Codebusters

Project 2 - Final Presentation

General task overview

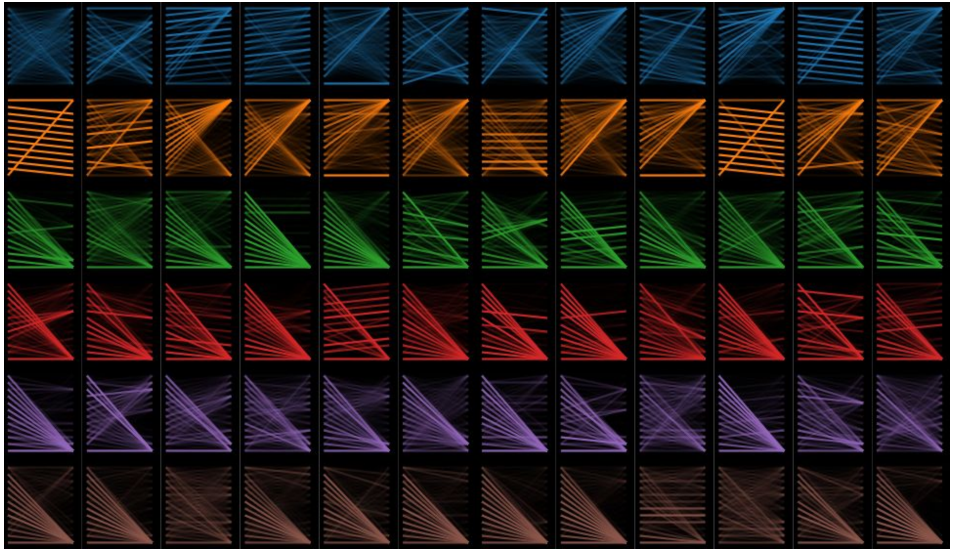
- Our chosen **model**:
  → DistilbertForSequenceClassification

- What **dataset**?
  → Imdb, containing information if review is negative or positive

- ‘Actual’ Bert **task**:
  → Sentiment Analysis

Visualize Attention Difference

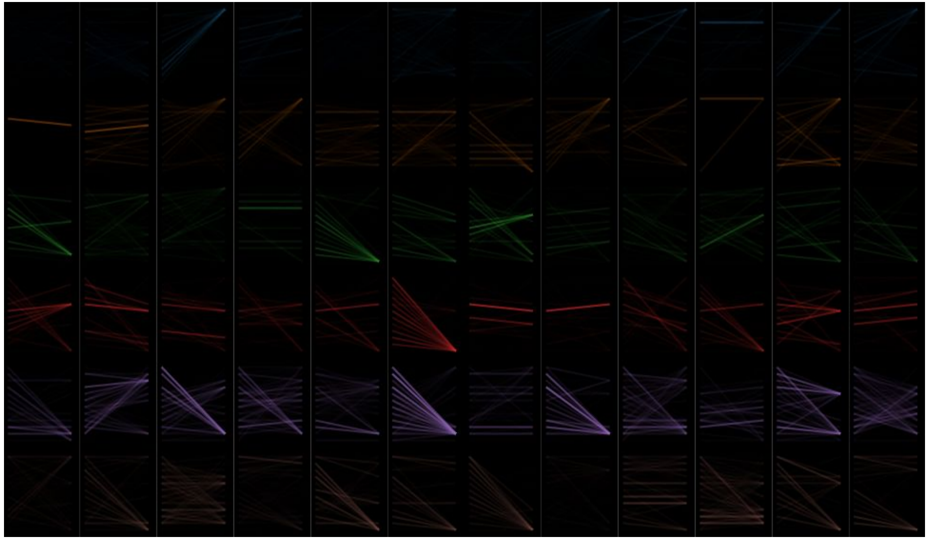
- Visualizing the **attention** of each head (columns) and each layer (rows)
- Used library: BertViz
- Visualize the difference of the attention between a trained and untrained model to explain what has been learned by the model
- Only positive attention is visualized



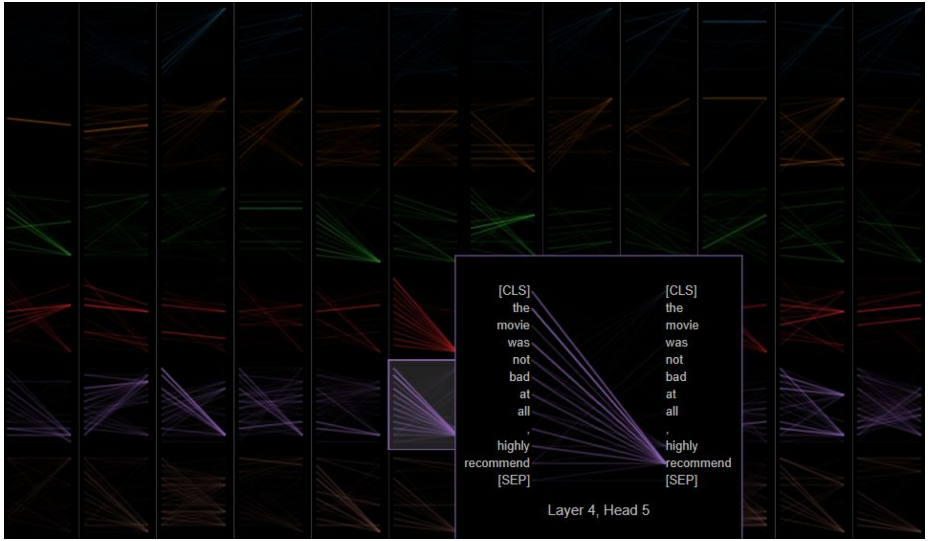
Img 1: Attention of Trained model on positive sentence
“The movie was not bad at all, highly recommend”

Visualize Attention Difference

Sentence: The movie was not bad at all, highly recommend



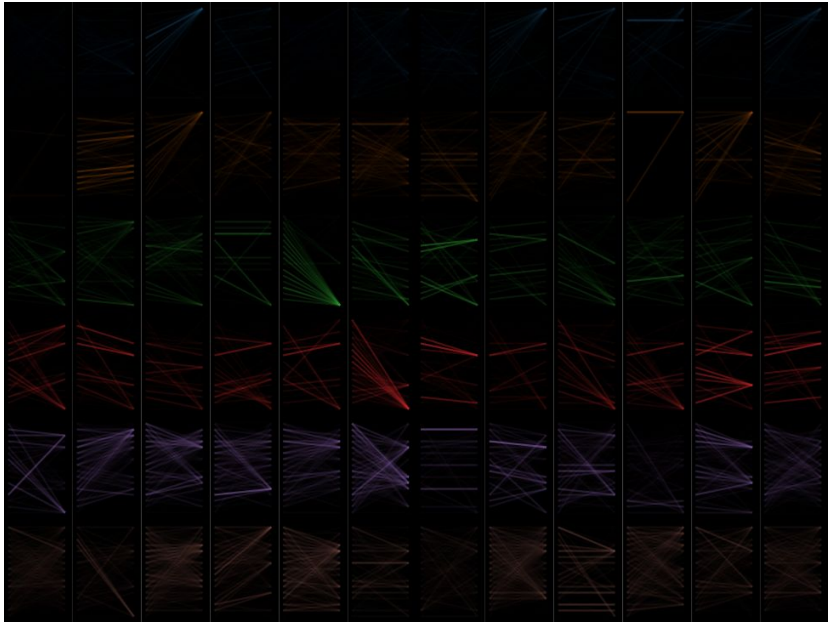
Img 2: Attention Diff between trained and untrained model



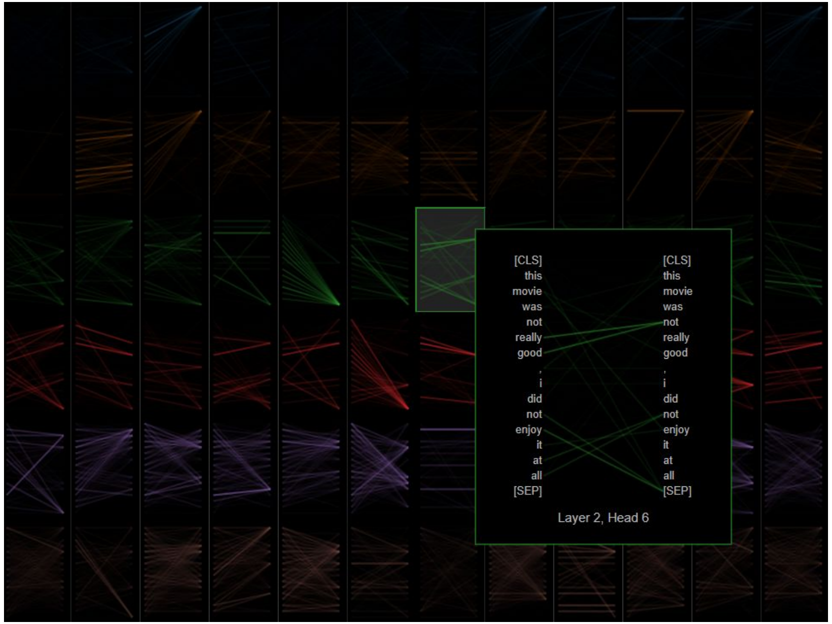
Img 3: Example of reinforced attention in trained model

Visualize Attention Difference

Sentence: This movie was not really good, I did not enjoy it at all



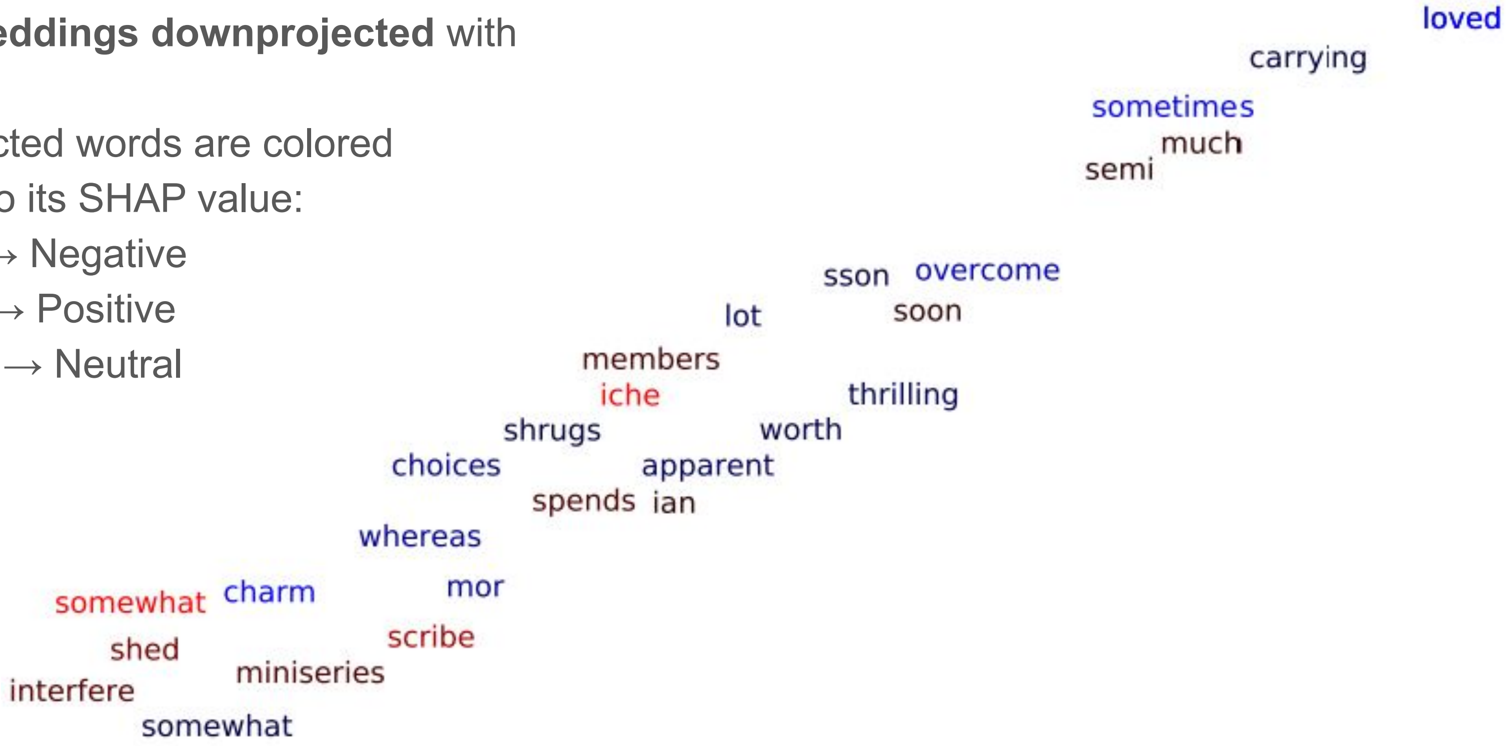
Img 4: Attention Diff between trained and untrained model



Img 5: Example of reinforced attention in trained model

Embeddings visualized

- Word **embeddings downprojected** with t-SNE
- Downprojected words are colored according to its SHAP value:
  - Red → Negative
  - Blue → Positive
  - Black → Neutral

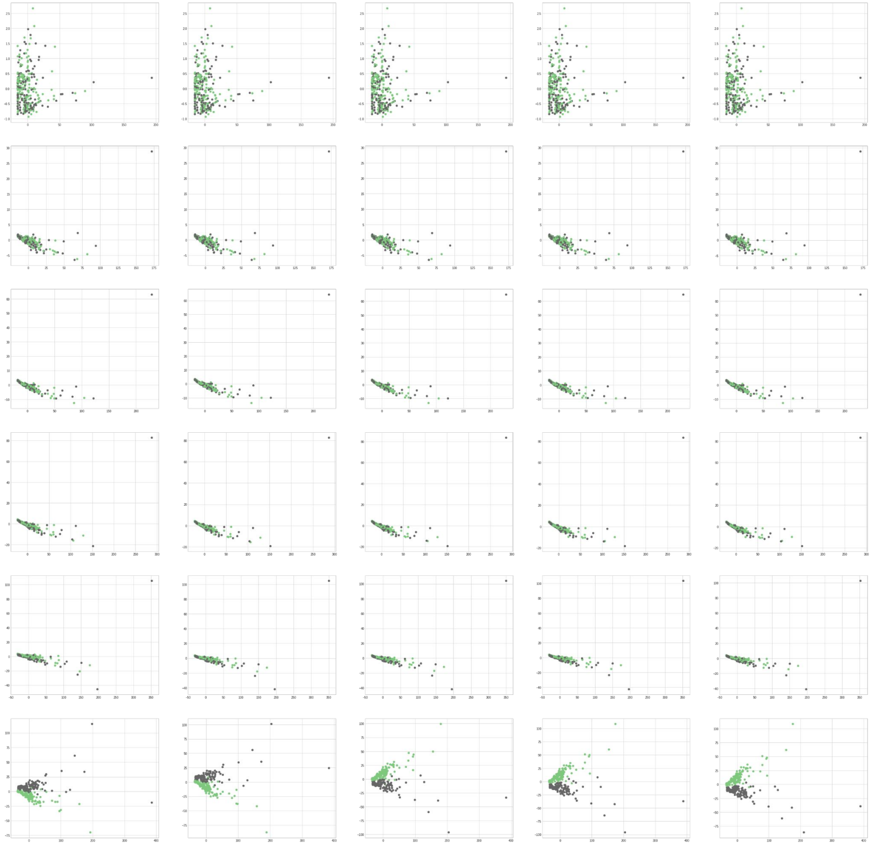


Sentiment Analysis: Hidden States per Layer and Epoch

- Trained DistilBERT model for 5 epochs

- Saved hidden states of model for 200 samples for each of the 6 layers (rows) and 5 epochs (columns)

- **Visualized hidden states** by Downprojection to 2 Dimensions with PCA



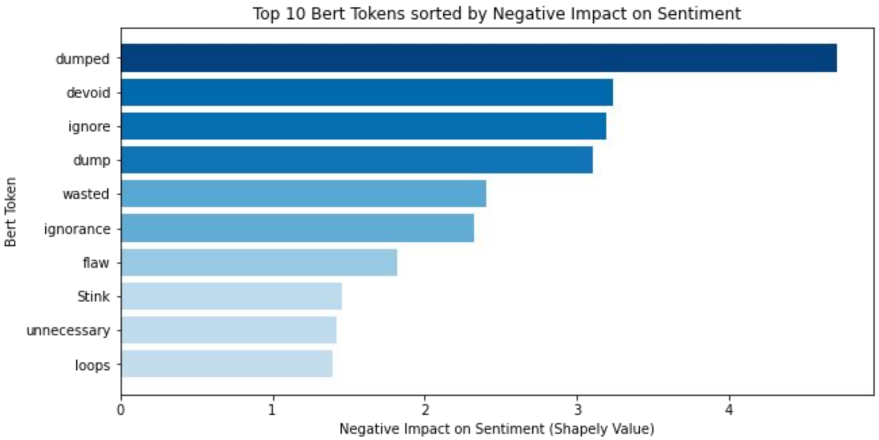
Sentiment Analysis: Visual Representation with SHAP

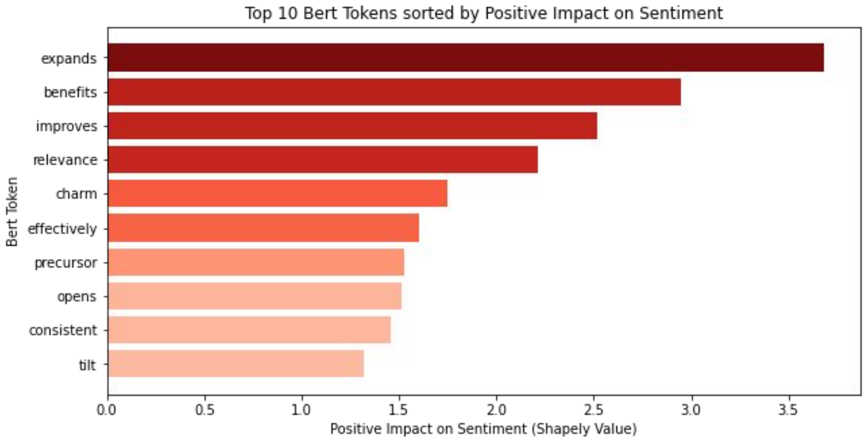
What is SHAP?

- SHAP uses **Shapley-Value** to determine marginal contribution of each token on output
- Can be used on **global** level & on **sentence** level

Observations:

- Analysis of global feature importance shows the tokens with the highest Shapely values
- Includes tokens that generally have a strong sentiment
- Also includes movie-specific vocabulary (‘precursor’, ‘loops’)





	Attention Visualization	Embedding visualization	Hidden states	SHAP
Why?	Interpretability & Explainability	Interpretability & Explainability	Interpretability & Explainability	Interpretability & explainability of the model/the input-output mapping, as well as for improving or comparing models
What?	Learned Model Parameters Computational Graph & Network Architecture	Aggregated Information	Aggregated Information	Learned model parameters, aggregated information
When?	After Training	After Training	During/After Training	After training
Who?	Model Developers & Builders	Model Users	Model Users	Model Developers & Builders / Model Users
How?	Algorithms for Attribution & Feature Visualization	Dimensionality Reduction & Scatter Plots	Dimensionality Reduction & Scatter Plots	Marginal contribution (Shapely Value), various plots (e.g. barcharts)
Where?	Interpretability of NLP Tasks	NLP Sentiment Research	NLP Sentiment Research	NLP, Image classification/object detection, Tabular data