

An Integrative Machine Learning Framework for Drug Repurposing Using Molecular Docking and Drug Sentiment Analysis in Type 2 Diabetes

Johnson Opeyemi Olayinka¹, Daniel Anuoluwapo Olarewaju², Gladys Chinyere Olumba³, Damilola Racheal Odubekun⁴, Anumo Francis Dominion⁵, Mark ogechi munachimso⁶, Oladayo Michael Ayoola⁷, Emmanuel Oladiran Amos⁸

¹ Department of Biochemistry,

University of Ilorin, Ilorin Kwara State, Nigeria,

² Department of Chemistry and Biochemistry

Florida Atlantic University, 777 Glades Rd, Boca Raton,

Florida, United States,

³ College of Ecology and Environment,

Chengdu University of Technology, China,

⁴ Department of Chemistry,

University of Ilorin, Ilorin Kwara State, Nigeria,

⁵ Department of Mechanical Engineering

Federal University Otuoke, Yenagoa, Bayelsa State, Nigeria,

⁶ Department of Accountancy

Federal Polytechnic Nekede Owerri, Nigeria,

⁷ Department of Biochemistry,

University of Ilorin, Ilorin Kwara State, Nigeria,

⁸ Department of Biochemistry,

University of Ilorin, Ilorin Kwara State, Nigeria.

Corresponding Author Email Address:

Johnsonopeyemiolayinka@gmail.com

<https://doi.org/10.56201/jbgr.vol.12.no5.2026.pg106.121>

Abstract

Type 2 diabetes (T2D) is a major global health challenge that requires continuous investigation of alternative therapeutic strategies. Drug repurposing is a strategic technique that involves identifying new therapeutic applications from existing drugs. This research developed a machine-learning framework that combines drug-review sentiment analysis and molecular docking to examine the interaction of selected non diabetic drugs with a type 2 diabetic protein target, and determined their potential to treat T2D under further investigations. Drug review data were obtained from the UCI machine learning repository. The data contains 4,143 records and were divided into training (containing 3,107 records) and testing (containing 1,036 records) sets. Using Term Frequency–Inverse Document Frequency (TF-IDF), textual information from the drug review datasets were converted into numerical features, suitable for model training. Three (3) different supervised learning machine learning algorithms; Logistic Regression, Linear Support Vector Machine, and Naive Bayes classifiers, were evaluated using five-fold cross-validation based on accuracy, precision, recall, F1 score, and macro-F1. Of the three (3) models, Logistic

Regression therefore had the best overall performance, achieving a macro recall of 0.567231 and a macro-F1 score of 0.558643, and was subsequently used for the drug-level sentiment analysis. The three non-diabetes drugs, bupropion, esomeprazole, and adapalene, with the highest sentiment scores were selected for molecular docking against dipeptidyl peptidase-4 (DPP-4). Also, Sitagliptin was selected as the control drug and docked against DPP-4. Using CB-Dock2, the predicted docking scores were -6.3 kcal/mol for bupropion, -8.8 kcal/mol for esomeprazole, -10.8 kcal/mol for adapalene, and -9.3 kcal/mol for Sitagliptin. Adapalene demonstrated the best predicted binding score, which suggests that adapalene may have better interaction with DPP-4 than Sitagliptin (control drug). This may require further investigations. The findings demonstrate the potential of integrating patient-generated textual evidence with structure-based molecular analysis to prioritize existing drugs for further investigation in T2D. However, the computational findings do not establish antidiabetic efficacy or clinical suitability and require further validation through additional computational and experimental studies.

Keywords: Type 2 diabetes; drug repurposing; machine learning; sentiment analysis; TF-IDF; molecular docking; CB-Dock2; DPP-4; Adapalene.

1. Introduction and Literature Review

Type 2 Diabetes (T2D) is a chronic disease that is characterized by metabolic disorders, which include insulin resistance and insufficient insulin secretion that is partly due to pancreatic β -cell dysfunction (Lestary et al., 2022). This can result in long-term elevated blood glucose levels in the body. According to the International Diabetes Federation (IDF), it has been determined that 589 million adults between the age of 20–79 years were living with diabetes worldwide in 2024, and this number will likely increase to 853 million by 2050 (Genitsaridi et al., 2026).

There are various strategies to managing T2D, they include: dietary and exercise modifications, medications, appropriate self-monitoring of blood glucose (SMBG), regular monitoring for complications, laboratory assessment, and more. The 2026 Standards of Care of the American Diabetes Association has recommended that T2D diabetes patients can focus on glucose-lowering therapy through the use of any of the approved T2D therapeutic drugs, including metformin, glucagon-like peptide-1 receptor agonists (GLP-1 RAs), sodium-glucose cotransporter 2 (SGLT2) inhibitors, dipeptidyl peptidase-4 inhibitors, sulfonylureas, and insulin, among others. Although there has been recent development in the treatment of T2D, there is a need for continuous investigation of alternative strategies to identify effective therapeutic candidates for managing T2D. Drug repurposing is one major strategy currently being used to identify non-T2D drugs with potential to manage T2D. Drug repurposing means investigating whether a clinical drug can be used to treat a disease for which it was not originally designed. There are various approaches to drug repurposing, but they can be classified into computational and experimental drug repurposing. Computational drug repurposing approaches employ biological databases, genomic profiling, structural biology, molecular docking, and machine learning (Cousins et al., 2024). These approaches can use different forms of evidence, including protein structures, genomic signatures, biological networks, and real-world clinical data, to identify and prioritize candidate drug-disease relationships. Machine learning models have demonstrated their ability to find complex patterns within large datasets and are now being used in computational drug discovery (Vamathevan et al., 2019). Another application of machine learning is natural language processing (NLP), which can be used for sentiment analysis of drug reviews to generate information that may contribute to the prioritization of drug candidates (Dubba et al., 2024). Recent developments in

machine learning models, particularly deep learning models, provide new insights into how drugs bind to targets and how their physicochemical characteristics relate to phenotypic alterations. In the context of drug repurposing, machine learning can be integrated into the molecular modelling of protein targets to identify therapeutic candidates.

In addition, molecular docking involves using computational algorithms to predict the interaction between small molecules and protein targets. Docking assists in identifying compounds that are likely to show favorable binding energies, making them suitable candidates for further research, by estimating the binding affinities and poses of these ligands within the active regions of target proteins. It is a process involving the creation and evaluation of numerous potential binding conformations through the use of scoring algorithms that assess the strength and stability of the interactions (Nivatya et al., 2025). Molecular docking can therefore provide structure-based evidence for prioritizing existing drugs as potential therapeutic candidates for T2D.

Despite advances in the use of machine learning, molecular docking, and sentiment analysis for evaluating therapeutic candidates, there remains an opportunity to investigate how molecular and patient-generated textual evidence can be incorporated within a unified framework for prioritizing potential therapeutic candidates for T2D. Therefore, the proposed study aims to develop an integrative machine-learning framework that combines drug sentiment analysis with molecular docking to identify and prioritize existing drugs as potential therapeutic candidates for type 2 diabetes. In this framework, machine-learning models will be used to analyse drug-related textual data and generate quantitative sentiment information, while molecular docking will be used to investigate the predicted interactions between selected candidate drugs and T2D-associated protein targets. The outputs from these two computational components will subsequently be integrated to support the prioritization of existing drugs for further investigation as potential therapeutic candidates against T2D. The resulting prioritization will provide computational evidence for further investigation rather than confirmation of therapeutic efficacy.

2. Materials and Methodology

2.1. Dataset

The dataset used for the sentiment analysis in this study is drug review data obtained from UCI Machine Learning Repository under the Drug Reviews (Druglib.com) dataset. The data contains 4,143 drug reviews, and was divided into training and testing data (3,107 and 1,036 records). The dataset comprises drug names, medical conditions, patient ratings, effectiveness and side-effect ratings, and three textual review components describing perceived benefits, side effects, and overall comments. To ensure accuracy and consistency of the data, outliers were removed and therefore, exploratory data analysis was conducted to examine the sentiment distribution (Figure 1)

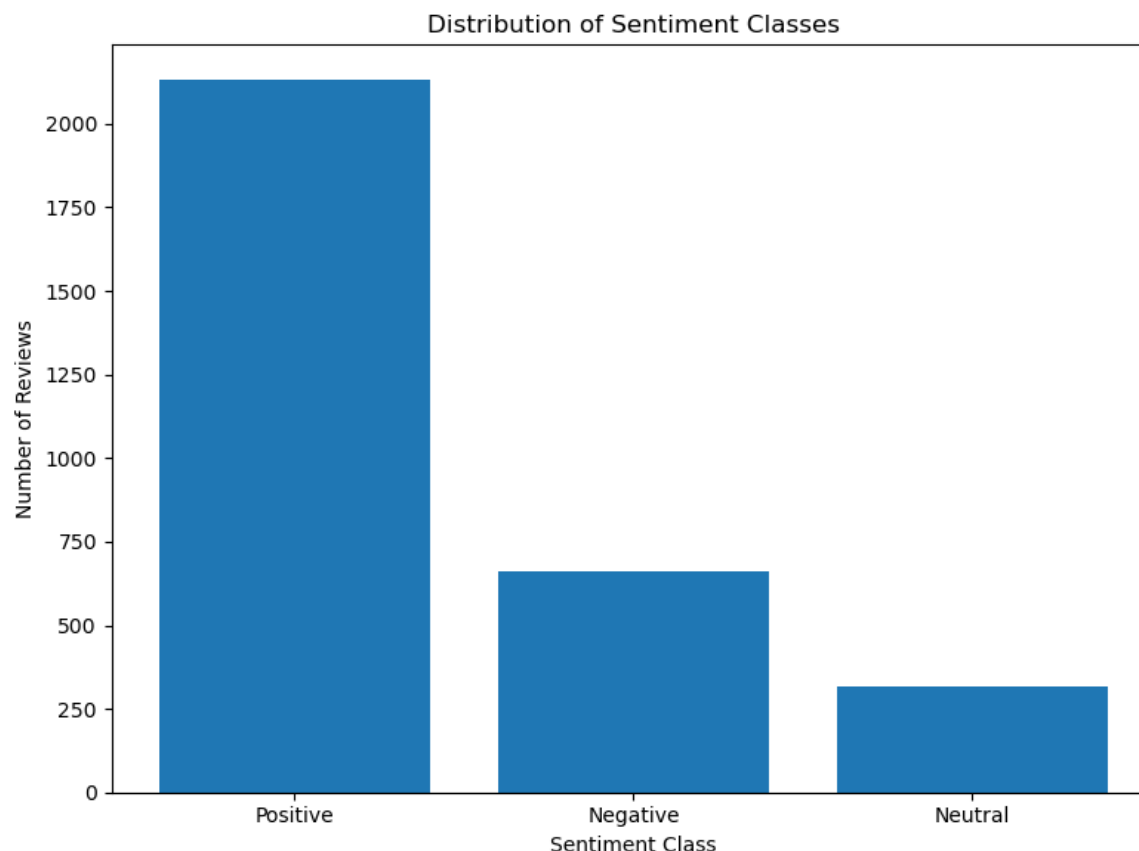


Figure 1: Distribution of Sentiment Classes

2.2. Machine Learning Framework

The drug review sentiment analysis was accomplished by a ten-step machine learning workflow (Figure 3) beginning with data preprocessing and ends with model evaluation and selection.

2.2.1. Data Preprocessing

Data preprocessing was carried out to prepare the textual data for machine learning analysis. Preprocessing involves a set of operations that include; data cleaning, transformation, normalisation, feature encoding, and dimensionality reduction (Dwivedi et al., 2015). The text fields; benefit Review, side Effects Review and comments Review were combined to form a new column called “review text”. The new column was standardized by removing punctuations, numbers, extra spaces and lower casings, and was named “Cleaned Review Text”. This step helps to reduce irrelevant variation in the text while retaining information relevant to sentiment classification.

2.2.2. Sentiment Labeling

The next step after data preprocessing is sentiment labeling. At this stage, the drug review ratings were classified into three groups; Negative (ratings from 1 to 3), Neutral (ratings from 4 to 6), and Positive (ratings from 7 to 10), under the column named “Sentiment”.

2.2.3. Feature Extraction

The major purpose of feature extraction in this analysis is to convert the preprocessed drug review textual data into numerical data, which is used for model training and classification. The text feature extraction technique that was used to quantify the importance of terms within individual reviews relative to their occurrence across the corpus is “Term Frequency–Inverse Document Frequency (TF-IDF)”, and the parameters were set to (stop_words="english", ngram_range=(1, 2), min_df=2, max_df=0.95, sublinear_tf=True).

Feature extraction allows the model to identify important words and terms in reviews, represent those words numerically, determine how relevant a word is within a review and across the dataset, and use those numerical patterns to learn the relationship between review text and sentiment.

2.2.4. Data Splitting and Model Training

The data was splitted into a training set (containing 75% of the data), which is for model training and validation), and a test set (containing 25% of the data), for final model evaluation. Three different models (supervised learning models) were used for sentiment analysis; they include: Logistic Regression, Linear SVM, and Naive Bayes. Each model was trained on the training data and predictions were made on the test data.

2.2.5. Model Validation and Evaluation

To prevent model overfitting, 5-fold cross validation was conducted on each model during evaluation (Figure 2). The models were evaluated using accuracy score, precision, recall, F1 Score, macro F1 score, and confusion matrix. The best model was selected based on higher recall and macro F1 score. Out of the three models, Logistic regression outperformed with the highest recall score of 0.567231 and macro F1 score of 0.558643 (Table 1)

	Model	Accuracy	Macro Precision	Macro Recall	Macro F1
0	Logistic Regression	0.738417	0.586906	0.567231	0.558643
1	Linear SVM	0.753861	0.630039	0.542205	0.545242
2	Multinomial Naive Bayes	0.646718	0.215573	0.333333	0.261821

Table 1 Model Evaluation Score

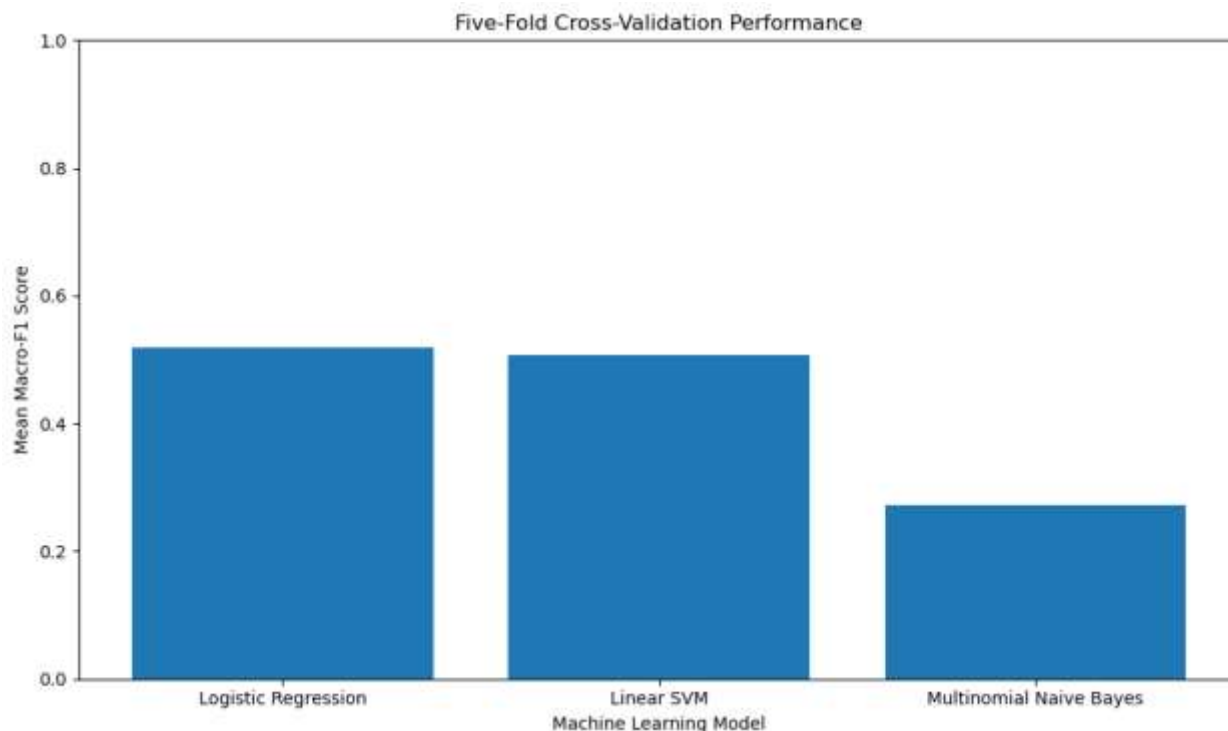


Figure 2 Five-Fold Cross Validation Performance

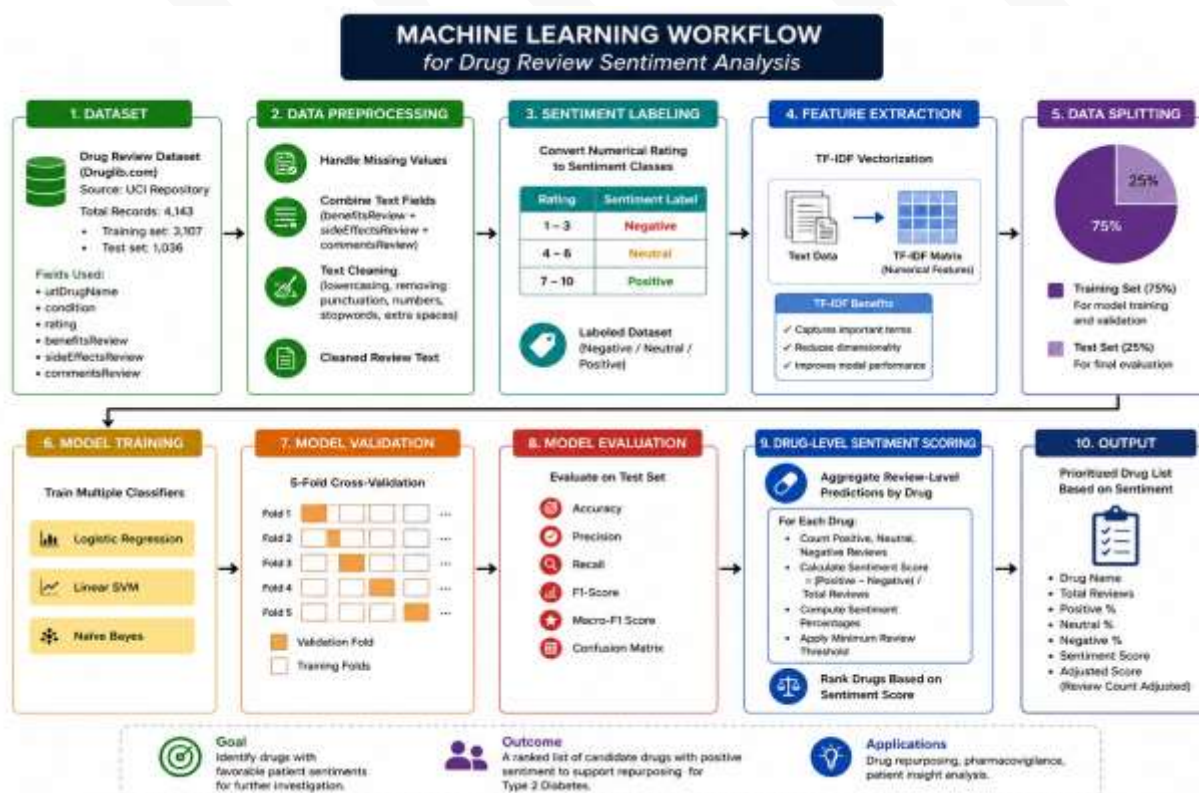


Figure 3 Machine Learning WorkFlow for Drug Review Sentiment Analysis

2.3. Molecular Docking

2.3.1. Ligand Selection and Preparation

Following the sentiment analysis, there are three drugs that had the best sentiment score of 1.0 and 0.84, which are selected as ligands for the molecular docking. The drugs include: Wellbutrin-XL (Bupropion), Nexium (Esomeprazole), and Differin (Adapalene). The drugs were confirmed to be non-diabetic drugs. The Ligands structures were downloaded from the ChEMBL database in sdf format. ChEMBL is an open-source database that contains bioactivity data from approximately 230 journals across the biomedical field, supporting drug discovery research (Mendez et al., 2019). The Ligands were then prepared on Pymol by adding hydrogen and confirmation of three-dimensional (3d) structure. In addition to these, Sitagliptin (a well-known drug that can bind and inhibit the activity of DPP4 in type 2 diabetes in adults), was also downloaded from ChEBML database, in SDF format and prepared in Pymol by hydrogen atom. Sitagliptin will serve as a reference or control drug for the molecular docking analysis. The prepared structures were then saved in PDB format for subsequent docking analysis.



Figure 4 Bupropion

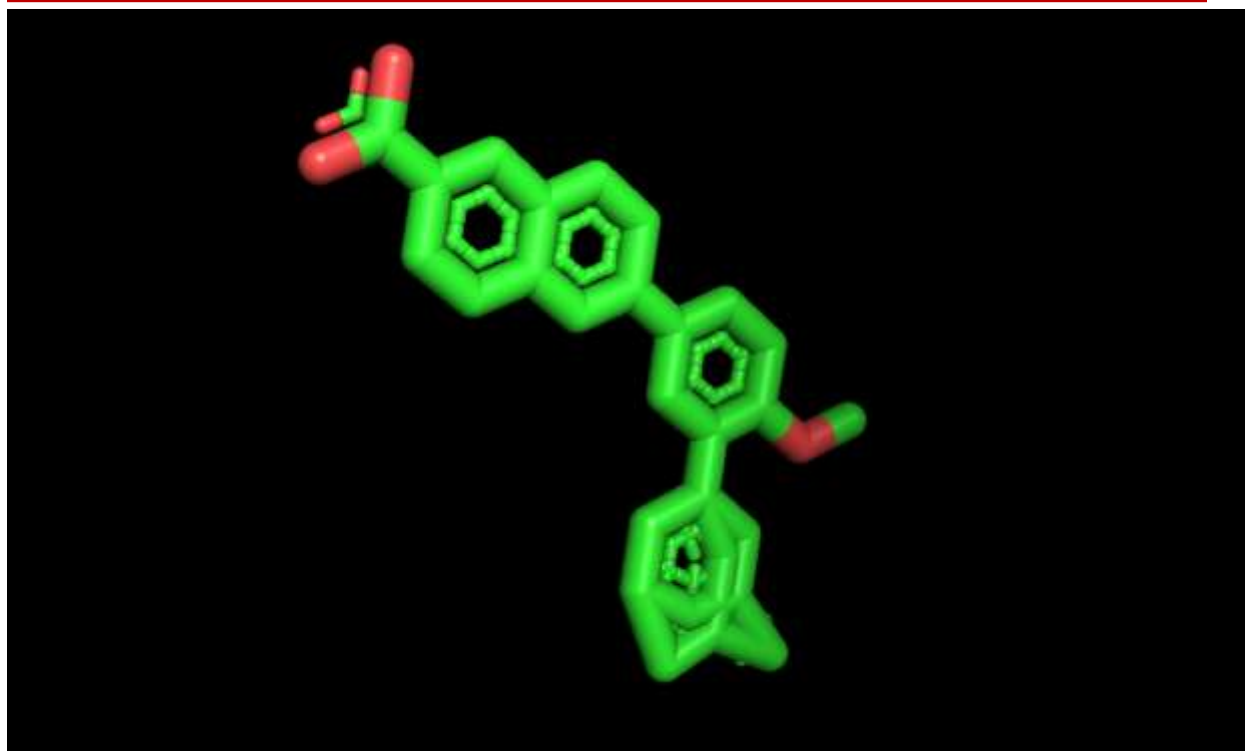


Figure 5 Adapalene

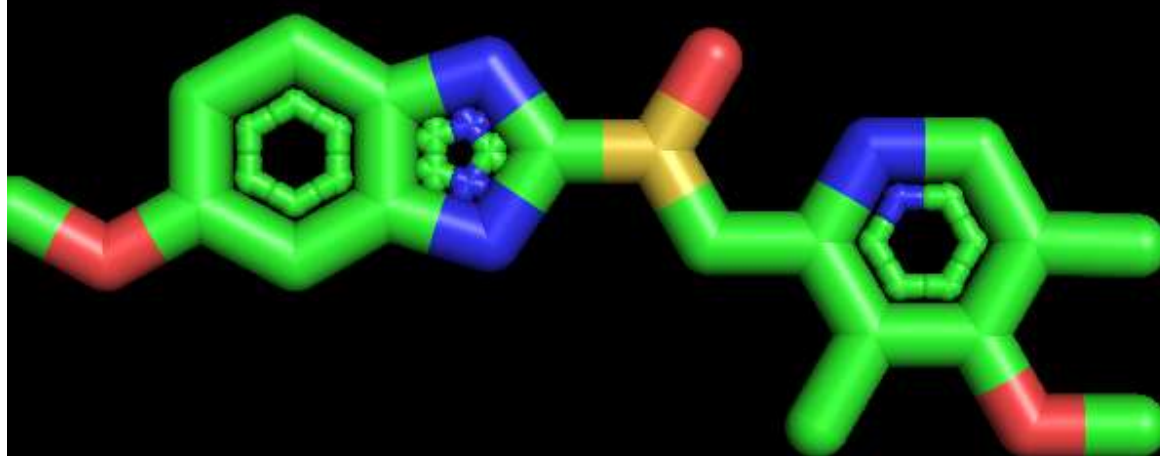


Figure 6 Esomeprazole

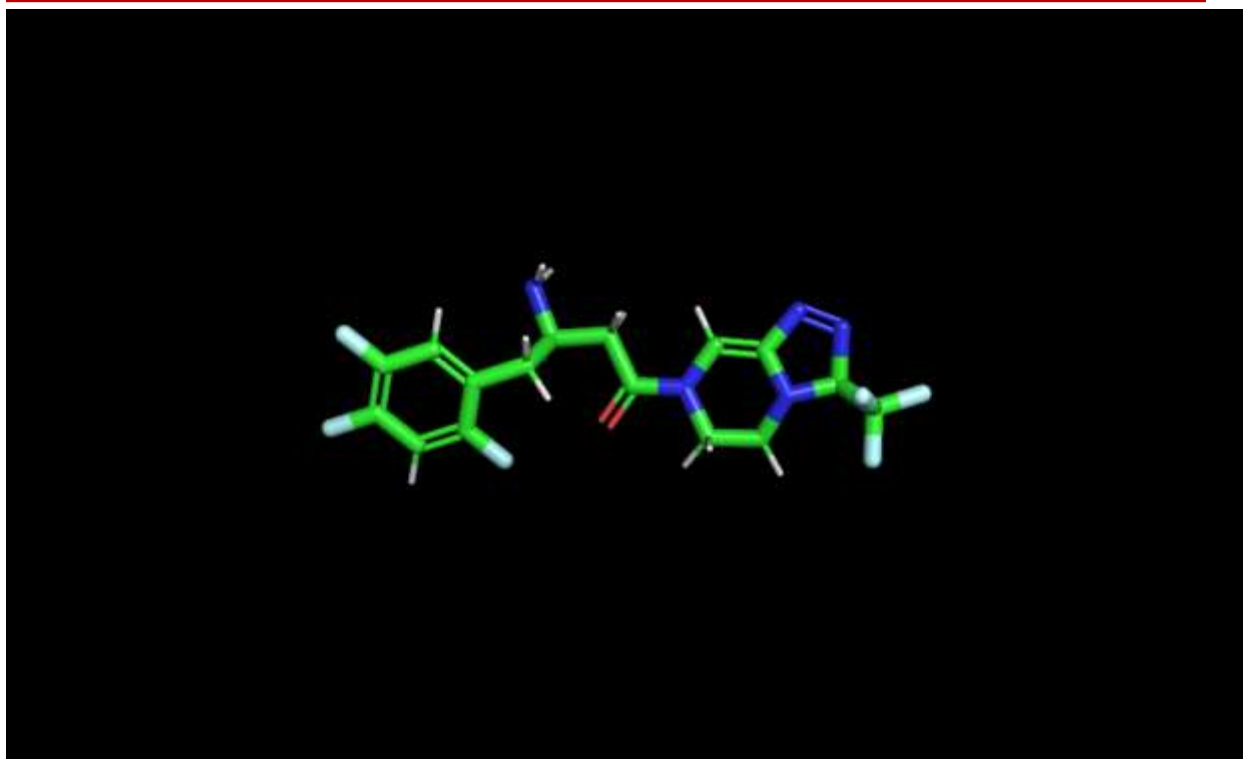


Figure 7 Sitagliptin

2.3.2. Protein Selection and Preparation

Dipeptidyl peptidase-4 (DPP-4), a well-known target in type 2 diabetes was selected as the protein target for the docking analysis. The activity of DPP-4 can be inhibited by some medications (DPP-4 inhibitors) used to manage high blood glucose levels in patients with type 2 diabetes mellitus. Some of these inhibitors are; Sitagliptin, Saxagliptin, Vildagliptin, Alogliptin, Linagliptin, Denagliptin, Dutogliptin and Omarigliptin (Mohd, 2024).

The three-dimensional (3d) structure of the protein target (DPP-4) was obtained from the Protein Data Bank (PDB). The structure was cleaned and prepared in PyMOL in the following six (6) steps:

1. Confirmation of 3d structure
 2. Removal of Water molecule
 3. Retaining Carbohydrate and Polymer structure
 4. Removal of N7F ligand
 5. Addition of hydrogen
 6. Download and save as pdb format, which is the acceptable format for docking on CB-Dock2
- Protein–ligand docking was performed using CB-Dock2. CB-Dock2 is an automated blind-docking platform that combines binding-cavity detection with docking and template-based information to identify potential binding sites and ligand poses (Liu et al., 2022). Its docking workflow incorporates AutoDock Vina, a widely used molecular-docking program that estimates ligand binding poses and provides a scoring function for ranking predicted complexes (Sarkar et al., 2024). For each protein–ligand complex, the best-ranked pose, predicted Vina docking score, and docking-box dimensions were recorded for comparison.

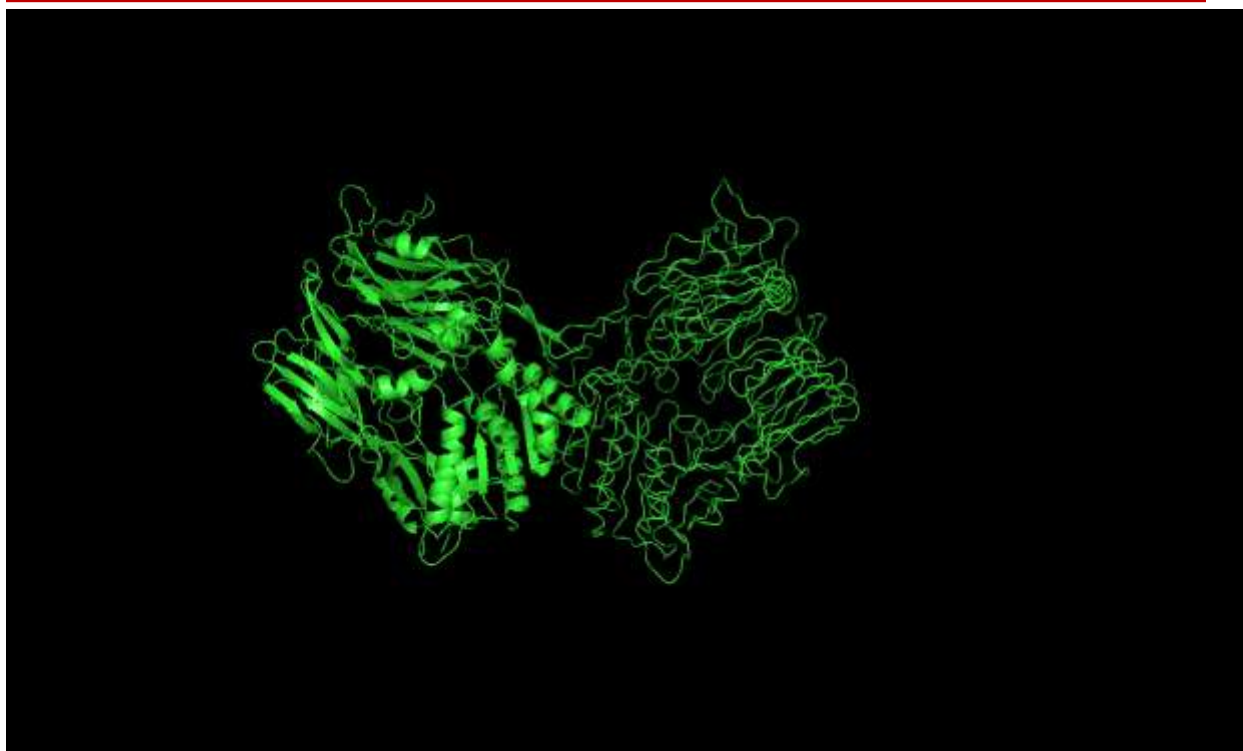


Figure 8 DPP-4

3. Results and Discussion

3.1 Sentiment Analysis

The Table 2 below shows the sentiment analysis score of the three selected drug candidates and molecular docking results for the three drug candidates and the control drug, Sitagliptin. The sentiment score of 1.0 was achieved for Welbutrin-XL (Bupropion), and Nexium (Esomeprazole), while a sentiment score of 0.84 was achieved for Differin (Adapalene). A positive sentiment score can be interpreted as a favourable patient reported experience with the drugs but not evidence of therapeutic efficacy for type 2 diabetes. Therefore, sentiment analysis of drug reviews alone cannot establish the antidiabetic potential of the drugs.

3.2 Molecular Docking Analysis

The Protein-ligand blind docking was performed on CB-Dock2, a protein-ligand blind docking server that integrates both structure-based cavity detection with template based molecular docking methods (Yang et al., 2022). The docking analysis shows that all the four ligands (including the control drug) had the best ranked pose in cavity C3. DPP4-Sitagliptin complex (control) had a vina score of -9.3 kcal/mol and dimension 25 x 25 x 25, DPP4-Bupropion complex had a vina score of -6.3 kcal/mol and dimension 19 x 19 x 19, DPP4-Esomeprazole complex had a vina score of -8.8 kcal/mol and dimension 23 x 23 x 23, and DPP4-Adapalene complex produced a Vina score of -10.8 kcal/mol and dimension 24 x 24 x 24. Sitagliptin recorded the largest dimension. The dimensions describe the docking search space and should not be interpreted as measures of binding strength. Sitagliptin is one of the prominent DPP-4 inhibitors, exhibiting potent hypoglycemic effects. Although it is associated with some side effects including; gastrointestinal issues, rashes, and more (Huang et al., 2020).

Compared with the reference drug, Sitagliptin, Adapalene produced a more negative predicted Vina docking score (−10.8 kcal/mol versus −9.3 kcal/mol). This suggests that Adapalene may have a favorable predicted interaction with DPP-4. However, additional computational and experimental studies would be required to determine whether this predicted interaction translates into DPP-4 inhibitory activity. In this study, Sitagliptin remains the established DPP-4 inhibitor used as the reference compound.

Drug/Ligand	Protein Target	Sentiment Score	Vina Docking Score (kcal/mol)	Best Pose	Docking Dimensions (X × Y × Z)
Sitagliptin (Control)	DPP-4	-	-9.3	C3	25 x 25 x 25
Wellbutrin-XL (Bupropion)	DPP-4	1.0	-6.3	C3	19 × 19 × 26
Nexium (Esomeprazole)	DPP-4	1.0	-8.8	C3	23 × 23 × 23
Differin (Adapalene)	DPP-4	0.84	-10.8	C3	24 × 24 × 24

Table 2 Sentiment Analysis and Molecular docking result

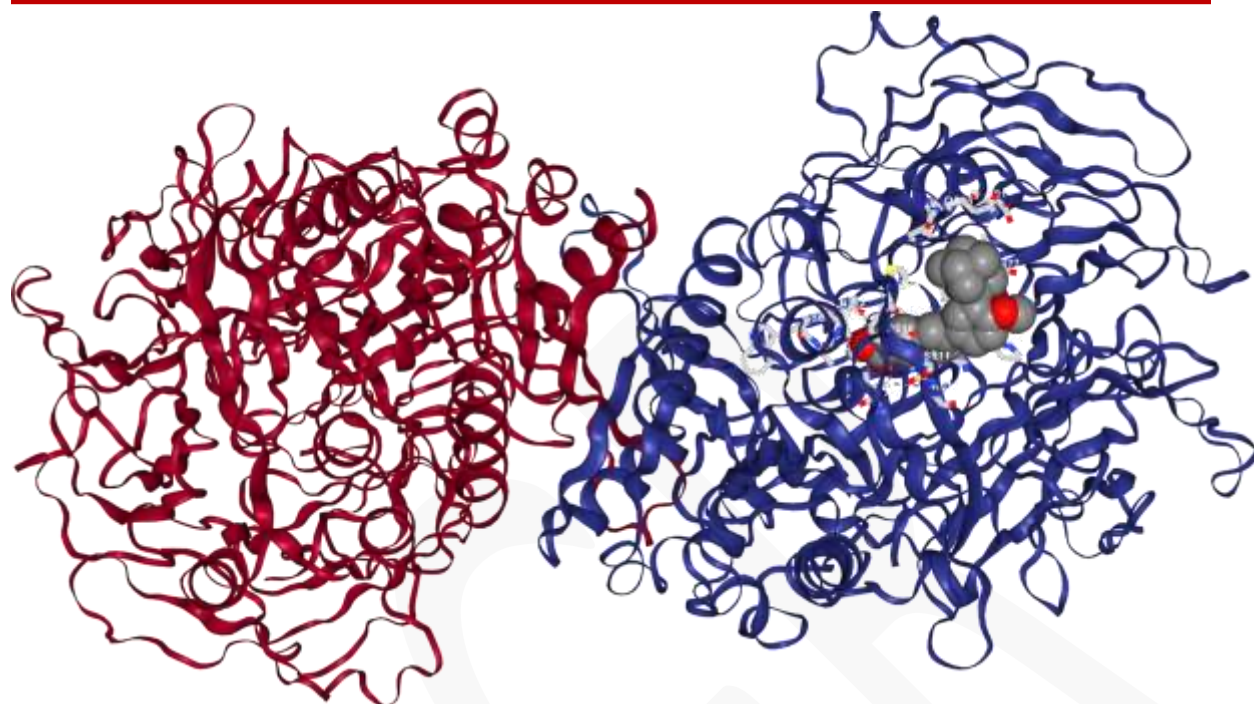


Figure 9 DPP-4-Adapalene Complex

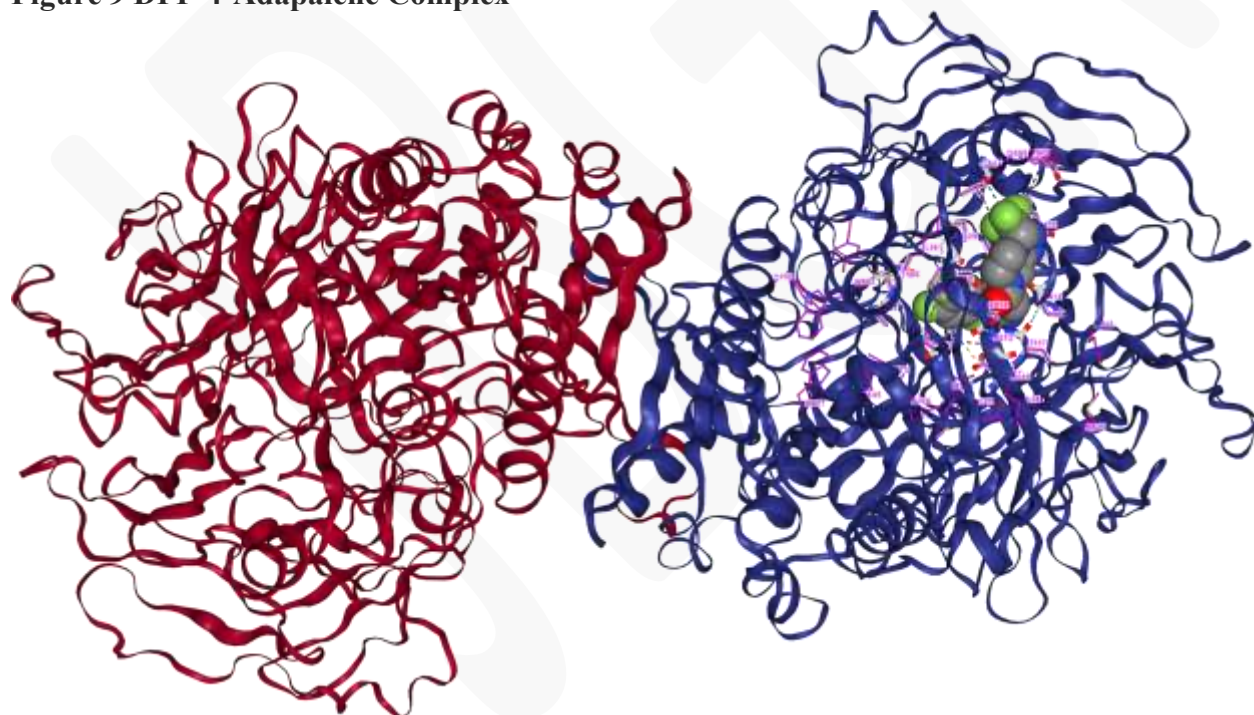


Figure 10 DPP-4-Sitagliptin Complex

4. Limitations of the Research

There are several limitations that should be considered when interpreting this research. First, the sentiment analysis was based on drug reviews obtained from the UCI Machine Learning Repository under the Drug Reviews (Druglib.com) dataset, and the experiences reported by patients may relate to different medical conditions and therapeutic indications. Therefore, the

sentiment scores do not necessarily represent patient experiences with the use of these drugs for type 2 diabetes. Second, molecular docking provides computational predictions of ligand–protein interactions and does not by itself establish target inhibition, pharmacological activity, safety, or clinical efficacy. Finally, the study did not include molecular dynamics simulations, biochemical assays, cell-based experiments, or clinical studies to experimentally validate the predicted interactions. Therefore, the identified compounds should be regarded as computationally prioritized candidates which require further investigation rather than confirmed therapeutic agents for type 2 diabetes.

5. Conclusion

This study developed an integrative computational framework that combined drug-review sentiment analysis with molecular docking to prioritize non-diabetes drugs as potential therapeutic candidates for type 2 diabetes. Logistic Regression was the best-performing sentiment-classification model based on macro recall (0.567231) and macro F1 score (0.558643). Three non-diabetes drugs; Bupropion, Esomeprazole, and Adapalene were subsequently selected based on their positive sentiment scores and evaluated against DPP-4 using CB-Dock2. Sitagliptin was used as the control drug and also docked with DPP-4 on CB-Dock2. Among the evaluated compounds, Adapalene produced the most favorable predicted Vina docking score (−10.8 kcal/mol), compared with −9.3 kcal/mol for Sitagliptin, −8.8 kcal/mol for Esomeprazole and −6.3 kcal/mol for Bupropion. These findings suggest that adapalene warrants further investigation as a computationally prioritized candidate for DPP-4-related studies. However, the results do not establish antidiabetic efficacy or clinical suitability. Further validation through repeated or consensus docking, molecular dynamics simulations, biochemical assays, and appropriate experimental studies is required before any therapeutic conclusion can be drawn. Overall, the framework demonstrates how patient-generated textual evidence and structure-based computational evidence can be combined to support the prioritization of existing drugs for future drug-repurposing research.

References

- American Diabetes Association Professional Practice Committee for Diabetes. (2026). 1. Improving care and promoting health in populations: Standards of Care in Diabetes—2026. *Diabetes Care*, 49(Supplement 1), S13–S26. <https://doi.org/10.2337/dc26-S001>
- Cousins, H. C., Nayar, G., & Altman, R. B. (2024). Computational approaches to drug repurposing: Methods, challenges, and opportunities. *Annual Review of Biomedical Data Science*, 7, 15–29. <https://doi.org/10.1146/annurev-biodatasci-110123-025333>
- Dwivedi, S. K., & Rawat, B. (2015). A review paper on data preprocessing: A critical phase in web usage mining process. In *Proceedings of the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT)* (pp. 506–510). IEEE. <https://doi.org/10.1109/ICGCIoT.2015.7380517>
- Genitsaridi, I., Salpea, P., Salim, A., Sajjadi, S. F., Tomic, D., James, S., Thirunavukkarasu, S., Issaka, A., Chen, L., Basit, A., Luk, A. O. Y., Ma, R. C. W., Mbanya, J.-C., Ramachandran, A., Wild, S. H., Duncan, B. B., Boyko, E. J., & Magliano, D. J. (2026). 11th edition of the IDF Diabetes Atlas: Global, regional, and national diabetes prevalence estimates for 2024 and projections for 2050. *The Lancet Diabetes & Endocrinology*, 14(2), 149–156. [https://doi.org/10.1016/S2213-8587\(25\)00299-2](https://doi.org/10.1016/S2213-8587(25)00299-2)
- Huang, J., Jia, Y., Sun, S., & Meng, L. (2020). Adverse event profiles of dipeptidyl peptidase-4 inhibitors: Data mining of the public version of the FDA adverse event reporting system. *BMC Pharmacology and Toxicology*, 21(1), 68. <https://doi.org/10.1186/s40360-020-00447-w>
- Liu, Y., Yang, X., Gan, J., Chen, S., Xiao, Z.-X., & Cao, Y. (2022). CB-Dock2: Improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting. *Nucleic Acids Research*, 50(W1), W159–W164. <https://doi.org/10.1093/nar/gkac394>
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., Hersey, A., & Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1), D930–D940. <https://doi.org/10.1093/nar/gky1075>
- Naim, M. J. (2024). A review of dipeptidyl peptidase-4 (DPP-4) and its potential synthetic derivatives in the management of diabetes mellitus. *Journal of Angiotherapy*, 8(1), 1–12. <https://doi.org/10.25163/angiotherapy.819417>
- Nivatya, H. K., Singh, A., Kumar, N., Sonam, Sharma, L., Singh, V., Mishra, R., Gaur, N., & Mishra, A. K. (2025). Assessing molecular docking tools: Understanding drug discovery and design. *Future Journal of Pharmaceutical Sciences*, 11, 111. <https://doi.org/10.1186/s43094-025-00862-y>
- Sarkar, A., Concilio, S., Sessa, L., Marrafino, F., & Piotto, S. (2024). Advancements and novel approaches in modified AutoDock Vina algorithms for enhanced molecular docking. *Results in Chemistry*, 7, 101319. <https://doi.org/10.1016/j.rechem.2024.101319>
- Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463–477. <https://doi.org/10.1038/s41573-019-0024-5>

Yang, X., Liu, Y., Gan, J., Xiao, Z.-X., & Cao, Y. (2022). FitDock: Protein–ligand docking by template fitting. *Briefings in Bioinformatics*, 23(3), bbac087.
<https://doi.org/10.1093/bib/bbac087>