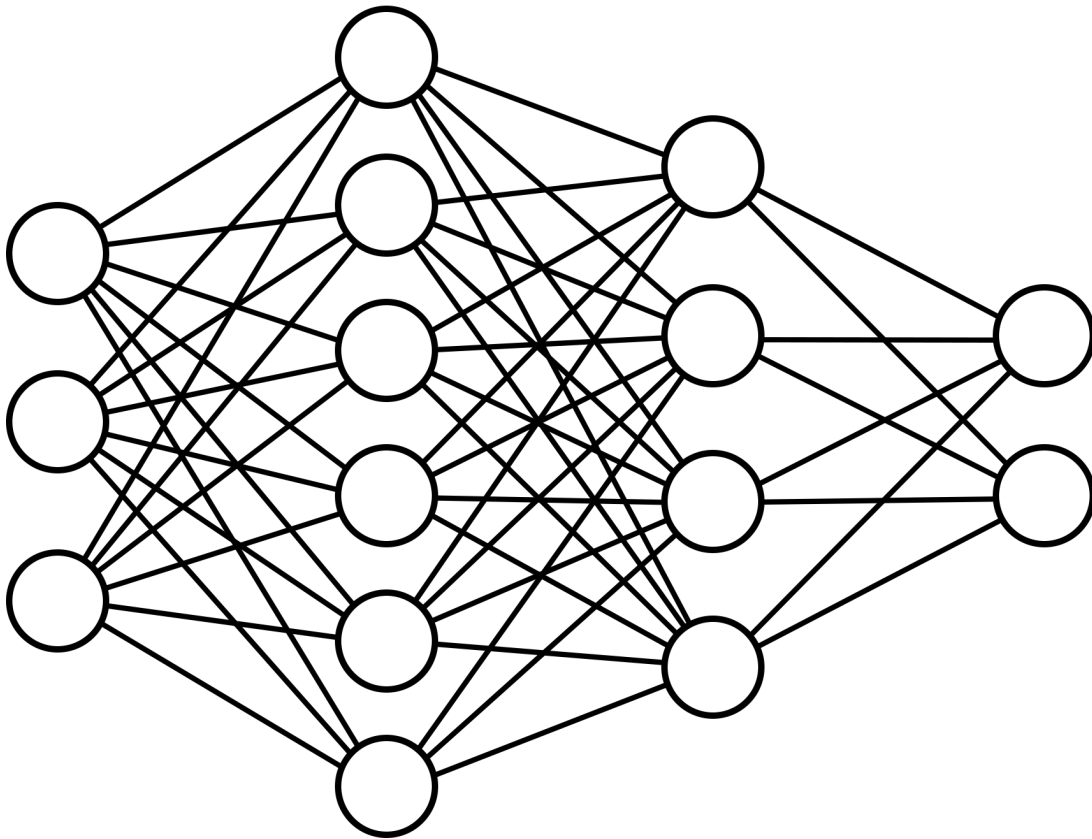


Figure 1

A Feedforward Neural Network with Four Layers



Navigating Disinformation: A Deep Learning Approach

Kimi Chen

Dominican International School

Final Paper

2023-2024

Advisors: Ms Janice Doyle

Ms Ma. Robeth Collado

Mentors: Mr. Tim Sampson

Mr. Ark Su

Abstract

This project explores the capabilities of artificial intelligence (AI) in detecting disinformation by using custom-built deep learning models and pre-trained language models. The purpose was to assess whether specialized and more general AIs can outperform humans in accurately identifying disinformation. Various ML models were built and utilized, including long short-term memory (LSTM) networks with and without GloVe embeddings, a fine-tuned DistilBERT model, and pre-trained models like Llama-2 and GPT-4 Turbo. The custom models (including DistilBERT) were trained on a large fake news dataset and evaluated on holdout sets. Their performance was then compared against human benchmarks established through surveys and quizzes administered to DIS high school students. The results indicate that AI models generally achieve higher accuracy than human benchmarks, with GPT-4 Turbo performing the best at 73.24% accuracy. However, their real-world applicability is limited by challenges such as overfitting, data quality, different information styles and formats, and the complexity of language nuances. This research highlights the potential of AI in combating disinformation but also emphasizes the need for continued advancements, especially in developing more comprehensive and diverse datasets related to fake news.

Keywords: Disinformation detection, natural language processing, long-short term memory, transformers

Navigating Disinformation: A Deep Learning Approach

As artificial intelligence (AI) has become more integrated into our day-to-day lives, many are afraid of its power and potential for misuse. This project seeks to alleviate the tension surrounding those claims by demonstrating that specialized AI can improve the wellness of society since it can outperform humans in specific tasks such as identifying disinformation (commonly known as fake news).

I will utilize the topic of identifying disinformation as a case study in this project. Disinformation impacts society at all levels; from simple daily news to public health to elections, they are prevalent in all aspects of society, so detecting them accurately is important for the growth and development of the general public. By utilizing various techniques, I seek to demonstrate that these advanced technological developments are more accurate than humans in these specialized tasks, and by building my own models and deploying a website to showcase this project, I also aim to demonstrate ethical AI development and usage.

The concept of AI has been around ever since the 1950s when Arthur Samuel developed a checkers program that learned from its own mistakes (Copeland, 2023). Since then, we have also constantly used AI-powered tools such as spell-checkers, translation tools, and search engines like Google. However, AI has become the talking point recently due to the upsurge of generative AI tools such as ChatGPT and DaLL-E. Due to these advancements, people have increasingly become worried about AI's impact on modern society, such as their potential to take over countless job fields or their abilities to portray sentient-like behavior (Ipsos, 2023; Lerner, 2023; McQuilling, 2022; Neuroscience News, 2023). Even though these notions are not unfounded, AI has also shown its usefulness in numerous fields. From education to product development to healthcare improvements, AI has improved the betterment of society: it allows for more one-on-one learning, reduces the workload of developers, and makes pattern recognition in the medical field more efficient than ever (Baidoo-Anu & Ansah, 2023; Dohmke, 2023; Esteva et al., 2021; Mello et al., 2023; Peng et al., 2023; Shani & GitHub Staff, 2023; Susanto et al., 2022). AI simply also does not have the ability to think, unlike what movies often portray. A common test to assess whether AI can demonstrate human intelligence is the Turing test, in which a human investigator asks an AI and a human respondent the same question without knowing which is the human and which is the AI. After a period of time, the investigator will then decide which response belongs to the robot (Copeland, 2023). However, this test was proposed in 1950 and is outdated; what AI really does is find patterns through advanced mathematics, statistics, and computer science, and it has no ability to think like a human would (Bhaimiya, 2023).

This project aims to not only explore the capabilities of AI in detecting disinformation but also to dispel common misconceptions about AI's role in society. The subsequent sections will detail more information on my secondary research, specifically the history of disinformation and traditional and modern methods of identifying disinformation, the process of machine learning from start to finish, the advancements in deep learning and natural language processing, and past research on the accuracy of identifying disinformation from human benchmarks and different AI techniques. I will then explain in detail the methodology of this research, the initial survey sent to DIS high school students, the development and evaluation of AI models, the process of making follow-up quizzes for students, and the analysis of data from the models and human benchmarks established at DIS.

History of Disinformation

In ancient times, disinformation was used in warfare to mislead enemies about strengths, weaknesses, and strategies, which was usually accomplished through spies, false messages, and deceptive tactics on the battlefield. In the modern era, disinformation has been employed in political propaganda, specifically during the 20th century, like in WWI, WWII, and the Cold War. Governments used disinformation to sway public opinion, weaken the enemy's morale, and gain geopolitical advantages (Shaer, 2017; Ward et al., n.d.). In addition, prior to the internet, the spread of fake news was primarily through traditional media forms like yellow journalism with a focus on eye-catching

headlines rather than factual reporting. Yellow journalism often included exaggerated stories about crimes, gossip, disasters, and satirical pieces. However, with the uprising of the internet, it is much easier for anyone to publish content, which has led to a rapid increase in both reliable and unreliable content. The challenge lies in the volume of content that needs to be checked as well as the time it takes to do the fact-checking act itself. Social media has also become a widely used tool for spreading fake news as it allows for rapid sharing of content, and the interconnected nature of social media means that information, whether true or false, can quickly go viral (De Beer & Matthee, 2020).

Disinformation's Impact

Disinformation significantly impacts society at all levels, with the most prominent effect on democracy. According to research from Stanford University, disinformation seriously threatens democracy as it manipulates public opinion and weakens the public's trust in democratic governments (De Witte, 2022). This manipulation is particularly effective in political contexts since disinformation campaigns often capitalize on preexisting societal divides to influence political events and public understanding, which affects people's perception of current issues and ultimately affects the results of elections (Faust, 2018).

In addition, in the health sector, disinformation about diseases and vaccines has led to public health crises, especially during the COVID-19 pandemic. False information deliberately created and spread about vaccine development, safety, and effectiveness resulted in people hesitating to receive the vaccine and ignoring health guidelines due to the lack of public trust, which worsened the pandemic's impact (*Tackling Coronavirus Disinformation*, n.d.).

Traditional Methods of Identifying Disinformation

The traditional methods for identifying fake news primarily relied on verification by human editors and expert journalists, which involved manually checking news content for authenticity and accuracy. However, as Galli et al. (2022) pointed out, it was ineffective due to the volume of content, limited scalability, subjectivity and bias, and the lack of automation—the volume of news generated on the internet is vast and continuously growing, meaning that manual verification cannot keep up with the rapid generation and spread of information; traditional methods are also not scalable since they require significant human resources and time; manual methods can also be subjective due to the personal beliefs, opinions, and biases of the journalists and the editors; and traditional methods lack the automation necessary for quick and efficient analysis of a large number of news sources, making them less effective in a digital environment where news spreads quickly.

Utilizing AI to Identify Disinformation

Several studies have shown positive results for utilizing AI to identify disinformation and have suggested its use in the future for their accuracy. The specific details will be mentioned and compared later in this paper. However, other research has also shown the impracticalities of deploying AI systems for identifying disinformation. One study conducted by He et al. (2023) suggests that even though AI is helpful in identifying disinformation, it is still in its early stages for practical applications due to the complexities of setting up these models. The ever-evolving nature of fake news and the complexity of multimedia also means that these models need to be updated consistently. Therefore, He et al. suggest a multidisciplinary approach that combines the usage of AI, media analytics, ethics, and social sciences for a more comprehensive strategy. Yuan et al. (2023) also showed that AI techniques are capable of efficiently processing large volumes of data, with newer models able to identify fake news accurately. However, research on fake news detection with large language models (like GPT) is still lacking, and explainable artificial intelligence methods for fake news detection are not yet fully established. Both research studies stated that AI eventually would have its place in identifying disinformation; the technology at the moment is not advanced enough just yet. In a later section, several previous studies on this field will be analyzed.

Machine Learning

Machine learning (ML), a crucial subset of AI that identifies patterns based on a set of inputs and outputs, is what allows all of the AI technologies mentioned so far to work. It is similar to the process of learning to cook from recipes. Imagine each recipe as a dataset, a collection of data points often in the format of a table. In each recipe, there are the ingredients, cooking steps, and the finished dish. The ingredients and cooking steps are the inputs, and the finished dish is the output. In this analogy, the ML model is an aspiring chef who is trying to learn how to cook a variety of dishes by understanding the patterns presented in the recipes. At first, the connection between understanding how each ingredient and step works is very weak, meaning that the ML model does not understand the relationship between different inputs and outputs. As a note, more formally, inputs are commonly referred to as *features* and outputs are referred to as *labels* (*Framing: Key ML Terminology*, n.d.). However, as it “learns” (will describe in detail later), it starts experimenting with different ingredients and steps based on its understanding of what makes a dish delicious. The more extensive and diverse the recipes (data), the better ML models become at predicting the outcome of new, unseen recipes. Specifically, in the case of identifying disinformation, ML models will be fed both real and fake news passages, as well as whether they are actually real or fake. They will learn from examples of both to establish an understanding and form relationships to correctly identify whether a new piece of news is real or not (Neagoie & Bourke, n.d.).

The heart of ML lies in the choice of algorithms, or rules and instructions that computers follow to learn from data and make predictions. In the chef analogy, an algorithm is the chef’s method and intuition for combining the ingredients and executing the cooking steps to create a dish. Just like a chef might use different techniques for baking, frying, or grilling, ML models rely on different algorithms tailored to the type of data they are processing and the specific task they are designed to perform (Wakefield, n.d.). An inappropriate algorithm in the context would result in ineffective predictions, similar to the chef using the wrong part of their arm to whisk the batter, which would result in food that does not taste as delicious (Undy, 2023). In identifying disinformation, there are specific algorithms that recognize patterns from text, and it would not be wise to recognize patterns with an algorithm designed to predict a continuous variable (like house prices).

A typical machine learning project involves data collection, exploratory data analysis, data preprocessing, model training, and model evaluation (Morgunov, 2023). In the following sections, all of these aspects will be explained briefly, with an emphasis on model training, as that is the core of machine learning, especially in the context of this project.

Data Preparation

Data is needed to build and train an ML model. Without data, these algorithms have nowhere to get information and build patterns. Data preparation is essential to make sure that the model trained is accurate, contains as little bias as possible, and can be generalized for inferencing from new data that it has not seen before. Data preparation contains multiple steps, each of which makes sure ML models receive high-quality, relevant, and well-structured data. This process includes collecting diverse and accurate data, exploring this data to understand its characteristics, and then preprocessing it to a form where it can be understood by the computer easily by removing any extraneous information. The importance of data preparation cannot be overstated—it is the key process that directly influences the accuracy and effectiveness of ML models; no ML algorithm can compensate for bad data processing (Redman, 2018). In the context of this project, proper data preparation ensures that the subtleties and complexities of language and content are accurately captured and represented, which allows ML models to have a strong foundation to learn patterns and make accurate predictions (The Pecan Team, 2023).

Data Collection. The quantity, quality, accuracy, and variety of data are essential to ensure that a model is “well fed”, much like a hungry child. This process, for the majority of people developing models, involves searching online for reliable datasets. However, for research areas that do not have

good datasets (or simply lack datasets), individuals and corporations might choose to compile their own datasets by scraping the web or through other physical media sources. An example of this is ChatGPT, which was trained on data from the majority of the web as well as other physical documents and texts (Ammu, n.d.). The objective is to build a dataset that is not only large but also representative of the variety of content in real-world scenarios. In this project, data will be collected from datasets publicly available online (more information under Methodology).

Exploratory Data Analysis (EDA). After collecting sufficient data, Exploratory Data Analysis (EDA) is conducted to gain insights into the data's structure and content. It involves analyzing the data descriptive statistics (such as mean, median, mode, and standard deviation) and different kinds of graphs and visuals (such as histograms, scatter plots, and box plots) to extract information that was hidden originally. As Purohit (2023) explained, the main goal of EDA is to extract and understand the bigger picture of the data as well as the relationships between different features and labels. Python libraries like Pandas for table viewing and data manipulation and Matplotlib and Seaborn for visualization are commonly used. In the context of identifying disinformation, EDA may include analyzing word frequencies, identifying common phrases, and understanding the context in which certain terms are used (more details later).

Data Preprocessing. Data preprocessing is the process of transforming raw data into a clean and organized format suitable for ML models to train from. It is usually done through the Pandas library in Python, and it involves data cleaning, data transformation, data reduction, data integration, and feature engineering. Data cleaning handles missing data, usually through methods such as dropping it from the dataset or imputation (replacing missing values with the mean, median, or mode). Data transformation scales the data with techniques such as normalization, which scales the data to a specific range, and standardization, which scales the data to have a mean of 0 and a standard deviation of 1 (Simplilearn, 2023). For categorical features (like the color of the house or the type of floor), they are encoded into numerical formats, usually by creating several new features and encoding 1 for the feature that matches with the original data and 0 otherwise (Brownlee, 2020). An example of one-hot encoding can be seen in Table 1 and Table 2, where Table 1 is before the encoding, and Table 2 is after the encoding. Data transformation is used so that the numbers inputted to the ML model are consistent and easier to train. Data reduction reduces the number of extraneous variables or irrelevant features. This process can be done through simple methods such as dropping the entire feature or complex methods like dimensionality reduction and Principal Component Analysis (PCA), which essentially extract information from related features and compress it into fewer features while maintaining most of the relevant and important information from the original features. Data reduction helps reduce the complexity of data, speed up the learning process, and improve model performance. Data integration combines data from multiple sources and datasets and ensures that the units stay consistent and that there are no duplicate data points. Feature engineering utilizes knowledge from the research field and analysis from EDA to extract features from raw data. Some common techniques involve creating new features and modifying existing ones to improve model performance (Azevedo, 2023). For example, an expert in housing price analysis might add a new feature for the living space per person as it may be useful and remove features such as the color of the toilet as it may be irrelevant. Finally, the data, usually presented in a table format for better readability, will be converted into matrices with libraries like NumPy so that the computer can read the information faster and easier. Text-specific preprocessing will be discussed in later sections, but the above steps are still essential and used for these types of preprocessing.

Table 1

Example of a Categorical Feature without One-Hot Encoding

Item	Color
1	Red
2	Green
3	Blue
4	Green
5	Red

Table 2

Example of a Categorical Feature One-Hot Encoded

Item	Color_Red	Color_Green	Color_Blue
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0
5	1	0	0

Fundamental ML Algorithms

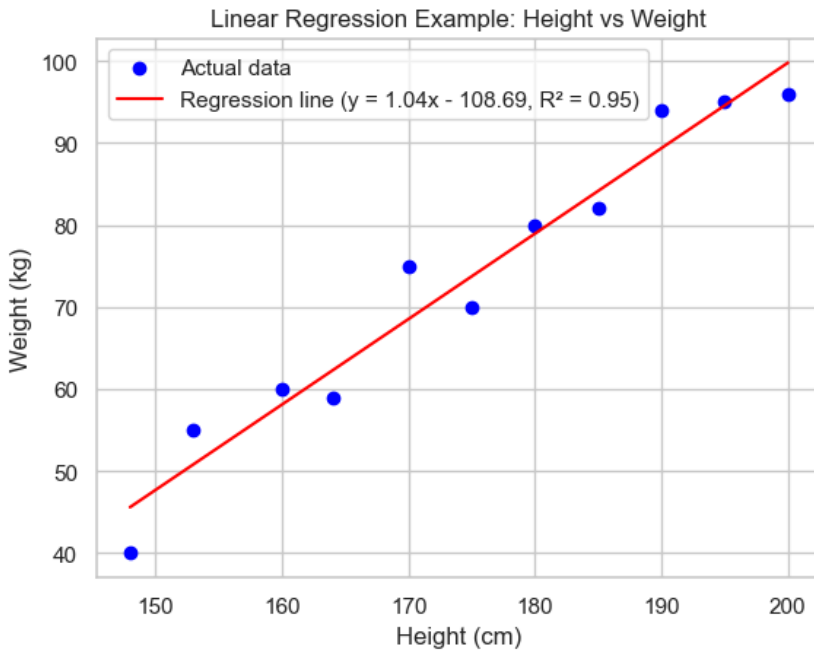
The heart of machine learning utilizes algorithms; as a reminder, algorithms are the set of rules and instructions that computers follow to learn from data and make predictions. These algorithms primarily fall into two categories: regression and classification, which are designed for different types of data analysis.

Regression. Regression is used for predicting continuous numerical values, such as forecasting tomorrow's temperature or housing prices. Some common regression models include linear regression and polynomial regression. Linear regression attempts to find a line of best fit, the process of fitting a linear equation to observed data so that future predictions can be made based on the generalization of the line (GeeksforGeeks, 2023). Figure 2 is an example of linear regression. The objective is to predict a dependent variable \hat{y} based on one or more independent variables x_j . As the name suggests, linear regression works well only when the relationship between the independent and dependent variables is linear. A common example of a regression task is predicting house prices based on features such as the size of the house, the number of rooms, and the floor material. As shown in (1), $f_{w,b}(x)$ is a function that takes an input x , depending on the values of w and b , and outputs a predicted value \hat{y} . This formula resembles a linear equation, often denoted by $y = mx + b$. In this formula, x_j are the values of the independent variable (feature) at index j . w_j is the weight (how much each independent variable x_j

should influence \hat{y} at index j . In the house example, x_1 might be the size, x_2 might be the number of rooms, x_3 might be the floor material, and so on. Correspondingly, the size might have the most “weight” in predicting house price, so w_1 might have a value of 5. The number of rooms might still be important, just not as important as the size, so w_2 might have a weight of 3. The floor material might only influence the prediction ever so slightly, so w_3 might only have a value of 0.5. Finally, b is the bias term (also known as the intercept), which represents the value of \hat{y} when all the features x_j are 0 (Burkov, 2019, pp. 21–22).

Figure 2

Example of a Basic Linear Regression with One Feature and One Label



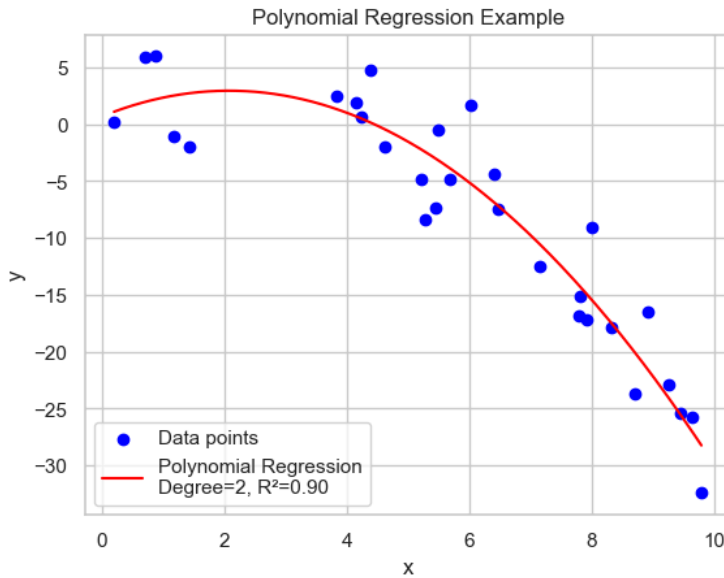
Note: The data points in this figure are made up.

$$f_{w,b}(x) = \hat{y} = w_1x_1 + w_2x_2 + \dots + w_nx_n + b \quad (1)$$

Polynomial regression is similar to linear regression, except that it can capture more complex relationships, specifically those that are polynomial (just as the name suggests). Figure 3 is an example of polynomial regression. Equation (2) is the general formula for polynomial regression, which is similar to (1) except it includes exponents. In polynomial regression, new features may be created from an original feature with a power of d if there is a polynomial relationship (like quadratic or cubic) between the original feature and the label (Ng et al., n.d.).

Figure 3

Example of a Basic Polynomial Regression with One Feature

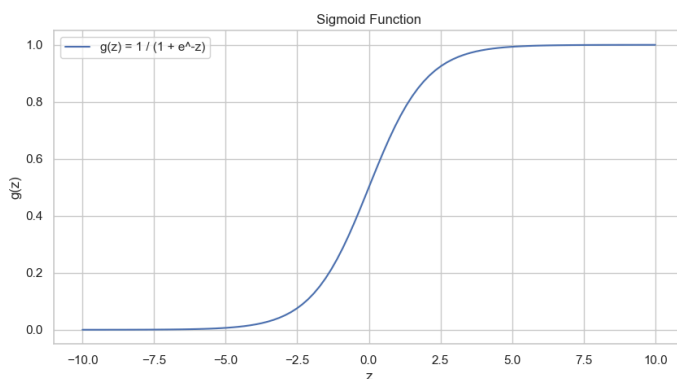


$$f_{w,b}(x) = \hat{y} = w_1x_1 + w_2x_1^2 + \dots + w_nx_1^d + w_1x_2 + w_2x_2^2 + \dots + w_nx_2^d + \dots + b \quad (2)$$

Classification. Classification is used to predict and categorize data into different groups or classes, such as identifying whether an email is spam or classifying the type of cat based on images. Classification is also where identifying disinformation falls under. The most fundamental classification technique that other methods branch off of is logistic regression. Logistic regression is used when the dependent variable is binary (like true or false). It predicts the probability of an event happening by fitting data to a logistic (also known as sigmoid) function as shown graphically in Figure 4 and defined in (3). The sigmoid function $g(z)$ takes a real number z and maps it onto a value between 0 and 1, which is used to predict the probability of an event happening. When z becomes a big negative value (as z approaches negative infinity), $g(z)$ becomes $\frac{1}{\infty} = 0$, and when z becomes a big positive value (as z approaches infinity), $g(z)$ becomes $\frac{1}{1+1/\infty} = \frac{1}{1+0} = 1$ (Ng et al., n.d.).

Figure 4

Graph of the Sigmoid Function.



$$g(z) = \frac{1}{1+e^{-z}} \quad (3)$$

Equation (4) is the formula for logistic regression. The arrows signify that the variable is a vector, which essentially means that it contains a list of numbers. \vec{w} means that it is a list of weights, and \vec{x} means that it is a list of features. Therefore, $\vec{w} \cdot \vec{x} + b$ is the same formula as linear regression defined in (1). The reason for substituting linear regression into the sigmoid function is to map the output to a probability between 0 and 1, which allows for a binary classification of the data points. A threshold can be set so that all probabilities over the threshold will be classified as positive and negative otherwise. A common choice is 0.50 since it represents equal probabilities of the event occurring (Burkov, 2019, pp. 25–26). For identifying disinformation, a value greater than or equal to 0.5 means that the news is real, and a value less than 0.5 means that the news is fake.

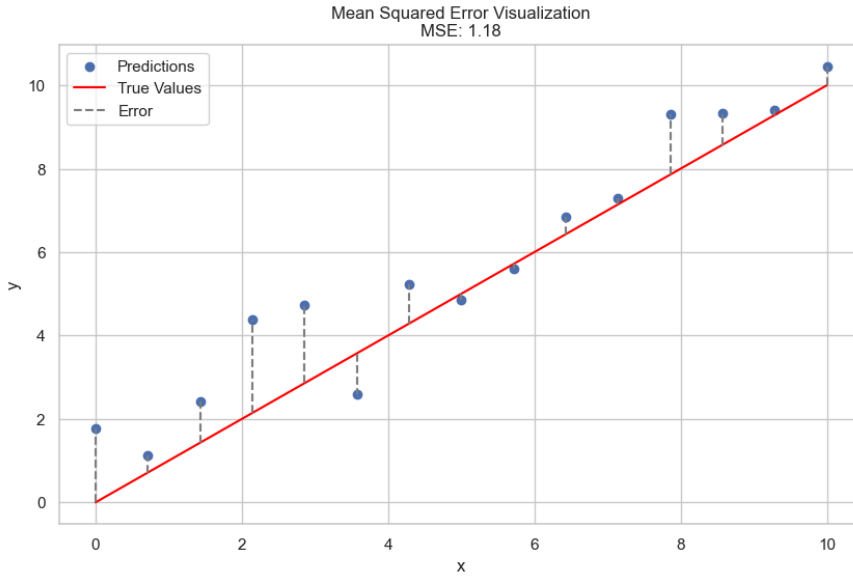
$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1+e^{-(\vec{w} \cdot \vec{x} + b)}} \quad (4)$$

Cost Function

The cost function is an important concept that helps ML models learn and improve their predictions. It is essentially a scoring system that grades how well the model is performing by minimizing the “cost”—the lower the score, the better the performance. The primary goal of the cost function is to quantify the error between the predicted values and the actual values to measure objectively how far off a model’s predictions are from the actual outcome.

There are various types of cost functions, but one of the most commonly used examples for regression tasks is the Mean Squared Error (MSE). Equation (5) is the formula for MSE, where $J(\vec{w}, b)$ takes weight w and bias b and outputs the cost. i represents the index of the current training example, and m represents the total training examples. Note that $y^{(i)}$ does not mean y to the power of i ; the parentheses around i means that it is the training example i . This equation calculates the difference between the predicted output at index i against the target value at index i . This difference is then squared to remove negative values and to exaggerate outliers. This error is then summed across all training examples, then divided by 2 times the number of training examples to calculate the average. Figure 5 is a visual representation of this concept; the errors represented by the gray dotted lines are squared and summed for all points. Then, it is divided to calculate the average error. The MSE essentially calculates the average of the squares of the errors, which is the average squared difference between the actual values and the values predicted by the model (Ng et al., n.d.).

$$J(\vec{w}, b) = \frac{1}{2m} \sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)})^2 \quad (5)$$

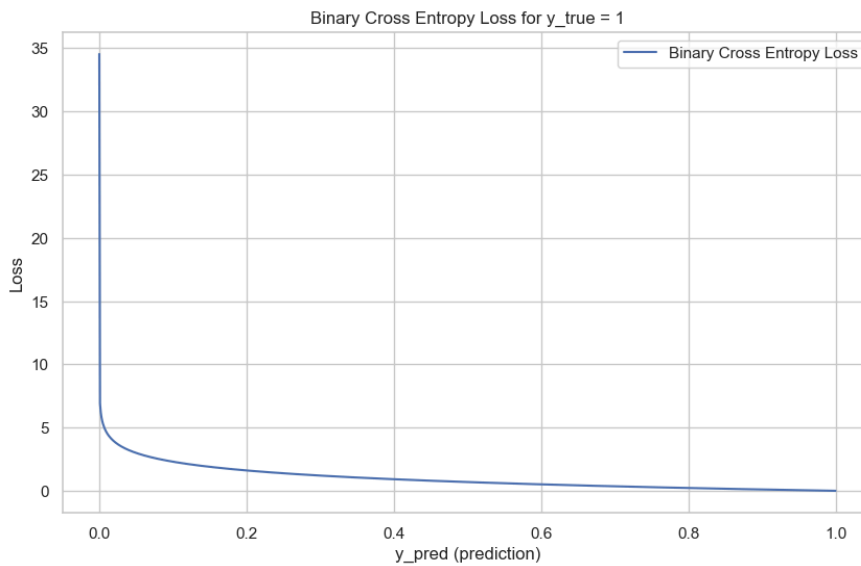
Figure 5*Visual Representation of MSE*

For classification tasks, Binary Cross-Entropy Loss (also known as Log Loss) is commonly used for models that categorize values into binary options. The formula is shown in (6). Similar to MSE, $y^{(i)}$ is the actual value for the i th training example, where it can only be 0 or 1 (since it's binary). $f_{w,b}^{\rightarrow(i)}$ is the predicted probability that the i th training example belongs to the class with label 1. This is the output of the logistic regression model. Since $y^{(i)}$ can only be 0 or 1, when $y^{(i)} = 1$, the equation inside the summation becomes $-\log(f_{w,b}^{\rightarrow(i)})$, which is shown graphically in Figure 6. Alternatively, when $y^{(i)} = 0$, the equation inside the summation becomes $-\log(1 - f_{w,b}^{\rightarrow(i)})$, which is shown graphically in Figure 7. As seen in Figure 6, when the predicted value $f_{w,b}^{\rightarrow(i)}$ approaches 1, the loss approaches 0, and as the predicted value approaches 0, the loss approaches infinity. Since $y^{(i)} = 1$ in this case, the loss decreases when $f_{w,b}^{\rightarrow(i)}$ approaches the true label $y^{(i)}$ at 1. On the other hand, in Figure 7, since $y^{(i)} = 0$, the loss is the lowest when $f_{w,b}^{\rightarrow(i)}$ predicted close to 0 (Ng et al., n.d.).

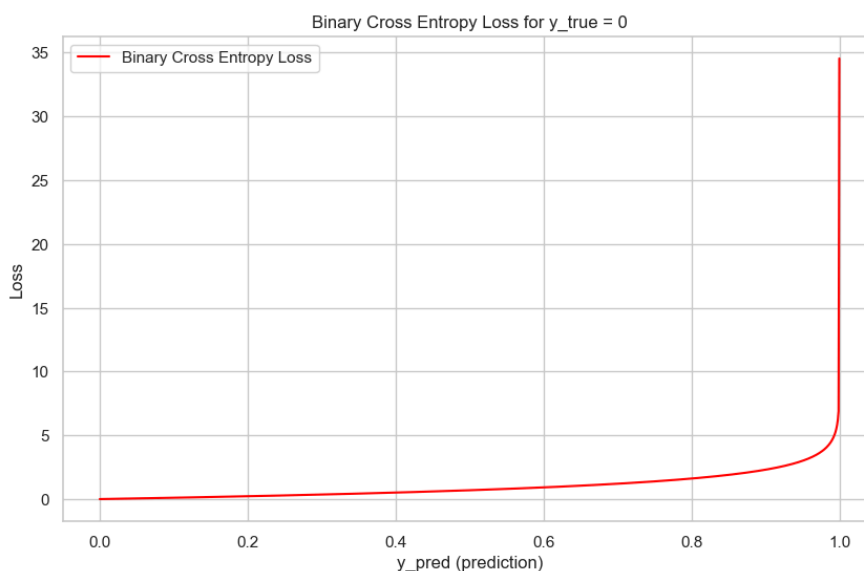
$$J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^m [-y^{(i)} \log(f_{w,b}^{\rightarrow(i)}) - (1 - y^{(i)}) \log(1 - f_{w,b}^{\rightarrow(i)})] \quad (6)$$

Figure 6

Graph of $L(f_{w,b}^{\rightarrow}(x^{(i)}), y^{(i)}) = -\log(f_{w,b}^{\rightarrow}(x^{(i)}))$ When $y^{(i)} = 1$

**Figure 7**

Graph of $L(f_{w,b}^{\rightarrow}(x^{(i)}), y^{(i)}) = -\log(1 - f_{w,b}^{\rightarrow}(x^{(i)}))$ When $y^{(i)} = 0$



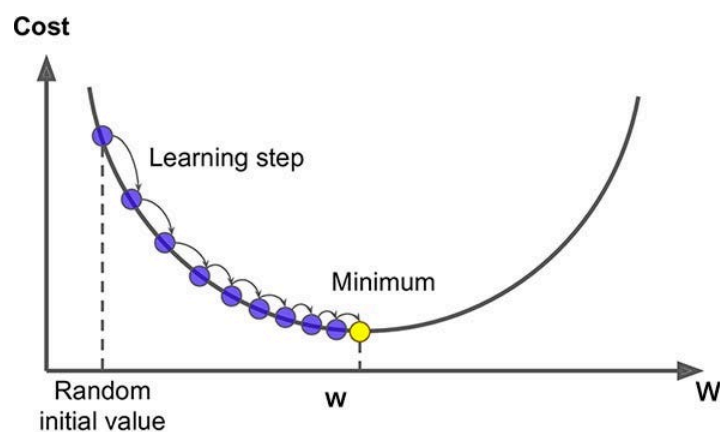
Gradient Descent

Gradient descent is an optimization algorithm to minimize the cost function, which, as previously discussed, measures how well a model is performing. As an analogy to gradient descent, imagine being on a mountain and needing to find the quickest path downhill. You cannot see the entire landscape at once, so you take steps in the direction that seems steepest, reassessing after each move. This similar process is what gradient descent does with a cost function: it iteratively adjusts the parameters (weights and biases) of the model, and then it can take steps toward the minimum value of the cost function (Ng et al., n.d.).

The calculation of gradient descent involves derivatives, which in the mountain analogy, represent the steepness of the hill. In technical terms, gradient descent first initializes the model parameters, usually with random values. It then calculates the gradient (derivative) of the cost function with respect to each parameter. The gradient is a vector that points in the direction of the steepest increase of the cost function. Finally, it updates the parameters by moving in the opposite direction of the gradient. The size of the step taken in each iteration is determined by the learning rate, a hyperparameter (essentially, a setting) that was prechosen. If the learning rate is too small, the ML model will take a long time to learn, but if the learning rate is too big, the gradients calculated will not lead the model to the desired minimum. Repeat these steps until the cost function stops decreasing drastically, which can be different depending on the problem at hand (Goodfellow et al., 2016, pp. 79–82). Figure 8 demonstrates the gradient descent of a basic model with only one feature. For multiple features, the graph would be in a multi-dimensional space rather than a 2D graph.

Figure 8

Visual Representation of Gradient Descent with One Feature



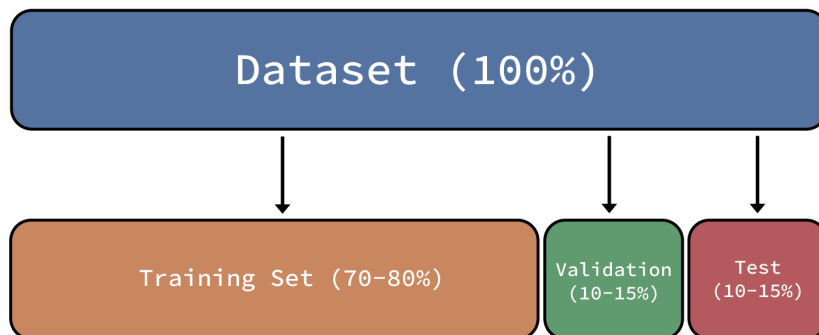
Training, Validation, Test, and Holdout Sets

The process of training ML models involves three major processes. Similar to students receiving homework, quizzes, and exams, ML models receive the same treatment. Homework teaches students how a concept works, just like the training set giving ML models data to learn from. Quizzes evaluate students' understanding before the exam, just like the validation test making sure that the model is ready for the final test. Exams assess students understanding, making sure they are ready to apply those concepts in the real world, just like the test set ensuring that ML models are ready for new inputs. Students receive grades for their homework, quizzes, and exams, and it is no different than ML models. It is essential to keep these three sets separate and use them for their intended purposes (Acharya, 2023). Just as students should not have access to exam questions ahead of time, a model should not be given the answers (test set) during its training phase. This separation ensures that the model can generalize well to new data rather than just repeating what it has seen. Géron (2019) also suggested that to get these sets of different data, a big dataset is usually split into three—the training set contains the majority of the data, usually around 70 to 80%, and the validation and the test sets contain the rest of the data, around 10 to 15% each (pp. 30–32). Figure 9 visually explains the dataset makeup. A holdout set is also sometimes used on top of the other three sets. A holdout set is a completely separate dataset from a different source that is only used at the end to make inferences about new data that might not be exactly the same (*What Is Holdout Dataset in Machine Learning?*, n.d.). (Note that the term “holdout set” is also used by some to describe a validation set or test set, but in the context of this paper, “holdout set” will always refer to testing on a different dataset). A bad performance on the holdout set might suggest

that the model is overtrained and needs readjustment. In the context of identifying disinformation where the language style is important, having a holdout set ensures that the model can also make inferences about a different style than it was originally trained and tested on.

Figure 9

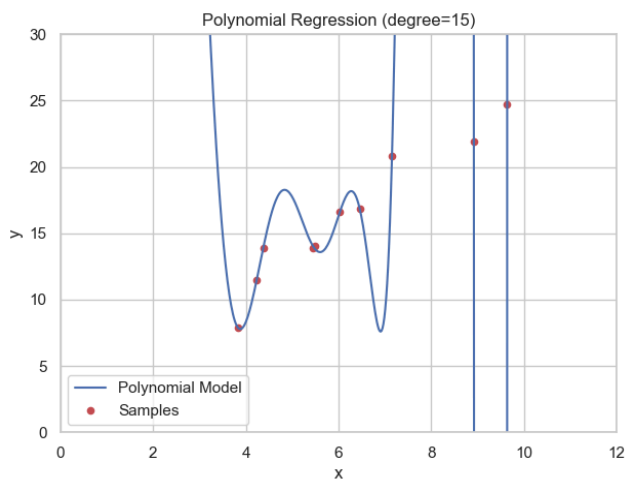
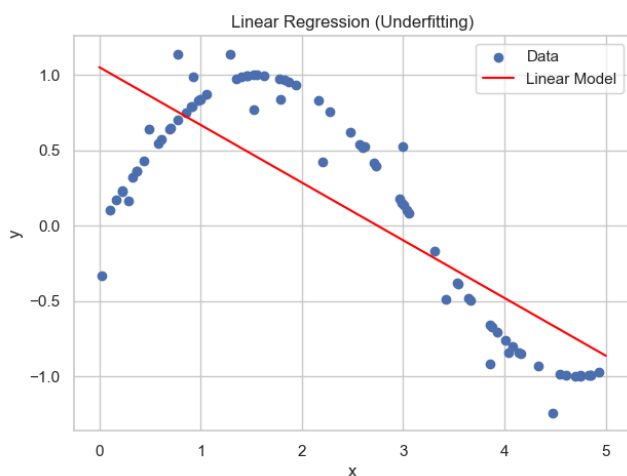
Visual Representation of Dataset Makeup



More specifically, the training set is the portion of the data used to train the model. It is where the model learns the relationships between the features and the labels by adjusting its parameters to minimize the cost function, just as students learn new material and improve their understanding over time. The larger and more representative the training set, the better the model can learn the underlying patterns in the data. The validation set (also known as the cross-validation or development set) is used to evaluate the model's performance during the training phase. It is not used for training the model but to tune the hyperparameters, like the learning rate or the complexity of the model (*What Is Hyperparameter Tuning?*, n.d.). The validation set helps prevent overfitting, which occurs when a model learns the training data too well (more details in the next section). Géron (2019) also stated that the test set is used to assess the model's performance after the model has been trained and validated; it provides the final measure of how well the model will perform on data it has never seen before when it's deployed in the real world (pp. 30–32). In the context of disinformation identification, this process helps to ensure that the model can accurately detect false information across different scenarios and sources.

Underfitting, Overfitting, and Regularization

It might seem intuitive to build a model that has a lot of parameters that fit every training feature perfectly. However, this would lead to overfitting, where the model is excellent at predicting data that it has already seen but horrible at predicting new data points (meaning that it performs well in the training set but horrible in validation and test sets). This would be like students memorizing answers to a test without understanding the subject: they might ace the practice questions but will struggle to apply the concepts in a real exam. This would result in similar graphs as shown in Figure 10, where the line of best fit is too complex and overcompensates each data point (*What Is Overfitting?*, n.d.). On the other hand, there is underfitting, which occurs when the model is too simple to understand the pattern of the data. Imagine teaching the basics of grammar without context or examples; the students might understand the rules, but they would not be able to construct or comprehend complex sentences. Géron (2019) also noted that underfitted models do not generalize to new data that well, and they do not know the training data that well either, which leads to poor performance across training, validation, and test sets (p. 29). An example of an underfitted model is shown in Figure 11.

Figure 10*Example of an Overfitted Model***Figure 11***Example of an Underfitted Model*

To combat overfitting, the size of the training data can be increased either by adding more data, reducing the size of the validation and test set, or artificially augmenting the data (for text, this might involve synonyms or rephrasing). As discussed in EDA, reducing the number of features in the dataset by removing irrelevant or extraneous features that may be causing the model to learn noise can also help with overfitting. In addition, regularization, a technique that penalizes complex models, can also be used to combat overfitting. It is typically done by reducing the weight of features that make the model complex, such as high-degree polynomials in polynomial regression (*Overfitting and Underfitting in Machine Learning*, n.d.). The amount of regularization is also a hyperparameter, meaning that it can also be controlled to maximize model performance.

To combat underfitting, using a more sophisticated model that can capture more information about the data can help, as well as adding more features or constructing polynomial features. If regularization is used, reducing its effect (or removing regularization) can allow the model to focus more attention on the training data. Sometimes, the model simply has not converged close to the minimum of the cost function yet, so letting the model train longer can help reduce underfitting (Géron, 2019, p. 29).

Measuring Performance

Assessing the performance of ML models is essential to understanding their effectiveness and ensuring they are making accurate predictions. This is similar to calculating students' grades at the end of the quarter; without tests, we will not know if the students understand the concepts or not. In ML, performance metrics act as these grading periods, which provide insights into how well the model is doing its job. Performance is usually measured by comparing the predicted values to the ground truth after the initial training on the training set to make sure that the training worked, and then again on the validation and test sets to assess its performance. The primary reason to measure a model's performance is to determine its accuracy and its ability to generalize to new data. Measuring performance helps to identify whether a model is underfitting or overfitting and whether further adjustments need to be made. Model performance is also often compared to human baseline performance, as certain tasks are inherently harder than others, so when a model's performance seems to be lacking in these tasks, it might still outperform humans; thus, it is still valuable to save the model and utilize it to a certain extent (Ng et al., n.d.).

Note: The following two paragraphs are highly technical, and there is not really a way to simplify the explanations; the main information is that there are a number of ways to measure performance in ML models for different scenarios.

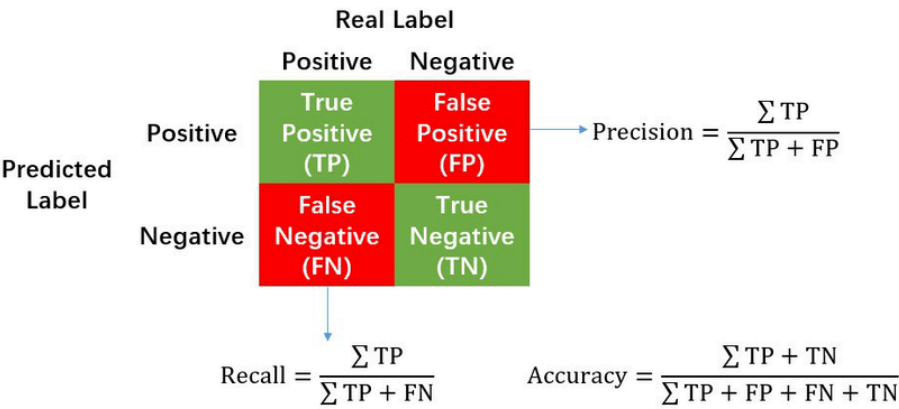
For regression models, common metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). Each of these metrics offers different insights into the model's performance and is used in different scenarios. MSE calculates the average of the squares of the differences between the actual and predicted values, similar to the previously mentioned regression cost function. It gives a higher weight to larger errors due to the squaring of each term, which is not ideal if there are significant outliers (James et al., 2023). RMSE, as the name suggests, is the square root of MSE. Agrawal (2023) also noted that RMSE is on the same scale as the target variable, making it easier to interpret, but like MSE, it gives more weight to larger errors. It is also a common metric used in deep learning. MAE takes the average of the absolute differences between the predicted and actual values. It is ideal when the goal is to treat all errors equally, meaning that, unlike MSE and RMSE, it is not sensitive to outliers. It is also easier to interpret than MSE since it's on the same scale as the original data (Singh, 2022). R^2 quantifies the proportion of variation in the dependent variable that can be explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit. However, it does not indicate whether the regression model is adequate; it only explains how well the model explains the variation in the data. R^2 is useful for a quick insight into how well the input and output are associated, as well as for explaining to non-technical people how much better the model is compared to the average (Gulati, 2023).

Accuracy, precision, recall, F1 score, and area under the receiver operating characteristics (AUC-ROC) are commonly used for measuring classification performance. Accuracy measures the proportion of correctly predicted values out of all instances. It is the simplest performance metric while offering easy-to-understand insights as to how well the model is performing (James et al., 2023). However, it might not be useful if the dataset is imbalanced, meaning that there are significantly more positive ground truths than negative, or vice versa. Precision measures the proportion of true positives (correctly predicted values in the positive class) to all positive predictions (both correct and incorrect predictions that belong in the positive class), which indicates the quality of the positive predictions. It is most useful when the cost of false positives is high, such as in spam detection (where it is important not to label non-spam as spam). Recall is almost the opposite of precision; it measures the proportion of true positives to the total actual positives (the sum of true positives and the ones missed minus false negatives). It is important in cases where missing true positives is costly, like in disease screening. Figure 12 gives a detailed explanation of precision and recall; the figure shown is also known as a confusion matrix. More often than not, however, precision and recall are used together to calculate the harmonic mean (a type of average that finds the balance between two ratios) between the two to form the F1 score. F1 scores are useful to balance precision and recall, especially in datasets that are

imbalanced (Shung, 2020). Without going into a lot of statistical details, AUC-ROC essentially measures how much the model is capable of distinguishing classes, which is similar to the F1 score except that it is less sensitive to imbalance data. However, it is important to note that both the F1 score and AUC-ROC are often used together to understand the model better (Czakon, 2023).

Figure 12

Visual Explanation of Precision and Recall (Also Known as a Confusion Matrix)



Every choice made in the ML process, from data collection and preprocessing to selecting an algorithm and tuning its parameters and hyperparameters, impacts the model's performance. Ensuring that data is clean, selecting relevant features, and ensuring that the model is not too complex or too simple (to avoid overfitting and underfitting) helps a model to be more accurate. In the context of disinformation identification, accurate performance measurement is essential to trust in an ML model's ability to discern true from false information. Each step in the ML pipeline contributes to the model's ability to perform accurately and reliably, which is ultimately reflected in the performance metrics.

Regression Example

This section will give an overview of an example ML project from start to finish. In this example, some developers are trying to predict the reach of an online article based on various features. The first step is to gather data. For this example, the developers might collect data on past articles from different online platforms. The data could include the article's title, the number of times an article was shared, the time of publication, the length of the article, and the number of images included. Once the developers have the raw data, they need to prepare it for analysis, which can be achieved by cleaning the data (such as removing and correcting errors and dealing with missing values), creating new features that might be useful (such as the day of the week the article was published), and normalizing or standardizing the data so that all features are on a similar scale. They will then conduct an EDA to explore the data and gain insights into relationships between each feature and their impact on the label. They might create visualizations such as scatter plots to see the relationship between shares and article length or use correlation matrices to identify which features are most strongly related to article reach. Next, they will divide their data into training, validation, and test sets. The training set is used to teach the model, the validation set is used to fine-tune the model's hyperparameters, and the test set is saved for the final evaluation. For the regression task, the developers might start with a simple linear regression model to set a baseline. They would use the training set to train the model to find relationships between different features and the label. They would then use the validation test to assess the model's performance through metrics such as MSE or RMSE to quantify how close the model's predictions are to the actual reach of the articles. A low MSE or RMSE score would indicate that the

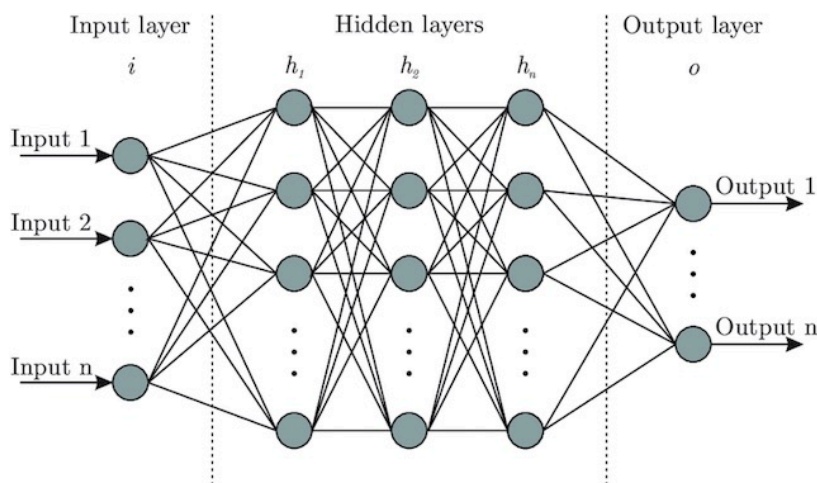
model is doing well at predicting the reach. If the model is not performing well, they would adjust the model's hyperparameters or try using more complex models like polynomial regression to capture more complex relationships. Finally, they would measure the model's performance using the test set through the same metrics they used in the validation set. Now, the developers can feed this ML model with new articles (with the same features they used when training, which are the article's title, the number of times an article was shared, the time of publication, the length of the article, and the number of images included) and ask it to predict the article's reach.

Deep Learning

Deep Learning (DL) is a subset of ML that builds upon the foundational concepts of ML to identify patterns from more complex structures. Unlike ML, which often requires feature extraction and relies on structured data (tables), DL is useful in environments where data is unstructured (meaning that it doesn't come in a table format, like images, audio, and text) and where the relationships within features and labels are not obvious. At its core, DL uses algorithms known as neural networks, which are inspired by the human brain. These networks consist of layers of nodes, or neurons, and each of them is designed to perform specific computations (Jaiswal, 2024). Figure 13 gives a visual representation of a neural network with layers of nodes. As data passes through these layers, the network learns to identify patterns, similar to a human brain. However, as Ng et al. (n.d.) mentioned, as DL evolves, new forms of neural networks have been developed, and they move away from the original brain-like structure. There are various forms of neural networks nowadays that specialize in different forms of data processing.

Figure 13

Visual Representation of a Dense Neural Network with 5 Layers



Deep Learning also extends the principles of regression and classification into more complex models. They can process information from both the bigger picture and the fine details of the data, which allows DL to perform tasks with a level of understanding that ML alone simply cannot achieve. In addition, most of the previously discussed techniques in ML are still used in DL, from data preparation to regression and classification to cost function to gradient descent to training, validation, and test set, to underfitting, overfitting, and regularization, and finally to measuring performance (Neagoie & Bourke, n.d.).

In the context of identifying disinformation, DL is useful as it can process and interpret unstructured text, whereas traditional ML techniques might struggle with the subtleties and complexities in language. It can find hidden patterns and meanings in data, making it a useful tool for identifying disinformation.

Neural Network

A Neural Network is an interconnected web of nodes, similar to the neurons in our brain. Each node in a network represents a tiny processing unit, working together to solve complex problems. The most fundamental type of network, Multilayer Perceptron (MLP), mimics the way human neurons signal to one another, where each connection is a pathway for data to flow from one another and each node processes the data and then passes it to the next layer.

MLP takes input data from the input, transforms it through a series of hidden layers, and outputs the result through the output layer, as shown in Figure 13 above. Each layer consists of multiple nodes, and each node combines input from the previous layer with a set of weights that amplify or dampen that input, which is then added to a bias. Each node in every layer except the output layer is connected to every node in the following layer, meaning that the information from every node is passed to all nodes that it's connected to, which allows the network to consider all aspects of the input data in each layer (Detrat, 2017). For simplicity, the more layers and the more nodes are in the hidden layers, the more complex the model is, meaning that it can perform more complicated tasks. For the input layer, the number of nodes is determined by the number of features. Continuing with the regression example, if the features of the article's title, the number of times an article was shared, the time of publication, the length of the article, the number of images, and the day of the week the article was published, the input layer would consist of six nodes, each node for each feature. For the output layer, the number of nodes is determined by the task that the model is trying to achieve (Sanderson, 2017a). For example, in a regression task like predicting a house price, there would only be one node since that is the numerical output of the house price. For binary tasks like identifying whether a piece of news is real or fake, there would be two nodes, where one would activate, signifying the output of the binary classification. For multiclass classification, like identifying the type of cat, there would be as many nodes as the number of types of cats. The corresponding node will activate if it is the said type of cat; otherwise, it will not.

Forward Propagation

Forward propagation (also called the forward pass) is the process that allows neural networks to make sense of the input data. Data flows from the input layer through the hidden layers and finally to the output layer. At each neuron, the data is processed by weighting the sums of the inputs from the previous nodes, and then an activation function is applied to this sum to determine the neuron's output. The exact values for inputs fed into the input layer are randomized, just like gradient descent discussed in ML (Starmer, 2020a).

Activation Function

Activation functions are located in every hidden layer and the output layer. The functions in the hidden layers are used to determine whether a neuron should be activated and to what extent, which influences each node's output to the next layer. They are crucial since they introduce non-linear attributes to the network. Without these functions, a neural network would essentially perform as a linear regression model, no matter how many layers it had. This would limit the network to only being able to solve problems that can be solved with a straight line through the data. The most popular activation functions used for hidden layers are rectified linear unit (ReLU) and hyperbolic tangent (Tanh). ReLU outputs the input directly if it is positive, otherwise, it outputs 0. It is popular since it allows for faster training and mitigates gradient issues that other functions run into (Starmer, 2020b). Refer to Figure 14 for the graph of ReLU. Tanh is similar to the sigmoid function discussed previously in ML, except that it, similar to ReLU, mitigates gradient issues that the sigmoid runs into. Just like sigmoid, Tanh is only used for classification tasks, but ReLU can be used for both regression and classification tasks. Refer to Figure 15 for the graph of Tanh. However, since tanh requires the computation of exponents, it is much slower than ReLU, meaning that ReLU is still by far the most used function among developers (Sharma, 2017).

Figure 14

Graph of a ReLU Activation Function

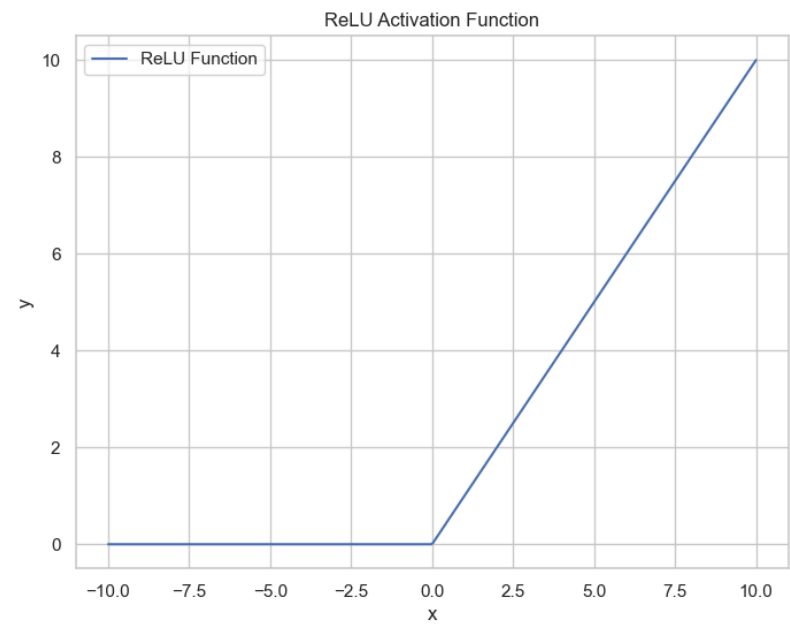
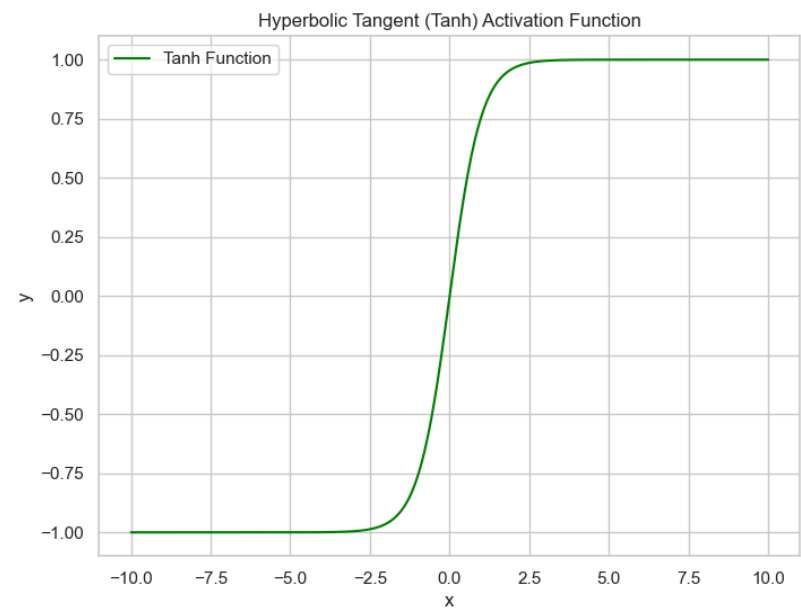


Figure 15

Graph of a Tanh Activation Function



The choice of activation functions in the output layer depends on the task the model is trying to perform. For regression tasks, a linear activation function is often used as it just sums the information from the previous layer without changing any information. For binary classification, a sigmoid function is often used (here sigmoid is used over Tanh since it has an output range between 0 and 1 rather than -1 and 1). For multi-class classification, a softmax function is used to produce a probability distribution across multiple classes. Softmax is similar to sigmoid, except that it can output multiple classes (Brownlee, 2021).

Backpropagation

Backpropagation is the training process where the network learns from its errors, similar to gradient descent in ML. After the forward pass, the network calculates the loss with a loss function like MSE for regression, binary cross entropy for binary classification, or categorical cross entropy for multiclass classification, similar to calculating the cost in ML (Bourke, 2022). During backpropagation, this calculated loss is then fed back through the network with chain rule (an important concept in differential calculus), and the weights are adjusted to reduce the loss. Activation functions play a key role here because their derivatives are used to calculate the gradient of the loss with respect to each weight (Sanderson, 2017b). It is similar to a teacher adjusting their teaching methods based on how well students perform on assessments.

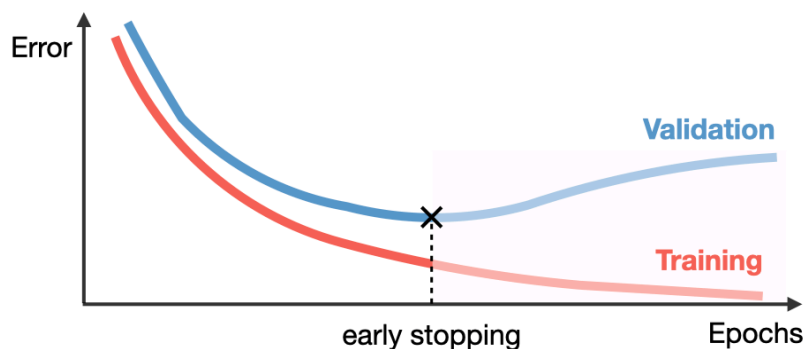
Batch, Epoch, Early Stopping, and Dropout

Similar to repeating gradient descent in ML for multiple iterations until the loss is minimal, deep learning also repeats the process of forward pass and backpropagation multiple times.

Batch. A batch is a subset of the data that is fed through the network. Instead of giving the model the entire training set at once, which can be computationally intensive, the data is divided into smaller chunks. Each chunk is fed through the network front and back and then repeated for each batch. The size of each batch, also known as the “batch size”, is a hyperparameter that can also be configured. Too small of a batch size can sometimes lead to overfitting since they tend to have more variance in their representation of the overall dataset, which can cause the model to make updates that are too specific. However, it uses much less computational power compared to larger batch sizes. On the other hand, larger batch sizes can lead to underfitting since they might smooth out the learning too much, which causes them to not capture the subtleties in the data (Géron, 2019, pp. 15, 326).

Epoch. An epoch is one complete pass of the entire dataset. It is essentially a collection of batches that goes through all the points in the dataset. During an epoch, the training is divided into batches, and the network iterates over these batches. After that, it validates the results of the training with the validation set. The number of epochs is also a hyperparameter. More epochs mean more iterations, which can lead to better-performing models, but it can also cause the model to overfit as it essentially memorizes specific features of the data rather than patterns that can be generalized. Less epochs can lead to underfitting as the model simply has not learned enough patterns yet (Géron, 2019, p. 290).

Early Stopping. Early stopping is often used to combat overfitting. It involves stopping the training process early when the validation performance stops increasing at a certain threshold. Hyperparameters such as the patience (how many epochs to wait for before early stopping occurs), the metrics type (such as the validation accuracy, validation loss, or validation F1 score), and the threshold (how much of the chosen metrics scores does the model not improve before early stopping occurs) can all be configured. Early stopping is a form of regularization, which, as previously discussed, ensures that the model doesn't overtrain (Géron, 2019, p. 141).

Figure 16*An Example of Early Stopping*

Dropout. Dropout is also another form of regularization that randomly selects neurons that are “dropped out” from the network, meaning that the selected neurons are temporarily ignored on both the forward pass and backpropagation. This technique helps neural networks to not be overly reliant on one specific feature, which helps with generalization and overfitting (Géron, 2019, pp. 365–368).

Natural Language Processing

Natural Language Processing (NLP) is the process that allows computers to analyze, understand, and derive meaning from language. It uses algorithms to identify and extract language rules so that the unstructured language data is converted into a form that computers can understand. Originally, NLP and deep learning were two separate entities with little association with one another. It was originally based on just set rules to extract meanings from them. The first prominent example of NLP was in the ELIZA program, a rule-based chatbot that essentially parrots the user (Jurafsky & Martin, 2023, p. 4). See Figure 17 for an example ELIZA conversation. As NLP evolved, however, it began to tie into machine learning and eventually deep learning. NLP is the technology that provides tools to identify disinformation; ML models can be trained to recognize patterns and anomalies in text that may indicate fake news.

Figure 17*Example of an ELIZA Conversation*

```
User: I am unhappy.
ELIZA: DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
User: I need some help, that much seems certain.
ELIZA: WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
User: Perhaps I could learn to get along with my mother.
ELIZA: TELL ME MORE ABOUT YOUR FAMILY
User: My mother takes care of me.
ELIZA: WHO ELSE IN YOU FAMILY TAKES CARE OF YOU
User: My father.
ELIZA: YOUR FATHER
User: You are like my father in some ways.
```

Since the human language is not only very complex from a grammar standpoint but also the lurking variables such as different tonalities, idioms, regional dialects, and new slang, developing NLP

systems that can detect subtle differences is a challenge. Therefore, data preparation needs to be taken with extra care as compared to typical ML tasks.

Text Preprocessing

Text preprocessing prepares the raw text for analysis and model training. This process is the equivalent of data preprocessing in ML, and similarly, the quality of the preprocessing will directly impact the quality of the model's output. It involves the following steps: initial text cleanup, tokenization, stop word removal, stemming and lemmatization, sequencing, and padding (Deepanshi, 2023).

Initial Text Cleanup. Initial text cleanup removes unnecessary elements from the text, such as HTML tags, URLs, special characters, numbers, and other non-linguistic elements. The goal is to strip the text down to normal language void of extraneous values. Since scraping text from the web usually results in ugly data, this text has to be cleaned to an extent; it is similar to cleaning the valuable items you found in a scrapyards (Deepanshi, 2023).

Tokenization. Tokenization is the process of breaking down text into smaller units called tokens. A token is typically a word, but it can also be a sentence in rare scenarios. The primary purpose of tokenization is to simplify the text data into pieces for analysis and processing. Word tokenization is the most common form, where each word is separated into its own "token", with punctuation marks receiving their own tokens (Jurafsky & Martin, 2023, p. 63).

Stop Word Removal. Stop words are common words like "the", "is", "in", and "which". These words often do not carry significant meaning and are used almost everywhere in language. Removing these words can help reduce the size of the data and focus on more meaningful words. Usually, this process is done by comparing against a predefined list of stop words, which can vary depending on the language and usage (Khanna, 2021).

Stemming and Lemmatization. Both stemming and lemmatization aim to reduce words to their base or root form. Stemming cuts off the prefixes and suffixes based on rules, such as simplifying "fishing", "fished", and "fisher" all to the word "fish". Lemmatization reduces words to their base form called the "lemma". Unlike stemming, lemmatization has contextual awareness, meaning that it considers the context and POS of a word. For example, "saw" would be lemmatized to "see" if it's used as a verb but remains as "saw" if it is a noun. However, the downside to lemmatization is that it is much slower than stemming as slicing pieces of words off is much faster than finding their roots (Jurafsky & Martin, 2023, pp. 5, 22–23).

Sequencing. Sequencing converts the tokens into a numerical value. The conversion is based on a word index, where frequent words are given lower integers, and less frequent words are assigned higher integers. This process is important as it transforms the textual data into a format that can be understood by the neural network (Moroney, 2020).

Padding. Padding is the final step of preprocessing. It essentially adds or truncates tokens so that every piece of text is the same length. This step is necessary as neural networks usually expect a certain number of input parameters (Moroney, 2020).

Basic Text Representation Techniques

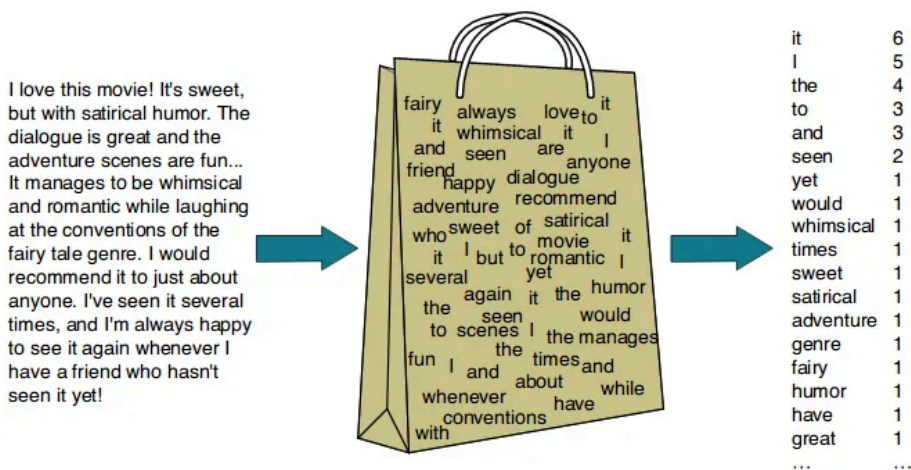
Even after cleaning up the text data, these texts still need to be transformed into a format that computers can understand. In basic ML, this is usually achieved by using NumPy to convert table formats into matrices. However, text is unstructured, so a different technique is used—text representation converts text data into a format that can be understood and utilized by machine learning

algorithms (Sarguroh, 2023). These techniques allow textual data to be interpreted and processed by machine learning models. In the context of identifying disinformation, how text is represented can significantly influence the effectiveness of the detection process. The goal is to represent text in a numerical format while preserving as much of the original information as possible. Some common techniques include Bag-of-Words (BoW) and N-grams.

Bag-of-Words (BoW). The Bag-of-Words model is one of the simplest forms of text representation. It involves representing text data as a bag (set) of words without considering grammar and word order. It essentially throws all the words it sees into a bag and then counts the number of occurrences. In more technical terms, it creates a vocabulary of all the unique words in the text and represents each document as a vector with counts of how often each word appears. Refer to Figure 18 for a visual representation of BoW. In disinformation detection, BoW can help in identifying frequently used words in misleading news versus credible ones. However, it disregards the order of words, meaning that it cannot capture the full context of sentences (Neubig, 2021).

Figure 18

Visual Representation of How Bag-of-Words Works



N-gram. An n-gram model considers a sequence of n items (words or characters) from the text. It is an extension of the BoW model that takes into account the context to a certain extent by capturing word co-occurrences. For example, a bigram (2-gram) model would look at pairs of consecutive words. Refer to Figure 19 for a visual representation of n-gram. In the figure, unigram groups word by one, bigram groups word by two, trigram groups word by three, and so on. Each of these groups is then thrown into a bag and counted for the number of occurrences. Using a larger n captures more contextual understanding of words and sentences but can lead to overfitting as the same sequence of words that are in the training set may never appear in the testing data. N-gram can be more effective in understanding language patterns compared to BoW, which is essential in identifying nuanced disinformation strategies (Jurafsky & Martin, 2023, pp. 31–37).

Figure 19

Visual Representation of Unigram, Bigram, and Trigram Models



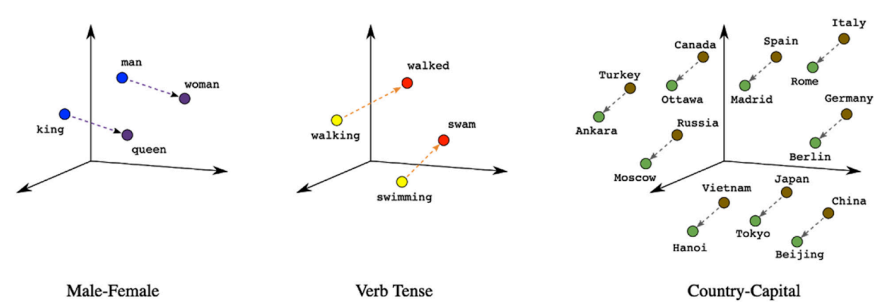
Advanced Text Representation Techniques

More advanced representation techniques are used to capture more details and intricacies of language, which basic methods like BoW and n-gram might miss. This is especially important in applications like identifying disinformation, where understanding context, sentiment, and subtle linguistic cues is essential. They recognize more hidden patterns within the language and understand the context of each word better, which helps models make more accurate and meaningful interpretations of words and sentences.

Word embeddings essentially map each word in a language to a unique, multi-dimensional point in space. Words with similar meanings or used in similar contexts are placed closer together in this space. This proximity captures their relationship in a way that a computer can understand and use. Embeddings convert high-dimensional data (like BoW) into a lower-dimensional space (essentially fewer layers of matrices). This conversion makes the data easier to process and analyze for machine learning models (Jurafsky & Martin, 2023, p. 103). Figure 20 provides a simplified visual representation of word embeddings. There are two common word embedding techniques: Word to Vector (Word2Vec) and Global Vectors for Word Representation (GLoVe).

Figure 20

A Simplified Visual Representation of Word Embeddings



Word2Vec, developed by a team at Google, is one of the most popular word embedding techniques. It involves two methods: Continuous Bag of Words (CBow), which uses a neural network to predict a target word based on its surrounding context, and Skip-Gram, which predicts the surrounding context given a word. Word2Vec models are trained using a large corpus of text, and once it is trained, it can detect synonymous words or suggest additional words for a partial sentence (Jurafsky & Martin, 2023, pp. 119–124). In the context of disinformation detection, using Word2Vec can help understand the contextual usage of words in different types of articles, which can help distinguish between genuine and misleading news.

GloVe, created by Stanford University, is another method for generating word embeddings. It is different from Word2Vec in the sense that it not only looks at the local context of words but also at their overall occurrence across the entire text. GloVe is useful at capturing both the general use of a word and how it's used in specific contexts, which provides a detailed word representation. It builds a matrix (called the co-occurrence matrix) where each cell value represents how frequently two words (one represented by the row and the other by the column) appear together in the text. For example, if 'coffee' and 'milk' are often seen together, their corresponding cell in the matrix will have a high value. The goal of GloVe is to learn a word vector (a list of numbers) for each word. These vectors are created in a way that their relationship with each other is similar to the relationship of the words in the co-occurrence matrix. Essentially, words that appear in similar contexts will have similar word vectors (Chakraborty & Roy, 2022).

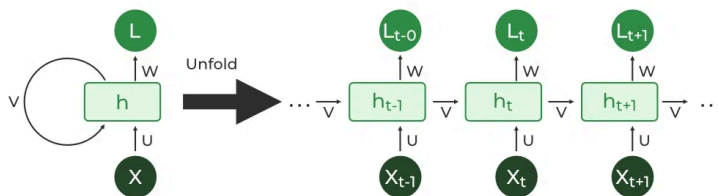
Advancements in Neural Networks

Before deep learning was developed to the current state, after text preprocessing and text representation, the cleaned-up and transformed text was fed into basic ML models like logistic regression for classification. However, basic ML models usually cannot capture the full context of words. Words in a sentence are not just a collection of independent entities; their meaning and interpretation largely depend on the preceding and following words. Hence, neural networks, specifically recurrent neural networks (RNN) and long short-term memory (LSTM), are now commonly used for NLP ("Language Models, Explained: How GPT and Other Models Work," 2023).

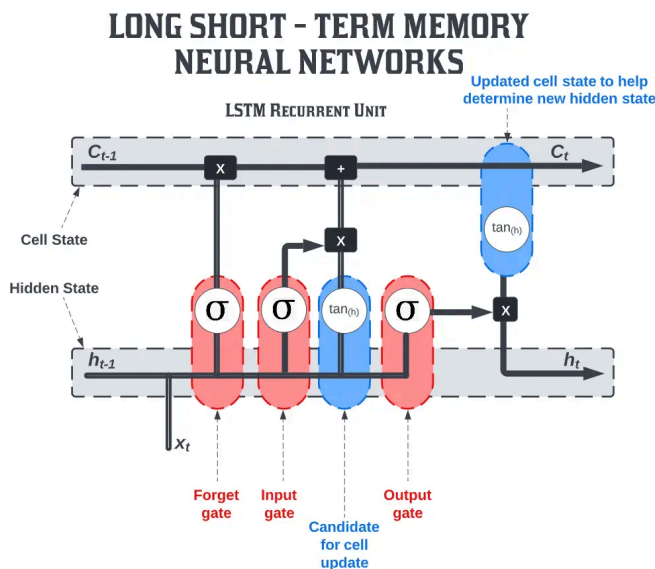
Recurrent Neural Network (RNN). Recurrent neural networks are a type of neural network specifically made to handle sequential data. In NLP, where the sequence of words is crucial in understanding context and meaning, RNNs have become a fundamental tool. Unlike traditional neural networks, where data flows in a straight line from input to output across layers, RNNs process information in a sequence by taking into account both the current and prior inputs. This sequential processing is possible since RNNs have connections that allow each cell to loop back on itself (Donges, 2024). Each "cell" (equivalent to a node in traditional neural networks) in an RNN layer acts as a mini-memory unit, and it updates its hidden state (a value that is essentially RNN's memory) with each time step (essentially the i th pass on itself) in the sequence. It loops back on itself until the end of a sequence, such as the end of a sentence, and then it passes the information onto the next layer. Figure 21 gives a visual representation of RNN. As an RNN processes a sequence, each input is tackled one at a time. For every new input (such as words), a cell in the RNN updates its hidden state. Then, it passes the hidden state along with the input to the next layer of cells, which would accomplish a similar task, except that it is done at a higher level, such as from understanding basic syntactic structure to more abstract semantic meaning (Nabi, 2019). Every cell in a layer does the same thing as the other cells in the same layer, but it may capture different aspects and features of the input sequence, similar to a basic feedforward neural network. This sequential processing allows RNNs to maintain the memory of previous data in the sequence, making them ideal for tasks like language modeling and text generation. However, RNNs often run into long-term dependencies, meaning that it is hard for RNNs to remember and utilize information from earlier in the sequence as the distance increases, which leads to difficulties in learning and remembering information over long text sequences (Starmer, 2022a).

Figure 21

Simple Representation of RNN Consisting of One Cell Feeding Itself the Hidden State



Long Short-Term Memory (LSTM). LSTMs, while being a type of RNN, were developed to solve the long-term dependency problems in RNNs. At the core of LSTMs is the cell state, which is similar to a conveyor belt and runs through the entire chain of the network. This cell state is important for maintaining information over extended periods, acting as the network's long-term memory. While the cell state acts as the long-term memory, the hidden state, similar to traditional RNNs, is a short-term memory that is used to make predictions and is passed to the next time step and the next layer. Similar to RNN, each layer in an LSTM network contains a certain number of cells, a hyperparameter that can be adjusted based on the complexity. However, as Baeldung (2023) noted, these cells have specialized structures known as gates: the input gate, the forget gate, and the output gate, which are used to regulate the flow of information in and out of the cell state. The input gate determines how much of the new information to store in the cell state, the forget gate decides what information to discard from the cell state, and the output gate determines what the next output should be, based on the current input and the memory of the cell state. In each LSTM cell, a "candidate" for updating the cell state is created. This candidate is formed by combining the current input with the previous hidden state, typically using a non-linear activation function like tanh; the candidate represents a potential new value for the cell state (Thakur, 2018). The actual update to the cell state then combines this candidate with the effects of the input and forget gates. This intricate mechanism allows the cell state to be updated in a way that retains important long-term information while introducing relevant new data (Starmer, 2022b). Figure 22 gives a visual representation of LSTM. LSTMs have become widely used due to their ability to understand and retain information over long sequences. In the context of this project, LSTM provides the necessary depth to analyze and interpret language where the differences between the tonality of real and fake news are extremely subtle.

Figure 22*Simple Representation of a Cell in LSTM*

Transformers

After the development of advanced neural networks like RNNs and LSTMs, the field of NLP was transformed significantly with the introduction of transformers. A transformer is a type of neural network architecture introduced by Google in a paper titled "Attention Is All You Need" by Vaswani et al. (2017). It is different from the sequential processing of RNNs and LSTMs and instead uses a mechanism known as "attention" to process all parts of the input data simultaneously, which makes processing text more efficient and effective, especially for longer sequences. Even though the original purpose was for translating languages, the transformer architecture was soon adopted for more NLP tasks, including sentiment analysis and text generation (Saleem, n.d.). Prominent transformers such as BERT, GPT, and LLaMA were developed, and they shaped the landscape of generative AI today.

Generative Pre-trained Transformers (GPT). GPT models are developed by OpenAI, with GPT-1, the first GPT model, released in June 2018. GPT is an autoregressive model, meaning it generates text by predicting the next word in a sequence given all the previous words through a probabilistic model. It is useful in tasks that require language generation, such as creating text based on a given prompt (OpenAI, 2022). Through countless improvements and versions over the years, GPT has become one of the most prominent examples of text-generative AI today, with ChatGPT amassing over 100 million users (Malik, 2023).

Bidirectional Encoder Representations from Transformers (BERT). BERT was developed by Google in October 2018 with the main purpose of understanding the context of a word in a sentence, as compared to GPT's text generation. Some prominent examples include sentiment analysis, question answering, text summarization, and filling in the blanks (fill-mask). It revolutionized the idea of transformers at the time with its bidirectional training approach, meaning that it could gain context from both the left and right of the word, which was unlike the original Transformer and GPT. This bidirectional approach allows BERT to gain a deeper understanding of language context and nuances, which made it better at doing the aforementioned tasks than GPT at the time of either's release. BERT can also be fine-tuned, where they are further trained for specific tasks, such as identifying disinformation in the context of this project. There are also various versions of BERT, which are optimized for different tasks

and environments; some prominent examples include ALBERT, RoBERTa, and DistilBERT (Muller, 2022).

Large Language Model Meta AI (LLaMA). LLaMA (or Llama) is a more recent development in the realm of transformer-based models developed by Meta. It's designed to be efficient while maintaining or even surpassing the performance of models like GPT-4. Similar to GPT, it is used in a variety of language generation and understanding tasks, specifically in environments where the balance between performance and computational efficiency is important (Meta, 2023). It gained a lot of attention for being a responsible and safe model and “is available for free for research and commercial use” (*Introducing Llama 2*, n.d.).

These transformers are usually pre-trained, meaning that they are ready to be used by the users for inferences without much modification. These transformers' efficiency and capabilities are also determined by the number of parameters. Parameters in transformers are similar to those in neural networks, such as weights and biases—they are essentially “settings” of the transformer. The number of parameters is a key factor in determining the model's ability to process and generate language, with larger models generally performing better. However, the larger models require more computational power and resources to train and run. For context, the default BERT has around 110 million parameters, Llama-2 comes in different sizes ranging from 7 billion to 70 billion, and GPT-4 reportedly has 1.7 trillion parameters (Muller, 2022; Touvron et al., 2023; Schreiner, 2023).

Past Works

In this final section, I will discuss previous research on the field of using AI to identify disinformation. After a long explanation of ML, DL, and NLP techniques above, this section serves to analyze some of the studies related to identifying disinformation with AI.

Human Benchmark

A study by Spezzano et al. (2021) compared the accuracy of humans and AI in identifying fake news under different conditions (news excerpt only; title and image; title and bias; title, image, and bias). Spezzano et al. found that the AI outperformed the humans statistically significantly across all four conditions. For the interest of this senior project, the news excerpt has an AI accuracy of 71% and a human accuracy of 53.3%. The AI in this study utilized only fundamental NLP techniques, such as extracting the average word count per sentence and rate of misspelling, which was analyzed using a simple logistic regression model. The respondents of this study have also revealed that the perception of professionalism and subjectivity in the excerpt influenced their judgment. It is also important to note that this model did not utilize a holdout set, meaning that the model might not perform well on other news sources.

Other Studies

Canale and Messina (2023) utilized both fine-tuned and non-fine-tuned large language models (LLMs) to identify disinformation in Italian news sources. The non-fine-tuned models were tested against three datasets (PagellaPolitica1, FEVER-it-small, and MinistryHealth), and the average of their accuracy was taken. Most prominently, GPT-4 performed the best, with an average accuracy of 78.15%. For fine-tuned models, DistilBERT trained on the combination of PagellaPolitica1 and FEVER-it-small and inferenced on MinistryHealth, performed the best, with an accuracy of 93.51%. However, the average accuracy for this model is only 67.18% due to their lower scores on the datasets that they were trained on. The authors did not give any explanations for this unexpected behavior.

Many other studies, such as ones from Chauhan and Palivela (2021), Azizah et al. (2023), and Mattern et al. (2021), all produced models that achieved excellent results, scoring 99.88%, 87.6%, and 98.1% accuracy scores, respectively. However, all three studies did not utilize any holdout sets and simply accessed the models from the same dataset that the training set was subset on, which may introduce biases that help increase the models' performance as they might have already seen and

memorized certain words or phrases that helps identify whether a piece of news is real or fake. This aspect will be discussed in more detail under the methodology section.

Methodology

For my research, I utilized different deep learning models and architectures to gauge each model's accuracy at detecting disinformation and compare the results between one another as well as the baseline found in DIS. I first sent an initial survey to DIS high school students to gauge their understanding of artificial intelligence and draw a baseline for their accuracy in identifying disinformation. I then built and utilized several deep learning models that detect whether a piece of information is real or fake. Finally, grade 9 students in CS classes completed five quizzes with five questions each on identifying whether a piece of news is real or fake; this process was used to collect more data points.

Initial Survey

I sent an initial survey (see Appendix A) through Google Forms to all high school students at DIS to gain insights into people's understanding of AI as well as to establish a baseline for DIS high school students' accuracy at identifying disinformation for ML models to compare against. In total, 101 respondents filled out the survey. The survey is split into three sections—getting the background of the respondent, gauging their understanding of AI, and identifying whether five pieces of media are real or fake. However, it is important to note that nonresponse bias may be present as it is a voluntary-based survey (with students later being pushed to fill it out), meaning that the data should be interpreted with caution.

General Background

This section delves into understanding students' basic background, which includes their grade level and whether they know what AI is.

Grade Level. Of the 101 high school students from DIS, grades 9 and 11 account for 20.79% each, with grade 10 accounting for 27.72% and grade 12 at 30.69% of the total sample (see Figure 23.1). In this survey discussion, both the entire sample and individual grade levels are analyzed.

Do You Know What Artificial Intelligence (AI) Is? From Figure 23.2, 96.04% of the respondents know what AI is. However, upon looking at the data points, one of the respondents certainly engaged in non-serious responding. In addition, given the school environment, it is more likely that all respondents know what AI is. Figure 23.3 gives a more detailed breakdown of the understanding of AI across different grade levels—two respondents from grade 9 and one respondent from grades 11 and 12 each responded “No” when asked the question. Given that answering “No” to this question skips the next section on the detailed understanding of AI, respondents may have answered “No” to skip over the section.

Understanding of AI

This section is only displayed to the 97 participants who answered “Yes” to “Do you know what artificial intelligence (AI) is?”. This section aims to gauge the level of understanding of AI among DIS high school students.

Which of The Following AI Tools Have You Used? This question is a multi-select question. Of the 97 responses, 95 responded that they had used ChatGPT, which is by far the most used AI tool. As shown in Figure 24.1, the difference between ChatGPT and the rest of the AI tools is extremely significant—the second most used tool is Bing AI with only 19 respondents. Diving deeper into each grade level, Figure 24.2 and Figure 24.3 show that grade 10 students have the highest proportion of AI

tool utilization across most AI tools (with the exception of Auto-GPT). Grades 11 and 12 students also utilized various tools, albeit at a lesser degree as compared to that of grade 10. Grade 9 students by far utilize AI tools the least across most AI tools, with multiple tools getting 0 responses.

Do You Think Generative Text Models (Like ChatGPT) Have the Ability to Think? 44.33% of the 97 respondents answered “Yes” to this question, as shown in Figure 24.4. As shown by the trend across grade levels in Figure 24.5, as the grade level increases, the proportion of students who think text models have the ability to think decreases (63.16% in grade 9 and only 26.67% in grade 12). As mentioned in the secondary research, generative text models do not have the ability to think. This trend suggests that older students generally have a better understanding of how generative models work.

Do You Think These Models are Sentient (Meaning That They Have the Ability to Experience Feelings)? Fewer respondents believe that this is the case, with only 12.37% of respondents answering “Yes” to the question (see Figure 24.6). Similarly to the previous question, there is also the same trend—as grade level increases, fewer respondents tend to believe that AI models are sentient (21.05% in grade 9 and 3.33% in grade 12), which also suggests a better understanding of how generative models work.

How Do You Think These Models Work? This question is a multiple-choice question; however, respondents may also choose to enter a custom response if they choose to. As shown in Figure 24.8, of the 97 respondents, 65.98% believe that the choice “They use a complex algorithm that learns from human behavior over time” is correct, and 21.65% believe that the choice “They are programmed with a fixed set of responses for every query” is correct. However, both of these statements do not reflect the true reality of AI. The correct response, “They use probability to gauge the next probable word choice”, is only chosen by 4.12% of respondents. Custom responses also generally do not reflect a true statement of how generative text models work. Figure 24.9 gives a deeper breakdown of responses across grade levels. Contradicting the two previous questions, grades 10 and 11 are the only grade levels that have respondents who selected the correct statement, with proportions of 7.14% and 10.00%, respectively.

Five-Question Quiz

All participants, regardless of the answer they picked for “Do you know what artificial intelligence (AI) is”, are required to fill out the five-question quiz. Respondents are given five passages, one after another, for them to indicate whether the piece of media is real or fake. The media pieces are pooled from multiple datasets and then manually picked to ensure that the passage is not too overly long. Namely, the datasets used are ones from Bharadwaj et al. (2020) and Verma et al. (2021). Since some datasets already come stemmed or lammetized, I first try to find the source of the original media to get the original text; however, if no media can be found, I modified them to be as close as possible to the original passage. This section serves as a human benchmark for AI models to compare against.

Question 1. Respondents were given the following passage:

A British man, accused of posting a picture of four-year-old Prince George and the address of his London school as part of a series of possible targets for Islamist militants, will face trial from April 30, a London court said on Wednesday. Husnain Rashid, 31, is accused of posting information on the Telegram messaging service to encourage jihadis to carry out attacks, along with information to help them with possible targets such as sports venues. He appeared at London's Old Bailey central criminal court charged with preparing acts of terrorism, and was remanded in custody. His four-week trial will take place at Woolwich Crown Court in south London.

The passage is real; 50.50% of respondents thought the passage is real, and 49.50% thought the passage is fake, as shown in Figure 25.1. Figure 25.2 also suggests an increasing trend of correctly

identifying the passage as the grade level increases, with only 47.62% of grade 9 students labeling the media as real as compared to the 61.29% correct labeling of the grade 12 students.

Question 2. Respondents were given the following passage:

Harvey Weinstein was among a host of Hollywood actors and producers that donated \$10,000 to former President Bill Clinton's defense fund during his 1998 impeachment trials. The Washington Post published a story in 1998 reporting Harvey Weinstein as one of 62 donors giving the maximum \$10,000. Weinstein donated along with a number of other notable names, like Tom Hanks, Barbra Streisand, Michael Douglas, Ron Howard, Norman Lear and Steven Spielberg. WaPo also reported that Clinton's defense fund collected \$2.2 million in its first six months, tapping into the resentment against the attorney that investigated Clinton as to whether he committed perjury or obstructed justice in a sexual assault lawsuit.

The passage is fake; 38.61% of all respondents correctly identified the media as fake, as shown in Figure 25.3. Similar to question 1, the trend of correctly identifying the legitimacy of the media increases as the grade level increases, with 23.81% and 51.61% correctly identifying the news for grade 9 and grade 12 students, respectively (see Figure 25.4).

Question 3. Respondents were given the following passage:

President Donald Trump celebrated the news that 227,000 jobs were created in January, meeting with his Economic Advisory Council at the White House on Friday morning. "Great spirit in the country right now," Trump said, speaking to the press. "So we're very happy about that. I think that it's going to continue big league. We're bringing back jobs. We're bringing down your taxes. We're getting rid of regulations." Trump promised "exciting times" ahead, previewing a tax bill "soon" and a healthcare bill "even sooner." "Amazingly, it's never happened before that politics has become a much bigger subject than the Super Bowl," he said. "This is usually Super Bowl territory." Trump told the business leaders that he wanted to hear from them as he tackled more reform proposals to make their businesses better, including getting rid of the cumbersome legislation.

The passage is real; 57.43% of all respondents correctly identified the media as real, as shown in Figure 25.5. A similar trend as the above two questions follows; however, grade 11 this time outperforms grade 12 by 5.38%, as shown in Figure 25.6.

Question 4. Respondents were given the following passage:

A Baltimore man was sentenced to four years behind bars for illegally trafficking food stamps. U.S. District Judge Richard D. Bennett sentenced Mohammad Shafiq, 51, to four years in prison and three years of supervised release after his sentence ends, in addition to making him pay \$3.7 million in restitution, the Baltimore Sun reported. Shafiq was the latest defendant to be sentenced in a series of prosecutions of 14 retailers in the Baltimore area. A federal grand jury indicted the retailers in August 2016 for food stamp and wire fraud. The 14 retailers stole a total of \$16 million from the U.S. Department of Agriculture by illegally exchanging food stamps for cash, according to the indictment. Twelve out the 14 defendants pleaded guilty, and two were sentenced this week, according to the Sun. Multiple stores across the country have been cited for millions of dollars in food stamp fraud. Investigators found more than \$20 million worth of food stamp fraud at retailers in Florida, and 140 stores in Chicago were cited for food stamp fraud.

The passage is real; 54.46% of all respondents correctly identified the media as real, as shown in Figure 25.7. However, for this passage, the grade 11 students outperformed the other grades, with the section

scoring an 80.95% accuracy, while the next closest group, grade 10, only has an accuracy of 50.00% (see Figure 25.8).

Question 5. Respondents were given the following passage:
The State Department misplaced and lost some \$6 billion due to the improper filing of contracts during the past six years, mainly during the tenure of former Secretary of State Hilary Clinton, according to a newly released Inspector General report. The \$6 billion in unaccounted funds poses a significant financial risk and demonstrates a lack of internal control over the Department's contract actions, according to the report. The alert, originally sent on March 20 and just released this week, warns that the missing contracting funds could expose the department to substantial financial losses. The report centered on State Department contracts worth more than \$6 billion in which contract files were incomplete or could not be located at all, according to the alert. The failure to maintain contract files adequately creates significant financial risk and demonstrates a lack of internal control over the Department's contract actions, the alert states.

The passage is fake; 55.45% of all respondents correctly identified the media as fake, as shown in Figure 25.9. Grades 11 and 12's accuracy scores of 76.19% and 67.74%, respectively, outperformed the 52.38% accurate grade 9, with grade 10's 28.57% significantly behind (see Figure 25.10).

Score Distribution. As shown in Figure 25.11 (histogram) and Figure 25.12 (box plot), the score distribution of all grade levels is unimodal and ever-so-slightly skewed to the left. The median of the distribution is 3 points, the mean is 2.564, and the mode is 3, which also suggests a slight skew to the left. With a Q1 of 2 and Q3 of 3, the interquartile range (IQR) is thus $3 - 2 = 1$ point, with a range of $5 - 0 = 5$ and a standard deviation of 1.252. As shown in Figure 25.12 with the calculation of $Q3 + 1.5 \cdot IQR$ and $Q1 - 1.5 \cdot IQR$, data points with scores 0 and 5 are all considered outliers. For individual grade level breakdowns, the mean, median, Q1, and Q3 of grades 9 and 10 are lower than that of grade 12, which may suggest that grade 11 and 12 students tend to perform better at identifying disinformation across the five questions (see Figure 25.13, Figure 25.14, and Table 3).

Table 3

Descriptive Statistics for the Five-Question Quiz Score

Grade	Median	Mean	Mode	Q1	Q3	IQR	SD	Range
Overall	3	2.564	3	2	3	1	1.252	5
9	2	2.143	3	1	3	2	1.108	4
10	2	2.179	1	1	3	2	1.335	5
11	3	3.000	3	2	4	2	1.095	4
12	3	2.903	3	2	4	2	1.221	4

Model Development and Utilization

For this project, I developed two models—LSTM and LSTM with GloVe embeddings, fine-tuned one—DistilBERT, and utilized two pre-trained LLMs—Llama-2 and GPT-4 Turbo, to assess each model's ability to identify disinformation. The models I built and fine-tuned were all trained on the WELFake dataset, where Verma et al. (2021) compiled the dataset from other well-known fake news datasets. There are a total of 72,134 news articles, 35,028 of which are real and 37,106 are fake. The only

columns utilized from the dataset are the title, text, and label. The title and text columns are combined and then preprocessed by removing specific words that might indicate the legitimacy of the media (such as “Reuters”) with the help of Regex. They are then lemmatized with the SpaCy library and split into train, test, and validate sets with a breakdown of 60%, 20%, and 20%, respectively. Each set is then tokenized with a maximum length of 500 tokens through the Keras library. Alongside the label as ground truth, they are fed to the models for training through PyTorch. The models were then tested on the test set from the same dataset and cross-inferenced on a holdout set compiled by Bharadwaj et al. (2020). The holdout set also underwent the same preprocessing treatment. The below sections detail the specifics of each model; for the source code of each model, please refer to Appendix B.

LSTM

The text, after preprocessing, is embedded into a 128-dimensional vector (first layer of the model). The data is then passed through 2 LSTM layers with 256 nodes each, and it passes through a linear layer with 1 node with a sigmoid activation function, which indicates the probability that the news is real. Throughout training, it uses a binary cross-entropy loss function to adjust the weights and biases. It is trained with a batch size of 32 and a learning rate of 0.001 throughout 10 epochs; however, early stopping based on validation accuracy is implemented, meaning that if the validation accuracy doesn't improve by the threshold set at 1% after a number of epochs (set as 3), the model will stop training. A dropout of 0.2 is also implemented to reduce overfitting. After the training process, the best model is saved based on the best validation accuracy score.

After several iterations of training and testing the model, the best one has an accuracy of 0.9665, an F1 score of 0.9685, and an AUC-ROC of 0.9938 on the test set. Early stopping was activated at epoch 9, and that also happens to be the best model (in terms of validation accuracy). As shown in Figure 26 and Figure 27, the training loss properly descended, and training accuracy improved as the epoch increased, while the validation loss slowly increased from around epoch 5, and the accuracy score plateaued at around 96%. In addition to the convincing metrics, the confusion matrix shown in Figure 28 also suggests a good performance for the model, with a high number of true positives and true negatives. However, the holdout metrics are less appealing: accuracy of 0.6095, F1 score of 0.1030, and AUC-ROC of 0.5110. Upon looking at the confusion matrix in Figure 29, there is a high number of true negatives and false negatives, meaning that the model does a great job of identifying fake news as fake but incorrectly marks real news as fake too. This effect is probably due to the different styles of writing that the holdout dataset has since the dataset includes not only news articles but also more colloquial types of media, such as emails and tweets.

Figure 26

LSTM Model Loss Over Epochs

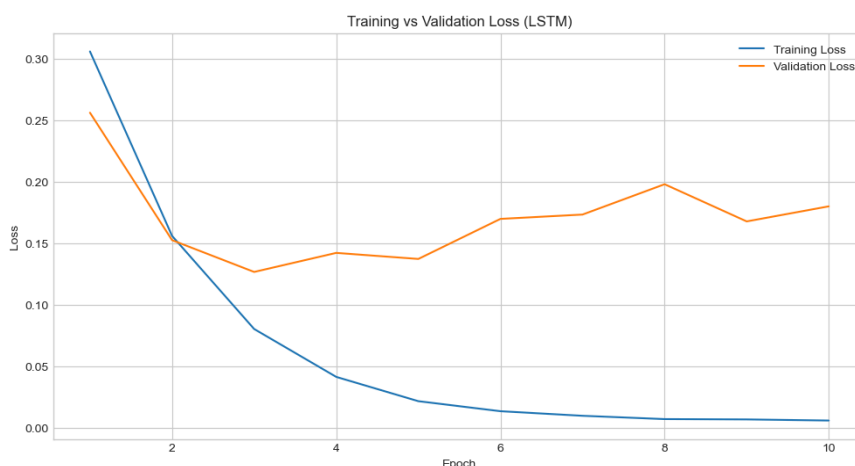


Figure 27

LSTM Model Accuracy Over Epochs

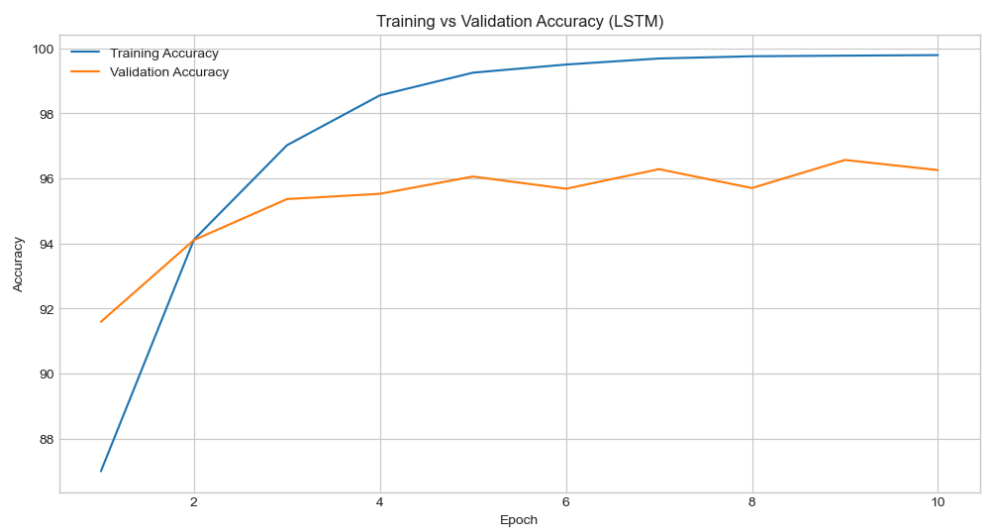


Figure 28

LSTM Confusion Matrix (Test Set)

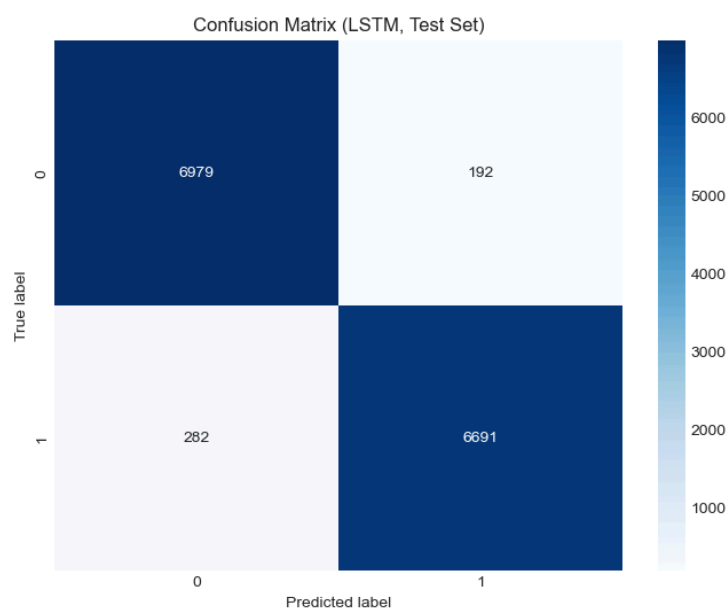
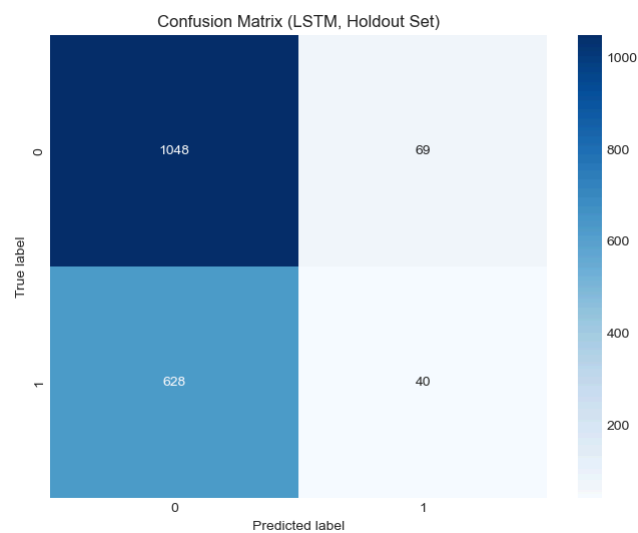


Figure 29

LSTM Confusion Matrix (Holdout Set)



LSTM with GloVe Embeddings

This model, after preprocessing, is embedded through a GloVe model that was pretrained from texts on Wikipedia (with 6 billion tokens as a 300-dimensional vector to be exact). Then, similar to the raw LSTM model, the data is fed through two LSTM layers with 256 nodes, each with a final linear layer at the end to map the data onto a sigmoid curve. The model is also trained through 10 epochs, with regularization techniques exactly the same as the raw LSTM model.

Early stopping occurred at epoch 7, with the model at epoch 4 saved as it has the best performance (see Figure 30 and Figure 31). It has an accuracy of 0.9712, an F1 score of 0.9707, and an AUC-ROC of 0.9957 on the test set. However, similar to the raw LSTM model, it does not infer well on the holdout set, with only an accuracy of 0.6246, an F1 score of 0.1667, and an AUC-ROC of 0.5409. As shown in Figure 32 and Figure 33, the model once again suggests that it performs well on the test set but predicts many false negatives on the holdout set, similar to the LSTM model. Further comparisons between models and human benchmarks will be analyzed later in the paper.

Figure 30

LSTM + GloVe Loss Over Epochs



Figure 31

LSTM + GloVe Accuracy Over Epochs

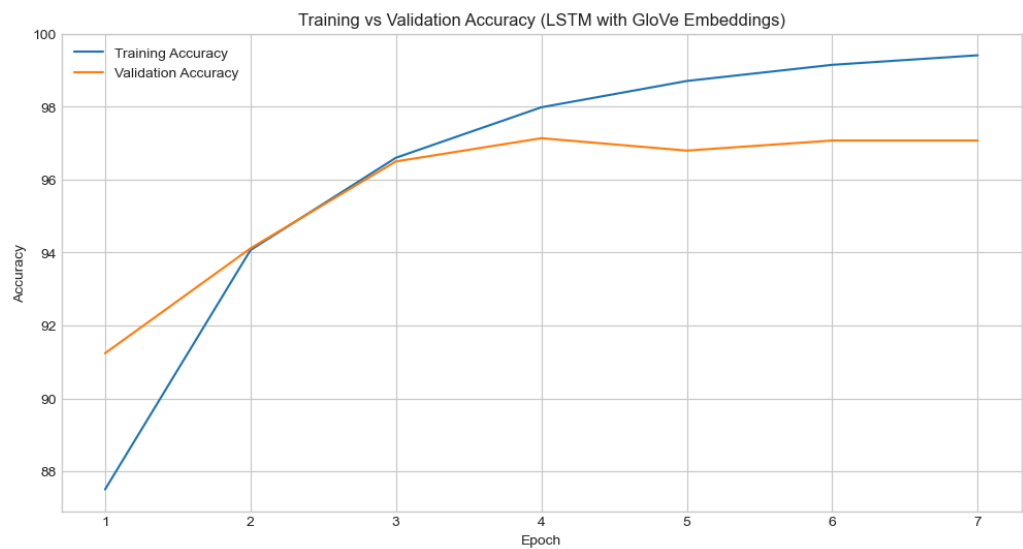


Figure 32

LSTM + GloVe Confusion Matrix (Test Set)

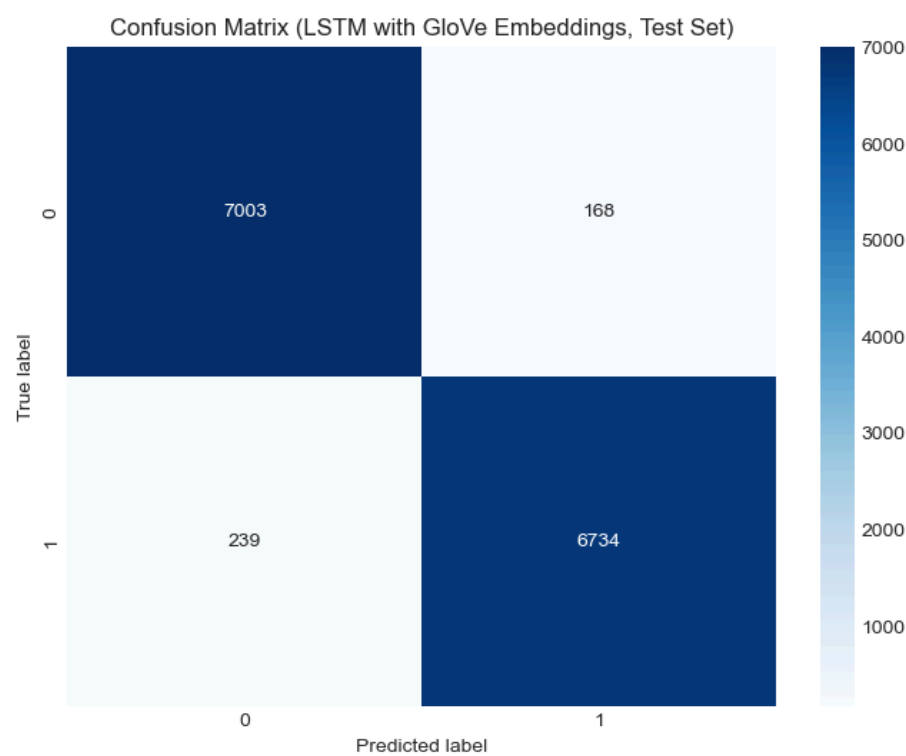
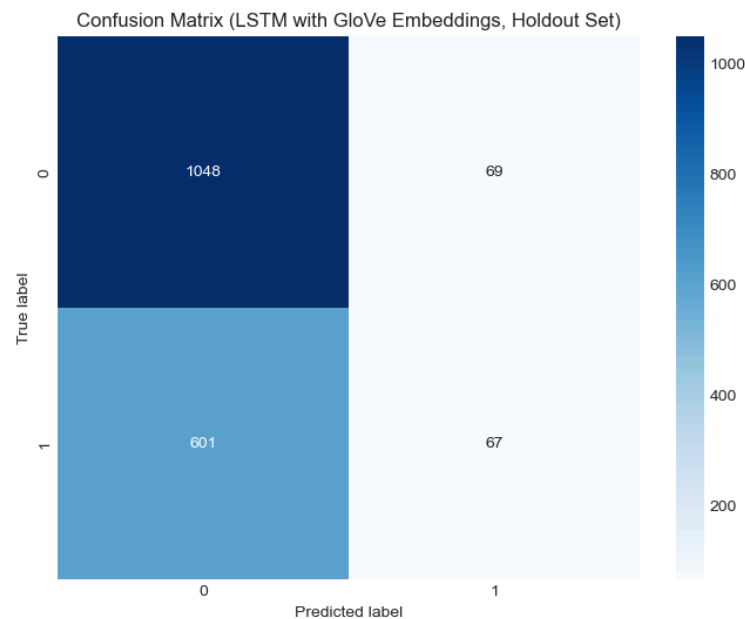


Figure 33

LSTM + GloVe Confusion Matrix (Holdout Set)



Fine-Tuned DistilBERT

For this model, I utilized the DistilBERT-base-uncased pre-trained model, meaning that this is the base version of the DistilBERT model (without fine-tunings) and is insensitive to casings. I trained the model on the same dataset as the ones from LSTM and LSTM + GloVe (WELFake). The fine-tuned training batch size is 16, a learning rate of 0.01, and the model is trained for 5 epochs. There is a weight decay of 0.01 (decreasing the weights slowly) as well as an internal dropout from the BERT model to regularize the model. The metrics of the model are logged every 50 steps (every batch pass-through is a step), with evaluation of the validation set occurring every 1000 steps to save time. The best model based on evaluation accuracy is saved at the end.

As shown in Figure 34, the training loss decreased drastically at first and decreased until it was nearly 0, while the validation loss hovered at around 0.5 (when it started to log at step 1000). The model was saved at the final step as it has the highest validation accuracy. This model achieved a test set accuracy of 0.9943, an F1 score of 0.9942, and an AUC-ROC of 0.9944. However, similar to the previous two models, this one also struggled with the holdout set, scoring an accuracy of 0.6156, an F1 score of 0.0130, and an AUC-ROC of 0.4478. Figure 35 and Figure 36 once again suggest the model performs extremely well on the test set but amasses a lot of false negatives on the holdout set.

Figure 34

DistilBERT Loss Over Steps

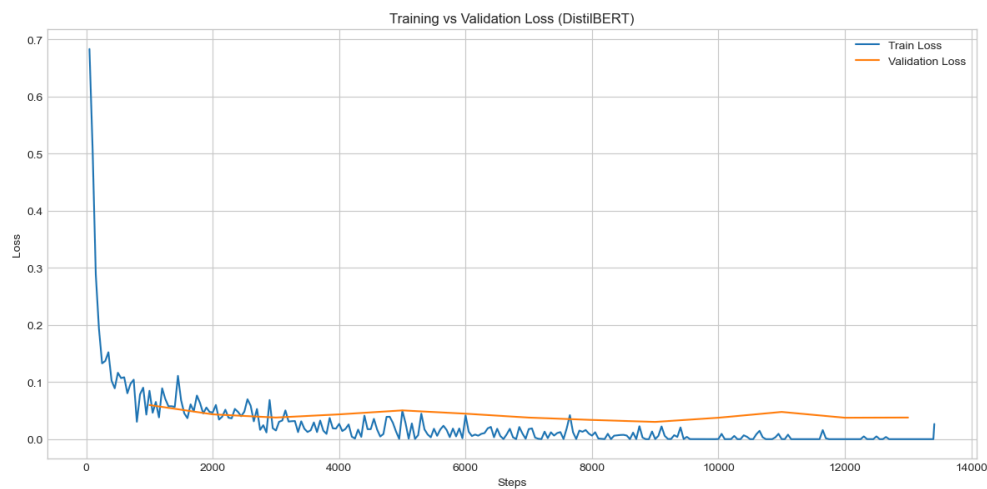


Figure 35

DistilBERT Confusion Matrix (Test Set)

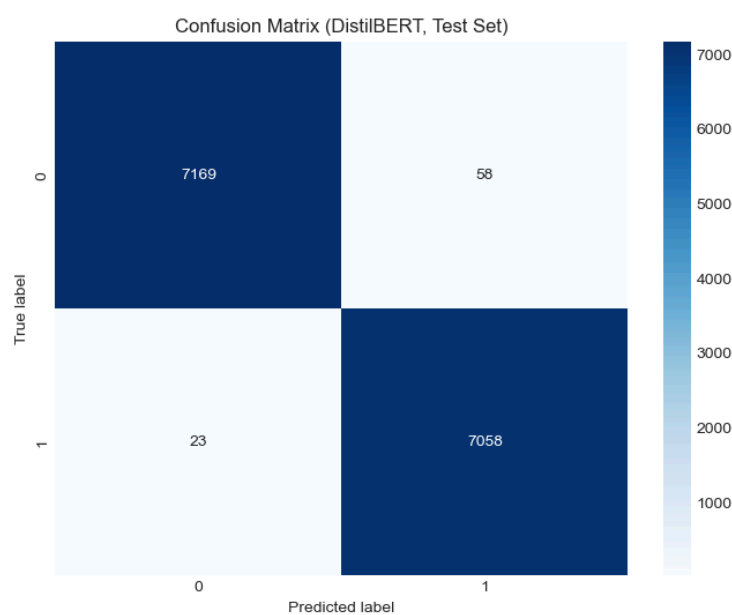
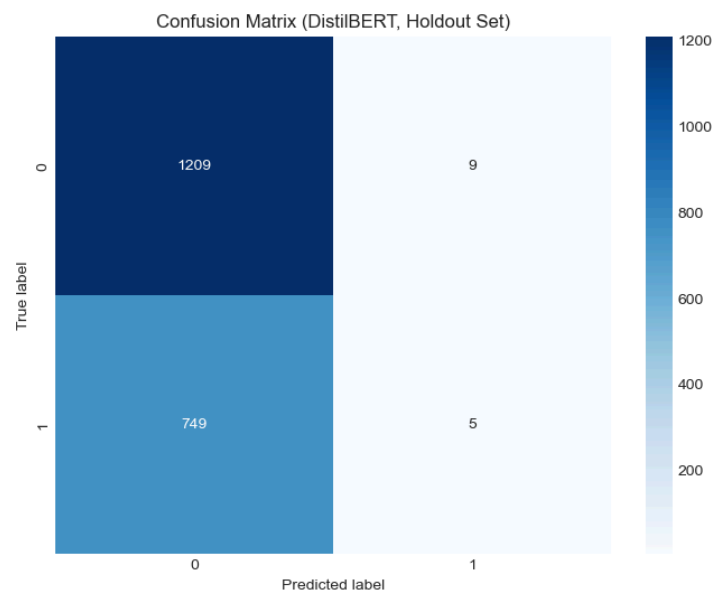


Figure 36

DistilBERT Confusion Matrix (Holdout Set)



Llama 2

For Llama 2 and GPT-4 Turbo, I randomly sampled 500 datapoints each from the WELFake dataset and the dataset from Bharadwaj et al. (2020), then removed any bias words from each datapoint. This process was done as inferencing with Llama-2 takes quite a lot of time, and GPT-4 Turbo could be costly. To be specific, this Llama model is the 13b-chat-hf version, meaning that it has 13 billion parameters in the chat completion “mode” and is downloaded through Hugging Face. The following is the prompt that was fed to Llama:

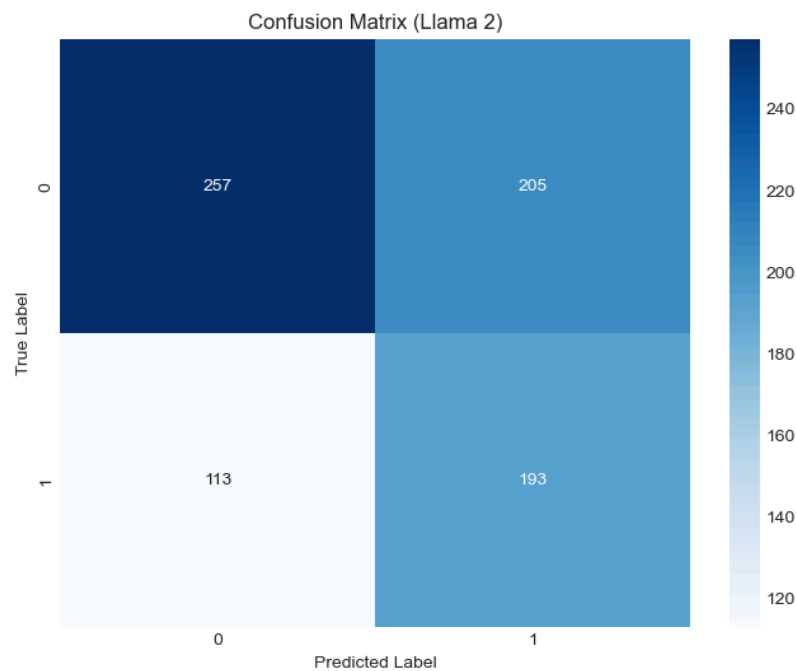
```
<s>[INST] <<SYS>>
You are an expert in identifying fake news and disinformation. Please identify whether the piece
of news is real or fake.
Please think step-by-step as you answer the question. However, please only respond with 'real' if
the news is real or 'fake' if the news is fake.
Do not respond with any other words or phrases.
If you are unsure if the news is real or fake, please still make an educated guess.
<</SYS>>
{{PASSAGE HERE}} [/INST]
```

The tags such as <<SYS>> in the above prompt come from the prompting guide for Llama by Schmid et al. (2023). These tags help ensure Llama answers the prompts more accurately. {{PASSAGE HERE}} indicates the location of where each of the 1000 data points belongs when prompting Llama. Each data point is fed to Llama one at a time until all 1000 are predicted. Since the max token output of Llama was limited to 1 due to time and computing constraints, some outputs from Llama are truncated, and all output from Llama that indicates “real” and “r” will mark the passage as real, and “fake” or “f” will mark the passage as fake. Of the initial 1000 data points, Llama predicted 768 of them.

Llama scored an accuracy of 0.5859 and an F1 score of 0.5482. As shown in Figure 37, Llama 2 does a great job at marking fake information as fake; however, it has a high count for false positives, meaning that it is incorrectly marking fake information as real.

Figure 37

Llama-2 Confusion Matrix



GPT-4 Turbo

After going through the same preprocessing steps as Llama 2, this model is called through OpenAI’s application programming interface (API). An API is essentially a call to the server that does some data retrieval and returns the output. In this case, calling the OpenAI API gives us the output for GPT’s response to the prompt. The model I used was specifically gpt-4-1106-preview, which is the model name for GPT-4 Turbo that was released on Nov. 06, 2023. This model was mainly chosen due to the cheaper price per token while showcasing a similar level of text comprehension as to the GPT-4 Turbo version accessed by users through the ChatGPT website. The following is the system instruction that was given to GPT:

You are an expert in identifying fake news and disinformation. Please identify whether the piece of news is real or fake.
Please think step-by-step as you answer the question. However, please only respond with 'real' if the news is real or 'fake' if the news is fake.
Do not respond with any other words or phrases.
If you are unsure if the news is real or fake, please still make an educated guess.

This prompt is similar to the Llama 2 prompt, except that it does not contain the special tags. The passage follows the system instructions, and the max output token was set to 10. Again, only outputs that are either “r”, “real”, “f”, or “fake” are marked, and GPT outputted 994 of those responses in total. GPT scored an accuracy of 0.7324 and an F1 score of 0.5970 across the 994 predicted outputs. As shown in Figure 38, ChatGPT predicted a high number of true negatives, a moderate number of true positives and false negatives, and a small number of false positives.

Figure 38

GPT-4 Turbo Confusion Matrix

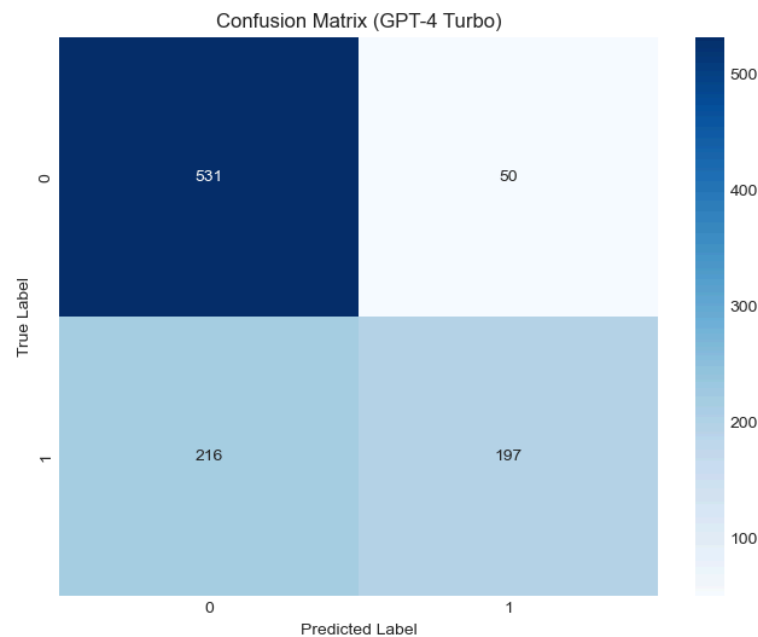


Table 4 gives a summary of the metrics across the above models. Further analysis between models and human benchmarks will be done under the Analyses section.

Table 4

Metrics Across ML Models

Model	Test Accuracy	Test F1 Score	Test AUC-ROC	Holdout Accuracy	Holdout F1 Score	Holdout AUC-ROC
LSTM	0.9665	0.9685	0.9938	0.6095	0.1030	0.5110
LSTM + GloVe	0.9712	0.9707	0.9957	0.6246	0.1667	0.5409
DistilBERT	0.9943	0.9942	0.9944	0.6156	0.0130	0.4478
Llama 2	N/A	N/A	N/A	0.5859	0.5482	N/A
GPT-4 Turbo	N/A	N/A	N/A	0.7324	0.5970	N/A

Note. For Llama 2 and GPT-4 Turbo, since they were tested on only the sampled 1000 data points, their results are put under the Holdout Accuracy and Holdout F1 Score columns. In addition, since I do not have access to the probability for their predictions on every data point, I am thus unable to calculate AUC-ROC for these two models.

Grade 9 Quizzes

Five quizzes were administered to grade 9 students to understand more about their abilities to identify whether a piece of information is real or fake. Appendices C, D, E, F, and G contain the

questions that were given to them on each occasion. Table 5 shows the descriptive statistics of the five quizzes, as well as the overall statistics with the five quizzes aggregated, which is shown visually throughout Figures 39.1 to 39.6. From Figures 39.1 to 39.5, even though the distribution of the quizzes seems significantly different from one another, they still roughly resemble a bell-shaped curve, albeit skewed in different directions, with mean and median lying somewhere between 2 and 3. As shown in Figure 39.6, the overall distribution of the five quizzes captures a more normalized curve with a slight skew to the left.

Table 5

Descriptive Statistics for the Five Quizzes

Quiz	Median	Mean	Mode	Q1	Q3	IQR	SD	Range
Overall	3	2.690	3	2	4	2	1.228	5
1	3	3.133	3	2	4	2	1.120	4
2	2	2.489	2	2	3	1	1.141	5
3	3	2.795	3	2	4	2	1.250	5
4	2	2.405	2	2	3	1	1.142	5
5	2	2.571	2	2	3.75	1.75	1.382	5

Discussion and Analysis

Due to the constraints of this project, a standard ANOVA test cannot be performed, as the distribution of some human benchmarks is not approximately normal and there is a lack of repetition of using ML models to inference data. In addition, the Kruskal-Wallis test can not be used as the human samples are not independent, and I do not have individual predictions of every data point for the ML models. Given these limitations, the best option is to conduct a qualitative assessment of the mean accuracy and figures across groups rather than a quantitative one. This approach involves directly observing and comparing these averages to see if there are differences between groups, which will provide insights that are not strictly inferential but instead offer a practical understanding of the data given the constraints.

The mean accuracies are used for this analysis, as it provides the easiest way to interpret the data. As shown in Table 6, all the ML models outperformed the two human benchmarks, especially GPT-4 Turbo, which performs significantly better than all other ML models and human benchmarks, setting an accuracy of 0.7324. This significant difference is most likely due to the large amount of text that GPT was trained on and the reasoning that it may be able to portray as compared to other models that was trained on a smaller set of data (especially the custom models). The human benchmarks all seem to be significantly worse than the rest, sitting at 0.5128 and 0.5380 for the initial survey and the 5 quizzes, respectively, whereas the worst ML model, Llama 2, has an accuracy of 0.5859. However, even though these ML models are better than the set human benchmarks, it is unclear whether it has practical use cases in the real world, as even an accuracy of 0.7324 might not be significant enough to replace human fact-checkers. In addition, as shown in the confusion matrices for each model above, LSTM, LSTM + GloVe, and DistilBERT have a very high number of false negatives with almost no true negatives, which is an indication that the models are purely marking everything as false, which would suggest that it is not any better than the human benchmark. This statement is also reflected in the holdout F1 scores as shown in Table 4, where all the custom models have a low score while Llama-2

and GPT-4 Turbo have a slightly above average score, which suggests that the two LLMs may be slightly better than average, but not necessarily human benchmarks. Again, it is important to mention that the above analysis should be taken with a certain level of caution and acknowledged as indicative rather than definitive.

Table 6

Mean Accuracy Scores Across Different Data

Data	Accuracy
Initial Survey	0.5128
5 Quizzes	0.5380
LSTM	0.6095
LSTM + GloVe	0.6246
DistilBERT	0.6156
Llama 2	0.5859
GPT-4 Turbo	0.7324

Note. The accuracy scores for ML models are from the holdout set, as it gives the best predictor of new data.

Conclusion

This research aimed to explore the capabilities of AI in detecting disinformation and compare the performance of AI models against human benchmarks. Through the development and evaluation of various deep learning models, including LSTM, LSTM with GloVe embeddings, and fine-tuned DistilBERT, as well as the utilization of LLMs like Llama-2 and GPT-4 Turbo, I found that AI models generally outperformed the human benchmarks established through surveys and quizzes administered to high school students at DIS. Among all the AI models tested GPT-4 Turbo demonstrated the highest accuracy of 0.7324 in identifying disinformation, significantly surpassing the other models and the human benchmarks through a qualitative analysis. This performance is most likely due to the amount of data that GPT-4 was trained on, as well as its advanced reasoning capabilities as compared to other DL models and LLMs. However, it is important to note that even an accuracy of 0.7324 may not be significant enough to fully replace human fact-checkers in real-world scenarios.

While custom-built models (LSTM, LSTM with GloVe, and DistilBERT) achieved high accuracies on the test sets from the training data, their performance on the holdout sets, which simulated new and unseen data, was less impressive. These models have a high number of false negatives, meaning that they tend to incorrectly label real news as fake. This observation suggests that these models may have overfitted to the specific characteristics of the training data and struggled to generalize effectively to new data.

Even though the analysis of the results also highlighted the potential of AI techniques in identifying disinformation, it also demonstrated certain challenges and limitations. While AI models can outperform human benchmarks in specific tasks, their real-world applicability may be limited by factors such as overfitting, data quality, and the complexity of language nuances. For instance, it is difficult to compile all the data needed for a good detector as there are many news sources in all kinds of formats

(typical news articles, tweets, videos, images, messages, etc.), and the datasets available currently come in different formats (some does not have any passage text, some only have a title, and some doesn't have either), which makes training models and inferencing data very difficult. Moreover, some datasets have the same data points, which makes inference useless. Once more datasets are publicly accessible, better models can be developed that encompass more forms of media.

Overall, this research demonstrates the potential contribution of AI to combat the spread of disinformation. However, there is still a need for continued research and development to improve the variety and quality of datasets related to disinformation, which would allow better disinformation detection models to be built.

Suggestions for Improvement

Because of time constraints, I was unable to build and utilize other models that I had missed in this research, such as Claude 2. The rapid development of generative models also means that I was unable to use newer technology (like Gemini Ultra and Claude 3), which might have outperformed GPT-4 Turbo. In addition, due to time, computational, and financial constraints, I was unable to repeat experiments for the models I built and utilized, meaning that there were fewer accuracy scores for rigorous testing and comparison. If the experiments were repeated, then statistical tests could be employed to validate the results of this experiment quantitatively rather than qualitatively.

In addition, since this research employed 5 quizzes on the same set of students, it is possible that students may have started guessing the answers in the final quizzes rather than trying to identify the correct answer to the best of their abilities. A better plan would have been to send out different quizzes to different populations for the most accurate representation of students' abilities.

Recommendations for Further Research

As discussed above, further research could be done on developing better datasets on disinformation so that better ML models can be developed. In addition, utilizing more up-to-date models is also a necessity as the field of AI continues to evolve at a rapid pace. I suggest further research to explore the depth of such technologies in the field of identifying disinformation through diversifying the data (if available), as well as incorporating multimodal data, such as images, videos, and audio, as they become more readily available in our lives. In addition, it would be beneficial to build a tighter definition as to what constitutes a piece of information as real and what as fake, as not everything in an article is completely true, and not everything in an article is completely false or either.

Word Count: 18534

References

Book Sources

- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: with applications in Python*. Springer Nature. https://hastie.su.domains/ISLP/ISLP_website.pdf
- Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

Other Sources

- Acharya, A. (2023, June 13). *Training, validation, test split for machine learning datasets*. Encord. <https://encord.com/blog/train-val-test-split/>
- Agrawal, R. (2023, October 9). *Know the best evaluation metrics for your regression model !* Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/>
- Ammu, B. (n.d.). *The importance of data preparation for machine learning*. Astera. Retrieved April 27, 2024, from <https://www.astera.com/type/blog/data-preparation-for-machine-learning>
- Azevedo, N. (2023, September 29). Data preprocessing techniques: 6 steps to clean data in machine learning. *Scalable Path*. <https://www.scalablepath.com/data-science/data-preprocessing-phase>
- Azizah, S. F. N., Cahyono, H. D., Sihwi, S. W., & Widiarto, W. (2023). Performance analysis of transformer based models (BERT, ALBERT and ROBERTA) in fake news detection. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2308.04950>
- Baeldung. (2024, March 18). *Prevent the vanishing gradient problem with LSTM*. Baeldung on Computer Science. <https://www.baeldung.com/cs/lstm-vanishing-gradient-prevention>
- Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.4337484>
- Bhaimiya, S. (2023, June 20). DeepMind's co-founder suggested testing an AI chatbot's ability to turn \$100,000 into \$1 million to measure human-like intelligence. *Business Insider*. <https://www.businessinsider.com/deepmind-co-founder-suggests-new-turing-test-ai-chatbots-report-2023-6>
- Bharadwaj, A., Ashar, B., Barbhaya, P., Bhatia, R., & Shaikh, Z. (2020). Source based fake news classification using machine learning. *International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)*, 9(6), 5147–5153. https://www.ijirset.com/upload/2020/june/115_4_Source.PDF
- Bourke, D. (2022, July 24). *Learn PyTorch for deep learning in a day. Literally*. [Video]. YouTube. https://www.youtube.com/watch?v=Z_ikDlimN6A
- Brownlee, J. (2020, June 30). *Why One-Hot encode data in machine learning?* Machine Learning Mastery. <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>
- Brownlee, J. (2021, January 22). *How to choose an activation function for deep learning*. Machine Learning Mastery. <https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/>
- Burkov, A. (2019). *The Hundred-Page machine learning book*. <https://themlbook.com/>
- Canale, L., & Messina, A. (2023, October 17). *Experimenting AI Technologies for Disinformation Combat: the IDMO Project*. arXiv.org. <https://arxiv.org/abs/2310.11097v4>
- Chakraborty, D., & Roy, A. (2022, November 28). *An introduction to the global vectors (GloVe) algorithm*. Weights & Biases. <https://wandb.ai/authors/embeddings-2/reports/An-Introduction-to-the-Global-Vectors-GloVe-Algorithm--VmldzozNDg2NTQ>

- Chauhan, T., & Palivela, H. (2021). Optimization and improvement of fake news detection using deep learning approaches for societal benefit. *International Journal of Information Management Data Insights*, 1(2). <https://doi.org/10.1016/j.jjime.2021.100051>
- Copeland, B. (2023, December 31). *artificial intelligence*. Encyclopedia Britannica. <https://www.britannica.com/technology/artificial-intelligence>
- Czaron, J. (2023, September 5). *F1 score vs ROC AUC vs accuracy vs PR AUC: Which evaluation metric should you choose?* neptune.ai. <https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc>
- De Beer, D., & Matthee, M. (2020). Approaches to identify fake news: A systematic literature review [Integrated Science in Digital Age 2020]. In *Lecture Notes in Networks and Systems* (Vol. 136). Springer, Cham. https://doi.org/10.1007/978-3-030-49264-9_2
- De Witte, M. (2022, April 13). What to know about disinformation and how to address it. *Stanford News*. <https://news.stanford.edu/2022/04/13/know-disinformation-address/>
- Deepanshi. (2023, April 26). *Text preprocessing in NLP with Python codes*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/>
- Dohmke, T. (2023, June 27). *The economic impact of the AI-powered developer lifecycle and lessons from GitHub Copilot*. The GitHub Blog. <https://github.blog/2023-06-27-the-economic-impact-of-the-ai-powered-developer-lifecycle-and-lessons-from-github-copilot/>
- Donges, N. (2024, February 28). *A complete guide to Recurrent Neural networks (RNNs)*. Built In. <https://builtin.com/data-science/recurrent-neural-networks-and- lstm>
- Dertat, A. (2017, August 8). Applied Deep Learning - Part 1: Artificial Neural Networks. *Medium*. <https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6>
- Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E. J., Dean, J., & Socher, R. (2021). Deep learning-enabled medical computer vision. *Npj Digital Medicine*, 4(1). <https://doi.org/10.1038/s41746-020-00376-2>
- Faust, R. (2018, May 29). *Issue brief: How disinformation impacts politics and publics*. National Endowment for Democracy. <https://www.ned.org/issue-brief-how-disinformation-impacts-politics-and-publics/>
- Framing: Key ML terminology*. (n.d.). Google for Developers. Retrieved April 27, 2024, from <https://developers.google.com/machine-learning/crash-course/framing/ml-terminology>
- Galli, A., Masciari, E., Moscato, V., & Sperli, G. (2022). A comprehensive benchmark for fake news detection. *Journal of Intelligent Information Systems*, 59(1), 237–261. <https://doi.org/10.1007/s10844-021-00646-9>
- GeeksforGeeks. (2023, December 21). *Linear regression in machine learning*. <https://www.geeksforgeeks.org/ml-linear-regression/>
- Géron, A. (2019). *Hands-On machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media, Inc.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <https://www.deeplearningbook.org/>
- Gulati, A. (2023, September 14). *How to interpret R-squared in regression analysis?* KnowledgeHut. <https://www.knowledgehut.com/blog/data-science/interpret-r-squared-and-goodness-fit-regression-analysis>
- He, L., Hu, S., & Pei, A. (2023). Debunking disinformation: Revolutionizing truth with NLP in fake news detection. *ResearchGate*. https://www.researchgate.net/publication/373552327_Debunking_Disinformation_Revolutionizing_Truth_with_NLP_in_Fake_News_Detection
- Introducing Llama 2*. (n.d.). Meta. Retrieved April 27, 2024, from <https://ai.meta.com/llama/>

- Ipsos. (2023, July 20). *AI is making the world more nervous*.
<https://www.ipsos.com/en-tw/ai-making-world-more-nervous>
- Jaiswal, S. (2024, February 7). *Multilayer Perceptrons in Machine Learning: A Comprehensive guide*. DataCamp. <https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning>
- Khanna, C. (2021, February 10). Text pre-processing: Stop words removal using different libraries. *Medium*.
<https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a>
- Language models, explained: How GPT and other models work. (2023, January 18). *AltexSoft*.
<https://www.altexsoft.com/blog/language-models-gpt/>
- Lerner, M. (2023, September 7). Worried about AI in the workplace? You're not alone. *American Psychological Association*.
<https://www.apa.org/topics/healthy-workplaces/artificial-intelligence-workplace-worry>
- Malik, A. (2023, November 7). *OpenAI's ChatGPT now has 100 million weekly active users*. TechCrunch.
<https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/?gucounter=1>
- Mattern, J., Qiao, Y., Kerz, E., Wiechmann, D., & Strohmaier, M. (2021). FANG-COVID: A new Large-Scale benchmark dataset for fake news detection in German. *Association for Computational Linguistics*, 78–91. <https://doi.org/10.18653/v1/2021.fever-1.9>
- McQuilling, D. (2022, August 26). *SlashGear survey: 43% of people polled are scared of the potential of sentient AI*. SlashGear.
<https://www.slashgear.com/944808/43-of-people-polled-are-scared-of-the-potential-of-sentient-ai/>
- Mello, R. F., Freitas, E. L. S. X., Pereira, F. D., De Souza Cabral, L., Tedesco, P., & Ramalho, G. (2023). Education in the age of generative AI: Context and recent developments. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2309.12332>
- Meta. (2023, July 18). Meta and Microsoft introduce the next generation of Llama. *Meta*.
<https://about.fb.com/news/2023/07/llama-2/>
- Morgunov, A. (2023, August 18). *The life cycle of a machine learning project: What are the stages?* neptune.ai. <https://neptune.ai/blog/life-cycle-of-a-machine-learning-project>
- Moroney, L. (2020, February 25). *Sequencing - Turning sentences into data (NLP zero to hero - part 2)* [Video]. YouTube. <https://www.youtube.com/watch?v=r9QjkdSJZ2g>
- Muller, B. (2022, March 2). *BERT 101 🤖 state of the art NLP model explained*. Hugging Face.
<https://huggingface.co/blog/bert-101>
- Nabi, J. (2019, July 12). Recurrent Neural Networks (RNNs). *Medium*.
<https://towardsdatascience.com/recurrent-neural-networks-rnns-3f06d7653a85>
- Neagoie, A., & Bourke, D. (n.d.). *Complete machine learning & data science bootcamp 2023* [MOOC]. Udemy. Retrieved April 27, 2024, from
<https://www.udemy.com/course/complete-machine-learning-and-data-science-zero-to-mastery/>
- Neubig, G. (2021, February 3). *CMU neural nets for NLP 2021 (1): Introduction* [Video]. YouTube.
<https://www.youtube.com/watch?v=vnX6M7N-ggs>
- Neuroscience News. (2023, June 23). Neuroscience, artificial intelligence, and our fears: A journey of understanding and acceptance. *Neuroscience News*.
<https://neurosciencenews.com/artificial-intelligence-fear-neuroscience-23519/>
- Ng, A., Ladwig, G., Bagul, A., & Shyu, E. (n.d.). *Machine learning specialization* [MOOC]. Coursera. Retrieved April 27, 2024, from
<https://www.coursera.org/specializations/machine-learning-introduction>
- OpenAI. (2022, November 30). *Introducing ChatGPT*. <https://openai.com/blog/chatgpt>
- Overfitting and underfitting in machine learning*. (n.d.). Javatpoint. Retrieved April 27, 2024, from
<https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning>

- Peng, S., Kalliamvakou, E., Cihon, P., & Demirel, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv (Cornell University)*.
<https://doi.org/10.48550/arxiv.2302.06590>
- Purohit, S. (2023, July 19). Role of exploratory data analysis in machine learning. *The Talent500 Blog*.
<https://talent500.co/blog/role-of-exploratory-data-analysis-in-machine-learning/>
- Redman, T. (2018, April 2). *If your data is bad, your machine learning tools are useless*. Harvard Business Review.
<https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless>
- Saleem, A. (n.d.). Transformer models: NLP's new powerhouse. *Data Science Dojo*. Retrieved April 27, 2024, from <https://datasciencedojo.com/blog/transformer-models/>
- Sanderson, G. (2017a, October 5). *But what is a neural network? | Chapter 1, deep learning* [Video]. YouTube. <https://www.youtube.com/watch?v=aircAruvnKk>
- Sanderson, G. (2017b, November 3). *Backpropagation calculus | Chapter 4, deep learning* [Video]. YouTube. <https://www.youtube.com/watch?v=tIeHLNjs5U8>
- Sarguroh, F. (2023, July 19). 04: Text Representation (Part-1). *Medium*.
<https://medium.com/@farhahahahahan/04-text-representation-part-1-b57659320e48>
- Schreiner, M. (2023, July 11). GPT-4 architecture, datasets, costs and more leaked. *THE DECODER*.
<https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/>
- Schmid, P., Sanseviero, O., Cuenca, P., & Tunstall, L. (2023, July 18). *Llama 2 is here - get it on Hugging Face*. Hugging Face. <https://huggingface.co/blog/llama2>
- Shaer, M. (2017, March 10). Fighting the Nazis with fake news. *Smithsonian Magazine*.
<https://www.smithsonianmag.com/history/fighting-nazis-fake-news-180962481/>
- Shani, I. & GitHub Staff. (2023, June 13). *Survey reveals AI's impact on the developer experience*. The GitHub Blog.
<https://github.blog/2023-06-13-survey-reveals-ais-impact-on-the-developer-experience/>
- Sharma, S. (2017, September 6). Activation functions in neural networks. *Medium*.
<https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>
- Shung, K. P. (2020, April 10). Accuracy, precision, recall or F1? *Medium*.
<https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>
- Simplilearn. (2023, January 18). *Normalization vs standardization - What's the difference?* Simplilearn.com. <https://www.simplilearn.com/normalization-vs-standardization-article>
- Singh, Y. (2022, October 14). *3 regression metrics you must know: MAE, MSE, and RMSE*. Proclus Academy. <https://proclusacademy.com/blog/explainer/regression-metrics-you-must-know/>
- Spezzano, F., Shrestha, A., Fails, J. A., & Stone, B. W. (2021). That's fake news! Reliability of news when provided title, image, source bias & full article. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1), 1–19. <https://doi.org/10.1145/3449183>
- Starmer, J., [StatQuest with Josh Starmer]. (2020a, August 31). *Neural networks pt. 1: Inside the black box* [Video]. YouTube. <https://www.youtube.com/watch?v=CqOfi41LfDw>
- Starmer, J., [StatQuest with Josh Starmer]. (2020b, November 23). *Neural networks pt. 3: RELU in action!!!* [Video]. YouTube. <https://www.youtube.com/watch?v=68BZ5f7P94E>
- Starmer, J., [StatQuest with Josh Starmer]. (2022a, July 11). *Recurrent neural networks (RNNs), clearly explained!!!* [Video]. YouTube. <https://www.youtube.com/watch?v=AsNTP8Kwu80>
- Starmer, J., [StatQuest with Josh Starmer]. (2022b, November 7). *Long short-term memory (LSTM), clearly explained* [Video]. YouTube. <https://www.youtube.com/watch?v=YCzL96nL7j0>
- Susanto, A. P., Winarto, H., Fahira, A., Abdurrohman, H., Muharram, A. P., Widitha, U. R., Efirianti, G. E. W., George, Y. a. E., & Tjoa, K. (2022). Building an artificial intelligence-powered medical image recognition smartphone application: What medical practitioners need to know. *Informatics in Medicine Unlocked*, 32. <https://doi.org/10.1016/j.imu.2022.101017>

- Tackling coronavirus disinformation.* (n.d.). European Commission. Retrieved April 27, 2024, from https://commission.europa.eu/strategy-and-policy/coronavirus-response/fighting-disinformation/tackling-coronavirus-disinformation_en
- Thakur, D. (2018, July 6). LSTM and its equations. *Medium*. <https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af>
- The Pecan Team. (2023, November 21). *Data Preparation for Machine Learning: The Ultimate Guide to doing it right*. Pecan AI. <https://www.pecan.ai/blog/data-preparation-for-machine-learning/>
- Touvron, H., Martin, L., Stone, K. H., Albert, P. J., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., . . . Scialom, T. (2023). Llama 2: Open foundation and Fine-Tuned chat models. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2307.09288>
- Undy, B. (2023, October 25). *The art and science of algorithm selection in machine learning*. LinkedIn. <https://www.linkedin.com/pulse/art-science-algorithm-selection-machine-learning-bryce-undy-3uryc/>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1706.03762>
- Verma, P. K., Agrawal, P., Amorim, I., & Prodan, R. (2021). WELFAKE: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4), 881–893. <https://doi.org/10.1109/tcss.2021.3068519>
- Wakefield, K. (n.d.). *A guide to the types of machine learning algorithms*. SAS UK. Retrieved April 27, 2024, from https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html
- Ward, M., Pierson, S., & Beyer, J. (n.d.). *Formative battles: Cold War disinformation campaigns and mitigation strategies*. Wilson Center. Retrieved April 27, 2024, from <https://www.wilsoncenter.org/publication/formative-battles-cold-war-disinformation-campaigns-and-mitigation-strategies>
- What is holdout dataset in machine learning?* (n.d.). Iguazio. Retrieved April 27, 2024, from <https://www.iguazio.com/glossary/holdout-dataset>
- What is hyperparameter tuning?* (n.d.). Amazon Web Services. Retrieved April 27, 2024, from <https://aws.amazon.com/what-is/hyperparameter-tuning/>
- What is overfitting?* (n.d.). Amazon Web Services. Retrieved April 27, 2024, from <https://aws.amazon.com/what-is/overfitting/>
- Yuan, L., Jiang, H., Shen, H., Shi, L., & Cheng, N. (2023). Sustainable development of information dissemination: A review of current fake news detection research and practice. *Systems*, 11(9), 458. <https://doi.org/10.3390/systems11090458>

Tables

- Table 1. Chen, K. (2023). *Example of a Categorical Feature without One-Hot Encoding*.
- Table 2. Chen, K. (2023). *Example of a Categorical Feature One-Hot Encoded*.
- Table 3. Chen, K. (2024). *Descriptive Statistics for the Five-Question Quiz Score*.
- Table 4. Chen, K. (2024). *Metrics Across ML Models*.
- Table 5. Chen, K. (2024). *Descriptive Statistics for the Five Quizzes*.
- Table 6. Chen, K. (2024). *Mean Accuracy Scores Across Different Data*.

Figures

- Figure 1. Chen, K. (2023). *A Feedforward Neural Network with Four Layers*.
- Figure 2. Chen, K. (2023). *Example of a Basic Linear Regression with One Feature and One Label*.
- Figure 3. Chen, K. (2023). *Example of a Basic Polynomial Regression with One Feature*.
- Figure 4. Chen, K. (2023). *Graph of the Sigmoid Function*.

Figure 5. Chen, K. (2023). *Visual Representation of MSE.*

Figure 6. Chen, K. (2023). *Graph of $L(f_{w,b}^{\rightarrow}(x^{(i)}), y^{(i)}) = -\log(f_{w,b}^{\rightarrow}(x^{(i)}))$ When $y^{(i)} = 1$.*

Figure 7. Chen, K. (2023). *Graph of $L(f_{w,b}^{\rightarrow}(x^{(i)}), y^{(i)}) = -\log(1 - f_{w,b}^{\rightarrow}(x^{(i)}))$ When $y^{(i)} = 0$.*

Figure 8. Rishi. (2023, May 3). Understanding of gradient descent: Intuition and implementation. *Medium.*

<https://blog.gopenai.com/understanding-of-gradient-descent-intuition-and-implementation-b1f98b3645ea>

Figure 9. Chen, K. (2023). *Visual Representation of Dataset Makeup.*

Figure 10. Chen, K. (2023). *Example of an Overfitted Model.*

Figure 11. Chen, K. (2023). *Example of an Underfitted Model.*

Figure 12. Ma, J., Ding, Y., Cheng, J. C. P., Tan, Y., Gan, V. J., & Zhang, J. (2019). Analyzing the leading causes of traffic fatalities using XGBOOST and Grid-Based analysis: A city management perspective. *IEEE Access*, 7, 148059–148072. <https://doi.org/10.1109/access.2019.2946401>

Figure 13. Shukla, L. (2023, September 19). *Fundamentals of neural networks.* Weights & Biases. https://wandb.ai/wandb_fc/articles/reports/Fundamentals-of-Neural-Networks--Vmldzo1NDQ0Mzk1

Figure 14. Chen, K. (2023). *Graph of a ReLU Activation Function.*

Figure 15. Chen, K. (2023). *Graph of a Tanh Activation Function.*

Figure 16. Thakur, A. (2022, December 13). *Examples of early stopping in HuggingFace transformers.* Weights & Biases.

<https://wandb.ai/ayush-thakur/huggingface/reports/Examples-of-Early-Stopping-in-HuggingFace-Transformers--Vmldzo0MzE2MTM>

Figure 17. Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

Figure 18. Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

Figure 19. Deshpande, A., & Kumar, M. (n.d.). *N-grams.* O'Reilly. Retrieved April 27, 2024, from <https://www.oreilly.com/library/view/artificial-intelligence-for/9781788472173/c11d01e5-bf13-488f-8922-16f6028254e4.xhtml>

Figure 20. Sigler, I. (2022, August 25). Vertex AI example-based explanations improve ML via explainability. *Google Cloud.*

<https://cloud.google.com/blog/products/ai-machine-learning/example-based-explanations-to-build-better-aiml-models>

Figure 21. aishwary. (2023, December 4). *Introduction to recurrent neural network.* GeeksforGeeks. <https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/>

Figure 22. Artificial Intelligence +. (2023, June 26). *Introduction to Long Short Term Memory (LSTM).* <https://www.aiplusinfo.com/blog/introduction-to-long-short-term-memory-lstm/>

Figure 26. Chen, K. (2024). *LSTM Model Loss Over Epochs.*

Figure 27. Chen, K. (2024). *LSTM Model Accuracy Over Epochs.*

Figure 28. Chen, K. (2024). *LSTM Confusion Matrix (Test Set).*

Figure 29. Chen, K. (2024). *LSTM Confusion Matrix (Holdout Set).*

Figure 30. Chen, K. (2024). *LSTM + GloVe Loss Over Epochs.*

Figure 31. Chen, K. (2024). *LSTM + GloVe Accuracy Over Epochs.*

Figure 32. Chen, K. (2024). *LSTM + GloVe Confusion Matrix (Test Set).*

Figure 33. Chen, K. (2024). *LSTM + GloVe Confusion Matrix (Holdout Set).*

Figure 34. Chen, K. (2024). *DistilBERT Loss Over Steps.*

Figure 35. Chen, K. (2024). *DistilBERT Confusion Matrix (Test Set)*.
Figure 36. Chen, K. (2024). *DistilBERT Confusion Matrix (Holdout Set)*.
Figure 37. Chen, K. (2024). *Llama-2 Confusion Matrix*.
Figure 38. Chen, K. (2024). *GPT-4 Turbo Confusion Matrix*.

Figures

Figure 23.1

Grade Level

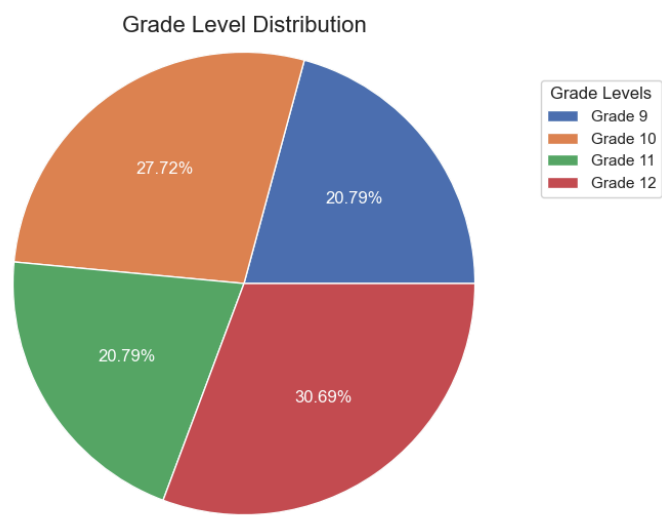


Figure 23.2

Do You Know What Artificial Intelligence (AI) Is?

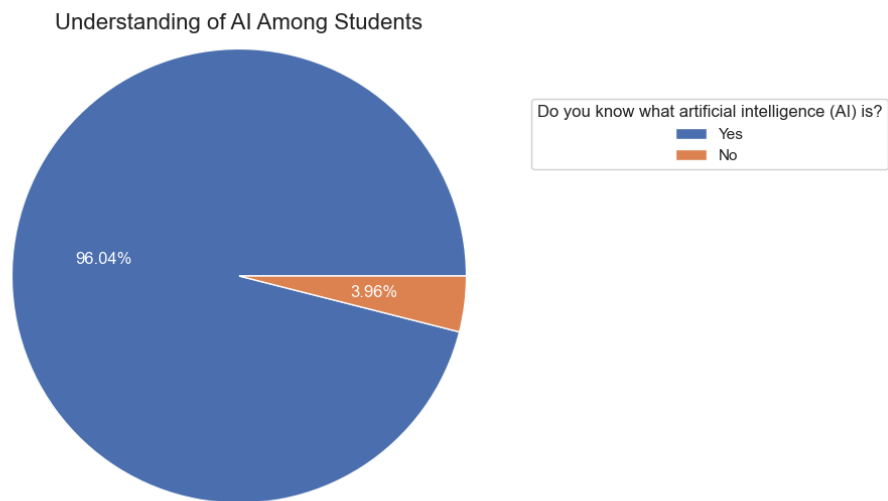


Figure 23.3

Understanding of AI Among Students per Grade Level

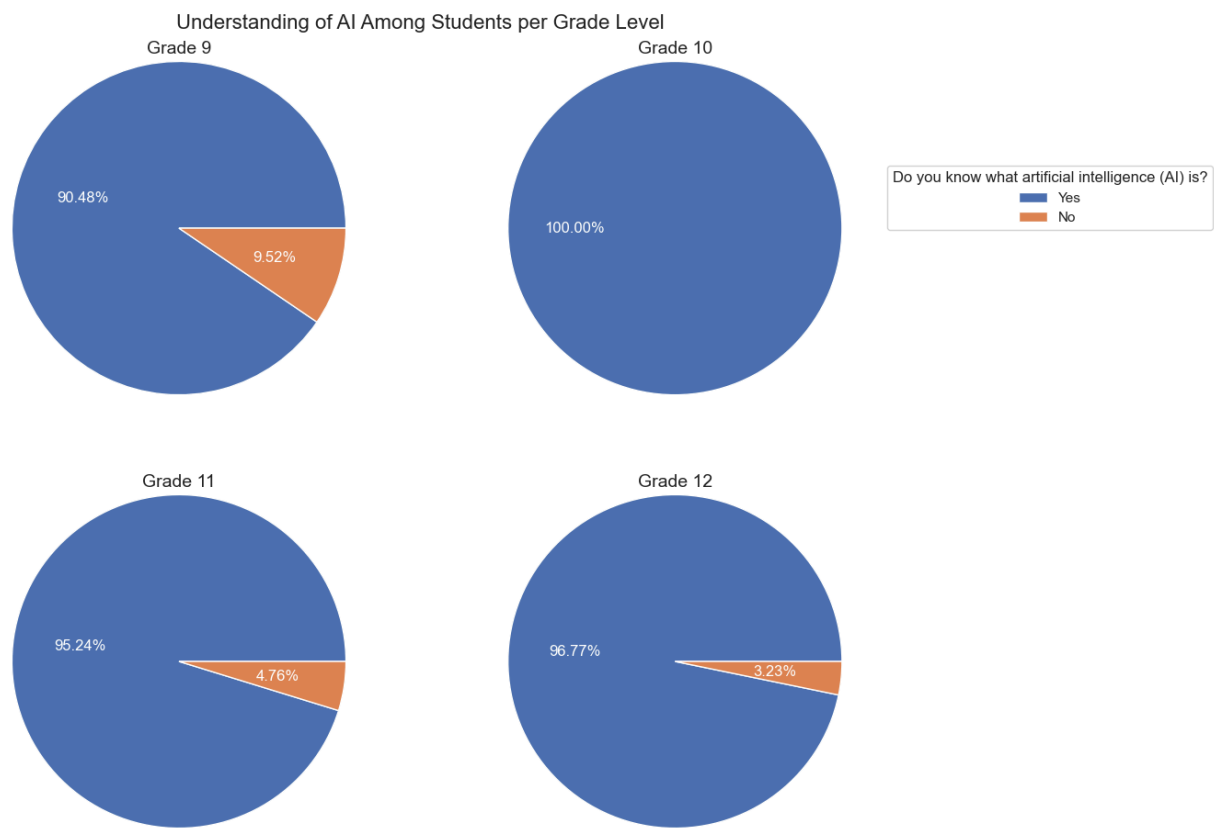


Figure 24.1

Which of The Following AI Tools Have You Used?

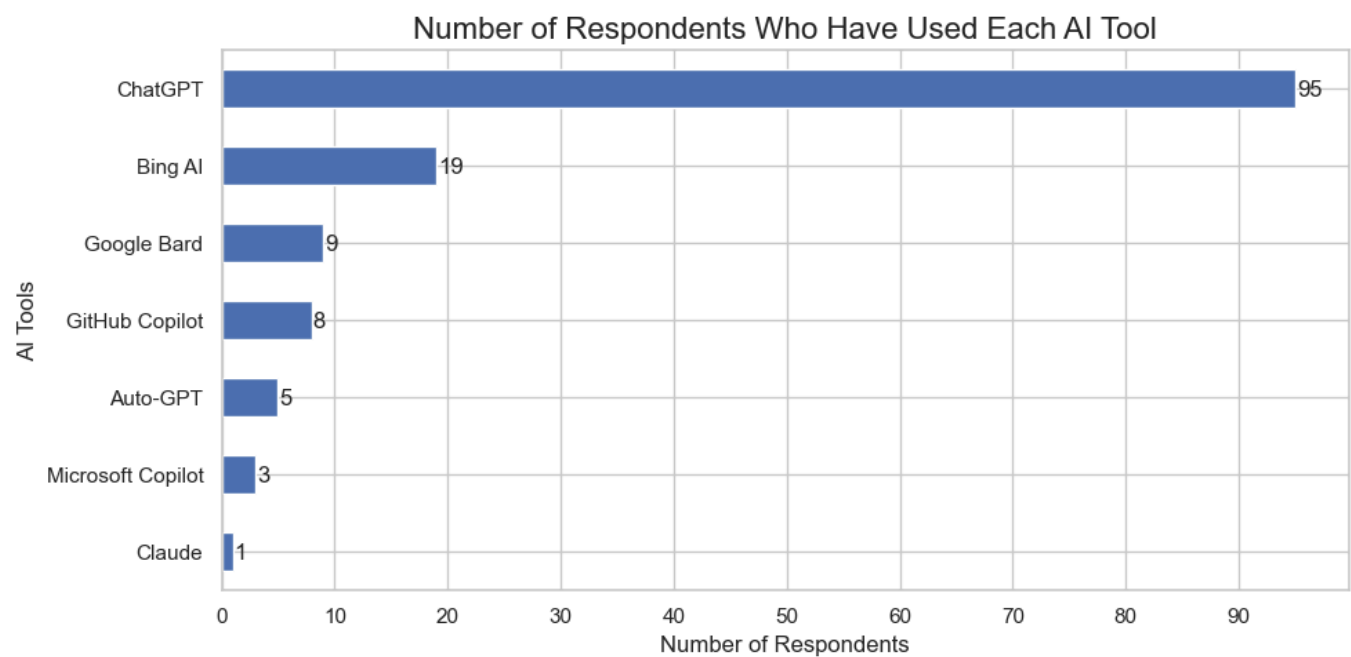


Figure 24.2

Number of Respondents by Grade Level for Each AI Tool

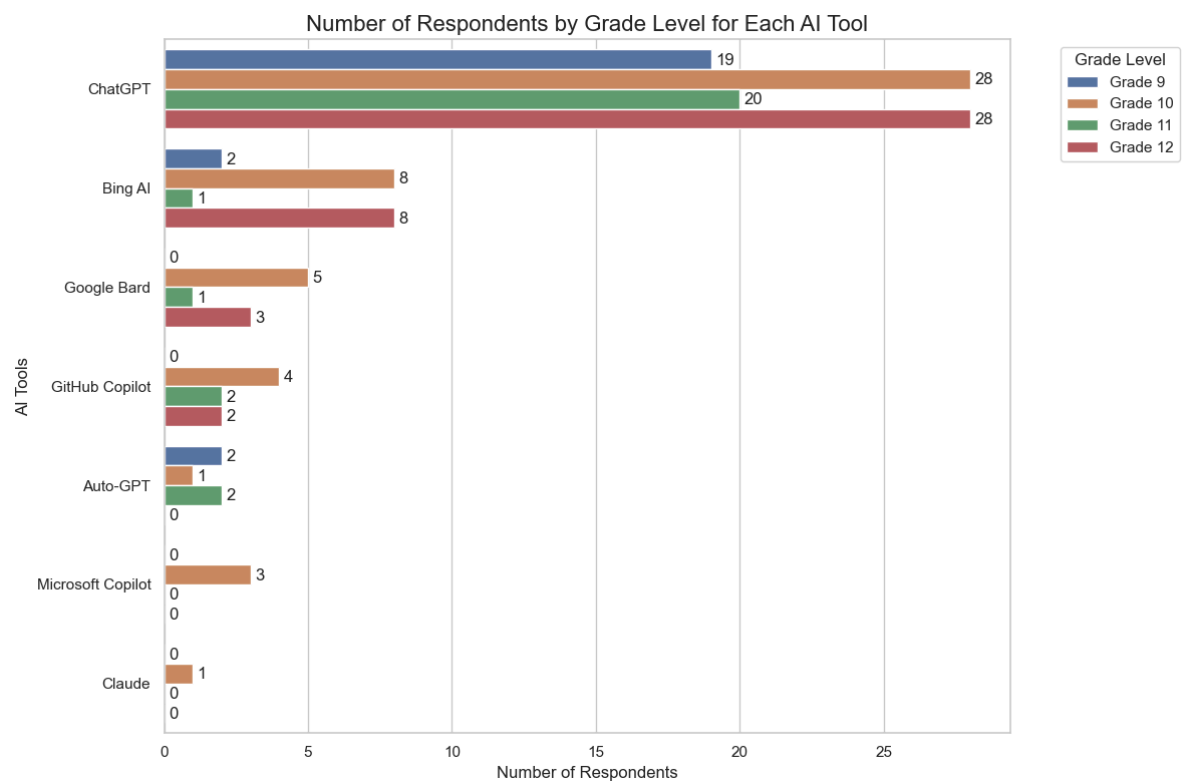


Figure 24.3

Percentage of Respondents by Grade Level for Each AI Tool

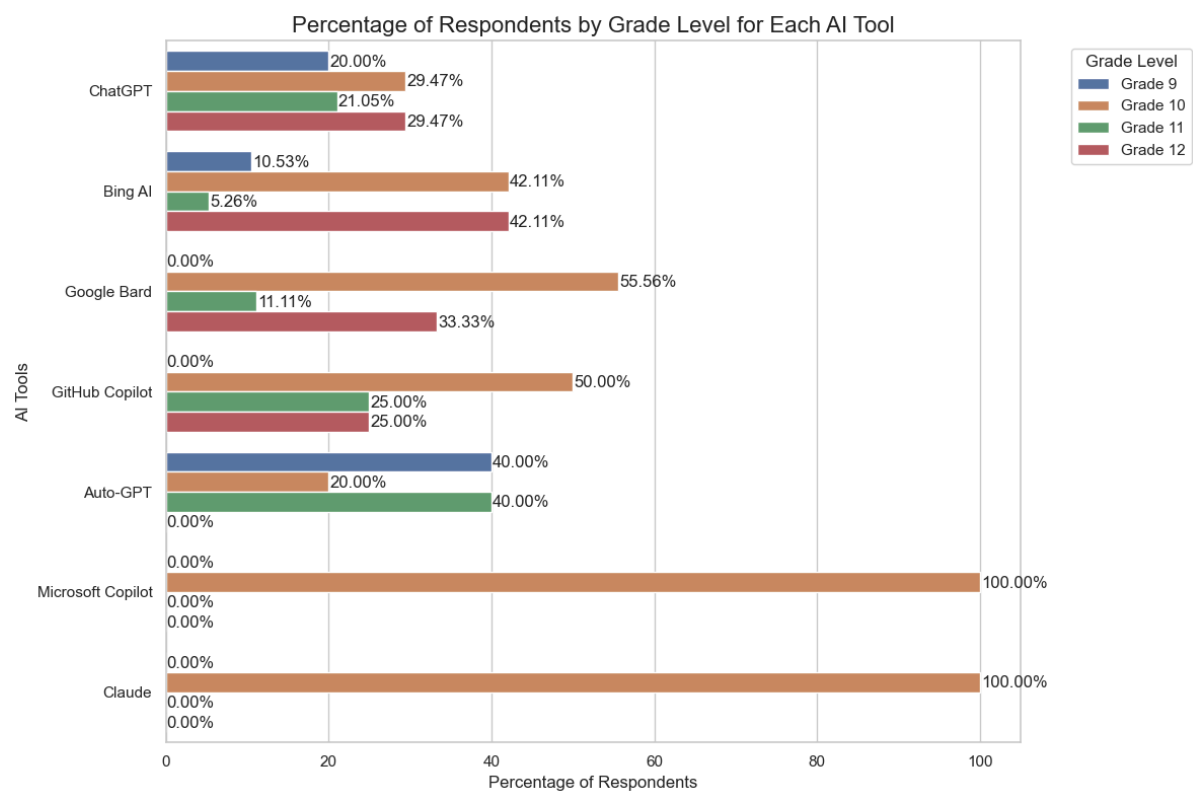


Figure 24.4

Do You Think Generative Text Models (Like ChatGPT) Have the Ability to Think?

Perceptions on the Thinking Abilities of AI Among Students

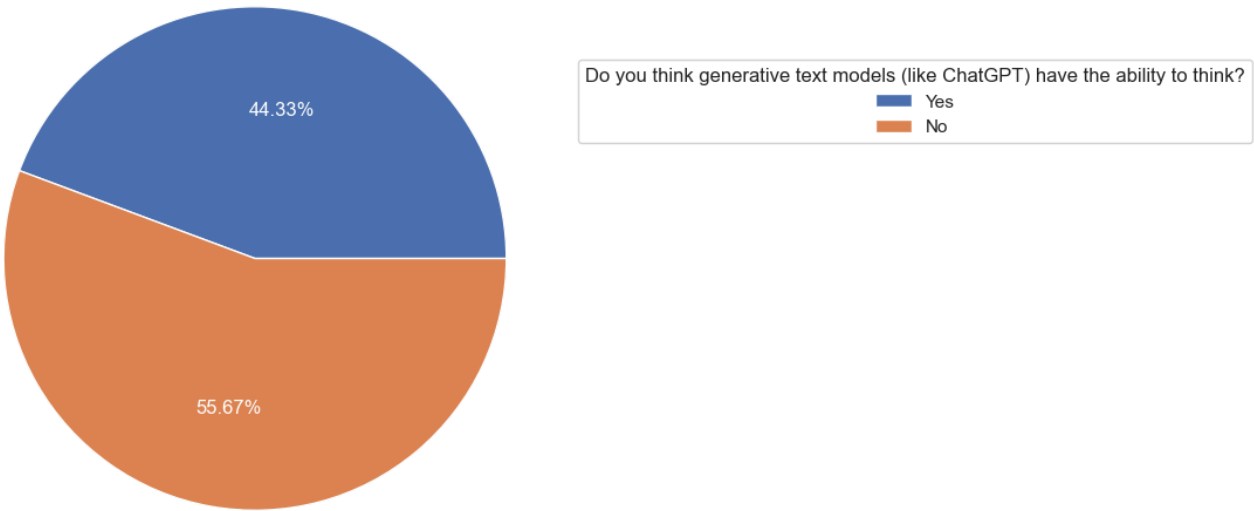


Figure 24.5

Perceptions on the Thinking Abilities of AI Among Students per Grade Level

Perceptions on the Thinking Abilities of AI Among Students per Grade Level

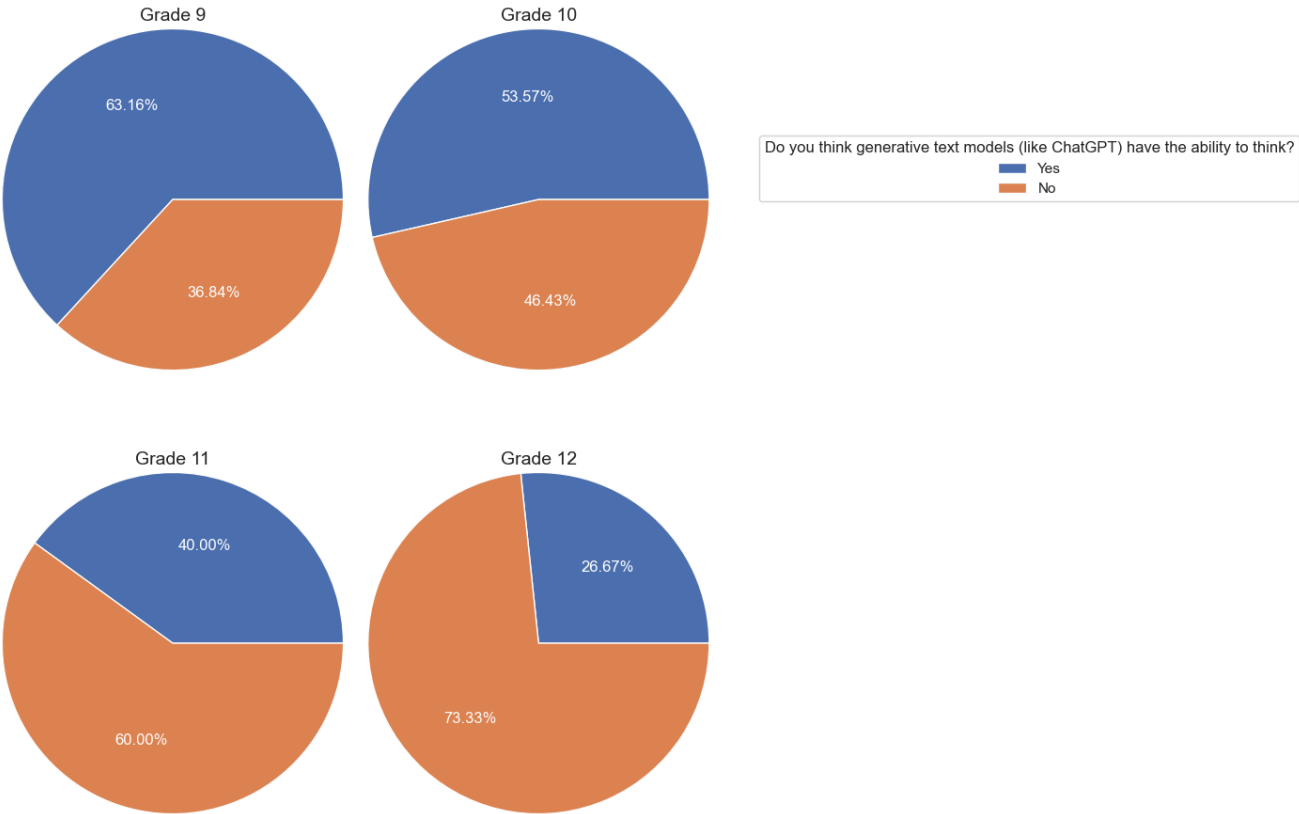


Figure 24.6

Do You Think These Models are Sentient (Meaning That They Have the Ability to Experience Feelings)?

Perceptions on AI Sentience Among Students

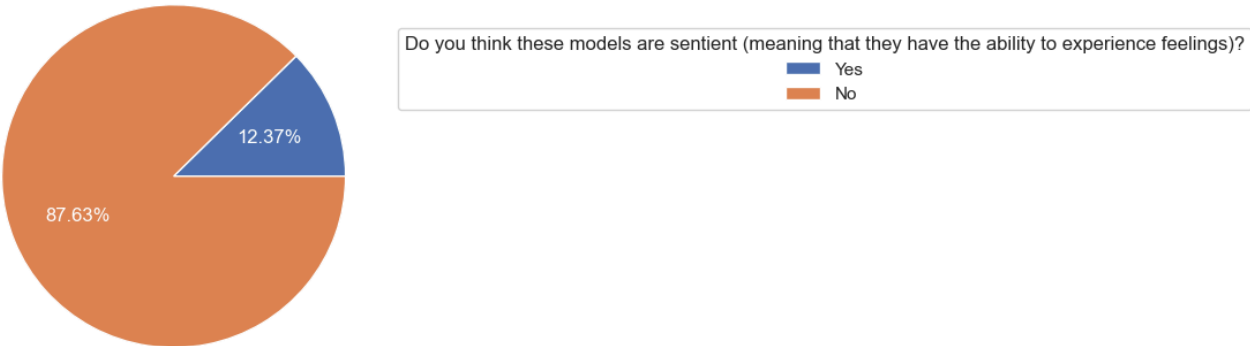


Figure 24.7

Perceptions on AI Sentience Among Students per Grade Level

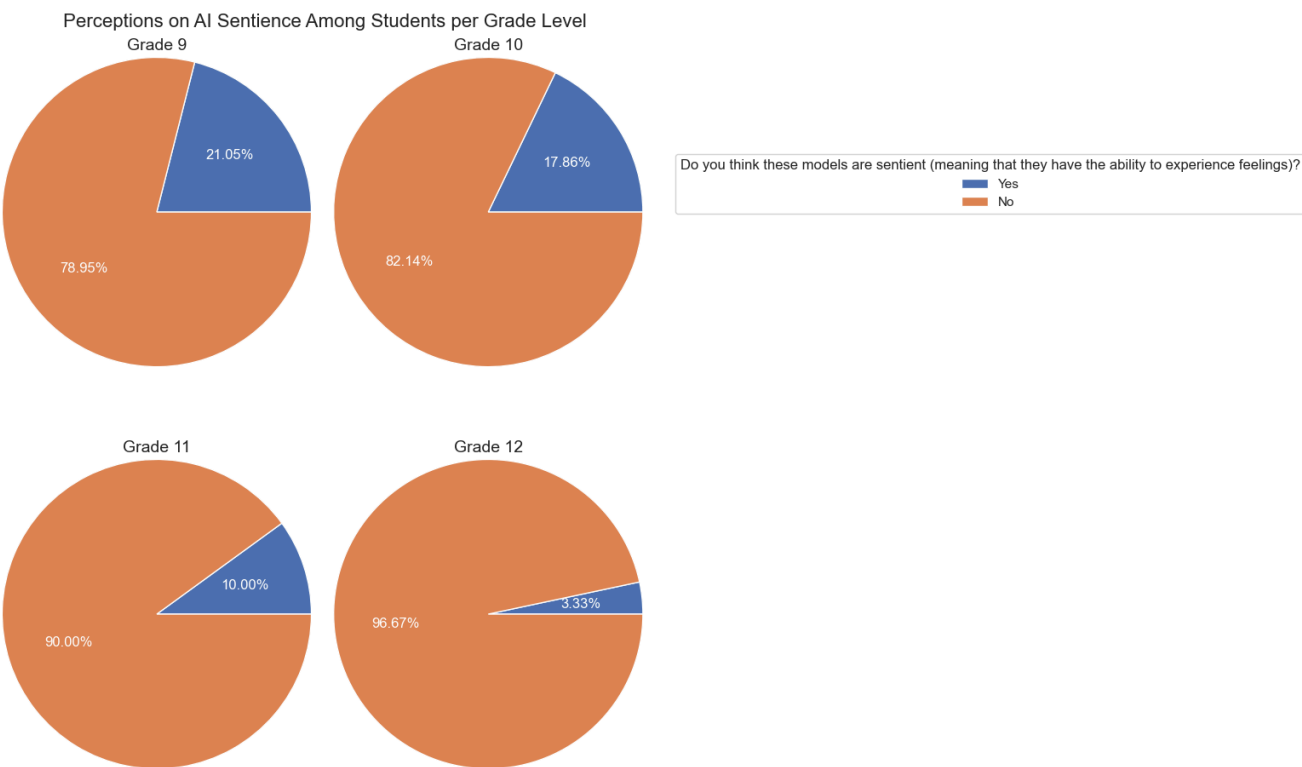


Figure 24.8

How Do You Think These Models Work?

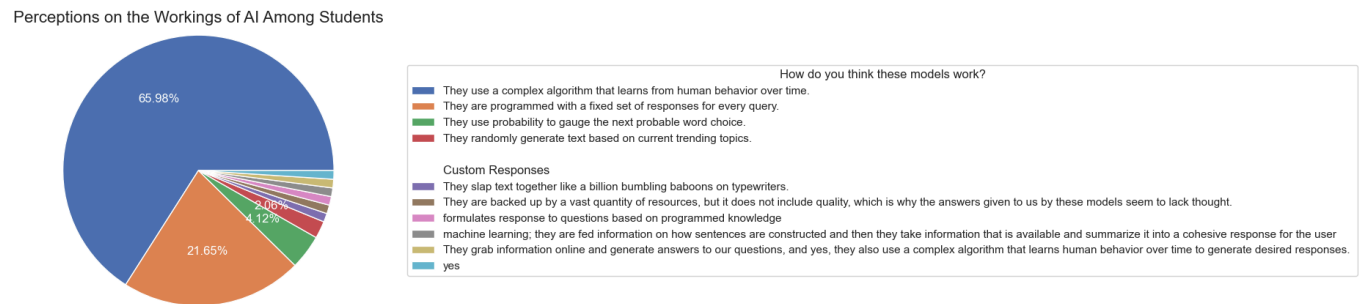


Figure 24.9

Perceptions on the Workings of AI Among Students

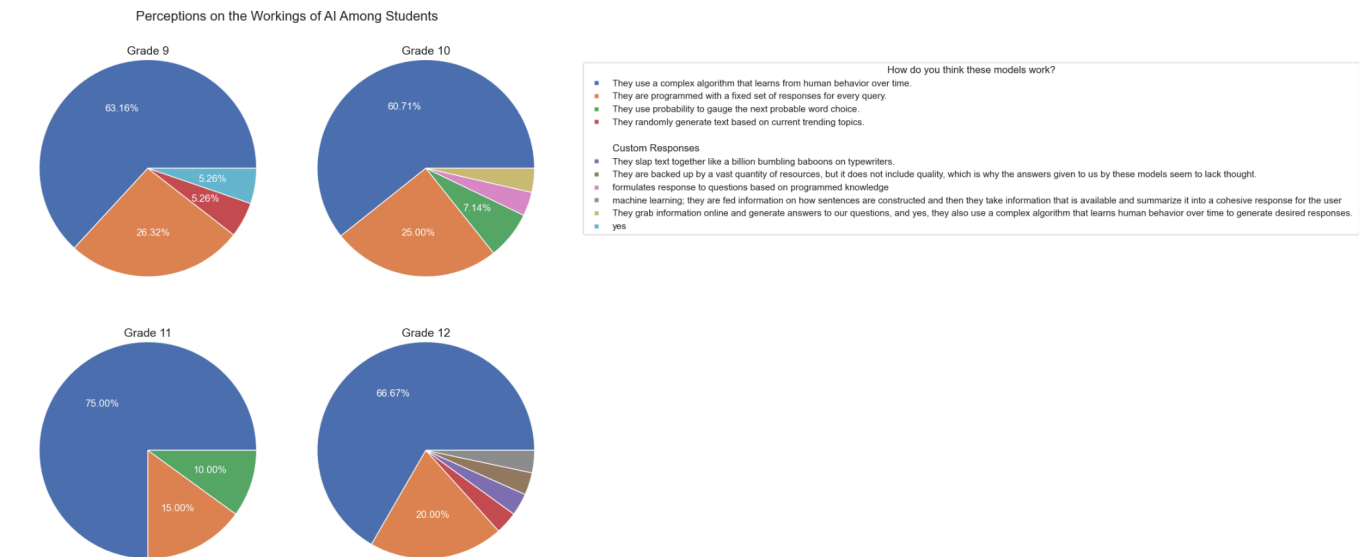


Figure 25.1

Real or fake:
A British man, accused of posting a picture of four-year-old Prince George and the address of his London school as part of a series of possible targets for Islamist militants, will face trial from April 30, a London court said on Wednesday. Husnain Rashid, 31, is accused of posting information on the Telegram messaging service to encourage jihadis to carry out attacks, along with information to help them with possible targets such as sports venues. He appeared at London's Old Bailey central criminal court charged with preparing acts of terrorism, and was remanded in custody. His four-week trial will take place at Woolwich Crown Court in south London.

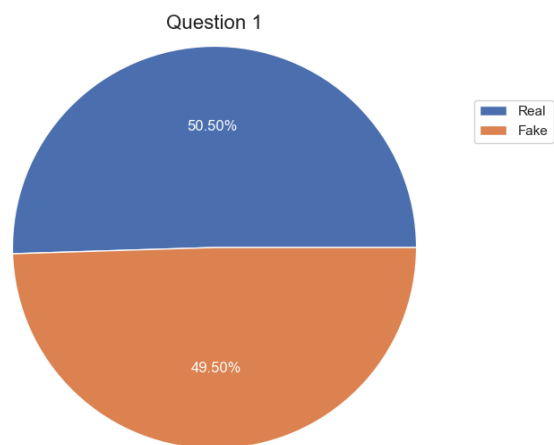


Figure 25.2

Question 1 Answer Distribution per Grade Level

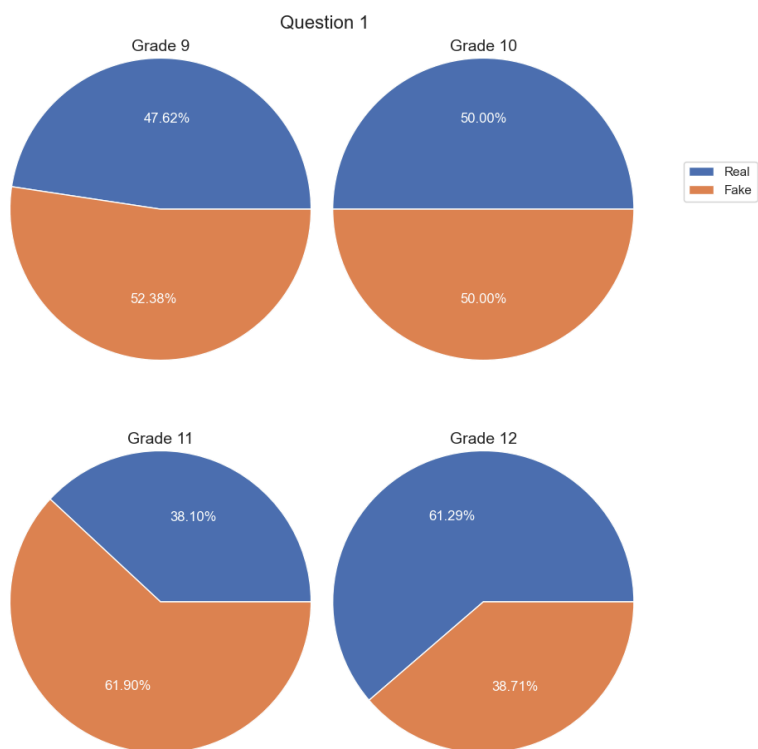


Figure 25.3

*Real or fake:
Harvey Weinstein was among a host of Hollywood actors and producers that donated \$10,000 to former President Bill Clinton's defense fund during his 1998 impeachment trials. The Washington Post published a story in 1998 reporting Harvey Weinstein as one of 62 donors giving the maximum \$10,000. Weinstein donated along with a number of other notable names, like Tom Hanks, Barbra Streisand, Michael Douglas, Ron Howard, Norman Lear and Steven Spielberg. WaPo also reported that Clinton's defense fund collected \$2.2 million in its first six months, tapping into the resentment against the attorney that investigated Clinton as to whether he committed perjury or obstructed justice in a sexual assault lawsuit.*

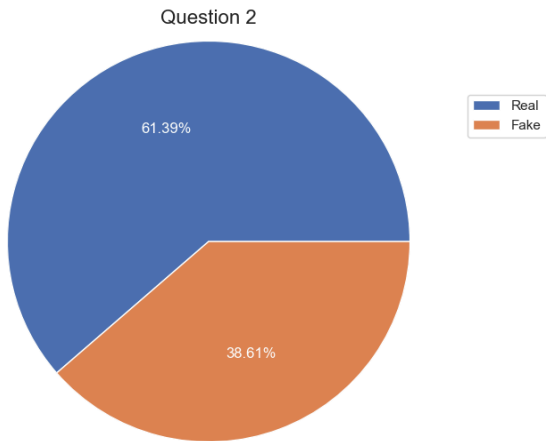


Figure 25.4

Question 2 Answer Distribution per Grade Level

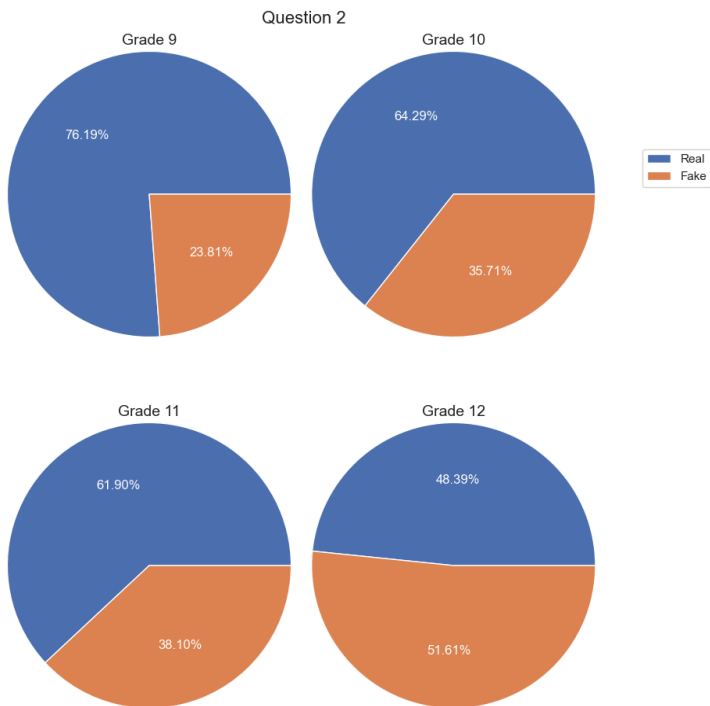


Figure 25.5

*Real or fake:
President Donald Trump celebrated the news that 227,000 jobs were created in January, meeting with his Economic Advisory Council at the White House on Friday morning. "Great spirit in the country right now," Trump said, speaking to the press. "So we're very happy about that. I think that it's going to continue big league. We're bringing back jobs. We're bringing down your taxes. We're getting rid of regulations." Trump promised "exciting times" ahead, previewing a tax bill "soon" and a healthcare bill "even sooner." "Amazingly, it's never happened before that politics has become a much bigger subject than the Super Bowl," he said. "This is usually Super Bowl territory." Trump told the business leaders that he wanted to hear from them as he tackled more reform proposals to make their businesses better, including getting rid of the cumbersome legislation.*

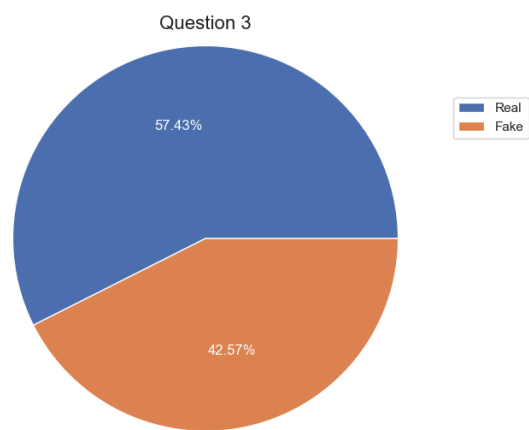


Figure 25.6

Question 3 Answer Distribution per Grade Level

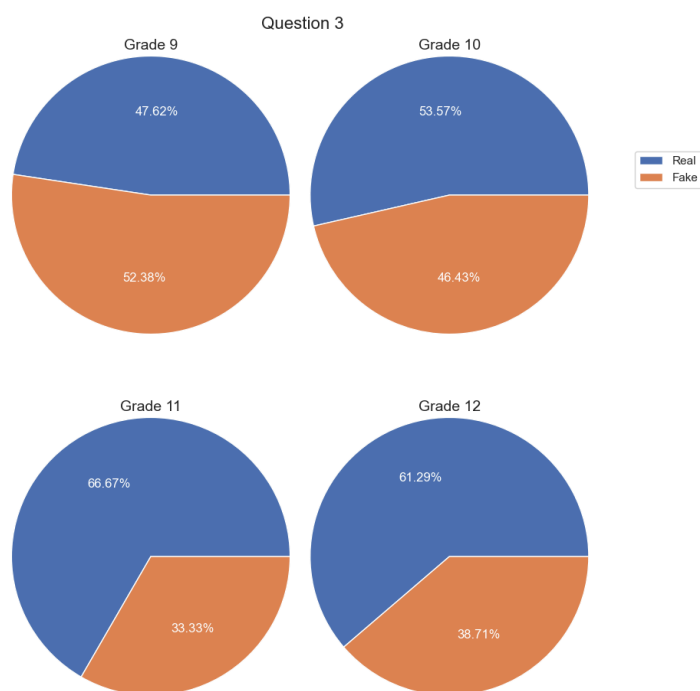


Figure 25.7

*Real or fake:
A Baltimore man was sentenced to four years behind bars for illegally trafficking food stamps. U.S. District Judge Richard D. Bennett sentenced Mohammad Shafiq, 51, to four years in prison and three years of supervised release after his sentence ends, in addition to making him pay \$3.7 million in restitution, the Baltimore Sun reported. Shafiq was the latest defendant to be sentenced in a series of prosecutions of 14 retailers in the Baltimore area. A federal grand jury indicted the retailers in August 2016 for food stamp and wire fraud. The 14 retailers stole a total of \$16 million from the U.S. Department of Agriculture by illegally exchanging food stamps for cash, according to the indictment. Twelve out the 14 defendants pleaded guilty, and two were sentenced this week, according to the Sun. Multiple stores across the country have been cited for millions of dollars in food stamp fraud. Investigators found more than \$20 million worth of food stamp fraud at retailers in Florida, and 140 stores in Chicago were cited for food stamp fraud.*

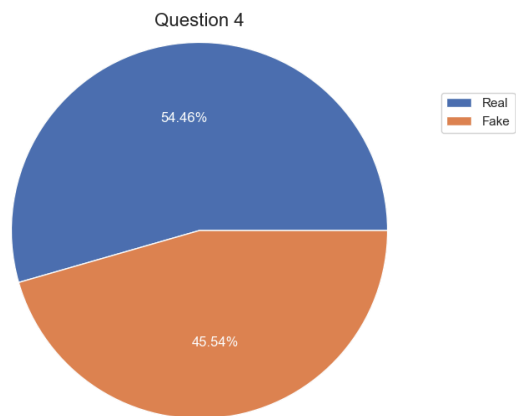


Figure 25.8

Question 4 Answer Distribution per Grade Level

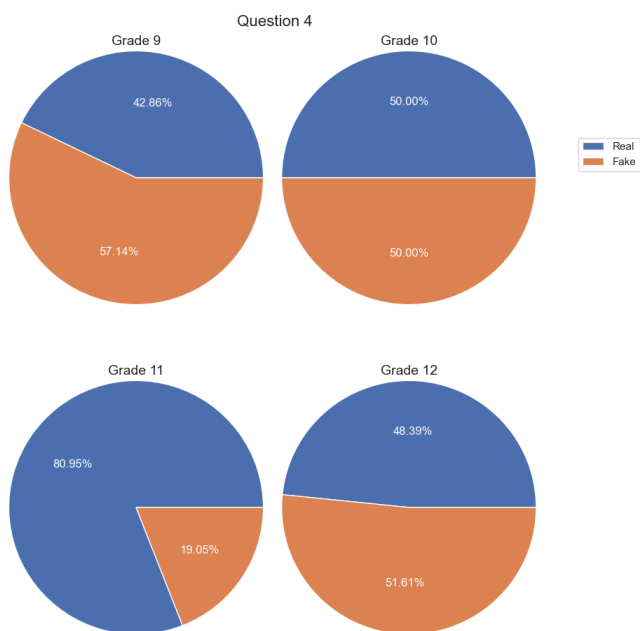


Figure 25.9

Real or fake:
The State Department misplaced and lost some \$6 billion due to the improper filing of contracts during the past six years, mainly during the tenure of former Secretary of State Hilary Clinton, according to a newly released Inspector General report. The \$6 billion in unaccounted funds poses a significant financial risk and demonstrates a lack of internal control over the Department's contract actions, according to the report. The alert, originally sent on March 20 and just released this week, warns that the missing contracting funds could expose the department to substantial financial losses. The report centered on State Department contracts worth more than \$6 billion in which contract files were incomplete or could not be located at all, according to the alert. The failure to maintain contract files adequately creates significant financial risk and demonstrates a lack of internal control over the Department's contract actions, the alert states.

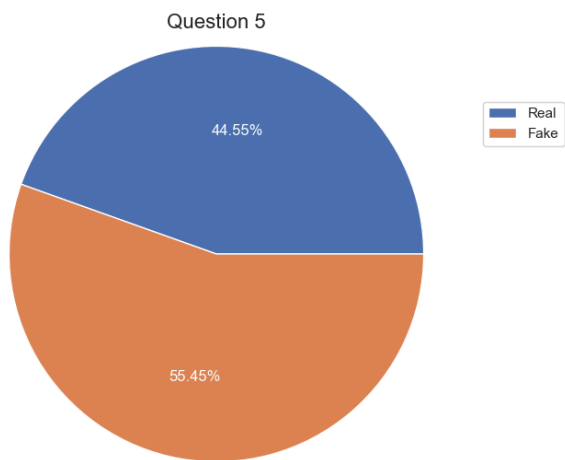


Figure 25.10

Question 5 Answer Distribution per Grade Level

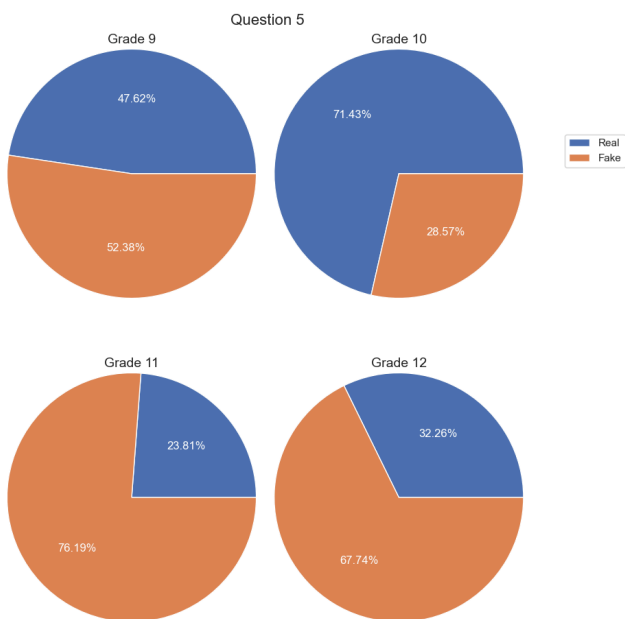


Figure 25.11

Total Score Distribution

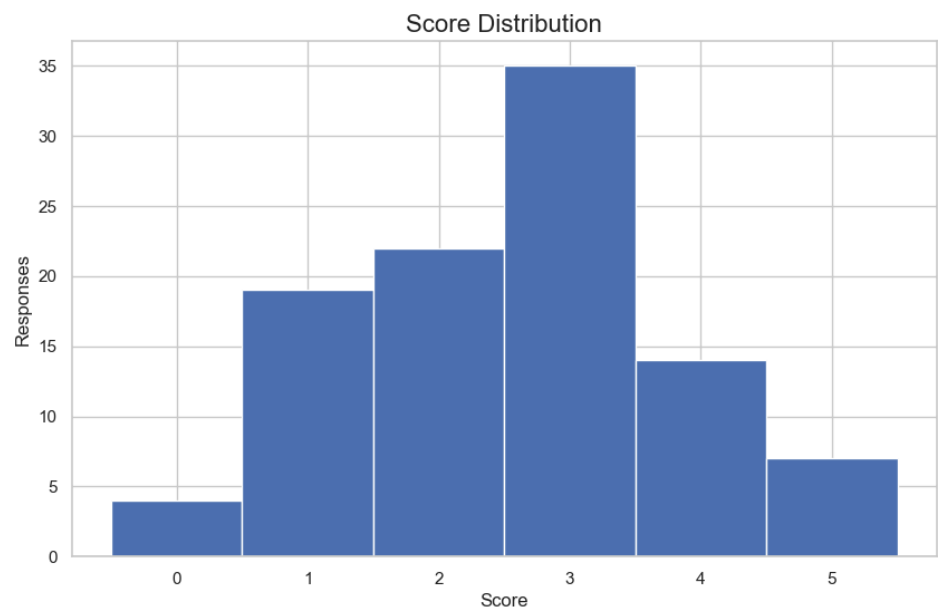


Figure 25.12

Total Score Distribution (Box Plot)

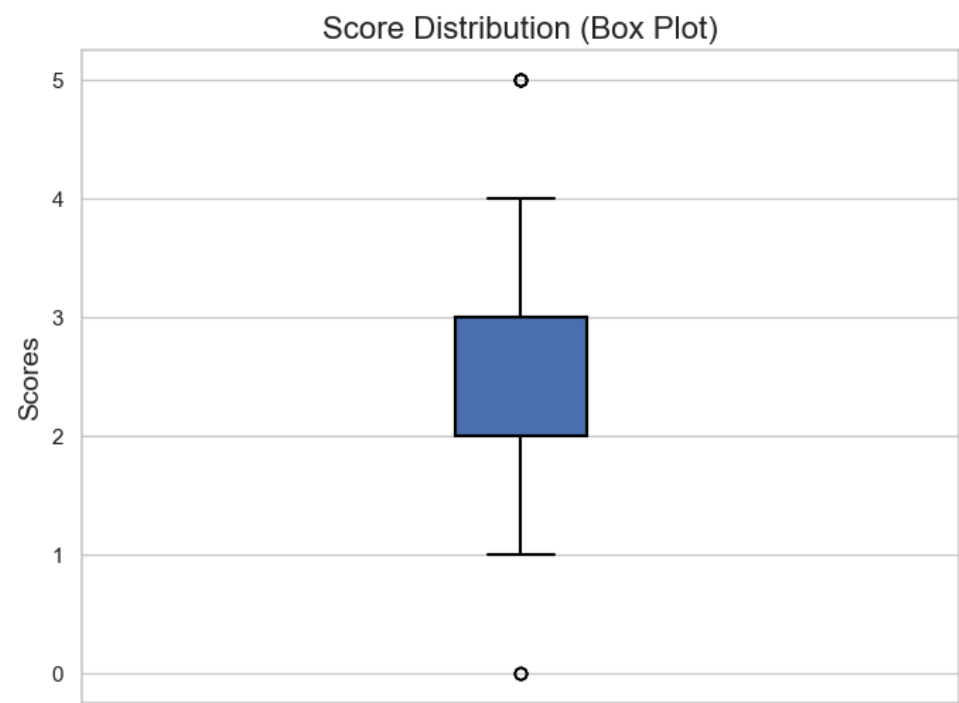


Figure 25.13

Score Distribution per Grade Level

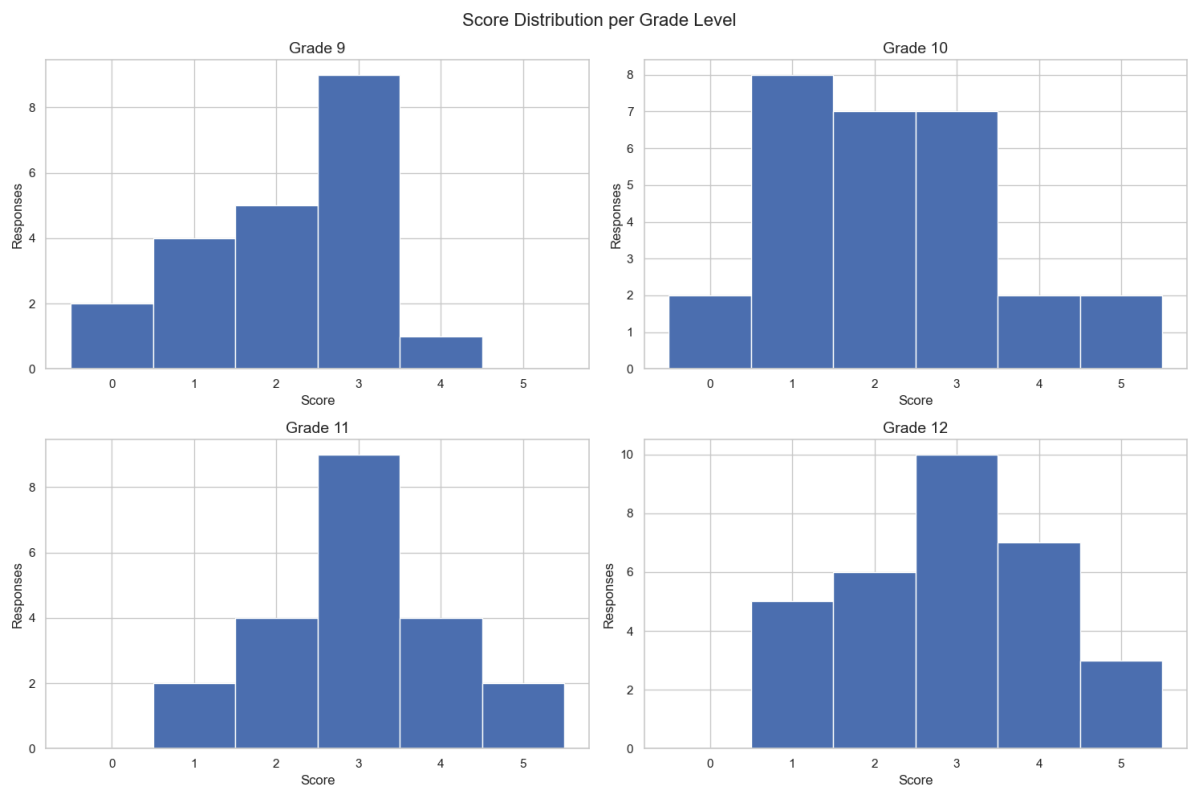


Figure 25.14

Score Distribution per Grade Level (Box Plot)

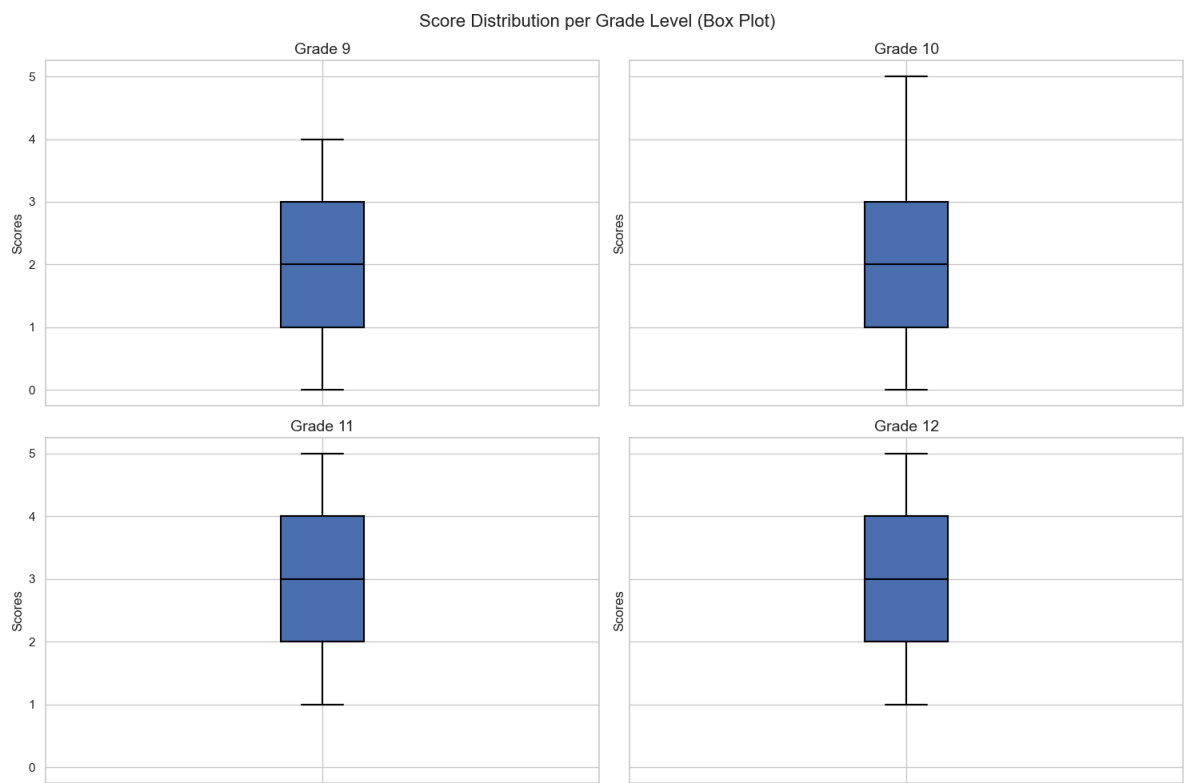


Figure 39.1

Score Distribution for Quiz 1

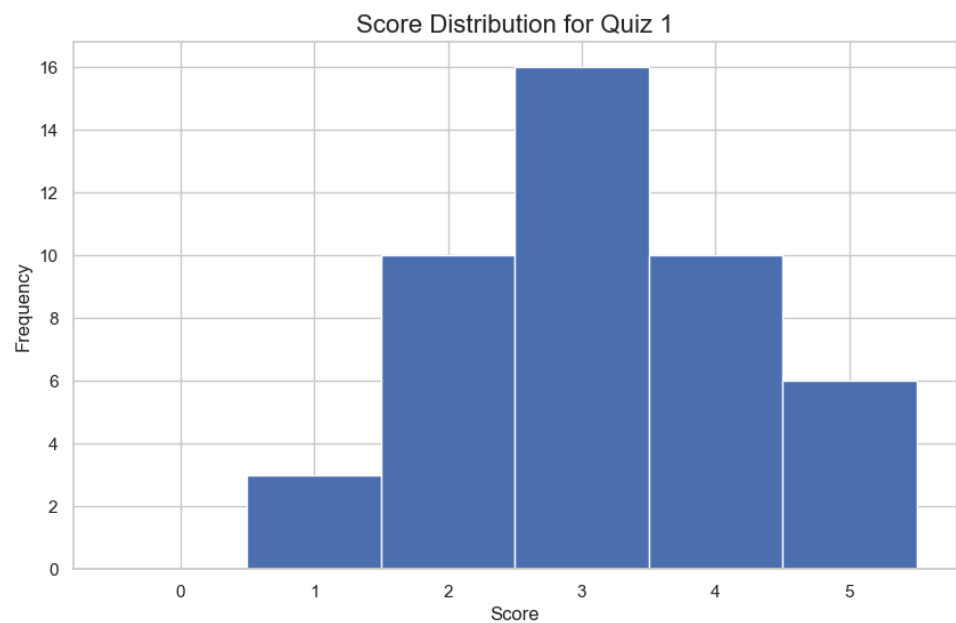


Figure 39.2

Score Distribution for Quiz 2

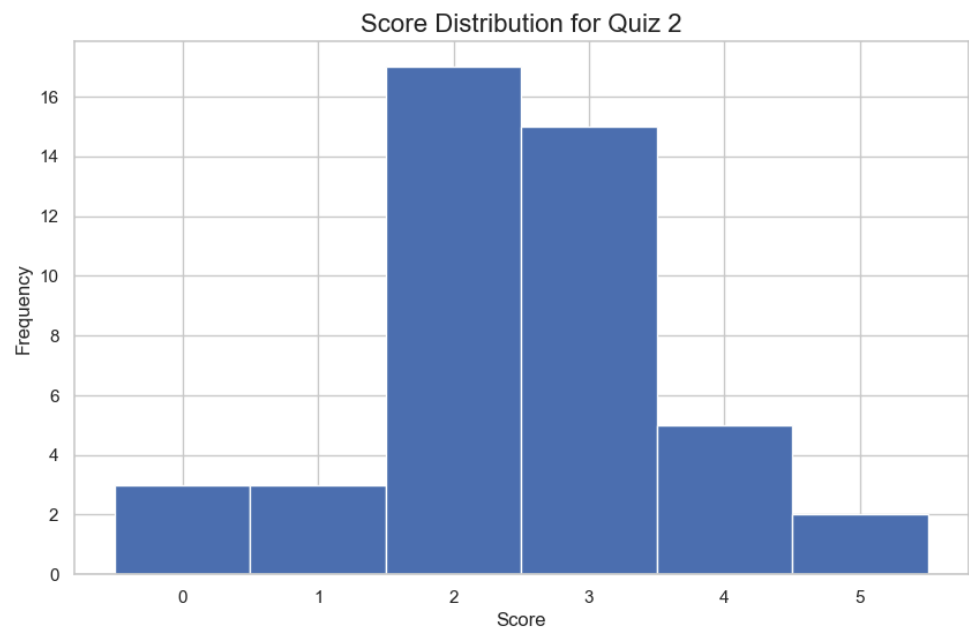


Figure 39.3

Score Distribution for Quiz 3

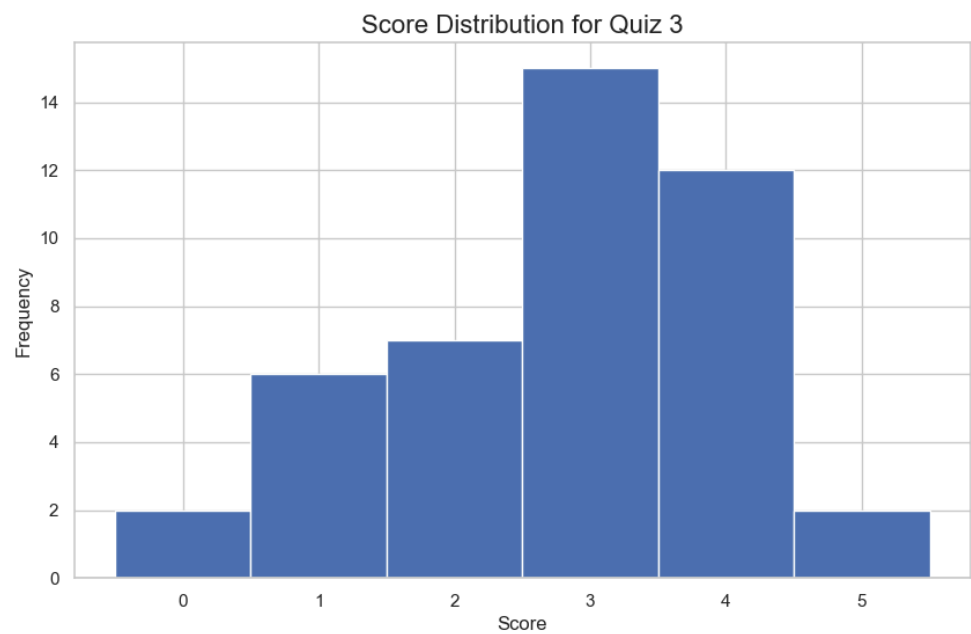


Figure 39.4

Score Distribution for Quiz 4

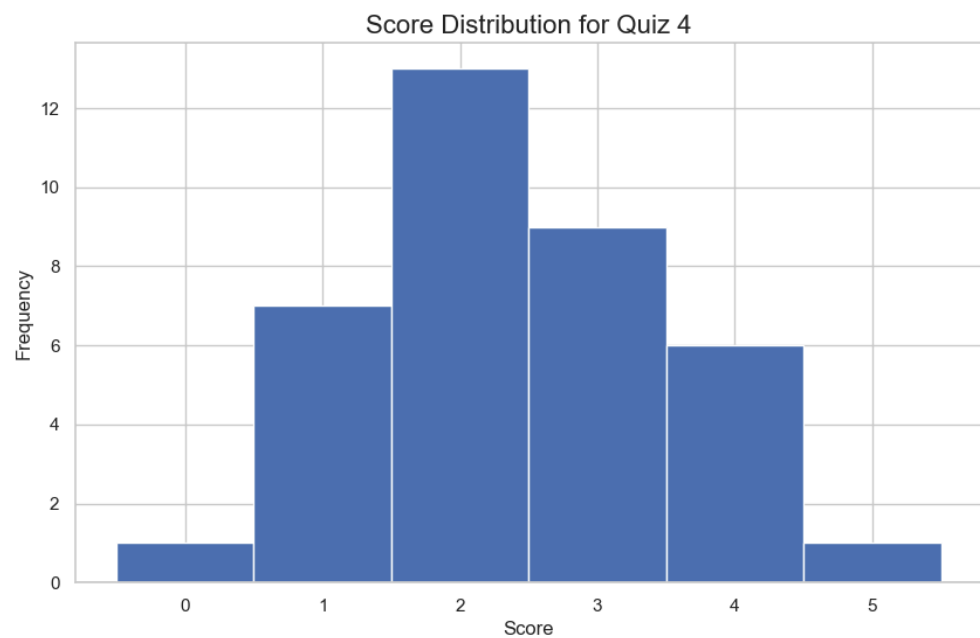


Figure 39.5

Score Distribution for Quiz 5

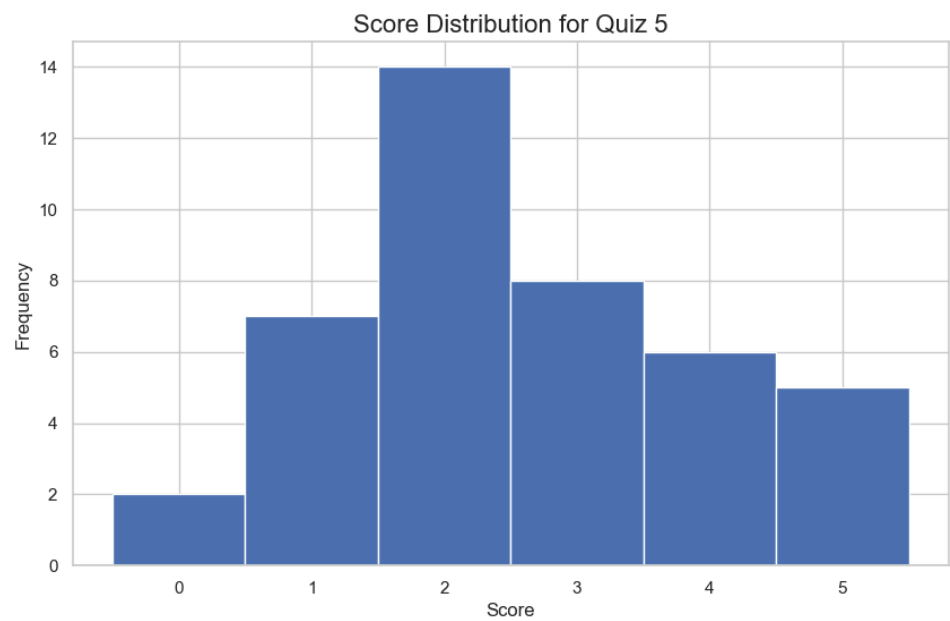
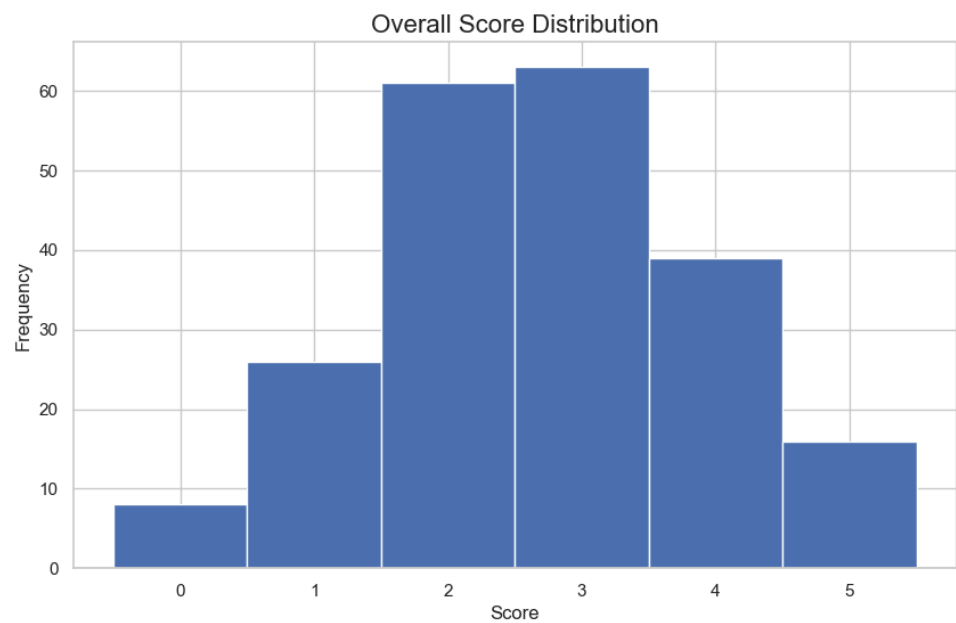


Figure 39.6

Overall Score Distribution Across the Five Quizzes



Appendices

Appendix A Initial Survey

Are you better than AI? A test of your disinformation- sniffing skills

Hey! I'm Kimi from grade 12 St. Pius.

My senior project aims to compare the accuracy of humans and AI at doing specialized tasks such as identifying disinformation. As you fill out this survey, **please don't Google anything!**

Everything is **multiple choice**, so please fill out this form!

** Indicates required question*

1. Grade level *

Mark only one oval.

- ☐ Grade 9
- ☐ Grade 10
- ☐ Grade 11
- ☐ Grade 12
- ☐ Teachers

2. Do you know what artificial intelligence (AI) is? *

Mark only one oval.

- ☐ Yes *Skip to question 3*
- ☐ No *Skip to question 7*

General background

3. Which of the following AI tools have you used?

Check all that apply.

- ☐ ChatGPT
- ☐ Bing AI
- ☐ Google Bard
- ☐ Claude
- ☐ Microsoft Copilot
- ☐ GitHub Copilot
- ☐ Auto-GPT

4. Do you think generative text models (like ChatGPT) have the ability to think? *

Mark only one oval.

- ☐ Yes
- ☐ No

5. Do you think these models are sentient (meaning that they have the ability to experience feelings)? *

Mark only one oval.

- ☐ Yes
- ☐ No

6. How do you think these models work? *

Mark only one oval.

- ☐ They are programmed with a fixed set of responses for every query.
- ☐ They use a complex algorithm that learns from human behavior over time.
- ☐ They use probability to gauge the next probable word choice.
- ☐ They randomly generate text based on current trending topics.
- ☐ Other: _____

The actual test

Please take a look at the following 5 news articles and determine if each of them is real or fake. These passages are all human-created; what you're looking for is whether the event is real or fake, and not whether it's generated by AI or human. The news must've had happened at some point in history for it to be considered "real"; it doesn't have to be current news. (Don't look this up on the internet, just try your best with your judgement.)

7. **Real or fake:**

*

A British man, accused of posting a picture of four-year-old Prince George and the address of his London school as part of a series of possible targets for Islamist militants, will face trial from April 30, a London court said on Wednesday. Husnain Rashid, 31, is accused of posting information on the Telegram messaging service to encourage jihadis to carry out attacks, along with information to help them with possible targets such as sports venues. He appeared at London's Old Bailey central criminal court charged with preparing acts of terrorism, and was remanded in custody. His four-week trial will take place at Woolwich Crown Court in south London.

Mark only one oval.

- ☐ Real
- ☐ Fake

8. **Real or fake:**

*

Harvey Weinstein was among a host of Hollywood actors and producers that donated \$10,000 to former President Bill Clinton's defense fund during his 1998 impeachment trials. The Washington Post published a story in 1998 reporting Harvey Weinstein as one of 62 donors giving the maximum \$10,000. Weinstein donated along with a number of other notable names, like Tom Hanks, Barbra Streisand, Michael Douglas, Ron Howard, Norman Lear and Steven Spielberg. WaPo also reported that Clinton's defense fund collected \$2.2 million in its first six months, tapping into the resentment against the attorney that investigated Clinton as to whether he committed perjury or obstructed justice in a sexual assault lawsuit.

Mark only one oval.

☐ Real

☐ Fake

9. **Real or fake:**

*

President Donald Trump celebrated the news that 227,000 jobs were created in January, meeting with his Economic Advisory Council at the White House on Friday morning. "Great spirit in the country right now," Trump said, speaking to the press. "So we're very happy about that. I think that it's going to continue big league. We're bringing back jobs. We're bringing down your taxes. We're getting rid of regulations." Trump promised "exciting times" ahead, previewing a tax bill "soon" and a healthcare bill "even sooner." "Amazingly, it's never happened before that politics has become a much bigger subject than the Super Bowl," he said. "This is usually Super Bowl territory." Trump told the business leaders that he wanted to hear from them as he tackled more reform proposals to make their businesses better, including getting rid of the cumbersome legislation.

Mark only one oval.

☐ Real

☐ Fake

10. **Real or fake:**

*

A Baltimore man was sentenced to four years behind bars for illegally trafficking food stamps. U.S. District Judge Richard D. Bennett sentenced Mohammad Shafiq, 51, to four years in prison and three years of supervised release after his sentence ends, in addition to making him pay \$3.7 million in restitution, the Baltimore Sun reported. Shafiq was the latest defendant to be sentenced in a series of prosecutions of 14 retailers in the Baltimore area. A federal grand jury indicted the retailers in August 2016 for food stamp and wire fraud. The 14 retailers stole a total of \$16 million from the U.S. Department of Agriculture by illegally exchanging food stamps for cash, according to the indictment. Twelve out the 14 defendants pleaded guilty, and two were sentenced this week, according to the Sun. Multiple stores across the country have been cited for millions of dollars in food stamp fraud. Investigators found more than \$20 million worth of food stamp fraud at retailers in Florida, and 140 stores in Chicago were cited for food stamp fraud.

Mark only one oval.

☐ Real

☐ Fake

11. **Real or fake:**

*

The State Department misplaced and lost some \$6 billion due to the improper filing of contracts during the past six years, mainly during the tenure of former Secretary of State Hilary Clinton, according to a newly released Inspector General report. The \$6 billion in unaccounted funds poses a significant financial risk and demonstrates a lack of internal control over the Department's contract actions, according to the report. The alert, originally sent on March 20 and just released this week, warns that the missing contracting funds could expose the department to substantial financial losses. The report centered on State Department contracts worth more than \$6 billion in which contract files were incomplete or could not be located at all, according to the alert. The failure to maintain contract files adequately creates significant financial risk and demonstrates a lack of internal control over the Department's contract actions, the alert states.

Mark only one oval.

☐ Real

☐ Fake

Appendix B*Source Code for All the Models***LSTM**

Chen, K. (2023). *Fake-news-detector-LSTM*. Hugging Face.

<https://huggingface.co/kimic/fake-news-detector-LSTM/tree/main>

LSTM + GloVe Embeddings

Chen, K. (2023). *Fake-news-detector-LSTM-GloVe*. Hugging Face.

<https://huggingface.co/kimic/fake-news-detector-LSTM-GloVe/tree/main>

DistilBERT

Chen, K. (2023). *fake-news-detector-DistilBERT*. Hugging Face.

<https://huggingface.co/kimic/fake-news-detector-DistilBERT/tree/main>

Llama-2

Chen, K. (2023). *Fake-news-detector-Llama-2*. Hugging Face.

<https://huggingface.co/kimic/fake-news-detector-Llama-2/tree/main>

GPT-4 Turbo

Chen, K. (2023). *Fake-news-detector-GPT-4-Turbo*. Hugging Face.

<https://huggingface.co/kimic/fake-news-detector-GPT-4-Turbo/tree/main>

Appendix C*Quiz 1 For Grade 9 Students*

Navigating (More) Disinformation Pt. 1

On the following page, you will see 5 media pieces (**news, forum posts, text messages, email, etc.**). Please determine if each of them is **real or fake**.

These passages are all human-created; what you're looking for is whether the event is real or fake, and not whether it's generated by AI or human. The media pieces must've had happened at some point in history for it to be considered "real"; it doesn't have to be relevant. (Don't look this up on the internet, just try your best with your judgement.)

** Indicates required question*

1. Email *

2. Which class are you in? *

Mark only one oval.

☐ St. Raymond

☐ St. Rose

Real or Fake

3. An eye-catching shoe piques your interest and draws you in. * 1 point
Upon closer inspection, the label leaps out at you — Ivanka Trump, in simple gold lettering — and you recoil as if stung. That's the kind of reaction behind a growing boycott of the products emblazoned with the brand of Republican presidential nominee Donald Trump as well as the popular, working women-targeted fashion line from his eldest daughter — who has arguably been his most influential and effective family member during the current election campaign.

Mark only one oval.

☐ Real

☐ Fake

4. The national Iranian women's inline hockey team has captured its second victory in a row at the preliminary round of the 17th edition of Asian Roller Skating Championships in China, overcoming the strong South Korean side. On Wednesday morning, the Iranian sportswomen gained a narrow 1-0 win over the East Asian contingent in a match staged in the eastern Chinese city of Lishui. Faezeh Modabber scored the sole goal of the game. The Iranian inline hockey players also got two other goals, but the referee disallowed them. The Iranian outfit was scheduled to face China later in the day. The 17th edition of Asian Roller Skating Championships started in Lishui on October 23 and will wrap up on November 1. The sports event has brought together dozens of male and female roller skaters from various Asian countries, including China, the Islamic Republic of Iran and South Korea. * 1 point

Mark only one oval.

☐ Real

☐ Fake

5. Boston Mayor Martin Walsh, a former labor leader, said on Monday that federal agents had not contacted him in connection with what local media has described as a wide-ranging investigation into whether the city's building unions used strong-arm tactics. The Boston Globe reported over the weekend, citing unnamed sources, that Walsh had been implicated in a federal investigation into whether city union officials threatened developers who hired nonunion workers on projects in and around the city. "I haven't been contacted myself," Walsh told reporters on Monday. A Democrat, Walsh is a former construction worker who led the city's Building and Construction Trades Council, a union grouping, for two years before his 2013 mayoral election victory. Walsh, who served as a state representative from 1997 through 2013, went on to say that even if there were an investigation into his former union, he was not concerned. "If there is an investigation, I'm assuming at some point there'll be indictments coming down. If that's the case, I will not be getting one of those," Walsh said. "Because I did nothing wrong." A spokeswoman for the U.S. Attorney's Office in Boston could not be reached for immediate comment.

★ 1 point

Mark only one oval.

☐ Real

☐ Fake

6. Former Speaker Newt Gingrich discusses the Russia investigation with Hannity. He drops a bomb when he says that Obama should testify under oath before Congress in the Russia probe. We totally agree! ★ 1 point
- Gingrich: "There's no question that Congress should call the former president in to testify under oath to explain what he was doing and why he was doing it," Gingrich said. "There's no question that they should build the case from the ground up. Who was doing the investigating? Who was reporting to the president? Who did it go through? What were the meetings like when they decided not to pursue it? I mean, talk about an extraordinary failure of national security. Guess what's about to happen, the Congress is about to have to call Barack Obama in to testify under oath about when he knew about Russia meddling. Who told him? Why did he do nothing? Who was in the meetings when he decided to do nothing? It's going to turn out, yes, there's a big Russian story. It's Barack Obama, not Donald Trump. And you cannot make this stuff up."

Mark only one oval.

☐ Real

☐ Fake

7. UFO over the east valley here in AZ last night. I couldn't take any pictures I live in West Mesa and the lights were way East. Over Queen Creek area. Lots of my friends and family started calling me right when the lights started happening so I went out to my balcony and there they were. Some people got some pretty good videos. I know some people will say they are flares from a plane but if they are flares why do they disappear right away. Typically if it's a flare it will stay ignited for a short period of time. Kinda nuts. I only witnessed the later 3 lights when it was dark out. It was pretty cool to see and stopped soon after it started. What does ATS think???
- ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

Appendix D*Quiz 2 For Grade 9 Students*

Navigating (More) Disinformation Pt. 2

On the following page, you will see 5 media pieces (**news, forum posts, text messages, email, etc.**). Please determine if each of them is **real or fake**.

These passages are all human-created; what you're looking for is whether the event is real or fake, and not whether it's generated by AI or human. The media pieces must've had happened at some point in history for it to be considered "real"; it doesn't have to be relevant. (Don't look this up on the internet, just try your best with your judgement.)

** Indicates required question*

1. Email *

2. Which class are you in? *

Mark only one oval.

☐ St. Raymond

☐ St. Rose

Real or Fake

3. U.S. Republican Senator Jeff Flake on Thursday expressed concern about the Senate Republican tax proposal's impact on the national debt. In a statement, Flake did not say how he might vote on the proposal unveiled earlier Thursday. "I remain concerned over how the current tax reform proposals will grow the already staggering national debt by opting for short-term fixes while ignoring long-term problems for taxpayers and the economy," he said. Both the Senate and House tax cut plans would add about \$1.5 trillion to the U.S. national debt over 10 years. Lawmakers should achieve tax reform in a fiscally responsible manner, Flake said, adding that he looked forward to working with his colleagues "to deliver on that goal."
- * 1 point

Mark only one oval.

☐ Real

☐ Fake

4. For the first time in centuries, scientists have exposed the original surface of what is traditionally considered the tomb of Jesus Christ. Located in the Church of the Holy Sepulchre in the Old City of Jerusalem, the tomb has been covered by marble cladding since at least 1555 A.D., and most likely centuries earlier. "The marble covering of the tomb has been pulled back, and we were surprised by the amount of fill material beneath it," said Fredrik Hiebert, archaeologist-in-residence at the National Geographic Society, a partner in the restoration project. "It will be a long scientific analysis, but we will finally be able to see the original rock surface on which, according to tradition, the body of Christ was laid."
- * 1 point

Mark only one oval.

☐ Real

☐ Fake

5. What was the legal rationale under which Hillary Clinton quite intentionally shared classified information with her lawyers, including David Kendall, Cheryl Mills, and Heather Samuelson? As I outlined in last weekend's column, we know that Clinton's e-mails were replete with classified information. According to the FBI, the classified e-mails included intelligence graded at the most closely guarded level: eight top-secret e-mails, and seven designated as "special access program" (SAP) information. (While FBI director James Comey's presentation understandably left this vague, the likelihood is that seven of the eight top-secret e-mails are SAP.) Under President Bill Clinton's 1995 executive order, top-secret intelligence is information the mishandling of which "could be expected to cause exceptionally grave damage to the national security." The SAP designation is added when the unauthorized disclosure of intelligence could compromise critical intelligence-gathering methods or imperil the lives of intelligence sources. That is why access to this information is so tightly restricted, and its unauthorized disclosure is routinely prosecuted.

★ 1 point

Mark only one oval.

☐ Real

☐ Fake

6. "We need more and more jobs. Unemployment is way up here," he said. "He's hired more employees, more people, than anyone I know in the world." I believe in him first of all because he's a businessman. I think jobs are badly needed. Hollywood star, Brad Pitt is usually catching some heat through supporters of President Obama after calling his agenda "anti-Christian." Pitt talked out about his own Christian beliefs during the 1999 interview. During the interview, he shared his story of how he or she uncovered religion. As the particular son of a preacher and spending a majority of his time in the church, he grew to become rebellious as a teen. After what Pitt calls a personal face along with God, he came back in order to the church in his 20s. "I had this particular tremendous physical and spiritual experience." His devotion to his faith has shocked liberal Hollywood plus angered President Obama supporters who support his alleged "anti-Christian agenda." As other celebrities live a liberal and free life-style, Pitt has explained that he is proud of his religion. Because of this, Brad has finally come out in support of Donald Trump and his run for Presidency. ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

7. Russia calls on Western countries to persuade Ukraine to carry out its obligations as part of the Minsk settlement process, the Russian Foreign Ministry said Friday in remarks announcing preparations for four-way talks next week. "Russia sees its task in encouraging Western partners who have influence on the leadership of Ukraine to urge Kiev to strictly follow the obligations assumed in the framework of the Minsk process," the ministry said. The ministry said it was preparing for Normandy format talks in the Belarusian capital of Minsk next Tuesday, November 29, after receiving an invitation from Germany and France.

★ 1 point

Mark only one oval.

☐ Real

☐ Fake

Appendix E*Quiz 3 For Grade 9 Students*

Navigating (More) Disinformation Pt. 3

On the following page, you will see 5 media pieces (**news, forum posts, text messages, email, etc.**). Please determine if each of them is **real or fake**.

These passages are all human-created; what you're looking for is whether the event is real or fake, and not whether it's generated by AI or human. The media pieces must've had happened at some point in history for it to be considered "real"; it doesn't have to be relevant. (Don't look this up on the internet, just try your best with your judgement.)

** Indicates required question*

1. Email *

2. Which class are you in? *

Mark only one oval.

☐ St. Raymond

☐ St. Rose

Real or Fake

3. If you've ever watched the Hunger Games movies you know that Effie Trinket is a psychotic woman who looks at murder as a game. She plays along with the Hunger Games and expects everyone else to just ignore the horror going on around them. She sends the young tributes out to play the game knowing all but one will be murdered. Hillary Clinton is our very own Effie Trinket playing along in a very sick game of politics and murder. Wonder if she's even thinking about the families of the murdered American heroes. * 1 point

Mark only one oval.

☐ Real

☐ Fake

4. Reddit users, also known as Redditors, are growing increasingly angry at the pro-Hillary Clinton Political Action Committee (PAC), Correct the Record, over its efforts to promote a pro-Hillary narrative on the site. Redditors are especially concerned that Correct the Record's paid internet commenters are suppressing revelations from WikiLeaks' release of Clinton campaign chair John Podesta's emails, which have produced a flood of damaging revelations about the Clinton campaign. Reddit, which describes itself as a source for what's new and popular on the web, is divided into thousands of subreddits, which the site describes as subcommunities, each focused on a specific topic. Whether or not a post does well on Reddit, and consequently, how large of an audience it reaches, largely depends on how many upvotes and downvotes it receives. * 1 point

Mark only one oval.

☐ Real

☐ Fake

5. Philippines President Rodrigo Duterte said on Sunday the South China Sea dispute, involving China and five other claimants including the Philippines, was "better left untouched". "We have to be friends, the other hotheads would like us to confront China and the rest of the world on so many issues," Duterte said at a ceremony on the eve of a summit of East and Southeast Asian nations in the Philippine capital. "The South China Sea is better left untouched, nobody can afford to go to war." China claims almost the entire South China Sea, with competing claims from Taiwan, Vietnam, the Philippines, Malaysia and Brunei. ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

6. President Donald Trump will sign an executive order on Tuesday "establishing discipline and accountability in the environmental review and permitting process for infrastructure projects," the White House said in a statement on Monday. Trump, who is visiting his residence at Trump Tower in New York City, will also participate in a discussion on infrastructure and give a statement on the subject at 3:45 p.m. (1945 GMT). The White House did not give additional details on the executive order. Trump, who was a real estate developer before becoming president, made rebuilding the country's crumbling infrastructure a top campaign issue. He has proposed leveraging \$200 billion in government spending into \$1 trillion of projects to privatize the air traffic control system, strengthen rural infrastructure and repair bridges, roads and waterways. In June, Trump said one of the biggest obstacles to new infrastructure projects was "the painfully slow, costly and time-consuming process for getting permits and approvals to build." ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

7. President-elect Donald Trump's national security adviser Michael Flynn spoke on the phone to the Russian ambassador to Washington on Dec. 29, the day the United States announced sanctions against Moscow, a Trump official said on Friday. Trump's team had previously said the phone call happened on Dec. 28. The official said it was their understanding that U.S. retaliation for Russia's interference in the presidential election did not come up in the conversation between Flynn and Russian ambassador Sergei Kislyak. Two people familiar with the issue said Flynn had several phone conversations with the ambassador on Dec. 29. * 1 point

Mark only one oval.

☐ Real

☐ Fake

Appendix F*Quiz 4 For Grade 9 Students*

Navigating (More) Disinformation Pt. 4

On the following page, you will see 5 media pieces (**news, forum posts, text messages, email, etc.**). Please determine if each of them is **real or fake**.

These passages are all human-created; what you're looking for is whether the event is real or fake, and not whether it's generated by AI or human. The media pieces must've had happened at some point in history for it to be considered "real"; it doesn't have to be relevant. (Don't look this up on the internet, just try your best with your judgement.)

** Indicates required question*

1. Email *

2. Which class are you in? *

Mark only one oval.

☐ St. Raymond

☐ St. Rose

Real or Fake

3. New scientific reports are once again forcing planetary scientists to rewrite the history of our own moon. A new study into the lunar surface contradicts the notion that cratering on the moon occurs incrementally over vast eons of time. A team of scientists studied several thousands of before and after images of the moons' surface, with the visual data covering nearly a million square miles. What they found is that lunar cratering appears to occur at a rate more than 100 times faster than the standard impact model has predicted. Could electrical processes on the moon be responsible for the cratering? ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

4. As western media outlets like the New York Times, Washington Post, CNN and the BBC continue to hurl viral images of 'child victims of Aleppo,' western and gulf-backed terrorists in East Aleppo continue to fire mortars, 'hell cannon' and use snipers to target civilians and children in government-protected West Aleppo. The level of information fraud and propaganda being perpetrated by the western mainstream media and politicians like John Kerry and Samantha Power is unprecedented – even by traditional low US standards. ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

5. There is a lot more than meets the eye to this mob, wait till you read this. " Soon thereafter I resigned from Yang and took another job at the Florida Department of Transportation (FDOT). While there, I discovered some other illegal activities of Yang, who had contracts with FDOT, involving over-billing and defrauding of the State of Florida regarding contracts that I had worked on at Yang." Nee was sentenced on October 7, 2004 by U.S. District Judge Gregory A. Presenell to three years of supervised probation and \$100 fine. An illegal Chinese alien admitting to what is paramount to corporate if not treasonable espionage, and the guy gets a slap on the wrist. ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

6. A refugee from Kenya killed one former co-worker, wounded three others and then killed himself Tuesday in a workplace shooting that authorities are still trying to unravel in Virginia, police said. Getachew Fekede, 53, had entered the U.S. through a refugee immigration program and worked for the railcar manufacturer FreightCar America before being fired in March when he stopped showing up for work, Roanoke Police Chief Tim Jones told reporters. A neighbor told the Associated Press that Fekede quit his job over being harassed by a co-worker. Clarence Jones said Fedeke would send money to his mother back in Kenya and had grown concerned about his finances. ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

7. At least 3,800 migrants have died or are missing in the Mediterranean Sea in 2016 - the deadliest year on record, the UN refugee agency (UNHCR) has said. It said this was despite a significant drop in migrant crossings compared with 2015, when 3,771 deaths were reported. Smugglers were now more often using flimsy boats and putting more people aboard, the UNHCR said. The agency said the most dangerous route had been between Libya and Italy, with one death in every 47 arrivals recorded. By comparison, another - much shorter - route from Turkey to Greece had a ratio of 1 in 88. Nearly 330,000 people have crossed the Mediterranean Sea this year, compared with more than one million in 2015.

★ 1 point

Mark only one oval.

☐ Real

☐ Fake

Appendix G*Quiz 5 For Grade 9 Students*

Navigating (More) Disinformation Pt. 5

On the following page, you will see 5 media pieces (**news, forum posts, text messages, email, etc.**). Please determine if each of them is **real or fake**.

These passages are all human-created; what you're looking for is whether the event is real or fake, and not whether it's generated by AI or human. The media pieces must've had happened at some point in history for it to be considered "real"; it doesn't have to be relevant. (Don't look this up on the internet, just try your best with your judgement.)

** Indicates required question*

1. Email *

2. Which class are you in? *

Mark only one oval.

☐ St. Raymond

☐ St. Rose

Real or Fake

3. The U.S. Senate unanimously confirmed President Barack Obama's nominee to be ambassador to Myanmar on Tuesday, after the administration assured lawmakers it would not rush to ease sanctions as the country moves from decades of military rule. The Senate voted 90-0 to confirm career diplomat Scot Marciel to the post in Myanmar, also known as Burma. Currently the principal deputy assistant secretary in the State Department's Bureau of East Asian and Pacific Affairs, Marciel has also been ambassador to Indonesia and served in Vietnam, the Philippines, Brazil and Turkey. Obama nominated Marciel to be his next envoy to Myanmar in October. His confirmation vote had been briefly delayed by concerns that the administration might ease sanctions before the country's human rights situation improved. ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

4. Iranian youngsters lost to Saudi Arabia 6-5 at the semi-finals of the AFC U-19 Championship. After putting on a superb performance which resulted in qualifying for next year's FIFA U-20 World Cup, Iranian youngsters have narrowly lost 6-5 to Saudi Arabia in the semi-finals of the AFC U-19 Championship. In what turned out to be an incredible 11-goal thriller, Saudi Arabia scored three times in the first half while the Persians found the back of the net twice. The second half was no less exciting as both teams were in attacking mode throughout, demonstrating an unprecedented showdown with both sides scoring three more times each to take the final score to 6-5 for the Saudis. Saudi Arabia will take on Japan who cruised past Vietnam 3-nil in the second semifinal. ★ 1 point

Mark only one oval.

☐ Real

☐ Fake

5. Obi-Wan Kenobi has reported a great disturbance in the Force, as if millions of voices cried out in terror and were suddenly silenced. The wise old man felt the disturbance whilst chatting to his young apprentice and felt a sudden need for a sit down, muttering 'something terrible has happened'. Jedi Master Kenobi, who fought with Bail Organa in the Clone Wars, is currently uncertain of the cause of the disturbance but suggests some sort of planet-busting doomsday event may have occurred. "I'm hoping it's just a giant space-laser obliterating an entire civilisation; that's probably the least-worst possible cause," he confirmed. "Failing that, it's likely something dark side-y. Some kinda of colossal use of the mind trick, maybe? "Something something dark side something," he added. Upon investigation, it turns out that's no moon, it's Donald Trump's bald patch. * 1 point

Mark only one oval.

☐ Real

☐ Fake

6. Hillary Clinton has canceled all upcoming campaign events following the FBI's announcement that they are reopening their email server investigation. The reopening of the case has sent the Clinton campaign into complete chaos, according to reports. According to "Citizens for Trump" Special Projects Director Jack Posobiec, Hillary is looking to get out of the media spotlight for a while. In a tweet, he stated: "Hillary has cancelled all campaign events in FL, OH, and NC." The Clinton campaign wants to focus on states that Hillary lost serious ground in – like Pennsylvania, New Hampshire, Virginia, and Michigan. Has the latest email scandal finally brought Hillary to her knees? Will there even be an election on November 8th, or will we be watching Hillary Clinton go on trial? It seems like anything is possible right now. * 1 point

Mark only one oval.

☐ Real

☐ Fake

7. A major new plan will be announced soon by the Hillary campaign. This will certainly solve the world problems that face humanity. She will tax the rich and pay hundreds of millions of dollars to states that fight bullies. It's the antibully agenda to keep everybody's free speech and free thought on heel of the government the greatness is astounding. Hillary Clinton is unveiling a plan to reduce bullying. Called "Better than Bullying," the plan released Thursday by the Democratic presidential nominee would provide \$500 million in new funding to states that develop comprehensive anti-bullying efforts. Clinton is expected to talk about the plan during an appearance with first lady Michelle Obama Thursday. She has called Republican Donald Trump a bully for his derogatory comments. The campaign said states must address verbal and cyber bullying and establish a process for addressing incidents. They must also ban bullying on the "basis of race, color, national origin, sex, disability, sexual orientation, gender identity, and religion." States could get \$4 dollars from the federal government for every \$1 they spend. It would be paid for through Clinton's proposed tax increases on the wealthy. * 1 point

Mark only one oval.

☐ Real

☐ Fake