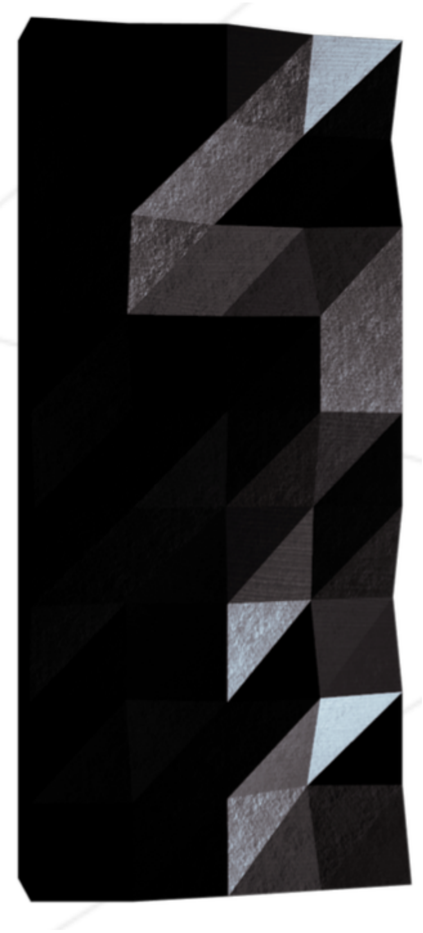
Transfer Learning in Sentiment Analysis

42 Urduliz AI Project

September 2024



Foreword

Having seen the last film of Nolan, I found myself plunged into a realm where narrative coherence is not merely absent but deliberately forsaken, where meaning itself appears to be a mirage, receding further as one strives to grasp it. The director, in what seems like an elaborate exercise in the simulacrum of intelligibility, orchestrates a sequence of events that, upon closer scrutiny, collapse into a void of semiotic indifference. Characters float in a liminal space of signifiers.

It is precisely here, in this liminal intersection of the real and the symbolic, where the film's attempt at creating a cohesive narrative disintegrates into a pure play of signifiers, a kind of Lacanian jouissance of interpretative paralysis. One might even argue that the very act of seeking meaning becomes the central theme of the film, a recursive loop of endless deferral, much akin to the structure of a Möbius strip, where the viewer, much like Lacan's subject, is suspended between desire and lack of desire.

In this manner, Nolan constructs a symbolic edifice so dense that the only viable interpretation is to relinquish any hope of interpretation altogether, surrendering to the pure opacity of the cinematic experience. And yet, in this surrender, a certain clarity emerges—not of the film's intended message, for such a message is irretrievably lost in the maze of signifiers—but of the viewer's own compulsion to decode a puzzle that was never meant to be solved.

I think I have made my point clear.

1 Introduction

Sentiment analysis is a powerful natural language processing technique used to determine whether a piece of text conveys a positive, negative, or neutral sentiment. It plays a vital role in understanding people's opinions and emotions from various forms of textual data.

It can be applied to film or book reviews. A review for a popular movie like Inception might read, "Nolan has truly outdone himself with this masterpiece, a mind-bending ride from start to finish," clearly expressing a positive outlook. On the other hand, a negative review for the same movie could state, "Inception was far too confusing, I couldn't make sense of it."

Another common use case is restaurant reviews. For example, a diner might leave feedback like "The food was delicious, and the service was excellent!" (positive) or "The food was cold, and the service was terrible" (negative).

This project is an introduction to **Transfer Learning**, a powerful machine learning approach that leverages pretrained models to solve new problems. In this case, you will use transfer learning to develop a binary sentiment analysis system. You will research various pretrained models, tokenization techniques, and hyperparameter configurations to find the best balance between performance and computational efficiency.

However, due to constrained resources, it may be necessary to choose smaller models and reduce the dataset size to ensure that the training process is feasible within your computing limitations. By doing so, you will gain hands-on experience with one of the most impactful advances in modern NLP.

2 Objectives

The primary objective of this project is to give you practical experience in transfer learning and natural language processing. You will:

- Utilize free-tier cloud computing or local resources (such as sgoinfre) to train machine learning models.
- Research and select pretrained models (e.g., BERT, RoBERTa, GPT-2, DistilBERT, Electra, XL-Net) and tokenization methods.
- Define and tune hyperparameters, including the number of epochs, learning rate, and batch size, to optimize model performance.
- Compare tokenization techniques and how they relate to the pretrained model.
- Evaluate the model's accuracy and generalization capability.

3 General Instructions

Note

We will see some basic steps of sentiment analysis. Of course, these are not the only techniques available or the only steps to follow. Each dataset and problem has to be approached in a unique way. You will surely find other ways to analyze your data in the future.

You are free to use any programming language, libraries, models, tokenization methods, and computing resources (cloud or local) of your choice. However, the following rules apply:

- You must decide the key hyperparameters: number of epochs, learning rate, and batch size.
- You are expected to perform your own research to select appropriate models, tokenizers, datasets, and platforms.
- While high-level libraries like HuggingFace's Transformers can be used, it is important to understand the underlying concepts and decisions.
- Projects will be evaluated on clarity, structure, and your understanding of the concepts involved.

4 Mandatory Part

Warning

It is highly recommended to perform the following steps in the prescribed order, especially when dealing with constrained resources.

Your project must include the following tasks:

- **Model and Platform Research and Selection:** Research various pretrained models (e.g., BERT, RoBERTa) and compare cloud platforms for training the model. Be mindful of the model size; given potential resource constraints, it may be necessary to choose smaller models that are computationally efficient.
- **Tokenization Research and Selection:** Explore different tokenization techniques and select an appropriate one based on the dataset and model.
- **Dataset Research and Selection:** You will be provided with a sample dataset called *sample_reviews.parquet* to give you a first approach to sentiment analysis. However, you must choose your own dataset for the final model (e.g., IMDB, Yelp, Amazon reviews) and split it into training and testing subsets. Be aware that due to potential resource constraints, you may need to reduce the dataset size.
- **Model Training:** Train your selected model on your chosen dataset, tuning hyperparameters such as learning rate, batch size, and number of epochs.
- **Model Evaluation:** Evaluate your model's performance on the test set and aim for an accuracy of at least 0.75, though this is a recommended benchmark rather than a strict requirement.

5 Bonus Part

If you have successfully completed the mandatory tasks, you may attempt the following bonus challenges:

- Test your model on additional datasets to assess its ability to generalize across different types of sentiment data. You must choose the other datasets and evaluate your model with these new datasets.
- Explore multiple tokenization techniques and compare their outputs, providing insights on their differences.
- Deploy your trained model on the web, allowing real-time sentiment prediction via a user interface.

6 Submission and Peer-Evaluation

This is a proof of concept, so there will be no Git repository submission. You must have a specific folder containing your code. Do not include the trained model or dataset in the folder.

- If you have used cloud computing resources to train your model, you must show your implementation online during the evaluation.
- If you have trained your model locally, you must run it locally during the evaluation.
- The training process should not take more than 15 minutes, allowing time for discussion and explanation of your implementation.

Ensure that your submission is clear, well-organized, and functional, as this will impact your evaluation.