JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

FAKE NEWS DETECTION USING MACHINE LEARNING

[1]Akashdeep Ghosh, [2]Priyanka Gupta,[3] Arijit Paul,[4] Debayan Dey, [5] Abhijit Ghosh

[1]Electronics and Communication Engineering,
[1]Narula institute of Technology, Kolkata, India

Abstract: The fake information on social sites and diverse different media is spreading and it is a matter of significant difficulty because of its capacity to cause a whole lot of social and country wide harm with negative impacts. A lot of studies are already targeted on detecting it. This paper makes an evaluation of the studies associated with fake information detection and explores the conventional system studying fashions to select the best, as a way to create a version of a product with supervised machine learning algorithm, which could classify fake information as real or false, with the aid of using the use of equipment like python scikit-learn, stemming for textual evaluation. This technique will bring about feature extraction and vectorization; we suggest the use of scikit-learn library to execute tokenization and feature extraction of textual content data, due to the fact this library includes beneficial equipment like Count Vectorizer and Tiff Vectorizer. Then, we can carry out feature selection methods, to test and choose the perfect fit features to acquire the best precision, in keeping with confusion matrix results.

Keywords– **Fake Information, Machine Learning, Count Vectorizer.**

I. INTRODUCTION

The Internet has come to be obligatory in our existence. It is now very smooth to use the Internet than it was before. Nowadays youngsters get most of the news through internet rather from newspaper. The Internet provides many possibilities for us, we are able to look for anything at the net to clean our doubt and for studies etc. As more humans are connecting to the Internet, they get most of the data through it. In a country like India where the internet is very cheap, a lot of human beings are having access to information through their virtual devices. But with regards to information publishing, it creates a lot of issues. If it's about the information, the net performs a essential role due to the fact that through the net, the information spreads very fast. It's because of faux information, many Innocents are put into risk. Fake information can be made deliberately or by chance to offer harms to a person or a group for any purpose, such as for political issues, for spiritual purposes and so on.

This paper proposes a method to create a model with the intention to check whether an information is genuine or faux, primarily based on its words, phrases, sources and titles, by making use of supervised machine learning algorithms on an annotated (labelled) dataset, which are manually categorized and guaranteed. Then, feature selection techniques are carried out to test and pick the high-quality match functions to achieve the very best precision. We advise to create the version the usage of extraordinary classification algorithms. The product version will check the unseen data, the effects could be plotted, and accordingly, the product could be a version that detects and classifies faux articles and may be used and integrated with any system for future purpose.

II. LITERATURE REVIEW

Automatic faux information detection has already been studied for a few years. Rubin, et.al in [1] gave a hybrid method which mixes the linguistic features of a language with the network evaluation approach. This approach continually might not appropriate because the network statistics can be constrained or now no longer available. In [2] as mentioned via way of means of Rubin, et. al. has analyzed rhetorical systems and the relation among the diverse different systems of faux and trustworthy information sample from NPR's "Bluff the Listener". They have carried out clustering to reap 63% accuracy. In [3], Mihalcea and Strapparava confirmed that via way of means of the use of deep learning it's far viable to distinguish among fake and real statistics to a few degrees.

Himank Gupta et. al. [4] gave a framework primarily based totally on unique machine learning technique that offers with numerous issues together with accuracy shortage, time lag (BotMaker) and excessive processing time to handle lots of tweets in 1 sec. Firstly, they have amassed 400,000 tweets from HSpam14 dataset. Then they similarly signify the 150,000 junk mail tweets and 250,000

non- junk mail tweets. They additionally derived some light-weight features together with the Top-30 phrases that are imparting maximum information gain from Bag-of- Words model. 4. They had been capable of reap an accuracy of 91.65% and handed the prevailing solution by approximately18%.

III.METHODOLOGY

The methodology section outlines the plan and method that how the study is conducted. This includes Universe of the study, sample of the study, Data and Sources of Data, study's variables and analytical framework. The details are as follows:

i) Data collection-

Data collection is the system of accumulating and measuring statistics from endless unique sources. In this project we have gathered the raw data from Kaggle. This process is very first step of machine learning.

The columns which our dataset contains are as follows: id, title, author, text, label.

ii) Data pre-processing-

Data pre-processing is the technique of reworking uncooked facts into a comprehensible format. It is likewise a critical step in facts mining as we can't paintings with uncooked facts. The great of the facts have to be checked earlier than making use of system mastering or facts mining algorithms.

At first, we check that whether any missing value is present in any of the columns of our dataset. The missing values if any, are then replaced by null string (as our project contains mostly text values). There are very few missing values so rather than dropping In the dataset, we are replacing those missing values.

After this the we merge the author and title column for prediction, as those columns are the most fundamental for predicting that whether a news is false or true. The label column is then dropped from the prototypical dataset and then stored into another variable for future purpose.

The dataset undergoes a process named stemming (a function that we create) , this process reduces a word to its root word. This step is very important in the project because we need reduce the words as much as possible to have better performance at our model.

Computers cannot understand text so we had to convert our dataset into meaningful numerical values to make the system understand it. This process is done by the function TfidfVectorizer.

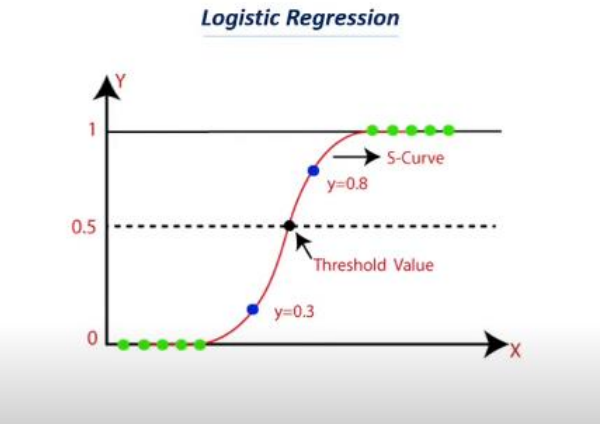
iii) Splitting Dataset-

On the dataset the splitting is done after pre-processing. The dataset is divided into train and test (80%-train and 20%-test) . Through the train part we train our dataset by feeding it to different algorithms and making it suitable to predict. The test part is used to check whether the our trained dataset is able to predict correctly or not.

Applying Machine Learning Algorithm:

In the project we applying Logistic Regression model to our dataset because it is a binary classification project and for binary classification problems, Logistic Regression is the best approach.

In statistics, the logistic model is a statistical model that models the probability of one event taking place by having the log-odds for the event be a linear combination of one or more independent variables.

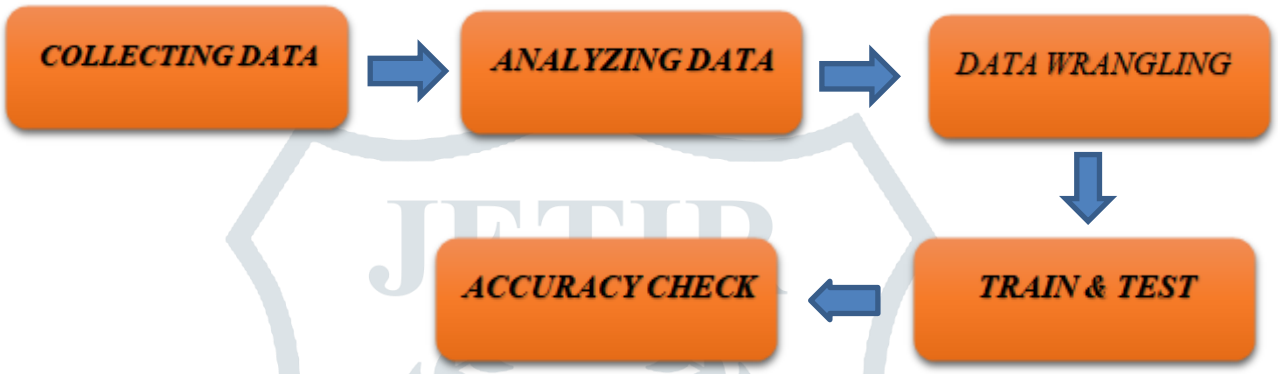


Y=1/(1+e^-z) [Sigmoid function]
Z=w.X+b

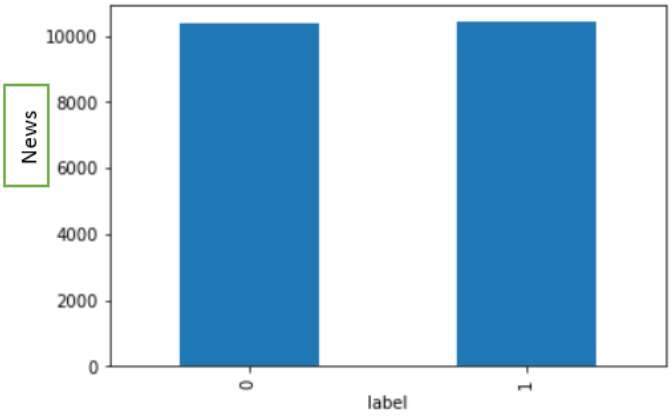
X= Input Features
Y=Prediction Probability
w=weights
b=biases

The model is trained and then fed to the test data, if the accuracy is satisfactory then the model is applied on new data ,so it can do prediction on it.

IV. WORKFLOW:



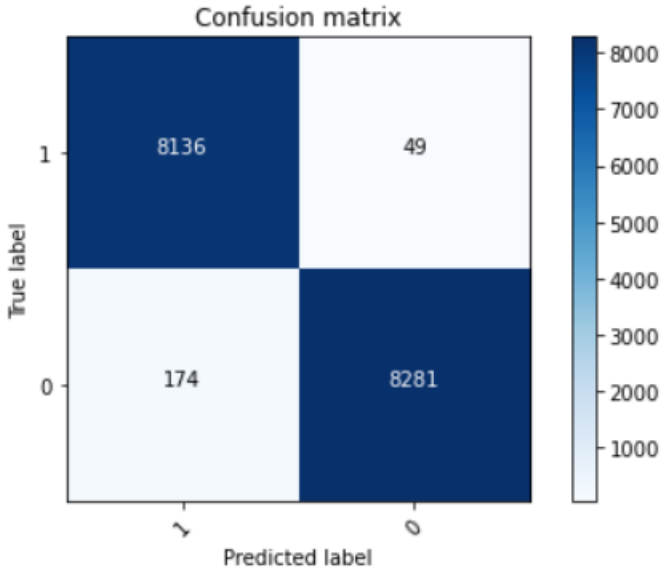
V. RESULT-



HERE 1 MEANS FALSE & 0 MEANS TRUE

The Dataset which we downloaded from Kaggle has both genuine and fake news. So we counted how many true and how many false news are there in the dataset. We have seen there are 10387 genuine news and 10413 fake news. From counting, we can see it is a good dataset because it has almost half true and half fake news.

After cleaning we trained that dataset with Logistic regression which predicts the news whether it is true or false. Then we checked the accuracy score of the training data, it has given 98.65% accuracy and 97.90% accuracy on the test data.



After Making Predictive system we created confusion matrix and labelled TF (True Positive), TN (True Negative), FP (False Positive), FN (False Negative). [Text Wrapping Break]

VI. CONCLUSION-

In this paper, we present fake news detection model. We have used a new publicly available fake news dataset(train.csv). The classification of fake news from the real news is very crucial task nowadays. It is becoming an imminent threat in some situation to not able to discern real and fake news. Our best performing models achieved accuracies that are comparable to the human ability to spot fake content. In future, we will try to make a live website where we can follow news and can check whether that news is correct or wrong**.[5]**

VII. REFERENCE-

[1]. V. L. Rubin, N. J. Conroy, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

[2]. Rubin, V., Conroy, N. & Chen, Y. (2015)A. Towards News Verification: Deception Detection Methods for News Discourse. Hawaii International Conference on System Sciences.

[3]. R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 309–312, Association for Computational Linguistics, 2009.

[4]/H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383

[5]. Lilapati Waikhom and Rajat Subhra Goswami," Fake news detection mode