TED Lens: Elevating TED Talk Discovery

Sai Rishi Mannava
Purdue University
Fort Wayne, Indiana
manns02@pfw.edu

Christopher.M.G
Purdue University
Fort Wayne, Indiana
geevc01@pfw.edu

Karthik Balaji
Purdue University
Fort Wayne, Indiana
balak02@pfw.edu

Abstract

TED Talks, known for their wealth of knowledge and diverse ideas, continue to serve as an invaluable resource for students and researchers across various domains. Whether used as a foundational resource or for the in-depth exploration of technical concepts, these talks remain a wellspring of inspiration and wisdom. In this project update report, we delve into our journey of data exploration and analysis on the TED Talks dataset. Our aim is to harness the power of Natural Language Processing (NLP) techniques to extract valuable insights from the transcripts of these talks. While the current model has demonstrated significant improvements compared to the initial baseline model, there is still ample opportunity for further advancements.

1 Introduction

1.1 Motivation

TED Talks have become a global platform for sharing innovative ideas, knowledge, and perspectives on a wide range of topics. These talks are transcribed, offering a rich source of textual data that can be harnessed for various NLP applications. Our motivation stems from the potential to unlock deeper insights, trends, and knowledge dissemination through NLP analysis of TED Talks transcripts. This project aims to bridge the gap between spoken content and the accessibility of information for a broader audience.

1.2 Problem Statement

The problem statement involves the development of a language model aimed at generating relevant tags or keywords for TED Talk transcripts. These generated tags will serve as short descriptors that enhance the organization, discoverability, and searchability of TED Talks. By automatically generating tags, users can find talks on specific topics of interest more efficiently. The model can also contribute to personalized recommendations and improved content organization. The goal of this project is to leverage NLP techniques to enhance the accessibility of TED Talk content.

2 Dataset

For this project, we will initially utilize the TED Talks transcripts dataset available from the official TED website. The dataset contains transcripts for a diverse range of talks spanning various topics and languages. Depending on the project's progress, we may also explore augmenting the dataset with additional sources or focusing on specific subsets. There is a previous version of the TED talk dataset available in kaggle, which has the data in a structured format. Looking at that dataset for reference, there are interesting features and columns in the dataset. The transcripts are available for 12 different languages with more than 4k talks in each of them. The columns transcripts, description, topics, authors, related content, occupations, about-speakers etc. can be a good source for experimenting with the model and generating meaningful insights. The link to the dataset is given below. [1].

3 Model/Algorithm

3.1 Tag Generation

In the initial baseline model, the simplest neural network Logistic Regression model had been used. Since Logistic regression could have only be used for binary classification, an additional layer called OneVsRestClassifier was employed to allow multilabel classification. The model used word2vec to convert natural language to word embeddings, utilizing the Google News pretrained model with 300 dimensions. At that time, the mean of all embeddings was taken, making the input size of a talk 300. This data was then fed into the model, with

[1] https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset

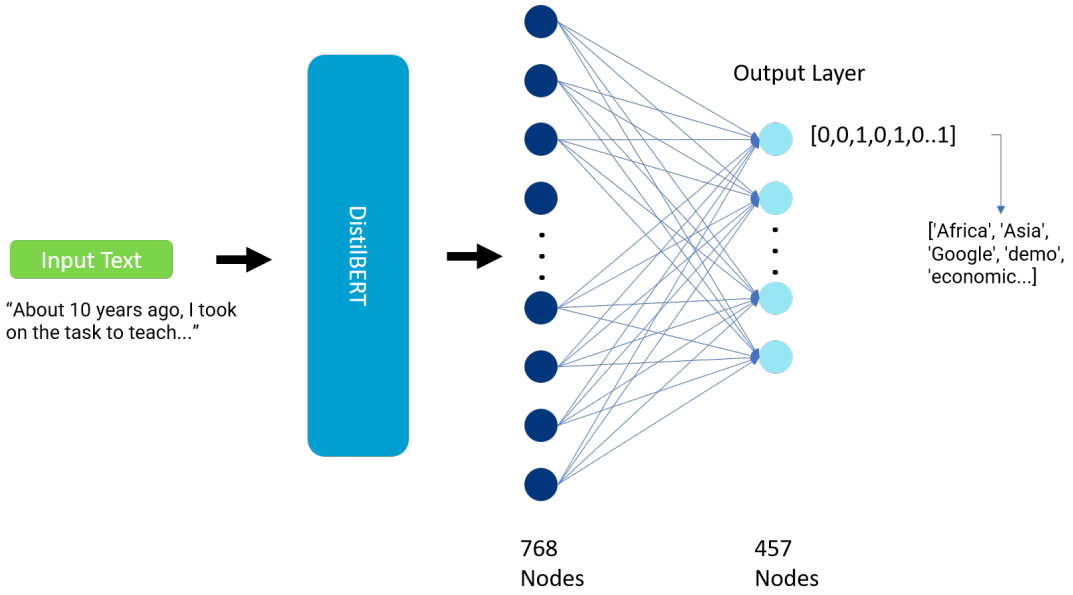

Figure 1: DistilBERT supported Architecture for TEDLens

the additional layer, and the binary representation of the column 'topic' was passed as the labels.

The latest architecture is supported by the DistilBERT model, which is a distilled version of the BERT (Bidirectional Encoder Representations from Transformers) model that is designed to be more computationally efficient while retaining substantial language understanding capabilities. Here, DistilBERT serves as the initial transformer-based layer, leveraging pre-trained contextualized embeddings that is expected to capture intricate patterns and semantic information from the input data.

Following the DistilBERT layer, a simple feed forward neural network is employed to further learn and process the representations from the transformer model into doing our classification task. This neural network operates as a feature extractor, transforming the high-dimensional output of DistilBERT into the corresponding tags associated with it. The output layer consists of 457 nodes, each representing a specific class or category. The integration of DistilBERT and the subsequent neural network forms is expected to form a powerful architecture capable of handling intricate language tasks including this multiclass classification problem.

3.2 Talk Recommendation

This model is still under development and the model utilized bidirectional LSTM to learn the features from the text. The evaluation criteria of the model is still under research.

4 Current Results and Analysis

The model has significantly improved compared to the simple Logistic regression model but still not satisfactory yet. The F1 score of the previous model 0.11, whereas here it has improved to 0.457. The model was trained over 100 epochs.

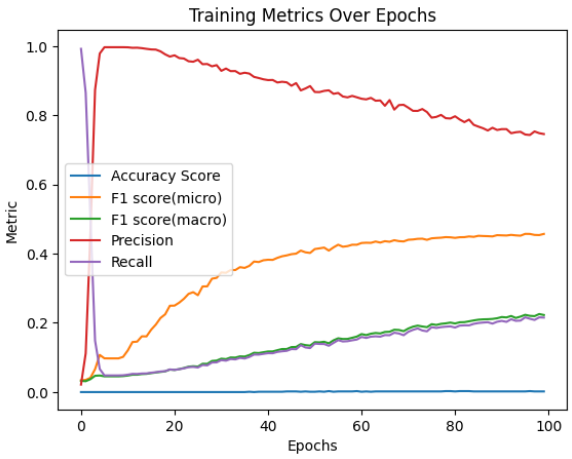

Figure 2: Visualization of Training Metrics over epochs

Looking at the metrics while training(Refer to Figure 2), it is evident that the F1 score is improving. But it is notable that the high value of precision is pushing the F1 Score. The recall score is what needs to be improved, which it does but slowly due to the lower learning rate. Further optimization efforts may be required to strike a balance between precision and recall, ultimately enhancing the model's ability to correctly identify instances

of interest. The average no of tags generated has reached 6.77 where the desired average is 7.97. So the number of tags generated have also been improved compared to the previous model. On the test dataset, the F1 score is 0.42, precision is 0.80 and recall of 0.55, which appears to be performing well but the metrics has to be re-verified.

The data imbalance is still a great concern and could be one of the reason behind the huge difference between precision and recall. The average function as well as the zero division parameters of these metrics might also be contributing to these errors. There is also a high chance that the model might be overfit on some of the features since there is no dropout layers included in the model as of now. The model was also experimented with multiple hidden layers which showed poor results compared to the current model. The loss function seems to be still minimizing after 100 epoch which could mean that there is still scope for improvement.

5 Upcoming Results and Analysis

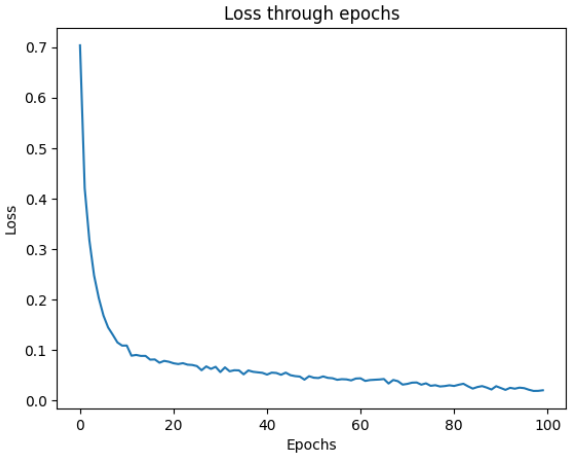

Figure 3: Visualization of Loss function over epochs

The results gives hope that upon correcting the mistakes and investigating more on the data imbalance, this could turn out to be better. The next set of tasks for this project includes, addressing the huge gap between precision and recall, trying out methods like weighted average in while metric calculations, trying to find a global minimum by experimenting with learning rates. The loss function of the current model seems to be minimizing well(Refer Figure 3) but it might not have reached the minima yet. Also convolutional layers is likely to do better than simple feedforward nets here. So the next task is of this project is to experiment with some convolutional layers and see if it improves the performance.

Also, if time permits plan to expand this to generating tags in the multiple domains. We plan to parallely implement the recommendation system that will recommended related content.

6 Conclusion

This project aims to harness the power of NLP to unlock the vast knowledge and insights contained within TED Talks transcripts. By developing innovative algorithms and tools, we aspire to make this valuable resource more accessible, insightful, and impactful for a global audience. The inclusion of a recommendation system will enhance the user experience, providing personalized and engaging TED Talk recommendations based on user preferences. We are excited to embark on this journey and look forward to the contributions this project can make to the fields of NLP, transformers, recommendation systems etc. In this update some of the basic approaches were tried out which gave us light to the future developments in the domain.