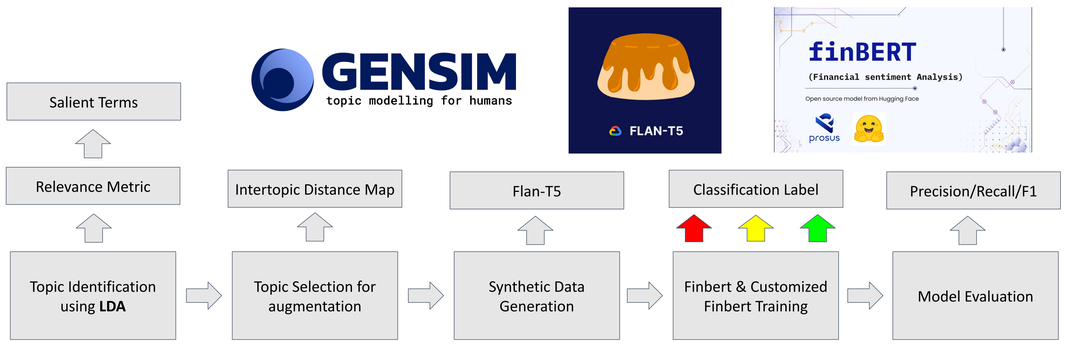
Data-Centric Financial Sentiment Analysis

VANDERBILT UNIVERSITY

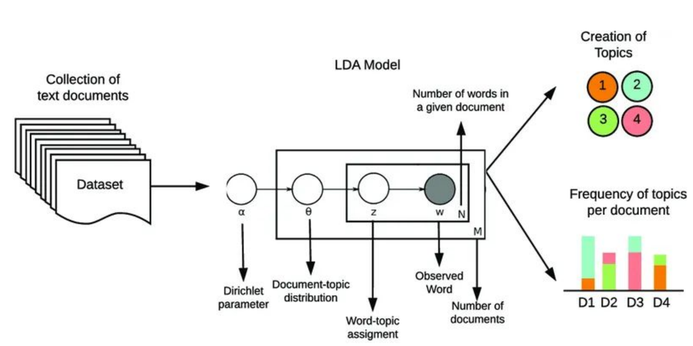
CS 3892/5892: Project in Data-Centric AI and Mining

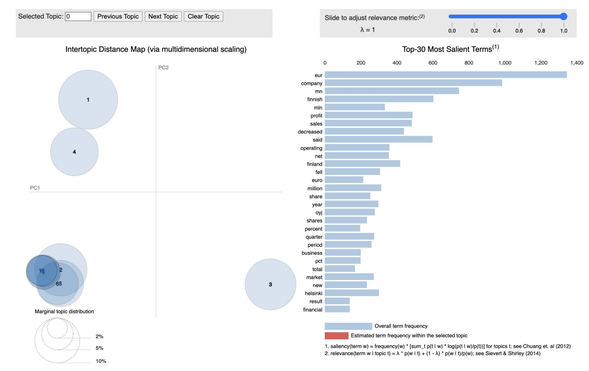
Yifan Wei, Silvie Isingizwe, Ziyao Zhang, Charlotte Zhao

Process Flowchart

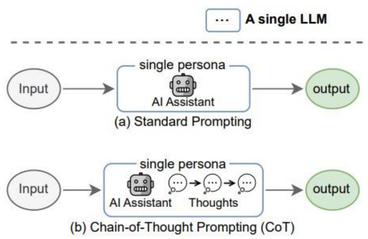


Latent Dirichlet Allocation

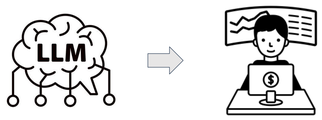




Chain of Reasoning



Persona

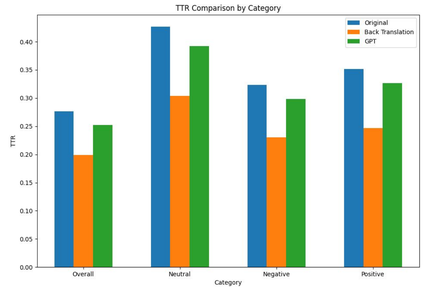


ChatGPT Prompt Generation

The above words are the resultant keywords for topic 3 when performing LDA. Act as a financial sentiment analysis expert. Create five positive, five negative, and five neutral financial headlines as prompts for future synthetic data generation, and explain why you present these headlines using the keywords for this topic.

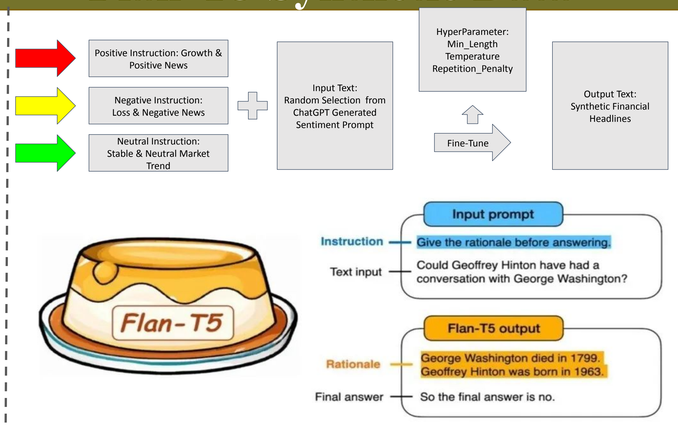
Positive Financial Headlines:

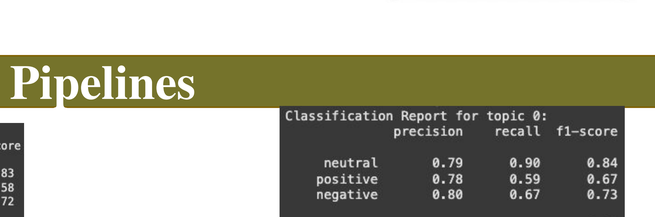
1. **"EUR 500mn Profit Reported as Sales Increase in Latest Quarter"**
   Reason: This headline indicates a profit of EUR 500mn and increased sales, reflecting a positive financial outcome for the company.
2. **"Operating Profit Surges in EUR as Net Sales Jump 15% Year-over-Year"**
   Reason: Words like "surges" and "jump" suggest strong financial performance, making this a positive headline.
3. **"EUR 200mn Net Profit Outpaces Expectations in a Strong Q1 Performance"**
   Reason: "Outpaces expectations" and "strong performance" indicate better-than-expected results, creating a positive sentiment.
4. **"Sales and Profit Both Increase in Latest EUR Quarterly Report"**
   Reason: An increase in both sales and profit is always a positive financial indicator.
5. **"Company Reports EUR 300mn in Net Profit, Highlighting Strong Year-End Growth"**
   Reason: This headline emphasizes a significant profit figure and highlights growth, which are positive indicators for investors.



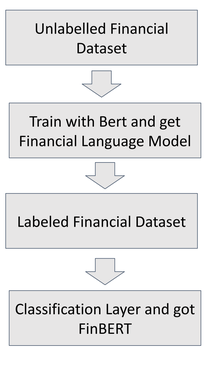
Measure of Lexical Diversity & Higher TTR result suggests greater variety in word usage and is beneficial for sentiment analysis.

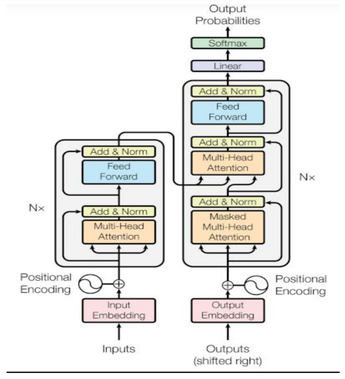
Flan-T5 Synthetic Data





FinBERT





Base Architecture:
- **BERT Core**: 12 layers, 12 attention heads per layer, 768 hidden dimensions, positional embeddings.
- **Pre-trained BERT Model**: FinBERT starts from BERT's pre-trained model and fine-tunes it for financial sentiment tasks, retaining BERT's structural advant
- ages.

Fine-tuning Process:
- **Financial Dataset**: Unlike generic datasets, FinBERT is trained and fine-tuned using financial texts like **company filings**, **earnings call transcripts**, and **financial news**.

Sentiment-Specific Output:
- **Classification Layer**: While the core transformer architecture is identical to BERT, FinBERT introduces a specialized **classification head** fine-tuned for sentiment analysis tasks.

Implementation Pipelines

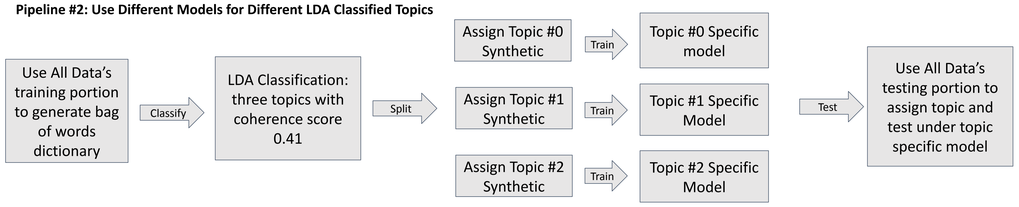
Pipeline #1: Train Augmented Dataset and Test after Dropping Stop Words


```
Classification Report:
              precision    recall  f1-score

     neutral       0.75      0.92      0.83
    positive       0.85      0.44      0.58
    negative       0.71      0.72      0.72

    accuracy                           0.76
```

Pipeline #2: Use Different Models for Different LDA Classified Topics




```
Classification Report for topic 0:
              precision    recall  f1-score

     neutral       0.79      0.90      0.84
    positive       0.78      0.59      0.67
    negative       0.80      0.67      0.73

    accuracy                           0.79
```

```
Classification Report for topic 1:
              precision    recall  f1-score

     neutral       0.78      0.93      0.85
    positive       0.83      0.49      0.62
    negative       0.72      0.59      0.65

    accuracy                           0.79
```

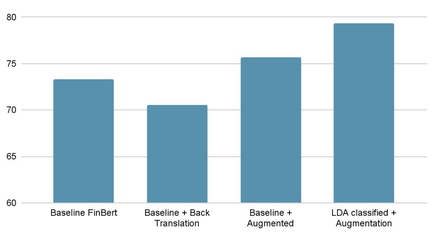
```
Classification Report for topic 2:
              precision    recall  f1-score

     neutral       0.66      0.95      0.78
    positive       0.97      0.74      0.84
    negative       0.92      0.79      0.85

    accuracy                           0.82
```

Result Comparisons

Test Accuracies



Under the augmentation steps that we took, our synthetic dataset used for training out-preformed the Baseline FinBERT by 2.5% and the clustered training approach's average testing accuracy outperformed Baseline FinBERT by 7%. The Back Translation Approach did not do well as it lowers the variability of training dataset.

Potentials and Future Work

Goal: Utilizing CleanLab to identify potential label issues in dataset. Result from the pretrained model is used to compare with the actual label of the data point, a larger difference indicates a lower confidence.

3 Trials CleanLab:

1. With built-in Linear Regression Model
   a. Issue labels identified, but the model is not well-trained.
2. With baseline Bert Model on augmented data
   a. Top 20% data points with low confidence are identified with potential label issues.
   b. After manual inspection, identified label issues are actually correct, proving the quality of the augmented data.
3. With small batch of kaggle data (10% of the original dataset)
   a. Finbert used as the pretrained model with 83.51% of accuracy.
   b. 0 label issues was identified.

Teacher Student Network and Sentence BERT

- Teacher Student Network is able to fine-tune a model based on incoming data and a baseline teacher model
- We can potentially apply Sentence BERT on the Student Network to let it understand more about the sentence embedding instead of only the tokens embedding. Pairs of words can be understood better

References

- FinBERT: Financial Sentiment Analysis with Pre-trained Language Models - Dogu Tan Araci
- Conditional BERT Contextualize Augmentation - Xing Wu
- Confident Learning: Estimating Uncertainty in Dataset Labels - Curtis G. Northcutt
- Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels - Curtis G. Northcutt
- What is Latent Dirichlet Allocation? - IBM
- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks - Nils Reimers