. The figures below show the Elbow plot and cluster formation.

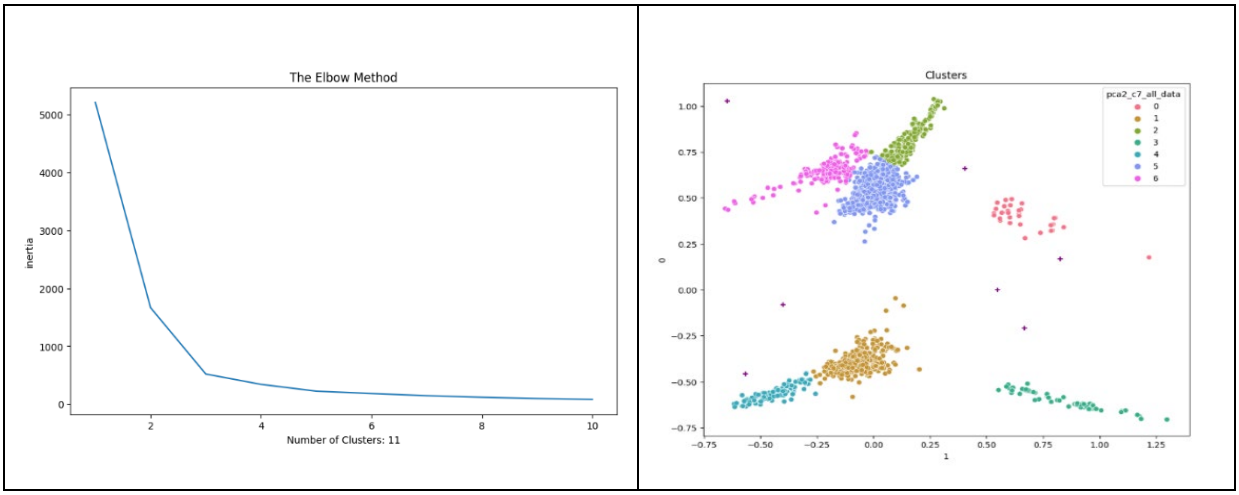
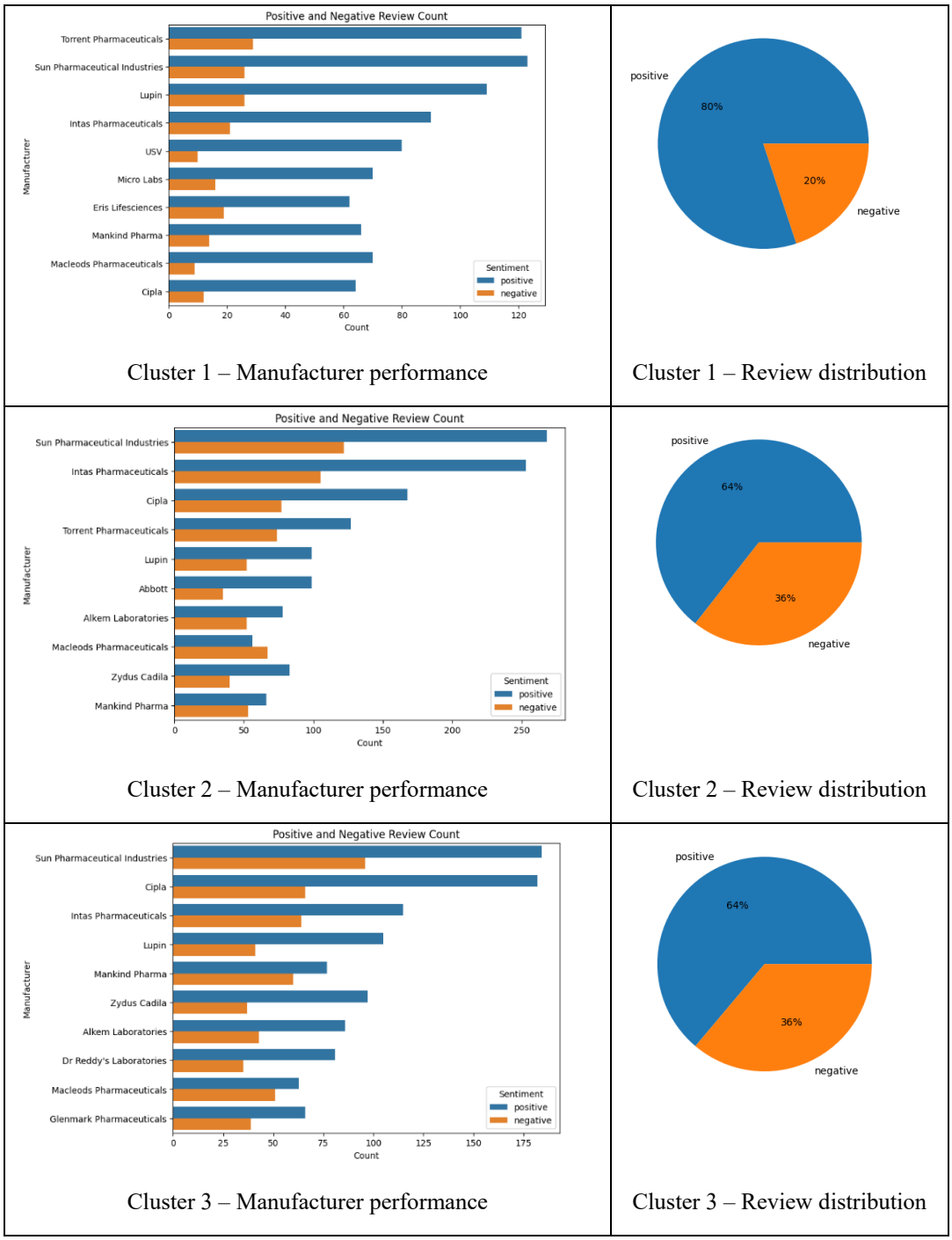



Figure 4: Elbow Plot and Cluster Visualization

Upon analysing each cluster, it was observed that diseases with similar characteristics were grouped together. This allowed for filtering the data based on clusters, which was then used for review classification. K-means cluster with labels 0, 3, and 4 were combined to form a dataset of 2,379 instances, which included medicine for diabetes and blood pressure, this will be referred as Cluster 1. Cluster with label value 1 contained 4,535 instances related to pain relief, allergies, and other medicinal uses, this will be referred as Cluster 2. A third data group, consisting of 4,827 instances, was created by combining label values 2, 5, and 6, which included medicines for various infections, this will be referred as Cluster 3. The Gaussian Mixture Model was also applied for cluster formation, yielding similar results.

After forming the clusters, the data was analysed based on positive and negative reviews within each cluster to identify the top medicine manufacturers. In the cluster containing diabetes and blood pressure instances (Cluster 1), Torrent, Sun, and Lupin were identified as the top manufacturers. In the cluster containing pain-related instances (Cluster 2), as well as those related to allergies, cough, and gastrointestinal issues (Cluster 3), Sun, Intas, and Cipla were identified as the top manufacturers. Cluster-wise target variable distribution check shows cluster 1 has data imbalance, whereas other two clusters show mild data imbalance. This data distribution is presented in Fig 5.




Cluster 1 – Manufacturer performance

Cluster 1 – Review distribution

Cluster 2 – Manufacturer performance

Cluster 2 – Review distribution

Cluster 3 – Manufacturer performance

Cluster 3 – Review distribution

Figure 5: Cluster wise Manufacturer distribution w.r.t. review labels and Review distribution

## 5.3 Train-Test Split

Train and test split is done using sklearn - train_test_split as 80-20 for each cluster group of data. Stratified train-test split is considered by adding parameter stratify to maintain same proportions of target samples as that of main dataset.

## 5.4 Vectorization

While building model with Random Forest, XGBoost and Easy Ensemble classifier – Three different transformer-based models are used to convert text data to vectors. Context based

vectorization is applied to columns disease, side-effects and component. Vectorization is obtained by using a reusable function which captures last hidden state of specified model. Model configurations are explained in detail below.

### 5.4.1 BioBERT Vectorization

Huggingface platform is used to get the BioBERT model (dmis-lab/biobert-v1.1) and tokenizer. Embeddings are obtained by using last hidden layer of the model. As mentioned in model configuration, BioBERT generates vector with 768 dimensions. When this model is applied to get vector representation, it is observed that computation time required to generate vector presentation is comparatively more than BioFormer-16. This observation is in line with the related studies mentioned in section 2.

### 5.4.2 Clinical-AI-Apollo Vectorization

Clinical-AI-Apollo model (Clinical-AI-Apollo/Medical-NER) details are obtained from Hugging face platform. Last hidden layer is used to get embeddings. This model produces vector with size of 768.

### 5.4.3 Bio-Former-16L Vectorization

Bioformer provides two variants one with 8L and 16L where 8L indicates 8 Hidden layers and 16L indicates 16 hidden layers are present in the model. For current study, 16L (bioformers/bioformer-16L) variant is used, it produces vector with size 384. This model is faster than above two models when used to generate the embeddings for current dataset.

### 5.4.4 TF-IDF Vectorization

Text data, including medicine name, manufacturer information, and type of medicine, are converted into vectors using TF-IDF vectorization.

### 5.4.5 Data concatenation

After applying vectorization to train and test, data is concatenated using numPy to create combined vector X_train and X_test.

## 5.5 Classification

Each cluster undergoes a comprehensive analysis using Random Forest, XGBoost, and Easy Ensemble Classifier with vectorization. To evaluate model performance, Stratified 5-fold cross-validation is initially conducted. Predictions are then made on the test samples, revealing that both Random Forest and XGBoost models tend to favor the majority class. The Easy Ensemble Classifier is constructed with XGB as the base estimator and configured with the objective set to ‘binary:logistic’ and a max_delta_step of 1. This model looks promising to handle majority and minority class.

As outlined in Section 2, an oversampling technique is used to address the data imbalance. In addition, the threshold-moving technique is employed to identify the optimal threshold, with model results recorded accordingly. For XGB classifier, n_estimators=100, objective='binary:logistic', random_state=42 parameters are used. The goal is to minimize both False Positives and False Negatives, ensuring the model accurately predicts products with true positive and true negative reviews. In this study, while both classes are equally important, the primary focus is on identifying products with negative reviews. Hence model capable to identify more count of True negatives is beneficial along with True positives. Model achieving the criteria would be considered as better model along with the evaluation metrics score.

# 6 Evaluation

This section provides detailed evaluation on each model designed with vectorization and classification. As mentioned in section 3.6, designed model outcomes are evaluated based on ROC AUC, F1-Score, precision, recall and accuracy.

## 6.1 Classification with BioBERT vectorization

Following table shows, Cluster-wise performance of Random Forest, XGB and Easy Ensemble classifier with BioBERT embeddings. Though XGB performance looks good for cluster two, it is biased towards majority class and misclassification was more for negative reviews. Easy Ensemble has provided more balanced outcome across all three clusters with ROC_AUC values as 0.58 and 0.59.

Table 1: Classification results with Bio-BERT

| Bio-BERT Vectorization | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Random Forest | | | XGB | | | Easy Ensemble with XGB | | |
| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 1 | Cluster 2 | Cluster 3 |
| **ROC_AUC** | 0.57 | 0.57 | 0.59 | 0.55 | 0.59 | 0.59 | **0.59** | **0.58** | **0.59** |
| **F1-Score** | 0.72 | 0.62 | 0.61 | 0.86 | 0.68 | 0.62 | 0.68 | 0.64 | 0.64 |
| **Precision** | 0.84 | 0.71 | 0.73 | 0.82 | 0.72 | 0.72 | 0.86 | 0.71 | 0.72 |
| **Recall** | 0.63 | 0.54 | 0.52 | 0.9 | 0.64 | 0.55 | 0.56 | 0.58 | 0.57 |
| **Accuracy** | 0.61 | 0.56 | 0.57 | 0.76 | 0.61 | 0.58 | 0.58 | 0.58 | 0.58 |

Confusion matrix for all three clusters using EasyEnsemble with XGBoost as base estimator is shown below, it clearly shows, model can identify positive and negative labels with good count of true positive and negatives.

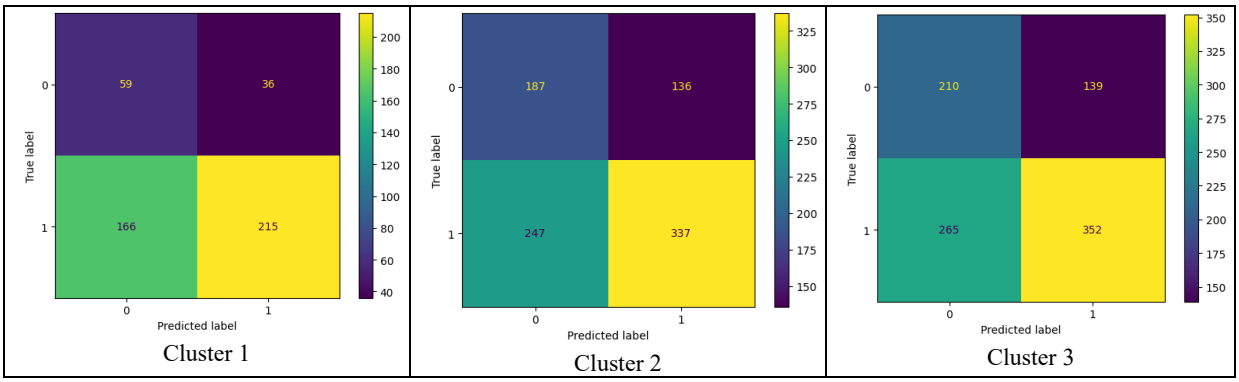



Figure 6: Confusion Metric - Easy Ensemble with XGB model + Bio-BERT

## 6.2 Classification with BioFormer-16L vectorization

As shown in following table, for all the three clusters Easy Ensemble classifier output for ROC-AUC ranges 0.58-0.6 which is better than Random Forest and XGB.

For cluster 1 - Easy Ensemble classifier with XGB as base estimator with parameters as (objective="binary:logistic", random_state=43, max_delta_step=1) tuned out better performer.

For cluster 2 and 3 Easy Ensemble classifier with XGB as base estimator and without XGB showed similar results.

Table 2: Classification results with BioFormer-16L

| BioFormer-16L Vectorization | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Random Forest | | | XGB | | | Easy Ensemble with XGB | | |
| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 1 | Cluster 2 | Cluster 3 |
| **ROC_AUC** | 0.57 | 0.58 | 0.6 | 0.56 | 0.57 | 0.59 | **0.6** | **0.58** | **0.59** |
| **F1-Score** | 0.77 | 0.69 | 0.67 | 0.85 | 0.58 | 0.63 | 0.68 | 0.65 | 0.65 |
| **Precision** | 0.83 | 0.7 | 0.72 | 0.82 | 0.72 | 0.73 | 0.86 | 0.72 | 0.72 |
| **Recall** | 0.72 | 0.68 | 0.63 | 0.88 | 0.48 | 0.55 | 0.56 | 0.59 | 0.59 |
| **Accuracy** | 0.66 | 0.61 | 0.6 | 0.75 | 0.55 | 0.58 | 0.57 | 0.59 | 0.59 |

When comparing the results with previous models, the confusion matrix shows that this model can identify a greater number of true positives and true negatives.

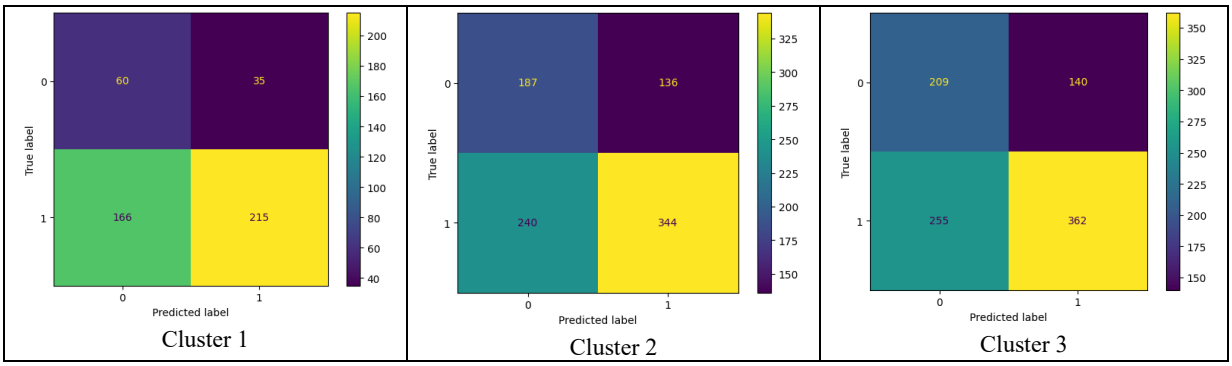



Figure 7: Confusion Metric - Easy Ensemble with XGB model + BioFormer-16L

## 6.3 Classification with Clinical-AI-Apollo vectorization

As shown in the following table, the Easy Ensemble classifier achieves an ROC-AUC score of 0.6 for all three clusters, outperforming the other models when configured with XGBoost as the estimator.

Table 3: Classification results with Clinical-AI-Apollo

| Clinical-AI-Apollo Vectorization | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Random Forest | | | XGB | | | Easy Ensemble with XGB | | |
| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 1 | Cluster 2 | Cluster 3 |
| ROC_AUC | 0.58 | 0.58 | 0.6 | 0.56 | 0.59 | 0.58 | **0.6** | **0.6** | **0.6** |
| F1-Score | 0.76 | 0.69 | 0.67 | 0.82 | 0.68 | 0.72 | 0.69 | 0.66 | 0.66 |
| Precision | 0.84 | 0.71 | 0.72 | 0.82 | 0.72 | 0.69 | 0.86 | 0.73 | 0.73 |
| Recall | 0.7 | 0.68 | 0.62 | 0.82 | 0.65 | 0.75 | 0.57 | 0.61 | 0.6 |
| Accuracy | 0.65 | 0.61 | 0.6 | 0.71 | 0.61 | 0.62 | 0.58 | 0.6 | 0.6 |

The confusion matrix for each cluster indicates that this model effectively identifies true positives and true negatives while minimizing false counts, in comparison to the other vectorized models discussed in Sections 6.1 and 6.2.

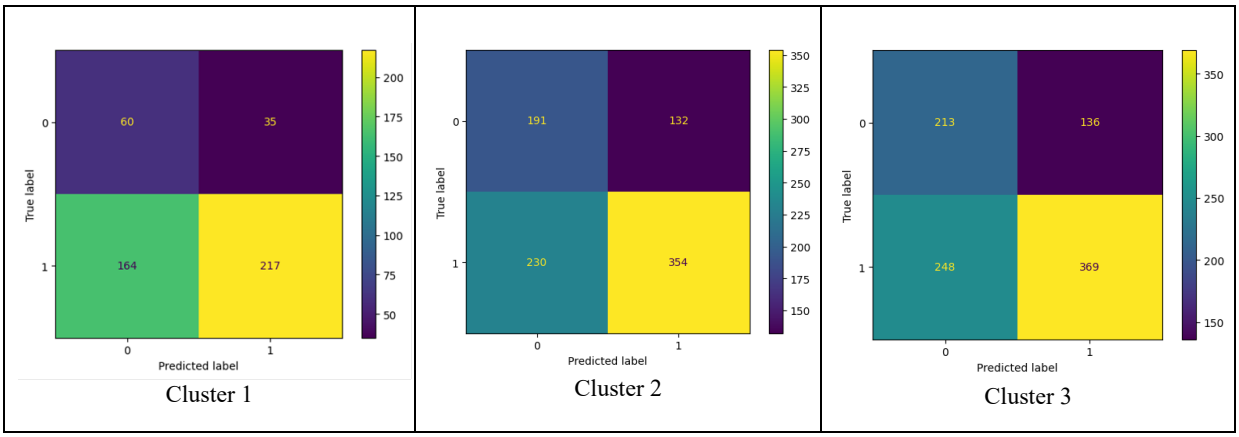



Figure 8: Confusion Metric - Easy Ensemble with XGB model + Clinical-AI-Apollo

## 6.4 Results & Discussion

The initial analysis of online pharmacy data using the K-Means clustering algorithm has provided valuable insights into the structure of the data and acted as groundwork for more advanced predictive modelling. Clustering identified underlying patterns and relationships, which proved essential in developing an effective model for predicting product reviews. This clustering approach, as highlighted in prior studies, helped guide the feature extraction and visualization. One of the key advancements in this study was extending clustering outcomes and integrate them with context-aware transformer models to predict product reviews. The use of advanced models such as BioBERT, BioFormer-16, and Clinical-AI-Apollo as part of the vectorization process demonstrated a clear performance improvement as we moved from one model to the next. This progression suggests that incorporating more domain-specific pre-trained models can enhance the model's ability to understand the medical context and generate more accurate predictions.

In terms of classification, experiments with Random Forest and XGBoost yielded promising results, but also revealed inherent biases towards the majority class. Both models showed a tendency to over-predict the majority class, which led to a higher number of false negatives, especially in predicting the minority class. This issue was significant in the context of the data imbalance. Easy Ensemble Classifier with XGBoost as the estimator, showed notable improvement in handling the class imbalance. The Easy Ensemble technique, despite its higher computational cost, effectively balanced the predictions between the majority and minority classes, resulting in better overall performance.

The Easy Ensemble Classifier (EEC) performed better in this research due to its ability to address class imbalance effectively. XGBoost, which is an advanced implementation of gradient boosting, improves upon the mistakes made by previous trees in a sequential manner, enhancing model accuracy over iterations. EEC, available in the imbalanced-learn library, is an ensemble technique that creates balanced samples by selecting all examples from the minority class and a subset from the majority class. This approach ensures that the model is trained on more balanced data, improving its performance with imbalanced datasets. The results were further improved when combined with the Clinical-AI-Apollo vectorization, which demonstrated superior ability in identifying medical entities. This suggests that using a vectorization method that better captures the relevant features in medical data can contribute significantly to model performance. The combination of EEC with XGBoost and Clinical-AI-Apollo vectorization collectively contributed to the improved results by increasing the identification of true positives (TP) and true negatives (TN), leading to a more robust and accurate model.

The presence of high-cardinality features in the dataset is another challenge. Features with a large number of unique values can hamper the model's ability to identify meaningful patterns, especially when these values occur infrequently. This sparsity can lead to overfitting or underfitting, as the model may struggle to generalize from such rare occurrences. In the case of text data, this issue becomes even more pronounced, as models may have difficulty distinguishing between rarely occurring words or terms.

A combinational approach to analysing online pharmacy data can uncover disease profiles within current populations, providing valuable insights for manufacturers to tailor their medicine production to meet present healthcare needs. Cluster-based analysis provides information on medicines, diseases, manufacturers, side effects, and product reviews, which can be valuable for healthcare professionals when prescribing treatments. This analysis also helps identify various forms, components and dosages of medicines used for different treatments. The model's outcomes can be analysed to identify underperforming products (TN), allowing manufacturers to focus on improving quality, modifying compositions, or enhancing formulations to better meet patient needs.

# 7 Conclusion and Future Work

The increase in the usage of online platforms to order medicine highlights the need to study product performance across various product categories. In this study, data from a leading online pharmaceutical company is used to classify product reviews. Using clustering techniques, data is categorized into three distinct clusters. To process the biomedical text, context-aware transformer-based models, namely BioBERT, Clinical-Apollo, and BioFormer, are used which effectively vectorized the text data. These vectorized representations were then used as inputs to supervised machine learning models, including Random Forest, XGBoost, and Easy Ensemble Classifier. The Easy Ensemble classifier with XGBoost estimator and Clinical-AI-Apollo vectorization, performed well in classifying reviews when compared with other models, especially in identifying non-performing products. By maintaining a low misclassification rate for negative reviews, the model is suitable to provide valuable insights to business, enabling the identification of underperforming products and informing strategic decisions.

Current study provides valuable insights into the classification of product reviews and the identification of non-performing products, in future, work can be expanded by adding the dataset with more samples, which can enhance the model's performance. Additionally, fine-tuning pre-trained transformer models with data and using deep neural networks could lead to even more accurate results. There is a scope to implement recommendation system by combining current data with consumer transaction-related data. Current work can also be used to create knowledge base for chat-bot systems.

# References

Ashabi, A., Sahibuddin, S. B. and Salkhordeh Haghighi, M. (2020) 'The systematic review of K-means clustering algorithm', in *ICNCC'20 Proceedings of the 2020 9th International Conference on Networks, Communication and Computing*. Tokyo, Japan, 18-20 December 2020, pp. 13-18. https://doi.org/10.1145/3447654.3447657

Bandyopadhyay, S., Thakur, S. S. and Mandal, J. K. (2021) 'Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: Benefit for the society', *Innovations in Systems and Software Engineering*, 17, 45–52. https://doi.org/10.1007/s11334-020-00372-5

Brownlee, J. (2021) 'Tour of Evaluation Metrics for Imbalanced Classification', *Machine Learning Mastery*, 1 May. Available at: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/ [Accessed 11 November 2024].

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA, 13-17 August 2016, pp. 785-794. https://doi.org/10.1145/2939672.2939785

Das, D., Katyal, Y., Verma, J., Dubey, S., Singh, A., Agarwal, K., Bhaduri, S. and Ranjan, R. (2020) 'Information retrieval and extraction on Covid-19 clinical articles using graph community detection and Bio-Bert embeddings', in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online, July 2020.

Davagdorj, K., Wang, L., Li, M., Pham, V. H., Ryu, K. H. and Theera-Umpon, N. (2022) 'Discovering thematically coherent biomedical documents using contextualized bidirectional encoder representations from transformers-based clustering', *International Journal of Environmental Research and Public Health*, 19(10), 5893. https://doi.org/10.3390/ijerph19105893

Dessi, D., Helaoui, R., Kumar, V., Recupero, D. R. and Riboni, D. (2021) *TF-IDF vs Word Embeddings for morbidity identification in clinical notes: An initial study*. https://doi.org/10.48550/arXiv.2105.09632

Fang, L., Chen, Q., Wei, C. H., Lu, Z. and Wang, K. (2023) *Bioformer: An efficient transformer language model for biomedical text mining*. https://doi.org/10.48550/arXiv.2302.01588

Fang, L. and Wang, K. (2022) *Multi-label topic classification for COVID-19 literature with Bioformer*. https://doi.org/10.48550/arXiv.2204.06758

Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z. and El Asri, B. (2021) 'BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis', *Computer Methods and Programs in Biomedicine Update*, 1, 100042. https://doi.org/10.1016/j.cmpbup.2021.100042

Hebbar, S. and Xie, Y. (2021) 'CovidBERT-biomedical relation extraction for Covid-19', *The International FLAIRS Conference Proceedings*, 34. https://doi.org/10.32473/flairs.v34i1.128488

Kormilitzin, A., Vaci, N., Liu, Q. and Nevado-Holgado, A. (2021) 'Med7: A transferable clinical natural language processing model for electronic health records', *Artificial Intelligence in Medicine*, 118, 102086. https://doi.org/10.1016/j.artmed.2021.102086

Kumar, P., Bhatnagar, R., Gaur, K. and Bhatnagar, A. (2021) 'Classification of imbalanced data: Review of methods and applications', *IOP Conference Series: Materials Science and Engineering*, 1099, 012077. https://doi.org/10.1088/1757-899X/1099/1/012077

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. and Kang, J. (2020) 'BioBERT: A pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, 36(4), pp. 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

Moradi, M., Dorffner, G. and Samwald, M. (2020) 'Deep contextualized embeddings for quantifying the informative content in biomedical text summarization', *Computer Methods and Programs in Biomedicine*, 184, 105117. https://doi.org/10.1016/j.cmpb.2019.105117

Naeem, S., Ali, A., Anam, S. and Ahmed, M. M. (2023) 'An unsupervised machine learning algorithms: Comprehensive review', *International Journal of Computing and Digital Systems*, 13(1), pp. 911-921. https://doi.org/10.12785/ijcds/130172.

Pes, B. (2021) 'Learning from high-dimensional and class-imbalanced datasets using random forests', *Information*, 12(8), 286. https://doi.org/10.3390/info12080286

Saputra, D. M., Saputra, D. and Oswari, L. D. (2020) 'Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method', in *Sriwijaya International Conference on Information Technology and its Applications (SICONIAN 2019)*. Palembang, Indonesia, 16 November 2019, pp. 341-346. https://doi.org/10.2991/aisr.k.200424.051

Selva Birunda, S. and Kanniga Devi, R. (2021), 'A review on word embedding techniques for text classification', *International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2020)*, pp. 267-281. https://doi.org/10.1007/978-981-15-9651-3_23

Sharaf, S. and Anoop, V. S. (2023) *An analysis on large language models in healthcare: A case study of BioBERT*. https://doi.org/10.48550/arXiv.2310.07282

Singh N. (2024) '11000 Medicine Details', *Kaggle,* Available at: https://www.kaggle.com/datasets/singhnavjot2062001/11000-medicine-details[Accessed 01 December 2024].

Song, M., Jung, H., Lee, S., Kim, D. and Ahn, M. (2021) 'Diagnostic classification and biomarker identification of Alzheimer's disease with random forest algorithm', *Brain Sciences*, 11(4), 453. https://doi.org/10.3390/brainsci11040453

Statista Research Department (2024) 'Online Pharmacy - India', *Statista,* Available at: https://www.statista.com/outlook/hmo/digital-health/digital-treatment-care/digital-care-management/online-pharmacy/india [Accessed 28 November 2024].

Susan, S. and Kumar, A. (2021) 'The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent state of the art', *Engineering Reports*, 3(4), e12298. https://doi.org/10.1002/eng2.12298

Szeghalmy, S. and Fazekas, A. (2023) 'A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning', *Sensors*, 23(4), 2333. https://doi.org/10.3390/s23042333

Wang, C., Deng, C. and Wang, S. (2020) 'Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost', *Pattern Recognition Letters*, 136, pp. 190-197. https://doi.org/10.1016/j.patrec.2020.05.035

Wang, T., Lu, C., Ju, W. and Liu, C. (2022) 'Imbalanced heartbeat classification using EasyEnsemble technique and global heartbeat information', *Biomedical Signal Processing and Control*, 71, 103105. https://doi.org/10.1016/j.bspc.2021.103105

Zhang, P., Jia, Y. and Shang, Y. (2022) 'Research and application of XGBoost in imbalanced data', *International Journal of Distributed Sensor Networks*, 18(6). https://doi.org/10.1177/15501329221106935.