Fake News Detection

Submitted in partial fulfillment of the requirements for the degree

of

Master of Science

by

Varshitha Choudary Vasireddy
113530813

Under the guidance of

Dr. Matthew Beattie

DATA SCIENCE AND ANALYTICS

UNIVERSITY OF OKLAHOMA

ACKNOWLEDGEMENT

I want to express my sincere gratitude to my professor **Dr. Matthew Beattie** for allowing me to complete my project from India and for providing me with constant guidance and support throughout the project. I am grateful for the invaluable knowledge and experience I have gained under his direction.

I would also like to extend my heartfelt thanks to my friends who helped me with the project and provided me with the necessary moral support. Their contributions and encouragement were instrumental in the successful completion of this project.

Once again, thank you to my professor and friends for their unwavering support and assistance.

Varshitha Vasireddy

ABSTRACT

Fake news has become a major issue in today's society, with its potential to cause significant harm to individuals and communities by influencing public opinion and impacting elections. Machine learning has emerged as a promising tool for combating fake news, and this project aims at developing a model to accurately identify potentially fake news articles using only their titles. A dataset was collected and preprocessed, and various machine learning and transfer learning techniques were employed to identify patterns and features that indicated the authenticity of an article. The results showed a high accuracy rate in detecting fake news articles using only their titles, highlighting the potential of machine learning in the fight against fake news.

Stopping the spread of fake news requires a multifaceted approach, including fact-checking, critical thinking, and media literacy education. Machine learning can play a significant role in this effort by providing automated tools to quickly and accurately detect fake news, allowing for timely interventions and corrections. The results of this study have the potential to benefit the general public, journalists and news organizations, and social media companies. By combatting the spread of fake news on social media platforms, the public can have access to more reliable and accurate information, improving civic engagement. This study provides a valuable tool for fact-checking and verifying information for journalists and news organizations, ultimately improving the quality and credibility of their reporting. Social media companies can also benefit by implementing machine learning algorithms to detect and remove fake news articles from their platforms, enhancing user trust and improving the integrity of their services. This project has demonstrated the potential of machine learning in identifying potential fake news articles and contributing to a more informed and trustworthy information ecosystem.

Table of Contents

List of Figures

List of Tables

Glossary

Accuracy: It measures the proportion of correct predictions among all predictions. A high accuracy score indicates that the model is able to correctly classify most of the instances.

Precision: It measures the proportion of true positives (correctly predicted positive instances) among all instances predicted as positive. A high precision score indicates that the model has a low rate of false positives.

Recall: It measures the proportion of true positives among all actual positive instances. A high recall score indicates that the model has a low rate of false negatives.

F1 Score: It is the harmonic mean of precision and recall, and provides a balanced measure of both. A high F1 score indicates that the model has a good balance between precision and recall.

ROC AUC Score: It measures the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate at different classification thresholds. A high ROC AUC score indicates that the model has a good ability to distinguish between the positive and negative instances.

TF-IDF tokenization: It stands for the Term Frequency-Inverse Document Frequency, which is a numerical statistic that reflects the importance of a word in a document or corpus. It is often used in text classification and information retrieval tasks. The TF-IDF value for a word increases proportionally to the number of times it appears in the document but is offset by the frequency of the word in the corpus. This helps to give higher weightage to words that are important in a specific document but not in the overall corpus.

Label encoding: Label encoding is a process of converting categorical data into numerical data. In this process, each unique category is assigned a numerical label.

Chapter 1

Introduction

Fake news has become a significant challenge in modern society, threatening the integrity of information and trust in media sources. False, misleading, or fabricated information can have severe consequences, including influencing public opinion, impacting elections, and causing harm to individuals and communities. In response to this issue, a multifaceted approach is needed to combat the spread of fake news.

One of the primary platforms for the spread of fake news is social media. Social media platforms have enabled the rapid dissemination of information but have also facilitated the spread of fake news. This is because social media allows anyone to create and share content, often with little oversight. As a result, social media has become a breeding ground for the spread of false and misleading information.

The effects of fake news are far-reaching and can have severe consequences. For example, during the 2016 US presidential election, false information was spread on social media, potentially influencing the election's outcome. Similarly, during the COVID-19 pandemic, the spread of misinformation on social media led to people refusing to wear masks, leading to increased infections and deaths. It is clear that fake news has the potential to cause significant harm, and steps need to be taken to combat its spread.

Stopping the spread of fake news is essential to ensure a more informed and engaged public, foster healthy political discourse, and maintain trust in journalism and other information sources. It requires a multifaceted approach, including fact-checking, critical thinking, and media literacy education. Journalists and news organizations play a critical role in this effort by fact-checking information and reporting on the spread of fake news. Social media companies can also combat the spread of fake news on their platforms by implementing policies to detect and remove false information.

One important benefit of combating fake news is the protection of democratic institutions and processes. In recent years, the influence of fake news on elections has become increasingly evident, with false information being used to sway public opinion and undermine the democratic process. By detecting and removing fake news from social media platforms, machine learning can play a critical role in protecting the integrity of elections and democratic decision-making. Additionally, detecting fake news can help prevent harm to individuals and communities. False or misleading information can cause panic, incite violence, and damage the reputation of individuals and organizations. By combating the spread of fake news, machine learning can contribute to a safer and more secure society.

Overall, stopping the spread of fake news is a significant challenge, and its effects can be far-reaching. A multifaceted approach may combat its spread, including fact-checking, critical thinking, media literacy education and machine learning. Taking steps to combat the spread of fake news can ensure a more informed and engaged public, foster healthy political discourse, and maintain trust in journalism and other information sources.

Chapter 2

Objectives

This report outlines the aim of developing a highly accurate machine-learning model for news classification that can differentiate between real and fake news on social media platforms. The specific problem addressed in this project is the rapid spread of false information on social media platforms that can mislead and influence users' beliefs and actions. The project aims to develop a model that can effectively classify news into either real or fake news, with the ultimate objective of raising awareness about the importance of fact-checking before sharing information on social media.

To accomplish this objective, the following specific objectives have been set:

- Developing a robust and accurate machine learning model: The first objective is to develop a machine learning model to identify patterns and features indicative of real or fake news. State-of-the-art technology, including Hugging Face language models, will be leveraged to achieve the best classification results. Various algorithms, feature engineering techniques, and hyperparameter tuning will be used to optimize the model's performance and refine it through iterative experimentation. Cross-validation will also be used to avoid over-fitting the models.

- Comparing the machine learning models: The second objective is to compare the machine learning models developed with the help of various metrics, such as accuracy, F1 score, recall, ROC curve, and AUC ROC score, to understand the best model for the task of news classification.

- Raising awareness about the importance of fact-checking: The third objective is to contribute to public awareness about the need to fact-check information before sharing it on social media. By accurately classifying news as real or fake, the project intends to encourage users to critically evaluate the integrity of information and promote responsible information-sharing practices.

- Advancing the field of natural language processing: The final objective is to contribute to the field of natural language processing by developing a highly accurate machine learning model for news classification. The project will explore state-of-the-art techniques and methodologies in natural language processing, like language models, to enhance the accuracy of news classification methods.

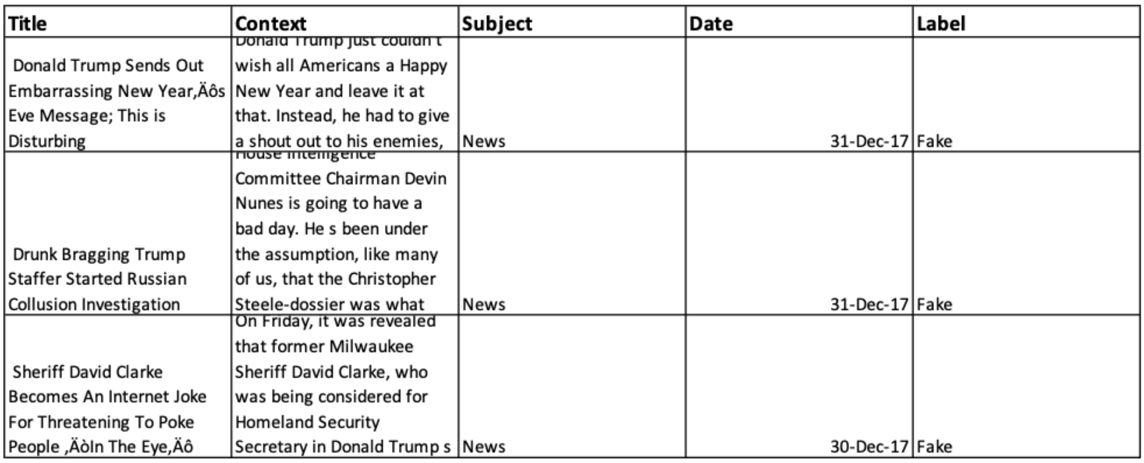
The success of the project will be evaluated based on achieving these specific objectives. The performance metric of the machine learning model in classifying news as real or fake news will serve as a critical factor in detecting the best model. The model should be able to identify fake news correctly.

In conclusion, this project aims to address a critical problem facing society today: the dissemination of fake news on social media. By developing a highly accurate machine learning model for news classification, the project aims to raise awareness about the importance of fact-checking and promote responsible information-sharing practices. Through this project, it is intended to contribute to the field of natural language processing and advance the accuracy of news classification methods.

Chapter 3

Data

The dataset is chosen from Kaggle. A vast dataset is chosen so machine learning models can learn well from the data. It consists of 5 columns with 44.9K rows. They are title, text, subject, date, and label. The title consists of the article's title; the text consists of the article's content. A date consists of the article's published date; the subject refers to the type of the article. Finally, the label consists of the classification of the article as either fake or real. The dataset has no missing values, although it has duplicates. After removing the duplicates, we have a total of 39K rows. The fake and real classification of the dataset is balanced, with 18k rows as fake news and 21k as real news. A few rows of raw data can be seen in figure 3.1



Title	Context	Subject	Date	Label
Donald Trump Sends Out Embarrassing New Year,Äôs Eve Message; This is Disturbing	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,	News	31-Dec-17	Fake
Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the assumption, like many of us, that the Christopher Steele-dossier was what	News	31-Dec-17	Fake
Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke People ,ÄòIn The Eye,Äô	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered for Homeland Security Secretary in Donald Trump s	News	30-Dec-17	Fake

Figure 3.1: Raw data

New features were created to facilitate the exploration of the dataset due to the limited number of columns available. Before conducting data exploration, the data underwent preprocessing. Initially, only the article titles were considered as inputs to train the model and evaluate its performance. This decision was motivated by the fact that the title contains relevant information from the article and is considerably shorter than the text field of the dataset, which contains multiple paragraphs. Preprocessing a large dataset can be a resource-intensive process; thus, focusing on the title alone was deemed appropriate.

The first step in the preprocessing phase was to convert the title into lowercase letters. This ensured that all characters were normalized and not overemphasized due to their capitalization. Consequently, the machine learning model would identify patterns in the dataset based on the given information rather than overlearning from capitalizing characters.

To preprocess the title column data, a method was implemented to segment paragraphs into sentences and further break them down into individual words. To improve the accuracy and efficiency of text analysis, all stop words present in the NLTK stop words vocabulary library was removed from the divided words. This technique helps to eliminate commonly occurring but insignificant words that do not add meaning to the text. Punctuation marks, such as commas, periods, and quotation marks, do not carry any semantic meaning and only serve to delimit sentences and clauses. Therefore, eliminating these punctuation marks can help to simplify the text data and make it easier to process by reducing unnecessary noise that could hinder the performance of NLP algorithms.

Subsequently, the next step in the data preprocessing phase involved the iteration of each character in every word in the title column, after which all non-alphanumeric characters and spaces were removed. The resulting preprocessed words were then joined together to form a sentence that served as the input variable for the machine learning model.

To obtain a deeper understanding of the dataset and identify patterns, a new column was created to count the occurrence of non-alphanumeric characters in each word of the preprocessed title sentence. The statistics of the non-alphanumeric characters can be seen in the table 3.1. It can be evaluated from the table result that fake news consists of many non-alphanumeric characters compared to real news. This finding can be justified because real news consists of formal writing, whereas fake news consists of improper writing which also involves unidentified characters.

Table 3.1: Non-Alphanumeric Count Comparison

Non Alphanumeric count	**Fake**	**Real**	**Total**	**%Fake**	**%Real**
0	1687	5433	7120	23.7	76.3
1	2675	5853	8528	31.4	68.6
2	3430	4722	8152	42.1	57.9
3	3354	2812	6166	54.4	45.6
4	2398	1248	3646	65.8	34.2
5	1629	514	2143	76.0	24.0
6	1159	180	1339	86.6	13.4
7	713	45	758	94.1	5.9
8	397	9	406	97.8	2.2
9	189	8	197	95.9	4.1
10	112	1	113	99.1	0.9
11	71	0	71	100.0	0.0
12	30	0	30	100.0	0.0
13	17	0	17	100.0	0.0
14	6	0	6	100.0	0.0
15	7	0	7	100.0	0.0
16	4	0	4	100.0	0.0
17	1	0	1	100.0	0.0
18	0	0	0	0.0	0.0
19	0	0	0	0.0	0.0
20	1	0	1	100.0	0.0
21	1	0	1	100.0	0.0

To better understand the dataset, I aimed to determine the sentiment of the preprocessed title, which can provide insight into the type of fake news being spread. To accomplish this task, I utilized an existing hugging face model, GPT2, to predict the sentiment of the preprocessed title. However, it is important to note that the accuracy of the model's sentiment prediction cannot be fully validated since it was trained on a large, unsupervised dataset rather than being fine-tuned on my specific dataset. The pre-training process allows the model to learn general language representations that can be fine-tuned for a variety of natural language processing tasks. As such, the results of the sentiment analysis graph may not be optimized for the used dataset. The sentiment analysis of the preprocessed title results can be seen in figure 3.2. It can be deducted from the figure that most of the news is of negative sentiment.

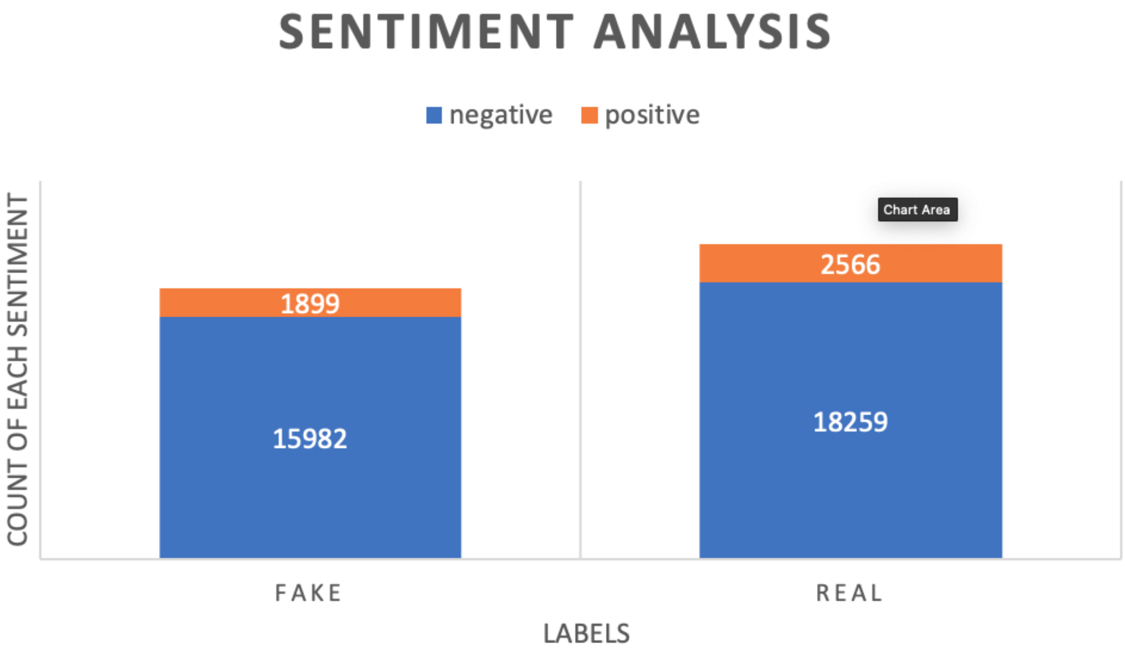


Figure 3.2: Title's predicted sentiment analysis graph

The process of manually labeling a substantial portion of the dataset is both laborious and time-intensive. The required amount of labeled data for achieving desirable machine learning model performance is difficult to estimate, making it a challenging endeavor. To circumvent these issues, alternative techniques that do not require manual labeling are deemed to be more practical and efficient. Again utilized the pre-trained GPT2 model for predicting the news genres of preprocessed titles such as politics, sports, entertain-

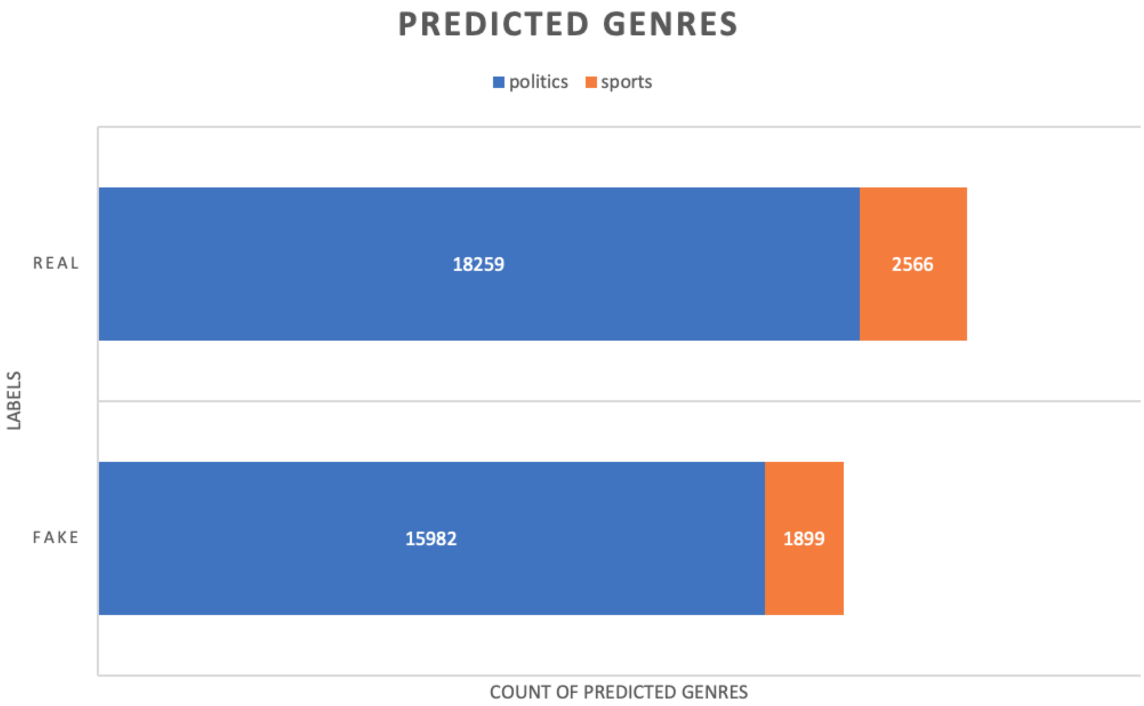


Figure 3.3: Title's predicted genre analysis graph

ment, education, and crime, among others. It is essential to acknowledge that the results obtained from this technique should be interpreted with care as the model has not been fine-tuned to the specific dataset being analyzed. Despite this limitation, this technique has the potential to provide valuable insights into the dataset, such as the identification of a significant proportion of news articles containing political content, as evidenced by the frequent mention of popular political personalities' names, such as Obama, Trump, and Modi. The preprocessed's title genre analysis graph can be seen in figure 3.3. It can be seen from the graph that most of the news is about politics and some part of news is talking about sports. This prediction can be made a little considering the names present in the title's. If politician names are mentioned then it is detected as politics genre, if sports personality name is present, then it is detected as sports genre.

As a part of the data exploration process, I investigated the presence of click-bait words in the unprocessed title of the news articles in my dataset. It is well-known that fake news often includes sensational and misleading headlines to attract readers. I created a list of commonly used click-bait words[1] to identify such words and then added two new

[1] Some of the click-bait words that I used were: ”shocking”, ”believe”, ”amazing”, ”secrets”, ”re-

columns to the dataset. The first column represents the presence or absence of click-bait words in the title sentence, and the second column represents the number of click-bait words present in the title sentence. This analysis of the data can help to verify the presence of click-bait words in the fake news articles in my dataset and also to determine the frequency of click-bait words in genuine news articles.

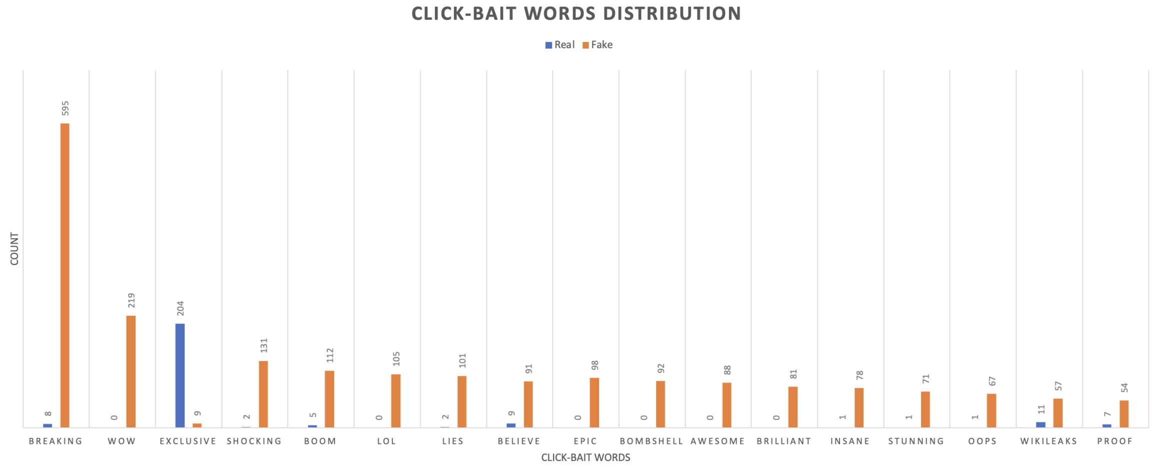
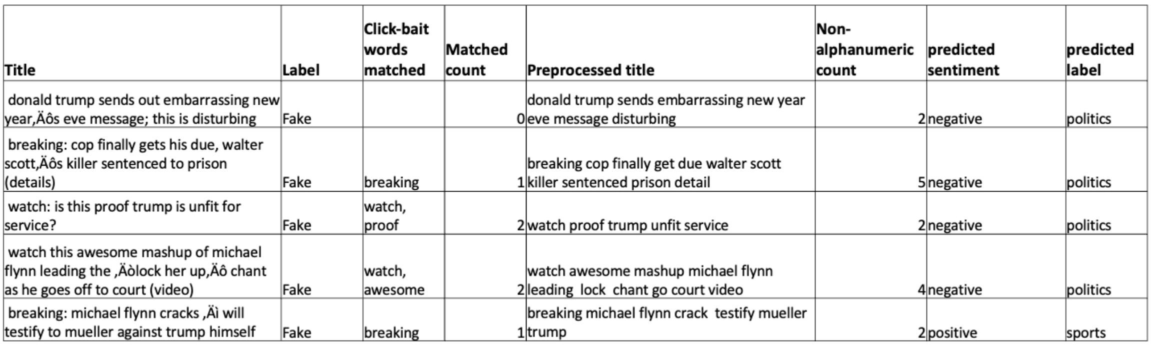


Figure 3.4: Click-bait words distribution

Figure 3.4 shows how a click-bait word is present in real and fake news. The word ”breaking” was seen mostly in fake news, whereas real news doesn’t have much. The word ”exclusive” was mostly present in real news. Could see that words like epic, bombshell, awesome, and brilliant were present in fake news and were not seen in real news. This figure is visualizing the truth that fake news contains more click-bait words than real news.

The data exploration techniques employed in this project have provided valuable insights into the characteristics and patterns present in the dataset. These findings can inform the development of models for classifying news articles as either fake or real. By identifying common features such as the prevalence of certain click-bait words, the dominance of political content, and the sentiment of the titles, we can better understand the data. The final preprocessed dataset can be seen in the figure 3.5

vealed”, ”mind-blowing”, ”unbelievable”, ”incredible”, ”insane”, ”must see”, ”exposed”



Title	Label	Click-bait words matched	Matched count	Preprocessed title	Non-alphanumeric count	predicted sentiment	predicted label
donald trump sends out embarrassing new year,Äôs eve message; this is disturbing	Fake		0	donald trump sends embarrassing new year eve message disturbing	2	negative	politics
breaking: cop finally gets his due, walter scott,Äôs killer sentenced to prison (details)	Fake	breaking	1	breaking cop finally get due walter scott killer sentenced prison detail	5	negative	politics
watch: is this proof trump is unfit for service?	Fake	watch, proof	2	watch proof trump unfit service	2	negative	politics
watch this awesome mashup of michael flynn leading the ,Äòlock her up,Äô chant as he goes off to court (video)	Fake	watch, awesome	2	watch awesome mashup michael flynn leading lock chant go court video	4	negative	politics
breaking: michael flynn cracks ,Äì will testify to mueller against trump himself	Fake	breaking	1	breaking michael flynn crack testify mueller trump	2	positive	sports

Figure 3.5: Preprocessed dataset

Chapter 4

Methodology

4.1 Techniques

This project aims to classify titles as either fake or real news, making it a classification problem. There are several machine learning models, each with unique characteristics and applications. For the classification problem of this project, we have labeled data, which makes supervised machine learning models better than other models. To send the dataset into machine learning models, the dataset was divided into train, test and validation. Train-test-split function from the Scikit-learn library was used to perform this task. 70% of the data reserved for training, 15% for testing and 15% for validation.

Before sending the data into models, it has to be tokenized/vectorized. Tokenization is a process in natural language processing (NLP) and machine learning tasks on text data that involves breaking down a text into individual words, phrases, or other meaningful elements, called tokens. It is essential because machine learning algorithms need to operate on individual words or tokens rather than the entire document or sentence as a single unit. Tokenization helps in reducing the data complexity and dimensionality of the text data, simplifying the machine learning task. It converts the text into a structured format that can be processed by algorithms, and reduces the dimensionality of the data. Used TF-IDF and label encoding for tokenization.

TF-IDF is used for vectorizing the inputs to the model, i.e. title of the article. Label encoding is used for labels of the dataset. Label encoding is a technique used for converting categorical variables into numerical variables. This allows the machine learning algorithm to understand and work with the categorical variable. Vectorization is the process of representing data as a vector or array of numbers, where each number corresponds to a particular feature or attribute of the data. By representing data as numerical vectors, machine learning models can identify patterns and relationships that would otherwise be challenging to detect.

Several widely-used machine learning algorithms were utilized for the task of classification, including Support Vector Classifier (SVC), Naive Bayes, Random Forest, Decision Tree, and k-Nearest Neighbors. These algorithms were selected based on their established effectiveness in handling text classification tasks, their capability to manage large datasets, and their interpretability. SVC is renowned for its ability to handle high-dimensional data and find optimal hyperplanes for separating data points, while Naive Bayes is a simple and efficient algorithm for text classification. Random Forest and Decision Tree are ensemble methods that can handle non-linear relationships in the data, and k-Nearest Neighbors is a lazy learner that can make predictions based on similarity to neighboring data points.

Hyperparameter tuning is done to optimize the performance of a machine learning model. Examples of hyperparameters include the learning rate, batch size, number of epochs, and the number of layers in the model. Optimizing hyperparameters is important because it can significantly improve the accuracy and generalization performance of a model. Cross-validation is performed to avoid overfitting and underfitting of a model to the training data. Cross-validation is a technique used in machine learning to evaluate the performance of a model. It involves partitioning the data into several subsets or folds, training the model on a subset of the data and testing it on the remaining subset, and then repeating this process with different subsets.

In conjunction with conventional machine learning algorithms, I investigated transformer-based language models[4] from the Hugging Face library. To check if they can perform better than other models. Language models are machine learning models that are designed to understand and generate human language. Transformer-based language models are a type of deep learning model that has revolutionized the field of natural language processing (NLP). The transformer architecture was introduced in June 2017. These models are based on a transformer architecture that allows them to process sequences of input data, such as text, in parallel, making them much more efficient and scalable than previous models.

A key feature of Transformer models is that they are built with special layers called attention layers. this layer will tell the model to pay specific attention to certain words in the sentence you passed it (and more or less ignore the others) when dealing with the representation of each word. In any task associated with natural language, a word by itself has a meaning, but that meaning is deeply affected by the context, which can be any other word (or words) before or after the word being studied. Hence transformer language models perform better. A popular example of a language model is GPT-3 developed by OpenAI.

The transformer architecture is composed of a series of transformer blocks, which are made up of two main components: an attention mechanism and a feedforward neural network. The attention mechanism is implemented through self-attention layers that compute the importance of each word in a sentence by comparing it to all other words in the sentence. This enables the model to learn the dependencies between words and capture long-range dependencies more effectively than traditional recurrent neural networks. The feedforward network performs non-linear transformations on the data. By stacking multiple layers of attention and feedforward neural networks, the Transformer model can learn increasingly complex representations of the input sequence. The transformer-model architecture is shown in Figure 4.1.

Transformer-based Language models can be accessed through the Hugging Face library[5]. Hugging Face is an open-source library that provides state-of-the-art machine-learning models for NLP. It offers a range of pre-trained models that can be used for various NLP tasks. Many organizations will make their pre-trained transformer-based language models available for free on the Hugging Face library. For transformer-based language models to be used onto a particular dataset, it has to be fine-tuned with that dataset, which is called transfer learning. Fine-tuning means optimizing the language model on a particular dataset to perform better. Any language model can be loaded from the Hugging Face library and can be fine-tuned, this model can be stored again in the Hugging Face library. Whenever the fine-tuned model is needed, it can be loaded from the Hugging Face library.

For this specific classification task on the project's data, the DistilBERT[3] and RoBERTa language models[2] were used. DistilBERT is a pre-trained transformer-based language

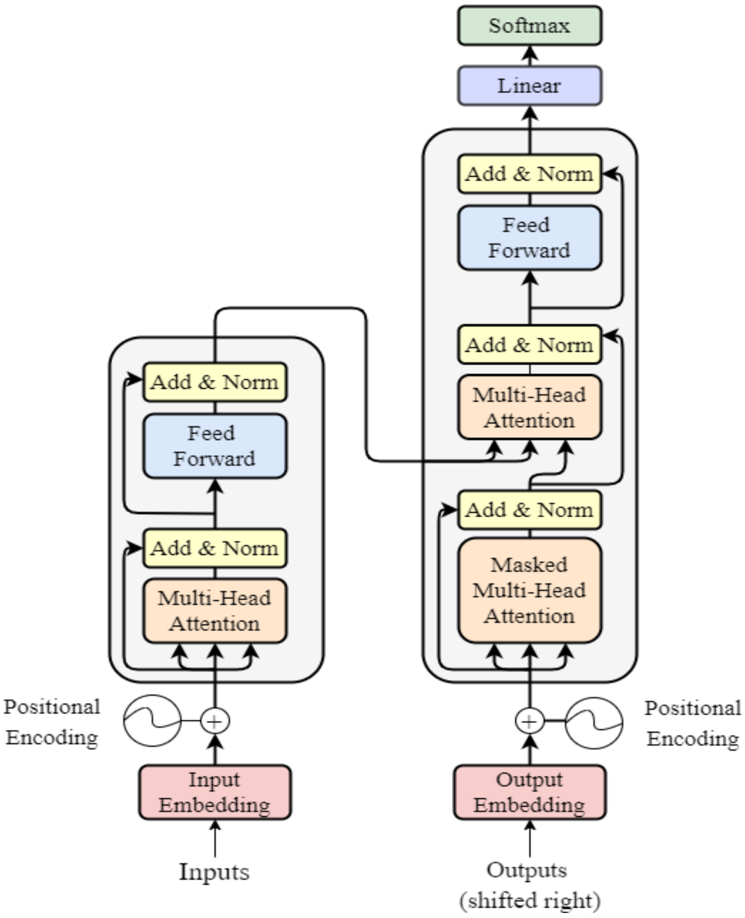


Figure 4.1: Transformer model architecture

model developed by Hugging Face. It is based on the architecture of BERT (Bidirectional Encoder Representations from Transformers), a popular pre-trained language model developed by Google. However, DistilBERT is smaller and faster than BERT, making it more efficient for certain natural language processing tasks. To pre-train DistilBERT, Hugging Face used a similar approach to BERT, which involves training the model on a large corpus of text data. Specifically, DistilBERT was pre-trained using a combination of two tasks: masked language modeling and next sentence prediction. In masked language modeling, the model is trained to predict missing words in a sentence, while in next sentence prediction, the model is trained to predict whether two sentences are consecutive in a text corpus. The pre-training process for DistilBERT involves several iterations of fine-tuning the model on large amounts of text data, with the aim of maximizing its ability to understand and generate coherent sentences. Once pre-trained, DistilBERT can be fine-tuned on specific natural language processing tasks.

RoBERTa (Robustly Optimized BERT approach) is a transformer-based language model that was developed by Facebook AI Research (FAIR) in 2019. RoBERTa was pre-trained on a large corpus of diverse text using a modification of the BERT pre-training procedure. The pre-training involved masking tokens, predicting the masked tokens, and predicting the next sentence in a pair of sentences. These models employ the cutting-edge transformer architecture and have been pre-trained on enormous quantities of text data, enabling them to capture intricate language patterns and representations. Refining these language models on the current project's dataset has the potential to yield superior accuracy and performance in categorizing fake news.

The selection of the aforementioned techniques was based on their applicability to the current problem at hand. Supervised machine learning models are widely used for text classification tasks and are an appropriate starting point for addressing the issue of fake news classification. The chosen algorithms strike a balance between accuracy and interpretability and their effectiveness has been widely acknowledged in relevant literature. The integration of pre-trained language models enables the utilization of knowledge obtained from large-scale language modeling tasks, potentially resulting in enhanced performance. The combined use of traditional machine learning algorithms and contemporary language models provides a comprehensive and robust approach for tackling the challenge of fake news classification on social media. The effectiveness of these techniques was evaluated via meticulous experimentation and model evaluation, with the results of said evaluations being extensively presented and discussed in the other sections of the project report.

4.2 Procedure

The process and methods employed in this project were designed to effectively address the problem of fake news classification on social media and achieve the stated objectives. The responsibilities were clearly identified, and the skills gained from core courses in Data Science and Analytics (DSA) were applied throughout the project.

The project commenced with data collection, where a large dataset of news classification was obtained from Kaggle. The data was pre-processed, including removal of stop words,

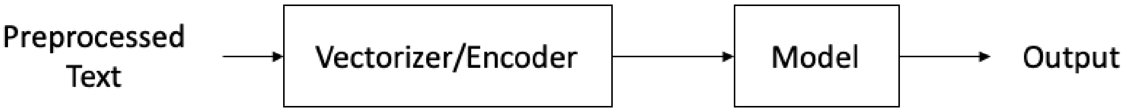


Figure 4.2: Supervised ML Algorithms high-level process flow

puntuations and non-alphanumeric characters. The pre-processed data was then split into training and testing sets using the train-test-split function from the Scikit-learn library, with appropriate stratification to ensure balanced class distribution.

Next, various techniques were applied to transform the text data into numerical features for machine learning algorithms. Term Frequency-Inverse Document Frequency (TF-IDF) vectorization was employed to convert the text data into numerical representations, capturing the importance of words in each title. Label encoding was applied to the output variable, representing the class labels of real or fake news.

Supervised machine learning algorithms were then applied to the training data for model training and evaluation. Support Vector Machine (SVM), Naive Bayes, Random Forest, Decision Tree, and k-Nearest Neighbors were implemented using their respective implementations in Scikit-learn. Cross-validation is used when training the models in order to avoid under-fitting and overfitting of the dataset. The project also involved iterative experimentation and refinement, including hyperparameter tuning to optimize the performance of the machine learning models. Multiple iterations were performed to fine-tune the models and select the best-performing approach. The process flow of the supervised machine learning algorithms as explained above is pictured in figure 4.2. The performance of each algorithm was assessed using various evaluation metrics such as accuracy, precision, recall, F1-score, AUC ROC score and ROC curve.

Additionally, language models from the Hugging Face library, namely DistilBERT and RoBERTa, were fine-tuned on the project's dataset. These state-of-the-art models were pre-trained on large-scale language modeling tasks and were fine-tuned using transfer learning to adapt to the specific task of fake news classification. Firstly the preprocessed text is tokenizer and then converted into tensors by the tokenizer of the language model, and then it it sent into model to get output. The tensors generated by the tokenizer can be of either pytorch, Tensorflow or plain NumPy, the type of tensor can be decided when passing the tokenization. The high-level process flow of the transformer-based language

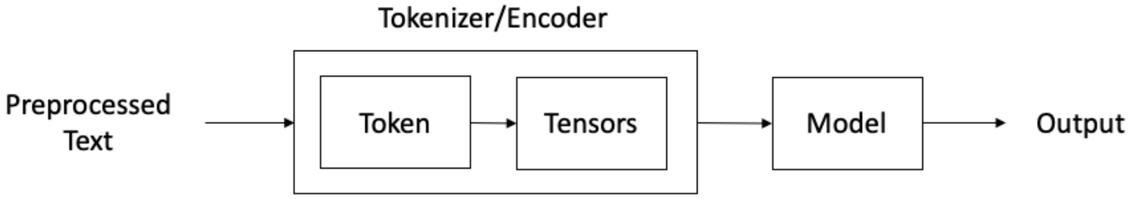


Figure 4.3: Transformer-based LM high-level process flow

models as explained above, is pictured in figure 4.3. The performance of these language models was also evaluated using the above-mentioned metrics.

Throughout the project, skills gained from core DSA courses were applied, including data preprocessing, data exploration, feature engineering, and model evaluation. Techniques learned in courses such as machine learning, natural language processing, and intelligent data analytics were effectively utilized. The entire process was well-documented, including the details of the techniques applied, the results obtained, and the lessons learned during the project.

In conclusion, the process and methods applied in this project followed a systematic approach, incorporating data preprocessing, feature engineering, model training, and evaluation using a combination of supervised machine learning algorithms and state-of-the-art language models. The skills gained from core DSA courses were effectively applied, and the responsibilities were clearly identified. The results and findings of this project will be presented in detail in another section of the project report.

Chapter 5

Results and Analysis

The table 5.1 shows the different models that are implemented along with the parameters used and their options. The selection column of the table shows the best model parameters. The Accuracy % column displays the results of the best models.

Table 5.1: Hyperparameter tuning

Model	Hyperparameter	Option	Selection	Accuracy(%)
SVC	tfidf-max-features	[1000, 5000, 10000]	10000	94.7
	C	[0.1, 1, 10]	1	
	Method	['sigmoid', 'isotonic']	sigmoid	
Naive Bayes	tfidf-nb-max-features	[1000, 5000, 10000]	10000	93.4
	alpha	[0.1, 1.0, 10.0]	1	
RF	tfidf-max-features	[1000, 5000, 10000]	10000	93.3
	n-estimators	[50, 100, 200]	200	
	max-depth	[None, 10, 20]	None	
	min-samples-split	[2, 5, 10]	5	
KNN	tfidf-max-features	[1000, 5000, 10000]	10000	90.8
	n-neighbors	[3, 5, 7]	7	
	weights	['uniform', 'distance']	distance	
	p	[1,2]	2	

From the table could see that SVC model got best accuracy compared to other models. The other metrics of the models are discussed in the later section.

The parameters of the transformer-based language model are present in the table.5.2

Table 5.2: Transformer-based language model parameters

Parameters	Selection
num-train-epochs	3
per-device-train-batch-size	16
per-device-eval-batch-size	64
warmup-steps	500
weight-decay	0.01
logging-steps	10
evaluation-strategy	steps
eval-steps	50
learning-rate	5e-5

The models and their evaluation metrics are shown in the table 5.3. All the models ROC curves are also attached below.

Table 5.3: Model Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC Score (%)
KNN	90.8	91.1	91	91	97.1
RF	93.3	93.4	93.3	93.3	98.2
Naive Bayes	93.4	93	93	93	N/A
SVC	94.7	94.8	94.8	94.8	98.8
DistilBERT	96.5	96.3	97.2	96.8	96.4
RoBERTa 2 features	96.8	96	98.1	97.1	96.7
RoBERTa 1 feature	96.9	95.8	98.5	97.1	96.7

In this project binary classification is performed where news is either divided into fake or real. This classification has 4 possible outcomes as below

- True Positive (TP): The model predicted the news article as ”real” (1) and it is actually ”real” (1).
- False Positive (FP): The model predicted the news article as ”real” (1) but it is actually ”fake” (0).

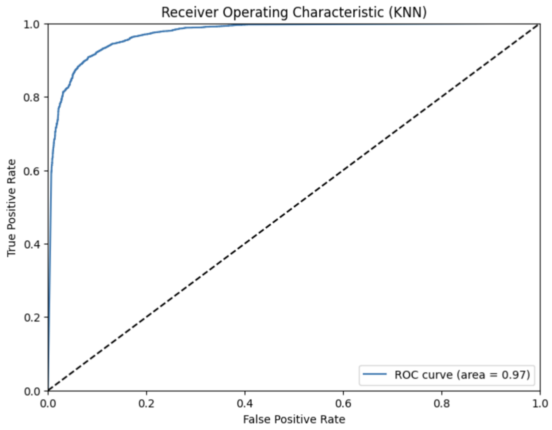


Figure 5.1: KNN ROC Curve

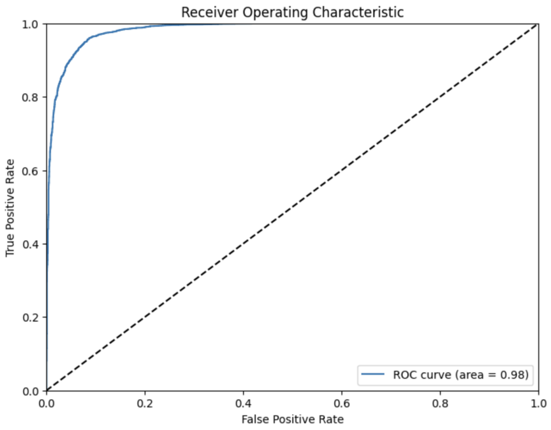


Figure 5.2: Random Forest ROC Curve

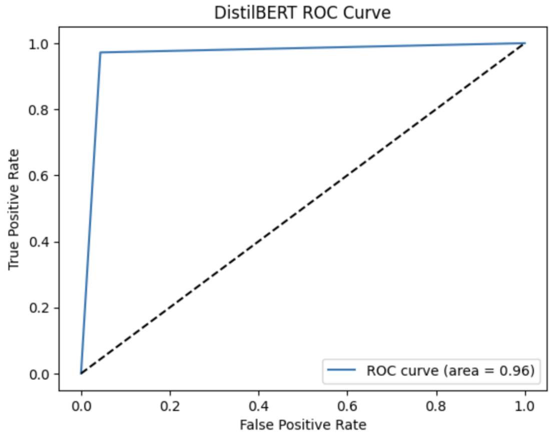


Figure 5.3: DistilBERT ROC Curve

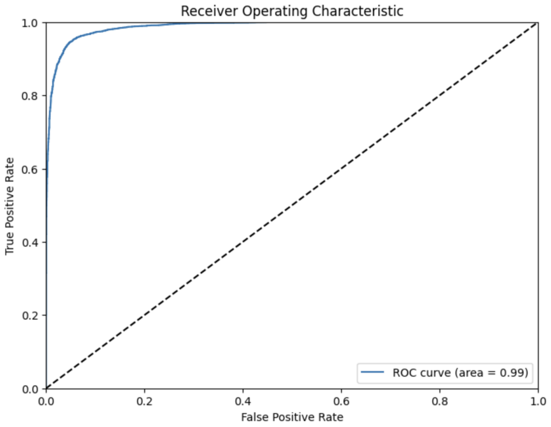


Figure 5.4: SVC ROC Curve

- True Negative (TN): The model predicted the news article as ”fake” (0) and it is actually ”fake” (0).
- False Negative (FN): The model predicted the news article as ”fake” (0) but it is actually ”real” (1).

Below explains the objectivity of each metrics when they are maximum:

Maximize accuracy: The model aims to correctly classify as many news articles as possible, regardless of the class distribution or the potential impact of misclassifications. This objective is useful when the cost of false positives and false negatives is roughly equal or when the classes are relatively balanced. In the project’s dataset, the class distribution is balanced so this metric is also considered as one of the important evaluations to determine

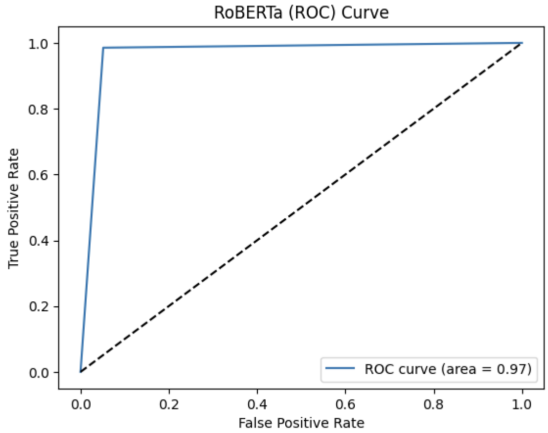


Figure 5.5: ROBERTa 1 feature ROC Curve

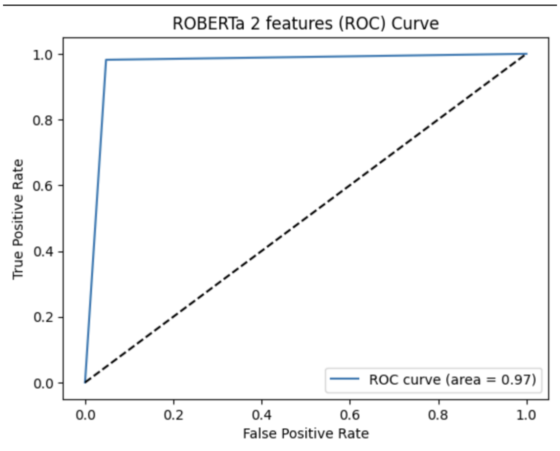


Figure 5.6: ROBERTa 2 features ROC Curve

the best model.

Maximize precision: The model aims to minimize the number of false positives, i.e., the instances where the model incorrectly labels a genuine news article as fake. This objective is useful when the cost of false positives is high, such as in critical applications where misclassifications can lead to legal, financial, or reputational consequences.

Maximize recall: The model aims to minimize the number of false negatives, i.e., the instances where the model incorrectly labels a fake news article as genuine. This objective is useful when the cost of false negatives is high, such as in applications where misclassifications can lead to public health risks, social unrest, or political manipulation.

Maximize F1 score: The model aims to balance precision and recall by maximizing the harmonic mean of the two metrics. This objective is useful when both false positives and false negatives are important and need to be minimized simultaneously.

Maximize ROC AUC score: The model aims to maximize the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate at different thresholds. This objective is useful when the cost of misclassification depends on the specific trade-off between sensitivity and specificity, or when the class distribution is highly imbalanced. In the project's dataset, the class distribution is balanced so this metric is not the priority metric for the model evaluation.

In the case of detecting fake news articles, the ultimate goal is to minimize the false positive, i.e to minimize the article's prediction as real when it is actually fake. This means the model's precision should be high. This helps in identifying the fake news correctly and helps in stopping the spread of misinformation. If considering precision as best metrics then DistilBERT and RoBERTa are considered as best models. If considering accuracy as best metrics then RoBERTa model is best. From the ROC AUC Score, could see that RF and SVC performed better than language models. Finally, considering all the metrics instead of considering only one metrics, language models performed better and RoBERTa models can be considered as best.

It can be seen that a good accuracy is achieved only when used ”Title” of the news instead of ”Context” of the news. Title of the news consists of only little information, but context of the news consists of all the information, yet predictions with just the title of the news gave good results. When researched about my good accuracy results of title, it was found from a research paper that fake news contains most of the information in title, whereas the body of the news is satire[1].

Chapter 6

Deliverables

The deliverables or outcomes of this project include a robust and accurate machine learning model for news classification that can effectively differentiate between real and fake news on social media platforms. This model was developed using state-of-the-art natural language processing techniques, algorithms, feature engineering, and hyperparameter tuning. The model's performance will be evaluated using various metrics, such as accuracy, F1 score, recall, ROC curve, and AUC ROC score.

In the context of the problem objectives, this project's outcomes can significantly impact the research and business objectives. By accurately classifying news as real or fake, the project can raise public awareness about the need to fact-check information before sharing it on social media. This can help promote responsible information-sharing practices, ultimately combating the spread of fake news.

Moreover, developing a highly accurate machine learning model for news classification can contribute to natural language processing by exploring state-of-the-art techniques and methodologies. The project's outcomes can also significantly impact the integrity of elections and democratic decision-making by detecting and removing fake news from social media platforms, thus protecting democratic institutions and processes.

In summary, the project's deliverables and outcomes can effectively address the critical problem of fake news on social media and promote responsible information-sharing practices. The project can also advance the field of natural language processing and contribute to protecting the integrity of elections and democratic decision-making.

Bibliography

[1] Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766, 2017.

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

[3] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[5] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Self-Assessment

My learning objectives were:

- To use data science for solving real-world problems.
- To understand the dataset well.
- To improve my problem-solving and critical-thinking skills.
- To use transformer-based language models for predictions.

I am pleased to report that I successfully achieved my individual learning objectives. Initially, I was able to identify a real-world problem that could be addressed through the use of data science. My project focused on classifying news as either fake or real, which would aid in detecting and preventing the spread of false information.

To achieve this objective, I employed several critical DSA skills. Firstly, I conducted a thorough analysis of the dataset to gain a comprehensive understanding of the information it contained. Furthermore, I created new features to determine if they would be useful in predicting my model's accuracy. Specifically, I analyzed the title's sentiment and evaluated its significance in achieving better predictions.

Lastly, I utilized transformer-based language models, such as DistilBERT and RoBERTa, to classify my dataset accurately. As a result, I gained valuable experience in utilizing machine learning algorithms and became proficient in working with language models.

The accomplishment of the individual learning objectives mentioned above requires various Data Science and Analysis (DSA) skills. Firstly, identifying a real-world problem and formulating it as a classification problem requires knowledge of classification algorithms.

Secondly, understanding the dataset requires various DSA skills, such as data cleaning, data pre-processing, data visualization, and data analysis. Thirdly, identifying the sentiment of the news title requires text processing and natural language processing (NLP) skills. Techniques like tokenization, stemming, lemmatization, stop-word removal, and others are required for text processing. Furthermore, the use of transformer-based language models like DistilBERT and RoBERTa requires knowledge of advanced NLP and deep learning algorithms.

The skills that I needed to learn individually to complete this project were:

- To do data analysis I had to learn Tableau.
- I had to do Hugging Face course to understand about transformer-based language models.
- I had to learn on how to use language models to do the classifications.

The practicum was for 4 credit hours. It is my own project, Dr. Matthew Beattie supervised my project. Name: Dr. Matthew Beattie Title: Adjunct Professor Contact: mjbeattie@ou.edu