# Classifying Online Misinformation with Cognitively-Informed Features from Transformer Language Models

Samuel Hutchinson

Columbia University Senior Thesis Project in Cognitive Science

Presented to the Program in Cognitive Science May 2023

Advisor: Dr. Christopher Baldassano

**Table of Contents**

**Abstract**

Online misinformation has crept into the public consciousness through discourse on topics ranging from presidential elections to the COVID-19 pandemic. Misinformation-containing statements spread both farther and faster than true statements on social networks, suggesting the need for an interpretable, algorithmic flagging mechanism. This investigation endeavors to devise such an algorithm based on cognitively-plausible explanations for misinformation's virality: informational novelty and emotional valence. I use next-word prediction error measures from a GPT-2 model fine-tuned on true news stories from Reuters to assess the novelty component of this problem and a pre-trained RoBERTa-based sentiment classifier for the emotional-valence component. To create a classification model, I calculated the joint distribution of these errors and sentiments over a subset of true news stories and misinformation-containing stories from the ISOT Fake News dataset. I then used this joint distribution to predict the likelihood that unseen news stories contain misinformation. This model classifies news stories with around 79.4% accuracy, furthering prior work showing similarity in next-word prediction between human readers and generative models like GPT-2. These results also indicate that online misinformation may be classifiable through computable and cognitively-interpretable natural-language metrics.

## *I. Introduction*

"They're lying to you," announced the poster-board sign interrupting coverage from First Channel, a Kremlin-backed outlet and one of the main sources of news for Russian citizens. The protestor—who had briefly appeared onscreen to suggest that the Russian state media was producing disinformation regarding the war in Ukraine—was arrested within seconds (Krause-Jackson, 2022). The COVID-19 pandemic proved to be another testing ground for the impact of online misinformation on our daily lives, with recent studies linking exposure to misinformation with higher rates of vaccine hesitancy across both Democrats and Republicans (Pierri et al., 2022; Loomba et al., 2022; Lee et al., 2022). With such high-profile examples in the media, the saliency and immediacy of mis- and disinformation spread online has quickly entered the international consciousness.

In their particularly alarming study, Vosoughi et al. (2018) find that tweets containing false information are over 70 percent more likely to be retweeted—spreading "significantly farther, faster, deeper, and more broadly"—than tweets containing true information. As of 2018, the authors observed three main spikes in the total number of false tweets: the 2012 and 2016 American presidential elections and the 2014 Russian annexation of Crimea, suggesting that false political information is particularly salient. What could explain this difference in virality between true and false information? Can we use data collected about misinformation to model its linguistic qualities? One place to start along this line of inquiry would be to consider the ways in which we already know misinformation-containing statements differ from those that do not contain misinformation. Responses to false statements online indicate that misinformation-containing statements disproportionately inspire reactions of "surprise" or "shock" and "anger" or "disgust" in comparison to true statements' inspiring of "sadness" and

"trust" (Vosoughi et al., 2018; Pennycook & Rand, 2021). This narrows the scope of our previous questions: from modeling misinformation, can we instead model statements that may inspire surprise or anger?

This investigation will endeavor to answer a part of this question through analyzing some of the ways in which misinformation is differentially communicated online. Aligned with the research above, I hypothesize that misinformation is communicated with a greater, more negative emotional valence and is more surprising to readers, more often violating their expectations of what words will be used. Further, I predict that these two factors are computationally measurable and the two classes of statements will vary significantly in these measures. Finally, I predict that this difference will allow for the construction of a model that accurately predicts whether a previously-unseen statement contains misinformation.

Such a result would prove both novel and significant for several reasons. As was revealed in leaked Facebook internal documents, human content moderators at large social media companies are struggling to stem the growing tide of misinformation on their platforms (Seetharaman et al., 2021). This highlights the need for alternative methods of flagging posts that potentially contain misinformation. Several studies (see Khan et al., 2021; Islam et al., 2020; Wang, 2017 for reviews) demonstrate that a variety of machine learning models perform well-above chance levels at identifying statements that contain misinformation, but the models are limited in explicability; many models are essentially black boxes, lacking clear justifications for each classification (but see Shu et al., 2019). This investigation attempts to leverage the power and speed of algorithmic classification while maintaining explicability: instead of the traditional machine-learning strategy of text-classification with labels only, the classifications

here will be done based on the pre-defined, cognitively-informed features of negative sentiment and surprise (as measured by next-word prediction error).

Recent advances in natural language processing (NLP) and large language models (LLMs) allow for this computational analysis of sentiment and surprise. Sentiment classifiers—machine-learning algorithms designed to detect the emotional content of natural-language statements—are a well-established aspect of NLP research today. This opens one algorithmic door for this investigation: do sentiment classifiers rate statements containing misinformation as significantly more negative than others, thereby echoing the subjective responses of anger or disgust? As for the surprise element of misinformation, Goldstein et al. (2021) find—in ways which I will unpack later in this paper—that LLMs make similar predictions as humans in next-word prediction tasks. Of paramount importance to this investigation, however, they also find that LLMs and humans overlap in their confidence ratings and predictability judgements (calculated as cross-entropy, see *Methods* section below) regarding the words that do appear next in these tasks. These similarities suggest a possibility that LLMs could reliably identify statements that humans find surprising—thereby potentially providing the second piece to our algorithmic puzzle.

## *II. Background*

### *A. LLMs*

As a foundation, let us begin by explaining what LLMs are and how they work. This section will start with a brief introduction to two key ideas in modern language modeling—neural networks and distributional semantics—and how those ideas contribute to the

development of particular transformer models like GPT-2 (Radford et al., 2019) and RoBERTa (Devlin et al., 2018; Liu et al., 2019).

### *1. Neural Networks*

Since the inception of the field now-known as 'Artificial Intelligence,' researchers in computer science and cognitive science have endeavored to find algorithmic accounts for the complex behavior exhibited by human beings. Much of the early research in this direction focused on 'symbolic' models, essentially programs that could—to a very limited degree—emulate some human behaviors through concatenating many pre-defined 'if-then' statements familiar to symbolic logic and computer science. One of these symbolic models relevant to the current investigation was Weizenbaum's (1966) ELIZA, a program that could converse with a user based on the presence of keywords in the user's input text. A clear limitation of these symbolic models, however, was their rigidity—in the absence of those certain keywords, a program like ELIZA loses its appearance as 'intelligent.' Naturally, scientists began to look for other solutions, some of which had been under development alongside symbolic models.

The key breakthrough in this new direction—'Connectionist' models—came with Rosenblatt's (1958) Perceptron model. Instead of operating by strict 'if-then' rules, the Perceptron gradually adjusts its output through a pre-defined system of operations over several iterations of input data. A quintessential problem that a Perceptron model is able to approximate is linear classification: given two classes of data ('blue dots' coded as 0s and 'red dots' coded as 1s, for example), what is the optimal straight line that divides the data cleanly into these two classes? Such a line would take something of the form:

$$y = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

Where the *x*'s represent the input data of each dot and the *w*'s are the 'weights', all of these products summing to the output *y*. This output *y* then undergoes an 'activation function,' which transforms the continuous output to a discrete one—either 0 or 1. Keeping with our above example, the problem now is: given a 'blue dot' represented by a vector of *x*'s, what are the appropriate weights that will transform that input data in such a way that the output becomes 0? In a different light, this process is a matrix multiplication:

$$X^T \times W = \begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_n \end{bmatrix} \times \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = [y]$$

The Perceptron starts with a random weights vector, and changes these weights depending on the accuracy of the output prediction after seeing each input example. After many trials, the weights vector is such that the resulting line separates the classes of dots as well as it can, minimizing a pre-defined loss function.

For more complicated outputs, such as confidence in *n* possible outputs represented by an $n \times 1$ vector, many of these perceptrons can be linked together—the output of each serving as an input to the next layer. The resulting matrix multiplication is thus able to capture much more nuance than the simple perceptron; these large models are known as neural networks and have become the dominant paradigm in artificial intelligence and machine learning research in recent years (Sejnowski, 2019).

### 2. *Distributional Semantics*

Stepping aside from computer science for a moment, let us examine a linguistic theory that intersects with the above discussion at the key junction of LLMs: distributional semantics. The central idea behind this theory of word meanings is well-phrased by Harris (1970):

> If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

Rephrasing the above in a way that will be useful for us later, a sufficiently-detailed accounting of the different contexts in which certain words appear comes to correlate with the differences in meanings those words may have from each other (what qualifies as 'sufficiently-detailed' is ambiguous, and we will see later that the size of this context-window has important implications for the current investigation). Such a claim necessitates empirical backing, however—can theories of distributional semantics account for human linguistic behavior?

Researchers vary substantially in their views on the degree to which the distributional hypothesis is a plausible cognitive or psychological account of word-meaning representation. Several experiments (e.g. McDonald & Ramscar, 2001; Mandera et al., 2017) find positive evidence that contextual variation and word co-occurrence impact semantic similarity judgements, semantic priming, and other related tasks. Additionally, humans appear to be extremely attuned to rates of word co-occurrence, with these rates reflected in adult gaze duration in reading tasks and infant gaze duration when presented with sequences of frequently and infrequently co-occuring words (Smith & Levy, 2013; Skarabela et al., 2021). On the other hand, Lake & Murphy (2021) critique interpretations of these results as being indicative of a distributional-semantic cognitive model, instead emphasizing the importance of embodiment and

multimodal knowledge on learning word meanings. However, Grand et al. (2022) more recently show that certain manipulations on the representations learned from word co-occurrence reveal more nuanced similarity judgements—such as similarity in 'size' as independent of similarity in 'dangerousness'—than previous experiments, perhaps countering Lake & Murphy's claim that multimodal knowledge is necessary for making these more nuanced judgements.

Regardless of the cognitive plausibility of distributional models, this understanding of a connection between word-context and word-meaning is what gives modern LLMs their predictive power. Recalling the dot-labeling example above, consider a different kind of classification problem: given a word (or sequence of words)—represented as a numerical vector, an *embedding*—in a certain text, what weights should be applied in the multiplication process such that the resulting vector represents the *next* word in the text? Scaling this process up to the level of larger and larger matrices, answering this question results in a weight-matrix that, given an input embedding, predicts the next word in the sequence solely based on the *context* in which that input has appeared before—the distributional hypothesis embodied algorithmically.

### *3. Transformers*

The combination of the above two ideas hinted at in the previous paragraph has proven fruitful ground for years of research in NLP, spawning a variety of model architectures (see Lake & Murphy, 2021 for a review). One of these architectures—the transformer (Vaswani et al., 2017)—has become a primary focus for NLP in recent years and is the variety employed by this investigation, so it is worth explaining here in some detail.

These models are trained, essentially, on countless fill-in-the-blank tasks from real-world sources, such as website pages and digitized books. Given a sentence like "I am thirsty, so I am

looking for a water _____" a model would then output a vector with a particular numerical value for each token (not quite a word, but analogous), which then can be transformed into a probability distribution over the next possible word. Suppose our model-in-training assigns a higher probability to the token "cat" than any other token, and thereby guesses that "cat" fills in the blank above. Given that the actual token was "fountain," our model's guess would be noted as incorrect and the weights-matrix adjusted accordingly. Over many, many trials of such guesses, we would hope that the resulting weights-matrix would result in more accurate predictions. In short, each guessing attempt would look something like:

$$P(w_i|w_{i-1}, w_{i-2}, \cdots, w_0)$$

For each word $w_i$ and its preceding context. However, if the preceding context is very long (a book, for example), such a calculation quickly becomes intractable. This is where the size of the context-window comes into play, as well as a key innovation of the transformer model: self-attention.

First, a bit of background and vocabulary is necessary. Recurrent neural networks (RNNs) made strides forward in NLP through allowing each next-word guess to be informed not only by the embedding of the previous word, but also by the hidden states (columns of the weights-matrix) computed in the previous guess (Tunstall et al., 2022). Therefore, as an RNN processes a string of text, contextual information about previous words and their relationship to each other gets passed along, allowing for a more nuanced prediction than if each guess was informed solely by the average embedding of previous words (which would be a bit like asking someone to predict the next word in a scrambled sentence). However, taking each of the previous steps into account when computing a next step again reaches an information bottleneck. Which steps are more important than others? The advent of *attention* for RNNs provides a way around

this bottleneck, as the models learn relative weights for each processing step along the way (Tunstall et al., 2022).

Another breakthrough came in both speed and performance when transformer models—eschewing the 'recurrent' aspect of RNNs—entered the scene with a modified version of the attention mechanism: *self-attention*. Self-attention transforms the initial token embeddings into weighted averages of all the embeddings in an input sequence—essentially pre-computing how relevant each token in the input is for the interpretation of the other tokens in the string (Tunstall et al., 2022). Such a process makes the task of disambiguating homonyms like "bank" much easier for algorithms to handle.

This line of research, from neural networks and distributional semantics to attention and self-attention, led to the meteoric rise in transformer LLMs like GPT and BERT. Both of these models (*Generative Pre-Trained Transformers* and *Bidirectional Encoder Representations from Transformers*, respectively) learn their weights-matrices through the kinds of repetitive fill-in-the-blank tasks as described above, although with a slight difference between them. BERT—as implied by the *bidirectional* aspect of its nomenclature—is trained on masked sequences like: "I am thirsty, so I am _____ for a water fountain." GPT, on the other hand, guesses the words in the string sequentially from left to right, more like the initial example above. GPT thus exhibits a closer proximity to human reading or listening behavior and displays significant similarities to human judgements in these next-word tasks in several experiments (Caucheteux et al., 2023; Goldstein et al., 2022; Wilcox et al., 2020; Golan et al., 2022; Schrimpf et al., 2021).

### *B. Misinformation*

Keeping these developments in mind, let us now turn to the cognitive science at play in exposure to misinformation, or language that lacks coherence with the actual state of the world. As briefly mentioned in the introduction, this paper will focus primarily on *political* misinformation. Additionally, there has been a flurry of recent studies exploring the psychology and cognitive science of misinformation and its viral effects due to a plethora of factors (see DeAngelis, 2023: the American Psychological Association's *Trends Report* on misinformation); the current investigation will focus on only two: novelty and negative emotional valence.

One especially important aspect of misinformation and how it is processed is its novelty. Recall Vosoughi et al. (2018)'s finding that tweets containing misinformation are 70 percent more likely to be retweeted than those that do not contain misinformation. By their very nature, statements containing misinformation will trend towards novelty from a cognitive perspective: we expect continuity with what we observe in the world around us, any deviation from that continuity—any deviation from that coherence—will register as a prediction error, as something novel. This novelty is part of what contributes to false tweets' vitality: novelty is attention-grabbing (consider visual oddball tasks, or the gaze-duration-versus-word-frequency tasks in Smith & Levy, 2013; Skarabela et al., 2021). Accordingly, Vosoughi et al. (2018)—in analyzing tweets in response to viral, false tweets—find that responses to misinformation are characterized by an expression of surprise.

Another key aspect in which misinformation infiltrates the information ecosystem is through its appeal to strong emotions. In concert with the effects of novelty, Brady et al. (2020) find that more emotionally-laden tweets draw earlier visual attention and garner more retweets than less-emotional posts. Vosoughi et al. (2018) and Brady et al. (2017) have identified the

increased emotional valence of viral misinformation through sentiment analysis of tweets in response to misinformation-containing tweets and noting the increased virality of tweets containing words categorized as "moral-emotional language." Additionally, Martel et al. (2020) find that readers' emotional state *before* they read the misinformation statement contributed to their belief—with more heightened emotions correlating with higher degrees of belief in false information—revealing a destructive spiral: users are exposed to large amounts of emotionally salient content, encouraging a heightened emotional state, and this valence as well as the inevitable repetition only serves to increase their belief in the content's truthfulness.

## *III. Methods*

### *A. Dataset*

A good place to start with a data-driven exploration of misinformation is with an appropriate dataset. I have chosen the ISOT Fake News Dataset compiled by Ahmed et al. (2018) due to its large number of news stories (44,898), focus on political stories, clear division into true and false news stories (with the true news stories all published by Reuters and the false news stories from sources rated untrustworthy by Politifact), and balanced coverage of news stories from both within the United States and around the world as well as from both left- and right-leaning sources. See Figure 1 (next page) for a detailed description of how I utilized the dataset in the various stages of this experiment, to be elaborated upon in the rest of this section.

### *B. GPT-2*

I used two GPT-2 (Radford et al., 2019) models in this experiment: one pretrained (downloaded from HuggingFace with no additional fine-tuning) and one that I fine-tuned on a

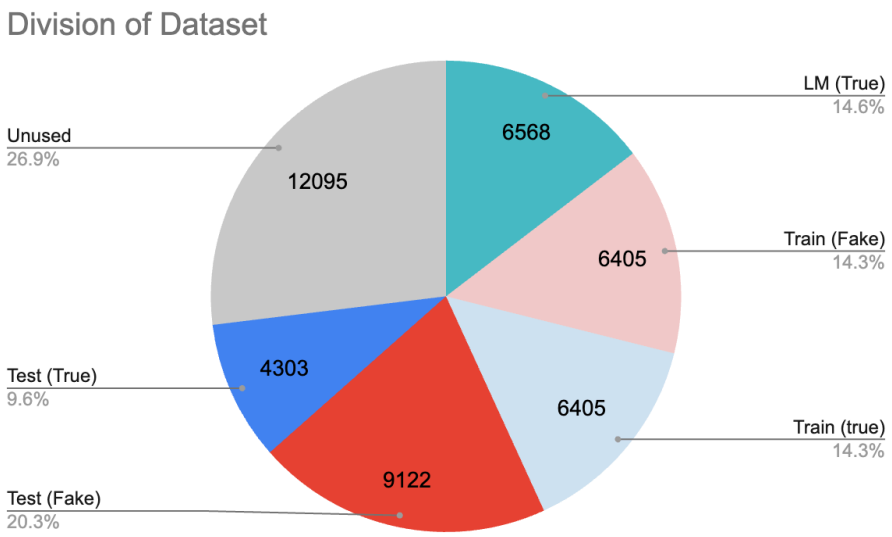



**Fig. 1**: Clockwise from top: LM (True): these stories were used to fine-tune the GPT-2 model used in calculating cross-entropy; Train (Fake): these stories were used to build distributions of cross-entropy and emotional content for false news stories; Train (True): same as previous, but distributions for true news stories; Test (Fake): these stories are sampled from to test the conditional Bayesian predictions from the generated distributions; Test (True): same as previous; Unused: these are fake stories that would otherwise bias the training data—there was a large excess of fake stories due to the use of true stories in the LM dataset—or stories where only the title and label were provided in the dataset, not the body of the text.

subset of the true news stories in the dataset. I chose GPT-2 over larger or more recent models due to the cognitive-scientific literature that explicitly compared GPT-2 and human next-word-prediction (Caucheteux et al., 2023; Goldstein et al., 2022; Wilcox et al., 2020; Golan et al., 2022; Schrimpf et al., 2021). Other model architectures have not been tested in similar contexts, or have been found to be less aligned with human judgements than GPT-2.

The fine-tuning process entails repeating the next-word-prediction task on new examples particular to the experimental circumstance: in this case, that means adjusting the weights-matrix so as to make more accurate predictions on true news stories. I chose to use a fine-tuned model to 1) generally test the effects of fine-tuning on news stories on model behavior and 2) mimic next-word-predictions of an ‘informed consumer’ in news stories.

The crucial measurement extracted from the predictions of both GPT-2 models was next-word prediction loss, or *cross-entropy*, an information-theoretic measure of the difference between two probability distributions. Following Goldstein et al. (2022), the two distributions measured here were the models' predicted probability distribution over the next word and a one-hot 'distribution' that assigns the *actual* next word a probability of 1 and all other next-words a probability of 0. The cross-entropy *H* of a particular prediction is calculated as follows:

$$H(p, q) = -\sum_{i=0}^{n} p(x_i) \cdot \log q(x_i)$$

For a predicted distribution *q* and one-hot distribution *p* over *n* possible next-words in the vocabulary ($x_i$). Below (see Figure 2, next page) is a toy example (with actual GPT-2 outputs) of a *q* (blue) distribution and *p* (red) 'distribution' on the water fountain example from earlier, showing the top ten most probable next tokens. Given the one-hot nature of the *p* 'distribution,' the above cross-entropy calculation is equivalent to the simpler:

$$H(p, q) = -\log q(w)$$

Where *q(w)* is the predicted probability of the actual next word *w*. This is the more familiar equation for surprisal, which is a term I will avoid using so as to not conflate the notion with the aforementioned aspect of subjective surprise. For whole news stories, an average cross-entropy was calculated by taking the mean of the cross-entropies (surprisals) of each prediction-word distribution pair across the story.

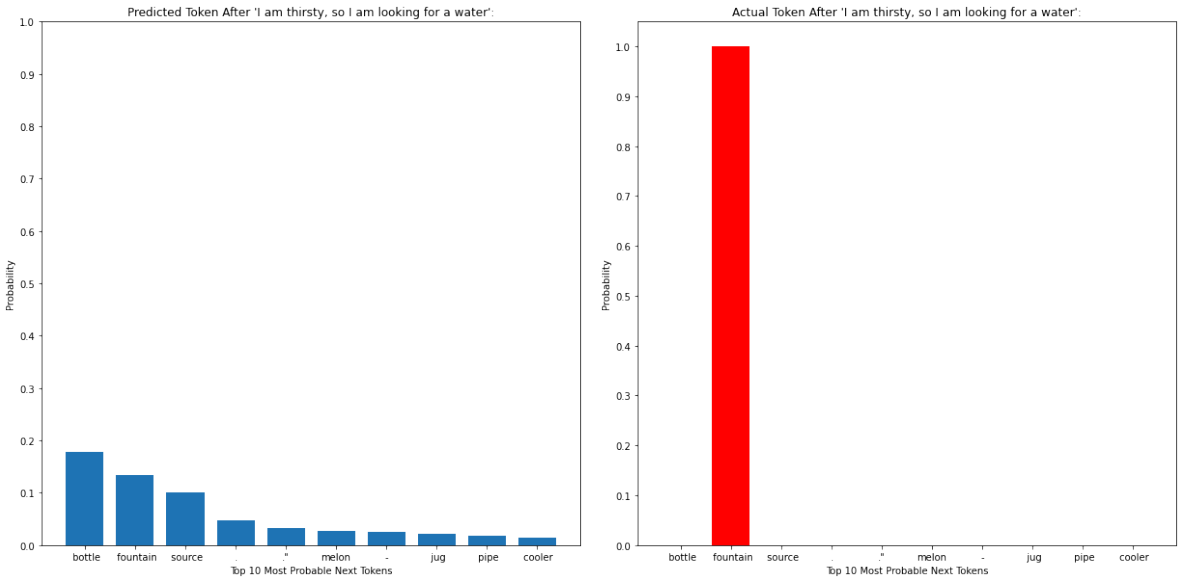



**Fig. 2**: Left: the ten tokens GPT-2 assigns the highest probability to following the input string; Right: the one-hot distribution representing the actual next token

### *C. RoBERTa*

Thus far, we have focused on the next-word-prediction capabilities of LLMs like GPT and BERT. However, Radford et al. (2017) introduced the idea that LLMs may harbor representations in their weights-matrices that reliably predict *other* meaningful features of the input text, namely the sentiment content of that text. This indicated that LLMs are successful with certain tasks of transfer-learning, or applying the same weights used in pre-training to an entirely new task. Barbieri et al. (2020)—taking advantage of this transfer-learning success—fine-tuned a RoBERTa model (Liu et al., 2019), a modified version of BERT, to classify the sentiment of tweets into positive, neutral, or negative. I used Barbieri et al. (2020)'s sentiment classifier model to analyze the sentiment of the news statements in this investigation, as many of the misinformation studies cited above focused on the characteristics of misinformation posted specifically on Twitter.

### *D. Building the Distributions*

As indicated in Figure 1, I took a subset of the news stories and calculated 1) the average-cross entropy of the pre-trained model, 2) the average cross-entropy of the fine-tuned model, and 3) the sentiment classification from the RoBERTa model for each story. For the sentiment classification, I transformed the discrete classification output into a continuous value by assigning each label an integer value (-1 for negative, 0 for neutral, 1 for positive), multiplying each integer label by the classifier's confidence in that label, and summing these products. Thus, instead of the model simply outputting 'neutral,' for example, it may output a value of -0.3 for a neutral-leaning-negative sentiment or a 0.3 for a neutral-leaning-positive sentiment.

After calculating these values for all of the training examples, I then created probability density functions to represent the distribution of these values by story veracity (see *Results* section below for plots).

### *E. Joint classifier*

Taking sentiment distribution as an example, the following conditional probabilities for a particular sentiment score *S* can be read from the density functions for true stories (*T*) and false stories (*F*):

$$P(S|T)$$

$$P(S|F)$$

Which can then be flipped (via Bayes rule) to the following:

$$P(T|S) \propto P(S|T) \cdot P(T)$$

$$P(F|S) \propto P(S|F) \cdot P(F)$$

These conditional probabilities, then, appear as a prediction of whether or not a given story is true or false based on the evidence from the story's sentiment value and prior values of the rate at which true and false stories appear.

However, as discussed in the *Background* section above, misinformation appears to not only vary on its emotional valence, but also how surprising it is to readers. I am accounting for this 'surprising' nature of misinformation with the average cross-entropy (CE) of the story, yielding the following joint conditional probabilities:

$$P(T|S, CE) \propto P(S, CE|T) \cdot P(T)$$

$$P(F|S, CE) \propto P(S, CE|F) \cdot P(F)$$

I then used *these* conditional probabilities to predict the category of unseen stories during the testing phase, with the story classified according to which probability comes out greater. I set the priors to the actual rate of true and false stories appearing in the testing dataset, roughly two-thirds false and one-third true.

### *F. Procedure*

Putting all of the above together, the experimental procedure proceeded as follows: 1) divide the dataset into language model, training, and testing partitions; 2) fine-tune a GPT-2 model on the language model partition using the Hugging Face *Trainer* class; 3) extract average cross-entropy and sentiment scores for each news story in the training partition; 4) calculate the distributions of average cross-entropy and sentiment for true and false news stories; 5) extract average cross-entropy and sentiment scores for each news story in the testing partition; 6) predict which category the test-set stories belong to through the Bayesian method outlined above; 7)

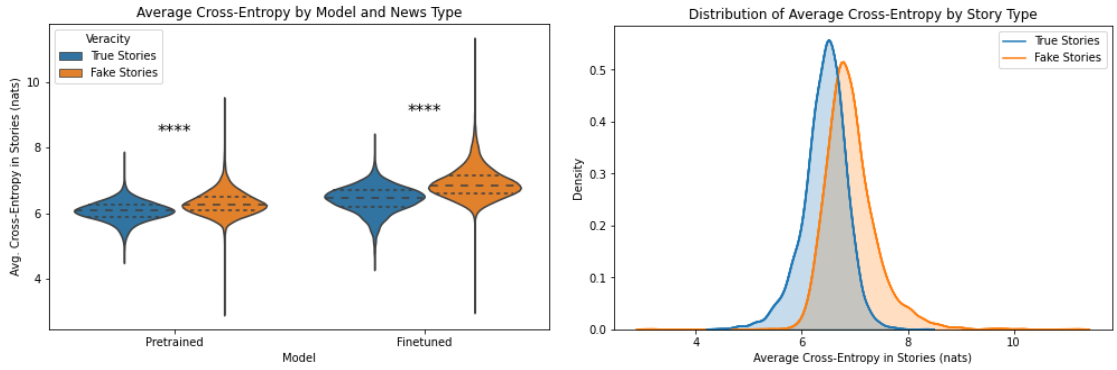
repeat steps 1-6 twice more with different dataset partitions; 8) aggregate results across each repetition.

## *IV. Results*

This investigation endeavored to discern any differences in two cognitively-plausible metrics between 'true' and 'fake' news stories: average cross-entropy in the stories and continuous sentiment score (labels multiplied by confidence and summed) for the stories. For calculating the average cross-entropy, two models were used: one pre-trained GPT-2 model downloaded from Hugging Face, and one GPT-2 model fine-tuned on a subset of true news stories. The differences in these distributions were then used to define a Bayesian classifier that would predict the label (veracity) of news stories in a held-out test set. The results shown in this section primarily describe the first of three cross-validations; see *Appendix* for more details of the other two cross-validations.

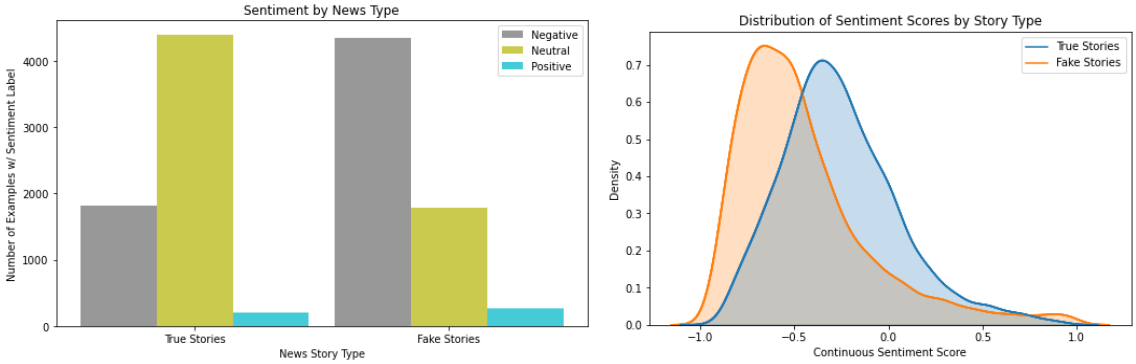
### *A. Differences in Cross-Entropy and Sentiment*

Both the pre-trained and fine-tuned GPT-2 models show significant differences in average cross-entropy between true and fake news stories (pre-trained: two-tailed independent t-test, $t = 42.03$, $p < 0.0001$; fine-tuned: two-tailed independent t-test, $t = 62.98$, $p < 0.0001$). However, the fine-tuned model shows a greater difference in average cross-entropy than the pre-trained model between the two news types (see Figure 3, next page). In both cases, the fake news stories exhibited higher average-cross entropies than the true news stories, aligning with the initial hypothesis.




| Model | Veracity | N | Mean | StDev |
|---|---|---|---|---|
| Pretrained | True | 6405 | 6.067 | 0.298 |
| Pretrained | False | 6405 | 6.316 | 0.368 |
| Finetuned | True | 6405 | 6.431 | 0.410 |
| Finetuned | False | 6405 | 6.927 | 0.478 |

**Fig 3**: Top Left: distributions of average-cross entropy by model and news type; Top Right: density distributions of average cross-entropy values for true and fake news stories as calculated by the fine-tuned model; Bottom: summary statistics for cross-entropy across model and news type




| Veracity | N | Negative | Neutral | Positive | Mean | StDev |
|---|---|---|---|---|---|---|
| True | 6405 | 1813 | 4394 | 198 | -0.257 | 0.313 |
| False | 6405 | 4354 | 1784 | 267 | -0.453 | 0.356 |

**Fig 4**: Top Left: counts of categorical sentiment label by news story type; Top Right: density distributions of continuous sentiment scores (calculated by summing the 'number' of each label (-1, 0, or 1) multiplied by the confidence of the sentiment model for that label); Bottom: exact counts and summary statistics for sentiment scores by news type

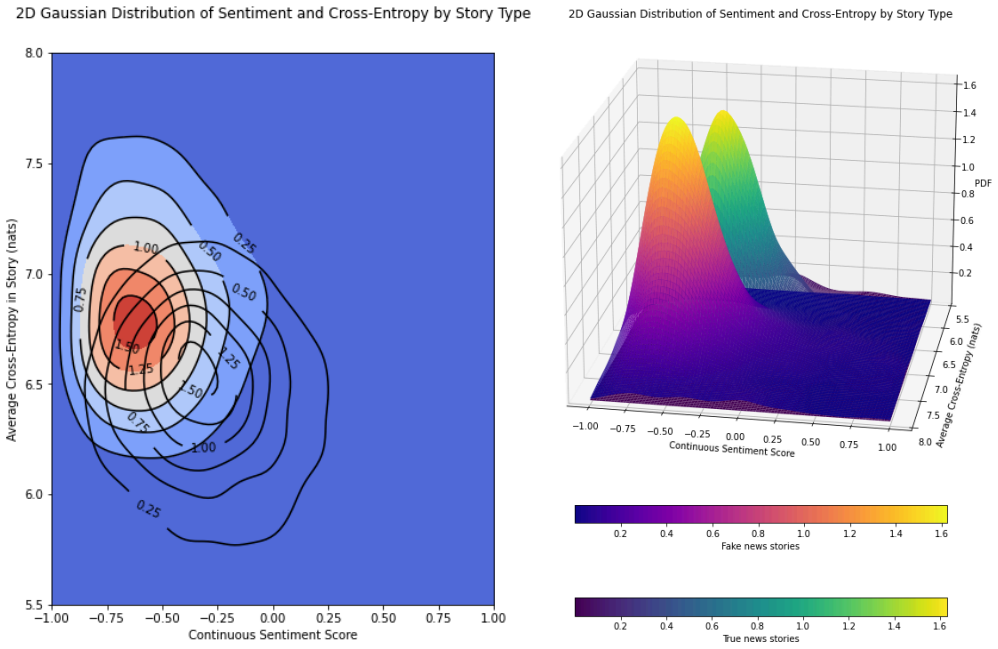



**Fig 5**: Left: joint distribution of continuous sentiment score and average cross-entropy by news type (the filled region represents fake news stories, while the empty region represents true news stories), the numbers around the rings represent the value of the probability density function for data in that region; Right: a three-dimensional view of the same plot, with the ring-numbers now represented by height along the z-axis

I tested for differences in sentiment score between the two news types in two ways (see Figure 4, previous page). First, I compared the counts of categorical sentiment labels by news type and found a significant difference in that regard (two-way $\chi^2$-test, $\chi^2 = 2159$, $p < 0.0001$). Second, I compared the distribution of *continuous* sentiment score (calculated by summing the 'number' of each label (-1, 0, or 1) multiplied by the confidence of the sentiment model for that label) by news type, and again found a significant difference (two-way independent t-test; $t = 33.02$; $p < 0.0001$). This test revealed that the sentiment of true news stories is more positive on average than fake news stories, while both skew towards the negative. This result again aligns with the initial hypothesis.

Given that these two metrics vary significantly according to news veracity, what may their joint distribution look like? Plotting the density of stories by category with continuous

sentiment score on the x-axis and average cross-entropy on the y-axis yields the left-most plot in Figure 5 (previous page). The fake news stories show a distinct cluster from the true news stories, with the fake cluster being higher-entropy and lower-sentiment than the true cluster. Both distributions are unimodal with approximately orthogonal axes. The values of these joint density distributions served as the likelihood factors in the Baeysian classifications.

### *B. Classification*

Do these distinct clusters facilitate predicting which category an unseen news story may fall in based solely on its average cross-entropy and continuous sentiment score? I tested three different classification methods to answer this question: one based solely on sentiment (Figure 4 Top Right), one based solely on cross-entropy (Figure 3 Top Right), and one using the joint distribution of sentiment and cross-entropy (Figure 5 Left). All cross-validations of all three methods exceed 66% accuracy (chance levels), with average accuracies of 69.8%, 76.8%, and 79.4%, respectively (see Tables 1-3, next two pages).

To evaluate the statistical robustness of these results, I performed permutation tests across all three joint-distribution cross-validations by randomly shuffling the labels and recalculating the distributions 1,000 times. I then tested each cross-validation's fine-tuned model using the respective set of new distributions. I then calculated the p-values as the fraction of these 1,000 shuffled-distribution testing runs that exceeded the accuracy of the initial un-shuffled testing. All of these p-values were less than 0.001, providing further evidence that true and fake news differ in their joint distributions of cross-entropy and sentiment and that these differences facilitate accurate class prediction.

| | **Average** | CV1 | CV2 | CV3 |
|---|---|---|---|---|
| Accuracy | **0.698** | 0.697 | 0.698 | 0.698 |
| True Positive Rate | **0.867** | 0.850 | 0.866 | 0.885 |
| True Negative Rate | **0.334** | 0.374 | 0.333 | 0.296 |
| False Positive Rate | **0.666** | 0.626 | 0.667 | 0.704 |
| False Negative Rate | **0.133** | 0.150 | 0.134 | 0.115 |

| | CV1 | | CV2 | | CV3 | |
|---|---|---|---|---|---|---|
| | Predicted Fake | Predicted True | Predicted Fake | Predicted True | Predicted Fake | Predicted True |
| Actual Fake | 7750 | 1372 | 7952 | 1226 | 8098 | 1057 |
| Actual True | 2693 | 1610 | 2832 | 1415 | 3004 | 1266 |

**Table 1:** Results from sentiment-only classification across all three cross-validations of the dataset; Top: accuracy and error measures; Bottom: confusion matrices

| | **Average** | CV1 | CV2 | CV3 |
|---|---|---|---|---|
| Accuracy | **0.768** | 0.752 | 0.772 | 0.781 |
| True Positive Rate | **0.910** | 0.926 | 0.910 | 0.894 |
| True Negative Rate | **0.465** | 0.383 | 0.474 | 0.539 |
| False Positive Rate | **0.535** | 0.617 | 0.526 | 0.461 |
| False Negative Rate | **0.090** | 0.074 | 0.090 | 0.106 |

| | CV1 | | CV2 | | CV3 | |
|---|---|---|---|---|---|---|
| | Predicted Fake | Predicted True | Predicted Fake | Predicted True | Predicted Fake | Predicted True |
| Actual Fake | 8445 | 667 | 8350 | 828 | 8180 | 975 |
| Actual True | 2657 | 1646 | 2236 | 2011 | 1967 | 2303 |

**Table 2:** Results from cross-entropy-only classification across all three cross-validations of the dataset; Top: accuracy and error measures; Bottom: confusion matrices

| | **Average** | CV1 | CV2 | CV3 |
|---|---|---|---|---|
| Accuracy | **0.794** | 0.782 | 0.795 | 0.805 |
| True Positive Rate | **0.893** | 0.895 | 0.892 | 0.891 |
| True Negative Rate | **0.583** | 0.542 | 0.587 | 0.619 |
| False Positive Rate | **0.417** | 0.458 | 0.413 | 0.381 |
| False Negative Rate | **0.107** | 0.105 | 0.108 | 0.109 |

| | CV1 | | CV2 | | CV3 | |
|---|---|---|---|---|---|---|
| | Predicted Fake | Predicted True | Predicted Fake | Predicted True | Predicted Fake | Predicted True |
| Actual Fake | 8163 | 959 | 8183 | 995 | 8160 | 995 |
| Actual True | 1970 | 2333 | 1756 | 2491 | 1626 | 2644 |

**Table 3:** Results from joint classification across all three cross-validations of the dataset; Top: accuracy and error measures; Bottom: confusion matrices

### *C. Exploratory Testing and Results*

The above results indicate that there is a significant difference in average cross-entropy and sentiment score between true and fake news stories in the ISOT Fake News Dataset, but to what extent does this result generalize to other datasets or news types?

To attempt to answer this question, I tested the method outlined above on the LIAR dataset (Wang, 2017). Like the ISOT Fake News dataset, the LIAR dataset contains examples composed of sentences from both true and fake news stories, with the labels again drawn from Politifact's human rating of the source. However, instead of a binary distinction between true and fake news, the LIAR dataset is broken down into six categories ranging from totally false to totally true; additionally, the examples are considerably shorter in the LIAR dataset compared to the ISOT.

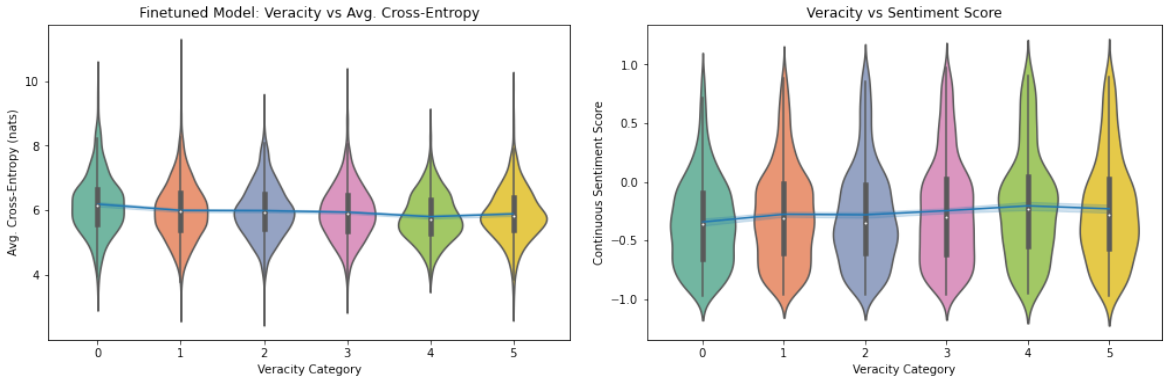



**Fig 6**: Left: distributions of average cross-entropy by news story veracity category in the LIAR dataset; Right: distributions of continuous sentiment score by veracity category

These differences limited the extent to which I could test the *classification* aspect of the original procedure, but I could still test for relationships between news story veracity on the one hand and cross-entropy or sentiment on the other. If the results above are indeed generalizable, then one would expect a negative relationship between higher veracity categories and average cross-entropy (as more-truthful stories will be less surprising), and a positive relationship between category and sentiment score (as more-truthful stories will be more neutral).

These hypotheses proved somewhat correct across the LIAR dataset. There was a slight negative correlation between increasing veracity category and average cross-entropy in a model fine-tuned on news stories of categories 0 and 1 (Spearman's rank correlation, $\rho = -0.092$, $p < 0.0001$) and a slight positive correlation between category and continuous sentiment score (Spearman's rank correlation, $\rho = 0.061$, $p < 0.0001$) (see figure 6, above). However, given how small these correlations are across this dataset (as opposed to the more discernible differences in the main dataset), it remains unclear *the extent to which* misinformation-containing news stories differ from their more truthful counterparts in terms of average cross-entropy and sentiment in general. This could be one fruitful direction for extensions of the current investigation.

## *V. Discussion*

The above results support the initial hypotheses that 1) generative language models have a significantly higher next-word-prediction error while parsing misinformation-containing news stories compared to more truthful stories and 2) misinformation-containing stories have a generally more negative sentiment than truthful stories, which tend to be more neutral in sentiment. Furthermore, these differences facilitate the classification of held-out news stories from the ISOT Fake News Dataset with a Bayesian classifier based on this prediction error (average cross-entropy) and sentiment score, and this classifier correctly identifies the label of these unseen stories with an accuracy of 79.4%, significantly above chance levels.

With these results in-hand, we can now revisit the theoretical concerns mentioned at the outset of this investigation. First, however, I must carve a distinction between two concepts I have thus far conflated: the truth of a statement and the way in which that statement is communicated. The results of this investigation do *not* imply that a statement's material correspondence with a state of the world can be accurately understood in terms of its average cross-entropy or sentiment; rather, we can conclude that when writing about things that are widely considered to be false, people tend to write in a way that can be characterized as more surprising and more negatively-valenced. Consider the breaking news of Queen Elizabeth's death: those stories would be written in such a way as to express surprise and negative sentiment, but the event itself did indeed happen in the world. However, if one finds their Twitter or Facebook feeds flooded with news stories of expiring monarchs, that may be a reason to think twice about at least some of those stories.

One conclusion we can draw from these results is a confirmation of earlier findings: the findings outlined above confirm the results that misinformation-containing news stories are both

more surprising and negatively-valenced than their truthful counterparts. This result takes on a cognitive dimension if we consider the complementary findings that humans and LLMs are generally aligned in their 'surprise' at a given word in-context. This investigation mutually confirms the findings of both strands of research while adding a new dimension: that one reason humans may rate misinformation as more surprising is the unpredictability of the words used in those news stories.

What could be underlying this linguistic unpredictability of misinformation? For LLMs, the answer is straightforward: they are trained on an enormous volume of text from the internet, most of which is generally aligned with the true state of the world (Wikipedia, for example). Thus, the kinds of words LLMs predict in a sequence reflect this training data's composition of generally true statements. Recall the distinction above, which may now appear somewhat more clear: LLM surprise does not reflect a conflict with the truth, but rather a conflict with the way true statements are generally written. This is confirmed by the greater difference in average cross-entropy in the fine-tuned model, as that model has re-weighted its parameters to reflect not only the kinds of things generally written on the internet, but more specifically the way *true news stories* are written. The elevated surprise in *fake* news stories follows naturally from this fine-tuning. Another possibility regarding LLM surprise could be that the fake news stories have a higher prevalence of syntactically malformed sentences, which would raise prediction-error rates. However, as these stories are meant for human consumption and appear generally well-formed in the dataset, I am making the assumption here that these syntactic errors play a relatively small part in the average cross-entropy calculations compared to the semantic violations discussed above.

The answer for what makes misinformation more surprising to humans is less straightforward. One possibility is that of a typical notion of evaluating the truth of a statement:

we consult our representation of the believed state of the world after reading or hearing a statement and then assess whether that statement and our representation are logically compatible. If they are, then we do not have to adjust this representation very much, and little surprise is registered (as is the case with most everyday statements). However, if there is a mis-match between the statement and our representation, then this registers to us as surprising (again, because most everyday statements do not cause such conflict). While some people certainly hold skewed or biased representations of the world that *do* align with the kinds of statements presented in misinformation-containing news stories, these viewpoints are generally not held by the majority. Thus, group-level results show an elevated surprise reaction to fake news stories.

Another possible interpretation of human surprise is closer to that of the LLM explanation. The results of human alignment with LLM surprise can be taken as indicating some similarity in the computational processes of LLM prediction and human language cognition; specifically that next-word prediction plays a large-enough role in human language cognition as to allow LLM hidden-state activation to reliably predict human fMRI activation when reading the same pieces of text (Schrimpf et al., 2021). Additionally, Goldstein et al. (2022)'s finding that GPT-2 cross-entropy correlates with the human N400 event-related potential response provides further evidence of a brain-LLM similarity. If we take these results as indeed indicating some computational similarity in next-word prediction, then humans may find misinformation surprising for largely the same reason LLMs register higher prediction errors: we are not used to seeing certain words combined in the ways they often are in fake news stories.

Note that these two interpretations need not be incompatible; the answer likely lies in some combination of the two. We could find misinformation surprising because we do not expect certain words in certain contexts, and we could fail to predict those words in those contexts

| Cross-Entropy | Tokenized Context |
|---|---|
| 20.360 | [' t', ' vote', **' LOL'**, '!', 'WE'] |
| 16.361 | [' otherwise', ' has', **' genital'**, 'ia', ' of'] |
| 22.042 | [' The', ' Donald', **' listening'**, ' to', ' the'] |
| 18.964 | [' State', ' Department', **' misplaced'**, ' and', ' lost'] |
| 18.967 | [' the', ' Russian', **' prostitute'**, ' story', ','] |
| 20.413 | ['Did', ' Hillary', **' die'**, ' after', ' leaving'] |

**Table 4:** A sample of context windows in misinformation-containing stories that contain the maximum cross-entropy word-prediction (bolded) in that story and the cross-entropy of that prediction; these indicate that elevated cross-entropy or surprise could arise from a stylistic conflict with how true news stories are written (first three rows) or a semantic conflict (last three rows)

because the statements are in conflict with our internal representation of the state of the world. LLMs, however, lack an internal representation of the state of the world, so their prediction errors are again a byproduct of their (exclusively linguistic) training data. I make this distinction here so as to not advocate for the view that humans and LLMs come to expect in-context words for the same reasons—or, even more extreme, that they learn language-use through the same mechanisms—which is a much more controversial claim than the one I am endeavoring to defend in this investigation.

Another conclusion we can draw from these results is of a more practical, computer-scientific nature: misinformation may indeed be classifiable through interpretable and computable natural-language metrics. This indicates a path forward for automated content-moderation on platforms such as Facebook and Twitter, which are sorely in need of such algorithmic moderation. However, the more modest results of the exploratory investigation with the LIAR dataset indicates that this problem may not be as simple as the results of the original investigation suggests. The LIAR data varies from the ISOT data in several ways, such as the

ordinal labeling of veracity and the length of the stories: 18 words versus 405 words, respectively. This may indicate that using cross-entropy and sentiment as guidelines for classification could require sufficiently-many words to reliably separate true and fake news stories, which would be a problem for the aforementioned platforms who deal mostly in bite-sized phrases. One workaround to this issue could be investigating the articles that are often linked in these posts, which would hopefully contain enough text for the classification algorithm to flag. Again, strategies for improved content moderation is a fruitful area for ongoing and future research.

Another possible impediment to implementing the classification strategy outlined by this investigation could be that misinformation-peddlers may find ways around this simple two-parameter classification system. This issue points to a question I raised at the outset: what makes misinformation go viral at such alarming rates? Could it be that the surprising nature and negative-valence of their content makes them inherently more attention-grabbing? This would seem to be the direct conclusion of studies like Brady et al. (2020), as well as the indirect conclusion of Smith & Levy (2013), which suggests that word surprisal correlates with gaze duration and therefore processing load. If this is indeed the case, then it may prove difficult for misinformation-writers to find a workaround to this classification system without sacrificing the very qualities that make misinformation effective. This does not rule out, however, some potential for an analogy to computer vision's problem of adversarial attacks (alterations to an image that are inconsequential to a human but lead a classification algorithm astray) in this domain, where writers may find combinations of words that, while still false, do not register as significantly unpredictable to an LLM (Szegedy et al., 2013).

## *VI. Limitations*

This investigation is not free from either theoretical or practical limitations, however. The ISOT Fake News Dataset contains only one source for true news stories: Reuters. This clearly introduces bias into the classification model, as the fine-tuned GPT-2 has re-weighted its parameters not to reflect true news *precisely*, but instead the writing style of Reuters. This could lead to true news stories from other reputable outlets being classified as likely containing misinformation, but I would expect this risk to be rather small due to the general similarities in professional news-reporting style. Regardless, this bias could be partially at play in the more moderate exploratory results, as the more-truthful stories in the LIAR dataset do not reflect a stylistic bias that could facilitate classification. Due to this limitation, a less-charitable interpretation of the results above may be that this classification model has merely learned to separate Reuters stories from those of obvious stylistic difference (see Table 4, page 30 for examples of these stylistic variances).

Additionally, while these metrics are indeed computationally tractable, calculating cross-entropy in particular is quite computationally *expensive*. Running a GPT-2 model—which uses 1.5 billion parameters in each prediction—through thousands of lengthy news stories is not feasible on a CPU. Accordingly, this experiment required hundreds of hours of compute time on a university computing cluster's GPUs. This hardware limitation could pose a problem for large-scale implementations of these classification metrics.

## *VII. Conclusion*

Misinformation is a scourge of our information age: false but convincing information floods the internet, and social media platforms have largely failed to stem this rising tide. What is to be done about this problem? What makes misinformation so attractive to internet users and correspondingly profitable for internet platforms? One potential path to answering both of these questions is to ask which cognitively-plausible, computationally-tractable natural-language metrics could clearly separate misinformation-containing stories from truthful ones. This investigation explored the potential of surprise (as measured by average cross-entropy calculated from a GPT-2 language model) and negative sentiment to serve as two of these metrics. I found that, in a dataset of over 40,000 news stories, true and fake stories were indeed classifiable according to these metrics with an accuracy of 79.4%. Despite limitations to the dataset and computational resources needed to carry out this classification, these results connect and further two strands of research: one on the cognitive features of misinformation, and the other on the similarities between language processing in large transformer language models and human brains. Both of these lines of work will be ever-more necessary as we become increasingly tied to the online world and language models become increasingly prevalent in our everyday lives.

References

Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1).

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of experimental psychology*, 149(4), pp. 746–756.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313-7318.

Caucheteux, C., Gramfort, A., & King, J. R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 10.1038/s41562-022-01516-2.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint, arXiv:1810.04805.*

Golan, T., Siegelman, M., Kriegeskorte, N., & Baldassano, C. (2022). Testing the limits of natural language models for predicting human language judgments. *arXiv preprint arXiv:2204.03592.*

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C.,

Fanda, L., Doyle, W., Friedman, D., Dugan, P., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3), pp. 369–380.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7), pp. 975–987.

Harris, Z. (1970). Distributional structure. In *Papers in Structural and Transformational Linguistics*. pp. 775-794.

Islam, M. R., Liu, S., Wang, Z., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(82).

Lake, B., & Murphy, G. (2021). Word Meaning in Minds and Machines. *Psychological Review*.

Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2022). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *arXiv preprint arXiv:2211.02001*.

Khan, J. Y., Khondaker, T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4.

Krause-Jackson, M. (2022, March 14). Putin's State Media News Is Interrupted: 'They're Lying to You'. *Bloomberg*. https://www.bloomberg.com/news/articles/2022-03-14/putin-s-state-media-news-is-interrupted-they-re-lying-to-you

Lee, S. K., Sun, J., Jang, S., & Connelly, S. (2022). Misinformation of COVID-19 vaccines and vaccine hesitancy. *Nat Sci Rep,* 12(13681).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graff, K., & Larson, H. (2022). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum Behav*, 5, pp. 337–348.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, pp. 57-78.

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications*, 5(1).

McDonald, S., & Ramscar, M. (2001). Testing the Distributional Hypothesis: The Influence of Context on Judgements of Semantic Similarity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 23(23).

Pennycook, G. & Rand, D. (2021). The Psychology of Fake News. *TiCS*, 25(5), pp. 388-402.

Pierri, F., Perry, B. L., DeVerna, M. R., Yang, K., Flammini, A., Menczer, F., & Bryden, J. (2022). Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Nat Sci Rep,* 12(5966).

Radford, A., Jozefowicz, R., & Sutskever, I. (2017). Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6), pp. 386-408.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, 118(45).

Seetharaman, D., Horwitz, J., & Scheck, J. (2021, Oct. 17). Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184?mod=article_inline

Sejnowski, T. (2019). The unreasonable effectiveness of deep learning in artificial intelligence. *PNAS*, 117(48),

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). dEFEND: Explainable Fake News Detection. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 395-405.

Skarabela, B., Ota, M., O'Connor, R. & Arnon, I. (2021). 'Clap your hands' or 'take your hands'? One-year-olds distinguish between frequent and infrequent multiword phrases. *Cognition*, 211.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), pp. 302-319.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tunstall, L., von Werra, L., & Wolf, T. (2022). *Natural Language Processing with Transformers, Revised Edition*. O'Reilly.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), pp. 1146–1151.

Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9(1), pp. 36-45.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

# Appendix

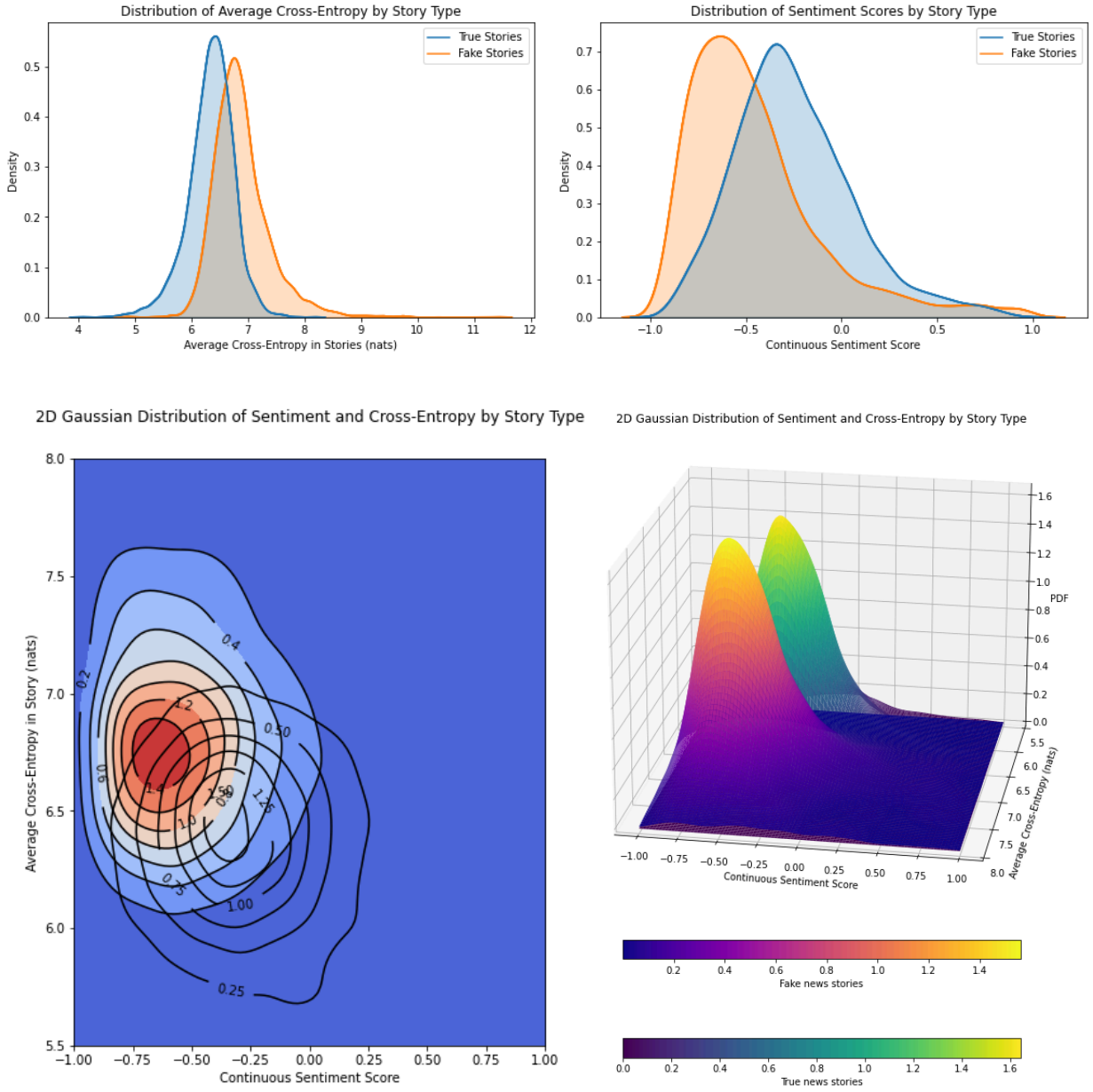



**Fig A1**: Same family of plots as described in *Results*; this data represents the distributions calculated from the second of three cross-validation folds.

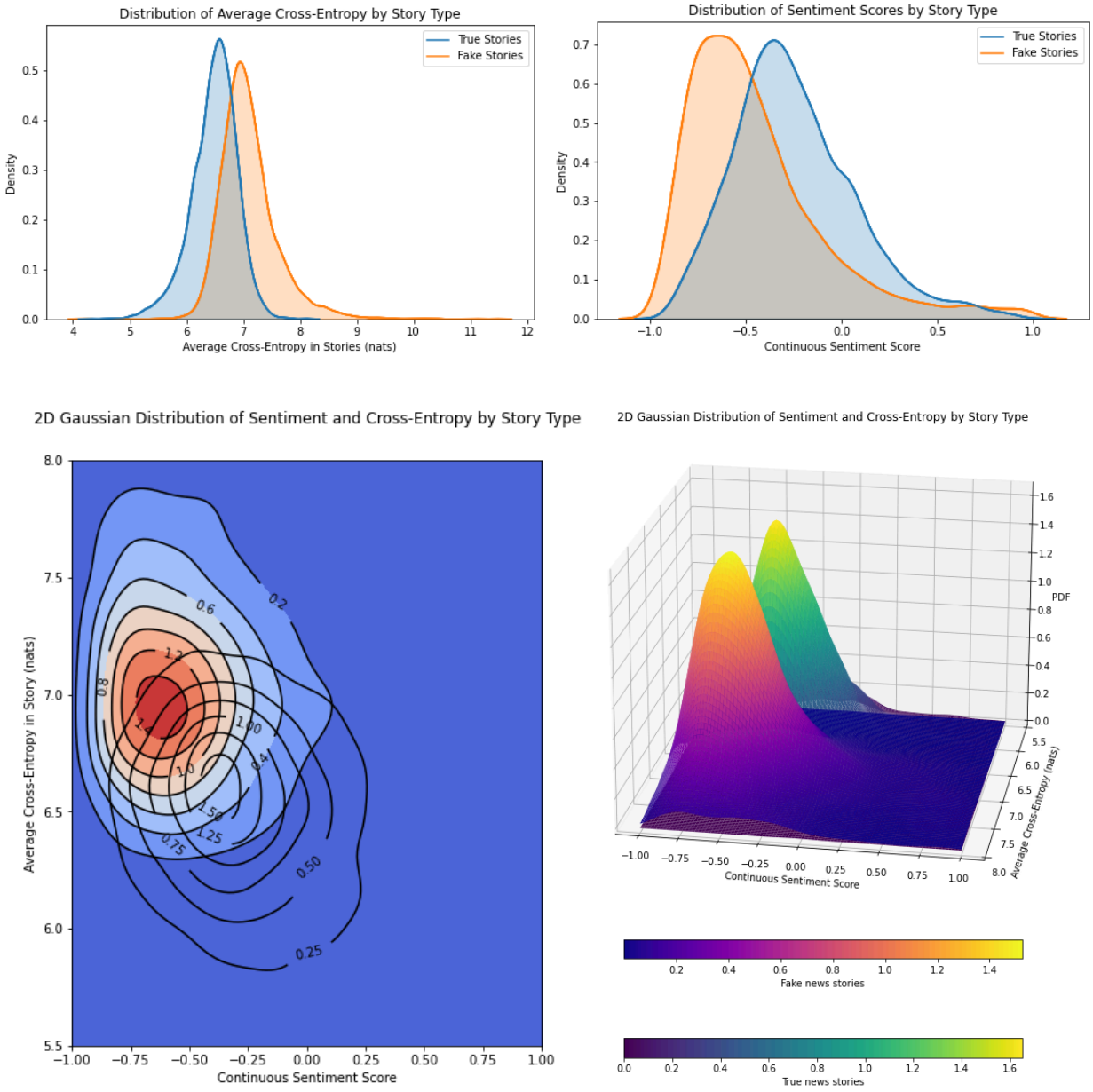



**Fig A2**: Same family of plots as described in *Results*; this data represents the distributions calculated from the third of three cross-validation folds.