Extracting Balanced Summaries from Politically Biased News using CNN and Pegasus Transformer

Yuliana Jasso, Sergio Valdez

Department of Computer Science, College of Engineering and Computer Science

The University of Texas Rio Grande Valley

INTRODUCTION

Extracting the “truth” from the news can be a challenging endeavor. With numerous news outlets and information sources, cross-referencing information is time-consuming and challenging.

This project aims to leverage classification and transformers to generate unbiased summaries of articles across the political spectrum.

Data Collection: Used webscraping techniques to obtain our data.

CNN for Classification Model: Built a CNN model to classify articles based on political bias.

Pegasus Transformer: Utilized the PEGASUS model for abstractive summarization, which masks important sentences and generates a cohesive summary.

GOALS

- **Develop a classification model** to identify the political bias of news articles.
- **Use Pegasus transformer** to generate unbiased summaries of these articles.
- **Evaluate the effectiveness** of the generated summaries in reducing bias.

PROJECT SUMMARY

Our project involves creating a system that classifies news articles based on political bias and then generates summaries that are intended to be unbiased.

We faced **challenges** such as limited data sources, summary evaluation, and modifying our classification model for better accuracy.

METHODS

Data Collection

We webscraped 300 articles: 100 each from left, center, and right-biased news sources.

Tools used: Requests, BeautifulSoup for scraping; CSV for data storage.

METHODS

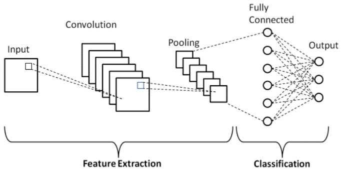
CNN for Classification Model

Convolutional Neural Network (CNN):

Purpose: Classify news articles into left, center, and right political biases.

Architecture:

- **Input Layer:** Processes the text data.
- **Embedding Layer:** Converts text into numerical vectors using embeddings.
- **Convolutional Layers:** Applies filters to capture essential features and patterns in the text.
- **Pooling Layers:** Reduces the dimensionality of the data, preserving important features.
- **Fully Connected Layers:** Integrates the features for final classification.
- **Output Layer:** Provides the probability distribution over the three classes (left, center, right).



METHODS

Pegasus Transformer

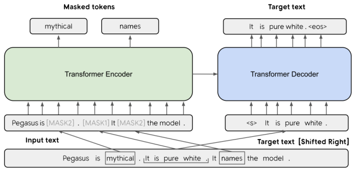
Purpose: Generate unbiased summaries of news articles, condensing them while maintaining the core message and minimizing bias.

Abstractive Summarization: Creates new sentences to summarize the main ideas, rather than copying parts of the text, producing more natural and coherent summaries.

Masked Sentence Generation: Masks important sentences during training and learns to predict them using the rest of the text, helping the model focus on key information for accurate summaries.

Encoder-Decoder Framework:

- **Encoder:** Converts input text into hidden states, capturing the text's meaning.
- **Decoder:** Generates the summary from these hidden states.


Masked tokens
mythical
names
Target text
It is pure white . <eos>
Transformer Encoder
Transformer Decoder
Pegasus is [MASK2] . [MASK1] It [MASK2] the model .
<s> It is pure white .
Input text
Target text [Shifted Right]
Pegasus is mythical . It is pure white . It names the model .

EXPERIMENT SETUP

Data Preparation: Collected and labeled 300 articles for bias.

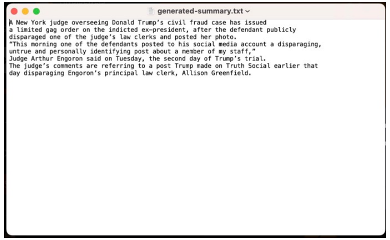
CNN for Classification Model: Trained the CNN model on the dataset to classify articles into left, center, and right bias.

Summarization Model: Used the transformer PEGASUS model to generate summaries of the classified articles.

Evaluation: Calculated accuracy for CNN model and used BERTScore metric for summarization evaluation.

EVALUATION AND RESULTS

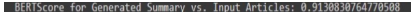
After training our CNN model with the dataset, we fed 3 different articles into our CNN model to classify them. After doing so, we used the Pegasus model to generate the following summary:



```
A New York judge overseeing Donald Trump’s civil fraud case has issued
a limited gag order on the indicted ex-president, after the defendant publicly
disparaged one of the judge’s law clerks and posted her photo.
“This morning one of the defendants posted to his social media account a disparaging,
untrue and personally identifying post about a member of my staff,”
Judge Arthur Engoron said on Tuesday, the second day of Trump’s trial.
The judge’s comments are referring to a post Trump made on Truth Social earlier that
day disparaging Engoron’s principal law clerk, Allison Greenfield.
```

Tested accuracy for our CNN model: 0.7

BERTScore for generated summary vs input article:



```
BERTScore for Generated Summary vs. Input Articles: 0.9130830764770508
```

NEXT STEPS

- **Improve Classification Model:** Enhance the CNN model to better handle nuanced bias and include more in-between affiliations.
- **Advanced Evaluation Techniques:** Implement more sophisticated methods to evaluate the neutrality of the summaries.
- **Expand Data Sources**: Gather more data from a broader range of news outlets to improve model training and evaluation.

REFERENCES

Zhang, Jingqing, et al. “PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization.” 2020.

“Pegasus.” Huggingface.co, huggingface.co/docs/transformers/model_doc/pegasus.

Kirenz, Jan. “Web Scraping in Python with Beautiful Soup and Requests.” Jan Kirenz, 2 May 2022, www.kirenz.com/post/2022-05-02-web-scraping-in-python-with-beautiful-soup-requests-and-pandas/.