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. Our aim is to harness the power of Natural Language Processing (NLP) techniques to extract valuable insights from the transcripts and descriptions of these talks. The project focuses on generating tags from transcripts. This paper presents our experiments and comparisons of three different models, ranging from simple logistic regression to custom transformer-based architectures. Additionally, the paper outlines the development of a recommendation system using tags generated by the model.

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. By doing so, we strive to make spoken content more accessible, ensuring that valuable information reaches a wider audience and addressing the challenges associated with information retrieval from speeches.

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, this capability can be extended on a large scale, allowing users to find talks on specific topics of interest more efficiently.

This problem is approached as a multilabel classification problem. Given a text, predict multiple labels for the given text. That is, given a list of tokens $T = \{t_1, t_2, \ldots, t_k\}$, the objective is to predict a set of labels $L = \{l_1, l_2, \ldots, l_n\}$ that best characterize the content of the token sequence. Each label l_i corresponds to a specific thematic or categorical aspect. The model is trained to learn the mapping function $f : T \rightarrow L$.

2 Related Work

This research overview explores recent progress in the classification of multiple labels within text content, covering various uses and methods. The first study, (Sadat and Caragea, 2022) introduces a way to classify scientific documents that takes into account the complex structures found in this type of text. They attempted to generate tags for scientific documents from 1,234 categories. (Ding et al., 2022) work, looks at classifying topics in a broad range of contexts, investigating strategies for handling different subjects. Their architecture impressive where they have trained it on wikipedia dataset and can even assign labels that the models have never seen. (Antypas et al., 2022) research, looks into classifying topics on Twitter, considering the unique aspects of social media content. They make use of BERT and implemented a custom architecture to classify the twitter content. Additionally, a project on Kaggle by (Afzal, 2023) explores using transformers to classify toxic comments, giving practical insights into how these advanced models can handle and classify toxic language. The model returns multiple labels for a given text data as well. Some of the code for TED Lens, including training script and resource allocation was taken from the code base of this project. Together, these studies provide valuable insights across different areas,

Dataset Name	Description	Transcript	Topics
	(Average Token Length)		(Average count)
TED Talks Dataset	58.20 ± 0.60	1796.85 ± 31.10	7.92 ± 0.12

Table 1: Average Token Lengths in the TED Talks Dataset

such as scientific literature, articles, social media etc. They highlight the diverse uses and methods in the ever-evolving field of classifying multiple labels in text content.

3 Dataset

For this project, the TED Talks dataset is downloaded from Kaggle[1], which is a structured version of data available from the official TED website. The dataset contains transcripts for a diverse range of talks spanning various topics and languages. In future, we may also explore augmenting the dataset with additional sources or focusing on specific subsets. 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. are rich in features which can be used for a variety of problems. For the current task of tag generation, we could make use of the columns *Description* and *Transcripts* for training the model against the column *Topics*. Considering the average number of tokens in both these columns, the Transcript column is too lengthy considering the maximum token input limit of the BERT model being 512. Refer to to Table 1 for the average token lengths of these columns. Hence, for the experiments in this paper we have used the Description column data. In future, we ll focus on adding layers to increase the input dimension to the model.

4 Methodology

4.1 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 the additional layer, and the binary representation of the column 'topic' was passed as the labels. The binary output was then decoded to get a list of topics for a given TED talk.

The next model was 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 final model(Refer Figure 1) was a variation of the previous model using DistilBERT. An additional convolutional layer was included in order to capture more features from the output of BERT model. This integration of a convolutional layer serves to augment the capabilities of the DistilBERT model by capturing intricate patterns and semantic relationships within the learned representations. The kernal size of the model was 3. The convolutional layer is followed by a feed forward network that will classify the learned representations to the output layer of 457 nodes(for 457 classes). By combining the strengths of both DistilBERT, convolutional layers and feed forward network, the final model aims to strike a balance between capturing global contextual information and extracting fine-grained features, ultimately improving its performance on this problem of multiclass classification.

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

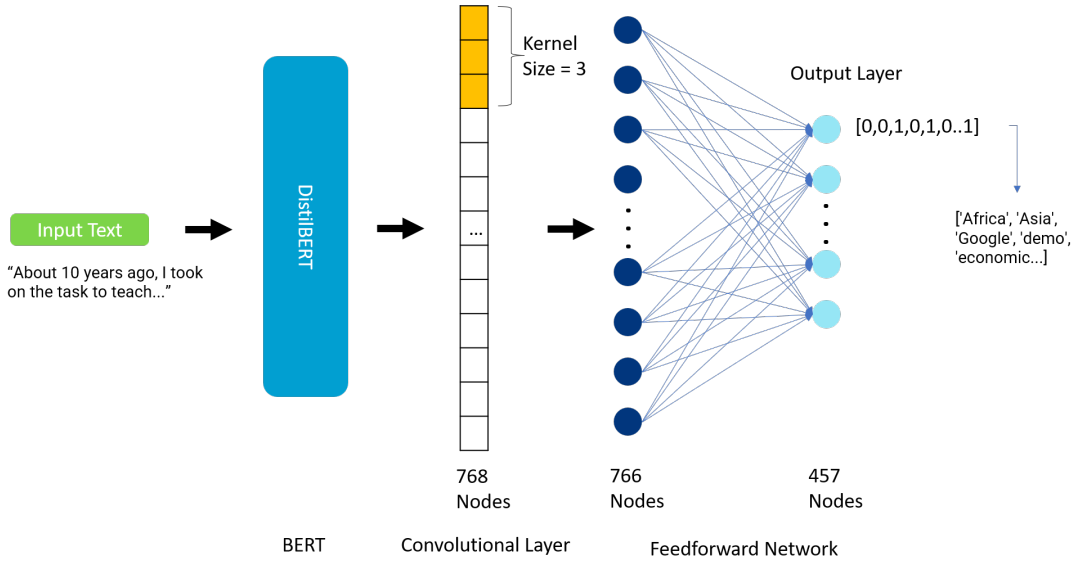

Figure 1: DistilBERT Supported Architecture for TEDLens

4.2 Talk Recommendation

Tag generation can serve various purposes, including content organization, intelligent querying, recommendation systems, and more. In this implementation, we have developed an unsupervised recommendation system based on the tags associated with each talk in the dataset. This algorithm represents a fundamental content-based recommendation system, intended purely as an example to illustrate the utility of tag generation in content recommendation.

Algorithm 1 Talk Recommendation

Input : TED talk id for recommendation
Output : List of 10 closest TED Talks
procedure RECOMMEND_TALK(talk_id)
    Get the tags based from *talk_id*
    $Similarity\_Scores \leftarrow [\ ]$
    $query \leftarrow$ Word2Vec(tags)
    for talk in talks **do**
        Extract tags from talk
        $embedding\_i \leftarrow$ Word2Vec(tags)
        $sim\_i \leftarrow$ Cosine similarity(query,embedding_i)
        Add sim_i, talk_id_i to Similarity_Scores
    end for
    return top 10 scores along with their Ids
end procedure

The system utilizes the Google News Word2Vec model to convert tags into their embedding format. Subsequently, cosine similarity is employed to identify the top 10 talks related to a given talk. Refer to Algorithm 1 for the pseudocode/algorithm for this recommendation system. This approach only uses only the tags/topics of the talks as of now. However this could be expanded by incorporating features like authors name, date, and much more. Currently since this is an unsupervised approach and hence validation can be challenging. But from a human perspective, whatever was recommended was related to talk. Refer to table 2 for a sample example were it demonstrates the title of the input talk and titles of top 5 recommended talks.

Input Talk: "Averting the climate crisis"
Recommended Talks:
1. "The Earth is full"
2. "How will we survive when the population hits 10 billion?"
3. "A new way to remove CO2 from the atmosphere"
4. "How we could change the planet's climate future"
5. "We need nuclear power to solve climate change"

Table 2: Example of Talk Recommendation

5 Experiments & Results

In the logistic regression model, the performance of the model, is not satisfactory. The model was trained for a 100 epochs and the model achieved

Evaluation Metrics					
Model	**Precision**	**Recall**	**F1 Score (Micro)**	**F1 Score (Macro)**	**Avg Tag Count**
LR_Model	0.06	0.01	0.11	0.01	0.6
DistilBert_FW	**0.81**	0.55	0.43	**0.52**	6.7
DistilBert_Conv	0.71	**0.60**	**0.46**	0.51	**7.7**

Table 3: Model Evaluation Metrics

Model	**Tags/Topics Generated**
Ground Truth	TED-Ed, animation, bacteria, health, health care, healthcare, human body, physiology, public health
DistilBERT_FW	TED-Ed, animation, health, history
DistilBERT_Conv	TED-Ed, animation, culture, health, health care, healthcare, psychology, public health

Table 4: Example of Tag Prediction from the test dataset

an F1-score(micro) of only 0.11. The average number of tags generated by the model was just 0.6, which is much less compared to tag count in the dataset. On an average the test dataset contain 7.9 topics/tags per talk. This suggest that on a lot of talks, tags are not getting predicted. Logistic regression model probably only works well the data is linearly separable and it turns out that the data is much complex, making a complex model necessary for this task.

The next model, DistilBERT_FW, the one with DistilBERT followed by a feed forward network, was trained for 100 epochs as well and the model achieved a precision of 0.81, recall of 0.55 and an F1-score(macro) of 0.52. The model was significantly better than the simple Logistic regression model. However the recall score of the model was too less compared to the precision. Also, the average number of tags generated was 6.7, where the average number of tags in the dataset is 7.9. This implies that the model is only confident about a predicting fewer tags compared to the ground truth. The decision threshold of the model was set to 0.25 after experimenting with multiple values to maximize the evaluation metrics.

The next model, DistilBERT_Conv, the one with DistilBERT followed by a convolutional layer which also has a feed forward network in the end, was also trained for 100 epochs and the model achieved a precision of 0.71, recall of 0.60 and F1-score(macro) of 0.51. The model was significantly better than previous model(DistilBERT_FW). The recall score improved to 0.60 making the enabling the model to generate more tags for a given text. The average number of predicted tags has raised to 7.3. This implies that model is now generating as much as tags that the ground truth has. The decision threshold was set to be 0.25 here too based on the experiments. Considering our problem of tag generation, false positives does not have a huge negative impact on this task. The expected outcome of the model should be able to capture a wide range of true positives with a decent precision. So from that perspective this model is considered to be the best of the 3 models that we have experimented with.

6 Additional Analysis & Future Work

The DistilBERT_Conv did a great job in analyzing the context and generating as many tags as it can and this model is considered to be the best for this task among the models part of the experiment. Looking at an example(Refer Table 4) from the the test dataset, we can see how well the BERT assisted models were able to capture the context of the text and then generate tags accordingly. But it can be observed that the DistilBERT_FW is only predicting fewer tags. Whereas the DistilBERT_Conv model was able to predict almost all the tags in the groundtruth. The number of tags generated by the model was very close to the ground truth.

However, we can observe that there are still a few important tags in the ground truth which are not predicted by the model like bacteria, human body, physiology. These are some important tags and it should have been recognized by the model which it did not. This could be because of the limited scope of the training data, where instances of these specific tags might be underrepresented or insufficient

for the model to generalize effectively. Additionally, the complexity of the semantic relationships associated with these tags may pose challenges, as the model might struggle to capture nuanced contexts within the training examples. It is also possible that the pre-trained DistilBERT embeddings may not adequately encapsulate these specialized concepts, requiring fine-tuning or adjustments to the model architecture. Another consideration is the impact of tokenization and the potential loss of information during the input encoding process. Fine-tuning hyperparameters, optimizing tokenization strategies, and exploring techniques to address class imbalances may contribute to enhancing the model's ability to recognize and predict these essential tags more accurately.

The future work of this project will include enhancing the recall and precision of the generated tags. It is important to note that the model currently generates tags only for those it has encountered during training. In the next phase of this project, we plan to expand it to open-domain text classification, allowing the model to predict labels it has never encountered before. Furthermore, the current model has been trained exclusively on English talks; however, we aim to broaden its language capabilities by leveraging more sophisticated models pretrained on various languages, such as GPTs. Finally, there is potential to extend this project to a broader domain, wherever video transcripts or text is maintained.

7 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. This paper experimented with a few custom models which as able to generate tags for a text precisely. However, certain limitations were identified and discussed. These drawbacks, now acknowledged, serve as guiding directives for refining and advancing the proposed models. The evolving nature of this research lays the foundation for continuous improvement, offering a pathway towards more sophisticated and effective NLP applications for extracting knowledge from spoken language across diverse domains. Code to our project is made available here [2]

[2]https://github.com/christopher-2000/TEDLens-NLP-Project

References

Bilal Afzal. 2023. Toxic comments classification using transformers.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hantian Ding, Jinrui Yang, Yuqian Deng, Hongming Zhang, and Dan Roth. 2022. Towards open-domain topic classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 90–98, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Mobashir Sadat and Cornelia Caragea. 2022. Hierarchical multi-label classification of scientific documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8923–8937, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.