CS 7641 Machine Learning Project Proposal

Christine Olds, Christian Lee, Jeongrok (Michael) Yu,
Tae Eun (Harrison) Kwon and Alexandra Pfleegor

I. Introduction/Background

A. Literature Review

There exists a significant body of work investigating how TED Talks popularizes topics [1] [2] [3]. The previous encoder-decoder model conducts fine-tuning using a dataset with human-annotated summaries to perform text summarization [4] [5] [6]. While the autoregressive model is more designed for text generation, some studies involve prompt engineering to inspect the few-shot or zero-shot learning abilities for text summarization [7]. Using nonlinear methods such as Support Vector Regression and Gaussian radial basis functions on a video's features, popularity pattern predictions can be improved to provide more precise prediction results [8]. Sentiment Analysis can be utilized to garner the perspective or attitude from a body of text, allowing analysis of negative, neutral, or positive feelings it may invoke in its reader [9]. Apriori association rule mining can cluster groups of commonly occuring keywords [10]. Features such as keywords, sentiment, etc. can be used to classify and predict the popularity of TedTalks [11].

B. Data Description

The dataset includes information about 5,662 TED Talks. The following information is included for each TED Talk: id, duration (seconds), event, likes, link, publishing date, recorded date, list of related videos, list of speakers, list of subtitle language options, summary, title, list of topics, views, full transcript, YouTube video code.

C. Dataset Link

The dataset can be accessed on Kaggle using this link.

II. Problem Definition

A. Problem

Create a regression model to predict the popularity of TED Talks based on association rule mining and sentiment of transcript, and compress these transcripts to summarize and generate an engaging title.

B. Motivation

Understand how the commonly used words and sentiment of a TED Talk affect its popularity.

Challenge state-of-the-art text summarization models to encapsulate a long transcript into one title sentence.

III. Methods

A. Data Preprocessing

1) Text Summarization:

- divide a text into a list of sentences to distinguish abbreviation words, collocations, and words that start using NLTK sentence tokenizer.
- split up the transcript to not exceed the token limit for the Transformer-based models.

2) Sentiment Analysis and Association Rule Mining:

- convert text to lowercase, remove stopwords, and create a feature vector [9].

B. Machine Learning Algorithms/Models

1) Supervised Methods:

- Create a support vector regression (SVR) model that predicts the popularity of a TED Talk based on features collected from association rule mining and sentiment analysis of the transcript.
- Look into the baseline results of text summarization models such as T5 [4], BART fine-tuned for summarization task [5], PEGASUS-x [6], and GPT-3.5 [7].
- Further pre-train the models with the summary from the TED dataset and compare the performance

2) Unsupervised Methods:

- Sentiment Analysis - Utilizing the NLTK library on Python, calculate the overall sentiment of each TED Talk's transcript [9].
- Association Rule Mining - Utilizing the PyCaret library on Python, generate clusters of frequently co-occurring words in TedTalk titles and transcripts [10].

IV. Results and Discussion

A. Quantitative Metrics

1) Regression Analysis: By splitting the dataset into training and testing, the Mean Squared Error (MSE), the average of the squared differences between predicted and actual values, can be used to measure the accuracy of the SVR model.

2) BLEU (Bilingual Evaluation Understudy): evaluates the machine-translated text by computing a score based on the precision of n-grams and it is formulated as follows [12]:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where:

$$\begin{aligned}
\text{BP} &= \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases} \\
c &= \text{candidate length} \\
r &= \text{reference length} \\
w_n &= \frac{1}{N}
\end{aligned}$$

3) ROUGE (Recall Oriented Understudy for Gisting Evaluation): evaluates the summaries generated by the machine, focusing on the shared n-grams, word sequences, and word sequences weighted by their frequency with respect to the reference [13].

$$\text{ROUGE}_{\text{N}} = \frac{\sum_{\text{summaries}} \sum_{\text{n-grams}} \text{overlap}_{\text{N-gram}}}{\sum_{\text{reference}} \sum_{\text{n-grams}} \text{count}_{\text{n-gram}}}$$

where:
$\text{overlap}_{\text{N-gram}}$ = shared N-grams between summary and reference
$\text{count}_{\text{n-gram}}$ = N-grams in reference

B. Project Goals

C. Expected Results

V. Contribution Table

Name	Contributions
Harrison	
Michael	
Christian	
Christine	
Alexandra	

References

[1] G. S. di Carlo, "The role of proximity in online popularizations: The case of ted talks," *Discourse Studies*, vol. 16, no. 5, pp. 591–606, 2014. [Online]. Available: https://doi.org/10.1177/1461445614538565

[2] K. MacKrill, C. Silvester, J. W. Pennebaker, and K. J. Petrie, "What makes an idea worth spreading? language markers of popularity in ted talks by academics and other speakers," *Journal of the Association for Information Science and Technology*, vol. 72, no. 8, pp. 1028–1038, 2021. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24471

[3] C. R. Sugimoto, M. Thelwall, V. Larivière, A. Tsou, P. Mongeon, and B. Macaluso, "Scientists popularizing science: characteristics and impact of ted talk presenters," *PloS one*, vol. 8, no. 4, 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0062403

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *CoRR*, vol. abs/1910.10683, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019. [Online]. Available: http://arxiv.org/abs/1910.13461

[6] J. Phang, Y. Zhao, and P. J. Liu, "Investigating efficiently extending transformers for long input summarization," *ArXiv*, vol. abs/2208.04347, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251442728

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[8] T. Trzciński and P. Rokita, "Predicting popularity of online videos using support vector regression," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2561–2570, 2017.

[9] N. Rathee, N. Joshi, and J. Kaur, "Sentiment analysis using machine learning techniques on python," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 779–785. [Online]. Available: https://doi.org/10.1109/ICCONS.2018.8663224

[10] C. M. Rahman, F. A. Sohel, P. Naushad, and S. M. Kamruzzaman, "Text classification using the concept of association rule of data mining," *CoRR*, vol. abs/1009.4582, 2010. [Online]. Available: http://arxiv.org/abs/1009.4582

[11] R. Obiedat, "Predicting the popularity of online news using classification methods with feature filtering techniques," *Journal of Theoretical and Applied Information Technology*, vol. 98, p. 8, 04 2020. [Online]. Available: https://www.researchgate.net/publication/359214966_PREDICTING_THE_POPULARITY_OF_ONLINE_NEWS_USING_CLASSIFICATION_METHODS_WITH_FEATURE_FILTERING_TECHNIQUES

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[13] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013