Enhancing Trust in News:
A Subjectivity Classification and Mitigation Approach

Omar Jamil, Ke Zhang
University of California, Berkeley, School of Information
ojamil@berkeley.edu, 2ez4kez@berkeley.edu

Abstract

In the age of digitized media and generative AI content, the availability of information has surged, leading to an increase in misinformation. This project aims to enhance trust in news by developing and fine-tuning models to identify and remove subjectivity in news articles. The process involves classifying sentences as subjective or objective and reconstructing articles with the objective content, followed by summarizing these articles using state-of-the-art models. Our results demonstrate that this approach effectively reduces subjectivity and provides more factual content, contributing to improved media literacy and reduced bias.

1 Introduction

With the rise of digitized media and the advent of Generative AI content, the availability of information has surged, but so has the spread of misinformation. Media literacy, the skill set that enables individuals to properly digest media, discern biases, and evaluate arguments from multiple perspectives before forming an opinion, is not widespread. This is particularly true in demographics with lower resources. For instance, a study from Nature[1] shows that individuals with lower socioeconomic status are significantly less likely to possess media literacy skills, making them more vulnerable to misinformation and its consequences.

The impact of limited media literacy is profound, contributing to a more polarized society. Without the tools to critically analyze information, people are more susceptible to confirmation bias, where they favor information that supports their preexisting beliefs. This can lead to the reinforcement of echo chambers and the escalation of divisive views. For example, a study shows how misinformation regarding vaccines has not only led to public health crises but also exacerbated societal divisions, highlighting the urgent need for improved media literacy.

2 Background

In response to the increased spread of misinformation and highly biased content, our goal is to help empower the reader with tools that will aid in their ability to inform themselves. Primarily, our aim is to develop and fine-tune a model that can accurately differentiate between subjective and objective statements in news articles and leverage this model to provide practical insights such as a 'subjectivity-score' and an objective summarization.

A recent study [2] demonstrates the effectiveness of transformers in sentence-level subjectivity detection within news articles. Transformers, with mechanisms for capturing long-range dependencies and contextual relationships, make them particularly adept at identifying the nuanced language differences in subjective and objective statements. A capability that serves as a useful way to accurately distinguish between content that is opinion-based and that which is fact-based.

Moreover, the use of an encoder BERT-like transformer's CLS token coupled with CNNs has been shown to enhance detection capabilities. This approach leverages the strengths of transformers in contextual understanding and CNN's [3]proficiency to identify patterns, syntactic structures or key phrases, enhancing the model's ability to detect nuances in subjectivity and objectivity, leading to more accurate and refined classifications.

[1] https://www.nature.com/articles/s41562-023-01641-6

[2] https://arxiv.org/pdf/2305.18034

[3] https://ar5iv.labs.arxiv.org/html/2209.06344

The literacy suggests that these model's can be effectively applied to assist individuals in recognizing highly suggestive and potentially biased content, thereby fostering a more informed public, capable of navigating the complexities of modern media.

3 Methods

3.1 Data Preparation

The dataset for this project was derived from the "All the News 2[4]" dataset, a comprehensive collection of news articles from a variety of publishers. The data extraction process focused on articles published between 2016 and 2020. Initial cleansing involved removing duplicate entries and articles lacking metadata such as publication dates and title. To capture a balanced perspective, we selected six prominent US publishers - Fox News and the Hill (right-leaning), CNN and The New York Times (left-leaning), and Reuters and Politico (centrist). This mix is aimed to represent a spectrum of political biases.

To focus on politically charged content, a set of keywords such as 'politic', 'election', 'president', 'congress', and 'government' was used to filter the articles. From the filtered articles, 150 samples per publisher were randomly selected to maintain a balanced dataset. The text from these articles was then segmented into individual sentences, sentences shorter than five words were excluded from training as they generally lacked the necessary context for reliable subjectivity classification

3.2 Data Annotation

The annotation process utilized the ChatGPT API, specifically the 'gpt-4o' model. Sentences were labeled according to the following criteria:

- **Subjective**: Sentences expressing personal opinions, sarcasms, exhortations, discriminatory remarks, or rhetorical figures.
- **Objective**: Sentences presenting facts, third-party opinions cited as such, and other non-emotive content.
- **Unclassifiable**: Sentences that could not be clearly identified as either subjective or objective often due to ambiguous language or insufficient context.

Each sentence was classified three times using the 'gpt-4o' model to ensure consistency and accuracy labeling. In cases where the labels differed across the three iterations, those sentences were flagged and subsequently excluded from the training dataset to maintain further accuracy of classified data. Sentences marked as 'Unclassifiable' were similarly excluded, as they did not provide clear indicators of subjectivity or objectivity. To further validate the annotations, a subset of 100 sentences were manually reviewed to ensure API's classifications were aligned with human judgment.

Our approach consists of two main components: subjectivity classification and objective summarization

3.3 Subjectivity Classification

3.31 Baseline Subjectivity

For the baseline model, we utilized a logistic regression classifier on top of sentence embeddings generated by the 'distilbert-base-uncased' model. This transformer based model, featuring an encoder only architecture with six layers, leverages self-attention and a hidden size of 768 to capture semantic information. DistilBert was chosen due to its efficient design which retains 97% of BERT's performance with significantly fewer parameters and reduced computational overhead.

	#Params	#FLOPS	Latency	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	GLUE
				8.5k	67k	3.7k	5.7k	364k	393k	108k	2.5k	
ELMo-BiLSTM-Attn	-	-	-	33.6	90.4	84.4	72.3	63.1	74.1/74.5	79.8	58.9	70.0
OpenAI GPT	109M	-	-	47.2	93.1	87.7	84.8	70.1	80.7/80.6	87.2	69.1	76.9
$BERT_{BASE}$	109M	22.5B	342 ms	**52.1**	**93.5**	**88.9**	**85.8**	71.2	**84.6/83.4**	90.5	66.4	78.3
$BERT_{BASE}$-6L-PKD*	66.5M	11.3B	-	-	92.0	85.0	-	70.7	81.5/81.0	89.0	65.5	-
$BERT_{BASE}$-4L-PKD†*	52.2M	7.6B	-	24.8	89.4	82.6	79.8	70.2	79.9/79.3	85.1	62.3	-
$BERT_{BASE}$-3L-PKD*	45.3M	5.7B	-	-	87.5	80.7	-	68.1	76.7/76.3	84.7	58.2	-
$DistilBERT_{BASE}$-6L†	62.2M	11.3B	-	-	92.0	85.0		70.7	81.5/81.0	89.0	65.5	-

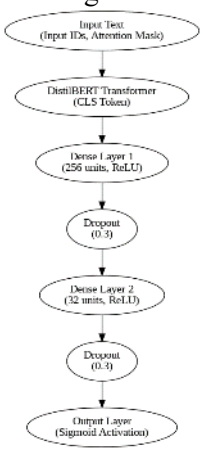
This table compares various models' performance on NLP tasks under the GLUE benchmark. DistilBERT offers a strong balance of performance and efficiency, making it an optimal choice for tasks requiring fast and accurate language understanding

3.32 Fine-tuning Process and CNN Layers

The fine-tuning process started with standard preprocessing, including tokenization using the DistilBERT tokenizer, ensuring alignment with the model's pre-trained vocabulary. Sentences were padded and truncated to a uniform length to match

[4] https://components.one/datasets/all-the-news-2-news-articles-dataset

DistilBERT's input requirements. Convolutional layers were then added to enhance feature extraction, with the first layer featuring 256 filters and the second 32. ReLU activations and max-pooling followed each convolutional layer, accompanied by dropout layers with a 0.3 rate to mitigate overfitting.



Initially, the training focused solely on the CNN layers, keeping DistilBERT's weights frozen. This phase aimed to fine-tune the CNN parameters while preserving the integrity of the pre-trained embeddings, thus minimizing the impact of noise from untrained CNN layers. This strategy ensured the initial training was directed towards improving the convolutional feature extraction capabilities without altering the established embedding space.

Once the CNN layers were adequately calibrated, the DistilBERT layers were unfrozen for further fine-tuning. A lower learning rate was used during this stage to carefully adjust both the transformer and CNN parameters, preventing the forgetting of valuable pre-trained knowledge. This approach enabled the model to effectively leverage both the deep contextual understanding from DistilBERT and the local feature extraction capabilities of the CNN layers, improving its ability to distinguish between subjective and objective content.

A focal loss function was implemented as the main loss function of the mode:

$$\mathrm{FL}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t)$$

where:

- p_t: is the model's predicted probability for the true class
- α is a weighting factor to balance the importance of positive and negative examples
- γ is the focusing parameter, which adjusts the rate at which easy examples are down-weighted, thus focusing learning more on hard, misclassified examples.

A low learning rate of $5\text{x}10^{-5}$ was used during fine-tuning after grid-search iterations using different magnitudes of learning rates. The AdamW optimizer was chosen for effective handling of weight decay. Fine-tuning was performed with a batch size of 16. The training lasted for 4 epochs without stopping early.

3.33 Subjectivity Classification Ranking:

The model assigns a score from 0 to 5 based on the percentage of subjective sentences:

- Score 5 (Extremely Objective): Less than 7.69% subjective sentences, indicating minimal subjectivity.
- Score 4 (Mostly Objective): 7.69% to 15.97% subjective sentences, with occasional subjective elements.
- Score 3 (Moderately Objective): 15.97% to 22.73% subjective sentences, showing a balanced mix.
- Score 2 (Moderately Subjective): 22.73% to 30% subjective sentences, indicating noticeable subjectivity.
- Score 1 (Mostly Subjective): 30% to 40% subjective sentences, with a dominant subjective tone.
- Score 0 (Extremely Subjective): Over 40% subjective sentences, indicating highly opinionated content.

These criteria were established after reviewing the 900 articles to understand the distribution of subjective content. The percentage of the subjective sentences (P) is calculated by dividing the number of subjective sentences (S) by the total number of sentences (N) and multiplying by 100. For example, an article with 120 sentences, where 20 are subjective, results in P = 16.67%, corresponding to a score of 3, labeled 'Moderately Objective'

3.4 Objective Summarization

Building upon the sentence-level subjectivity analysis presented earlier, we now focus on the task of generating objective summaries from news articles. This task is crucial in mitigating bias and providing readers with factual information from potentially subjective reporting. Our approach leverages the sentence-level subjectivity classifications to create summaries that prioritize

objective content. We evaluated both PEGASUS and T5 models for the summarization task.

In our project, we found PEGASUS to outperform their T5 counterpart in summarizing news articles using pre-trained models without additional fine-tuning in the overall F1 score. This observation contrasts with the findings of Goodwin et al. (2020), [5]who evaluated PEGASUS, T5, and BART in zero-shot and few-shot learning settings across multiple datasets. While their study indicates that PEGASUS did not consistently outperform T5, our results highlight the variability of model performance depending on the dataset and specific task requirements.

Our final model is a PEGASUS-based architecture with dynamic output length adjustment. This model demonstrated superior F1 scores across most summary length categories, with improved recall indicating better capture of relevant information.

4 Results and Discussion

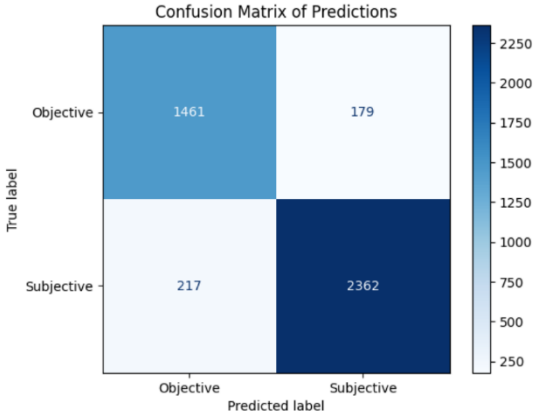
4.1 Subjectivity Classification

The evaluation of the models' performance focuses on comparing the baseline and fine-tuned versions in terms of accuracy, precision, recall, and F1 score. The results demonstrate a noticeable improvement with fine-tuning.

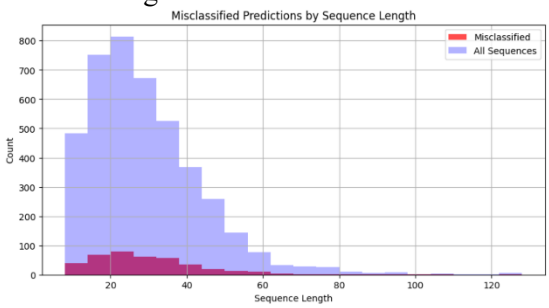
Model	Accuracy	Precision	Recall	F1 Score
Baseline	0.889	0.898	0.923	0.910
Fine-tuned	0.904	0.890	0.962	0.925

These metrics show that the fine-tuned model achieved higher recall and F1 scores, indicating improved sensitivity in identifying subjective content while maintaining a high level of overall accuracy. The use of convolutional layers and the fine-tuning process enhanced the model's ability to capture nuanced differences in text, leading to more precise and reliable classifications. Notably, while the fine-tuned model's precision slightly decreased compared to the baseline, the increase in recall suggests the model became better at correctly identifying subjective instances, which is crucial for minimizing the risk of overlooking biased content.

The results support the conclusion that incorporating additional layers and fine-tuning can significantly enhance the model's performance in distinguishing between subjective and objective content.



The confusion matrix reveals that while the model is effective at identifying subjective content, with a high number of true positives (2362), it occasionally misclassifies subjective sentences as objective (179 cases). This type of error is particularly concerning because it risks misleading readers by presenting opinionated statements as factual, which can be more harmful in disseminating misinformation.



The histogram of misclassified predictions by sequence length, alongside the confusion matrix, highlights areas for improvement. The model tends to struggle with shorter sequences and subtle subjective cues, suggesting a need for refined detection in these cases. Future efforts should focus on enhancing the model's sensitivity to subjective content, especially in short and ambiguous texts, to reduce the likelihood of inaccurately classifying subjective information as objective. This will be crucial in preventing the misinterpretation of subjective content as factual.

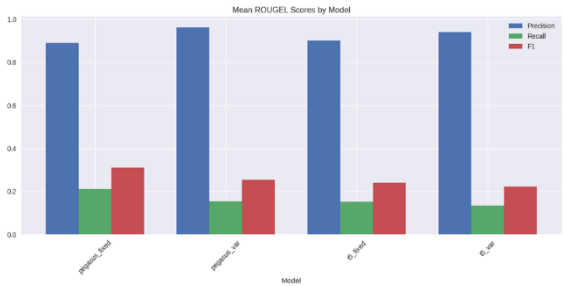
4.2 Objective Summary

Our evaluation of summarization models yielded interesting results, particularly in comparing PEGASUS and T5 models. PEGASUS models consistently outperformed T5 models in overall F1 score. When looking deeper, the precision scores

[5] https://aclanthology.org/2020.coling-main.494.pdf

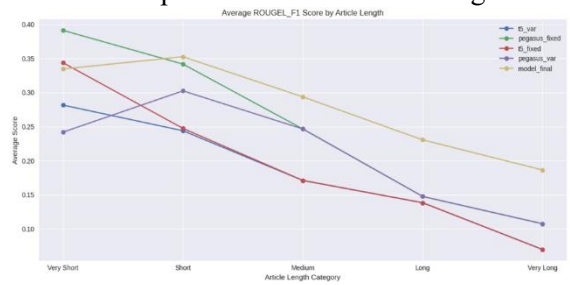
are similar between PEGASUS and T5, but recall scores are marginally higher for PEGASUS (shown below) We suspect this due to PEGASUS model employs a novel pre-training objective, Gap Sentences Generation (GSG), where important sentences are removed from an input document and generated together as one output sequence from the remaining sentences. This approach closely aligns with our task of abstractive summarization, encouraging whole-document understanding and summary-like generation. Furthermore, PEGASUS was pre-trained on a large corpus of news articles [6], which contributes to its effectiveness in summarizing news content by leveraging domain-specific knowledge during the pre-training phase.

Table shows the ROUGE-L metrics for two PEGASUS and two T5 models:



Our final PEGASUS-based model with dynamic output length adjustment showed superior performance across most summary length categories.

Table shows the ROUGE-L F1 for our final model compared to other configurations:



The final model outperformed other configurations in most categories, with particularly strong performance in longer summaries. This suggests that our approach is especially effective for more complex or detailed articles. The model's strength lies in its improved recall scores, which led to higher overall F1 scores despite slightly lower precision compared to some configurations. This balance indicates superior capture of relevant information from the original text, aligning with our goal of producing concise, objective summaries that effectively distill key facts from potentially subjective news articles.

We optimized various generation parameters, including maximum and minimum output lengths, to balance content coverage and brevity. The dynamic length adjustment mechanism adapts to the input article length, allowing for more flexible and context-appropriate summarization. This approach proves particularly valuable for objective summarization, as it ensures comprehensive coverage of important information without introducing excessive irrelevant content.

4.3 Objective Summary

Our algorithm for determining subjectivity scores allowed for a quantitative comparison between original articles and generated summaries. This evaluation demonstrated a consistent reduction in subjectivity in our generated summaries, validating the effectiveness of our two-stage approach.

Subjectivity Score	Before Summary Generation	After Summary Generation
0	356	21
1	133	21
2	80	23
3	83	20
4	96	8
5	149	804

The objective results show a significant increase in the number of articles rated with a subjectivity score of 5 (most objective) after summarization. This change suggests a clear reduction in subjective content, indicating the effectiveness of our summarization approach in producing more objective summaries.

4.4 Limitations and Future Work

Our study revealed several limitations that present opportunities for future research. A key challenge is improving context preservation when removing subjective content, ensuring coherence in the final summary. The model's performance on very short articles also needs enhancement to better handle concise texts without losing critical information. We aim to optimize the precision-recall balance, addressing the current trade-off where higher recall comes at the cost of slightly lower precision. Future work should focus on acquiring larger, more

[6] https://arxiv.org/pdf/1912.08777v3

diverse datasets to improve model training and generalizability. Additionally, exploring the potential gains from fine-tuning summarization models on domain-specific data could yield significant improvements. Extending our approach to multiple languages would broaden its impact on global media literacy. Lastly, conducting user studies to assess the real-world effectiveness of our system in reducing bias interpretation will be crucial for validating its practical utility. Addressing these limitations will enhance the model's performance and applicability across various contexts, furthering our goal of objective news summarization.

5 Conclusion

Our research introduces a novel two-stage approach that combines subjectivity classification and abstractive summarization to reduce bias in news articles. We developed a fine-tuned subjectivity detection model and an improved summarization model with dynamic length adjustment, along with a method for quantifying subjectivity reduction in summaries. This approach shows promise in promoting media literacy, mitigating misinformation, and supporting objective reporting. By effectively distilling factual content from potentially biased sources, our work contributes to fostering a more informed public capable of navigating complex media landscapes.

References

Andrew Thompson. 2020. All the News 2.0: 2.7 Million News Articles and Essays from 27 American Publications. *Components One*. https://components.one/datasets/all-the-news-2-news-articles-dataset. Updated July 9, 2022

Arechar, A.A., Allen, J., Berinsky, A.J. et al. Understanding and combatting misinformation across 16 countries on six continents. Nat Hum Behav 7, 1502–1513 (2023). https://www.nature.com/articles/s41562-023-01641-6

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *arXiv preprint arXiv:1912.08777v3*. https://arxiv.org/pdf/1912.08777.

Charaf Eddine Benarab and Shenglin Gui. 2022. CNN-Trans-Enc: A CNN-Enhanced Transformer-Encoder on Top of Static BERT Representations for Document Classification. *arXiv preprint arXiv:2209.06344v1*. https://ar5iv.labs.arxiv.org/html/2209.06344

Travis R. Goodwin, Max E. Savery, and Dina Demner-Fushman. 2020. Flight of the PEGASUS? Comparing Transformers on Few-shot and Zero-shot Multi-document Abstractive Summarization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5640–5646, Barcelona, Spain (Online). https://creativecommons.org/licenses/by/4.0/

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria. PMLR 119. https://arxiv.org/abs/1912.08777