National College of Ireland

Sentiment-Driven Credit Risk Analysis: Hybrid Framework with Country-Level Fusion and Explainable AI

MSc Research Project
MSc. Data Analytics

Ojas Shivadatta Acharekar
Student ID: x23268492

School of Computing
National College of Ireland

Supervisor: Prof. Qurrat Ul Ain

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	OJAS SHIVADATTA ACHAREKAR
Student ID:	x23268492
Programme:	MSc. Data Analytics **Year:** 2024-2025
Module:	MSc. Research Project
Supervisor:	Prof. Qurrat Ul Ain
Submission Due Date:	15th September 2025
Project Title:	Sentiment-Driven Credit Risk Analysis: Hybrid Framework with Country-Level Fusion and Explainable AI
Word Count:	8081 **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.

OJAS SHIVADATTA ACHAREKAR

Signature: .....

Date: 15th September 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):	

Sentiment-Driven Credit Risk Analysis: Hybrid Framework with Country-Level Fusion and Explainable AI

Ojas Shivadatta Acharekar
x23268492

Abstract

The statistical probability that the borrower may not meet the financial obligations, alerting the lenders and financial institutions before making a decision, is known as credit risk. Traditional credit risk models that were developed mainly depended on structured data such as income, credit history, and employment status, but often neglected the dynamic behavioral signals that can act as an indicator in the early stages of financial instability. In this study, this gap is being addressed by introducing a hybrid framework that integrates the structured financial attributes with the sentiment data extracted from unstructured financial tweets. The Deep Learning models (LSTM and BiLSTM) were used with GloVe embeddings for sentiment classification, which achieved an accuracy of 92.28% suggesting validation on user link dataset. A novel aspect in this study is that the structured and unstructured data are integrated based on the country-level mapping, which ultimately enriches the individual credit profiles by adding the aggregate national sentiment scores. The merging of the two datasets was done by using the country column in both datasets. In addition to this, Explainable AI (LIME) was used to interpret the predictions of the model and find the correlation between sentiment and creditworthiness. As per the findings, the Stacking ensemble model achieved an accuracy of 83.66% having the best performance. Overall, considering the financial sentiments, an approach that comprises a scalable and explainable way of assessing individual credit profiles was introduced, which is suitable and more useful in real-time financial applications.

Keywords: Credit Risk, Sentiment Analysis, LSTM, Machine Learning, Financial Tweets

1 Introduction

1.1 Aim of the study

The main aim of this study is to propose a framework for credit risk analysis that is scalable and explainable, this is achieved by fusion of two datasets that is structured financial attributes with unstructured sentiment data derived from the financial tweets' dataset. Traditional credit scoring systems rely on demographic and historical variables like income, credit history, employment status, salary and debt ratios, which in turn neglects the effect of real-time economic sentiment. Using advanced Natural language processing (NLP) techniques and deep learning models like Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) sentiment scores were extracted from social media tweets dataset for addressing the limitations identified in the traditional approaches. The key novelty in this study is the integration of sentiment data at country level, where average sentiment score is calculated for each country and is mapped to individual credit profiles based on their country of residence. This integration aims to capture the effect of macro-level sentiments that impacts the borrower's behavior. This study hypothesizes that sentiments when combined with the financial attributes can have impact on individual credit profiles, considering an additional

parameter while assessing a credit profile. In addition to this, Explainable AI (XAI) method such as LIME is used for exploring and visualizing the correlation between country-level sentiment and structured financial attributes. Moving ahead if any such correlations exist, then further analysis on how sentiment shift impacts the default probabilities can be carried out, which will help in improving predictive accuracy in credit risk in real-time.

1.2 Research Question

To what extent does integrating the sentiment scores extracted from financial tweets using deep learning models affect the accuracy, precision, recall, and F1 score of ML models in credit risk classification compared to models that only used structured financial data?

1.3 Objectives of the Research

The research objectives for this report are:

1. Applying advanced natural language processing techniques and using the sentiment analysis tools such VADER, GloVe, and TF-IDF embeddings sentiments data is to be extracted from the unstructured financial data taken from social media platforms, specifically Twitter and integrating this data with the structured credit risk attributes.

2. To develop and compare deep learning models (LSTM and BiLSTM) for performing sentiment classification of financial tweets, and determining the performance by generating the sentiment scores that can help in enriching the data for credit risk.

3. Merging the datasets on the basis of country and evaluating the performance of various machine learning algorithms like SVM, Random Forest, CatBoost, XGBoost, and Stacking for classification in credit risk, and then integrating it with Explainable AI for model interpretation and scalability for real-world financial applications.

1.4 Outline of the Report

The work carried out in this research is presented in this report in 7 sections. In section 1, the aim, objective and motivation of why integrating sentiments with credit risk prediction is stated. Section 2 covers relevant work carried out in this domain by reviewing the traditional credit scoring techniques, sentiment analysis in finance, and application of machine learning in risk modelling. In section 3 details of research methodology is defined which includes dataset description, data preprocessing, sentiment labeling, and merging of structured and unstructured data. Section 4 presents design of the system proposed as well as the architecture diagram and how data flows from sentiment classification to credit risk prediction phase. Section 5 discusses the implementation of deep learning models for sentiment extraction and merging two datasets and finally implementing various machine learning techniques for credit risk predictions. In section 6, evaluation of developed model is carried out on the basis of accuracy matrix and confusion matrices and model interpretation is done by integrating tools like LIME. In the very end, section 7 is conclusion where in all the research work carried out is summarized and key findings are mentioned also proposing direction for future work which includes improvements during model interpretability and large-scale deployment with real-time integration of sentiments.

2 Related Work

2.1 Credit Risk Analysis

2.1.1 Definition of Credit Risk

The analysis of credit risk is the principal technique of the financial and banking sector, which tries to assess the probability of failure among the debts of the borrower (Bhattacharya, et al., 2023). It is the key to sustaining financial viability, lending portfolio, and the reduction of the risk on loan issuance. Historically, the process of credit risk assessment was based on structured data, including the past credit history, income rates, debt balance, and occupation/employment status (Malik, et al., 2024), and therefore, statistical and machine learning models can be used to correlate such factors and forecast creditworthiness of a given borrower. Banks and other organizations depend on such analyses in order to make good decisions in terms of issuing loans, interest rates and capital buffers (Patel, et al., 2020). As digital banking and other non-conventional sources of data continue to grow, credit risk analysis has taken a different turn to now comprise simpler real-time evaluations that consider behavioural and psychological data (Kuna, 2025). Over the past few years, many researchers have embarked on incorporating the unstructured data, including social media sentiment and financial news updates, into credit risk models to improve them in predicting credit risks. This hybrid model reveals a superior way of observing the potential risk assigned to a borrower as it takes into consideration both quantitative variables and qualitative behavioural ones. Thus, the analysis of credit risk based on artificial intelligence, deep learning, and hybrid data modelling is further developed to implement more accurate and explicable risk assessment.

2.2 Sentiment Analysis in Finance

Opinion mining or sentiment analysis is a natural language processing (NLP) method to detect, extract, and divide those opinions or feelings that are presented in a text data (Jayasudha & Thilagu, 2022). When this method is applied to a financial context, sentiment analysis is concerned with identifying the feeling underlining the textual information in terms of positive/ negative/ neutral (Mir, et al., 2025) emotions that may include, the financial news, analyst reports, social media posts, investor forums. With the help of sentiment scores traditional financial model will have an additional parameter, which will be useful having real-time responses to the current market behaviour and investors psychology. The benefits of sentiment analysis in finance are the fact that it is able to provide an early indication in case of a possible risk. Take, as an example, the adverse sentiment on social media, reflecting the leadership, governance or product withdrawal of a certain firm, which may result in the decline of stock price or creditworthiness (Peivandizadeh, et al., 2024). Sentiment analysis models such as VADER or TextBlob or deep learning models can be used to perform analysis on tweets, headlines which are short in length and might have some noise in it. Sentiment analysis used in credit risk assessment can improve the assessment of a borrower by including the effect of economic news and in turn predicts the default. There are high correlations between financial sentiment analytics, market volatility, credit events, and loan frauds which proves to be useful in predictive modelling (Fu, et al., 2019). There are few issues which may arise if we consider sentiment analysis like ability to identify the sarcasm, use of jargon and ambiguity of context, all this may cause inaccuracy. Despite of these problems, sentiment analysis has proved to be useful in finance sector.

2.3 Traditional Credit Risk Analysis Techniques

The study proposed by (Hossain, 2023) is a comparative analysis for determining the efficiency of traditional and modern methods in credit risk assessment. This study aimed at comparing

the manual memo-based traditional credit risk assessment system with the newer emerging technologies which includes the usage of Artificial Intelligence (AI) and Machine Learning (ML). This research used combined method from personal experience in internship and organizational records which was carried out based on the existing research which used AI for credit scoring. The main motive of the proposed solution was to have a approach which depends on data to reduce the errors occurred by humans, which occur in traditional credit assessments and make it tough for the Micro, Small, and Medium Enterprises (MSMEs) to get a fair evaluation. The main problem faced was structuring the lack of real-time, organized data in existing systems, as well as not adopting the new technology-based models in th organization. The findings provided evidence that new types of AI/ML models could improve the predictive accuracy of credit risks and inclusiveness on a small-scale, such as MSMEs, with a high opportunity to scale. However, these models have high efficiency when exposed to high quality data and an excellent policy framework and it is found that both these things are not available in the emerging financial systems.

Recently, (Dastile, et al., 2020) suggested of have a comparison between simple statistical model that uses logistic regression with the more advanced models that are being used in credit scoring. The aim of the article was to identify the gaps present in the literature based on interpretability of models as well as data imbalance and in addition to these, propose a framework that could be followed in future when working with machine learning in credit risk assessment. This research followed one path and evaluated 74 original sources such as journals and conference papers which were published in 2010-2018. According to the meta-analysis, models that are trained individually and then combined in a particular way performs much better than a single model when it comes to accuracy. Additionally deep learning models were not used as per the studies reviewed at that time, but there is a high chance of having a good impact of those on credit scoring. Some key challenges occurred are that explaining how machine learning models make decision that can cause problem in financial sectors as well as another problem is the data imbalance which can have impact on credit scoring. One drawback of the study is that it only reviews the papers published before 2018 and may not have information about the latest trends. Also, it solely focuses on academic literature and this may not cover the real-world scenarios that are practised by financial institutions or third-party providers.

In the work of (Levy & Baha, 2021), they have proposed a comparative study of two traditional statistical models Logistic Regression (LR) and Linear Discriminant Analysis (LDA) so as to evaluate how better they perform in credit risk prediction as well as determining the borrower's reliability in small and medium sized enterprises (SMEs) in private sector of Algeria. This study aimed at finding which model's performance was found to be superior on the real-financial data in that region. The real-financial data was collected from SMEs in Algeria and then step by step applied LR and LDA models to classify the creditworthiness of the firms. The result of the study was that both models may be used for prediction of credit risk, but Logistic regression model was found to be performing well in most of the cases where there was no presence of non-linearity in data. The limitation of this research was not having enough of the quality data with respect to the SME finances so that model could be generalized and used. This research was not only limited to data but also traditional methods of analysis were used instead of using advanced methods of analysis that could have increased the accuracy. These models may not perform well on more diverse and larger dataset because these models were trained on specific dataset and usage of traditional methods of analysis.

2.4 ML Techniques in Credit Risk Analysis

(Arora & Kaur, 2020) suggested an effective method for credit risk assessment wherein they used Bolasso (Bootstrao-Lasso) for uniformity and proper feature selection along with various algorithms. Two research question focus on increasing the predicting accuracy as well as reducing the complexity of credit risk classification models by removing the features which are not relevant for further analysis. Bolasso is employed in this research to select important and relevant features from three different benchmark datasets, namely the Lending Club, Kaggle Bank Loan Status, and German Credit dataset, with 70:30 train test ratio. The results of the experiment shows that Bolasso-enabled Random Forest (BS-RF) has the better classification performance than other models in both the accuracy and AUC. This strategy was also tested against basic feature selection algorithms like Chi-Square, Gain Ratio and Relief and it was found that when Jaccard Stability Measure (JSM) was used as an evaluation criterion, Bolasso achieved more stability in terms of feature selection. The research trying to state that proper feature selection has an effect on the predictive performance. However, there is a limitation to this study that we may not get same results on different types of data. This is because the linear methods like lasso do not capture all the interactions between the dependant variables, which can have a negative impact on complex and imbalanced datasets.

The research proposed by (Uddin, et al., 2022) suggested to use Random Forest (RF) algorithm for interpretability of credit risk prediction on Chinese micro-enterprises' dataset because RF algorithm faces issue with the bias variance. The goal of this research is to perform credit risk analysis on groups formed using the dataset so model can be evaluated on different types of data, and thereafter examining how groups of related predictors have an impact on the result. Furthermore, the analysis evaluates how traditional financial variable and non-traditional factors such as non-financial indicators affect the predictions. The efficiency of the RF model is determined on the basis of five different performances and evidence shows that model works well in all the criteria. By including non-traditional indicators in dataset, the accuracy of classification is enhanced which is better and more sustainable approach. However, this study has certain limitations that is this study uses regional data, which may affect the results making it less relevant. Despite of the Random Forest algorithms capability of showing variable importance, it still does not offer deep insights or explanations and may not handle extreme class imbalances well unless extra tuning is done on the model.

In the study of (Teles, et al., 2021), it has been suggested to use ML-based algorithms like Support Vector Machines (SVM) and Random Forests (RF) to solve the corporate problems and manage the credit risk issue faced by the corporates using a large-scale financial data. This research focuses on bankruptcy issue face by the corporates in the financial global market and also discusses about how the managers make use of the quantitative and qualitative data to solve the problems being caught up early. In this study the researchers have used the latest tools and technologies of machine learning for credit scoring keeping the collateral variable aside and predicting how much loan value can be recovered. The study shows the comparison between the efficiency of SVM and RF and it is found that RF is more efficient in all perspective considering speed and running the model and therefore it can be used in real-time processing of credit risks, but SVM was found to perform well in terms of classification precision. The model is tested on various algorithmic tests and finding proved the effectiveness of the two approaches in financial risk analysis. A drawback of this study is that one has to be selected from efficiency of RF or accuracy of SVM, so it is quite clear to use combination of the two models. Also, the dependence on the 'collateral' variable could limit the model, making it less capable on different sets of data which do not have standard data.

According to (Abbasov, 2023) Gradient Boosting Machine (GBM) and Random Forest (RF) models can handle complex data interactions and can be used for predictions in the banking sector for achieving a high accuracy. This paper focuses on improving the accuracy of default loan prediction by using structured machine learning approaches that involves cleaning, preprocessing and preparing the data including demographic and financial indicators. During the training and validation stage performance of both the models was measured for better results. The analysis shows that such ensemble methods perform well and play a crucial role in making a model for better and informed decision systems lowering the default rates and helping the financial or banking sectors. The findings show the results that predictive performance of both the models was very good, especially the GBM which handled the complex interactions very well. But on the contrary, there is a limitation in terms of interpretability, although these models are very accurate and very useful in decision making process in regulatory situations.

The study proposed by (Schmitt, 2022) shows a comparative analysis of Deep Learning (DL) and Gradient Boosting Machine (GBM) to find out which model is effective for prediction across various sets of data. As from previous papers discussed in the literature review performance of model depends on the nature of data so in this study the two algorithms are compared based on the three different types of datasets which is divided based on the structure. The aim is to have a system that is scalable, efficient and is able to work well in the real-world scenarios. Experiments carried out during this research shows that GBM performs well most of the time in comparison to DL in terms of predictive power, training time and being able to implement it easily, all these points make the GBM model stand out and a preferred choice in most of the cases with respect to structured credit scoring tasks. However, we know that every coin has two sides the same is with GBM model it always does not perform the best. Sometimes it's performance may degrade based on the characteristics of dataset. To conclude with, both DL and GBM models are advanced techniques for a binary classification task, but GBM model is preferred because of its ability to handle complex and intricate data patterns, less training time and comparatively high accuracy.

The study conducted by (Sakri, 2022) offers the comparative analysis of Deep Neural Networks (DNN) and Gradient Boosting Machines (GBM) with objective of evaluating their superiority in terms of classification and determine which model is suitable for usage in financial terms. Addressing the gap in the existing research direct comparison of these two leading algorithms is done on three different datasets and also applied hyperparameter tuning to both GBM and DNN models by using the activation functions like ReLU, Maxout, Tanh. To analyse the models in a perfect and ordered manner that will guarantee good evaluation, the AUC ROC is used with both 80:20 and 70:30 train-test splits. The result of the experiment shows that both GBM and DNN achieve high accuracy in classification, but GBM was found to perform well than DNN in terms of training speed, computational efficiency and ease in implementation. These are the reasons why GBM is considered for credit scoring applications. The study however observes that although GBM has the upper hand in dataset that are structured, DNN might hold the edge in a more detailed or unstructured data environment which is not covered by the study. Therefore, one of the main drawbacks is the fact that the analysis is limited to structured credit data only, there is a possible risk of underestimating the overall competence of DNN architectures.

3 Research Methodology

3.1 Dataset Description

In my research I have used two different datasets but both the datasets are complimentary to each other and I have carried out credit risk analysis on it with the primary aim of integrating real-time sentiment analysis. Discussing about the first dataset it is being taken from Kaggle, is a structured financial risk dataset with having all the variables that are needed for credit risk analysis. These variables include income, age, occupation, number of dependents, existing debts, and loan repayment history of the borrower. As we know these attributes are basic for judging a person on the credit part and many of the financial institutions are using this information for credit risk analysis.

The second dataset is about the financial tweets, which is unstructured text data, this data is also taken from Kaggle but the tweets in this dataset are from twitter which includes information related to finance industry, market events and company statements. This data was used for performing sentiment analysis and getting information about financial markets and institutions and what are the opinions of public. Unlike the structured data as in the first dataset, tweets are dynamic and talks about real-time behaviour and psychology, this information can serve as an early indicator for anything found to be a risk in financial terms. Additionally in this dataset country column was derived based on the already existing columns namely company and source columns.

To move ahead in the research, I have merged two datasets based on the country column. With this integration, sentiment data for social media publications with respect to the country is being added to every record of credit risk. The tweets data was pre-processed by using the techniques like noise removal, stop words, eradication, punctuation treatment and embedding words in GloVe and Word2Vec. The sentiment labels namely positive, negative, and neutral were created using the VADER sentiment analyser.

Such hybridized dataset solution, which consists of two parts of structured financial data and non-structured sentiment scores, helps in creating a more sustainable machine learning model. It satisfies the purpose of the study, which is to determine the effectiveness of adding the sentiments of the people that may help in increasing accuracy of credit risk prediction models. The developed data is, based on both conventional risk indicators and real-time emotional conditions that provide a multidimensional insight into creditworthiness analysis.

3.2 Merging two datasets

In order to improve the predictive powers of credit risk assessment, this study incorporates behavioral sentiment data with financial features through the merging of two separate datasets, including credit risk data and financial tweets data. The tweets dataset after a thorough preprocessing and sentiment analysis with the help of VADER sentiment analyzer gave us sentiment scores that were positive, neutral, or negative per country. These ratings would then be summarized and created as country_sentiment which becomes a new feature which indicates the prevailing financial mood of a country. In order to do the merge efficiently a dictionary (sentiment_dict) was created which mapped with the country with the corresponding sentiment score using the tweets dataset. This enabled a mapping function. The most important column that was to be applied in merging the datasets was the Country column, which was shared between them. The records in the credit risk dataset were then enhanced with sentiment score of the corresponding country by using a lambda function which used the country key to look into the sentiment dictionary. When a country was not found in the twitters data, a default

sentiment score, usually the score of 'Other' or a neutral backup, was given so that data completeness is achieved and no null answer can occur. The practical outcome of that merging procedure was the unified data set with traditional demographics, financial, and behavioral variables, and the new column country sentiment. This augmented set of data became essential to the hybrid modeling since it includes quantitative attributes such as income, credit score and debt-to-income ratio along with qualitative sentiment markers that would help to identify the risk associated with the borrower. A pictorial representation of the combined data demonstrates the congruence of the sentiment polarity with systematic variables such as employment status, education, and risk rating, which has the potential of providing a new dimension of behavior into credit scoring.

3.3 Phase 1: Sentiment Extraction from Tweets

3.3.1 Text Cleaning

In Phase 1 of the research, cleaning of text was done on raw financial tweets using multiple steps that include various inbuilt python libraries such as nltk, re, and wordninja. The transformation of noisy, unstructured tweet data into clean, analysable text was carried out so that text can be made suitable for sentiment analysis. First step was to remove missing values were handled then URL and HTML entity was removed from each tweet which was then followed by conversion of tweet data into lowercase using a regex-based function. To maintain semantic nature of the words "can't" and "won't" they be changed to "can not" and "will not". Using regular expression punctuations and characters which were not printable were removed. Moreover, words containing characters and numbers (like "USD123") were separated. To reduce redundancy, consecutive repeated characters (e.g., "goooood") were shortened to a maximum of two characters using a user defined regular expression function. Words joined together without spaces were intelligently split using wordninja library. In the end, stop words (e.g., “the,” “is,” “at”) and single-character or irrelevant short words were removed from the token stream. These steps were performed on the entire tweet dataset using the lambda functions and generating a new column named clean_tweet. This clean textual data served as a strong foundation for effective vectorization, embedding, and sentiment classification using deep learning models.

3.3.2 Preprocessing

After cleaning the tweets using multiple libraries of python the next step was to do preprocessing through vectorization and embedding to convert textual data into numerical values. Sentiment labels were also encoded by creating a mapping and applying it on the sentiment column, negative = 0, neutral = 1, positive = 2. For extraction of the features from the text, Term Frequency – Inverse Document Frequency (TF-IDF) vectorization was applied on the clean text using a maximum of 300 features. The Tfidf Vectorizer learned from the vocabulary and transformed the text into numerical matrix, resulting in a structured array which was used for training a deep learning model such as LSTM and BiLSTM.

3.3.3 Labelling

The VADER (Valence Aware Dictionary and Sentiment Reasoner) Sentiment Analyzer from the NLTK library was used for labelling the tweets due to its effectiveness in analysing short, informal text. Cleaned text was passed through the VADER analyser, which returned sentiment score ranging from -1 (most negative) to +1 (most positive). Based on thresholds defined, tweets were classified into three types: negative (score $\leq$ -0.05), neutral (between -0.05 and 0.05), and positive ($\geq$ 0.05). These sentiment labels were then used as the target variable for training and evaluating classification models in the next steps.

3.3.4 Visualization

Figure 1 is a horizontal bar chart that talks about the number of positive, negative and neutral sentiment labels assigned to the clean financial tweets which were analysed using the VADER algorithm. The bar chart illustrates that the neutral tweets dominate the dataset with more than 12000 records, while 5000 records assigned to positive and around 2000 tweets were assigned negative. This creates class imbalance in terms of classification problem with neutral values dominating the dataset. Such uneven distribution should be considered while training a model, because it can lead to biased results unless methods like oversampling or SMOTE is used on the dataset to balance it.

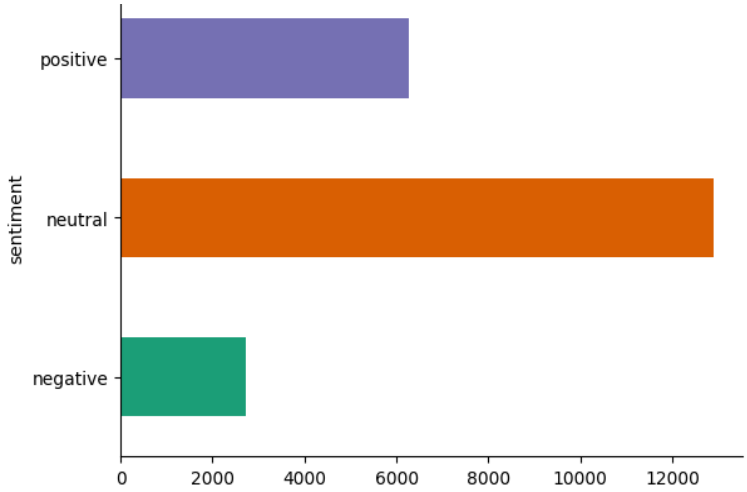

Figure 1: Count Plot

Figure 2 is a WordCloud visualization that highlights most common keywords in the textual data, and it is grouped by sentiment labels. So we can clearly see that which are the frequent words used in a particular label. In the positive sentiment group words like "binance", "max", "new insider", "options" appear most often showing sense of optimism in stock trading terminology. Negative label words are “per share,” “earnings per quarter,” and “SEC filing” which tells us about the problems regarding company performance. Neutral tweets include words like “join us,” “register,” “bonus,” and “investment research” which is kind of general information.

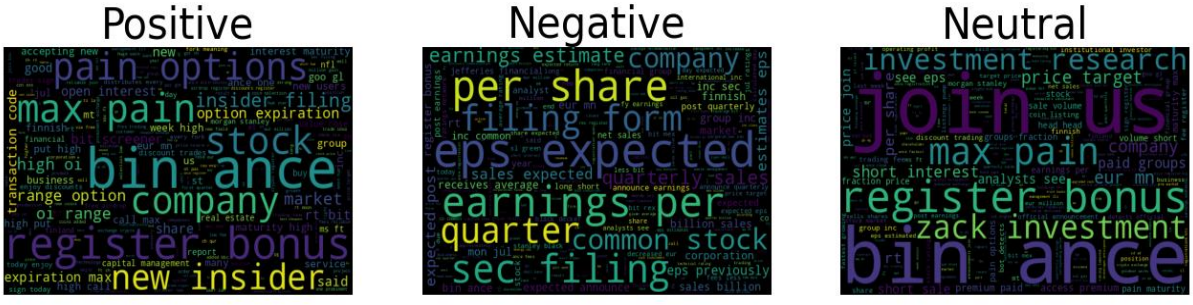

Figure 2: Word Cloud

Figure 3 shows the length of the tweets which is calculated based on the count of words. From the figure it is clear that most of the tweets have length between 30 to 90, with majority to be seen at 60 words in a tweet, which is around 4000 tweets. As we move to the right in the histogram figure frequency of the words per tweet decreases drastically. It can also be stated that very few tweets go beyond 150 words and no tweet exceeds than 250. This gives a good piece of information regarding the financial tweets that they are short which also fits twitter character limit. This information can be used to have a maximum length of 500 words.

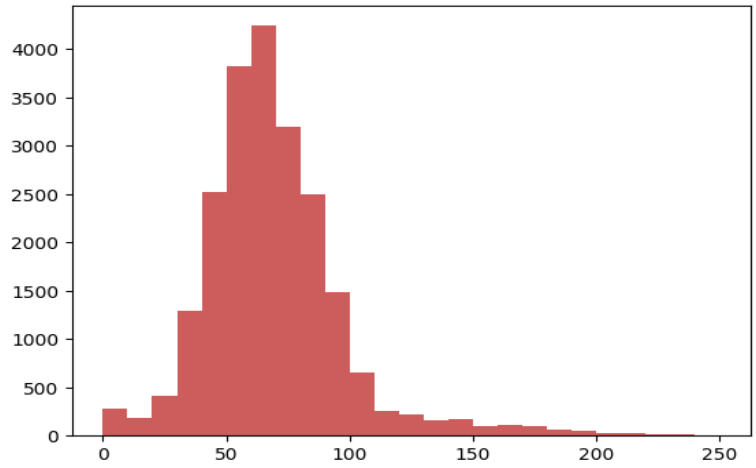

Figure 3: Histogram

3.4 Phase 2: Credit Risk Prediction using Sentiment Scores

3.4.1 Preprocessing

In Phase 2 of the research, data preprocessing was required because I have to merge the dataset, which is created using structured credit risk data and sentiment scores. Firstly, the missing values in numerical columns handled by using statistical mean of the values in that particular column to keep data consistency throughout the dataset. Data type of columns like “Previous Defaults” and “Number of Dependents” were converted to integer as the column in this value will always be a whole number. Next step was to apply encoding technique on the categorical variables using the label encoding. Then MinMaxScaler technique was used to scale the values within the range 0 to 1 making sure data is treated overall the same. Then, conducted elementary EDA on the data in case of null values. These were the steps carried out in phase 2 on the merged dataset.

3.4.2 EDA

Figure 4 refers to a bar chart that indicates the number of credit profiles related to risk categories of the dataset. The X-axis includes the three categories namely High, medium and low while Y-axis tells us about the number of records per category. From the bar chart it is clear that most of the credit profiles fall under 'low' risk category with around 9500 records followed by 'medium' with 5000 records and least were associated with 'high' risk accounting 2000 entries.

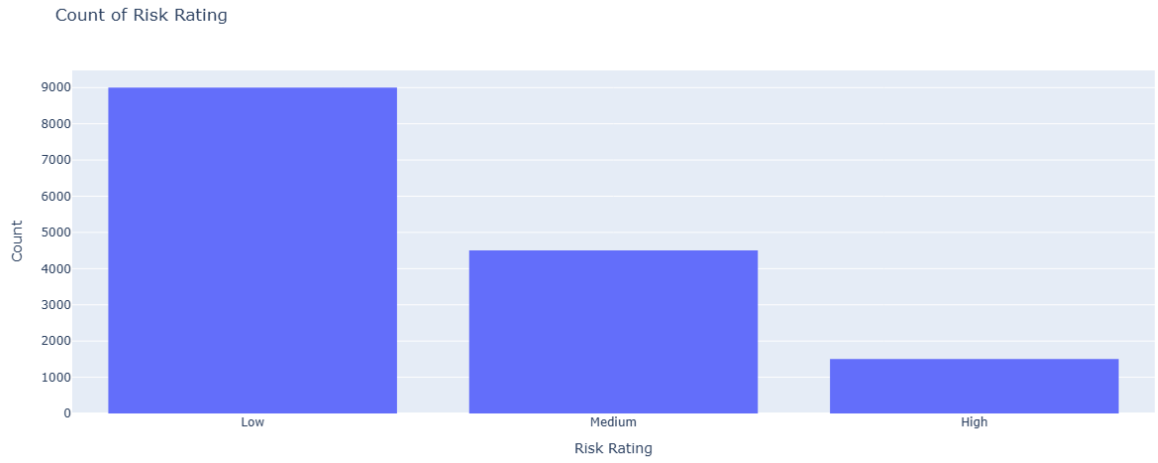

Figure 4: Bar Chart

Figure 5 is a boxplot which depicts difference between income levels across three different categories of risk. Each box is a interquartile range with the medium income falling between 80K to 85K. The total earnings are between 20K and up to 125K maximum.

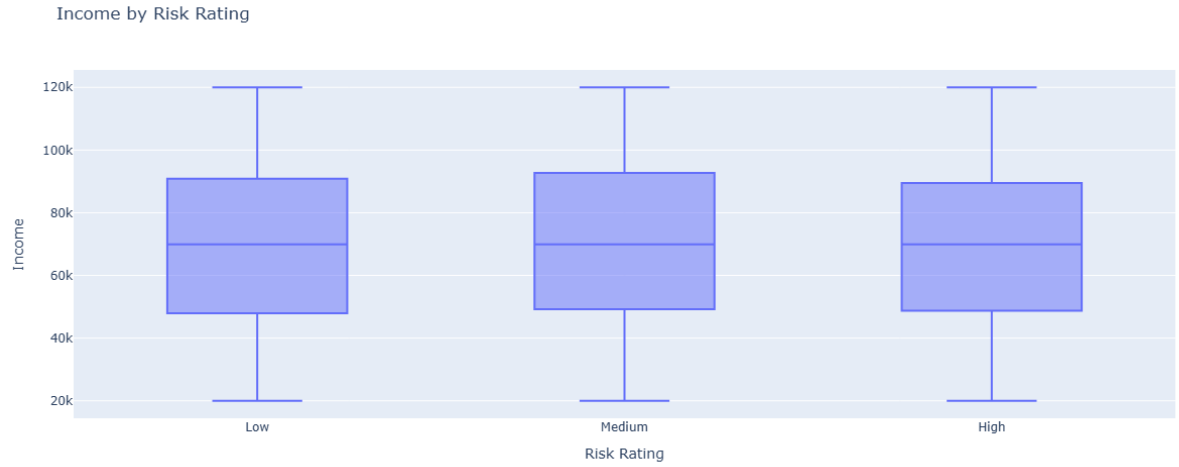

Figure 5: Box Plot

3.4.3 Balancing

From previous preprocessing steps we know that there exists class imbalance in the dataset so to tackle this problem, Synthetic Minority Over-Sampling Technique (SMOTE) was used. In financial data, high-risk applicants are fewer so this might create a bias towards the end result in the model. SMOTE technique generates new records that are real for minority classes, in this case, for high-risk applicants, without repeating the existing records. SMOTE is applied on the data after splitting it into features and labels in order to make sure that all three risk categories have equal number of entries.

Figure 6 illustrates the count of all three categories after applying SMOTE technique. The bar chart shows the number of records in each of category that have approximately 9000 entries. The use of different colors (blue, pink and yellow) makes the chart more readable and easier to interpret and shows that classes are equally represented. Before the SMOTE, the data was imbalanced and may have affected the end results, leading to bias in the low-risk and medium-risk categories. Now with the balanced dataset, this research has a chance to make fair and unbiased predictions.

Figure 6: Value Counts of Each Risk Label After SMOTE Application

4 Design Specification

Figure 7 is the complete system architecture diagram for the proposed credit risk analysis framework integrating sentiment scores. The system is divided into two major phases. Phase 1 is Tweet Sentiment Classification begins with cleaning the tweet data by removing stop words, punctuation, symbols, and handling white spaces. The cleaned and embedded tweets are then labeled using the VADER sentiment analyzer. The next step was text embedding, which is performed using GloVe, Word2Vec, and WordStore. These labelled sentiments are used to train deep learning models such as LSTM and BiLSTM, with data split in an 80:10:10 ratio. Evaluation is done using a confusion matrix and accuracy.

The output sentiment scores are averaged and then merged with the credit risk dataset based on country, creating a sentiment-enhanced dataset. In Phase 2: Credit Risk Analysis, the integrated dataset undergoes preprocessing, null value removal, label encoding, and normalization using MinMaxScaler. Exploratory Data Analysis (EDA) is performed using correlation and visual plots. To handle class imbalance, SMOTE oversampling is applied. Multiple machine learning models, SVM, Random Forest, XGBoost, CatBoost, and Stacking, are trained, followed by model interpretability using LIME, which highlights important features. The final output is an enhanced credit risk prediction incorporating sentiment insights.

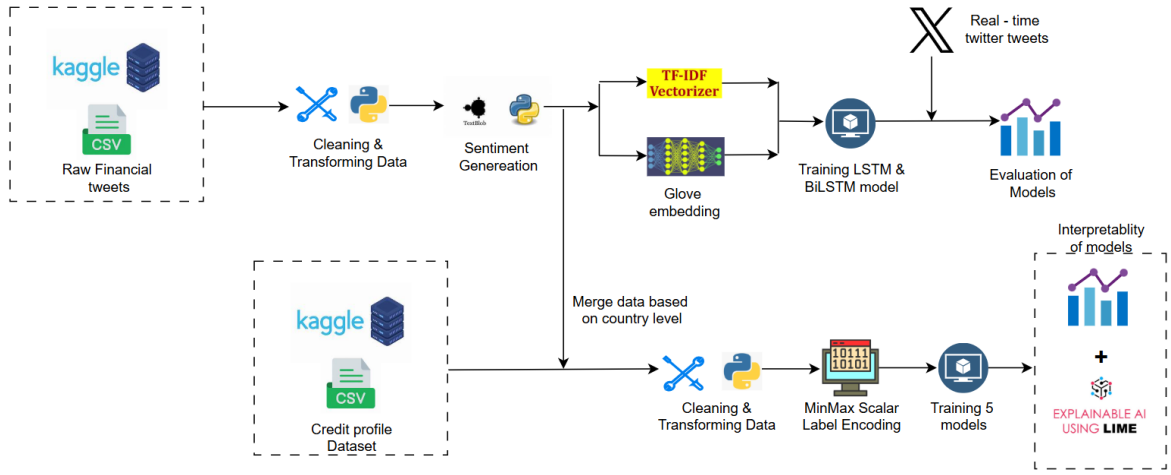

Figure 7: System Architecture Diagram

5 Implementation

5.1 Model Training – Phase 1: Sentiment Classification Models

In Phase 1 of the implementation, LSTM and BiLSTM models were employed using two different embedding techniques one is the TF-IDF vectorization and other is GloVe embedding. These models were trained to classify the tweets into three different types. LSTM captured the sequential dependencies while BiLSTM processed the text to make the text richer in context. Pre-trained embeddings were used by both the models for classification.

5.1.1 LSTM

To implement LSTM for sentiment classification in Phase 1, GloVe embeddings were used to create a pre-trained embedding matrix. Keras Tokenizer was used to tokenize the text in every tweet and after the tokenization process these tokens were converted into a sequence which had a maximum length of 500 words. For every word in the dataset, its 300-dimensional GloVe vectors were added to the matrix and this remained as a fixed layer and was used to preserve the semantic meaning of the word during the model training. The architecture for the trained

model had the very first layer as GloVe embedding layer which was followed by LSTM layer with 32 units that helped in capturing the information present in the sequence. The output of this layer was flattened and then passed through a Dense Layer which had 16 units and was activated by ReLU function. which was then followed by dropout layer with rate of 0.5 to help prevent overfitting. In the end, softmax layer was used to classify each tweet into three sentiments. Adam optimizer was used along with the loss function to compile the model which is suitable for multi-class classification. The dataset was split into training (81%), validation (9%) and 10% for testing. The final model comprised of 5.6 million parameters out of which 5.3 million came from GloVe layer and were not changed during the training.

5.1.2 BiLSTM

The BiLSTM model for sentiment classification was implemented using a pre-trained GloVe embedding matrix with 300-dimensional vectors. Tweets were first tokenized and padded to a maximum sequence length of 500. The Keras Tokenizer generated word indices used to build a GloVe-based embedding matrix, which was loaded into a non-trainable Embedding layer. The model began with this embedding layer, followed by a Bidirectional LSTM layer with 32 units and return_sequences=True to capture context from both forward and backward directions in the text. The output was then flattened and passed through a dense layer of 16 neurons using ReLU activation, followed by a dropout layer set to 0.5 for regularization. The final dense layer used softmax activation to classify tweets into three sentiment categories: positive, negative, or neutral. The model was compiled using the Adam optimizer and sparse categorical crossentropy loss. Dataset splitting was maintained as 81% training, 9% validation, and 10% testing. The model architecture included approximately 5.94 million parameters, of which 5.34 million were non-trainable from GloVe embeddings and 597,315 were trainable. This bidirectional setup improved context learning and produced higher performance than standard LSTM by effectively handling complex sentiment patterns in financial tweet data.

5.2 Model Training – Phase 2: Credit Risk Classification Models

This section is about the implementation of five different supervised machine learning models for classification of credit risk using structured financial attributes merged with sentiment data extracted from the unstructured tweet data. Implementation also involves steps like tuning of hyperparameters, configuring model, training and prediction steps for each classifier.

5.2.1 SVM

To create the Support Vector Machine (SVM) model, the SVC () classifier was used along with the hyperparameter tuning via the GridSearchCV. A parameter grid was defined with varying values for regularization parameter C (0.1,1,10), kernel coefficient gamma (1,0.1,0.01) and kernel types (linear,rbf). Along with these parameters fivefold cross-validation was used for identifying best model configuration. The best parameters for SVM model were C=10, gamma = 1 and kernel = 'rbf'. Using these parameter final SVM model was trained on balanced dataset. Predictions that were generated were tested and results were evaluated using standard classification metrics.

5.2.2 Random Forest

The Random Forest model was implemented using RandomForestClassifier with hyperparameter tuning based on a predefined parameter grid. The grid included variations in the number of estimators (10 to 50), maximum tree depth (None, 10, 20), and splitting rules (min_samples_split and min_samples_leaf), along with bootstrap settings. Based on evaluation, the optimal configuration was found to be n_estimators=50, max_depth=20,

min_samples_split=2, min_samples_leaf=1, and bootstrap=False. Using these parameters, the final model was trained on the processed training data. Predictions were then generated on test set using predict(), evaluated using accuracy, confusion matrix, and classification report.

5.2.3 CatBoost

The CatBoost model was implemented using CatBoostClassifier with hyperparameter tuning performed over a grid that included iterations (25, 50), learning_rate (0.01, 0.1), and depth (8, 10, 15). The best configuration identified was depth=15, iterations=50, and learning_rate=0.1, which offered a good balance between model complexity and generalization. The model was initialized with verbose=0 to suppress intermediate output and trained on the processed training dataset. Predictions on the test data were generated using the predict() function. The CatBoost model performed efficiently and was particularly effective in handling non-linear relationships in credit risk features.

5.2.4 XGBoost

The XGBoost model was implemented using XGBClassifier with hyperparameter tuning performed on a parameter grid including n_estimators (10–50), learning_rate (0.01–0.1), and max_depth (3–10). The best parameters selected were learning_rate=0.1, max_depth=10, and n_estimators=50, which balanced model complexity and generalization. The final model was initialized using these optimal values and trained on the credit risk dataset using the fit() function. Predictions were then generated on the test set using predict(). XGBoost performed robustly, handling class imbalances and complex relationships well, and delivered high accuracy with excellent precision-recall trade-offs in multi-class classification.

5.2.5 Stacking

The stacking model was implemented by combining multiple strong learners to enhance prediction performance. The base classifiers included an SVM with RBF kernel, a CatBoost model with depth 15 and 50 iterations, and a Random Forest with 50 estimators and max depth 20. These diverse models captured various data patterns individually. An XGBoost classifier, tuned with learning_rate=0.1, max_depth=10, and n_estimators=50, was used as the meta-classifier to learn from the base models' outputs. The StackingClassifier was then trained on the processed training data. Predictions were generated on the test set using predict(), yielding the highest overall accuracy among all tested models.

5.2.6 Explainable AI with LIME for Credit Risk Classification

In this section, LIME is implemented to interpret the predictions of the credit risk classification model. The LimeTabularExplainer is initialized using the training feature matrix (X_train.values) and corresponding class labels (y_train). Feature names are dynamically extracted from the dataset for clarity. The explainer is configured in classification mode, and continuous features are discretized to simplify interpretation. Once initialized, LIME can generate local explanations for individual predictions by approximating the model's behavior around a specific data point using a simple, interpretable surrogate model. This helps in understanding why the model classified a case as high or low credit risk.

6 Evaluation

6.1 Case Study 1: Results of Sentiment Classification Models

Figure 8 shows the accuracy comparison of four deep learning models used for financial tweet sentiment classification in Phase 1. Two embedding techniques were applied namely TF-IDF and GloVe, across both LSTM and BiLSTM architectures. The BiLSTM with GloVe

embedding achieved the highest accuracy of 92.28%, slightly outperforming LSTM with GloVe, which scored 91.73%. In contrast, models using TF-IDF embeddings performed comparatively lower, with LSTM at 80.82% and BiLSTM at 80.22% as shown in Table 1. These results indicate that semantic-rich embeddings like GloVe significantly enhance classification performance.

Figure 8: Model Accuracy Comparison

Table 1 shows the performance of four deep learning models trained using different embedding techniques for sentiment classification of financial tweets. We can clearly see that models using GloVe embedding outperform the TF-IDF vectorization by achieving an accuracy of 0.91 and 0.92. BiLSTM with GloVe embedding achieved highest accuracy but tied in terms of F1 score. TF-DF models indicated low recall value, telling us that limited ability to detect all sentiment classes effectively. The results confirmed that sentiment rich GloVe embedding enhanced the classification performance with respect to TF-IDF vectorization technique.

Table 1: Performance Comparison of Sentiment Classification Models

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1 (Avg)
BiLSTM + TF-IDF	0.8	0.87	0.67	0.72
LSTM + TF-IDF	0.81	0.85	0.68	0.73
BiLSTM + GloVe	0.92	0.9	0.87	0.88
LSTM + GloVe	0.91	0.9	0.86	0.88

```
================================================================================================
Text : Margin calls at @Robinhood can crush your #credit score if you're over-leveraged - stay liquid or face the music, folks!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 72ms/step
Predicted label for GLV_BiLSTM Model: negative
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 76ms/step
Predicted label for GLV_LSTM Model: positive
```

Figure 9: Model Test on Real-time data from Twitter

As we have accuracy scores for each model that were trained, but selecting a model only on the scores is not a good decision, we need to have a solid reason to choose a model. The model needs to be tested on data, and then a final decision should be made. So, for that, real-time tweets from Twitter were passed to the trained models and then evaluated. Figure 9 shows the result of the two models on the tweet taken from twitter, as we read the tweet it tells that, the credit score is impacted and may fall, so which is a negative tweet and it is quite clear from the

result that BiLSTM model predicted it correctly and testing on the other tweets showcased the ability of BiLSTM model to capture more insights from the text data, which shows BiLSTM model with GloVe performs slightly better than LSTM model with GloVe embeddings.

6.2 Case Study 2: Results of Credit Risk Classification Models

Figure 10 displays the accuracy scores of five machine learning models applied during Phase 2 for credit risk classification. The Stacking ensemble model emerged as the best-performing model, achieving the highest accuracy of 83.66%, slightly ahead of XGBoost, which scored 83.22%. Ensemble models clearly outperformed standalone algorithms. CatBoost and Random Forest followed with 76.44% and 76.11%, while SVM performed the lowest with 73.07% as shown in Table 2. The results confirm that combining multiple models through stacking enhances predictive power, making it the most effective approach for accurate credit risk analysis.

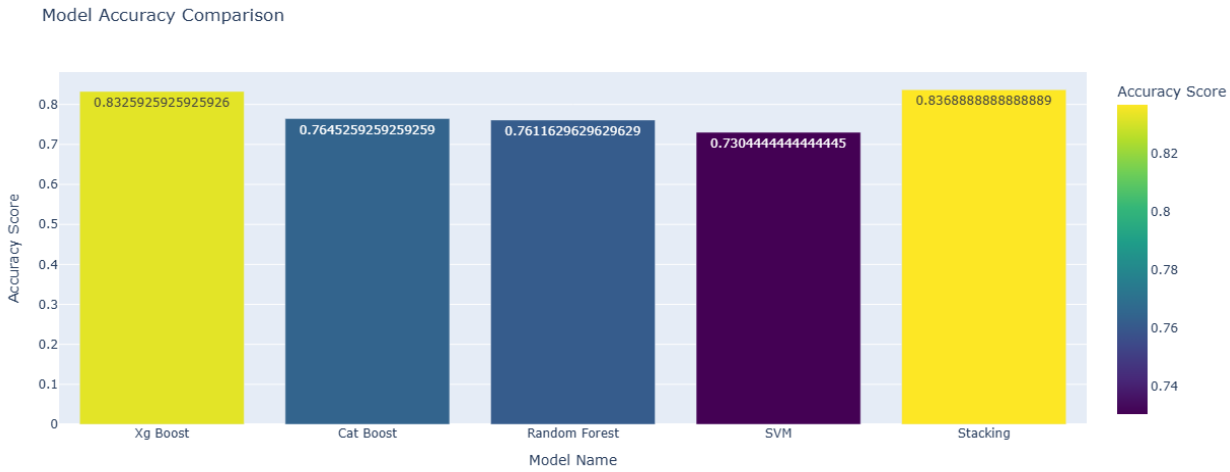

Figure 10: Model Accuracy Comparison

Table 2 depicts the statistics of five ML models trained for credit risk classification using the classification metrics. It can be seen that the stacking model achieved the highest accuracy of 0.84 and precision of 0.85. While XGBoost close to the readings. SVM model score low in terms of in most of the metrics. The results show that ensemble approaches particulary stacking, provide good predictions over individual classifiers.

Table 2: Performance Comparison of Credit Risk Classification Models

Model	Accuracy	Precision (Avg)	Recall (Avg)	F1 (Avg)
SVM	0.73	0.76	0.73	0.72
RandomForest	0.76	0.79	0.76	0.76
CatBoost	0.76	0.78	0.76	0.75
XGBoost	0.83	0.83	0.83	0.83
Stacking	0.84	0.85	0.84	0.83

6.3 Case Study 3: Results of Credit Risk Classification on Test Data

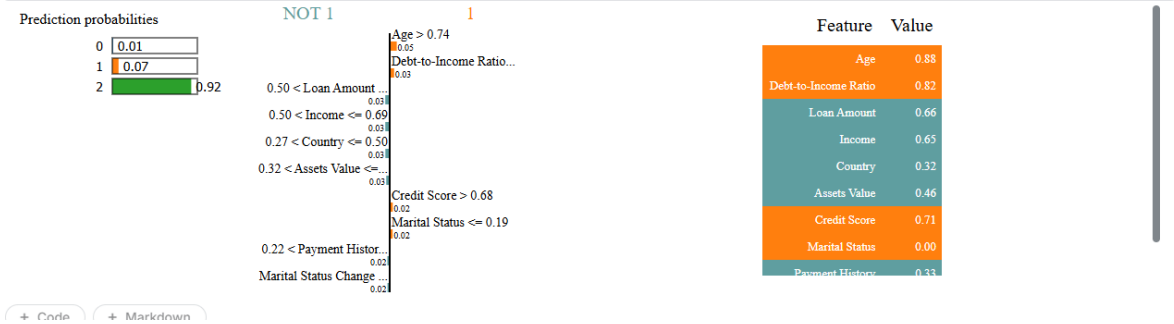

Figure 11.a: Lime Explainability of Credit Profile

Testing on a few test data sets was done using the stacking model, and LIME was used to learn about which features have the highest impact on a particular credit profile. Lime helped to explain the interpretability of each credit profile, and explained which factor is affecting the credit risk parameter. From Figure 11.a and 11.b we can clearly see that the records have different probabilities based on the sentiment, where in 11.b tells that most of the factors aligned with that profile contribute to default based on tweets' sentiment probability of 99% (0 = negative).

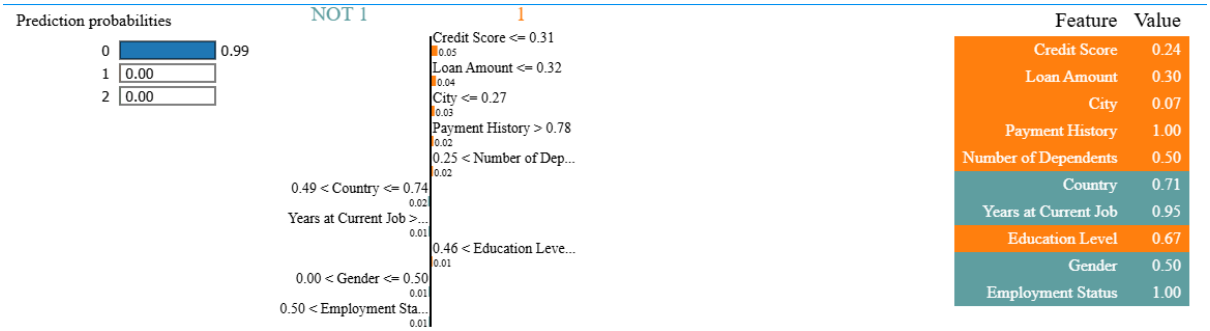

Figure 11.b: Lime Explainability of Credit Profile

Critical Analysis of why sentiment analysis ?

The use of sentiment data in credit risk analysis has the upper hand in measuring investor psychology and market-wide responses to events in the economy or a company, which can often be traced in real-time using social media. These indicators can help give early warning signals of financial instability that would otherwise not be known based on traditional attributes of credit. The country-level sentiment aggregation, however, comes with a lot of limitations. Although it places an emphasis on macroeconomic patterns, there is a risk that it smears out individual-level behavioral indicators, rendering the influence of sentiment on individual borrower profiles less definitive. Also, emotion based on Twitter data is subject to noise, sarcasm and demographic bias, which could lead to a decrease in reliability. The fact that there is no additional unique identifier to associate structured borrower data with individual sentiment further restricts validation, preventing the chance to extract strong causal claims. Therefore, despite the results that sentiment-enriching models can be useful in terms of improving predictive performance, the results should be regarded with caution due to the potential and limitations of such an approach.

7 Conclusion and Future Work

In this study, a novel credit risk score prediction using both the structured traditional financial data and the unstructured sentiment data using the financial tweets has been illustrated. In Phase 1, sentiment classification was carried out based on deep learning model (LSTM and BiLSTM)

and (Glove and TF-IDF) embedding tools. BiLSTM in BiLSTM with GloVe has accomplished an optimal accuracy of 92.28 percent, which provides evidence for the feasibility of semantic-rich embeddings in learning financial sentiment. Phase 2 further elaborated a process involving the training of various machine learning models leveraging structured financial data with country-level sentiment scores. Out of them, the Stacking ensemble model produced the most accurate events of credit risk classification of 83.66% which is larger than the other models like XGBoost and Random Forest. An effort was made to increase the transparency and confidence associated with the models by applying LIME to explain the predictions, the model makes on an individual basis so that stakeholders can decide. It is a sustainable and explainable scalable framework that the financial institutions can adapt to. In this study so far, it was observed that by using LIME, model interpretability was achieved, and LIME also helped in explaining the features associated with the credit profiles and how much they contribute to credit risk. Very few credit profiles were found to have an impact because of the sentiments. This tells us that there is some relation between sentiments and credit profiles. This work could be extended in the future by having sentiments related to individuals credit profiles and also to use multilingual tweet analysis, real-time sentiment streams, and domain-specific lexicons to get a better signal of behaviour. Furthermore, using sophisticated explainability methods of AI, such as SHAP or counterfactual explanations, can enhance the interpretability of a model as well. Finally, the delivery on a cloud hosting framework and API support to provide real-time evaluation can bring the solution to the industrial level and enable it to conduct large-scale credit evaluation.

References

Abbasov, R., 2023. Revolutionizing risk management in banking: Implementation of AI/ML-based gradient boosting machines (GBM) and random forest models for credit risk management. *International Journal of Research in Finance and Management,* 6(1), pp. 441--444.

Arora, N. & Kaur, P. D., 2020. A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing,* 86(1568-4946), p. 105936.

Bhattacharya, A., Biswas, S. K. & Mandal, A., 2023. Credit risk evaluation: a comprehensive study. *Multimedia Tools and Applications,* 82(12), pp. 18217--18267.

Dastile, X., Celik, T. & Potsane, M., 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing,* Volume 91, p. 106263.

Fu, X. et al., 2019. A Sentiment-Aware Trading Volume Prediction Model for P2P Market Using LSTM. *IEEE Access,* Volume 7, pp. 81934-81944.

Hossain, N., 2023. A comparative analysis of conventional and modern methods of credit risk assessment in financial institutions: implications for micro, small and medium enterprises (MSMEs).

Jayasudha, J. & Thilagu, M., 2022. A survey on sentimental analysis of student reviews using natural language processing (NLP) and text mining. In: Springer, ed. *International conference on innovations in intelligent computing and communications.* s.l.:Springer, pp. 365--378.

Kuna, A., 2025. AI-Driven Behavioral Risk Profiling in Digital Lending Platforms: A Cross-Disciplinary Framework for Dynamic Risk Assessment. *Journal of Computer Science and Technology Studies,* 7(6), pp. 746--751.

Levy, A. & Baha, R., 2021. Credit risk assessment: a comparison of the performances of the linear discriminant analysis and the logistic regression. *International Journal of Entrepreneurship and Small Business,* 42(1-2), pp. 169--186.

Malik, P. et al., 2024. Credit risk assessment and fraud detection in financial transactions using machine learning. *Journal of Electrical Systems,* 20(3s), pp. 2061--2069.

Mir, M. N. H. et al., 2025. Joint topic-emotion modeling in financial texts: A novel approach to investor sentiment and market trends. *IEEE Access,* Volume 13, pp. 28664-28677.

Patel, S. B., Bhattacharya, P., Tanwar, S. & Kumar, N., 2020. Kirti: A blockchain-based credit recommender system for financial institutions. *IEEE Transactions on Network Science and Engineering,* 8(2), pp. 1044 -1054.

Peivandizadeh, A. et al., 2024. Stock Market Prediction With Transductive Long Short-Term Memory and Social Media Sentiment Analysis. *IEEE Access,* Volume 12, pp. 87110-87130.

Sakri, S., 2022. Assessment of Deep Neural Network and Gradient Boosting Machines for Credit Risk Prediction Accuracy. In: IEEE, ed. *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN).* s.l.:IEEE, pp. 1-7.

Schmitt, M., 2022. Deep Learning vs. Gradient Boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring. *arXiv preprint arXiv:2205.10535.*

Teles, G., Rodrigues, J. J., Rab{\^e}lo, R. A. & Kozlov, S. A., 2021. Comparative study of support vector machines and random forests machine learning algorithms on credit operation. *Software: practice and experience,* 51(12), pp. 2492--2500.

Uddin, M. S., Chi, G., Al Janabi, M. A. & Habib, T., 2022. Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. *International Journal of Finance \& Economics,* 27(3), pp. 3713--3729.