What makes a presentation great?

Milestone 2

# 

# State of the Art

In “Language that Captivates the Audience: Predicting Affective Ratings of TED Talks in a Multi-Label Classification Task” [1], authors Elma Kerz, Yu Qiao and Daniel Wiechmann devised a technique to automatically predict ratings assigned by viewers and explored what types of features drive the classification accuracy for each of the 14 rating categories utilised in the paper. These ratings are: Beautiful, Confusing, Courageous, Fascinating, Funny, Informative, Ingenious, Inspiring, Jaw-dropping, Long-winded, Obnoxious, OK, Persuasive and Unconvincing. A total of 2392 TED Talk transcripts were used, all chosen so that they featured only one speaker and were centred around music performances.

The speech transcripts were automatically analysed using CoCoGen (Complexity Contour Generator) [2]. It implements a sliding-window technique to compute a series of measurements for a given language feature (contours). Unlike other software for automated text analysis that relies on aggregate scores representing the average value of a feature in a text, CoCoGen can track the distribution of the feature scores within the text by advancing sentence by sentence and computing indicator values for each window. The output of CoCoGen for each window of a document is a 119 dimensional vector.

The rating prediction was done by training a Recurrent Neural Network (RNN) classifier that exploits the sequential information in the generated contours. Since the lengths of the feature vector sequences depend on the number of sentences in each text in the corpus, a dynamic RNN was chosen.

The RNN takes the sequence of all vectors outputted by CoCoGen as input and passes it through 5 bidirectional Long Short-term Memory (LSTM) layers with 400 hidden units in each cell. The hidden variable of the last LSTM cell and the hidden variable of the last LSTM cell in the backward direction are then concatenated and transformed through a feed-forward Neural Network.

The NN consists of two fully connected (dense) layers with a Batch Normalisation layer, a Parametric Rectifier Linear Unit (PReLU) layer and a dropout layer between them. The output dimensions of the dense layers are 400 and 14.

The dataset was split in 80% training and 20% testing and a 5-fold cross-validation was applied. The chosen loss function for training was binary cross entropy.

The result was an average total accuracy of 69% across the 14 rating categories. The highest accuracy was reached for the Persuasive category (77%) and lowest for the Long-winded category (62%).

In “Prediction of Standing Ovation of TED Technology Talks” [3] the authors use a Convolutional Neural Network with two convolutional layers with max pooling and three fully-connected layers. Their dataset consists of 173 speeches with standing ovations (more than 10% of the audience offered a standing ovation), and 299 speeches without standing ovations.

The data was split into 75% training data and 25% test data and the NN was trained using batch learning. The result was an accuracy of 77.11% and a 0.63 F-measure.

In “A Causality-Guided Prediction of the TED Talk Ratings from the Speech-Transcripts using Neural Networks” [4] two NN architectures were examined: Word Sequence Model and Dependency Tree-based Model.

The Word Sequence Model is a “Bag-of-Sentences” model in which each sentence is modelled using the LSTM and the outputs are averaged for predicting the scaled and binarized rating counts. The averaged output vectors are then passed through a feed-forward network to produce a 14-dimensional output vector corresponding to the rating categories. Finally, an element-wise sigmoid activation function is applied to the output vector. This model achieved an average recall of 0.76.

The Dependency Tree-based Model uses the SyntaxNet dependency parser to represent each sentence as a hierarchical tree of dependent words. The trees are then processed using the child-sum TreeLSTM [5]. The network learns parameters such as dependency type embedding and parts-of-speech embedding through back-propagation. They are then concatenated with the GloVe [6] word vectors. This model achieved an average recall of 0.77.

The dataset consisted of 2231 talks, 150 of which were used for testing. The training dataset was split into 90% for training the NNs and 10% for development. The development set was used for tuning the hyper-parameters, adjusting the learning rate and controlling the regularisation strength, and to select the best model for final evaluation by examining the loss.

Regularisation for both models was done using a weight-dropping technique: the dropout operation was applied to the hidden-to-hidden weight matrices.

# Corpora

1. [TED-LIUM](https://lium.univ-lemans.fr/en/logicielscorpus/#TED-LIUM-Release-3):
   * 3 iterations
   * 2351 automatic transcripts in the latest one
2. [TCSE (Ted Corpus Search Engine)](https://yohasebe.com/tcse/):
   * 4938 talks with English transcripts
   * annotations for laughter and applause
   * segments and keywords
   * scores for talk speed and readability
3. [Talk Corpus](http://talkcorpus.com/):
   * 2051 talks with English transcripts
   * Felsch Reading Ease score
   * talk length and words per minute
   * number of words from New Academic Word List and New General Service List
4. WIT 3 ([Web Inventory of Transcribed and Translated Talks](https://wit3.fbk.eu/home)):
   * 1556 English transcripts in XML format
   * annotations for laughter and applause
   * number of words and of characters
   * segments and keywords
5. Kaggle datasets such as [TED Talks Transcripts for NLP](https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset), [TED Talk Transcripts (2006 - 2021)](https://www.kaggle.com/datasets/thedatabeast/ted-talk-transcripts-2006-2021) or [TED Talks](https://www.kaggle.com/datasets/rounakbanik/ted-talks)
6. Crawling the TED Talks website with Selenium and BeautifulSoup

**Notes:**  
  
**The Flesch Reading Ease Score** is a measure of the readability of a written text. It's calculated based on the average number of syllables per word and the average number of words per sentence. The resulting score ranges from 0 to 100, where a higher score indicates that the text is easier to read. A score of 60-70 is considered standard, and a score of 90-100 is considered very easy to read. Used in education, journalism, and other fields where the readability of a text is important.

The **New Academic Word List (NAWL)** and the **New General Service List (NGSL)** are two lists of English words that are widely used in language teaching and learning. The NAWL contains 963 words that are frequently used in academic writing, while the NGSL contains 2800 words that are commonly used in everyday communication. These lists were developed based on extensive research into the most frequently used words in English, and they are widely used by language teachers and learners to help improve their vocabulary and language proficiency. The inclusion of NAWL and NGSL word counts in corpora like Talk Corpus can help researchers and educators better understand the language used in TED talks and other speech data, and can inform language teaching and learning strategies.

**TED-LIUM, TCSE, Talk Corpus,** and **WIT 3** are all high-quality corpora that can be used for a wide range of natural language processing tasks, such as speech recognition, sentiment analysis, and topic modelling.

**TED-LIUM** is a speech database that contains automatic transcripts of TED talks, making it useful for research on speech recognition and related applications. The latest version of TED-LIUM has a large number of transcripts, making it a valuable resource for training and evaluating speech recognition models.

**TCSE** and **Talk Corpus** both contain English transcripts of TED talks, with annotations for laughter and applause, segments and keywords, and scores for talk speed and readability. These features make them useful for research on speech analysis and natural language processing. In addition, Talk Corpus also contains the Flesch Reading Ease score, which can be considered as an additional feature when training our models.

**WIT 3** is a web inventory of transcribed and translated talks, making it a useful resource for research on machine translation and cross-lingual speech analysis.

# Approach

Data gathering is an essential aspect of the project as it is the foundation upon which all subsequent analysis and modelling is based. To ensure the quality and relevance of our data we have decided to limit our TED talk transcripts to only those in English that meet a couple of specific criteria.

For this approach, for practical reasons we are considering making use of only TED talks in English with a duration of at least 8 minutes and no more than 18 since this is just enough time and space to explore a topic in depth and short enough to keep the audience engaged. This time constraint forces speakers to be concise and direct with their message delivery.

We plan on avoiding the talks that have more than one speaker and any talks centred around music performances due to practical constraints. The presence of multiple speakers can introduce complexities in the terms of speaker identification, multiple opinions on the topic at hand, turn-taking and co-reference resolution which in turn can make it more difficult to accurately extract meaningful linguistic patterns from the data. By focusing on a single speaker, we can better analyse their use of language, structure and rhetorical techniques.

Finally, we’re excluding talks with music performances and other distractions since this allows us to focus solely on the speaker’s words and presentation delivery without the influence of other sensory stimuli the audience might be picking.

By gathering this specific type of data, we aim to provide a more focused and comprehensive analysis of the language used in TED Talks. This will allow us to better understand the underlying patterns and structures in successful presentations, and potentially apply this knowledge to improve public speaking and communication skills.

To simplify the task of prediction, we will standardise the ratings in order to mitigate the influence of the extraneous factors and undesirable variables like the speaker’s notoriety, media coverage, contemporary hot topics and upload date, since newer presentations will have naturally fewer views which could cause some biases.

Our main focus will be on several key metrics and features of interest which will be including the views, comment numbers, ratings, key words and key phrases, reading ease, words per minute and audience engagement through laughter and applause. This list will be further refined as we begin our experimentation. None of the corpora identified so far provides all these metrics, so we plan to use web crawling and automation to collect the missing ones.

We plan to perform several tasks based on the transcripts of TED Talks. Firstly, we will need to annotate each sentence in the text with its corresponding audience response. This will help us analyse the impact of the speaker's statements on the audience, and how certain phrases or topics may elicit a more significant response.

Secondly, we need to remove stop words from the text. These stop words are words that occur frequently in a language but they do not provide any significant meaning to the text. Removing these words will help us focus on the essential phrases and keywords.

Thirdly, we will tokenize the text, this should help us analyse the structure of the text and identify significant patterns.

Lastly but not least, we need to identify emerging words and key phrases which are non-stop words but are of common use among all discourses. This will help us identify the most important and relevant topics discussed in these speeches and provide more insight into the overall themes and trends of TED talks in general.

The aim is to gain a deeper understanding of the language and structure used in successful presentations, and provide insights for improving the quality of lectures and public speaking.

In order to evaluate the success and effectiveness of a presentation from a feature engineering perspective. We plan on using the following factors as metrics.

Audience rating which is going to provide us with the feedback given by the audience which in turn can provide a valuable insight into how well the message was conveyed and received.

The number of views and comments can in turn be used to measure the popularity and engagement by indicating the extent to which the presentation resonated with its intended audience.

The audience responses such as laughter, applauses can be used to measure the engagement and enjoyment, indicating how well the presenter was able to connect with the audience and keep their attention.

Keywords and word families are also important as they can reveal the most common and important concepts discussed in a presentation, providing an insight into the main themes and concepts conveyed by the presenter.

Words per minute can be used to scale the pacing of a presentation and to determine whether the presenter was able to effectively convey their message within the reasonable time frame and how this impacts the engagement of the audience.

Finally, comprehensibility can be evaluated through measures such as reading ease or the use of complex language, indicating how well the presenter was able to communicate their message in a way that was accessible to their audience.

Taken together, these factors provide a comprehensive picture of the success and effectiveness of a presentation, and can be used to evaluate and compare different presentations in a meaningful way.

We will attempt an approach similar to the Dependency Tree-based Model presented in [4], to analyse the syntactic and semantic structures of the transcript, along with other features. The model uses natural language processing techniques to extract linguistic features from the transcript, such as word frequency, sentence length and part-of-speech tags. These features should be used to train a machine learning algorithm to predict the success of a presentation based on various outcome measures previously discussed and enumerated. The Dependency Tree-based Model offers a promising approach for predicting what makes a good presentation.

We will use the [Syntaxnet Parsey McParseface](https://github.com/spoddutur/syntaxnet) wrapper for dependency learning because of its use of neural network architecture to analyse the grammatical structures of sentences.

We will perform sentiment analysis with [Tree-LSTM](https://github.com/dmlc/dgl/tree/master/examples/pytorch/tree_lstm). Sentiment analysis is another important technique that can be used to identify the tone and emotions expressed in the talk. Tree-LongShortTermMemory Networks are an effective way of carrying out sentiment analysis, as it is capable of understanding the syntax and context of the language used. This allows for more accurate analysis for the expressed sentiments. The implementation of the pytorch library TREE-LSTM algorithm found on github appears to be able to achieve a test accuracy of 51.72 which is comparable with the result reported in the original paper.

To carry out both sentiment analysis alongside a deep analysis of the relationships between words in a sentence, word vectors are necessary. In order to achieve this, we will create word vectors using [GloVe](https://nlp.stanford.edu/projects/glove/), which stands for Global Vectors for Word Representation), GloVe creates a mathematical representation of words based on their appearance alongside with other words. This allows for a more accurate understanding of the relationships between words and their importance in the context of the talk.

We plan to use PyTorch and python’s NLTK to create our models and write our networks following the pseudocode presented in the papers.

# References

[1] Elma Kerz, Yu Qiao, and Daniel Wiechmann. (2021). [Language that Captivates the Audience: Predicting Affective Ratings of TED Talks in a Multi-Label Classification Task](https://aclanthology.org/2021.wassa-1.2/). In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 13–24, Online. Association for Computational Linguistics

[2] Ströbel Marcus, Elma Kerz, Daniel Wiechmann, and Stella Neumann. (2016). [CoCoGen - Complexity Contour Generator: Automatic Assessment of Linguistic Complexity Using a Sliding-Window Technique](https://aclanthology.org/W16-4103/). In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), pages 23–31, Osaka, Japan. The COLING 2016 Organizing Committee

[3] Maeno, S., Maeshiro, T. (2018). [Prediction of Standing Ovation of TED Technology Talks](https://link.springer.com/chapter/10.1007/978-3-319-92043-6_53). In: Yamamoto, S., Mori, H. (eds) Human Interface and the Management of Information. Interaction, Visualization, and Analytics. HIMI 2018. Lecture Notes in Computer Science, vol 10904. Springer, Cham. <https://doi.org/10.1007/978-3-319-92043-6_53>

[4] Md Iftekhar Tanveer, Md Kamrul Hassan, Daniel Gildea, M. Ehsan Hoque. (2019). [A Causality-Guided Prediction of the TED Talk Ratings from the Speech-Transcripts using Neural Networks](https://arxiv.org/abs/1905.08392). <https://doi.org/10.48550/arXiv.1905.08392>

[5] Tai, K. S., Socher, R., & Manning, C. D. (2015). [Improved semantic representations from tree-structured long short-term memory networks](https://arxiv.org/abs/1503.00075). arXiv preprint arXiv:1503.00075.

[6] Pennington, J., Socher, R., & Manning, C. D. (2014, October). [Glove: Global vectors for word representation](https://aclanthology.org/D14-1162.pdf). In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[7] Xu, Hao & Jiang, Chengzhi & Huang, Chuanfeng & Chen, Yiyang & Yi, Mengxue & Zhu, Zhentao. (2022). [Key Information Extraction and Talk Pattern Analysis Based on Big Data Technology: A Case Study on YiXi Talks](https://www.researchgate.net/publication/358754921_Key_Information_Extraction_and_Talk_Pattern_Analysis_Based_on_Big_Data_Technology_A_Case_Study_on_YiXi_Talks). Electronics. 11. 640. 10.3390/electronics11040640.

[8] Pooja Kherwa, Poonam Bansal. (2019). [Topic Modeling: A Comprehensive Review](https://eudl.eu/doi/10.4108/eai.13-7-2018.159623). In European Union Digital Library. DOI: 10.4108/eai.13-7-2018.159623

### Corpora:

[C\_1] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève, [TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation](https://arxiv.org/abs/1805.04699), submitted to the 20th International Conference on Speech and Computer (SPECOM 2018), September 2018, Leipzig, Germany

[C\_2] Hasebe, Yoichiro. (2015). [Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks](https://www.sciencedirect.com/science/article/pii/S1877042815044353). Procedia: Social and Behavioral Sciences 198(24), 174–182

[C\_3] Raine, Paul. (2020). [Talk Corpus: A Web-based Corpus of TED Talks for English Language Teachers and Learners](https://www.researchgate.net/publication/338691522_Talk_Corpus_A_Web-based_Corpus_of_TED_Talks_for_English_Language_Teachers_and_Learners)

[C\_4] M. Cettolo, C. Girardi, and M. Federico. (2012). [WIT3: Web Inventory of Transcribed and Translated Talks](https://aclanthology.org/2012.eamt-1.60.pdf). In *Proc. of EAMT*, pp. 261-268, Trento, Italy