Enhancing Trust in News: A Subjectivity Classification and Mitigation Approach

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 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 demonstrates the effectiveness of transformers in sentence-level subjectivity detection within news articles. Transformers, with mechanisms for capturing long-range dependencies and contextual relationships, makes 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 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.

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. Methodology:

Dataset Preparation:

The dataset for this project was derived from the “All the News 2” 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 amix 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.

Data Annotation:

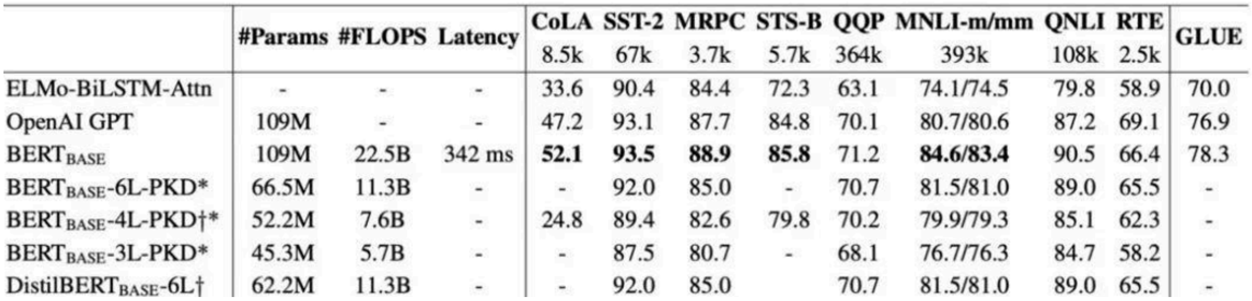
The annotation process utilized the ChatGPT API, specifically the ‘gpt-4o’ model. Sentences were labeled according to the following criteria:

- **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.

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 transformed 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.

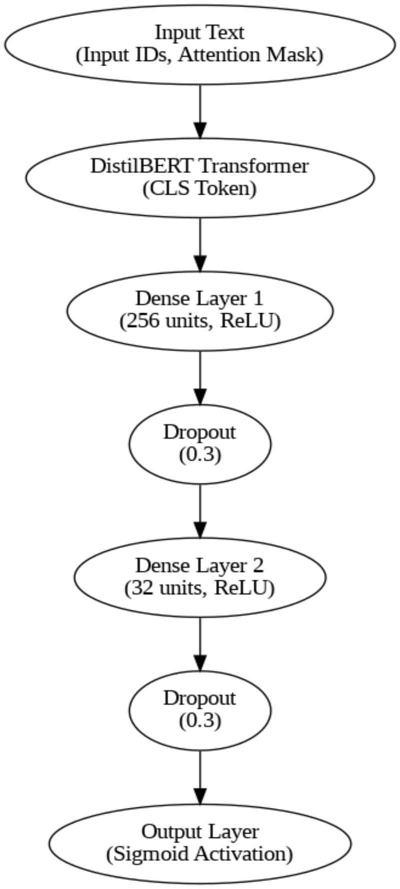


	#Params	#FLOPS	Latency	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m/mm	QNLI	RTE	GLUE
				8.5k	67k	3.7k	5.7k	364k	393k	108k	2.5k	
ELMo-BiLSTM-Attn	-	-	-	33.6	90.4	84.4	72.3	63.1	74.1/74.5	79.8	58.9	70.0
OpenAI GPT	109M	-	-	47.2	93.1	87.7	84.8	70.1	80.7/80.6	87.2	69.1	76.9
BERT$_{\text{BASE}}$	109M	22.5B	342 ms	**52.1**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

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 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 5×10^{-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.

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'