International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Fake News Detection Using ML

A.Sushma[1], G.V.Lahari[2], S.Himaja[3]

[1]UG Student, Department of Computer Science and Engineering, CBIT, Proddatur, YSR, A.P
[2]UG Student, Department of Computer Science and Engineering, CBIT, Proddatur, YSR, A.P
[3]Assoc.Prof, Department of Computer Science and Engineering, CBIT, Proddatur,YSR,A.P
*Corresponding Author E-mail: aravasushma2@gmail.com

Abstract

In the digital era, social media and online platforms have become primary sources of news consumption. However, the rapid spread of fake news has created serious social, political, and economic issues. Fake news refers to false or misleading information presented as legitimate news. This project focuses on detecting fake news using Machine Learning techniques. The system collects news datasets, preprocesses the text, extracts meaningful features, and applies classification algorithms to determine whether a news article is real or fake. Algorithms such as Logistic Regression, Naïve Bayes, and Support Vector Machine are used for classification. The proposed model helps in automatically identifying misinformation with high accuracy and reduces the manual effort required for verification.

Keywords: Fake News, Machine Learning, Text Classification, Natural Language Processing, Logistic Regression, Naïve Bayes.

1. Introduction

Fake news has become a major concern in recent years due to the rapid growth of social media platforms. False information spreads quickly and influences public opinion. Manual verification of news articles is time-consuming and inefficient. Therefore, an automated system using Machine Learning is required to detect fake news effectively.

Machine Learning provides algorithms that can learn patterns from data and classify news articles as real or fake. By analyzing textual features such as word frequency, sentence structure, and context, the system can predict the authenticity of news articles.

1.1 Subsection Sample

Data Collection

The dataset consists of labeled real and fake news articles collected from reliable online sources. The dataset is divided into training and testing sets.

Data Preprocessing

Data preprocessing involves:

Removing punctuation and special characters

Converting text to lowercase

Removing stop words

Tokenization

Stemming or Lemmatization

This step ensures clean and structured input for the Machine Learning model.

2. Literature Review (Existing System & Proposed System)

There exists a large body of research on the topic of machine learning methods for deception detection, most of it has been focusing on classifying online reviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining “fake news” has also been the subject of particular attention within the literature. Conroy, Rubin, and Chen outline several approaches that seem promising towards the aim of perfectly classify the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars have been shown to be particularly valuable in combination with n-gram methods. Feng, Banerjee, and Choi are able to achieve 85%-91% accuracy in deception related classification tasks using online review corpora.

In this paper a model is build based on the count vectorizer or a tfidf matrix (i.e) word tallies relatives to how often they are used in other articles in your dataset) can help. Since this problem is a kind of text classification, Implementing a Naïve Bayes classifier will be best as this is standard for text-based processing. The actual goal is in developing a model which was the text transformation (count vectorizer vs tfidf vectorizer) and choosing which type of text to use (headlines vs full text). Now the next step is to extract the most optimal features for countvectorizer or tfidf-vectorizer, this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as “the”, “when”, and “there” and only using those words that appear at least a given number of times in a given text dataset.

3. Methodology (Architecture & Modules)

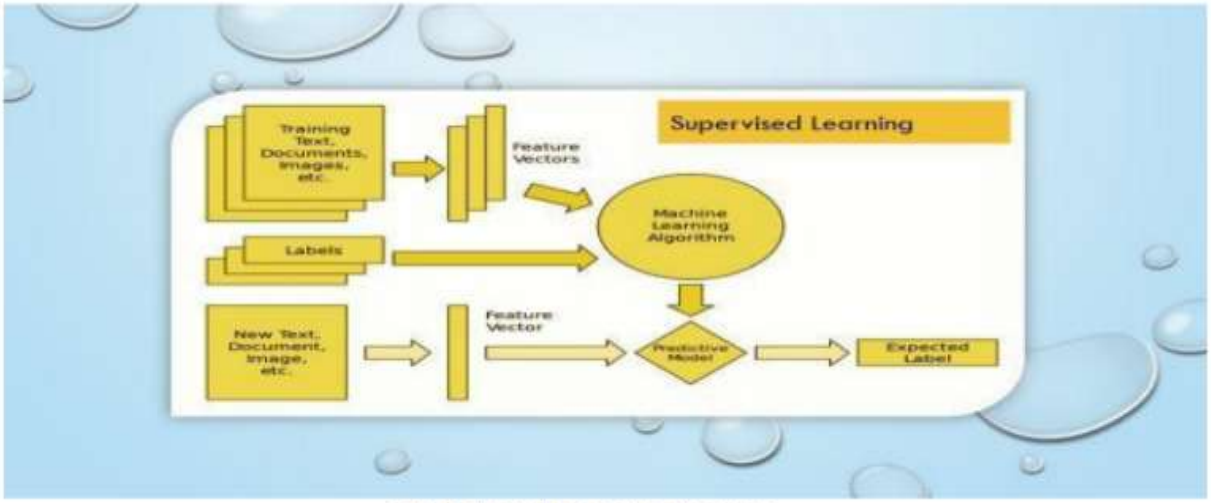

Fig:3.1 Architecture diagram

MODULES

Data Use

Preprocessing

Feature Extraction

Training the Classifier

MODULES DESCRIPTION

A. Data Use

So, in this project we are using different packages and to load and read the data set we are using pandas. By using pandas, we can read the .csv file and then we can display the shape of the dataset with that we can also display the dataset in the correct form. We will be training and testing the data, when we use supervised learning it means we are labeling the data. By getting the testing and training data and labels we can perform different machine learning algorithms but before performing the predictions and accuracies, the data is need to be preprocessing i.e. the null valueS which are not readable are required to be removed from the data Set and the data is required to be converted into vectors by normalizing and tokening the data SO that it could be understood by the machine. Next step is by using this data, getting the visual reports, which we will get by using the Mat Plot Library of Python and Sickit Learn. This library helps us in getting the results in the form of histograms, pie charts or bar charts.

B. Preprocessing

The data set used is split into a training set and a testing set containing in Dataset 1

-3256 training data and 814 testing data and in Dataset ll- 1882 training data and 47i testing data respectively. Cleaning the data is always the first step. In this, those words are removed from the dataset That helps in mining the useful information .Whenever we collect data online, it sometimes contains the undesirable characters like stop words, digits etc. which creates hindrance while spam detection. It helps in removing the texts which are language independent entities and integrate the logic which can improve the accuracy of the identification task.

C. Feature Extraction

Feature extraction s the process of selecting a subset of relevant features for use in model construction. Feature extraction methods helps in to create an accurate predictive model. They help in selecting features that will give better accuracy. When the input data to an algorithm is too large to be handled and i ts supposed to be redundant then the input data will be transformed into a reduced illustration set of features also named feature vectors. Altering the input data to perform the desired task using this reduced representation instead of the full-size input. Feature extraction is performed on raw data prior to applying any machine learning algorithm, on the transformed data in feature space.

D. Training the Classifier

As In this project I am using Scikit-Learn Machine leanring ibrary for implementing the architecture-Scikit Learn is an open source python Machine Learning library which comes bundled in 3rd distribution anaconda. This just needs importing the packages and you can compile the command as soon as you write it. If the command doesn't run, we can get the error at the same time. I am using 4 different algorithms and I have trained these 4 models i.e. Naive Bayes- Support Vector Machine, K Nearest Neighbors and Logistic Regression wic are very popular methods for document classification problem. Once the classifiers are traned, we Can c heck the performance Of the models on test-set. We Can extract the word count vector for each mail in test-set and predict it class with the trained models.

4. RESULTS AND DISCUSSION

- Algorithm"s accuracy depends on the type and size of your dataset. More the data, more chances of getting correct accuracy.
- Machine learning depends on the variations and relations.

- Understanding what is predictable is as important as trying to predict it.
- While making algorithm choice , speed should be a consideration factor.

REQUIREMENT ANALYSIS

Requirement analysis, also called requirement engineering. is the process of determining user expectations for a new modified product. It encompasses the tasks that determine the need for analysing, documenting, validating and managing software or system requirements. The requirements should be documentable, actionable, measurable, testable and traceable related to identified business needs or opportunities and define to a level of detail, sufficient for system design.

FUNCTIONAL REQUIREMENTS

It is a technical specification requirement for the software products. It is the first step in the requirement analysis process which lists the requirements of particular software systems including functional, performance and security requirements. The function of the system depends mainly on the quality hardware used to run the software with given functionality.

Usability

It specifies how easy the system must be use. It is easy to ask queries in any format which is short or long, porter stemming algorithm stimulates the desired response for user.

Robustness

It refers to a program that performs well not only under ordinary conditions but also under unusual conditions. It is the ability of the user to cope with errors for irrelevant queries during execution.

Security

The state of providing protected access to resource is security. The system provides good security and unauthorized users cannot access the system there by providing high security.

Reliability

It is the probability Of how Often the software fails. The measurement is Often expressed in MTBF (Mean Time Between Failures). The requirement is needed in order to ensure that the processes work correctly and completely without being aborted. It can handle any load and survive and survive and even capable of working around any failure.

Compatibility

It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience**.**

Flexibility

The flexibility of the project is provided in such a way that is has the ability to run on different environments being executed by different users.

Safety

Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one"s personal information.

NON- FUNCTIONAL REQUIREMENTS

Portability

It is the usability of the same software in different environments. The project can be run in any operating system**.**

Performance

These requirements determine the resources required, time interval, throughput and everything that deals with the performance of the system.

ACCuracy

The result of the requesting query is very accurate and high speed of retrieving information. The degree Of security provided by the system is high and effective.

Maintainability

Project is simple as further updates can be easily done without affecting its stability. Maintainability basically defines that how easy it is to maintain the system. It means that how easy it is to maintain the system, analyse, change and test the application. Maintainability of this project is simple as further updates can be easily done without affecting its stability.

SYSTEM DESIGN AND TESTING PLAN

INPUT DESIGN

The input design is the link between the information system and the user, It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in Such a way so that it provides security and Of use with retaining the privacy. Input Design considered the following things:

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results Of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct

source information to the user. Efficient and intelligent Output design improves the System"s relationship to help user decision-making.

The output form Of an information system should accomplish one or more Of the following objectives.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised SO that he is also able to make Some constructive criticism, which is welcomed, as he is the final user of the system.

S python

Python2.4.3(#1,Nov112010,13:34:43)

[GCC 4.1.220080704(RedHat4.1.2-48)] on linux2

Type"help" ,"copyright" , credits"or"license"for more information.

Sr .No	Methods & Description
	GET Sends data in unencrypted form to the server. Most common method.

GET, but send HTML server.

2 HEAD Same as without response body

all current 15

3 POST used to not cached form data to server. Data received by POST method is by representations Of the target resource with the content.

all current resource

4 Replaces uploaded

5 DELETE Removes representations ofthe target given by a URL

user = request.formCnm']

It is passed to Silsuccess" URL as variable part. The browser displays a welcome message in the window.

Fig:3.3 Dashboard

Change the method parameter to „GET" in login. html and open it again in the browser. The data received on server is by the GET method. The value of „nm" parameter is now obtained by —

```
user =request.args.get(„nm")
```

5. CONCLUSION

Many people consume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which has negative impacts on individual people and society- In this paper, an innovative model for fake news detection using machine learning algorithms has been presented. This model takes news events as an input and based on twitter reviews and classification algorithms it predicts the percentage of news being fake or real.

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company, For feasibility analysis, some understanding of the major requirements for the system is essential.This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

6. References

[ll. Parikh, S. B., & Atrey, P. K. (2018, April). Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.

[2]. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November). Automatic deception detection: Methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (p. 82). American Society for Information Science.

[3. Helmstetter, S. , & Paulheim, H. (2018, August). Weakly supervised learning for fake news detection on Twitter. In 2018 IEEE/ACM International Conference on

Advances in Social Networks Analysis and Mining (ASONAM) (pp. 274-277). IEEE. [4] Stahl, K, (2018), Fake News Detection in Social Media.

[51. Della Vedova, M. L.. Tacchini, E. , Moret, S. , Ballarin, G., DiPierro, M. , & de Alfaro, L. (2018, May). Automatic Online Fake News Detection Combining Content and Social Signals. In 2018 22nd Conference of Open Innovations Association (ERUCT) (pp. 272-279). IEEE.

[6] Tacchini, E., Ballarin, G. , Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:1704.07506.

[7] Shao, C. , Ciampaglia, G. L, Varol, O. , Flammini, A.. & Menczer, F. (2017). The spread offake news by social bots. arXiv preprint arXiv: 1707.07592, 96-104.

[8] Chen, Y. , Conroy, N. J., & Rubin, V. L. (2015, November). Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (pp. 15-19). ACM.

[9] Najafabadi, M. M.. Villanustre, F„ Khoshgoftaar, T. M„ Seliya, N., Wald, R. , & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1.

[101. Haiden, L, & Althuis, J. (2018). The Definitional Challenges of Fake News.