Fake Posts Detection on Reddit using Supervised learning model

Abhishek Devidas Pawar*, Rakesh Thota†, Surya Vijayakumar‡ *abhishekdevidaspawar@trentu.ca
†rakeshthota@trentu.ca ‡suryavijayakumar@trentu.ca

***Abstract*—Detecting fake posts on social media platforms like Reddit has gained importance due to the proliferation of satirical and fake news. Classification of fake and genuine posts is not an easy task considering the volume of data generated and the various techniques used by news generators to conceal the fake or satirical aspects. This paper takes into consideration various attributes of Reddit posts like the body of the post, title, number of comments, score, author, and time at which the post is created to classify the post as fake/satirical or true/genuine. Building on previous works in fake news detection, we propose 3 classification algorithms trained on Reddit posts scraped from 2 subreddits; TheOnion and NotTheOnion, to identify whether a post is fake or genuine. We have attained an accuracy of 97.94% and an F1 score of 0.978 with Random Forest which is an ensemble of Decision trees.**

I. Introduction

Fake news/posts spreading across the web affects individuals, organizations, governments etc in both positive and negative ways. But the negative effects exceed the positive ones. UCF Lecturer Chrysalis Wright was appointed to the board of directors for the United Nation's Communications Coordination Committee, which aims at improving communications across cultures globally during the 2016 US elections. Her research revealed that opposition parties and other countries were involved in circulating disinformation in social sites on hot topics and they were trying to create internal conflicts that could have changed the results of the elections. Similarly, there are many examples where it is affecting the finances of organizations, the life of individuals and the credibility of governments.

Digital platforms have given a new powerful medium to spread fake news to millions of users. Fake news arises due to poor fact-checking, sloppy writing, and mainstream media bias. The flow of fake news happens mainly from Originators to high-level spreaders and then to low-level spreaders. Originators are the ones who create the news to spread. They could have a motive like political propagandists or they could be people who rely on it to make a living. It could also be targeted at an individual or an organization to defame them. The high-level spreaders are the main link in the process who spread the news easily through their Facebook groups, blogs, or fake news websites. Consumers who like or share the news act as low-level spreaders. Readers are manipulated and they take important decisions with prejudice.

As Fake news/posts have such widespread and long-lasting effects, we are building a reliable machine learning model that can help to detect fake posts. We chose the social media platform called Reddit as it has about 52 million daily active users and 430 million users who use it at least once a month.

II. Related Work

Fake news detection as general and especially related to social media has been discussed in literature from different perspectives. There are many studies about how fake news or controversial discussions spread through social media platforms like Facebook, Twitter, and Reddit, and the factors that contribute to its proliferation. For instance, the work done by [1] elaborates that the factors that make users vulnerable to fake news are: naive realism and Confirmation bias. These cognitive biases are inherent in human nature and fake news is often considered as real by consumers.

Fake news happens when publishers go for short term monetary benefits rather than long term benefits. And when consumers go for psychological benefits rather than information utility. The argument was confirmed based on human behavioral patterns by [2]. They explained that controversy within a discussion on social media attracts more users and promote the spread of the material further and faster. The researchers used methods like descriptive analysis and network analysis to monitor the amount of activity produced by a cascade. They also employed k-core and temporal analysis of the evolution of contributors. It was concluded that controversial contents were associated with higher degree of activity compared to non controversial ones and it attracted more new users into the cascade.

As fake news travel faster than genuine news, many propagandists use this as an opportunity to widen their supporters and many accounts on social media seeking to create more influence tend to do it by creating and spreading fake news. These articles gave us an insight into the importance of detecting fake news and posts and its importance in tackling fake news proliferation.

A number of projects have been implemented to perform fake news detection using different Machine learning algorithms and yielding results with varying accuracy. The questions like how the internet and social media have made access to news easy and how this has made the spread of the false information easy and undetected was answered by [3]. The authors tried to solve the problem of fake news detection using Naïve Bayes classifier for which they used a manually labelled news data-set. They tested their model on a newly acquired data-set and evaluated the performance of their model. They shared some similarities between spam messages and fake news.They often tend to have a lot of grammatical mistakes, are often emotionally coloured, use similar limited set of words and try to affect reader's opinion on some topics in a manipulative way. These similarities makes them almost indistinguishable and thus same methods can be used to detect fake news as are used for spam messages. They have used the

data-set collected by Buzzfeed news to test their NB classifier model. The authors also shared some ways to improve the performance of the classifier. Some of them being, removing stop words, using long length news articles and using words in their base form by removing any prefix or post-fix from them. They concluded that the NB classifier worked great with a simple model they built.

A framework called TriFN was developed by [4] in 2017 to utilize the tri-relationship(the relationship between publisher, news pieces and users) inherent in social media to verify the authenticity of posts.This framework exploited user-news interactions and publisher-news relationships. Tri-relationship embedding consisted of 5 components including news content embedding using the Non negative matrix factorization for document representations, user embedding, user news interaction embedding, publisher-news relation embedding and a linear classifier. Their study showed that social context based features are more effective than news content based features and combining them produced better results compared to stand alone techniques.

The issue of publishers spreading content based fake news commonly referred to as "Clickbait's" was addressed by [5]. Clickbait's are nothing but phrases those attract the attention of the users, who upon clicking the link will be redirected to a new web page where the content that they find is irrelevant and most of the users feel these clickbait's as extremely distracting. The authors used WEKA tool which was designed specifically to detect and eliminate web pages that contained the misinformation intended to mislead the readers. Their first step was to identify the credible clickbait's database and later compute the attributes and load the file into WEKA. However, they also crawled the web to collect the URLs for the clickbait's by focusing on social media platforms like Facebook, Forex, and Reddit. After gathering the URLs into a file, they wrote a Python script that computed the attributes from the title and the content from web pages. They extracted the features like keywords in Arabic, English, titles that started with numbers, all uppercase words, containing question and exclamation marks, if user left the page immediately, and content related to title. Authors used classifier algorithms like Bayes Net, Logistic, Random Forest and Naïve Bayes in WEKA to evaluate the given data. They compared the classifiers performance based on the metrics like Precision, Recall, F-Measure, and ROC. The Logistic and Random Forest classifiers gave the best Precision, Recall and F-measure scores whereas Bayes Net and Naïve Bayes gave the highest area under the ROC curve.

A solution for fake news detection was proposed by [6] in 2020. They shared two strategies for automatic identification of fake news posts : facts and style of content or stance and propagation of social. They used Naïve Bayes classifier for detecting fake news (facts and style of content) as it has high level of accuracy, requires small training data-set and is fast at creating relations and modelling the data-set. To improve the accuracy of the model, they used weights for the words. Some words were given more weight than others. The authors used TFIDF vector to count the number of words and number of unique words in the news post and allotted weight to each of them to reduce the prediction time of the model and increase accuracy.

A method for Fake news Detection on Reddit was proposed by [7] the following year. The authors of the paper shared a problem of how recently fake news spreaders have started to mix real/authentic news with fake news to make it look more authentic and real. This has made the already tough task of fake news detection more difficult. They used 72000 posts from Reddit subreddits as their data-set and used some NLP and ML algorithms for fake news detection. The authors built 6 different models to find which combination of NLP and ML algorithms worked best with the collected data. The data-set used by the authors contained the title and subreddit from which it was taken. The model used only title to classify if the post is fake or real. They have used Counter Vectorizer, Logical Regression, TFIDF, MultinominalNB, and SVM in their research. The best combination of algorithm they got was from CountVectorizer and MultinominalNB in which they achieved 88.73% accuracy and a misclassification rate of 10.44%.

Another team of researchers also emphasised the importance of title in classifying news articles as fake or genuine. [8] proposed an approach in 2021 to classify news based on title and then comparing the results obtained with classification based on the entire news(text), a method that balanced data analysis time and quality. They used Natural Language Processing (NLP) in their work by following methods like tokenization, stop words, stemming and lemmatization, and word weighting measures which helped to prepare text for analysis aimed at standardizing, cleaning, and reducing the variability in text. Authors used machine learning algorithms like Support Vector Machine (SVM), Classification and Regression Trees (CART), and Random Forests to build the model and they tested the model on a real data-set containing both political and world news. The model built with SVM algorithm gave an accuracy of 0.9419 with a higher running time whereas the model built with random forest algorithm gave good balance between time and quality of classification with accuracy of 0.9269 and a running time of 7.82 seconds. The authors confirmed that their proposed model produced good results after the initial analysis of news titles.

Michał Choraś and his team published a systematic mapping study of Advanced Machine Learning techniques for fake news detection [9] in 2021. The authors have done a systematic study of research papers, projects, and current initiatives worldwide to fight against disinformation. They mainly addressed systematic overview of ML approaches for fake news detection and those were: Text Analysis, Natural Language Processing Based Data Representation (Psycho-linguistic Features, Syntax Based, Non-linguistic Methods), Reputation Analysis, Network Data Analysis, and Image Based Analysis and Detection of Image Manipulations.

The paper stated that text analysis is the most common way to recognize fake news in NLP and it is done in the form of Bag-of-words or N-grams. Whereas during NLP based data representation, certain challenges have to be addressed by methods like, psycho-linguistic features, syntax based, and non-linguistic methods. In reputation analysis, reputation of an

individual, organization, or a place is based on social assessment and thus the authors analysed the type of comments, feedback and discussions happening on social media platforms. Network data analysis requires studying graphs which are representation of either symmetric or asymmetric relations between discrete objects and it examines both homogeneous and heterogeneous networks. During image-based analysis and detection of image manipulation, images are analysed using Convolutional Neural Network (CNN), and Structural Similarity Index Measures (SSIM). The authors also pointed out more challenges in the fake news arena like streaming nature of fake news, lifelong learning solutions, explain-ability of ML based fake news detection systems, and emergence of deep fakes.

From reviewing the literature, it is evident that multiple teams of researchers and big data enthusiasts have developed various approaches and ML algorithms for detecting fake news in social media and across the web in various domains. Many of them have used publicly available data-sets and some have gone an extra stretch to scrape their data. Various psychological and human relationship based studies have also been used to develop effective methods to tackle this ever growing problem of fake news detection. But, we have noticed that most of the studies focus on mainly one feature like the title of the article/post to determine the genuineness of a post. For instance, [8] mentioned that the analysis of title alone gave them good results, but further analysis of the text of articles allows for a very good classification of true or fake news. News content features and Social context features could be used for detecting fake news too. News content features like linguistic based features, domain specific features and visual based features are good indicators of fake news. Social context features like users, generated posts and networks can also be utilized for classifying a piece of news as fake or genuine. We are also incorporating likes, dislikes, awards and other available attributes of Reddit posts to improve the existing models taking inspiration from [3]. While developing the model, we will also explore methods to use post body and title together and also the advantages and disadvantages of using an unsupervised method [7].

III. Project Objectives

Fake news includes news articles that are intentionally false and could mislead readers. The idea of fake news has been in existence even before the Internet arrived. Previously publishers used to spread fake news through mainstream media channels but nowadays there are millions of users across the web which makes spreading fake news across social media networks very easy. Social networks like Reddit, Facebook, and Twitter provided a platform to share content freely but not without negative effects like polarisation of opinions and controversial discussions. So, we have decided to take some steps towards solving the problem of fake news in social media. As fake news detection is a vast topic that requires inputs and collaboration from many disciplines, we are going to focus on detecting fake posts in one of the most widely used social platforms, Reddit.

Reddit is a popular online platform that helps users to share opinions and knowledge through posts that are either text or images. Other users can respond to these posts by commenting or by voting for or against them. These votes are called up-votes or down-votes. Like-minded users tend to join spaces called 'subreddits' where focused discussion on a particular topic or field occurs. As of September 2021, Reddit ranks as the 19th most visited website in the world and 7th most visited in the US according to Alexa Internet.

Fake news detection is usually considered as a binary classification problem and it is essentially distortion bias on the information manipulated by the publisher. Multiple teams of researchers have developed various approaches and ML algorithms for detecting fake news in social media and across the web in various domains. But, most of the studies focus on mainly one or two features like the title of the article/post and language features to determine the genuineness of a post. To solve the problem of identifying fake news, we need to employ more sophisticated methods which can be used effectively in real-time. Taking more features of social media posts into consideration can help considerably improve the efficiency of the fake post-detection model.

Fake news detection generally involves 2 phases. The first one is feature extraction in which the news content or the post itself and its auxiliary information is represented in a formal mathematical structure. By focusing on detecting fake posts in Reddit, we will be able to use the data obtained with Reddit API efficiently. In addition to the title of the post, we will consider the body of the post, score, number of comments, etc. which makes our method of solving this problem different from the former approaches in this area. The second phase builds machine learning models to better differentiate fake news and real news based on the feature representations. The large data set prepared by scraping the subreddits gives us clearly labeled data that can be used effectively for training and testing the model.

IV. Description Of Study

A. Data Source

Even though there are many data sets that are available for Reddit fake-post detection, most of them do not have the attributes that we are using in our model. So, we have web scraped 40k rows of data from two subreddits by using Reddit API. r/nottheonion [10] is a subreddit for posting true news. the subreddit description says: "For true stories that are so mind-blowingly ridiculous that you could have sworn they were from The Onion". This subreddit has around 20.7 million members. r/TheOnion [11] is a subreddit in which fake news is posted. By scraping 20K recent posts each from r/TheOnion [11] and r/nottheonion [10], we have up-to-date training and testing data set that is labeled. We have also scraped another 20K rows of data from the subreddit 'r/worldnews' [12] using the Reddit API for testing purposes after we have a model with high accuracy.

B. Data Description

We cleaned the data and created a data frame with 22 attributes. Upon further investigation that led to a better

understanding of the data, we removed the attributes that were highly correlated. After the initial pre-processing, we have the following 6 attributes in the final data-set: author, created_utc, num_comments, score, title, and URL. 'author' is the username of the author, 'created UTC' is a timestamp that gives information about the time at which the post was created, 'num_comments' is the number of comments obtained on a post, and 'score' is the difference between the number of upvotes and the number of downvotes. The title gives the title of the post. 'URL' is the URL of the post which can be used to access the body of the post.

TABLE I
TABLE SHOWING DATA ATTRIBUTES AND THEIR DESCRIPTION.

No	Attribute name	Attribute Description
1	label	Classification label with fake posts as 1 and genuine posts as 0
2	title	Title of the post
3	score	Difference between the number of upvotes and downvotes
4	author	Name of the user who submitted the post on Reddit
5	num_comments	Number of comments received for a post
6	created_utc	Time at which the post was created
7	post_body	Content of the post

C. System Model

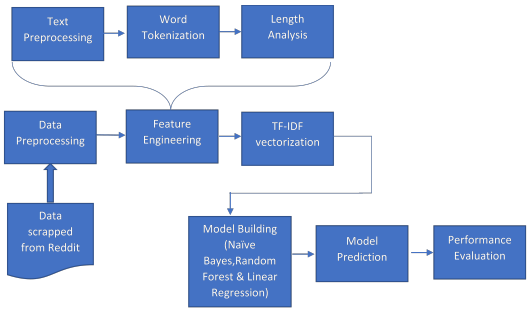
The system that we propose to solve the problem of fake news involves stages like data collection, data pre-processing, feature engineering, model building, and performance evaluation. Figure 1 represents our system model.

The first step, data collection using Reddit API was done with PUSHSHIFT function from PSAW library to scrap Reddit posts. We obtained our post data from three subreddits named TheOnion, NotTheOnion, and WorldNews.

Data pre-processing involved removing missing values and removing duplicate rows, using URL from scraped data to get post body and then removing the URL column, converting UTC to local time, combining data from TheOnion and NotTheOnion and adding a class label with 1 for fake posts from TheOnion and 0 for genuine posts from NotTheOnion. post_body was obtained by extracting body of the post using a library called Newspaper3k. It crawls through the web page and extracts the text within HTML tags and returns it as a string. The relevant attributes that we are using for building the model are listed in Table I. Further pre-processing included converting UTC to local time and removing the created_utc attribute.

The next stage is Feature Engineering which includes processes like Text Preprocessing, Word tokenization and Length Analysis. Text Preprocessing is the stage in which removing stop words and punctuation, converting all text to lowercase, and lemmatization of the text is done. Lemmatizing is swapping a word with its base or dictionary form of the word. Word tokenization is to split a sentence into a list of words. Tokens consist of 'N' number of consecutive words known as n-grams. Finally, length analysis is the stage in which we find average word length and sentence length to get better insights. Next, we assign weight to the remaining words which actually provide meaning to the text body. We are using TF-IDF (Term frequency-inverse document frequency) which is a text vectorizer that transforms the text into a usable vector, where each word is given a weight to show how important it is in the text. This will be the main attribute in model building where the model will understand how to differentiate fake and authentic news posts.

Fig. 1. System Model



After completing the pre-processing steps, the final data set is fed to the model. The model with the best accuracy is used in the label prediction process. This step is followed by matrix evaluation to check the overall performance of the model. The final testing of the model is performed using data from WorldNews that doesn't have assigned labels. This helps to see how our model works with real-world data.

D. Formulation

The four main algorithms those are used to build the model are TF-IDF Vectorizer, Multinomial Naïve Bayes, Random Forest, and Logistic Regression.

1) TF-IDF : To process natural language, the raw text needs to be transformed into a vector form which has been done using the TF-IDF algorithm(term frequency-inverse document frequency) and it is a text vectorizer that transforms the text into a usable vector. The main challenge is to choose the best method for numerical representation of the text strings. We considered the two techniques Count vectorizers and TF-IDF. But it seems that the latter is a better choice as it not only focuses on the frequency of the words present in the corpus and it also provides the importance of words. Removing the least important words makes easier in building the model.

The term frequency(TF) refers to the number of times a term appears in a document and it displays each text in the data as a matrix with the number of documents in the rows and the number of distinct terms in the columns. The number of documents that include a specific term is known as document frequency. The weight of a term is determined by the inverse document frequency (IDF), which attempts to minimize the weight of a term if the term's occurrences are scattered throughout all documents. IDF and TF-IDF scores can be calculated as follows:

$$idf_i = log(n/df_i) \tag{1}$$

Where, idf_i is the IDF score for term i, df_i is the number of documents containing term i and n is the total number of documents.

The lower the IDF for a term, the higher the term's DF. Because $log(1)$ is 0, the IDF will be zero when the number of DF equals to n, which means the term appears in all documents. This term can be added to the stop-word list because it doesn't provide much information.

$$w_{i,j} = tf_{i,j} * idf_i \tag{2}$$

Where, $w_i,_j$ is TF-IDF score for term i in document j. $tf_{i,j}$ is term frequency for term i in document j. idf_i is the IDF score for term i.

2) Classification: Classification is a form of machine learning that uses a supervised learning approach. It consists of two steps; a learning step and a classification step. In the first step, the algorithm builds a model by learning from a data set that has class labels. The model is then used to predict the class labels for the test data set. This prediction can be used to determine the accuracy of the classifier.

3) Naive Bayes: The Naive Bayes [13] is a strong algorithm for analyzing text data and solving problems with numerous classes. As the Naive Bayes algorithm is based on the Bayes theorem, it is necessary to first comprehend the Bayes theorem notion. The Bayes theorem, which was developed by Thomas Bayes, calculates the likelihood of an event occurring based on prior knowledge of the event's conditions. It's based on the formula below which calculates the probability of class A when predictor B is already provided.

$$P(A|B) = P(A) * P(B|A)/P(B) \tag{3}$$

$P(B)$ is the prior probability of B, $P(A)$ is the prior probability of class A and $P(B|A)$ is the occurrence of predictor B given class A probability.

4) Random Forest: Random Forest is a supervised learning model and it uses labeled data to learn how to classify unlabelled data. It is an ensemble classifier that estimates the output based on the combination of multiple decision trees which are in general weak learners. When performing Random Forests based on classification data, we are using the Gini index, or the formula used to decide how branches form from each node.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{4}$$

This formula calculates the Gini of each branch on a node based on the class and probability, deciding which branch is more likely to occur. The relative frequency of the class we're observing in the data-set is represented by pi, and the number of classes is represented by c. We will also use entropy to determine how nodes branch in a decision tree.

$$Entropy = -\sum_{i=1}^{c} p_i * log_2(p_i) \tag{5}$$

Entropy is used to determine how the node branches based on the probability of a specific outcome. It is more mathematically complex than the Gini index because of the logarithmic function used to calculate it.

5) Logistic Regression: Logistic regression has been widely used for binary classification problems. In the logistic model, the independent variables can be either binary or continuous and the dependent variable can take 2 values, 0 or 1. The model gives a probability between 0 and 1 for each class.

E. EXPERIMENTS

The solution to our problem of identifying fake news on Reddit posts is to use the post body as well as other identifiers such as the number of comments, author name, time at which the news was posted, and the score of the post. The experimental settings and methods, evaluation parameters that we used, and performance evaluation and comparison of the performance of different models are explained below.

1) Experimental Settings: IDLE and Jupyter notebook environments were used for most of the tasks involved. We have used Pushshift API for scraping the Reddit posts and a library named Newspaper3k for downloading the body of posts with the post URL as a parameter. The Scikit-learn library in python was used extensively for performing various steps like $tf-idf$ vectorization, train test splitting, and model building.

Different NLP techniques were employed to obtain the features required for model building. The 'langdetect' library was used for language detection to filter out English posts. The 'nltk' library was used for performing NLP preprocessing such as removal of stopwords, and tokenization for creating a bag of words. Also, we have lemmatized the post body to make the data more useful and simpler for the model to understand. The author names have been encoded using frequency encoders. To reduce the multicollinearity and eliminate any label leak that might lead to undesirable models, we have used VIF (Variable Inflation Factor). The Variance Inflation factor is an indication of the multicollinearity of a variable and other predictors in the model. If a variable has a VIF of 10, it indicates that there is a serious case of multicollinearity. A value between 5 and 10 is acceptable but it still indicates that there is some relation between the predictor variables whose value is less than 2.5 which is desirable.

While performing feature engineering, we introduced more columns such as word count, character count, and average word length for both the post body and title of the post. TF-IDF vectorizer was utilized to provide weights to the words by using both uni-grams and bi-grams in a range of occurrence of words greater than 5 percent and less than 80 percent and thus it helps the model in differentiating the news(fake/real). Post vectorization, we had built different models for different scenarios by using Naïve Bayes classifier, Random Forest, and Linear logistic regression algorithms.

We faced some difficulties during the data collection and cleaning process as the Reddit API has restrictions such as 100 records per minute and many attributes/features were added/removed over the last few years. Because of this, we had to put extra effort to extract the data from subreddits

TABLE II
EVALUATION MATRICES

Features	Precision			Recall			F1 score		
	RF	NB	LR	RF	NB	LR	RF	NB	LR
Post body(tf-idf)	0.935	.821	.898	.907	.835	.892	.927	.828	.894
Title(tf-idf)	.714	.564	.696	.393	.764	.376	.507	.649	.488
Post body + Meta *	**.993**	.842	.955	.963	.904	.907	**.978**	.872	.930

(TheOnion, NotTheOnion & Worldnews). Upon extracting the data we had a huge task of cleaning and processing it for NLP functions and model building. This task became more difficult due to the unstable and unstructured Reddit API.

2) Evaluation Parameters: We have measured the performance of supervised machine learning models by using confusion matrix on the test data where it mainly has 4 elements True Positive(TP), True Negative(TN), False Positive(FP), and False Negative(FN). The following scores were been calculated using these elements.

$$Precision : TP/(TP + FP) \tag{6}$$

$$Recall : TP/(TP + FN) \tag{7}$$

$$F-Score : 2*Precision*Recall/(Precision+Recall) \tag{8}$$

3) Performance Evaluation: The F1 score results we obtained for the Random Forest, Naïve Bayes classifier, and Linear logistic regression are 0.978, 0.872, and 0.930, with Random Forest performing the best. Table II shows the comparison of the evaluation matrices. We took the F1 score because it is the harmonic mean of precision and recall and is a better measure than accuracy. In our context, a high precision means that we have few false positives which indicate that the number of real posts that are predicted as fake is few. It also means that we can correctly predict a large portion of true positives; the number of fake posts that are predicted as fake. A false negative is considered undesirable in fake news prediction. It indicates that fake posts are predicted as true, which can result in the uncontrolled spread of such news. But our high F1 score proves that we have fewer false negatives than false positives.

We tested our model on posts that we obtained from the subreddit r/worldnews. The majority voting technique; an ensemble method was used for classifying the posts. In this technique, all the 3 models make a prediction(vote) and the majority prediction is considered the final classification label of the post. After classification, we randomly picked a few posts from the data set and did some research to find out if the post is fake or real. We read the comments and verified the news on reliable websites. For example, one of the posts in the world news data set is: "People are attacking each other for food as Russian forces bombard Ukrainian city of Mauripol". Our model classified it as fake news. We verified that it was actually fake news by reading comments of people living in the place mentioned in the post and doing fact checks. We manually verified a few posts out of which almost 85% were correctly classified.

V. RESULTS AND DISCUSSION

This study demonstrates that a classification approach using post body, number of comments, title word count, and score can be used to accurately predict the genuineness of a Reddit post.

TABLE III
TRAINING AND TESTING ACCURACY

Features	Training Accuracy			Testing Accuracy		
	RF	NB	LR	RF	NB	LR
Post Body(tf-idf)	99.59	84.20	90.85	92.70	83.81	90.17
Title(tf-idf)	69.38	61.87	63.72	64.30	61.36	63.15
Post body + Meta *	100	87.46	94.10	**97.94**	87.59	93.65

* Attributes including score, no. of comments, author and time of posts

The results indicate that ensemble classification algorithms are generally more accurate than individual classifiers. Random forest is an ensemble of decision trees and uses a bagging approach. It was more accurate than other classifiers like Naive Bayes and logistic regression that we employed. We got an accuracy of 97.94% and an F1 score of 0.978 using Random Forest and TF-IDF. This accuracy is considerably higher than the 88.73% accuracy achieved by [7] in a similar study using Countvectorizer and Multinomial NB. Table III shows the training and testing accuracy for different models that we used.

The TriFN method which uses both news content and social contexts to detect fake news had attained only 86.4 % accuracy with an F1 score of .870 on the BuzzFeed data set. Another study that employed an FND-NS model that included both features of news and the social context of news exhibited an accuracy of 74.89 %.

VI. CONCLUSION

This study provides a model for how machine learning and NLP can be used to understand whether a post is fake or real. We have designed reliable models by using additional attributes like post body, scores, author name, and post time along with other commonly used attributes like the title of the post. We have also used suitable text pre-processing and feature engineering methods to build accurate models. The accuracy and F1-score improvement in our approach, when

compared to the approaches followed in previous studies are attained by the use of these additional attributes. These attributes provide a lot of insight into the post and help us predict if a given post on Reddit is fake or genuine. The prediction on the world news data set also produced remarkable output considering the fact that we tested our model on a data set obtained from a source that is different from our training data set. The set of words that we obtained after vectorization is slightly different in the data sets. There is ample scope for improving the model by using techniques like ELMo(Embeddings from Language Models) which uses contextual embedding instead of word embedding.

REFERENCES

1 Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H., "Fake news detection on social media: A data mining perspective," 2017.

2 Jasser, J., Garibay, I., Scheinert, S., and Mantzaris, A. V., "Controversial information spreads faster and further than non-controversial information in reddit," in *Journal of Computational Social Science*, 2021.

3 Granik, M. and Mesyura, V., "Fake news detection using naive bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017, pp. 900–903.

4 Shu, K., Wang, S., and Liu, H., "Exploiting tri-relationship for fake news detection," *CoRR*, vol. abs/1712.07709, 2017.

5 Aldwairi, M. and Alwahedi, A., "Detecting fake news in social media networks," *Procedia Computer Science*, vol. 141, pp. 215–222, 2018.

6 Chitra, A. J., "Fake news detection using naive bayes classifier," *Journal of Analysis and Computation (JAC)*, 2020.

7 Patel, A. and Meehan, K., "Fake news detection on reddit utilising countvectorizer and term frequency-inverse document frequency with logistic regression, multinominalnb and support vector machine," in *2021 32nd Irish Signals and Systems Conference (ISSC)*, 2021, pp. 1–6.

8 Probierz, B., Stefański, P., and Kozak, J., "Rapid detection of fake news based on machine learning methods," *Procedia Computer Science*, vol. 192, pp. 2893–2902, 2021.

9 Chorąś, M., Demestichas, K., Giełczyk, A., Álvaro Herrero, Ksieniewicz, P., Remoundou, K., Urda, D., and Woźniak, M., "Advanced machine learning techniques for fake news (online disinformation) detection: A systematic mapping study," *Applied Soft Computing*, vol. 101, p. 107050, 2021.

10 Not the onion. [Online]. Available: https://www.reddit.com/r/nottheonion/

11 Reddit. The onion. [Online]. Available: https://www.reddit.com/r/TheOnion/

12 World news. [Online]. Available: https://www.reddit.com/r/worldnews/

13 Nir, F., Dan, G., and Moises, G., "Bayesian network classifiers," *Procedia Computer Science*, vol. 29, no. 1573-0565, pp. 131–163, 1997.