Subjectivity Detection in English News using Large Language Models

Mohammad Shokri
Graduate Center
CUNY

Vivek Sharma
Graduate Center
CUNY

Elena Filatova
City Tech College
CUNY

Shweta Jain
John Jay College
CUNY

Sarah Ita Levitan
Hunter College
CUNY

Abstract

Trust in media has reached a historical low as consumers increasingly doubt the credibility of the news they encounter. This growing skepticism is exacerbated by the prevalence of opinion-driven articles, which can influence readers' beliefs to align with the authors' viewpoints. In response to this trend, this study examines the expression of opinions in news by detecting subjective and objective language. We conduct an analysis of the subjectivity present in various news datasets and evaluate how different language models detect subjectivity and generalize to out-of-distribution data. We also investigate the use of in-context learning (ICL) within large language models (LLMs) and propose a straightforward prompting method that outperforms standard ICL and chain-of-thought (CoT) prompts.

1 Introduction

We live in a world dominated by information where we observe an unprecedented pace of news and opinion propagation. There is an increased demand for fact-checking, as inaccurate stories are disseminated constantly. Opinion pieces and news stories play an important role in shaping individuals' ideologies and beliefs. The rise of subjectivity in news reporting has become increasingly evident in recent years, particularly in online publications (Blake et al., 2019). In addition, fake news and misleading articles often rely heavily on subjective language Jeronimo et al. (2019).

According to estimates, only 41% of publishers categorize their articles by type (e.g., editorial, review, analysis), and among those that do, there is a lack of consistency (Harris, 2017). Opinions are usually conveyed through subjective language and detecting such language accurately is crucial for effective fact-checking. Subjective language includes utterances that communicate emotions, opinions, and beliefs. In addition, many NLP fields (e.g., sentiment analysis) benefit from successfully detecting subjectivity in text.

Subj. Score	Example Sentence
0.93	No punishment could ever be enough for him.
0.55	While what happened to Arthur is rare, the NPSCC has raised concerns about the risks to children during lockdown.
0.45	But while countries from Latin America to Europe are now ordering batches of Sputnik, the rollout in Russia itself has been slow, as people prove deeply reluctant to be injected.
0.04	Jones was found guilty of fatally shooting Mr. Howell, as insurance executive, during a 1999 carjacking on his driveway.

Table 1: Examples of sentences from News-2 with their subjectivity levels: higher subjectivity scores correspond to higher subjectivity level within text.

Most studies focus on identifying subjectivity within three scopes: Document-level, sentence-level, and aspect-level. While document-level and sentence-level tasks differ in the length of their textual input, aspect-based subjectivity analysis aims to identify opinions toward specific aspects in a particular sequence. In this study, we focus on detecting subjective clues in text within sentences. This aligns perfectly with our broader goal of analyzing news articles to identify potential techniques for manipulating readers' interpretations of reported events.

One of the main challenges for learning subjective language arises from the nature of the task. Subjectivity exists on a spectrum, where sentences at the extreme ends are easier to categorize, but as you move towards the center, it becomes increasingly challenging and reliant on personal interpretation to assign a single label due to the nuanced blend of perspectives (see Table 1). In most existing datasets, finding the ground truth on sentence subjectivity is done via majority voting among a group of annotators. However, this could lead to extremely noisy labels due to the low inter-

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 215–226

annotator agreement (Davani et al., 2022). Humans often disagree in their assessment of controversial topics due to a variety of reasons such as socio-demographic factors, political stance, environment, and culture (Luo et al., 2020).

Despite language models' strong performance on various benchmarks, they still lack human-level performance in semantics-related tasks. While fine-tuning a language model on a specific dataset/task could lead to a higher score for that particular dataset/task, it often does not generalize well outside of the training distribution.

Online news articles exhibit a range of writing styles, word choices, and sentence structures. This diversity creates a challenge for model robustness. As large pre-trained language models are trained on huge data collections from a wide range of text distributions, they perform relatively robustly when confronted with different datasets, making them a useful tool for our problem. In this work, we investigate how three different language models detect subjectivity in the news domain and where they fail. Our main research questions are:

RQ1. To what extent does fine-tuning a language model like BERT generalize to out-of-distribution data from the news domain?
RQ2. How well do pre-trained state-of-the-art large language models such as GPT-3.5, GPT-4, and Gemini detect subjectivity in news?
RQ3. How can we improve LLM performance using different prompting methods?

This work contributes empirical studies and insights about the efficacy of language models in detecting subjectivity in news and addressing generalization challenges. We propose and evaluate prompting methods to enhance the performance of LLMs at detecting subjectivity in news.

2 Related Work

Subjectivity Analysis. Various methodologies have been explored for subjectivity analysis. Early work on fine-grained subjectivity detection focused on developing subjectivity lexicons and developing hand-crafted rules to learn subjectivity clues and opinion-bearing terms in sentences (Yu and Hatzivassiloglou, 2003; Gordon et al., 2003; Riloff et al., 2005; Riloff and Wiebe, 2003; Kim and Hovy, 2005). These methods, while simple, often struggle with nuanced expressions and lack generalizability. As machine learning techniques matured, SVMs, Naive Bayes classifiers, and decision tree classifiers emerged as prominent choices. These models leverage features like n-grams, Part-of-speech tags, and syntactic structures for classification, demonstrating improved performance and flexibility (Harb et al., 2008; Goldberg and Zhu, 2006; Zhang et al., 2007). With the advent of deep learning, RNNs and LSTMs gained significant attention due to their capability to capture intricate contextual dependencies in textual data (Irsoy and Cardie, 2014). However, recent advancements in language models and transfer learning reshaped the field. Transfer learning, in particular, allows pre-training models on massive corpora to learn a general representation of words and expressions. Followed by fine-tuning, models can outperform all the previous feature-based and lexicon-based techniques.

In-Context Learning. In-context learning refers to a situation where a frozen language model performs a task by only conditioning on the prompt task. A study by McCann et al. (2018) is a foundational framework for the concept of in-context learning, where multiple NLP tasks are treated as a unified question-answering problem. In addition, the first GPT paper (Radford et al., 2018) paved the way with some tentative prompt-based experiments with the model. However, it was not until GPT-3 (Brown et al., 2020) that the full potential of in-context learning was realized. The seminal GPT-3 paper demonstrates the unprecedented capability of large-scale language models to perform various NLP tasks with minimal task-specific fine-tuning, relying solely on the context provided in the prompt. With the scaling of the model size and data size, large language models demonstrate in-context learning (Dong et al., 2022; Chowdhery et al., 2023). As in-context learning provides interpretable ways for communicating with LLMs, its performance is sensitive to many factors in the prompt, such as the order of examples, length of the examples, and the semantic similarity of the examples to the test set. (Dong et al., 2022; Wang et al., 2023; Zhao et al., 2021; Min et al., 2022). This work evaluates LLMs for subjectivity detection and explores prompting methods for improving generalizability.

3 Datasets

We use multiple datasets to ensure the generalizability of our approach outside of the training domain: MPQA, a classic dataset in the subjectivity

domain Wiebe et al. (2005) and two recently introduced datasets consisting of subjective sentences in the news domain. One of the news datasets (News-1) is focused on political news (Antici et al., 2023); the second dataset (News-2) is focused on crime and COVID-19 Savinova and Del Prado (2023). This diversity in news topics provides a distribution shift within the news domain in our experiments.

MPQA. The MPQA (Multi-Perspective Question Answering) dataset Wiebe et al. (2005) is a significant dataset in sentiment analysis and opinion mining research. This dataset is designed to address the multifaceted nature of subjective language and offers a diverse collection of text segments annotated with sentiment polarity and subjectivity information. It comprises a variety of sources, including news articles, product reviews, discussion forums, and social media posts, and reflects the varied contexts in which subjective expressions manifest. To exclude variability across text genres, we only include the MPQA sentences from news articles in our experiments. We work with MPQA opinion corpus version three. After preprocessing steps and removing sentences with less than 5 words, we are left with 1,707 sentences, 954 subjective and 753 objective.

News-1. We use a recently introduced News dataset (Antici et al., 2023), a collection of subjective and objective sentences extracted from 8 different online political news outlets. This dataset focuses on controversial political topics such as civil rights, politics, law, and economics. We refer to this dataset as "News-1". It consists of 1049 sentences extracted from 23 news articles, out of which 638 are labeled objective and 411 are labeled subjective.

News-2. Our third dataset is collected by Savinova and Del Prado (2023). This dataset contains sentences from news articles and Facebook posts about "crime" and "COVID-19" published by four major UK news sources with a total size of 7,751 sentences. We filter out the Facebook posts since they are shorter and possibly not written by journalists. Hence, all our experiments throughout the paper are carried out using only news sentences with a total count of 2,973 sentences containing 1013 subjective sentences and 1960 objective sentences. An important characteristic of this dataset is that its labels are continuous numbers in the range $[0, 1]$, with 1 being the most subjective and 0 being the most objective. The annotators are instructed to evaluate the sentence subjectivity on a 7-point scale, and they set the mean as the final label. A portion of the dataset is manually labeled and the rest is labeled with the model trained on the manually-labeled set. We refer to this dataset as "News-2" in the rest of this paper. Several examples from the News-2 dataset are presented in Table 1.

4 Methods

4.1 Lexical Features

We first examine the linguistic features that are traditionally used for distinguishing subjective language from objective language. We select lexical features helpful to distinguish the subjective language in news articles from mere news reporting (Krüger et al., 2017) and add 9 lexical richness features (McCarthy and Jarvis, 2007) to form our linguistic features set for this study. The features from Krüger et al. (2017)'s study are claimed to be robust against change in topic and domain and we explore their effectiveness in this our study. We train a logistic regression model with these features to establish our baseline.

4.2 Fine-tuning

First, we study how fine-tuning a language model like BERT helps generalize to out-of-distribution data from the news domain. We fine-tune several popular language models to asses the adaptability of each for our datasets (Section 5.2). A problem often associated with fine-tuning is over-fitting: the model adapts to the training dataset and cannot generalize to out-of-distribution data. However, as the goal of our study is to design a system that can be used in real-time, it is expected to run on data from different distributions than the training data distribution. Hence, we analyze how well a model trained on each dataset generalizes to the other two datasets. We fine-tune a model on each of our datasets and test on the remaining pair as out-of-distribution data (OOD).

4.3 Re-formulating the task

Next, we examine the effect of re-formulating the problem as an entailment task (Section 5.3). As demonstrated by Wang et al. (2021), language models become better few-shot learners as they benefit from transforming the classification problem into a language entailment task. Therefore, we transformed the problem into a language entailment problem. We convert the sentences in all

three of our datasets into pairs of hypotheses and premises and use a RoBERTa-large model for entailment classification. The RoBERTa model is already trained on the MNLI dataset (Williams et al., 2017), so it has learned whether a sentence (hypothesis) entails another sentence (premise). We additionally train it on a small set of the MPQA dataset which has high-quality labels to teach the model the specifics of our task.

4.4 In-Context Learning

In-context learning has become an increasingly popular paradigm for adapting large language models to different tasks (Brown et al., 2020; Kojima et al., 2022). To answer how well the pre-trained state-of-the-art large language models such as GPT-3.5, GPT-4, and Gemini detect subjectivity in the news domain, we examine three large pre-trained language models in both zero-shot and few-shot settings and study how different prompting strategies affect in-context learning performance. We work with Google's Gemini (Team et al., 2023), GPT-3.5-turbo (Brown et al., 2020), and GPT-4 (Bubeck et al., 2023). We access Gemini through Vertex AI API, GPT models through Openai API, and the RoBERTa model through Hugging Face Hub. As few-shot examples in the prompt teach the model the nuances of the task, models demonstrate high sensitivity to the training examples in the prompt. Mitigating this issue requires manual inspection for high-quality relevant examples. Hence, we address research question 3: how could we improve LLM performance using different prompting methods? We examine more general prompting strategies to explain the task and reasoning process to the model without relying on hand-picked examples.

5 Experiments

5.1 Baseline

As our baseline, we use a logistic regression model with 36 linguistic features. We train a logistic regression separately on each dataset and test it on the remaining two datasets. Table 2 presents macro average scores for the logistic regression model across the three datasets. We compute scores for (1) training and testing within the dataset; and (2) using one of the datasets for training and the other two datasets for testing. The Logistic Regression model trained on the MPQA dataset yields the highest score on out-of-distribution data (OOD), exhibiting the highest out-of-distribution generalization. Although the logistic regression model does not achieve high scores, it provides a great deal of interpretability and one can easily figure out what features contributed to the model's predictions. This could be done by analyzing the largest coefficients of the model or by using SHAP values (Lundberg and Lee, 2017) to explain every prediction and quantify the feature contributions.

Result for Dataset	**Logistic Regression (Baseline)**		
	trained on MPQA	trained on News-1	trained on News-2
MPQA	0.54	0.30	0.34
News-1	0.50	0.39	0.44
News-2	0.42	0.48	0.65
OOD Avg	**0.46**	0.39	0.39
	BERT FT		
MPQA	0.86	0.38	0.53
News-1	0.62	0.79	0.65
News-2	0.66	0.65	0.90
OOD Avg	**0.64**	0.51	0.59

Table 2: Classification results for the baseline (Logistic Regression) and BERT FT. For each column, OOD avg is the average of the two rows corresponding to out-of-distribution data.

5.2 Fine-Tuning

We fine-tune several pre-trained language models: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and Llama-2 (Touvron et al., 2023) on each dataset separately to compare performance with zero-shot and few-shot learners. We tune the hyperparameters for each model using grid search and save the best model based on the validation set evaluation. As expected, fine-tuning achieves high F-1 scores for every dataset as the model fully adapts to the training dataset. These results are presented in Table 3. In addition, we run the best model for each dataset on the two other datasets to measure its OOD generalization. Although fine-tuning achieves high F-1 scores for every dataset, its performance drops significantly when tested on OOD data points. Therefore, with the current size and state of available datasets, fine-tuning does not offer a robust solution for classifying subjectivity. Table 2 shows that BERT trained on either of the news datasets has OOD generalization power comparable to the logistic regression model trained on MPQA.

5.3 Reformulating as Entailment

We use the RoBERTa-Large model trained on the MNLI dataset from the Hugging Face Hub. The

Dataset	Baseline	BERT	Llama-2	RoBERTa-L
MPQA	0.54	0.86	0.77	0.82
News-1	0.65	0.79	0.72	0.76
News-2	0.39	0.90	0.69	0.87

Table 3: Fine-tuned models on each dataset. Baseline is a logistic regression model trained on lexical and syntactic features.

model has learned to classify a hypothesis sentence as entailing, contradicting, or neutral towards a premise sentence. We train the model on 20 sentences from the MPQA dataset to further adapt it for the task. For every sentence S_1 in the datasets, we add a premise sentence, S_2 = *"This sentence is Subjective"*. Every *entail* label is translated to subjective, and every *contradict* label is translated to objective. We feed the model $< S_1, S_2 >$ pairs from each dataset. The results are shown in Table 4.

Dataset	Models	
	RoBERTa MNLI*	RoBERTa MNLI trained on MPQA
MPQA	0.43	0.86
News-1	0.38	0.66
News-2	0.36	0.72

Table 4: RoBERTa-MNLI* model has been fine-tuned on 20 examples from MPQA dataset to learn the structure of the task, outputting only 'entail' or 'not entail' without considering 'neutral' for any sentence.

After training on the MPQA dataset, the RoBERTa model performs well on the News-1 and News-2 datasets. Its out-of-distribution (OOD) generalization outperforms the best BERT model fine-tuned on MPQA from Section 5.2 and the logistic regression models' OOD generalization. However, to evaluate its capabilities in a zero-shot setting, we train the model on 20 sentences from the MPQA dataset to further teach it our task. When tested on new data, it does not perform well.

5.4 Zero-Shot Inference

In this section, we describe our experiments with four large language models. In the Zero-shot setting, we prompt the language models to assess the subjectivity of the test sentences without giving them any examples (see Table 9). We use a temperature value of 0 for all our experiments with all three models. We also test the RoBERTa-MNLI model in the zero-shot setting, as explained above in Section 4.3.

As displayed in Table 5, the three large language models vary in their performance across different datasets, but on average across all datasets, GPT-3.5 and GPT-4 score slightly higher than Gemini. Further, compared to the previous sections, the models show more robust performance across all datasets, reducing the gap between best and worst scores.

Dataset	Zero-Shot Models			
	GPT-3.5	GPT-4	Gemini	RoBERTa MNLI
MPQA	0.68	**0.77**	0.62	-
News-1	0.68	0.62	**0.71**	0.38
News-2	**0.78**	0.74	0.73	0.36
Average	0.71	0.71	0.69	0.39

Table 5: LLM's macro f1 score in zero-shot setting on each dataset. As RoBERTa-MNLI is fine-tuned on 20 sentences from MPQA, its score on MPQA test set is not considered under a Zero-shot test setting.

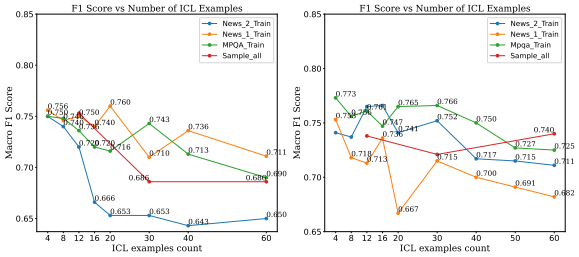

Figure 1: The left graph presents the scores with random seed set to 42; the right graph presents the scores with random seed set to 100.

5.5 Few-Shot In Context Learning

In Section 5.4, we showed that LLMs perform well and robustly in zero-shot settings without seeing any examples from the target text distribution. To answer RQ3, we investigate whether different sampling strategies for in-context learning examples can increase performance. We experiment with varying factors in the prompts and evaluate the impact of each factor on performance. The variants include the number of ICL examples in the prompt, the random seed for sampling sentences from the data, the subjective-to-objective ratio in ICL examples in the prompt, and the dataset from which we draw the ICL examples to account for in-distribution and OOD sentences. In Table 6, we report the average macro F1 scores over five experiments for each set of variants.

Count of ICL Examples. The first factor we study is the count of example sentences in the prompt. As previously proven in supervised ma-

chine learning, more labeled data could lead to better performance. However, this does not seem to be true with ICL examples (Min et al., 2022). We use a fixed random seed for sampling sentences from our datasets. We test the model with prompts containing {4, 8, 12, 16, 20, 30, 40, 50, 60} ICL examples. For each test, we add new sentences to the previously existing ones; for example, the 8-ICL examples prompt adds 4 new sentences to the 4-ICL examples prompt, and so on. Similar to Min et al. (2022)'s findings, we do not see any clear increasing trend in performance with a higher number of examples (Figure 1). Moreover, the best performance is achieved with fewer than 20 labeled examples. Figure 1 shows that the choice of random seed can substantially affect performance.

Input Data Distribution. An intuitive assumption is that using in-distribution-data in the prompt should help the model conditioning on the input reach better performance. However, unlike to Min et al. (2022), we observe that in many cases, sampling from OOD data outperforms a prompt with in-distribution training examples. This finding aligns with the rest of their findings, in that the model learns more information about the task and the input-output structure than the data itself. In addition, their work shows that assigning random labels to input sentences does not hurt performance, suggesting that the model does not learn substantial information about the data. Furthermore, we observe that sampling equally from all three datasets performs competitively in $k = 12$ ICL examples, however, we can not hypothesize more generally due to the limited number of experiments.

Subjective to Objective Ratio. To learn the effect of the majority labels on the performance, we set up several experiments where we changed the subjective-objective ratio in ICL examples. Unlike (Zhao et al., 2021) we do not observe a strong correlation between majority labels and the model's predictions. As shown in Figure 2, two out of three of our experiments suggest that increasing the subjective-objective ratio in training examples marginally hurts the performance.

5.6 Chain of Thought Prompting

Due to the instability and unpredictability of standard few-shot in context learning, we switch to *Chain of Thought* prompting (Wei et al., 2022) expecting higher performance and stability. Standard few-shot prompting has shown promising results in many tasks, except for the tasks that require reasoning. With Chain of Thought (CoT) prompting, we break down the task into smaller steps, which the model is more likely to solve, hence teaching the model to reason about the task the same as how humans do. This method requires manual task-aware examples curated by experts in the prompt. However, it has been demonstrated Kojima et al. (2022) that one can bypass that step and require the model to think step by step and achieve competitive results with standard CoT prompts. This approach, called *Zero-shot Chain of Thought*, is task-agnostic and comparatively simple to implement. We extend this method by adding instructions for classifying our sentences in Figure 3. We do not provide any examples for the model but explain a general framework for classifying sentences as subjective or objective based on the annotation scheme done by Wiebe et al. (2005). We refer to this prompting method by *ZCoT-Inst* in the rest of this paper. Table 6 depicts our results for each prompting strategy. In all three models, ZCoT-Inst leads to best average performance across all datasets. We also observed that the biggest gain of standard CoT prompting happens for the MPQA dataset, which might be due to the reason that our chain of thought instructions aligns well with MPQA's annotation procedure.

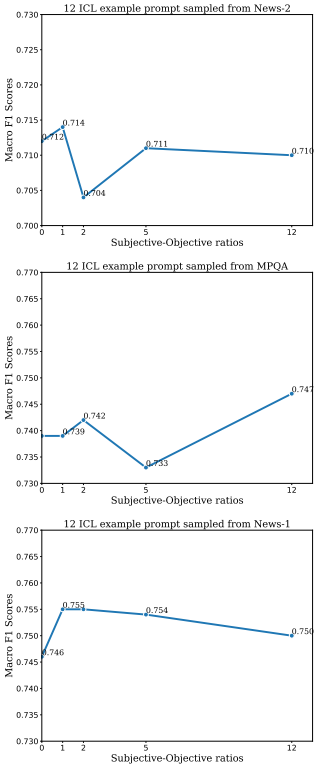

Figure 2: Changing the ratio of subjective to objective sentences in the 12-example prompt.

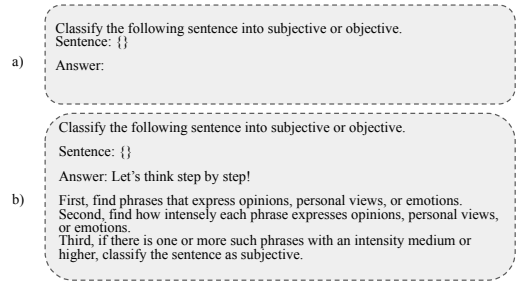

Figure 3: (a) Zero-Shot prompt (b) Zero-Shot CoT prompt with manual instructions.

The results in Table 6 indicate that the Chain of Thought prompting does not work well on GPT-3.5 as it scores higher with standard zero-shot prompts (see Table 5. This could be because reasoning abilities increase with model size. Comparing Table 5 and Table 6 shows that even though the highest score for News-1 and News-2 is achieved with zero-shot prompting, the highest score for MPQA is achieved with GPT-4 with ZCoT-Inst prompt. The highest average score across all three datasets is achieved by GPT-4 with ZCot-Inst prompt, which gains a 2.3% increase compared to the Zero-shot setting. This demonstrates the efficacy of our method and effectively addresses RQ3.

5.7 Ensemble Model

Next, we explore an ensemble of our three large language models. We feed every sentence to each of our models while prompting them to just output the final label. For each sentence, we get three predictions by the models and we use the majority vote as the final verdict. We test the ensemble model in a Zero-shot setting(all three models are given the same zero-shot prompt) and with a CoT prompt(all three models are given the same CoT prompt). In addition, we also run an ensemble model of the three prompting strategies (Zero-shot, ZCoT-Inst, and Zero-shot CoT) and we refer to it by all-prompts. All-prompts setting is an ensemble of 9 different predictions (3 prompt settings for each model) and we use the majority vote as the final verdict. Table 7 summarizes the results. The ensemble model with a Zero-shot prompt achieves the highest scores on News-1 and News-2 datasets among all the non-fine-tuned models in our experiments. The All-prompts setting achieves the highest average score of all our experiments. However, it is less practical than the other settings because it captures each model's predictions under three different prompting settings.

Dataset	Prompting Methods		
	Zero-Shot CoT	ZCoT-Inst	Standard CoT
MPQA	0.70	0.69	**0.76**
News-1	0.65	**0.67**	0.59
News-2	**0.73**	**0.73**	0.68
Average	0.693	**0.696**	0.676

(a) GPT-3.5 average macro F-1 scores over 3 runs.

Dataset	Prompting Methods		
	Zero-Shot CoT	ZCoT-Inst	Standard CoT
MPQA	0.76	**0.80**	0.75
News-1	0.66	**0.67**	0.55
News-2	**0.77**	0.73	0.66
Average	0.73	**0.733**	0.653

(b) GPT-4 average macro F-1 scores over 3 runs.

Dataset	Prompting Methods		
	Zero-Shot CoT	ZCoT-Inst	Standard CoT
MPQA	0.67	0.73	**0.76**
News-1	0.69	**0.72**	0.60
News-2	0.73	**0.74**	0.70
Average	0.696	**0.73**	0.686

(c) Gemini average macro F-1 scores over 3 runs.

Table 6: Comparison of (a) GPT-3.5 and (b) GPT-4 and (c) Gemini average macro F-1 scores on different datasets.

Dataset	Ensemble Model			
	Zero-Shot	Zero-Shot CoT	ZCoT-Inst	All-prompts
MPQA	0.70	0.75	0.75	**0.76**
News-2	**0.80**	0.75	0.76	0.78
News-1	**0.72**	0.67	**0.72**	0.70
Average	0.74	0.723	0.743	**0.746**

Table 7: Ensemble model performance on three datasets.

6 Error Analysis

In this section, we analyze false negatives and false positives predicted by the best model from the previous section. As discussed in Section 5.6, *ZCoT-Inst* outperforms all the other prompting techniques across all models. Therefore, the analysis in this section is with regard to the models' prediction in that setting. We look at the predictions by the models for test sets in each dataset. There are 220 sentences in the MPQA test set, 219 sentences in the News-1 test set, and 298 sentences in the News-2 test set.

Table 8 summarizes the classification results of the models on each of our datasets. GPT-4 gener-

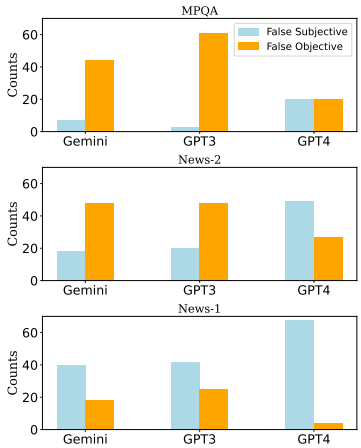

Figure 4: False objective and false subjective prediction counts by each model.

Model	MPQA		News-1		News-2	
	Subjective	Objective	Subjective	Objective	Subjective	Objective
	P / R	P / R	P / R	P / R	P / R	P / R
GPT-3.5	0.94 / 0.50	0.60 / 0.94	0.67 / 0.77	0.72 / 0.59	0.74 / 0.53	0.78 / 0.90
GPT-4	0.83 / 0.84	0.79 / 0.77	0.62 / 0.96	0.90 / 0.36	0.61 / 0.74	0.84 / 0.74
Gemini	0.93 / 0.51	0.60 / 0.94	0.70 / 0.83	0.78 / 0.60	0.76 / 0.54	0.78 / 0.91

Table 8: Precision (P) and Recall (R) for Different models and datasets.

ally has a higher recall for subjective class compared to the other models. Gemini and GPT-3.5 exhibit similar behavior across all three datasets with generally higher precision and lower recall in subjective class compared to GPT-4. These differences might justify the advantage of the ensemble model as compared to the individual models. Figure 4 demonstrates the counts of false subjective (sentences annotated as '*objective*'), and false objective (sentences annotated as '*subjective*'), instances across each model within every dataset.

Sentiment. We analyze the misclassified sentences and assess positive and negative sentiment patterns across datasets using the RoBERTa-based sentiment analysis model trained on tweets (Loureiro et al., 2022). We aim to inspect if the models struggle with sentences carrying strong sentiments (positive or negative), which intuitively should be easier to identify. First, we run the sentiment analysis model on every sentence in all three of our datasets, to understand their sentiment distribution. Figure 5 summarizes the information. In general, subjective sentences in all three datasets, range from high positive to high negative sentiment with more than half of the instances carrying neutral or negative sentiment. This is the case for objective sentences in both News-1 and News-2 datasets, whereas MPQA's objective sentences mostly contain neutral sentences which could speak for the distribution shift among the datasets.

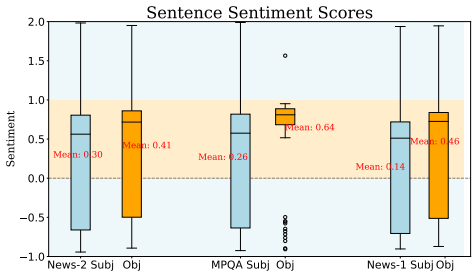

Figure 5: Sentiment scores across subjective and objective sentences in each dataset. [-1,0) in the y axis represents negative sentiment, [0-1] represents neutral sentiment, and(1,2] represents positive sentiment.

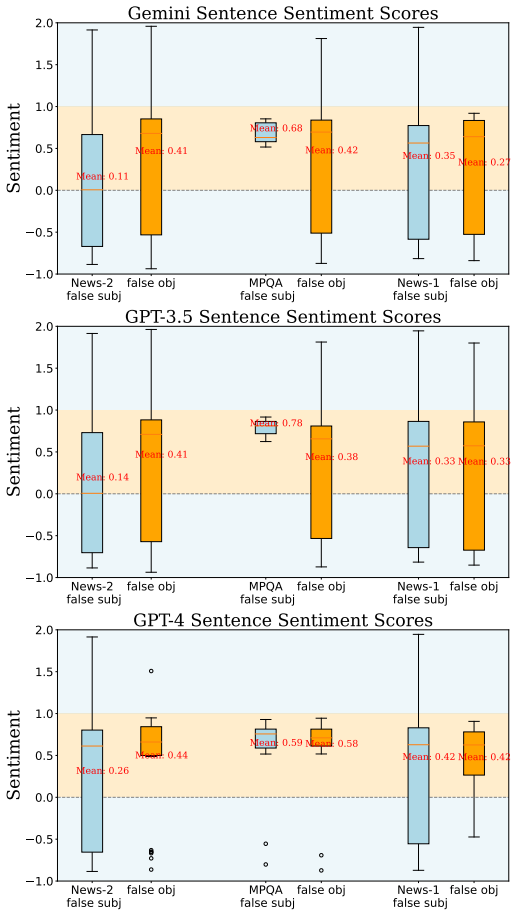

Figure 6: Different models sentiment score across false subjective and false objective sets.

Next, we proceed to examine the sentiment scores of false subjective and false objective sentences predicted by all three models. Similar to the previous analyses for general classification reports of the models, Gemini and GPT-3.5 exhibit very

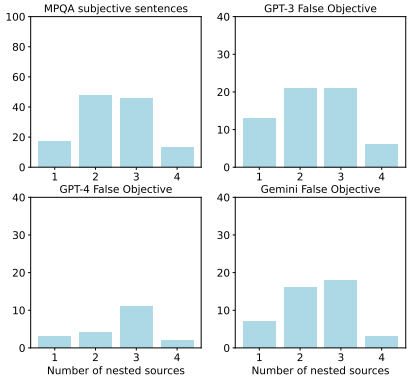

Figure 7: False objective instances in each model versus subjective sentences in MPQA dataset.co

similar behavior in their false predictions across all three datasets. Their false predictions only slightly differ in mean sentiment scores, whereas GPT-4's false predictions tend to have higher mean sentiment scores across all datasets. GPT-4's false subjective set seems to be shifted more towards neutral sentiment as its having higher scores for mean and median. However, GPT-4's false objective prediction sets are quite different than other models in terms of sentiment. In contrast to Gemini and GPT-3.5, it is noticeable that for all three datasets, GPT-4's false objective predictions mainly lay in the neutral zone of the sentiment graph. This might signal that a model can detect easier subjective signs, such as high-sentiment words, and fails to detect subjectivity in neutral sentences. Another reason could be due to the inductive bias of the prompt (Figure 3) that relies on the sentiment of extracted subjective terms.

Opinion Holders. According to our initial examination of false objective sentences, none of the instances have obvious, significant clues of subjectivity. Therefore, classifying these sentences correctly requires identifying nested opinion holders in them. We examine the misclassified sentences of each model on the MPQA dataset, as MPQA has fine-grained annotations for subjective terms and their nested sources (opinion holders) in every sentence. The source of a subjective frame is defined as the person or entity that is expressing the opinion. Consider the following example from Wiebe et al. (2005) work on annotating subjective texts:

> "China criticized the U.S. report's criticism of China's human rights record."

In the sentence above, the U.S. report's criticism is the target of China's criticism. Thus, the nested source for *criticism* is <writer, China, U.S. report>, as writer of the text is a default source of subjectivity in all written texts. Hence, the sentence above has 4 nested sources. Figure 7 summarizes our findings: both Gemini and GPT-3 fail in adhering to the original distribution of nested opinion holders. However, GPT-4 diverges from this trend, primarily failing in statements containing three nested opinion holders.

7 Conclusion

In this work, we investigate how language models learn and classify subjective language across three different datasets from the news domains. We examine how well different models generalize to out-of-distribution data. In addition, we analyze how LLMs detect subjective language with different prompts. Based on our experiments, we conclude that the standard in-context learning does not guarantee robust classification as it introduces a great deal of sensitivity to the examples provided in the prompt. In future work, we plan to investigate how different prompting techniques, such as explaining how to detect potentially subjective terms and analyzing sentiment intensity, can lead to better, more robust performance across different datasets.

Limitations

There are several algorithms for domain adaptation when the source and target data distributions are known, such as sample re-weighting. There also exist algorithms for cases when the target distribution is unknown, usually referred to as domain-generalization. In our study we mainly focused on fine-tuning and did not explore domain generalization algorithms for our smaller models.

Acknowledgement

We would like to express our sincere gratitude to the Rosen Center for Advanced Computing(RCAC), Purdue University Laboratory for the computational resources, Anvil. Their lab was supported by the National Science Foundation (NSF) under grant [2005632]. This work was supported by an unrestricted gift from Google through the Google CyberNYC Initiative.

References

Francesco Antici, Andrea Galassi, Federico Ruggeri, Katerina Korre, Arianna Muti, Alessandra Bardi, Alice Fedotova, and Alberto Barrón-Cedeño. 2023. A corpus for sentence-level subjectivity detection on english news articles. *arXiv preprint arXiv:2305.18034*.

Jonathan S Blake et al. 2019. *News in a digital age: Comparing the presentation of news information over time and across media platforms*. Rand Corporation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Andrew B Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of TextGraphs: The first workshop on graph based methods for natural language processing*, pages 45–52.

Andrew Gordon, Abe Kazemzadeh, Anish Nair, and Milena Petrova. 2003. Recognizing expressions of commonsense psychology in english text. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 208–215.

Ali Harb, Michel Plantié, Gerard Dray, Mathieu Roche, François Trousset, and Pascal Poncelet. 2008. Web opinion mining: How to extract opinions from blogs? In *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, pages 211–217.

Laurie Beth Harris. 2017. Helping readers tell the difference between news and opinion: 7 good questions with duke reporters' lab's rebecca iannucci.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 720–728.

Caio Libanio Melo Jeronimo, Leandro Balby Marinho, Claudio EC Campelo, Adriano Veloso, and Allan Sales da Costa Melo. 2019. Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, pages 15–24.

Soo-Min Kim and Eduard Hovy. 2005. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Katarina R Krüger, Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687–707.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Philip M McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112.

Ellen Riloff, Janyce Wiebe, and William Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *AAAI*, pages 1106–1111.

Elena Savinova and Fermin Moscoso Del Prado. 2023. Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 305–314.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*, page 3.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136.

Wei Zhang, Clement Yu, and Weiyi Meng. 2007. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 831–840.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Additional Experimental Details

A.1 Prompts

Here we provide more details about the prompts used in our experiments to make models predict the subjectivity of our test datasets. Table 9 presents all four prompts that we used in our experiments. As seen in the table, models prompted with standard CoT and Zero-shot CoT with instructions, generate longer answers required post-processing to extract the final label. To extract the final answer from the longer answers, we feed the answer to a Gemini model and prompt it to extract the final label from the answer.

Prompt Name	Prompt Text
Zero-shot	Classify the following sentence into Subjective or Objective. Just output the label. Sentence: {} Label:
Zero-shot CoT	Classify the following sentences into Subjective or Objective. Let's think step by step. Sentence: {} Label:
ZCoT-Inst	Classify the following sentence into Subjective or Objective. output reasoning for each step. Sentence: {} Answer: Lets think step by step! First, find phrases that might express opinions or personal views. Second, find out how intense each phrase is expressing opinions or personal views. Third, if there is one or more phrases with expression intensity medium or above, classify the sentence as Subjective.
Standard CoT	Classify the Sentence into Subjective or Objective. Sentence: Meanwhile, some other countries, including Japan and Germany, already issued statements on Bush's new climate change policy in rather different tones. Answer: First, the phrases that might express opinions or personal views are 'rather different'. The expression intensity of the phrase is medium. Since there is one or more phrases with expression intensity medium or above, classify the sentence as Subjective. Sentence: {} Answer:

Table 9: Different prompt used in our experiments. We only include one example for standard CoT for demonstration purposes, but our experiments are done with 6 examples in standard CoT prompt setting.

A.2 Logistic Regression Features

Here we list the features used for training a logistic regression model as our baseline. The features were taken from the work by Krüger et al. (2017). The set of features are claimed to be robust for classifying opinion vs news report.

Feature	Description
SentLength	sentence length measured in tokens (inverted)
TokenLength	Avg. token length measured in characters (inverted)
Negation	Norm. frequency of lemmatized negation words
NegationSuffix	Norm. frequency of negation suffix *n't*
Complexity	Norm. frequency of finite verbs per sentence
Questions	Ratio of question marks
Exclamations	Ratio of exclamation marks
Commas	Ratio of commas
Semicolons	Ratio of semicolons
Temporal Conn.	Ratio of temporal connectives
Causal Conn.	Ratio of causal connectives
Contrastive Conn.	Ratio of contrastive connectives
Expansive Conn.	Ratio of expansive connectives
Citations	Ratio of citations
CitationLength	Avg. number of tokens per citation
Past	Ratio of past tense outside quotes
Present	Ratio of present tense outside quotes
VoS	Ratio of lemmatised communication verbs outside quotes
Modals	Ratio of lemmatised modal verbs outside quotes
Future: Will	Ratio of verb 'will' outside quotes
1st person	Norm. frequency of 1st person pronouns outside quotes
2nd person	Norm. frequency of 2nd person pronouns outside quotes
1st/2nd person	Norm. frequency of 1st and 2nd person pronouns outside quotes
Digits	Norm. frequency of digits
Interjections	Norm. frequency of interjections
Sentiment	Norm. text polarity outside quotes
Sentiment Adj	Norm. text polarity outside quotes in adjectives only

Table 10: Features and Descriptions

We supplement the above list of features with 9 lexical richness features from *lexicalrichness* python library. These form the set of 36 features that we use to train a logistic regression model.