Fake News Detection

Aiesha Ayub, Aidan Murphy, Anna Nguyen, Jialene Westcott
Binghamton University

Abstract

In the age of digital media, the spread of fake news has become a major societal issue, particularly influencing politics. This study used a variety of machine learning models, including Support Vector Machines, transformers, Naïve Bayes classifiers, and Multilayer Perceptron classifiers, to accurately differentiate between real and fake news stories. Our results demonstrated high training accuracies, with the transformer model scoring the best at 99.6%. The ensemble model efficiently used each model's capabilities to improve classification accuracy and reliability, providing a useful tool for minimizing the effects of disinformation.

1. Introduction

Fake news is a more recent issue in history with the rise of social media and the internet, but it is a serious issue which constantly requires new solutions to mitigate its detrimental effects on society worldwide. Fake news consists of articles that are distributed, often to a certain target audience, which purposefully spread disinformation. This is typically done in a malicious way to influence the thoughts of the public and further the goals of a specific party. This doesn’t always mean a political party, however one of the biggest contexts in which fake news is used is the political context. Any media outlet, company, group, or individual which may benefit from having a certain candidate or party in power can use fake news to generate negativity and outrage towards the opposing side. Figure 1 shows how widespread fake news is with 42% of respondents in the USA being exposed to some sort of false information pertaining to politics or COVID-19.

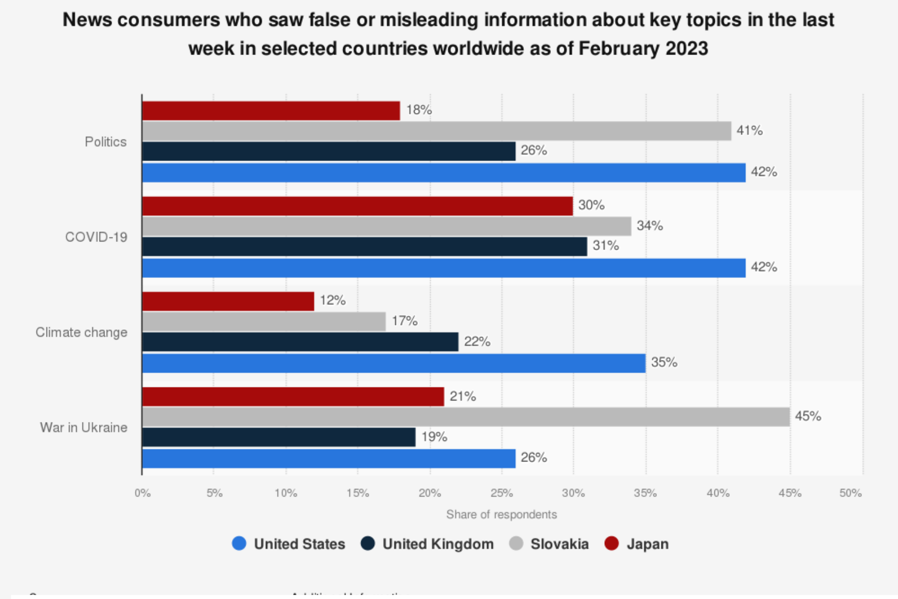
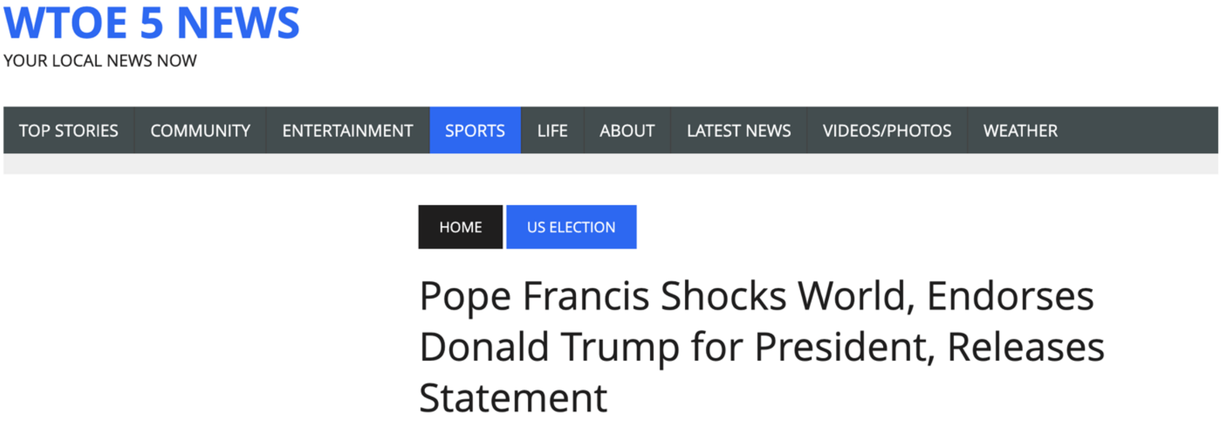

News consumers who saw false or misleading information about key topics in the last week in selected countries worldwide as of February 2023

Figure 1: Graph showing the prevalence of fake news in different countries through a survey that was conducted in 2023.

The reason fake news can be so detrimental to our society is because it takes away our freedom to think and make choices for ourselves. “If you fall into the trap of believing fake news, your beliefs and your decisions are being driven by someone else’s agenda,” (Smith). Fred Smith points out the most dangerous consequence of fake news. It misleads us to form ideas based on information that isn’t true, that was fed to us intentionally. When we form these ideas, we are under the misconception we arrived at these conclusions naturally and that they are true, when really, we are being used as part of a machine to further the goals of others. If our ideas and values are all built on malicious information, then we are essentially under the control of whoever controls this flow of information. Fake news had a significant impact on the 2020 U.S. Presidential election. Disinformation campaigns aimed to take advantage of social divisions and sway voters' opinions about candidates Donald Trump and Joe Biden on a variety of platforms. False narratives concerning their policies, electoral fraud, and personal character attacks were widespread and were oftentimes highlighted by well-known individuals and media sources, making it more difficult for the general public to distinguish fact from fiction. The challenges surrounding disinformation and fake news are still at the forefront as the 2024 election draws near. It is likely that the tactics employed in 2020 will only grow stronger, utilizing more advanced methods to disseminate misleading information. The AI technology that has exploded onto the scene since the last election, gives the general public and bad actors behind the scenes, access to powerful tools they did not previously have access to in any past election. The growth of this threat presents serious obstacles in maintaining an impartial electoral process. To ensure the integrity of the democratic system, these challenges must be resolved.



WTOE 5 NEWS

YOUR LOCAL NEWS NOW

Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement

Figure 2: Popular fake news headline from a 2017 article published to a network of disinformation sites, which then circulated on social media (Fox 26)..

Machine learning can help to free us from the control of fake news and combat the threat it poses to democracy. By allowing models to analyze thousands and thousands of real and fake articles and identify the subtle patterns and cues in the information that we as humans would not be able to recognize, we can differentiate factual content from disinformation. Once trained, they can accurately and automatically classify new content, possibly warning moderators or users of potential false information before it spreads. When built and trained correctly, these types of models can operate at a near perfect rate of success, without the need for fact checking or having any context of the information within the articles at all. These tools could greatly reduce the negative consequences of fake news on the democratic process, especially when used in conjunction with widespread educational initiatives on media literacy.

2. Data

We sourced our datasets from a GitHub repository created by Bhavik Jikadara, a machine learning enthusiast and recent graduate of Gujarat Technological University. The data in the two sets was collected by Jikadara by scraping news articles between 2016-2017. The two datasets were categorized as fake and real. The fake dataset contained a title column, a text column, a subject column, and a date column for each article. All the articles in this dataset had already been classified as fake news. The real dataset contained the same columns, however all the articles in this set had already been classified as real news. We created a label column containing the value 1 to every article in the real dataset, and a label containing the value 0 to every article in the fake dataset. Then we combined the datasets into one single dataset with over 44,000 articles total.

title	text	subject	date	label
As U.S. budget fight looms, Republicans flip their fiscal script	The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts, called himself a "fiscal conservative"…	politicsNews	December 31, 2017	1
U.S. military to accept transgender recruits on Monday: Pentagon	Transgender people will be allowed for the first time to enlist in the U.S. military starting on Monday as ordered by federal courts, the Pentagon said on Friday, after President…	politicsNews	December 29, 2017	1
Senior U.S. Republican senator: 'Let Mr. Mueller do his job'	The special counsel investigation of links between Russia and President Trump's 2016 election campaign should continue without interference in 2018, despite calls from some…	politicsNews	December 31, 2017	1

Table 1: Final combined dataset after cleaning.

To prepare the data for the models, we had to remove certain characters such as punctuation, symbols, and spaces. Every article in the real dataset contained the header 'Reuters--Location', which was affecting the classification. We used the Python replace method with a regular expression to find the word Reuters and remove it as part of pre-processing. Once these initial pre-processing steps are completed, the data was split 80/10/10 for training, testing, and validation. The MLP classifier has a parameter which sets aside a portion of the training data as validation data for early stopping, so the training and testing split for that model is simply 80/20.

We had to use different libraries to tokenize and vectorize the data for the different models. For the transformer model, we used AutoTokenizer. This tokenizer takes the words from the dataset and converts them into tokens which represent dictionary indices. Padding and truncation are set which adds artificial tokens to ensure all the text sequences are the same length and limit the sequences to a specific size. The max length is set to 50 so any sequence longer than 50 will be truncated. Special tokens are added to the sequences, so the model is able to understand the sequences properly, and finally there is a return parameter used to set the output type. For the MLP classifier model, text is automatically pre-processed to highlight important words and phrases while disregarding very common and uncommon terms. Similarly, text is pre-processed by the Support Vector Machine and Naïve-Bayes classifiers, which streamline the data for improved predicted accuracy by prioritizing essential terms and determining their significance across all texts.

3. Methods

We trained four different models to complete the classification task at hand. We used a Support Vector Machine, a transformer model, a Naïve Bayes classifier, and a Multilayer Perceptron classifier. Support Vector Machines, Naïve Bayes classifiers, and MLP classifiers are all examples of supervised learning models. Transformers and MLP classifier models are types of neural networks, but transformers contain supervised and unsupervised elements based on the task. In general, a neural network is a type of model which mimics the functions of a human brain. These models take raw data that has been pre-processed into an input layer, the input layer passes the data through several different hidden layers in which calculations are performed on the data and the output is passed on to the next layer. After the data passes through the hidden layers, the final output is passed to the output layer. Neural networks can be supervised, unsupervised, or a combination of both like the transformer used in this ensemble model.

Supervised learning models require more intervention from humans to create the desired output. They are based on data which already has an output label for each input. In this model, the output labels are 1 or 0 for real and fake, while the inputs are the article texts. These models may be easier to interpret as you are better able to follow the steps they take to classify the information. On the other hand, it is harder to track the inner workings of an unsupervised model due to the high level of complexity. This is because unsupervised models operate on unlabeled data. Instead of using labels, these models seek to find patterns within the data during training to create their own classifications. The pros of supervised models include their increased ability to evaluate and measure their effectiveness, and their predictability which makes it easier to determine what changes can be made to improve the accuracy of the model and achieve the desired results. The results of unsupervised models are harder to understand and evaluate due to their complexity, which makes fine tuning them and achieving the desired outcome more of a challenge. However, their ability to operate on data without labeled outputs, uncover much more complex patterns in the data, and produce highly accurate results if trained properly, makes them very useful for many tasks that unsupervised models may not be suited for. While both have their pros and cons, they all perform well at classification tasks, especially the task we were looking to complete.

The SVM model turns each text into a data point in a multi-dimensional space. The dimensions represent words or features from the texts. The hyperplane is a boundary in this

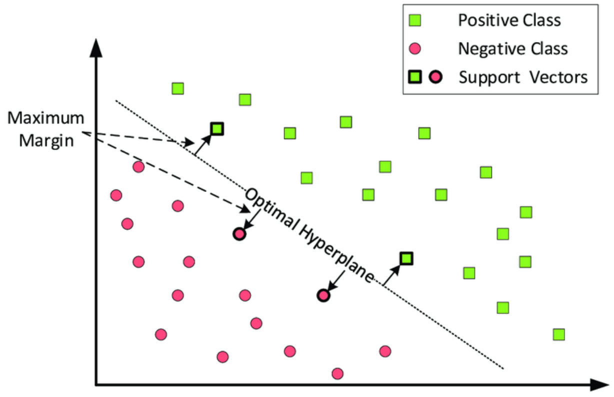


Figure 3: Graphic example of an SVM binary classification task (Ferre, Ruben & Fuente, Alberto & Lohan, Elena Simona.

space which divides the points in the different dimensions into two classes, in this case either real or fake news. The goal of the model is to position the hyperplane in such a way that it maximizes the margin between the data points closest to the hyperplane. After the model is trained, any new article can be vectorized and placed within this multi-dimensional space.

The process by which the transformer model is trained is more complicated than the other models. It involves a careful approach to build, train, and fine tune for desired results. Before training, there are several different arguments that must be set. These arguments provide the model with parameters to follow as it is training. This involves setting a location for the results to be saved, the gradual increase of the learning rate of the model, the decay of the weights, the number of training epochs, and the optimization method.

Once the arguments are set, the training of the transformer can take place. After vectorization, the data is divided into batches. The bert-base-uncased model is initialized with pre-trained weights assigned. The model processes the data in batches. Each batch is passed through the layers of the model and outputs predictions. These predictions are then compared with the true values and then the model uses a function to calculate the loss, which is a measure of how well the predictions match the true values. This is called a forward pass. The next step that takes place is a backward pass which calculates how the loss would change if certain changes were made to the weights. The weights are then updated based on the arguments. This process is repeated for each batch until there are none remaining at which point one epoch has been completed and a new one will begin.

In an ensemble model for binary text classification, the transformer model plays a crucial role. Its unique features help to improve the overall performance and accuracy in ways the other models cannot. First, the transformers' architecture allows for parallel text processing, which dramatically accelerates both the training and prediction speed, making them extremely efficient when dealing with big datasets. This efficiency is further improved by the model's ability to carry out transfer learning, which allows pre-trained transformers to be fine-tuned with minimal extra input, utilizing their broad pre-learned linguistic patterns for specific tasks. This not only improves the model's adaptability but also its scalability, (Turing). In addition, transformers excel at extracting features and contextual knowledge with their self-attention mechanism. This process assesses the value of each word in respect to others in the same context, allowing the model to capture complex meanings and subtle cues that are necessary for accurately categorizing texts as real or fake. Transformers' feature extraction capabilities ensure that they produce a comprehensive analysis of text, which is critical for producing precise results in binary text classification tasks.

The Naïve Bayes classifier model is one of the easier models to understand out of the four used. This multinomial Naïve Bayes model is based off Bayes' Theorem. The equation states that the probability of A when B is true, is equal to the probability of B when A is true multiplied by the independent probability of A, and then divided by the independent probability of B. In our model, the probability of y given x is true is what is being calculated. This is done by multiplying the likelihood of the words given the class by the probability of real or fake, which is taken directly from the training set. The article is then classified as real or fake depending on which label had the higher probability.

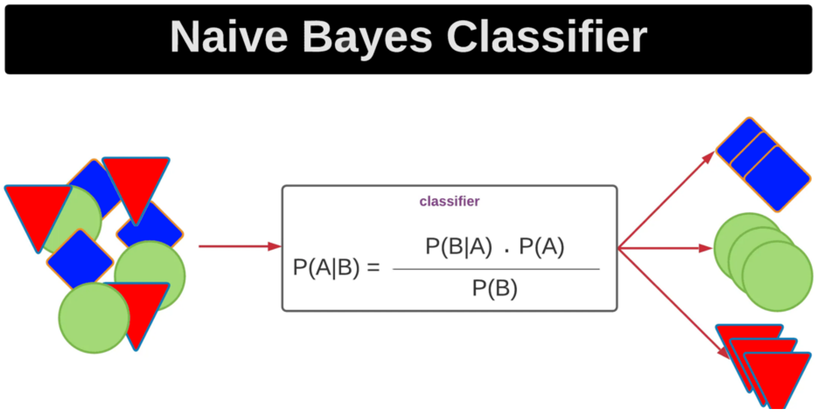
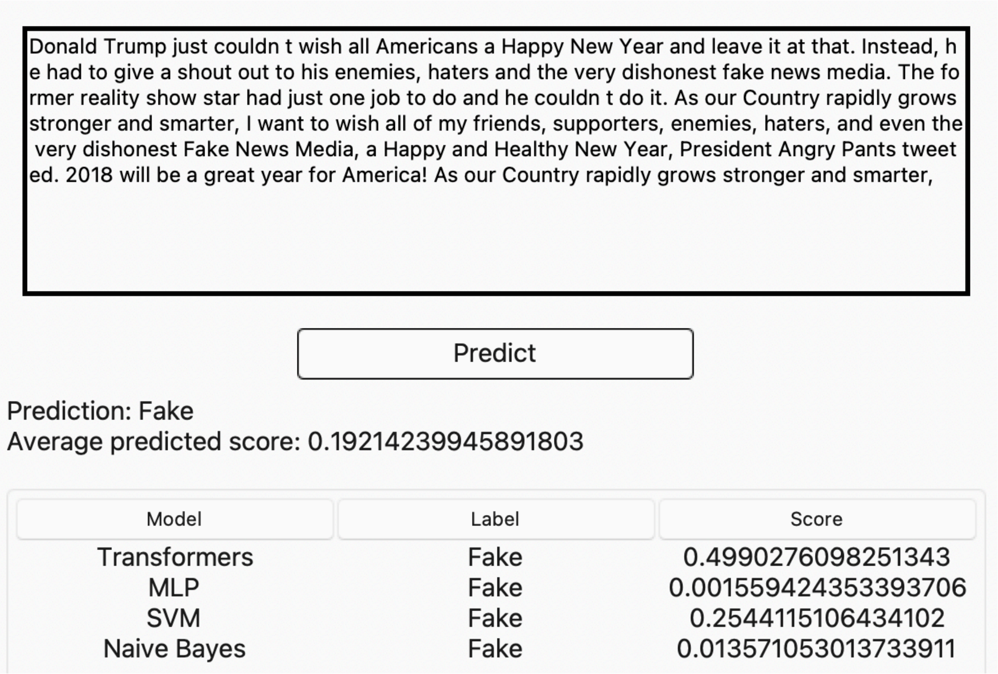


Figure 4: Naïve Bayes Classifier model visualization and function (Elzeiny).

To train an MLP classifier for binary classification, preparing the data is the first step in making sure the input features are formatted and normalized correctly for learning. Unlike certain models, like transformers, which might start with pretrained weights, the MLP initializes its weights randomly, laying the groundwork for training. Using the ReLU activation function, the MLP's layers allow the model to recognize more intricate patterns in the data. The model goes through a number of iterations during training in which it uses the gradient descent method to modify the weights after making predictions, calculating the loss to evaluate prediction accuracy. To prevent overfitting, the training data is randomized after each iteration. Furthermore, the training includes methods such as early stopping, which stops training if the improvement falls below a predetermined threshold after several iterations, saving processing power and preventing overfitting. The way this classifier model is trained is very similar to the transformer model, but it uses random weights and relies on labeled data.

Finally, the four models are saved after training using the Python pickle library. These models are then loaded into a new notebook where they are used in an ensemble model class.

The class defines the functions necessary to process input text into the proper format that each model accepts, and functions to pass this new data to each model and calculate a predicted score.



Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn t do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year, President Angry Pants tweeted. 2018 will be a great year for America! As our Country rapidly grows stronger and smarter,

Prediction: Fake
Average predicted score: 0.19214239945891803

Model	Label	Score
Transformers	Fake	0.4990276098251343
MLP	Fake	0.001559424353393706
SVM	Fake	0.2544115106434102
Naive Bayes	Fake	0.013571053013733911

Figure 5: Ensemble model GUI showing the predicted scores for a fake news article.

Figure 5 shows the final version of the ensemble model GUI, which takes new, user-inputted article text from the text box and calls the ensemble model class when the button is pressed.

4. Results

Model Name	Training Accuracy
Transformers	99.577%
MLP	99.22%
SVM	98.57%
Naïve Bayes	94.07%
Combined	**97.86%**

Table 2: Final training accuracies of the four models and the average training accuracy.

After training our models, we received high accuracy results across the board which can be seen in Table 2. Aside from the training accuracies, there were important takeaways from the relationship between words from the text and the log ratio within the dataset. The log ratio measures the correlation of meaningful words within the dataset and the label they were associated with.

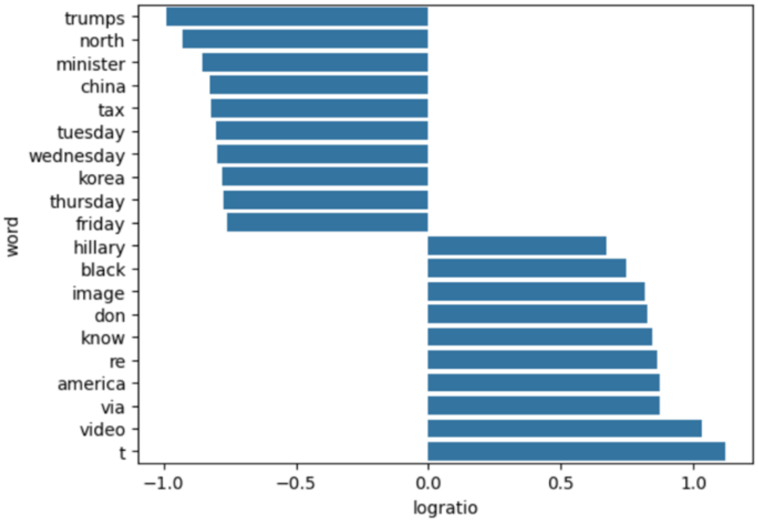


Figure 7: Log ratio visualization showing the relationship between words and labels.

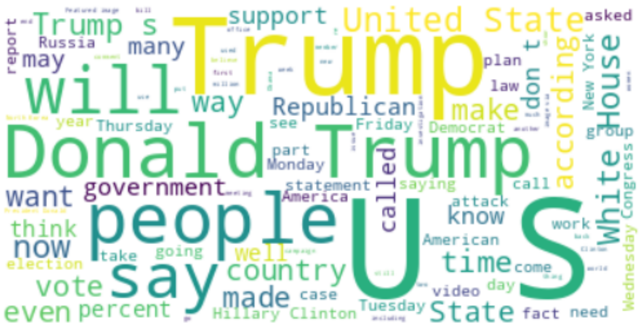


Figure 6: Word cloud showing the most common words in the dataset's text column.

In Figure 6, it's clear that there is a strong correlation between the fake label and words like Trump, China, and tax, and a strong relationship between the real label and words like video, America, and black. There are also correlations with words that may not have much use for classification like the days of the week. The graph provides important insights into trends in the dataset that could affect training. By cleaning and preprocessing the data in different ways and experimenting with the removal of certain terms, new outcomes and results can be achieved that may increase or decrease the model's accuracy and effectiveness.

Another useful visualization to achieve the goal of improving results, is a word cloud. This visualization shows the most common words in the news articles across the entire dataset. Again, the name Donald Trump is extremely common throughout the fake and real portions of the articles. By using the word cloud in conjunction with the log ratio graph, there are many types of hypotheses that can be created and tested by removing certain words and training the models based on the updated data. It is also helpful to see how the presence of certain common words in an input article may affect the predictions of the ensemble model.

Overall, our model showed high training accuracies as seen in Table 2, however its applications are still somewhat limited due to the training data. Something we noticed during the final testing of the ensemble model is that it has trouble making predictions on articles written outside of 2016-2018. The most likely cause of this issue is that the model is trained on news articles that span from 2016-2017. A large portion of the terms used in those articles and their subject matter contained issues specific to those years. To solve this issue, more article data from the years before 2016 and after 2017 would be necessary. After training the model on updated information, it would have more widespread applications. The problem is datasets that contain fake and real news articles are scarce and not usually updated with current news. The two years of data used for this model's training contained over 44,000 rows, so to include a decade's worth of data would require an enormous amount of processing power for training. For the purposes of this project, the 2016-2017 data was sufficient to create a model and the goal of distinguishing between real and fake news was reached, however, there is room to expand this model and increase its functionality.

5. Conclusion

Our goals were to effectively create an ensemble model and to create a program in which users can input text from a news article, run it through an ensemble model, and then have the model predict whether this is real or fake news. Utilizing the features of several machine learning models—Support Vector Machines, transformers, Naïve Bayes classifiers, and Multilayer Perceptron classifiers, we sought to develop an effective system that not only accurately classifies news but also improves the accuracy of information shared with the public. By merging the predictive powers of several models, we were better able to counteract the drawbacks of individual models and increase overall accuracy and dependability.

Our approach's effectiveness was demonstrated by the promising results we obtained. With an accuracy of 99.4%, the Multilayer Perceptron model outperformed the Support Vector Machine model, which came in at 99%. Even with its 94% accuracy, the Naïve Bayes model performed impressively considering its simplicity and speed. The transformer model outperformed expectations with its advanced neural network design, reaching an impressive 99.6% accuracy in our testing.

The process of achieving these outcomes started with the careful preparation of the data and training parameters, where we modified datasets to train our models efficiently without bias from data features like the frequent recurrence of a particular term corresponding to real news sources. To ensure that every model could effectively learn from the training data, we used a number of preprocessing techniques, including tokenization, stop word removal, and vectorization. In order to maximize each model's learning process without overfitting, the training phase was meticulously controlled using parameters including regularization, early stopping, and learning rate adjustments. These steps contributed significantly to the high accuracies that were achieved.

In conclusion, the ensemble model we developed not only successfully classified news stories, but also has the potential to improve the quality of information shared with the public by training with new data. This study has created a strong foundation for our future research and development in the fight against fake news, with the potential for future upgrades including longer-term data collection and a wider range of news sources to boost the model's applicability and reliability. Our results aim to contribute to ongoing efforts to combat disinformation, particularly as we approach future elections.

Works Cited

Elzeiny, Mahmoud. “The Ultimate Guide to Naive Bayes.” *Machine Learning Archive*, ML Archive, 18 June 2023, mlarchive.com/machine-learning/the-ultimate-guide-to-naive-bayes/.

This article was used to understand Naïve Bayes models better but mainly for the visualization it provided of the Naïve Bayes function and how it is used in a machine learning model.

Ferre, Ruben & Fuente, Alberto & Lohan, Elena Simona. (2019). Jammer Classification in GNSS Bands Via Machine Learning Algorithms. Sensors. 19. 4841. 10.3390/s19224841.

This article was used solely for the support vector machine visualization it provided. The support vector machine visualization in the article was the best one that I was able to find in terms of its similarity to the model that we used. This visualization specifically shows an SVM being used for a binary classification task.

Fox 26. Schools across the country are teaching kids how to fight fake news. *Facebook*, 11 Jan. 2017, 3:39 p.m., https://www.facebook.com/watch/?v=1253495491383840. Accessed 9 May 2024.

This social media post was used as an example image of a fake news article from the period of time in which our dataset comes from, and the clear detriment effects it can have to play on people’s emotions.

Smith, Fred. “The Dangers of Fake News.” *The Elm*, University of Maryland, Baltimore, 11 Nov. 2020, elm.umaryland.edu/elm-stories/Elm-Stories-Content/The-Dangers-of-Fake-News.php.

This article investigates the negative impacts of fake news on society, focusing on how it influences public perception and decision-making. This story was utilized to highlight the societal effects of false news, and it offered a real-world context for our research's motivation, showing the importance and relevance of building effective fake news detection technologies.

Turing. *The Ultimate Guide to Transformer Deep Learning*, Turing Enterprises Inc, 11 Feb. 2022, www.turing.com/kb/brief-introduction-to-transformers-and-their-power#feed-forward-network.

This guide provides a detailed explanation of transformer models, focusing on their mechanics, uses, and advantages over other models. This material was useful in understanding the transformer model's sophisticated design and capabilities in general and in our ensemble model, particularly its efficiency and effectiveness in processing language data for fake news detection and binary classification, and its differences from the other models.

Watson, Amy. "News Topics and False Information Worldwide 2023." *Statista*, 7 Mar. 2024, www.statista.com/statistics/1317019/false-information-topics-worldwide/.

This report provides statistics on the prevalence and types of false information circulated in multiple countries. Used to provide statistical evidence of the widespread nature of fake news, supporting the introduction's discussion on the scope and scale of the problem our research aims to address.