Counterfeit News Detection Using Machine Learning

SHANI P.R
Lecturer Department of Computer Science
FMKMCC Madikeri

Abstract:- World is advancing rapidly. Doubtlessly we have different advantages of this Digital world anyway it has its impediments moreover. There are different issues in this cutting-edge world. One of them is fake data. Someone can easily spread fake news. Fake news is spread to hurt the remaining of an individual or an affiliation. Fake news is counterfeit information that is formed and conveyed by dishonest person. Clients are uninformed that the information that they got is deluding information. Using Machine learning that can orchestrate whether the news is substantial or deceiving through setting up the model. There are different web based stages where the individual can spread the fake news. This consolidates Twitter, face book, Instagram, Whatsapp, etc.

ML is the piece of man-made awareness that helpers in making the structures that can learn and perform different exercises. Simulated learning computations will recognize the fake news thus at whatever point they have arranged. A collection of machine learning computations are available that consolidate the controlled computer based intelligence estimations like Decision Tree, Random forest , Stochastic gradient Descent, K Nearest Neighbor. As a rule simulated intelligence estimations are used for assumption reason or to perceive something hidden away.

General Terms:- Counterfeit, Algorithms, Datasets, Patterns, Graphs, Fake news, Real News

Keywords:- Machine Learning, Sentimental Analysis, Social Media, Decision Tree, Random Forest , Stochastic Gradient Descent, K Nearest Neighbor, Cross Validation.

I. INTRODUCTION

Current life has become very steady and individuals need to thank the tremendous commitment of the web innovation for transmission and data sharing. This is a development in mankind's set of experiences and yet it unfocused the line between evident media and malevolently manufactured media. Today anybody can distribute content tenable or not that can be consumed by the Internet. Tragically phony news collects a lot of a consideration over the web particularly via online entertainment.

World is evolving quickly. Most likely we have various benefits of this computerized world however it has its detriments also. There are various issues in this advanced world. One of them is phony information. Somebody can undoubtedly get out counterfeit word. Counterfeit word is gotten out to hurt the standing of an individual or an association. Counterfeit news is bogus data that is composed and distributed by untrustworthy individual. Clients are ignorant that the data that they got is misleading data. Utilizing AI that can arrange whether the news is valid or misleading via preparing the model. There are different internet based stages where the individual can get out the phony word. This incorporates Twitter, face book, Instagram, Whatsapp and so forth.

Machine Learning is the piece of man-made consciousness that aides in creating the frameworks that can learn and perform various activities. . AI calculations will identify the phony news consequently whenever they have prepared. An assortment of AI calculations are accessible that incorporate the regulated AI calculations like Choice Tree, Irregular Backwoods, Stochastic Inclination Plunge, K Closest Neighbor. More often than not AI calculations are utilized for expectation reason or to recognize something stowed away.

Online stages are useful for the clients since they can undoubtedly get to news. Be that as it may, the issue is this offers the chance to the digital crooks to get out counterfeit word through these stages. This news can be demonstrated unsafe to an individual or society. Recognizing the phony news is a major test since it's anything but a simple undertaking .In the event that the phony news isn't distinguished early then individuals can spread it to other people and every one individuals will begin trusting it. People, associations, entertainers or ideological groups can be affected through the phony news. Individuals conclusions and their choices are impacted by the phony news.
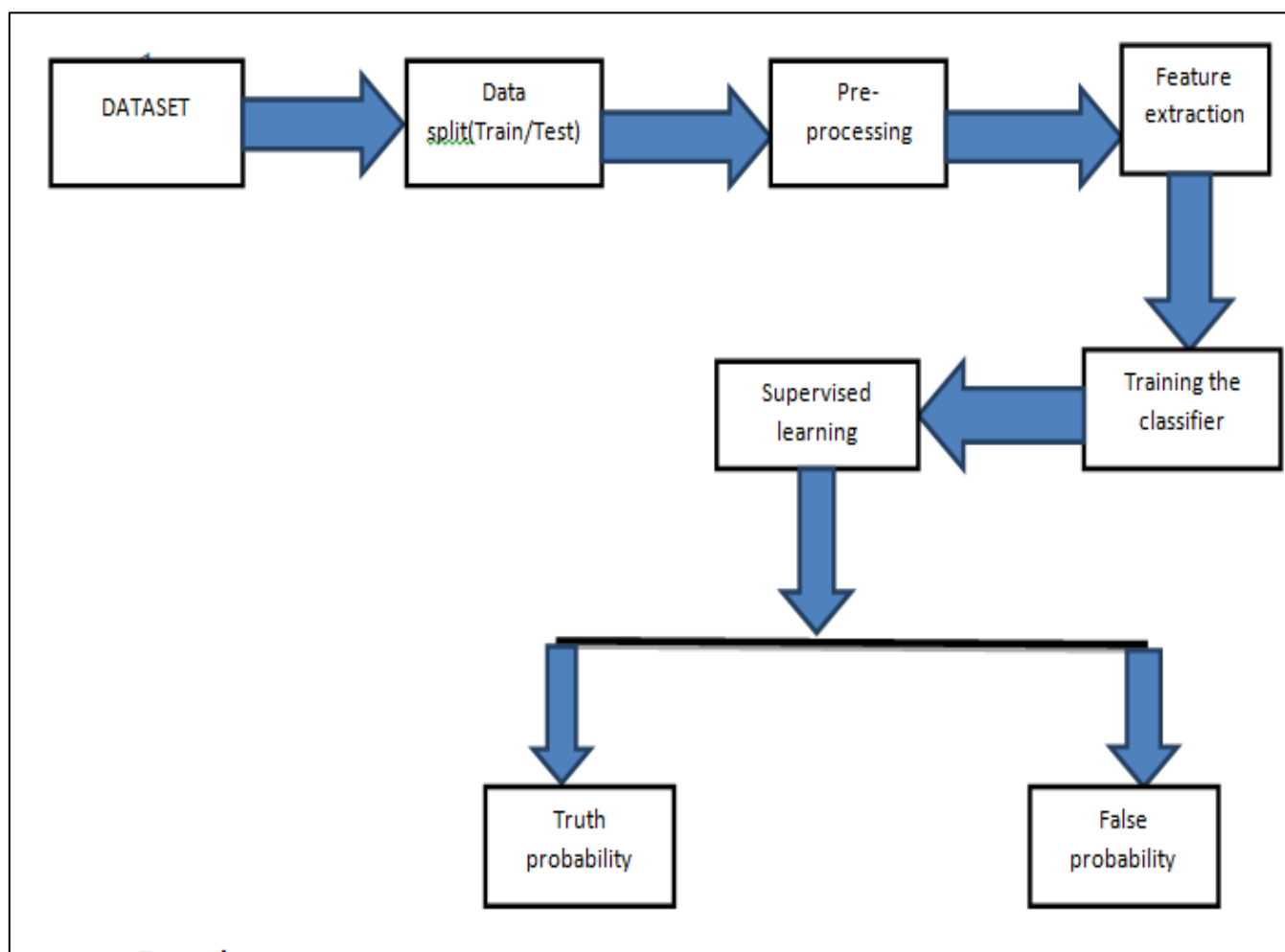
II. SYSTEM ARCHITECTURE

Fig 1: System Architecture

➢ *Data Set:*

The datasets used for this endeavor news.csv, is a broad social event of reports got from Kaggle. It mixes two specific datasets: one containing Genuine news and the other Phony news. The accompanying subtleties gives us a detail Depiction of dataset :

➢ *Dataset Sythesis*
- Fake News Dataset:
- Entries: 23,481
- Label: 1 (indicating fake news)

➢ *True News Dataset:*
- Entries: 21,417
- Label: 0 (indicating true news)

We are consolidated both datasets utilizing panda's inherent capability. The dataset we'll use for this Machine Learning project is news.csv. This datasets has a state of 7796*4. The main segment distinguishes the news, the second and third are the title and text, and the fourth section has marks signifying whether the news is Genuine or Counterfeit. The dataset occupies 29.2 MB .

```
target
fake      23481
true      21417
Name: text, dtype: int64
```


Fig 2 How many fake and real articles

➢ *Word Cloud*

Word Cloud is an information representation strategy utilized for addressing message information in which the size of each word shows its recurrence or significance. Huge text based information focuses can be featured utilizing a word cloud. Word mists are generally utilized for breaking down information from interpersonal organization sites.

Fig 3 Word Cloud for Fake News

Fig 4 Word Cloud for Real News

➢ *Confusion Matrix*

A confusion matrix is a table that is utilized to characterize the exhibition of an order calculation. It imagines and sums up the exhibition of an order calculation. It comprises of 4 fundamental qualities that are utilized to characterize the estimation measurements of the classifier. These four attributes are:

- TP(True Positive): The model has anticipated indeed, and the genuine worth was likewise obvious.
- TN (True Negative): The model has anticipated no, and the real worth was likewise no.
- FP (False Positive): The model has anticipated indeed, however the real worth was no. It is likewise called a Sort 1 mistake.
- FN (False Negative): The model has anticipated no, yet the real worth was yes. It is likewise called a Sort 2 mistake.

Fig 5 Confusion Matrix

➢ Algorithm
- Step 1: Extract Data from source.
- Step 2: Pre-process the text based information.
- ✓ Testing data.
- ✓ Eliminate stop words.
- ✓ Standardize text based information to shape vector grid utilizing tf-idf/count vectorizer.
- ✓ Create Feature matrix.

- Step 3: Train_Test Split.
- Step 4: Train the classifier.
- ✓ Random Forest.
- ✓ Decision Tree.
- ✓ Stochastic Gradient Descent.
- ✓ K Nearest Neighbour.

- Step 5: Test the algorithms.
- Step 6: Evaluate Algorithm.

➢ *Decision Tree Classifier*
This class is initialized with parameters such as criterion='entropy',max_depth=20,splitter='best',random_state=101.

The result shows that our DecisionTree Classifier algorithm was able to classify the test set with 99.62% accuracy in the particular dataset.

```
Classification Report for Decision Tree Classifier :
              precision    recall  f1-score   support

       fake       0.99      1.00      1.00      4648
       true       1.00      0.99      1.00      4332

   accuracy                           1.00      8980
  macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980

Confusion Matrix of Decision Tree Classifier:
 [[4642    6]
 [  27 4305]]
accuracy: 99.63%
```

Fig 6 Decision Tree Classifier

➢ *Random Forest Classifier*
This class is initialized with n_estimator 100, max_depth =8, random_state =42,verbose = 1, class_weight="balanced".

The result shows that our Random Forest classifier algorithm was able to classify the test set with 0.98 accuracy in the particular dataset.

```
Classification Report for Random Forest Classifier:
              precision    recall  f1-score   support

       fake       0.98      0.97      0.98      4677
       true       0.97      0.98      0.97      4303

   accuracy                           0.98      8980
  macro avg       0.98      0.98      0.98      8980
weighted avg       0.98      0.98      0.98      8980

Confusion Matrix of  Random Forest Classifier:
 [[4525  152]
 [  69 4234]]
```

Fig 7 Random Forest Classifier

➢ *Stochastic Gradient Descent.*

This class is initialized with parameters such as loss=modified_huber, shuffle = True, random state= 101.The result shows that our SGDClassifier algorithm was able to classify the test set with 0.74 accuracy in the particular dataset.

```
Classification Report for sgd Classification:
              precision    recall  f1-score   support

        fake       1.00      0.99      1.00      4648
        true       0.99      1.00      1.00      4332

    accuracy                           1.00      8980
   macro avg       1.00      1.00      1.00      8980
weighted avg       1.00      1.00      1.00      8980


Confusion Matrix of sgd Classification:
 [[4624   24]
 [  16 4316]]
accuracy: 99.55%
```

Fig 8 Stochastic Gradient Descent

➢ *K-Nearest Neighbors Classifier*

This class is initialized with one parameter n-neighbors. This is basically values forth k

The result shows that our K-Neighbors classifier algorithm was able to classifythe test set with 0.98 accuracy in the particular dataset.

```
Classification Report for knn Classification:
              precision    recall  f1-score   support

        fake       0.61      0.99      0.76      4693
        true       0.97      0.32      0.48      4287

    accuracy                           0.67      8980
   macro avg       0.79      0.65      0.62      8980
weighted avg       0.78      0.67      0.62      8980


Confusion Matrix of knn Classification:
 [[4652   41]
 [2930 1357]]
accuracy: 66.92%
```
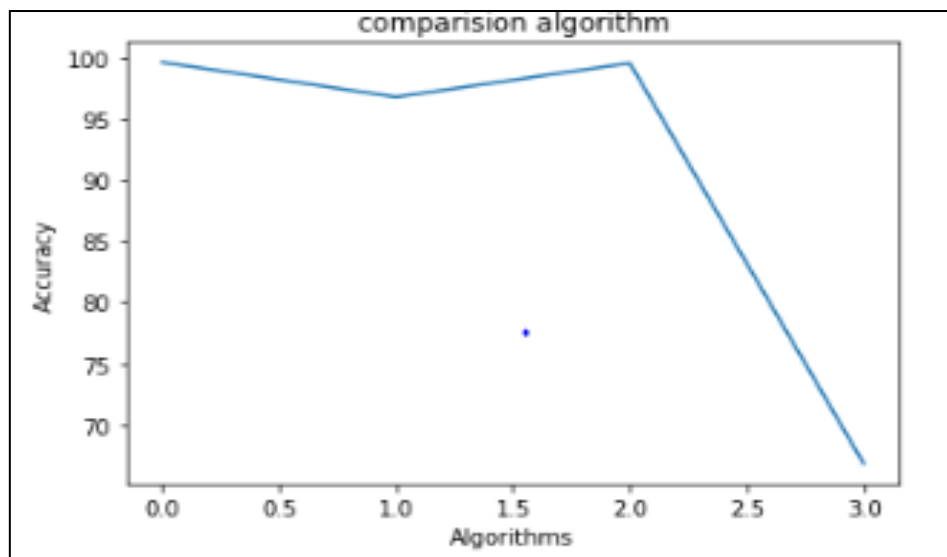
Fig 9 K-Nearest Neighbors Classifier

➢ *Machine Learning Model Comparison*

Fig 10 Machine Learning Model Comparison

I presented the classification report visualize displays the Comparison of Accuracy of Classification.

III. CONCLUSION

In this Research Machine learning techniques have been checked for functionality. Models are prepared utilizing information from news to confirm the framework adequacy on different stages to lessen the commonness of counterfeit news. These modules help in the correlation of genuine and counterfeit news. Twitter, face book, snap chat for example are well known stages that highlight news pages and applying these models there could help clients separate genuine or counterfeit news. I presented the classification report visualize displays the Comparison of Accuracy of classification, precision, recall, F1 and support scores, Combined graphical representation of Confusion Matrix for the model. In order to support easier interpretation and problem detection.

REFERENCES

[1]. Kecman, Support Vector Machines-An Introduction in "Support Vector Machines: Theory and Applications", Springer, New York City, NY, USA, 2005.

[2]. Kaggle, Fake News Detection, Kaggle, San Francisco, CA, USA, 2018, https://www.kaggle.com/jruvika/fake-news-detection.

[3]. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, pp. 127–138, Springer, Vancouver, Canada, 2017.

[4]. Evaluating Machine Learning algorithms for fake news detection. -Shloka Gilda

[5]. **"Fake News Detection on Social Media: A Data Mining Perspective"Authors:** Anil Kumar, S. Balaji, et al.

[6]. **"Detecting Fake News with Deep Neural Networks"Authors:** Y. Zhang, M. A. Elaziz, et al.

[7]. **"Fake News Detection Using Machine Learning Algorithms"Authors:** Santhosh Kumar, V. K. Dhiraj, et al.

[8]. **"A Survey on Fake News Detection with Deep Learning"Authors:** Aman Deep, Abhinav Moudgil, et al.

[9]. **"Combating Fake News with Machine Learning: A Survey"Authors:** M. Gupta, A. Singh, et al.

[10]. **"Fake News Detection: A Novel Approach Using Bert-Based Pretrained Language Models"Authors:** A. G. Tiwari, P. Kumar, et al

[11]. **"Towards Robust Fake News Detection: A Multimodal Approach"Authors:** C. Zhang, S. Yao, et al.

[12]. H. Jabeen, "Stemming and Lemmatization in Python", DataCamp Community, 2020. [Online]. Available:https://www.datacamp.com/community/tutorials/stemminglemmatization-python. [Accessed: 14-Jul- 2020].