A Business-Focused Deep Learning Project (Assigenment 4)

MGT5301 – Predictive Analytics

Professors:

Dr. Jonathan li and Dr. Rafid Mahmood

Prepared by: Morteza Emadi

Student Num. : 300387478

University of Ottawa

Telfer School of Management

Dec 2023

The length of the report is short because of the structures defined by the professor. This proposal covers various sentiment analysis and text generation approaches, with part of the solution implemented, as explained in the Appendix.

Business Problem Investigation

The project targets the dynamic domain of social media interactions, specifically on Twitter, where understanding user sentiments is crucial for effective customer engagement. In the context of customer service, promptly identifying and appropriately responding to user sentiments bipolar – positive or negative – or multi-aspect emotions in their messages can significantly enhance customer experience and brand perception. This is particularly relevant in industries like airlines, where customer feedback on social media can dramatically influence public opinion and business success. This AI module, serving as a CRM tool, is versatile enough to analyze various platforms including tweets, Reddit posts, and review sites, tailoring its functionality to the specific regional needs of the business.

Project Challenges

The system operates in two stages: initially, it conducts sentiment analysis to discern customer emotions from tweets, using these insights to direct responses in the text generation module. The main challenge is accurately interpreting complex sentiments in tweets, which may include nuanced language, slang, and emoticons. Effectiveness hinges on the model's ability to comprehend and generate language matching the emotional tone. Additionally, Topic Modeling can be integrated between the steps to identify key customer concerns or satisfaction areas.

Dataset Requirement and Limitation

For our analysis on platforms like Twitter and Reddit, we need a labeled dataset for each. Initially, the text and bi-polar sentiment of tweets should suffice. Our primary dataset, Sentiment140[1], offers a wealth of sentiment-tagged Twitter data. To adapt the model for specific sectors, like airlines, I plan to use Tweepy[1], an open-source Python library, to scrape tweets through the Twitter API (a Tweepy code snippet is included in the appendix). The main challenges involve

[1] https://www.tweepy.org/

labeling this data, especially for multiple emotions. In the method section, I'll address these challenges by implementing/suggesting novel ideas in "Method and Justification" section.

Method and Justification

Sentiment Analysis:

For the sentiment analysis component, my project leverages the advanced capabilities of Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) applied to DistilBERT. DistilBERT is a streamlined version of the larger BERT model, requiring fewer computational resources, which makes it well-suited for processing large volumes of Twitter data and fits within the resource constraints of my course project.

PEFT is a technique that enables the fine-tuning of large-scale pre-trained models like DistilBERT in a resource-efficient manner. It involves selectively updating a subset of the model's parameters, rather than retraining the entire network. This approach is particularly beneficial in my project, where there are variations in the nuances of specific tweets across different industries. PEFT's capability to make precise, nuanced adjustments is crucial for accurately capturing sentiments without the need for extensive retraining on large datasets. LoRA (as a PEFT approach) targets the Transformer layers, where it strategically inserts trainable low-rank matrices. These matrices expand the model's representational capacity, allowing it to adapt to new tasks with minimal additional parameters. As depicted in my code, even though only 15,000 tweets out of the 1.6 million available in the dataset were loaded due my time constraints, the precision of DistilBERT increased from 0.5 to 0.76 in the PEFT model. Aside from the appendix, the model is also accessible in a repository on my Hugging Face profile.

Text Generation:

For the text generation aspect, considering DistilBERT's limitations for Seq2Seq tasks and time constraints in leveraging another LLM, I utilized an RNN-GRU architecture combined with an N-gram model. This combination is ideal for Twitter's short, context-rich texts, with GRUs effectively capturing sequential word patterns and N-grams aiding in word order understanding. This approach strikes a balance between complexity and performance.

Further Enhancement Scenarios

1. *Semi-Supervised Learning*: In situations where specifically labeled tweets are lacking, especially in unique industry contexts like Miracle Flights in air travel, where conventional Lexicon methods are inadequate, adapting the learning model with limited labeled data becomes crucial. Zou and Wang's study[2] advocates for a semi-supervised approach. This technique blends a portion of labeled data with unlabeled, industry-specific tweets. It utilizes data enhancement methods, notably back translation, and employs MixMatchNL for label prediction. This combination of labeled and unlabeled data optimizes sentiment analysis efficiency.

2. *Generative Models for Data Augmentation*:A diffusion model in NLP consists of a forward process that adds noise to data, and a reverse process that reconstructs original data from this noise, thereby generating new content. It can capture complex data patterns. In our project, this model can be effectively used for topic modeling of tweets, adeptly categorizing them into distinct topics, like Urgent Complaints and Normal Appreciation.

3. *Innovative Transformer Learners*: If we have enough labeled data of some emotions, but for other ones, we have less labeled data -unbalanced data- we can employ innovative transformer learning mechanisms as demonstrated by Jie & Xing.[3] The approach described in the paper involves a multi-layered process using BERT and XLNet models for multi-label classification. First, raw input text is transformed into embeddings. These embeddings are fed into a 24-layered transfer learning model, where each layer computes multiple self-attentions to encode the embeddings. The final encoded output is then used to map onto the classes. The method employs the logistic sigmoid function for output activation and a modified loss function for multi-label classification. This pipeline effectively handles the complexity of multi-label sentiment analysis, even with limited data.

Implementation and Evaluation

The project will be implemented in two phases. The first phase involves training the sentiment analysis model using the Sentiment140 dataset, which includes data preprocessing, fine-tuning DistilBERT with LoRA, and model validation through accuracy, precision, recall, and F1 score. The second phase focuses on developing the text generation model using RNN-GRU and N-gram methodologies. This phase encompasses structuring textual data into trigrams for training,

followed by the development of a Recurrent Neural Network (RNN) model with Gated Recurrent Unit (GRU) layers. The RNN-GRU model, once trained, interprets and utilizes complex text sequences to produce contextually relevant and sentiment-aligned text. Evaluation of this phase includes assessing linguistic quality using the BLEU score and conducting qualitative analysis for contextual relevance and emotional appropriateness.

References:

[1] A. Go, R. Bhayani, and L. Huang, "Sentiment140 dataset with 1.6 million tweets," *Kaggle*, 2009. https://www.kaggle.com/datasets/kazanova/sentiment140.

[2] H. Zou and Z. Wang, "A semi-supervised short text sentiment classification method based on improved Bert model from unlabelled data," *J. Big Data*, vol. 10, no. 1, p. 35, Mar. 2023, doi: 10.1186/s40537-023-00710-x.

[3] J. Tao and X. Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," *J. Big Data*, vol. 7, no. 1, p. 1, Dec. 2020, doi: 10.1186/s40537-019-0278-0.

Appendix:

This section demonstrates the codes I used for the first and second parts of the project, also, including a "Tweepy" code snippet for tweet scraping in the future development as the last part.

Also, the trained DistilBERT model with PEFT tuning is accessible from a repository on my Hugging Face profile.

Our dataset, the Sentiment140 dataset, contains 1.6 million tweets annotated for sentiment. It includes fields like tweet polarity, ID, date, query flag, user, and text. I used only 15,000 tweets for training the DistilBERT model due to my time constraints, which increased the precision from 0.5 to 0.76 within only eight epochs of training in the PEFT model. This suggests the potential for credible results of this short implementation for further research.