e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

DETECTION OF FAKE NEWS (USING MACHINE LEARNING)

Omkar S Bhosale[*1], Prajwal K Sarnobat[*2], Rajat R Shahapure[*3], Pratik S Revanna[*4], Pallavi R Desai[*5]

[*1,2,3,4]Dr.J.J.M.C.O.E Shivaji University,Kolhapur Jaysingpur,India.

[*5]Assistant Professor, Dept. Of Information Technology Dr. J.J. Magdum College Of Engineering Jaysingpur, India.

DOI : https://www.doi.org/10.56726/IRJMETS41309

ABSTRACT

Fake news is pervasive on social media and in other forms of media, and it is a serious worry since it has the potential to devastate society and the country. On finding it, a lot of research has already been done. This analyses the research on fake news detection and investigates the best traditional machine learning models in order to build a model of a product with supervised machine learning algorithm that can categories fake news as genuine or false. Features will be extracted as a result of this approach. As this library includes practical methods like Count Vectorizer and Tiff Vectorizer, we suggest utilizing Python to carry out tokenization and feature extraction of text data.

I. INTRODUCTION

The majority of information that is considered to be fake news comes from reports that are fabricated and spread on purpose with the intention of deceiving and hurting Internet users. Clickbait, which entices visitors to click links with appealing headlines or designs in order to enhance ad income, readership, or profit, is a prevalent practise for social media and news websites. As a result, some rumours or fake information are created and disseminated online, leading to other internet users believing and spreading these rumours or incorrect information. When people get false information, it may cause them to hold false beliefs and take false actions. It takes a lot of effort to spot fake news when it initially surfaces for negative motives, whether they be political, economic, or social.

Most of the information that could be fake news comes from stories deliberately produced and disseminated to mislead and harm internet users. Social media and news sites often use clickbait to entice users to click on links with catchy titles or designs in order to increase ad revenue, readership, or profits. As a result, some misinformation or rumors are generated and spread on the internet, causing other internet users to believe and spread the lies themselves. When people receive incorrect information, it can lead to erroneous beliefs and behaviors. Detecting where fake news first appears requires a lot of attention for the wrong reasons, whether political, economic or social. Machine learning and other techniques were widely used to identify 4,444 fake news items. In the context of "fake news", any information produced with the intention of to contradict the facts is considered false and therefore reported as false. Fake news on the Internet has two main characteristics: first, it is not true; second, they are not credible. Fake news in mainstream media refers to publications such as newspapers, radio and television. Be careful when trying to determine the legitimacy of despite receiving news from different sources, readers tend to believe what they believe to be true.

In today's technologically advanced world there are many ways to tell if something is true, but fake news is always popular because there are many ways to tell. English has used various methods to check if something on the internet is true, but the language community has fallen far behind. This is why this study proposes to model features that contribute to the dissemination of fake news. Before building a model that can solve the problem of the spread of fake news, it is essential to understand the problems that arise from it. More specifically, it is important to understand how Internet users are misled and harmed by fake news. Typically, this involves spreading disinformation on social media platforms such as Facebook, Twitter and Instagram. Spreading fake news on social media is easy by simply creating an account designed to share fake news and keeping it active

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

II. METHODOLOGY

Data preprocessing

This is the process of preparing raw data and fitting it to machine learning models. This is the crucial first step in creating a machine learning model. When building machine learning projects, we don't always come across clean and formatted data. When preprocessing the machine learning data, we split the dataset into training and test sets. This is one of the key steps in data preprocessing because by doing so we can improve the performance of our machine learning models. Suppose we train our machine learning model through a dataset times and test it through different datasets. It then poses a problem for our model to understand the correlations between the models. If we train our model well, its training precision is also high, but we feed it with a new set of data, it will degrade the performance. So we are still trying to build a machine learning model that performs well on both the training set and the test data set. Data should be preprocessed for best results.

This involves removing URLs, radicals, punctuation, and stop words. At this time, the natural language processing method is used. Natural language processing allows us to extract key information from data. The pre-processing mainly takes place at three levels. The first part is static and suitable for machine learning classifiers. We studied and trained the model using 4 different classifiers and selected the best classifier for the final run of the. The second part is dynamic, it takes user keywords/texts and finds the true probabilities of online news. The third part provides the authenticity of the URL entered by the user. Python has a large number of libraries and extensions, which can be easily used in machine learning. Python has a large library with such built-in classification techniques.

Data Preprocessing Steps

Data Quality Assessment:

Carefully review your data for overall quality, relevance, and consistency with your program. There are many data anomalies and inherent problems in nearly all datasets.

Data type mismatch:

When you collect data from many different sources, it can appear in different formats. Although the end goal of this whole process is to reformat the data for the machine, you should always start with data in a similar format. For example, If part of your analysis looks at household incomes from multiple countries, you will need to convert each amount of income into a single currency.

Mixed data values:

It's possible that different sources use different descriptors for a characteristic - for example, male or male. These value descriptors must all be unified.

Data outliers: Outliers can have a significant impact on the results of data analysis. For example, if is the average test score for a class and a student has not answered any of the questions, their 0% could greatly skew the results.

Missing Data:

Check for missing data fields, spaces. In the text, or unanswered survey questions. This may be due to human error or incomplete data. To deal with missing data, you will need to perform data cleansing.

Data Cleansing: Data cleansing is the process of adding missing data from a dataset and correcting, recovering, or deleting incorrect or irrelevant data. Appointment cleanup is the most important step in preprocessing because it ensures that your data is ready for your downstream needs. Data cleansing will correct any data inconsistencies found in your data quality assessment. Depending on the type of data you are using, you should run the data with possible cleaners.

Data Cleansing Techniques

Your choice of data cleansing techniques depends on many factors. First, what type of data are you dealing with ? Are hey numeric or string? Unless you're dealing with too few values, don't count on cleaning your data with just one technique either. You may need to use more than one technique for best results. The more types of data a has to deal with, the more cleanup techniques must be used. The methods, which we will discuss, are some of the most common data cleansing methods in the field of data mining. Thanks to them, you will be able to understand

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

how to clean the data before starting the scanning process. Knowing all these methods will help you fix errors and get rid of unnecessary data.

Removing Irrelevant Values

The most basic data cleansing method in data mining is to remove irrelevant values. The first and most important thing to do is to remove unnecessary data from your system. Any useless or irrelevant data is something you don't want. This probably doesn't fit the context of your question. This is usually a data type that is not suitable for the issue you are trying to analyze Maybe you just need to measure the average age of your salespeople. Then their email address will not be needed. Another example is you could look at how many customers were contacted in a month. In this case, you do not need the data of the people with whom you have been in contact in the last month. However, before deleting specific data, make sure it's not relevant because you might need it later to check the associated values (to check consistency).If you can get a second opinion from a more experienced specialist before deleting your data, do so. When using data cleansing algorithms, be sure to only remove information that is not relevant to your data set. You don't want to delete a value and then regret the decision. But once you are sure the data is no longer relevant, delete it. Removing irrelevant data will make your data set more manageable and efficient. This is why data cleansing in Data Mining is so important.

Remove duplicate values

Duplicate values are similar to useless values - you don't need them. They just increase the amount of data you have and waste your time. Duplicate values are the most common weak data type in the data set. You can get rid of it with a simple search. Duplicate values can appear in your system for various reasons. Perhaps you combined data from multiple sources. Or the person who submitted the data mistakenly repeated a value.

Some users clicked "enter" twice when filling in the online form. You should remove duplicates as soon as you find them. The process of removing duplicate data is called deduplication, and it is one of the most important data cleansing methods in data mining

Avoid spelling mistakes (and similar mistakes)

Spelling mistakes are the result of human error and can be found everywhere. You can fix typos with over algorithms and techniques. You can map the values and convert them to the correct spelling of. The typo needs to be corrected because the model handles different values of differently. Strings are highly dependent on their spelling and capitalization. "George" is not the same as "george", even though they are spelled the same. Similarly, 'Mike' and 'Mice' are different from each other even though they have the same character count. You will need to find these typos and correct them accordingly. Another error similar to a typo is the size of the string. You may need to fill them to keep them in the same format. For example, your data set may require you to have only 5 digits . So if you have a value with only four digits, like "3994", you can add leading zeros to increase its number of digits. Its value will remain the same as '03994' , but your data will remain consistent. Another error with strings is white spaces. Be sure to remove them from the string for consistency.

Beware of missing values

Will always have missing data. You cannot avoid it. So you need to know how handles it to keep your data clean and error free. You may have too many missing values for a particular column in your dataset. In this case, it would be wise to drop the entire column as it doesn't have enough data to work with. Note: You should not ignore missing values. Ignoring missing values can be a major mistake because it will pollute your data and you won't get accurate results. There are several ways to handle missing values.

Imputation of missing values

You can impute missing values, which means that approximate values are assumed. You can use linear regression or the median to account for missing values. However, this method has implications because you can't tell if it's a real value. Another way to impute missing values is to replicate data from similar data sets. This method is called "hot-deck imputation". You add the value to the current record as, taking into account some constraints such as data type and range.

Execution

Using the code, the data is finally converted into the selected format. Data is pulled from sources ranging from structured to streaming, from telemetry to log files. Then perform transformations such as aggregation,

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

conversion to format, or merge on the data according to the plan of the mapping step. The transformed data is then sent to the target system, which may be a data set or a data warehouse.

Data Transformation Techniques

Aggregation: Data aggregation combines all of your data into a unified format.

Normalization:

Normalization scales your data within a regularized range so you can make more accurate comparisons. For example, if you want to compare the employee losses or gains of companies (some with only a dozen employees, others with more than 200 employees), you would scale them to in a specified range, such as -1.0 to 1.0 or 0.0 to 1.0.

Feature Selection: Feature Selection is the process of deciding which variables (features, attributes, categories, etc.) are most important to your analysis. These features will be used to train the ML model. It is important to remember that the more features you choose to use, the longer the training process will take and sometimes the less accurate your results will be, as some features may overlap with features or be less present in the data.

Discreditization: Discreditization collects data at shorter intervals. This is somewhat similar to clustering, but usually occurs after data cleaning. For example, when calculating an average of daily workouts, you can aggregate the data to over periods of 0-15 minutes, 15-30, etc., rather than using exact minutes and seconds.

III. ALGORITHM

To build the Logistic classification algorithm to classify the news as fake or real,

Logistic Regression:

Logistic regression is one of the most popular machine learning algorithms, and is a supervised learning technique. It is used to predict a categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the result must be categorical or discrete. It can be Yes or No, 0 or 1, True or False, etc. But instead of giving the exact value of 0 and 1, gives the probability that is between 0 and 1. Logistic regression is very similar to linear regression, except it is used differently. Linear regression is used to solve regression problems, while logistic regression is used to solve classification problems. In Logistic Regression, instead of fitting a regression line, we fit a logistic function of "sigmoid" shape, predicts two maxima (0 or 1).

The logistic function curve represents probabilities such as whether cells are cancerous, whether the mice are obese (based on their body weight), etc. Logistic regression is an important machine learning algorithm because of its ability to provide probabilities and classify new data using continuous and discrete data sets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

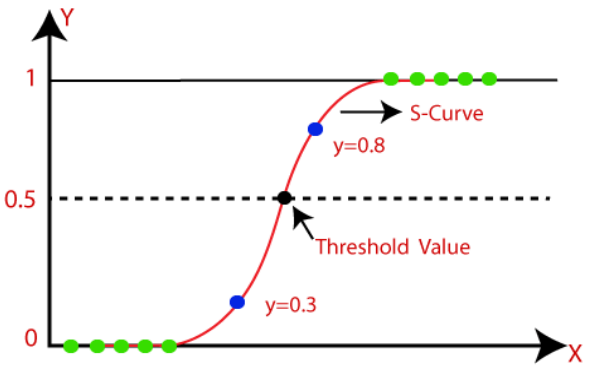


Fig.1 Logistic Regression

Logistic Regression Equation

The Logistic regression equation can be obtained from the Linear Regression equation.

The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

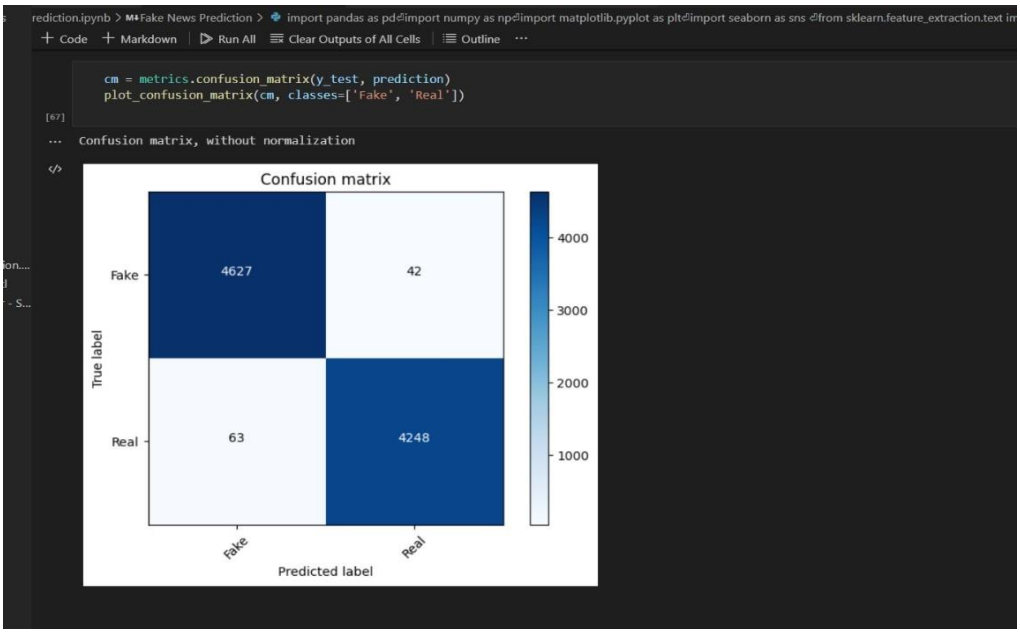
But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

IV. RESULT

Implementation was done using the above algorithms with Vector features- Count Vectors and Tf-Idf vectors at Word level and Ngramlevel. Accuracy was noted for all models. We used K-fold cross validation technique to improve the effectiveness of the models. A. Dataset split using K-fold cross validation This cross-validation technique was used for splitting the dataset randomly into k-folds. (k-1) folds were used for building the model while kth fold was used to check the effectiveness of the model. This was repeated until each of the k-folds served as the test set. I used 3- fold cross validation for this experiment where 67% of the data is used for training the model and remaining 33% for testing. B. Confusion Matrices for Static System After applying various extracted features (Bag-of-words, Tf-Idf. N-grams) on three different classifiers (Naïve bayes, Logistic Regression and Random Forest), their confusion matrix showing actual set and predicted sets.

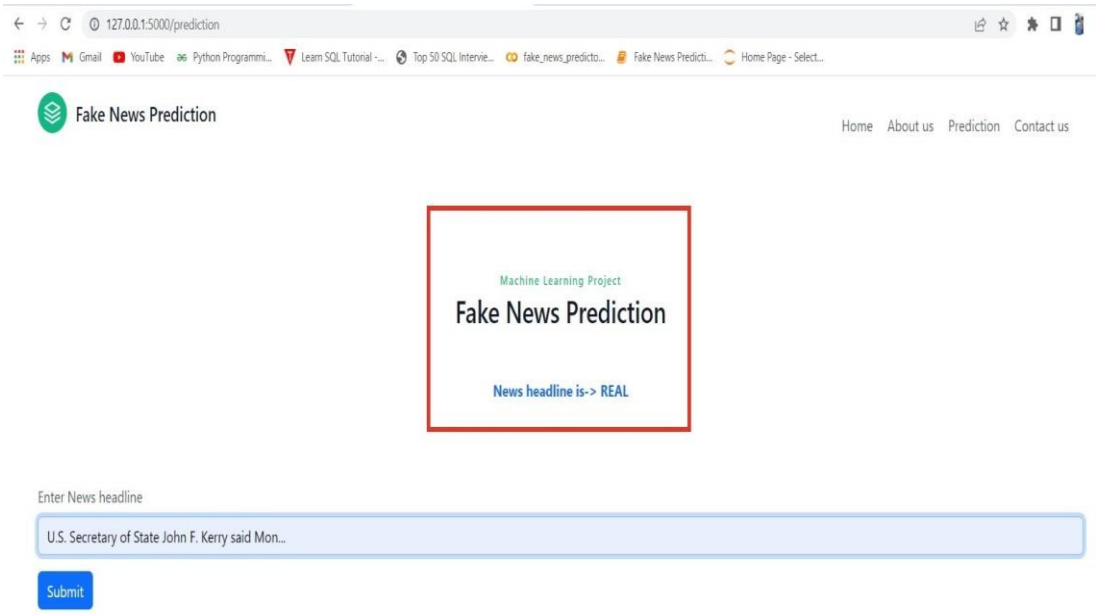


In the above FIG-1 plotting the fake and real news which has more accurate, here there is more fake news as compared to true news.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

```
[64]
X_train,X_test,y_train,y_test = train_test_split(data['text'], data.target, test_size=0.2, random_state=42)

dct = dict()
from sklearn.linear_model import LogisticRegression

pipe = Pipeline([('vect', CountVectorizer()),
                 ('tfidf', TfidfTransformer()),
                 ('model', LogisticRegression())])

# Fitting the model
model = pipe.fit(X_train, y_train)

# Accuracy
prediction = model.predict(X_test)
print("accuracy: {}%".format(round(accuracy_score(y_test, prediction)*100,2)))
dct['Logistic Regression'] = round(accuracy_score(y_test, prediction)*100,2)
[66]
... accuracy: 98.83%
```

From the above FIG 2 dataset is imported for getting accuracy of fake news and as a result there is 98.83% of accuracy of fake news.



From the above FIG-3 Final Output of the Project it shows News is Fake or Real.

V. CONCLUSION

It is important that we have a mechanism to detect fake news, or at least realize that everything we read on social media and other sites is not true. Several projects have targeted fake news since the 2016 US presidential election. A popular method we use to measure machine learning data is to compare a given URL to the OpenSources.co dataset. Open source. The co dataset is a list of reliable and unreliable news sites maintained by researchers at Merrimack College. A popular Chrome plugin, B.S. Detector, uses only the OpenSources.co dataset to determine if a given URL is fake news. In this project, the technique used is superior to the aforementioned method because it uses machine learning to perform statistical analysis on a given news article, while does not rely on a "blacklist" of articles press release from OpenSources.co. As with all blacklists, fake news sites that the Merrimack Academy team has never seen before will not be correctly identified as fake news by the B.S. detector.

So this project will make people more informed. This will help start a new revolution against one of the most common dangers, fake news. This will serve the root and branch of the same eradication.

VI. REFERENCES

[1] M. Granik and V. Mesyura, "False news recognition using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.

[2] H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam recognition in Twitter," 2018 10th International Conference on Communication Systems Bengaluru, 2018, pp. 380-383

[3] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online False News Recognition Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2018, pp. 272- 279.

[4] GoelAnant, Nabanita De, Qinglin Chen, and MarkCraft.

[5] ”Anantdgoel/HackPrincetonF16.” GitHub. N.p., 30 Jan. 2017. Web. 06 Feb. 2017.

[6] https://github.com/anantdgoel/HackPrincetonF16.

[7] Cutler A, Zhao G. Pert-perfect random tree ensembles[J]. Computing Science and Statistics, 2001, 33: 490-497.

[8] M. Granik and V. Mesyura, "False news recognition using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.

[9] H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam recognition in Twitter," 2018 10th International Conference on Communication Systems Bengaluru, 2018, pp. 380-383

[10] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin,M. DiPierro and L. de Alfaro,"Automatic Online False News Recognition Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2018, pp. 272- 279.