**Mini Project Report on**



**FAKE NEWS DETECTION**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name: DISHA JAIN** **University Roll No.:2021878**

***Under the Mentorship of***

**Dr. Neha Tripathi**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**

GEU logo

**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Fake News Detection”** in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Neha Tripathi,** Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name: Disha Jain University Roll no:2021878

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 4-5 |
| Chapter 2 | Literature Survey | 6-7 |
| Chapter 3 | Methodology | 8-10 |
| Chapter 4 | Result and Discussion | 11-12 |
| Chapter 5 | Conclusion and Future Work | 13 |
|  | References | 14 |

**Chapter 1**

**Introduction**

In the following sections, a brief introduction and the problem statement for the work has been included.

* 1. **Introduction**

This outgrowing technical industry is overpowering the brains of many people out there. Everyone is becoming more and more accustomed to the need and usage of technology. For obvious reasons technology has been introduced in this era but it comes with some dark side as well which needs to be discussed with the time. There are many cyber criminals who use this technical for various corruptions . One of those is spreading fake news among their readers. Online forums help users because they can easily access information. But the problem is that it gives cybercriminals an opportunity to spread fake news through these platforms. This issue can be harmful to individuals or society. The popularity of digital technology contributed to the problem of information distortion. It has become increasingly common through fake news. Harmful, offensive or illegal or malicious messages. Content can be misleading, misunderstood and cause social unrest, affect social stability. It is a serious concern to look into it as once the news is out people start believing it without its verifications which can also lead to spreading of hatred, misconceptions and a lot more. In the 2016 US election, people’s opinions and decisions are influenced by fake news (Dewey, 2016). With the rapid development of big data and information sharing technologies, fake news spreads through social media, posing a serious threat to sustainability and community development. In this today’s world full of brainy people , they know the art to make the fake news to be trusted by the people more. They will try their best that their effort of creating false news does not get out of hand and it should solely affect the society. This is a serious issue which needs to be taken into concern as these fake news always have some agenda behind it, either against any political party, some organising community, against any nation or to create panic and anxiety in the society.

This technology also introduces various ways to resolve this issue. There are institutions, such as the Commission and The Crosscheck project, that attempt to address issues that assert that the authors are responsible. However, the feasibility is very limited because it relies on human hand recognition, with millions of records worldwide .Whether removed or displayed minute by minute, this cannot be calculated manually. The solution could be, either by creating a system to provide reliable automated index scoring, or recognition from various publishers, and related media content. So, Machine learning comes as a saviour to this problem of fake news. Machine learning is the subcategory of artificial intelligence which focuses on training the model according to the various algorithms and thus testing it to produce the result. There are various subcategories of machine learning, that is , supervised learning and unsupervised learning. They have various defined algorithms which help in building models. It is a very enhancing field which is coming as a solution to many. Creating fake news models is not that easy as it comes with a lot of complications and case scenarios. But to a large extent machine learning algorithms proved to be the correct solution in detecting the fake news.

This paper proposes the methodology for modelling verification based on its words, phrases, sources and topics, using a supervised machine learning algorithm. The methods are used to test and select the most suitable features to achieve the highest accuracy. According to confusion matrix results. We propose to develop a model using different machine learning algorithms.

* 1. **Problem Statement**

We are provided with a large dataset consisting of over 1000 statements of news articles headlines. Some of them are fake and some of them are true ones. We are required to build a model that is capable of detecting if the statement in the article is fake or true . For this we are using various machine learning algorithms to train the model so we can predict the result. Fake news is a big concern which is no longer hidden from this world. And people are becoming more and more aware of this problem. So, the responsibility increases to build the most precise model for detecting fake news.

**Chapter 2**

**Literature Survey**

In this chapter some of the major existing work in these areas has been reviewed.

Fake news detection is not a completely unknown field. This is the most concerning field which already has the hands-on by some researchers. But since it is not that easy to make a complete precise model to predict fake use . So, the research is ongoing till now. With the constant repetition of artificial intelligence, the researchers developed news data mining and automated and intelligent analysis based on it. However, the current study lacks knowledge and multidisciplinary research to use on how the relevant processes can be interpreted. This paper includes information about the dataset of fake and true news. In recent years, different work has been recorded in this area of fake news detection. Different methods have been recorded by the researchers . Some of them are:

1- Zhang et at. [3] , describes what are the negative impacts of this fake news that is continuously being spreader through this online platform. They also shortlisted the latest techniques which were available at that time.

2- Athira et al. [4] , organised a proper systematic explanation in the usage of artificial intelligence in detecting fake news.

These reviews lack some respect to the theoretical knowledge, new introduced methods and techniques which can be used in this field. Since they have some or more deficiencies in their reviews about fake news detection. So, the search for the precise one is still being continued. Based on the above analysis, in order to take the responsibility to solve these deficiencies by making our model more prone to be believed by people out there. There is a compressed overview of the research situation in this field of fake news detection. The model which is being represented by this review paper will definitely be able to make people more prone to be more aware of this fake news . Some specific contributions are as follows:

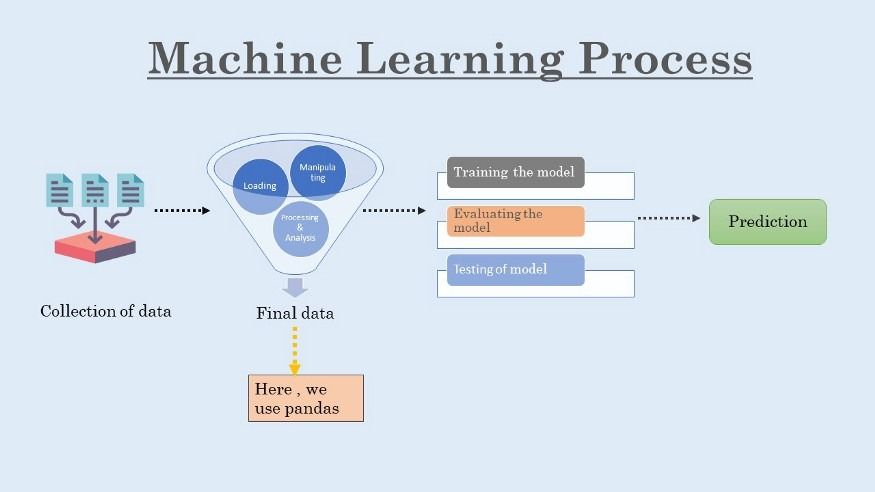
* We have used the latest technologies and methods to make our model more prone to today’s world. We investigate previous research based on fake news detection and put the theoretical knowledge as well in our model, so that it can be understood by the people as well.
* We have made use of machine learning algorithms. The extra point is that we have added multiple classification algorithms of machine learning and on the bases of their accuracy we have trained our model.
* We have also used the confusion matrix while predicting the accuracy of each algorithm which makes it more understandable to another person about the model. Confusion matrices make our model more precise, readable, and understandable.

This paper uses the dataset from Canadian Institute for Cybersecurity as a source to acknowledge the various trends in fake news detection over years. This dataset contains news article headlines from a lot of references. Thus, this model focuses on making the model which can predict the fake or real news of any kind, a model which is more readable and reliable, a model which uses the new technologies aspects and methods to be defined which can overcome the deficiencies which previous researchers lack.

**Chapter 3**

**Methodology**

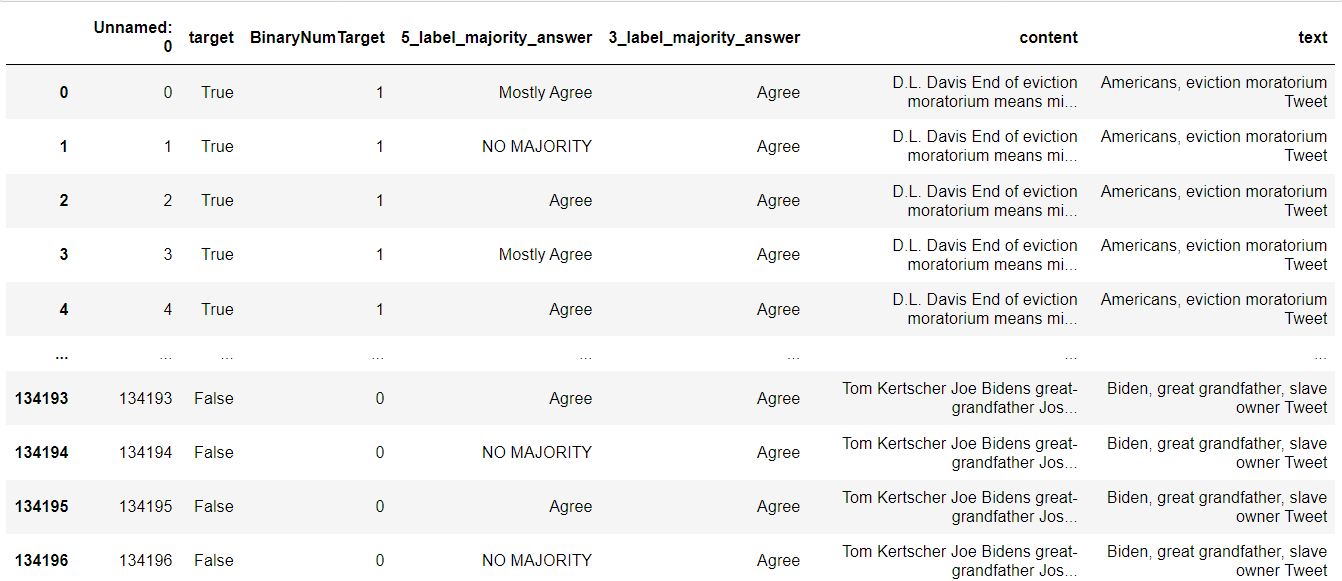
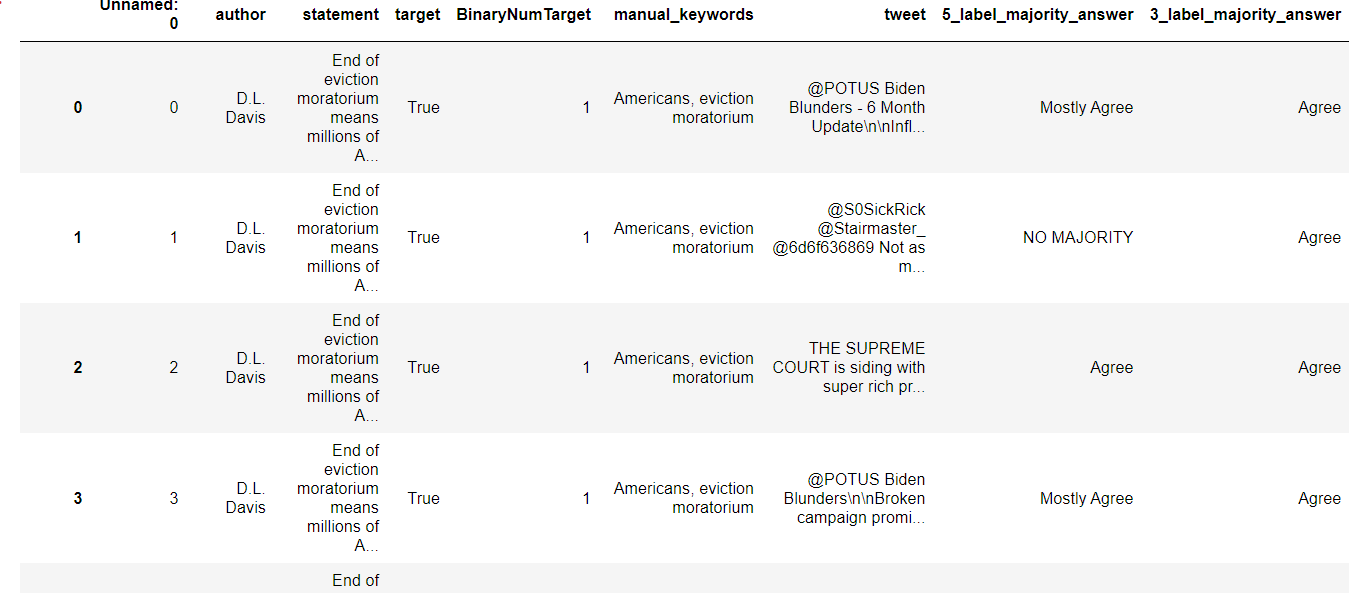
This section presents the methodology which is used in this model to build it from the very start to the end where it can be declared as a model to detect fake news. It contains all the algorithms, methods, figures used to make this model. It contains detailed information about each step from data collecting to data testing.

A systematic process is being followed to build a model which we will be discussing in this section. The first step in this is the dataset collection phase, followed by preprocessing, then implementing selection features and then training and testing the dataset and then finally running the model to predict news and declare it to be true or false. We have taken the use of machine learning to build our model.

**Fig. 3.1 Procedure to build a model**

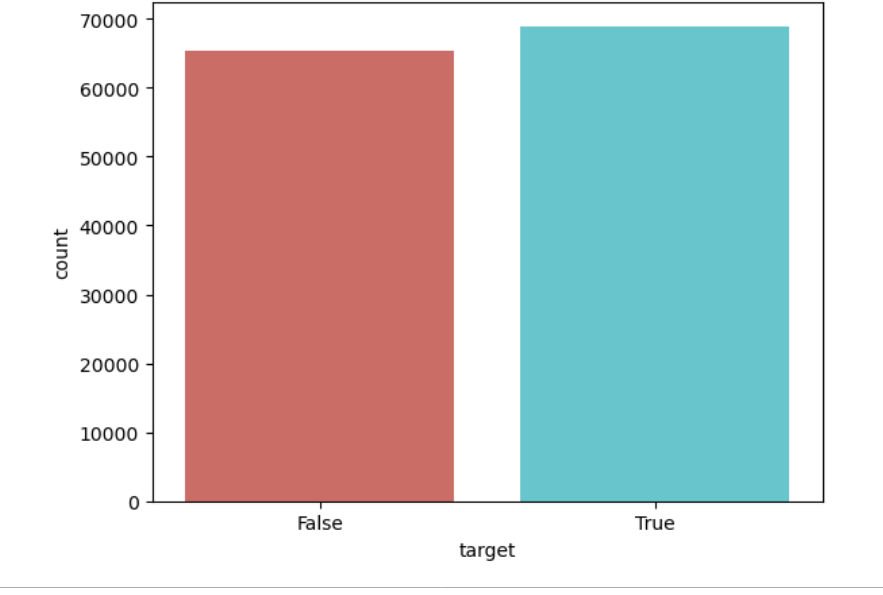
The very first step which we performed is the collection of data from the cic site which provides a dataset having a collection of thousands news article statements containing 0 or 1 to specify whether the news is true or false. Numeric value 1 is used to specify that the news statement is true and 0 is used to specify that the statement is false. After collection of data, we used pandas, a library of python to input the dataset. After that all the work of preprocessing, collecting and mining of data is done by the pandas library only of python.

Data preprocessing is one of the major steps in this which further include the steps to be carried out.



**Fig. 3.2 Data before pre-processing Fig. 3.3 Data after preprocessing**

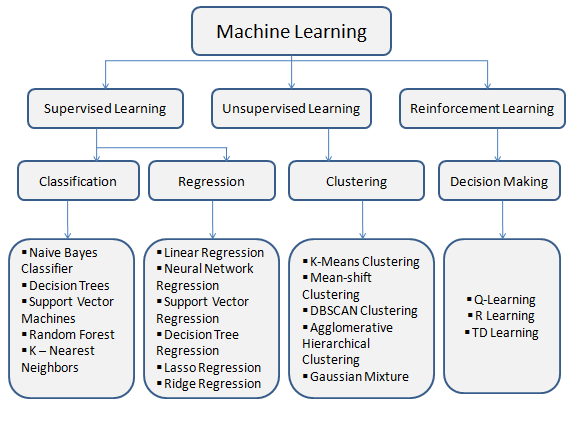
In data preprocessing, we use various techniques for data cleaning where we find the missing or null values, merge some columns and drop some to reduce the complexity in the dataset and make it more reliable and readable. These steps make data more precise and simpler for the machine to understand. The more the data is short and precise, the more it will be better trained by algorithms and then better be tested.



**Fig. 3.4 True and Fake news in dataset**

After data preprocessing, we use NLTK (Natural language toolkit) for importing stopwords and performing the stemming process in which we ask the machine to reduce some unwanted words, it can be some helping verbs, prepositions or anything else. Removing these words help in minimising the content which will be trained further. After this data cleaning procedure, we convert complete text data into the numeric data using the count Vectorizer because the machine can understand the digits only.

Then we use machine learning to train the model. They are categorised into two, that is, supervised and unsupervised. We have used supervised machine learning in classification of the dataset. We have used supervised machine learning because we want a precise end result that is whether the news is fake or not. So, supervised machine learning contains those algorithms which make sure that the model will be trained to that extent till we get the precise result. Further in supervised learning we have used the classification algorithms since our model is about true answers , it can be either true or false. So, for this purpose we used the various classification algorithms to predict its accuracy. To prove this point, we have also included the accuracy rate of linear regression which is a regression supervised algorithm and the result produced from it was definitely not satisfactory. Thus, the classification algorithm proved the one to be used for training the model.



**Fig. 3.5 Machine Learning Algorithms classification**

We used various supervised algorithms to train our model like decision tree, random forest, logistic regression, linear regression and gradient boosting classifier.

Majority of them are classification algorithms but two of them are regression algorithms. But all of them are supervised machine learning algorithms. After training the dataset with these algorithms and then predicting the accuracy based on them, we will decide which algo will be used for final training of the model. After this process of training, finally the testing stage comes where it is being tested whether the news is true or fake. If it gives the correct result, that means we have trained the dataset properly.

In this way, using a dataset we made a model to predict the fake news which is spreading like a virus in our societies.

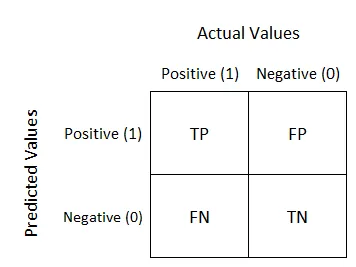
**Chapter 4**

**Result and Discussion**

This section contains the results that are produced after preprocessing and then training of the data. We have used various supervised machine learning algorithms to train our model. We have five algorithms to predict the accuracy and to train our model on the basis of it. These algorithms are:

* Logistic regression
* Decision tree
* Random forest
* Gradient boosting classifier
* Linear regression

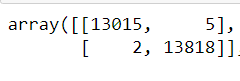
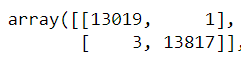
We have used mainly classification and also regression algorithms to prove classification algorithms to be more precise in this case. We have used the accuracy rate and also confusion matrix to select the best algorithm to be used in training.



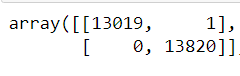
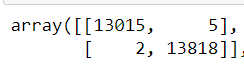
**Fig. 4.1 Confusion matrix**

Confusion matrix is ​​a performance analysis tool in machine learning, representing the accuracy of classification models. It displays the number of true positives, true negatives, false positives, and false negatives.

So, we have displayed the confusion of these four algorithms. Since linear regression is used for algebraic data, it does not create a confusion matrix. Confusion matrices of rest four algorithms as shown as follows:

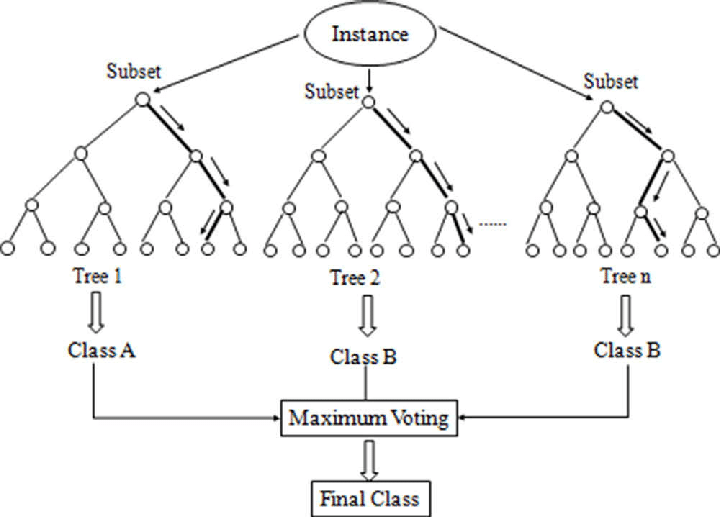
**Fig. 4.2 Logistic regression confusion matrix Fig. 4.3 Decision tree confusion matrix**

**Fig. 4.4 Random Forest confusion matrix Fig. 4.5 Gradient boosting confusion matrix**

As it can clearly be depicted by the confusion matrix that on training the model with random forest supervised machine learning algorithm, it gives 0 false negative result which can be concluded to be the most precise of them all.

Random forest is a known supervised machine learning algorithm that is used to solve the cases related to regression and classification as well. In this case it is used as a classification algorithm in which it has to differentiate whether the news is true or false.



**Fig. 4.6 Random Forest working process**

Random forest thus proves to be a better algorithm since it makes it easy to visualise the data and is a further classified version of decision tree. So, it is a more improved one. It works according to the voting procedure, this works for the majority. Thus, with an accuracy score of 99.99% random forest algorithm proved to be the best to train our model.

Thus, this classification supervised machine learning algorithms proved that they can solve this issue of fake news and make such models to predict them.

**Chapter 5**

**Conclusion and Future Work**

The Internet is a thing which is not under the control of anyone out there. It comes with a vast area of network which connects this world. It is very easy today to either save someone from a network or to destroy someone with the same network. Spreading news or we can say making people aware of what is happening around the world, this all work takes a few minutes to be uploaded on the internet. So, it is very important to verify whether the news that is being spreaded is correct or if it is someone’s purpose to destroy someone. For this supervised machine learning proved to be of great help to make a model which can predict.

We have seen that random forest proved to be the one algorithm which helped in building a model which we needed. Confusion matrix was used in making our predictions into assurance. Making a correct, accurate, precise, reliable, human understandable fake news detection model is not that easy. Because this is a real-life problem which needs a lot of considerations and scenarios to be considered to get a model whose chances of being failed is less than 1%. Till the data, not all the technologies and methods are used in a proper way to make this model. Some used modern technologies but lacked precision, some used only specified and defined ones and lacked the probabilities.

In this model, we have tried our best to make our model more precise, understandable and easy to read. We have collected a dataset from the cic site providing most searched information in it. It contains 50% fake and 50% true dataset. Dataset is very precise and clean with no missing values. So, there are less chances that it may lack accuracy by training with this dataset. And then we have used the advanced method of machine learning consisting of supervised machine learning algorithms to train the model. We have used both classification and regression algorithms to check accuracy. We have used matrices and graphs to make our model more readable to the readers. This model can be used to predict fake news.

In this field, a lot more is left to be analysed and be looked into. Regularly emerging technologies and methods can be searched for. Searching into unsupervised machine learning algorithms and considering them can be a point that can be taken in the future. Using those methods which will help in the sustainable development of the society will be considered. Other python libraries can also be used for graphics, statistics to make the model more readable. So, this topic does not end here, it has a wide scope to be followed in the coming future.

**References**

[1] Alim Al Ayub Ahmed (alim@jju.edu.cn) School of Accounting, Jinjiang University, Jinjiang, Jiangxi, China.

[2] Z. Khanam1, B. N. Alwasel1, H. Sirafi1 and M. Rashid, College of Computer and Information Sciences, Saudi Electronic University, Dammam, KSA, School of Computer Science and Technology, Lovely Professional University, 1999, Jalandhar, India

[3] Zhang, X.; Ghorbani, a. Observing online fake news: Identity, analysis, and discussion. Inf. approach. Management ,2020, 57, 102025. [CrossRef]

[4]Athira, A.B.; Kumar, S.M.; Chakko, A.M. A systematic review of semantic AI was used to detect false positives. Eng. Request. Artificial Intelligent. 2023, 122, 106087.

[5] State Key Laboratory of Media Convergence and Communication, Communication University China, 1999.Beijing 100024, China; 311909011312@home.hpu.edu.cn (H.J.); shenhao@cuc.edu.cn (H.S.). chengnanchang@cuc.edu.cn (NC).