universität freiburg

IDENTIFYING OPINIONS IN INFORMATIVE TEXT

Supervisor : Leonhard Sommer

Sejal Mutakekar, Shanthini Malarvizhi, Sripriya Sridath

MOTIVATION

Know What's Fact and What's Opinion !

- In an era where information is everywhere, it's important to identify whether the content being read is opinionated or informative.
- Opinionated content has personal viewpoints, while informative content provides factual data.

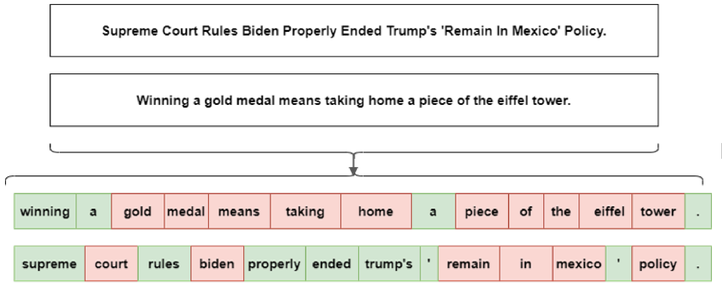
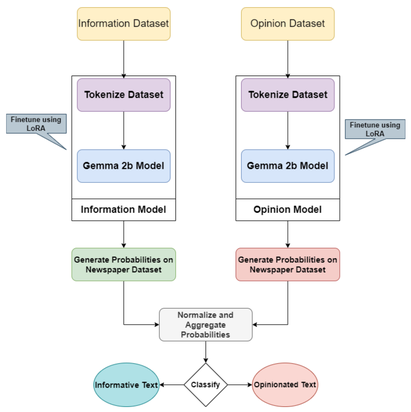


Fig 01: Information vs. Opinion Highlights

METHOD



- Input Text : Two datasets have been extracted from the Hugging Face Platform.
- Dataset1 : VoxPopuli – Speech dataset extracted from European Parliament recordings.
- Dataset2 : CommonVoice – Large-scale open source speech data contributed by volunteers worldwide.
- Model: Gemma-2b

Fig 02: Methodology

- Fine-tuning a large language model (LLM) can be computationally expensive and resource-intensive.

Why LoRA?

- **Efficient Fine-Tuning**: Reduces the number of trainable parameters.
- **Memory and Computational Efficiency**: Enables fine-tuning on smaller hardware.
- **Improved Generalization**: Helps prevent overfitting on small datasets.

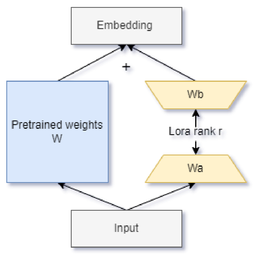


Fig 03: Lora Fine Tuning

How to compare Probabilities?

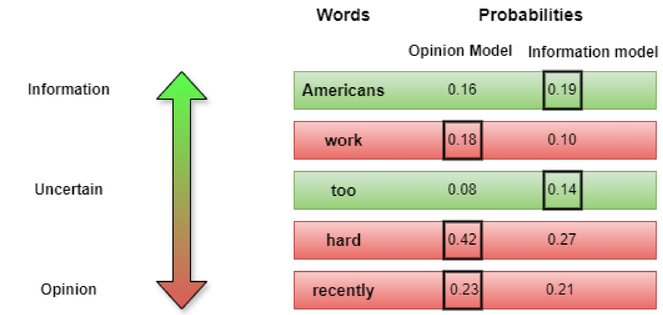


Fig 04: Comparison of probabilities

- Normalized probabilities are compared using the max probability method. (Chosen probabilities are averaged to get the prediction).
- Output : Classified input sentence as either Informative or Opinionated.

Additionally, tokens defining each category are highlighted with specific color to indicate whether they are Informative or Opinionated.
This approach efficiently classifies and visualizes text, highlighting key tokens for better interpretability.

EXPERIMENTAL RESULTS

Datasets : D1 : VoxPopuli - 890115 Transcriptions D2 : CommonVoice - 177091 Transcriptions

Sl No.	Learning Rate	Max Gradient Norm	Lora_r	Lora_alpha	Lora_dropout	Max Sequence Length
1.	1e-4	0.5	8	8	0.2	128
2.	2e-4	0.3	4	16	0.1	128
3.	3e-5	0.5	8	8	0.2	128
4.	3e-5	0.5	8	8	0.2	16

Table 01: Different hyperparameters used for training

- We optimized LoRA for opinion identification by fine-tuning hyperparameters, finding that a learning rate of 3e-5 stabilized training and reduced loss instabilities.
- We selected a LoRA rank of 8 and an alpha of 8 to balance parameter efficiency with model expressiveness.
- For maximum sequence length, we experimented with both 128 and 16 to handle varying input sentence lengths and optimize processing efficiency.

Training Curves :

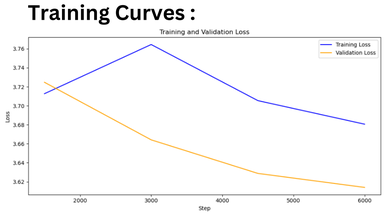


Fig 05: Training and validation loss for CommonVoice Dataset for Max Sequence length 16

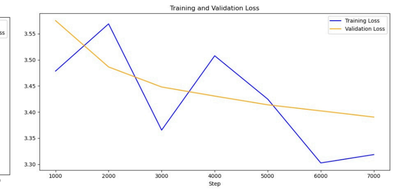


Fig 06: Training and validation loss for Voxpopuli Dataset for Max Sequence length 16

DISCUSSION

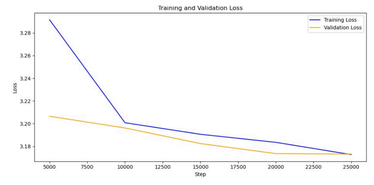


Fig 07: Training and validation loss for Voxpopuli Dataset for Max Sequence length 128

Fig 08: Training and validation loss for CommonVoice Dataset for Max Sequence length 128

- The model attained 90% accuracy rate with a maximum sequence length of 128 and was able to extract the information and context required for an accurate classification.
- Comparatively lower accuracy with a maximum sequence length of 16 due to the loss of important context, leading to less accurate predictions.

Classification Results :

Headline: Justice must be served hot and fresh, like a perfect pizza.

Justice __must __be __served __hot __and __fresh , __like __a __perfect __pizza .

Model Output: *OPINION*

Headline: California Drought Tests History Of Endless Growth.

California __Drought __Tests __History __Of __Endless __Growth .

Model Output: *INFORMATION*

Fig 09: Classification Results with Color-Coded Highlights for Opinions and Informative Text

CHALLENGES AND FUTURE WORK

Challenges :

- Dataset Quality - Tough to find a clean data to enable exact classification.
- Manual data labelling

King, Big Joe Williams, and Ace Atkins.

Future work :

Enhance data quality and automate test data labelling to improve model evaluation.

REFERENCES

- LORA: https://arxiv.org/abs/2106.09685
- Gemma-2b Model: https://huggingface.co/google/gemma-2b
- Fine-tuning:https://huggingface.co/blog/gemma-peft