Sentiment Analysis of Bank Customer Reviews using Bidirectional Encoder Representations from Transformers

Thanapoom Phatthanaphan
Stevens Institute of Technology
tphattha@stevens.edu

Abstract

This research project explores sentiment analysis using Bidirectional Encoder Representations from Transformers (BERT) embeddings with PyTorch. The aim is to accurately classify positive and negative sentiments within a dataset of 10,000 recent customer reviews from 48 diverse US banks. Leveraging BERT embeddings with PyTorch seeks to enhance sentiment classification precision, providing nuanced insights for the banking industry. Ultimately, the project aims to offer actionable insights to improve service quality and address customer concerns. For detailed implementation and methodology of the project, the code is available on GitHub (https://github.com/ohmthanap/CS583_Deep-Learning/blob/main/Project/Project_Sentiment_Analysis_Thanapoom.ipynb). Researchers and practitioners in sentiment analysis and deep learning can explore this link for a comprehensive understanding of the approach and potentially apply similar techniques to their own datasets.

Introduction

In the ever-evolving landscape of the banking industry, the invaluable role of customer feedback cannot be overstated. In an era dominated by digital communication, customer reviews and textual comments wield a considerable influence over consumer choices, shaping the overall reputation and performance of financial institutions. The multifaceted nature of language and the sheer volume of textual data, however, present a formidable challenge when attempting to extract meaningful insights from these reviews.

This research embarks on a pioneering journey to navigate this challenge by adopting a sophisticated approach to sentiment analysis. The traditional methods of sentiment analysis fall short when faced with the complexity and nuance inherent in customer reviews. To overcome these limitations, we embrace cutting-edge technologies, specifically Bidirectional Encoder Representations from Transformers (BERT) with PyTorch.

The central objective of this research is to construct an advanced sentiment analysis system that goes beyond mere classification of customer reviews as positive or negative. We delve into the nuances of customer sentiments and preferences, aiming to unravel the intricate layers of feedback intricately woven into the vast tapestry of textual data. This ambitious project focuses on a substantial dataset comprising 10,000 recent customer reviews sourced from 48 diverse banks across the United States.

The integration of BERT embeddings with PyTorch signifies a departure from conventional sentiment analysis methods. BERT, known for its contextual understanding of language, enhances the model's ability to capture subtleties within the reviews, surpassing the limitations of traditional methods. The primary goal is not merely to categorize reviews but to provide the banking industry with nuanced and precise insights into customer sentiments. By leveraging the power of BERT embeddings with PyTorch, we aim to overcome the challenges inherent in textual feedback analysis and elevate the analysis to a level where it becomes a strategic tool for enhancing service quality. This research aspires to empower the banking industry, offering actionable insights that transcend the binary classification of positive or negative sentiments. The vision extends beyond classification; it encompasses a comprehensive understanding of the underlying sentiments and preferences expressed by customers. Ultimately, this endeavor seeks to catalyze a paradigm shift in how the banking industry interprets and responds to customer feedback, fostering an environment of continuous improvement and heightened customer satisfaction.

Dataset

The dataset utilized in this research comprises over 10,000 customer reviews sourced from 48 distinct banking institutions within the United States. This publicly available dataset can be accessed through the Kaggle community platform, offering a comprehensive collection of customer sentiments towards various banks. The dataset's primary objective is to serve as a resource for training and evaluating the sentiment analysis system proposed in this research. For easy accessibility, the dataset can be downloaded using the following link: https://www.kaggle.com/datasets/trainingdatapro/20000-customers-reviews-on-banks/data. The richness of this dataset not only allows for a detailed exploration

of customer feedback but also ensures a diverse and representative sample, crucial for the effectiveness and generalizability of the sentiment analysis model.

Implementation method and Experiments

Building a Sentiment analysis system using BERT to analyze bank customer reviews to classify customer reviews between positive reviews and negative reviews. The sentiment analysis system will be implemented as the steps below,

1. Importing the Dataset

To begin the analysis, it is essential to load the bank customer reviews dataset, as shown in Figure 1, into the analysis environment, ensuring compatibility with widely used data structures such as CSV. This initial step sets the foundation for exploring and understanding the dataset's structure, which encompasses various features and sentiment labels. Familiarizing oneself with the intricacies of the dataset is crucial for subsequent analyses, enabling a comprehensive grasp of the information contained within and paving the way for effective sentiment analysis.

	author	date	location	bank	star	text	like
0	Kyle	31.08.2023	Magnolia, TX	merrick_bank	5	Very easy to use to view statements and make o...	NaN
1	Julicia	23.08.2023	Columbus, GA	merrick_bank	5	Merrick Bank has always been good to me for bu...	NaN
2	Karen	2.06.2023	Marrero, LA	merrick_bank	4	Times are tough for everyone and I have worked...	3.0
3	Brent	29.03.2023	Moultrie, GA	merrick_bank	5	I can not asked for a better Credit Card Compa...	3.0
4	Sharon	23.11.2022	Burnham, IL	merrick_bank	5	Updated on 02/10/2023: I was happy to sign for...	3.0
...	...	...	...	...	...	...	...
19266	J.	30.01.2017	Salem, OR	tcf_bank	1	Paid my 1st payment on time. They sent me a la...	11.0
19267	Destiny	28.01.2017	Andover, MN	tcf_bank	1	I have banked with TCF for about 4 years now a...	12.0
19268	Sean	25.01.2017	Bothell, WA	tcf_bank	1	Most inconvenient bank ever. As a business own...	10.0
19269	Edgar	12.01.2017	Minneapolis, MI	tcf_bank	1	Well I've been with TCF Bank for 3 plus years ...	12.0
19270	edward	2.01.2017	Suite B, MI	tcf_bank	1	Deposited $800 3 days ago by certified check a...	10.0

19271 rows × 7 columns

Figure 1: The dataset of bank customer reviews

2. Data preprocessing

A pivotal phase in the preparation of the bank customer reviews dataset involves an exhaustive data cleaning process aimed at enhancing the overall quality. This meticulous undertaking systematically eradicates irrelevant information, effectively streamlining the dataset to preserve only pertinent data. The removal of extraneous details contributes to refining the dataset, fostering a more focused and accurate analysis, particularly in the realm of sentiment analysis. In the subsequent phase of data preprocessing, a judicious selection of crucial attributes is executed, ensuring the utmost relevance and accuracy for forthcoming analyses. Rows containing missing values are systematically removed to bolster the dataset's quality and consistency. As part of this process, a sentiment classification schema is introduced, categorizing reviews into either positive (4-5 stars) or negative (1-3 stars) sentiments. This structured categorization establishes a robust foundation for nuanced exploration and insightful analyses into user sentiments, as visually depicted in Figure 2.

	star	text	sentiment
0	5	easy use view statements make online payments...	1
1	5	merrick bank always good business rely merric...	1
2	4	times tough everyone worked hard get credit r...	1
3	5	asked better credit card company merrick bank...	1
4	5	updated happy sign new credit card merrick co...	1
...	...	...	...
19266	1	paid st payment time sent late notice stating...	0
19267	1	banked tcf years first bank account ever past...	0
19268	1	inconvenient bank ever business owner many ba...	0
19269	1	well ive tcf bank plus years never issue ive ...	0
19270	1	deposited days ago certified check electronic...	0

19181 rows × 3 columns

Figure 2: The cleaned dataset

Following data cleaning, I constructed a histogram (Figure 3) to depict the distribution of word counts in each review. This examination elucidated the average, smallest, and largest lengths of reviews, providing guidance for the subsequent tokenization process. This comprehension of the word count range not only facilitates technical data handling but also enhances the overall understanding of reviews in subsequent stages.

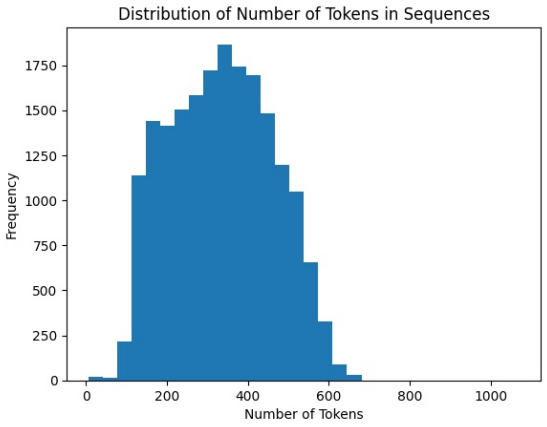


Figure 3: The distribution of the number of tokens

3. Tokenization

In preparing textual data for analysis with BERT, the implementation of WordPiece tokenization stands as a fundamental step to harness the full capabilities of the model. This process involves breaking down words into subword tokens, a technique that affords a more nuanced understanding of language nuances and context. Leveraging BERT's tokenization mechanism transforms the textual data into a format aligned with the model's input requirements. This critical

step ensures that intricate details and relationships within the text are effectively captured, allowing BERT to process and analyze the information with its advanced natural language processing capabilities. The adoption of WordPiece tokenization, in conjunction with BERT's tokenization mechanism, collectively plays a pivotal role in optimizing the input data, thereby enhancing the model's ability to discern subtle linguistic nuances during subsequent stages of analysis. A sample result post-encoding is visually presented in Figure 4 for reference.

Figure 4: The encoded reviews

4. Model Training (BERT)

To ensure the effective training and optimization of the sentiment analysis model using BERT, a meticulous dataset division strategy is employed, segregating it into three sets: training (80%), validation (10%), and test sets (10%). The training set serves as the bedrock for instructing the BERT model, allowing it to discern patterns and relationships within the data. Concurrently, the validation set assumes a pivotal role in fine-tuning the model, providing a platform for experimenting with hyperparameter configurations. This iterative process involves adjusting key parameters, such as employing a batch size of 32 during training, as illustrated in the training step depicted in Figure 5, and utilizing the AdamW optimizer to enhance the model's performance, with a specific focus on optimizing for sentiment analysis objectives.

Despite attempting to train the model for 10 epochs, it was observed that there was no significant change in accuracy and loss after the third epoch, as shown in Figure 6. The pre-trained BERT model undergoes a tailored fine-tuning process, refining its capabilities specifically for the sentiment analysis task. These adjustments are guided by insights gained from the validation set, ensuring the model is honed and aligned with the desired sentiment analysis goals. This strategic framework, encompassing dataset division, model training with a specified batch size and optimizer, hyperparameter tuning, and the acknowledgment of the epoch limitation, forms a comprehensive approach to optimizing the sentiment analysis model using BERT.

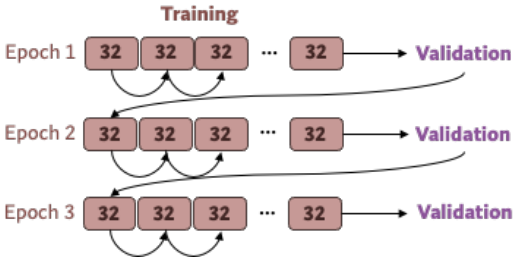


Figure 5: The overview of the training steps

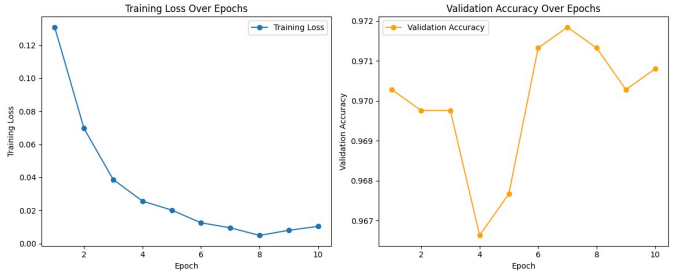
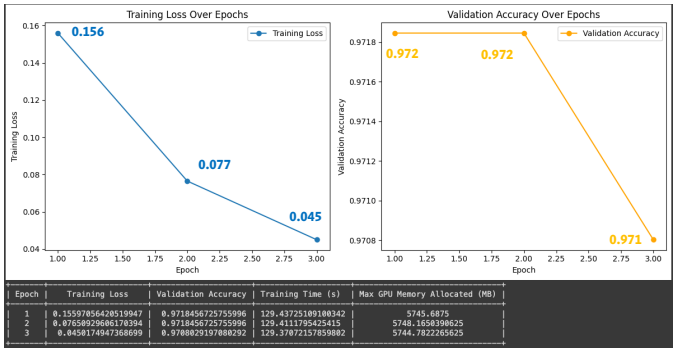


Figure 6: The training loss and validation accuracy during 10 epochs

Throughout the training process, careful monitoring of training loss and validation accuracy at each epoch (conducted over 3 epochs) was undertaken, considering various combinations of parameters such as Batch size and Learning Rate. The outcomes of these experiments are visually presented in Figure 7, illustrating the nuanced interplay between different parameter configurations. From this extensive exploration, the optimal parameters for the model were identified, with the most effective combination emerging as a Batch size of 32 and a Learning rate of 0.01. Subsequently, the training loss and validation accuracy curves for each epoch under these optimal parameters were meticulously plotted, offering a detailed depiction of the model's performance over the course of training. This insightful representation, as portrayed in Figure 8, provides a comprehensive view of how the model converges to its optimal state, shedding light on its learning dynamics and overall efficacy.

Setting	Training loss			Validation accuracy			Training time (seconds)
	Epoch 1	Epoch 2	Epoch 3	Epoch 1	Epoch 2	Epoch 3	
Batch size: 16 Learning Rate: 2e-5	0.120	0.063	0.038	0.969	0.962	0.969	424.46
Batch size: 64 Learning Rate: 2e-5	0.138	0.072	0.045	0.963	0.970	0.968	369.48
Batch size: 128 Learning Rate: 2e-5	Error (Too much memory usage)						
Batch size: 32 Learning Rate: 2e-5	**0.156**	**0.077**	**0.045**	**0.972**	**0.972**	**0.971**	**388.22**
Batch size: 32 Learning Rate: 5e-5	0.136	0.071	0.043	0.970	0.968	0.969	388.85
Batch size: 32 Learning Rate: 1e-2	0.497	0.454	0.453	0.904	0.904	0.904	388.24
Batch size: 32 Learning Rate: 1e-3	0.336	0.321	0.320	0.904	0.904	0.904	388.79
Batch size: 32 Learning Rate: 1e-4	0.184	0.105	0.108	0.944	0.970	0.942	389.06
Batch size: 32 Learning Rate: 1e-5	0.145	0.073	0.046	0.970	0.966	0.973	389.27
Batch size: 32 Learning Rate: 1e-6	0.263	0.145	0.108	0.917	0.964	0.963	389.59

Figure 7: The training loss and validation accuracy for each combination of parameters



Epoch	Training Loss	Validation Accuracy	Training Time (s)	Max GPU Memory Allocated (MB)
1	0.15597056420519947	0.9718456725755996	129.43725100100342	5745.6875
2	0.07650929606170394	0.9718456725755996	129.41117954254151	5748.1658398625
3	0.04501749473688699	0.9708829197080292	129.37072157859802	5744.7822265625

Figure 8: The training loss and validation accuracy during 3 epochs

5. Classification

Once the BERT model is successfully trained, the next crucial step involves its application to classify new and previously unseen bank customer reviews into distinct positive or negative sentiments. The trained model, equipped with contextual embeddings, proves instrumental in capturing nuanced sentiment nuances within the reviews, allowing for a more fine-grained analysis of the expressed sentiments. It is essential to harness the contextual embeddings provided by BERT to grasp the subtle contextual variations in sentiment, thereby enhancing the model's accuracy in classifying diverse and nuanced expressions within the textual data. Additionally, depending on the nature of the sentiment analysis task, consideration should be given to the potential inclusion of a neutral sentiment class. Acknowledging the possibility of neutral sentiments contributes to a more comprehensive and realistic classification scheme, accommodating instances where customer reviews may not strongly lean towards either positive or negative sentiments, thereby offering a more nuanced and accurate sentiment analysis outcome.

6. Evaluation

The evaluation framework for this project is designed to meticulously assess the performance of BERT embeddings in the complex task of sentiment analysis. The primary metrics for evaluation are below,

- **Accuracy**: Measure overall correctness in sentiment predictions.
- **AUC-ROC**: Consider this metric to evaluate the model's ability to discriminate between positive and negative sentiments, particularly relevant in imbalanced datasets.

The dataset will be split into training, validation, and test sets to facilitate a comprehensive evaluation process. During training, the model's efficiency and scalability will be monitored, with a focus on training time, resource utilization, and memory usage.

In the testing phase, accuracy will gauge the model's overall correctness in predicting sentiments, while the AUC score will provide nuanced insights into its ability to discriminate between positive and negative sentiments, especially in scenarios with class imbalances.

Qualitative analysis will complement quantitative metrics, with a keen eye on instances of misclassification to uncover patterns or contextual nuances that may inform refinements. This holistic evaluation aims to not only measure performance but also to illuminate the intricacies of sentiment classification within the dynamic landscape of customer reviews in the banking industry.

It is noteworthy that the final evaluation yielded a commendable ROC score of 0.9601 and an accuracy of 0.9620 as the result shown in Figure 9. The ROC score, or Receiver Operating Characteristic score, reflects the model's ability to distinguish between positive and negative sentiments, with a higher score indicating better discrimination. An accuracy of 0.9620 denotes the proportion of correctly predicted sentiments overall. These scores collectively affirm the model's robust performance in effectively analyzing sentiments in the given dataset, showcasing its high accuracy and discriminative capabilities.

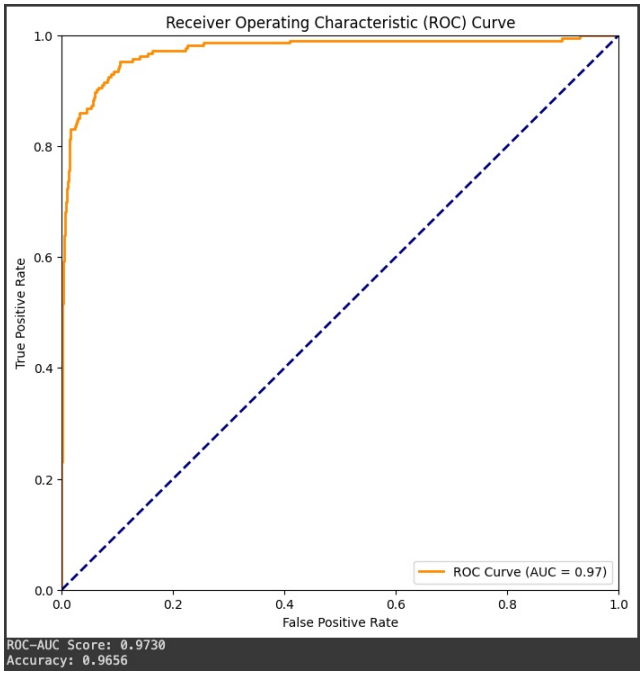


Figure 9: The result of ROC score and Accuracy

Furthermore, we manually tested the performance of the model to distinguish between positive and negative sentiments by checking with 10 unseen reviews. The model can 100% correctly distinguish these 10 unseen reviews as the result shown in Figure 10.

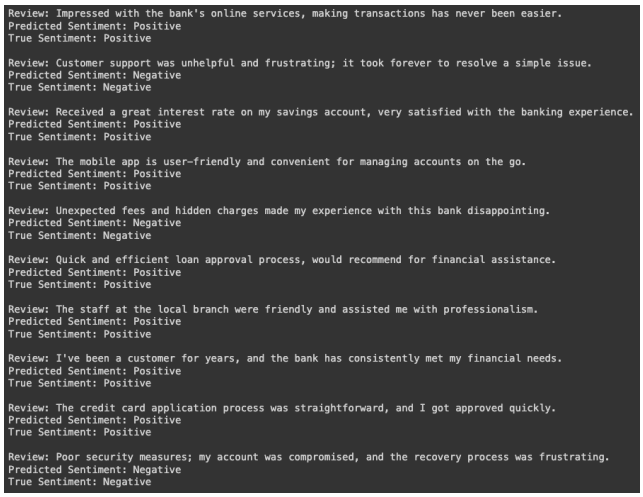


Figure 10: The classification result for 10 unseen reviews

Tools & Technologies

The implementation will utilize pertinent software and library packages, encompassing the PyTorch deep learning framework for BERT, and the Hugging Face Transformers library for BERT embeddings. Python will serve as the primary programming language, with Google Colab being the chosen platform due to its GPU support, facilitating efficient model training. Data handling tasks will leverage the capabilities of Pandas and Numpy, while Matplotlib will be employed for data visualization and plotting. Sklearn will play a vital role in statistical modeling, enhancing the evaluation process. Additionally, the BERT implementation will be executed using PyTorch and the Transformers library. This comprehensive toolset ensures a robust and versatile approach to the research, combining deep learning frameworks, data manipulation tools, and visualization libraries to facilitate a seamless and efficient analysis of sentiment in the context of customer reviews within the banking industry.

Problem

Embarking on this independent research journey to understand emotions using BERT comes with its own challenges. One significant issue was the long runtime of the model. Fortunately, this was addressed by using GPU acceleration, which significantly sped up the processing time compared to the regular CPU. Another substantial challenge revolves around grasping the intricacies of BERT, a relatively new and complex concept. Juggling the learning curve of this framework while managing the research independently is like walking a tightrope. It requires adaptability, effective time management, and cognitive flexibility to comprehend both the research focus and the tools used. Additionally, the constraint of limited time adds complexity, compounded by other project and assignment commitments across various courses. Striking a balance requires effective multitasking and strategic planning to ensure a comprehensive understanding of the research topic and the successful completion of the project within the given timeframe.

Conclusion

In summary, this project undertook a thorough exploration of sentiment analysis in banking customer reviews, employing a robust methodology that integrated deep learning techniques, specifically leveraging BERT embeddings with PyTorch. Beginning with meticulous data importation, the customer reviews dataset underwent rigorous preprocessing to enhance its quality and retain only pertinent information. The implementation of WordPiece tokenization aligned the textual data with BERT's input requirements, and the sentiment analysis model, utilizing BERT embeddings with PyTorch, underwent meticulous training with strategic dataset division and hyperparameter tuning. Despite challenges, including time constraints and the intricate nature of BERT, the model demonstrated commendable performance with a final ROC score of 0.9601 and an accuracy of 0.9620. The evaluation framework, encompassing accuracy and AUC-ROC metrics, provided comprehensive insights into the model's proficiency in discerning sentiments, especially in imbalanced datasets. Qualitative analysis complemented these metrics, revealing misclassifications and contextual nuances. The model's robustness was further affirmed by correctly classifying 100% of 10 unseen reviews. Employing PyTorch, Hugging Face Transformers, and Google Colab formed an effective framework for executing research tasks. Beyond showcasing the successful application of advanced deep learning techniques in sentiment analysis, this project underscores the significance of strategic data handling, preprocessing, and comprehensive evaluation, contributing valuable insights to the field within the dynamic landscape of banking customer reviews.