A Sentiment Analysis of Thread App Reviews and Ratings with Explainable AI

1st Ishrat Jahan Easha
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
ishrat.jahan.easha@g.bracu.ac.bd

2nd Mosa. Rabeya
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
rabeyashammi145@gmail.com

3rd Alvi Anowar
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
alvi.anowar@g.bracu.ac.bd

4th Md Alkawser Rahman
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
md.alkawser.rahman@g.bracu.ac.bd

***Abstract*—In the era of mobile applications, user opinions are conveyed through reviews and ratings. It plays a significant role in reforming both user perceptions and developer strategies. This study explores sentiment analysis specifically in the context of the Threads social networking app. This is a prominent messaging app which employs advanced machine learning algorithms to understand user sentiments embedded in app reviews and ratings. Going beyond conventional sentiment analysis, this research focuses on the enhancement of interpretability and transparency through the implementation of Explainable Artificial Intelligence (XAI) techniques. In the ever evolving world dominated by black box models for sentiment analysis, this study raises concern regarding trust, accountability, and biasness. The objective is to illustrate the inner workings of the sentiment analysis model. In our research we have proposed an interpretable sentiment analysis using five machine learning models: Decision Tree, logistic regression, Support Vector Machine, Random Forest and LGBM. SVM has the highest accuracy of 90%. We have also used the popularXAI models, SHAP to show the interpretability of our used models.**

***Index Terms*—Semtiment Analysis, Thread App, Machine Learning, XAI**

I. INTRODUCTION

In the dynamic era of different mobile apps, user critique, expressing opinion through app reviews and ratings, has become a vital part of influencing users opinions and developers manifestation. The epidemic growth of app ecosystems has highlighted the necessity for robust tools to extract relevant observation from the vast collection of user-generated content. In order to extract sentiments, opinions and approaches expressed in textual data, sentiment analysis which is a part of Natural Language Processing(NLP) can be considered a powerful tool.

This study focuses on sentiment analysis of Threads, a leading messaging app, leveraging machine learning algorithms to distinguish user sentiment from app reviews and ratings. Beyond the scope of traditional sentiment analysis, this research also strives to improve the interpretability and transparency of the analysis through the incorporation of explainable artificial intelligence (XAI) techniques. Finding explainability in AI models is becoming increasingly urgent, especially in applications that directly impact user experience and developer strategy.

The popularity of black box models in sentiment analysis raises concerns about trust, accountability, and bias. To address these challenges, our research aims to shed light on the inner workings of the sentiment analysis model used, providing users, developers, and researchers with an understanding more clarity on how to infer sentiment from text reviews. By combining sentiment analysis with explainable AI, we aim to bridge the gap between model complexity and human understanding, fostering a trusted and reliable application ecosystem with more information.

II. RELATED WORKS

This research [1] aims at the delivery system where they used Deep Learning models as well as XAI algorithm. Here they gathered samples and used different DL algorithms for instance Bi-GRU-LSTM-CNN, LSTM, Bi-LSTM and performed a sentiment analysis. They also explained the projections by using the XAI model called SHAP. Here the Bi-GRU-LSTM-CNN gained the best result which is 96.33

The author [2] of this paper worked on a sentiment analysis where they used XAI models. They key sentiment analysis models and XAI models that are used currently are surveyed for the first time here

This paper was [3] focused on varied NLP and Deep learning models that have created an impact on large amounts of data. They also emphasize sentiment analysis. In order to find what motivates to predict the outcome of DL when the XAI is injected to show perception. Here the famous models SHAP, LIME, Anchor etc were used

The main purpose [4] of this study is to focus on illustration of data in an amount of time. They used Explainable AI

models LIME and SHAP and conducted sentiment analysis on the medical sector. According to them the Deep learning model has worked better for recognizing the thoughts and emotions of the test subjects. However as deep learning can not be explained by humans, there is a distressing matter that it may lack necessary capabilities of interpreting the model.

Ref. [6] In this particular research the author analyzed the sentiment of a series of reviews for a particular hotel. They used Machine Learning techniques and explainable AI for analyzing. However , no prominent result was found after using the XAI model in order to inspect customer reviews.

In another important study, Noor [7] evaluated vast range of studies and as a result a contrast of outcomes among Support Vector Machines (SVM), NLP,Lexicon and other text mining revealed a significant enhancement of 87.33% in favor of Lexicon when juxtaposed with alternative methodologies.. We initiate that it has the highest accuracy. However, Conducting sentiment Analysis in languages other then English is tough. Furthermore, he must consider alteration of the domain when building the model, as words in one domain may have different meanings in another domain. For example, “lightweight” is a positive emotion word for electronics, but a negative emotion word for kitchen gadgets. Machine Learning or Deep Learning techniques can address domain conversion challenges by training models on the same domain dataset. Customer reviews can have multiple words for the same characteristics. For example, in the context of mobile phones, the terms screen or Liquid Crystal Display(LCD) refer to the same meaning. In the realm of cinema, the terms ”images” and ”movies” are often used receprocally, however when it comes to cameras, they denote similar substances. Moreover, it can be found that the people related to the camera industry use the terms ”photo” and ”image” in a similar tone which is indicated in the reference (b8). Traditional dictionary-focus training methods are impaired by their incarceration to a limited number of words, but these challenges can be addressed efficaciously and navigated properly by machine learning and deep learning models. [9].

III. METHODOLOGY

The main goal of this research is to conduct sentiment analysis of reviews and ratings in the context of Thread Apps, using explainable AI techniques to improve interpretability and transparency. Transparency of the sentiment classification model. This research aims to explore and understand users' emotions as they evaluate the Thread app, providing insight into the app's perceived strengths and weaknesses. Using explainable AI, the paper seeks to go beyond prediction accuracy and provide clear and understandable reasoning for the sentiment predictions made by the model. This not only improves the reliability of sentiment analysis, but also provides valuable explanations to stakeholders and users, contributing to more informed decision-making about Thread applications.

A. Dataset Description and Data Preprocessing

In our study, we conducted a comprehensive analysis using a dataset of app thread reviews and ratings from Google Apps and the Play Store. Our approach involved labeling the dataset using unigram feature extraction techniques. This included applying a preprocessor to the raw sentences to improve interpretability. The dataset was then trained using the feature vectors using various machine learning techniques. The semantic analysis component then provided extensive synonyms and similarities, which helped determine the polarity of the content. In the following subsections, we explain our methodology in detail and graphically represent the entire process

preprocessing stage, we focus on two main columns, review description and rating of the dataset. To handle missing values, we first replace scores 1 and 2 with 0, representing negative sentiment, and scores 4 and 5 with 1, representing positive sentiment. Rating 3 has been temporarily set to None, indicating missing or neutral values. To ensure a clean dataset, we removed rows containing any missing values using a dropna operation. We then converted the notes column to an integer type for consistency and deeper analysis. The resulting dataset includes 33,987 cases. Of these, 20,588 cases received a rating of 1, indicating positive sentiment, while 13,399 cases received a rating of 0, indicating negative sentiment. This preprocessing lays the foundation for subsequent sentiment analysis, providing a balanced dataset for meaningful insights into user sentiment related to app reviews by Chain.

Next, we clean as part of preprocessing the text data. This function orchestrates a sequence of operations to refine and standardize entered text for later analysis. To begin with, all characters are converted to lower case to ensure uniformity of case presentation. The feature then systematically removes mentions, URLs, references to images or media on Twitter, as well as any non-alphabetical or designated accented characters. Additionally, it filters single-character words and removes remaining punctuation. Next, the text was coded into individual words and common stop words in English were excluded to focus on content terms. The final step involves removing excess leading and trailing spaces, ensuring consistent spacing between words. We applied this function in the review description column in the dataset. This application ensures that each review description goes through defined preprocessing steps. The overarching goal is to create a refined and standardized dataset, preparing it for subsequent natural language processing and analysis efforts. This meticulous pre-processing improves the quality and relevance of text data, laying the foundation for more accurate and meaningful insights in later stages of analysis.

We then fitted the augmented text dataset with an important preprocessing step that improves the consistency and suitability of the dataset for analytical efforts and natural

language language processing tasks. Finally, the function returns the modified data set, now supplemented with modifier versions of the original text.

Finally, we used the TF-IDF vector to convert the text data into numeric vectors, thereby capturing the significance of the terms in the context of the dataset. The transformed representations of the training and test sets will then be available for use in machine learning models that require numerical input.

We then split the dataset into train and test section and ran several machine learning models. The accuracy was measured using confusion matrix, accuracy matrix, recall and precision. The algorithms that we used to produce the results are Decision Tree, Logistic Regression, SVM, Random Forest and LGBM.

B. Description of Classification and XAI Models

1) Decision Tree: Decision tree analysis is a widely used machine learning method for classification and regression tasks. The basic idea is to recursively divide the data set based on feature values, creating a tree structure where each internal node represents a decision based on a particular feature, and each leaf node corresponds with a predictable result. The two primary criteria for node splitting are Gini Impurity and Entropy, both measuring the impurity or disorder of a dataset. Gini Impurity is calculated using the formula The decision tree is constructed by iteratively selecting splits to minimize impurities. The resulting model provides interpretability that helps better understand the decision-making process, while its predictive capabilities make it a valuable tool in various machine learning applications.

2) Logistic Regression: Logistic Regression is usually used for estimating the criterias of a logistic model. It basically anticipates and analyse. Logistic Regression encircles a more complex cost function called 'Sigmoid function.'Logistic regression is used to predict the output of catagorical points. It can be true or false, 1 or 0, yes or no, and so on. However it produces probabilistic values in between 0 and 1, instead of exact numbers like 0 and 1 .This is a well known machine learning alogorithm as it can bring about classification of data from discreate datasets.

3) SVM: SVM is a famed method in Supervised Learning. It stands for Support Vector Machine. SVMs intercept regression and classification problems in machine learning. The Support Vector Machine method divides n-dimensional space into sectors to categorise data points explicitly. Best choice dissolution is a hyperplane. The top points and vectors of the hyperplane is chosen by SVM. SVMs helps us to figure out complicated data combination without needing many changes. It's a good choice when the dataset is small datasets but has many attributes. Because little and convoluted data is best handled by them. Moreover in most times they give more accurate results than other models. SVM need less processing power than random forests and logistic regression as they employ only the class borderline data points. SVMs only recruit class-border data points. The model uses merely classification-border data.

4) Random Forest: Random Forest is a robust and versatile ensemble learning algorithm widely used in machine learning for classification and regression tasks. This method constructs multiple decision trees during training, each utilizing a random subset of features and training data. This randomness introduces diversity, mitigating overfitting and enhancing the model's resilience to noise. By aggregating predictions from these varied trees, Random Forest achieves high predictive accuracy and proves effective in handling high-dimensional datasets with both categorical and numerical features. Known for its minimal hyperparameter tuning requirements, it is user-friendly and applicable across diverse domains, offering insights into feature importance and finding application in finance, healthcare, and natural language processing.

5) LGBM: LightGBM (Light gradient Boosting Machine) is a gradient boosting framework that stands out for its efficiency and speed in training large datasets. It was developed by Microsoft, LightGBM is specifically designed to handle distributed and efficient training for large-scale, high-dimensional datasets. It uses tree-based learning, similar to other gradient boosting methods, but is distinguished by its use of graph-based learning. LightGBM builds decision trees by grouping continuous features into distinct groups, thereby reducing the computational cost associated with finding optimal distributions. This technique allows for faster training times and efficient memory usage. Additionally, LightGBM supports parallel learning and GPUs, making it well-suited for processing large data sets. The ability to provide accurate predictions with reduced computing resources has made LightGBM a popular choice in various machine learning applications, especially in scenarios where high-dimensional and scale data processing is difficult.

6) Interpretable model SHAP: SHAP is a powerful, model-independent technique used to interpret the output of machine learning models. Developed based on the principles of cooperative game theory, SHAP values provide a comprehensive and intuitive understanding of feature contributions to model predictions. The SHAP model assigns a Shapley value to each feature, representing its marginal contribution to the difference between the actual model output and the average model output for all possible feature combinations . This approach allows for nuanced interpretation of the impact of individual characteristics on predictions, making it easier to understand the model's decision-making process. SHAP values are especially useful in situations where interpretability and transparency are important, allowing stakeholders to understand the importance of each feature in achieving specific model results. Additionally, SHAP values provide a unified framework for interpreting predictions across various machine learning algorithms, thereby improving their applicability in different analytical contexts.

IV. RESULT ANALYSIS AND DISCUSSION

In this paper we worked with Google Colab an online based IDE that provides free access to high-performance GPUs and TPUs. After data preprocessing, we have applied five ML models for the classification purpose of our research. Table 1 represent that the accuracy for both Negative and Positive classes is 87% while using the Logistic Regression model. The Precision, Recall and F1 Score of the Negative class are 89%, 85% and 87% respectively whereas for the Positive class these values are 86%, 89% and 89% respectively. Moreover, For Decision Tree we can see that the accuracy, precision, Recall and F1 Score for both negative and positive are 87%. Regarding Support Vector Machines, the following values apply: for negative classes, they are 90%, 92%, 88%, and 90% for accuracy, precision, recall, and F1 Score; for positive classes, they are 90%, 88%, 92%, and 90%. Additionally, we obtain a similar result of 89% for Random Forest, and lastly, for LGBM, the positive values are 85%, 89%, 78%, and 84%, and the negative values are 85%, 81%, 91%, and 85%. This indicates that the SVM model provided the highest accuracy.

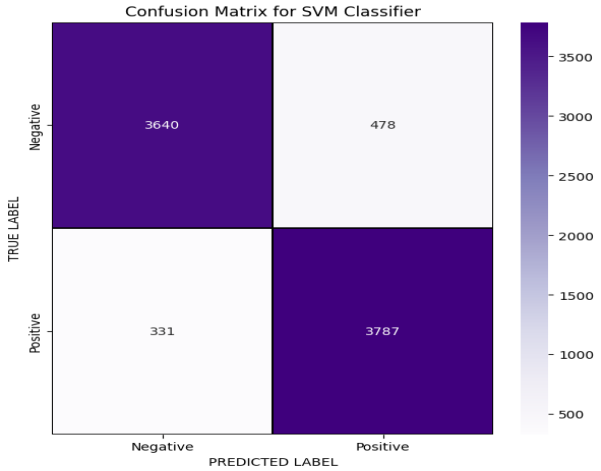

Fig. 1. Confusion Matrix of SVM Classifier

As evident in Figure 1, the confusion matrix provides remarkable insights into the SVM classifier's admirable performance. We can see that, 3640 data points were accurately classified as negative when they were indeed negative, alongside 3787 data points were correctly recognised as positive for each class, highlighting the model's adroitness, especially in conceding the second class. The poitns of misclassification are minimal, with only 331 data points inaccrately labeled as negative when they were actually positive, and 478 data points were misclassified as positive insted of negative which was the real value. These minimal numbers of misclassifications underscore the overall robustness of the model, highlighting its effectiveness in correctly classifying instances across different catagories.

As we can see in the Figure 2 above,We used SHAP on Logistic Regression model. SHAP values are based on game theory concept and it assign an importance value to

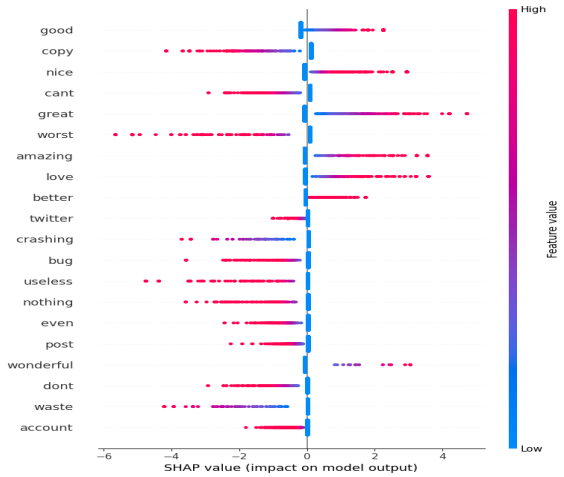

Fig. 2. Summary of SHAP values across all features for Logistic Regression

every feature in a model. Features with positive SHAP values positively impact the prediction just like the feature 'good', 'nice', 'great', 'amazing', 'love', 'better' and 'wonderful' and the magnitude or in other other the feature value represents the measure of how strong the effect is. The above mention features provide a significant positive contribution to the forecast because of their high feature value and positive SHAP value. Whereas features with negative SHAP values have a negative impact to model's prediction just like the features 'copy', 'worst', 'useless', 'nothing', 'waste' etc. and their high feature value represent that they have a strongly negative contribution to the prediction of the model.

V. CONCLUSION AND FUTURE WORK

1) Conclusion: We use the Kaggle dataset to train our model, and grip multiple layers to increase accuracy. By integrating sentiment analysis and explainable artificial intelligence (XAI) features into our machine learning algorithms, we aim to upgrade model prediction. We use different classifiers such as Decision tree, Random Forest, Logistic Regression, SVM, LGBM and get the best accuracy from the SVM model which is 90By exploiting sentiment analysis, stakeholders can identify areas for improvement, address user concerns, and grasp positive feedback. Integrating explainable AI further upgrade, the transparency and reliability of sentiment analysis, allowing for a deeper understanding of the model's decision-making process. To move forward, continuous improvement of sentiment analysis models will be important to adapt to changing language usage and user expectations. Cooperation between data scientists, linguists, and domain experts is essential to address domain-specific linguistic nuances and ensure accurate sentiment interpretation. Eventually, sentiment analysis in app reviews serves as a powerful tool for making informed decisions, driving a user-centric approach to product development and alteration service improvement. As technology continues to advance, the synergy between Explainable AI and sentiment analysis promises to deliver a deeper and

TABLE I
VALUES GIVEN OF ACCURACY, PRECISION, RECALL AND F1-SCORE RESPECTIVELY FOR POSTIVE AND NEGATIVE VALUES

	Negative				**Positive**			
Model Name	accuracy	precision	recall	F1 Score	accuracy	precision	recall	F1 score
Logistic Regression	0.87	0.89	0.85	0.87	0.87	0.86	0.89	0.88
Decision Tree	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
Support Vector Machine	**0.90**	**0.92**	**0.88**	**0.90**	**0.90**	**0.88**	**0.92**	**0.90**
Random Forest	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
LGBM	0.85	0.89	0.78	0.84	0.85	0.81	0.91	0.85

more dependable understanding of user emotions, helping to improve overall app quality ability and customer gratification.

2) Future Work: In future, more combinatorial models and more data layers can be used. Develop more refined models that can capture the subtle nuances of emotions, going beyond simple classifications of positive, negative. Detailed sentiment analysis can involve classifying emotions into specific emotions. Integrate analysis of text into the app review. Multimodal emotion analysis provides a more comprehensive understanding of user emotions by taking into account visual and auditory signals. Enlarge sentiment analysis capabilities to multiple languages, allowing businesses to gain insights from a global user base. Multilingual sentiment analysis is pivotal for businesses with users with various demographics. Acknowledge how sentiment changes over time in response to app updates, new features. Analytics over time can furnish a dynamic perspective, helping businesses adapt to changing user expectations. Go on the far side of sentiment classification to understand users' purpose and goals in their reviews. This can require recognizing whether the user is looking for help, suggesting improvements, or expressing acknowledgement. Further improving explain capability in sentiment analysis models. The capability to provide clear explanations of model detections increases confidence and facilitates better understanding of analysis results. Expand models that can adapt and learn from continuous user feedback and changing language trends. Continuous learning makes sure that sentiment analysis models remain effective and relevant over time. Generate more direct connections between sentiment analysis information and development. Integrating sentiment analysis into the user feedback loop can lead to faster responses to user disquietudes and more agile product development. Investigate the ethical imputation of sentiment analysis, considering issues such as privacy, bias, and responsible use of user data. Future work should focus on making sure fairness and transparency in sentiment analysis applications. Contribute to the development of standardized standard datasets and evaluation metrics for sentiment analysis. This will smooth fair comparisons between different models and approaches.

REFERENCES

[1] Adak, A.; Pradhan, B.; Shukla, N.; Alamri, A. Unboxing Deep Learning Model of Food Delivery Service Reviews Using Explainable Artificial Intelligence (XAI) Technique. Foods 2022, 11, 2019. https://doi.org/10.3390/foods 11142019

[2] A. Diwali, K. Saeedi, K. Dashtipour, M. Gogate, E. Cambria and A. Hussain, "Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2023.3296373.

[3] Adi, D.; Nurdin, N. Explainable Artificial Intelligence (Xai) Towards Model Personality in Nlp Task. IPTEK J. Eng. 2021, 7, 1. [CrossRef]

[4] Manjunatha, V.; Anneken, M.; Burkart, N.; Huber, M.F. Validation of Xai Explanations for Multivariate Time Series Classification in the Maritime Domain. J. Comput. Sci. 2022, 58, 101539.

[5] de Souza, L.A., Jr.; Mendel, R.; Strasser, S.; Ebigbo, A.; Probst, A.; Messmann, H.; Papa, J.P.; Palm, C. Convolutional Neural Networks for the Evaluation of Cancer in Barrett's Esophagus: Explainable Ai to Lighten up the Black-Box. Comput. Biol. Med. 2021, 135, 104578. [CrossRef] [PubMed]

[6] So, C. What Emotions Make One or Five Stars? Understanding Ratings of Online Product Reviews by Sentiment Analysis and Xai. In International Conference on Human-Computer Interaction; Springer: Berlin/Heidelberg, Germany, 2020

[7] Sakinah, S.N.; Mohamed, A.; Mutalib, S. Customer Reviews Analytics on Food Delivery Services in Social Media: A Review. IAES Int. J. Artif. Intell. 2020, 9, 691. [CrossRef]

[8] Adak, A.; Pradhan, B.; Shukla, N. Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review. Foods 2022, 11, 1500. [CrossRef] [PubMed]

[9] Chakriswaran, P.; Vincent, D.R.; Srinivasan, K.; Sharma, V.; Chang, C.-Y.; Reina, D.G. Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues. Appl. Sci. 2019, 9, 5462. [CrossRef]