Sentiment Analysis of Bank Customer Reviews using

Bidirectional Encoder Representations from Transformers

Thanapoom Phatthanaphan

Stevens Institute of Technology  
tphattha@stevens.edu

Abstract

This research delves into the realm of sentiment analysis, focusing on the application of Convolutional Neural Networks (CNN) with Bidirectional Encoder Representations from Transformers (BERT) embeddings to decipher the sentiments embedded in a substantial dataset comprising 10,000 recent customer reviews from 48 diverse US banks. The primary objective is to develop an advanced sentiment analysis system capable of accurately classifying customer feedback as positive or negative. The utilization of CNN with BERT embeddings aims to enhance the precision of sentiment classification, offering the banking industry actionable insights to elevate service quality and effectively address customer concerns.

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 Convolutional Neural Networks (CNN) augmented with BERT embeddings.

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 Convolutional Neural Networks (CNN) with BERT embeddings signifies a departure from conventional sentiment analysis methods. CNNs, known for their prowess in image recognition, are adapted to capture intricate patterns and relationships within the textual data. BERT embeddings, a state-of-the-art language representation model, further enhance the contextual understanding of the reviews, capturing subtleties that elude 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 CNN with BERT embeddings, we aim not only to overcome the challenges inherent in textual feedback analysis but also to 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.

A screenshot of a black and white screen

Description automatically generated

Figure 1: The dataset of bank customer reviews

2. Data preprocessing

A critical phase in preparing the bank customer reviews dataset for analysis involves data cleaning to enhance the overall quality. This process revolves around the systematic removal of irrelevant information, aiming to streamline the dataset and ensure that only pertinent data is retained. By eliminating extraneous details, the dataset becomes more refined, facilitating a more focused and accurate analysis. Data cleaning serves as a crucial step in optimizing the dataset for subsequent tasks, contributing to the reliability and effectiveness of the forthcoming analyses, particularly in the context of sentiment analysis.

In the data preprocessing phase, a meticulous process was undertaken to discern and extract the dataset's crucial attributes, a crucial step in ensuring the relevance and accuracy of subsequent analyses. This involved a careful selection of features that would play a pivotal role in the analysis, contributing significantly to the overall understanding of the dataset. Moreover, to enhance the dataset's quality and consistency, rows containing missing values were systematically removed. This strategic curation of the data aimed to foster a cleaner and more reliable foundation for subsequent tasks. Additionally, a sentiment classification schema was implemented to categorize each review distinctly as either positive or negative. Reviews adorned with 4-5 stars were unequivocally classified as positive (1), while those garnering 1-3 stars were unequivocally labeled as negative reviews. This categorization, illustrated in Figure 2, not only facilitates a more nuanced exploration of the dataset but also lays the groundwork for insightful analyses into the sentiments expressed by users.

A screenshot of a computer screen

Description automatically generated

Figure 2: The cleaned dataset

After cleaning up the data, I checked how many words were in each review. I made a histogram (like a bar chart) called Figure 3 to show how the word counts are spread out. This helped me figure out the average (middle), smallest, and largest lengths of the reviews. Knowing this is useful because it guides how I break down the text into smaller parts during the tokenization process. Understanding the range of word counts not only helps with the technical side of handling data but also makes it easier to make sense of the reviews later on.

A graph of a number of tokens

Description automatically generatedFigure 3: The distribution of the number of tokens

3. Tokenization

To prepare the textual data for analysis using BERT, it is imperative to implement WordPiece tokenization, a fundamental step in leveraging BERT's capabilities. This process involves breaking down words into subword tokens, a technique that allows for a more nuanced understanding of language nuances and context. By adopting BERT's tokenization mechanism, the textual data is transformed into a format that aligns with the model's input requirements. This crucial step ensures that the intricate details and relationships within the text are effectively captured, enabling BERT to process and analyze the information with its advanced natural language processing capabilities. The utilization of WordPiece tokenization and BERT's tokenization mechanism collectively plays a pivotal role in optimizing the input data, enhancing the model's ability to discern subtle linguistic nuances during the subsequent stages of analysis. The sample result after encoding shown in Figure 4.

A screenshot of a computer program

Description automatically generated

Figure 4: The encoded reviews

4. Model Training (BERT)

To ensure the effective training and optimization of the sentiment analysis model using BERT, the dataset is meticulously divided into three sets: training (80%), validation (10%), and test sets (10%). The training set serves as the foundation for instructing the BERT model, enabling it to discern patterns and relationships within the data. Simultaneously, the validation set plays 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 using a batch size of 32 for training, as illustrated in the training step depicted in Figure 5, and employing 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 the result shown in Figure 6. The pre-trained BERT model undergoes a tailored fine-tuning process, honing its capabilities specifically for the sentiment analysis task. These adjustments are guided by insights gained from the validation set, ensuring the model is refined 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.

A screenshot of a training program

Description automatically generated

Figure 5: The overview of the training steps

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure 6: The training loss and validation accuracy

during 10 epochs

During the training process, the training loss and validation accuracy at each epoch (3 epochs) for several combination of parameters of Batch size and Learning Rate as the results shown in Figure 7. The best parameters that we discovered for this model are Batch size at 32 and Learning rate at 0.01. The training loss and validation accuracy of the best parameters for each epoch were plotted as shown in Figure 8.



Figure 7: The training loss and validation accuracy

for each combination of parameters



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 the Convolutional Neural Networks (CNN) with 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.

A graph of a positive curve

Description automatically generated with medium confidence

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.

A screenshot of a computer

Description automatically generated

Figure 10: The classification result for 10 unseen reviews

Tools & Technologies

The implementation will be conducted utilizing pertinent software and library packages, encompassing the TensorFlow or PyTorch deep learning frameworks for CNN, 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 Convolutional Neural Networks (CNN), and the transformative capabilities of BERT embeddings. 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, 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 TensorFlow, 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.