Fake News Detection in Python

Nitin Sinha[1], Dr. Devesh Katiyar[2]
Student of MCA[1], Asst. Professor[2]
Department of Computer Science and Information Technology
DSMNR University
Lucknow, Utter Pradesh

I. ABSTRACT

The paper "Fake News Detection" is an application of computer security that is developed by machine learning, "a subclass of Artificial Intelligence". This paper will help to spot the fraud news that is spread by some non-reputable resources. My project aim is to represent a prototype model "fake news detector" based on machine learning so that it could be helpful in detecting misleading news for security purpose on different social platforms like Twitter, Facebook, Whatsapp, etc. in future.

II. KEYWORD

Artificial Intelligence, machine learning, tf-idf vectorizer, Passive Aggressive Classifier.

III. INTRODUCTION

Fake news state the wrong information of anything that could be a person, news, article, etc. These days, faux news has been a major issue in society. Misinformation is spread by spiteful actors for prohibited purposes or illegal motives. Their motives are to achieve the advantages by spreading online advertisements to harm the people. Mostly the news is broadcast by social media.

These days, distributing the mock news has been as a great threat in society for democracy, journalism, and liberty of expression. For example, in 2018, it was examined that "as said by Buzzfeed's report" millions of people shared around 50k wrong news. However Facebook is trying to overcome with these all. Social sites have become a vital platform to spread fake news. These social-sites provide a setting for the general people to express their thoughts about the theme and that's why faux news is becoming popular because it can spread easily. According to the research, it came to know that US election 20116 was heavily impacted by fake news.

Technology like Artificial intelligence provides developer to point out the methodologies to identify the bogus news. However, it's a challenge to classify hoax(fake) news and detect them even now. Yet many researchers and scholars have given a lot of ways to detect fake news by using Different AI technologies. But it will be very helpful for pulling out the feature of the news.

In this, the paper organization is as following: the first part we will discuss about the terminologies used to detect wrong news like Artificial Intelligence(AI), Machine learning(ML), classifiers, etc. After that we'll discuss about the wrong information problem how it is a threat to society. After that I'll represent different approaches done to detect misinformation. Finally, I'll represent my methodology to detect misinformation with programming algorithm and flow chart and finally conclusion.

IV. TERMINOLOGY

A. FAKE NEWS

Fake news/ Hoax news refers untruthful statements that is published under the authenticate person. The information is spread by some malicious persons to mislead the people. They spread the news by social media, magazines, news-papers, etc. It has different synonyms as rumors, misleading information, hoax news, and so on.

B. ARTIFICIAL INTELLIGENCE

In uncomplicated words "A computer intelligence that can copy human behavior" is called artificial intelligence. It's a wide branch in computer science related with developing such smart mechanisms that can take decisions like a human. It has vital applications like speech recognition, face detection, spam filter, fraud detection, etc. Mostly it is concerned with robotic development.

Example:- Manav(Humanoid Robot) and Lakshmi.

C. MACHINE LEARNING

Machine learning is related to scientific study that makes computers able to learn. With the help of this, a computer can perform the task without human interference. It is done in three distinct ways: Supervised learning method(knowledge based), unsupervised learning and reinforcement learning. The learning process is done by observation, instructions or future based examples that are fed in system.

D. CLASSIFICATION

In machine learning, classification is a supervised learning method in which the computer program learns from the input data and uses this learning method to classify new statement. There are following classifiers used for learning as linear classification, nearest neighbor, support vector machine, neural network, etc.

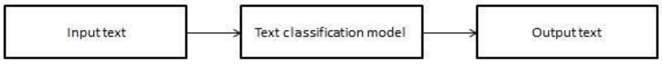
Text classification is a task that assigns an arranged of predefined classes to text. It can be used to classify, structure and categories. A classifier takes text as input, explore its content, and then automatically hand over relevant tags.

How does it work?

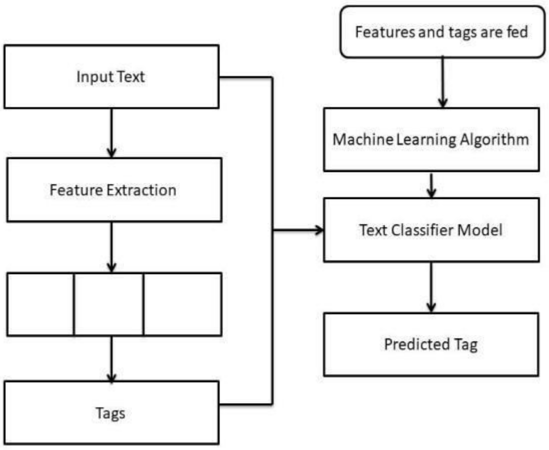
Text classification is done using two methods: manual classification and reflex classification(auto classification).

In the manual approach, text categorization is done using the human annotator decodes but it is expensive. Automatic method classifies the text using machine learning, natural language processing which is faster and cost valuable way.

There are many tactics for reflex text classification as rule-based system, machine learning built system, hybrid system.



Ruled based approach is constructed on manually. It classifies text into prearranged group by using a set of handcrafted linguistic rules. Machine learning based procedure classifies text using observation by learning methods. First classifier extracts the feature of the text and represents it in the mode of numerical representation in way of vector (bag of words approach). Then machine learning system is fed with trained data that depend on pairs of features set and tags to turn out classification model.



After the set is trained, text classification model starts to make true prediction.

E. TF-IDF

TF stands for "Term Frequency" and IDF stands for "Inverse Data Frequency". TF is expanded o find out the frequency or expression in each document. It is the ratio of the number of single word appears many times in the document match up to the total number of words in that file.

$$\mathbf{TF_{ij} = n_{ij} / \sum_k (n_{ij})}$$

IDF computes the weight of valuable words occur in the file in the corpus. The rare word has high IDF value.

$$\mathbf{Idf(w) = log(N/df_t)}$$

Conjoining these two formulas turn up with the TF-IDF score (w) for a word in a document.

$$\mathbf{w_{ij} = tf_{ij} * log(N/df_t)}$$

Where:-

$\mathbf{W_{ij}}$ = quantity of occurrence of i in j

$\mathbf{df_j}$ = amount of document containing i

N = total sum of document

Example:-

Sentence 1: This is a cat.

Sentence 2: Cat is eating fish.

WORD	TF		IDF	TF*IDF	
	A	B		A	B
THIS	¼	0	LOG(2/1)=0.3	0.075	0
IS	¼	¼	LOG(2/2)=0	0	0
A	¼	0	LOG(2/1)=0.3	0.075	0
CAT	¼	¼	LOG(2/2)=0	0	0
EATING	0	¼	LOG(2/1)=0.3	0	0.075
FISH	0	¼	LOG(2/1)=0.3	0	0.075

So as you can see that TF_IDF for the common words is zero and with high significance words are non-zero.

F. PYTHON PROGRAMMING LANGUAGE

This is the trendiest program writing language in these days. It is used to develop desktop application, websites and system scripting. It was developed by G V Rossum in 1991.

FEATURES:

I. It is cross platform supportive programming language.
II. It is easy to learn.
III. Lines of codes are fewer than other computer programming language.
IV. It runs on interpreter system.
V. It provides different IDE like Anaconda Navigator, Pycharm, Jupiter, etc.
VI. It is assembled with advance library packages.
VII. Mostly it is used to develop AI mechanisms.
VIII. Speech recognition, Patterns matching such applications can be developed very easily.
IX. It is object-oriented programming language.

APPLICATION:

I. Web application
II. Scientific development application
III. Network programming
IV. Video games
V. Software development

V. RELATED WORKS

There are lots of approaches done to distinguish the fake news. Discovering the fake news on social media like Facebook, twitter, etc. is up to. For example, Hussein Imotlagh developed a bogus news identifier technique using unsupervised tensor model. It was established on the TF and Spatial relationship in the middle of terms and articles. Guacho developed a semi-supervised model via tensor embedding. It uses spatial contextual figures about news article. Gupta purposed a classifier for estimating tweet credibility from features like words, URL, hash tags, pronouns, etc. This model indicates the relation between news and publisher and interaction of user with news. It takes some simultaneous parameters that come from altered sources of information. Rubin mentioned that there are diverse natures of fake news with different textual indicators. Thus the existing hand crafted features are not only tough but dependent on unambiguous dataset and availability of knowledge to figure apt features.

Ma purposed a re-current neural network (RNN) prototype to sense the fake news that uses linguistic features.

Castillo showed that a little pattern of user response to news article play an important role. With the help of the user interaction, it helps to absorb the features of contents. It is done by measuring the response an article received by the study its propagation. CSI model captures all three faces of fake news those are: suspicious user, few parameters for more accurate classification.

Major platforms as Facebook give priority the trust worthy sources and shut down all the accounts linked with fake information. Some users use tools like NewsGuard, Hoaxy, websites like snopes and politifacts. There has been also different purposed prototype to figure out the fake news using feature extraction as filtering methods like word frequency, mutual information and information gaining by using technique like deep learning, convolution neural network, fuzzy-neurons, etc.

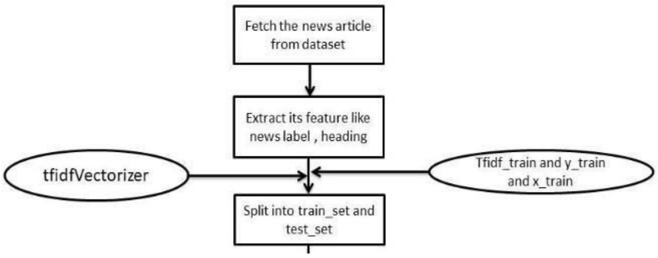
VI. RESEARCH METHODOLOGY

Fake news discovery is most important drawback in our civilization in these days. The people who are unknown with the fake news, they belief them very quick. Different people share different posts, news, articles, etc. that may be real or fake.

Let's understand my research methods by simple algorithm and flowchart but before that you need to get some news article related to previous and current days and make a dataset of news. You can collect news from newspaper, magazines and internet.

After that follow the following instructions:-

I. Install Anaconda Navigator Pycharm.
II. Fetch the articles from the prepared dataset.
III. Get the labels from the data frame. This will happen by feature extraction methods.
IV. Split the dataset into training and testing sets.
V. Fit and alter the labels on the train_set and then test_set.
VI. Next we'll apply a passive aggressive classifier.
VII. Above classifier will acceptable on tfidf_train and y_train.
VIII. Then I'll predict the above test_set from the tfidfVectorizer and evaluate the accuracy with accuracy_score() from sklearn-matrix.


Fetch the news article from dataset
Extract its feature like news label , heading
tfidfVectorizer
Tfidf_train and y_train and x_train
Split into train_set and test_set

FLOW-CHART

VII. CONCLUSION

My research paper aim was to represent all necessary information related to expose fake/bogus news. All given information is extremely necessary for the development of detection system. Also my represented algorithm can be applied in different programming language so that development can be done easily. In the future, there may be some advance technology to overcome with the bogus news and it will be applied at all social platforms so that no one can harm other peoples with the misleading articles.

VIII. REFERENCES

[1]. Rown Zeller, Ari holtezmen, Hanah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin choi “Defending Against Neural Fake News” cs. CL International Conference on IEEE, 20119

[2]. Frosso Papanastasious, Georgious katsimpras, Georgios paliouras, “Tensor Factorization with Label Information for fake News Detection” cs. si, 2019.

[3]. Jaiawei Zhang, Bowen Dong, Phillip S. Yu “Fake Detector Effective Fake News Detection with Deep Diffusive Neural Network” cs. si 2019

[4]. Xinvy Zhou, Reza Zefarani, “Fake News: A survey of Research Detection Methods and opportunities” cs.CL 20118