IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005065|

Fake News Detecting Web app Using Passive Aggressive and TF-IDF Vectorizer (Supervised Machine Learning Model)

Nayan Zope, Uzaif Shaikh, Kartike Dhote, Yash Sangole, Prof. Komal Bijwe

UG Student, Department of Computer Science Engineering, P R Pote College of Engineering and Research, Amravati, India

UG Student, Department of Computer Science Engineering, P R Pote College of Engineering and Research, Amravati, India

UG Student, Department of Computer Science Engineering, P R Pote College of Engineering and Research, Amravati, India

UG Student, Department of Computer Science Engineering, P R Pote College of Engineering and Research, Amravati, India

Assistant Professor, Department of Computer Science Engineering, P R Pote College of Engineering and Research, Amravati, India

ABSTRACT: With the advent of mobile technology and growing social media platforms, information is readily available. Mobile applications and social media have distorted the traditional media in spreading the news. Next to the climb in the use of online media categories such as Facebook, Twitter, etc. News spreads quickly among large numbers of clients who have a very limited ability to focus on time. Machine learning and methods based on knowledge and methodology are the two methods used to investigate content authenticity. Public and private testing in a wide area a variety of subjects are transmitted and disseminated continuously through various media outlets. Many ways are used, for example, controlled AI. The proliferation of fake news has far-reaching effects such as the bias of one-sided emotions influencing the results of the political race to support certain applicants. Additionally, spam senders use attractive news features to generate revenue using notifications by clicking on obstacles. In this paper, we aim to create a consistent group of variant news accessible online with the help of ideas associated with Artificial Intelligence, Natural Language Processing, and Machine Learning. The result of the project determines the availability of false information on social networks using a machine read and check the authenticity of the publishing news website.

I. INTRODUCTION

With the growing popularity of social media, more and more people use the news on social media instead of traditional media. False stories are now considered one of them the greatest threats to democracy, and freedom of speech. It has undermined public trust in government. Access to false news will always be enhanced. The widespread spread of false news has the potential to have the worst effects on the individual and the community. A kind of yellow journalism, false stories include pieces of stories that can be false and often spread on social media and other social media. This is usually done in order to forward or enforce certain ideas and is often achieved with political agenda. Such stories may contain lies and / or exaggerated claims, and may end up being so-called algorithms, and users may end up filtering its bubble.

- **Aim:**
  The main objective behind the development and upgradation of existing projects are the following smart approaches:
  - Beware of false topics, stories while passing them on to others
  - Telling true stories
  - Avoid false alarming incidents To Be Information

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005065|

- **Objective:**

  The main purpose of this project is to learn the problem of false news (including tweets, false posts, articles) on online forums and to make people easily understand the difference between false and real stories. Based on a variety of sources, including both the content of the text / profile / descriptions and the author's relationship to the subject matter, we aim to identify false stories from the online social network at the same timenews.This paper aims to create a systematic framework for in-depth research of false stories.

- **Motivation**:

  Machine learning (ML) is a form of Artificial Intelligence (AI) that allows software applications to be more accurate in predicting results without explicitly planning to do so. Machine learning algorithms use historical data as input to predict new output values. The spread of false stories can have a detrimental effect on individuals and communities. First, false news can distort the authenticity of the news ecosystem for example. Understanding the truth of the news and the message of discovery can have a positive impact on society.

II. LITERATURE REVIEW

[1] Fake News Detection on Social Media: A Data Mining Perspective. Author: - Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang and Huan Liu. In this paper to detect fake news on social media, a data mining perspective is presented that includes the characterization of fake news in psychology and social theories. This article looks at two main factors responsible for the widespread acceptance of fake messages by the user which are naive realism and confirmatory bias. It proposes a general two-phase data mining framework that includes 1) feature extraction and 2) modelling, analysing data sets, and confusion matrix for detecting fake news.

[2] MykhailoGranik et. al. in their paper shows a simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News). They achieved classification accuracy of approximately 74%. Classification accuracy for fake news is slightly worse. This may be caused by the skewness of the dataset: only 4.9% of it is fake news.

[3] Himank Gupta et. al. gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag and high processing time to handle thousands of tweets in 1 sec. Firstly, they have collected 400,000 tweets from HSpam14 dataset. Then they further characterize the 150,000 spam tweets and 250,000 non- spam tweets. They also derived some lightweight features along with the Top-30 words that are providing highest information gain from Bag-ofWords model. 4. They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately 18%.

[4] Social networking sites read news mainly in three ways: The (multilingual) text is analysed with the help of computational linguistics, which semantically and systematically focuses on the creation of the text. Since most publications are in the form of text, a lot of work has been done on analysing them. Multimedia: Several forms of media are integrated into a single post. This can include audio, video, images, and graphics. This is very attractive and attracts the viewer's attention without worrying about the text. Hyperlinks allow the author of the post to refer to various sources and thus gain the trust of viewers. In practice, references are made to other social media websites, and screenshots are inserted.
[5]Evaluating Machine Learning algorithms for Fake News Detection. Author: - Shloka Gilda. In this article, the author introduced the concept of the importance of NLP in stumbling across incorrect information. They have used time frequency-inverse document frequency (TF-IDF) of bigrams and probabilistic context-free grammar detection. Shloka Gilda introduced the concept of the importance of NLP in stumbling over incorrect information. They used BiGram Count Vectorizer and Probabilistic Context-Free Grammar (PCFG) to detect deceptions. They examined the data set in more than one class of algorithms to find out a better model. The count vectorizer of bi-grams fed directly into a stochastic gradient descent model which identifies non credible resources with an accuracy of 71.2%.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

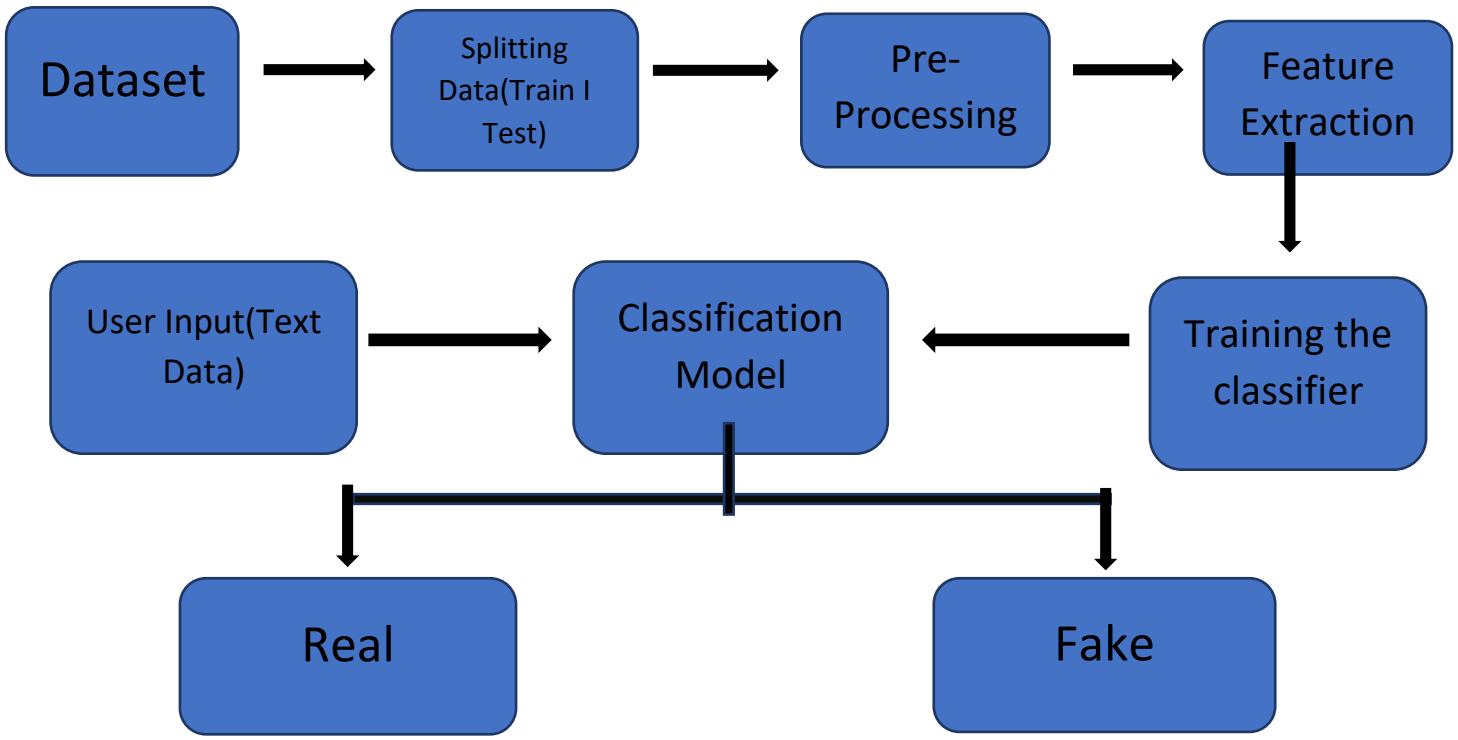
| DOI: 10.15680/IJIRCCE.2022.1005065|

- **Summary:**
We took a news dataset from Kaggle, implemented a Tf-IdfVectorizer, initialized a Passive Aggressive Classifier, and fit our model. We ended up obtaining an accuracy of 94% in magnitude.

III. PROPOSED METHODOLOGY

This project will help to find a way to use Natural Language Processing (NLP) to identify and classify fake news articles. The main purpose is to find false stories, which is a problem to separate the old text. We will collect our data, process text, and convert our articles into features that will be used on supervised models. We will use the Passive-Aggressive section to train data sets and tests in news articles. For this project, we will be using Python and Sci-kit libraries. Python has a wonderful collection of libraries and plugins that you can use in machine learning. The Sci-Kit Learn library is the best tool for machine learning algorithms, almost all types of machine learning algorithms readily available in Python, so easy and quick testing of ML algorithms is also possible. We used flash to extract the model and the HTML help, CSS finally.

IV. SYSTEM DESIGN



V. IMPLEMENTATION

- **Data Collection:**

In the first step of operation is data collection. The machine learning algorithm used in this project is called supervised learning. Learning is said to be monitored if the model is trained in a data set consisting of both input and output components. In supervised learning, the model is trained using a data set containing both input and output parameters. To train the model we took a database from kaggle.com The database has 20000 news articles and 5 attributes.Adjective names are 'id', 'title', 'author', 'text' and 'label'. In all four there are independent input or alternative parameters which are 'id', 'title', 'author', and 'text'. The adjective 'label' is a variable variables or output parameter. The 'label' means that a news article is 'True' or 'False'.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005065|

- **Data Processing:**

This step is to pre-process the text. The function of the text-separating model depends largely on the corpus words and the characteristics created in those words to form the model. In preprocessing we are omitting the stopwords from the news article.We use lemmatization that will erase common morphological words and produce the origin of translated words and all of this is done by using nltk.Therefore this process will help to reduce the size of the feature and increase the efficiency of the model.

Feature Extraction:

Machine learning algorithms work with numerical values to convert text into something the machine can understand. We take advantage of Natural Language processing which translates text into numerical vector. In Natural Language processing, there are two methods for extracting one element, the count vectorizer and the TF-IDF (Term frequency-inverse document frequency). For this project, we used the TF-IDF strategy.

TF-IDF Vectorizer

TF (Term Frequency): The frequency with which a word appears in a document is its Term Frequency. A higher value means that one term occurs more often than others, so the document fits well if the term is part of the search terms.

IDF (Inverse Document Frequency): Words that occur many times in a document, but also occur many times in many others, maybe irrelevant. IDF is a measure of how important a term is in the entire corpus. IDF can be calculated as follow:

$$idf_i = \log(\frac{n}{df_i})$$

The TF-IDF score as the name suggests is just a multiplication of the term frequency matrix with its IDF, it can be calculated as follow:

$$w_{i,j} = tf_{i,j} \times idf_i$$

Where w_{ij} is TF-IDF score for term i in document j, tf_{ij} is term frequency for term i in document j, and idf_i is IDF score for term i.

Passive Aggressive Classifier:

Passive-Aggressive algorithms are generally used for large-scale learning. It is one of the few 'online-learning algorithms. In online machine learning algorithms, the input data comes in sequential order and the machine learning model is updated step-by-step, as opposed to batch learning, where the entire training dataset is used at once. This is very useful in situations where there is a huge amount of data and it is computationally infeasible to train the entire dataset because of the sheer size of the data. We can simply say that an online-learning algorithm will get a training example, update the classifier, and then throw away the example.

Passive: If the prediction is correct, keep the model and make no changes. That means the data in the example is insufficient to effect a change in the model.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005065|

Aggressive: If the prediction is incorrect make a change to the model i.e. some change to the model may correct it. After that, a model is formed which is trained on the data of the training set and will be applied to the testing dataset to evaluate the performance of this classifier.

Evaluating Metrices:

Evaluate the performance of algorithms for false information detection problems, using a variety of test metrics. In this section, we review the metrics most commonly used to detect false stories. Many existing methods view the issue of false news as a divisive issue that predicts whether a news article is fake or not:

True Positive (TP): when predicted fake news pieces are actually classified as fake news;
True Negative (TN): when predicted true news pieces are actually classified as true news;
False Negative (FN): when predicted true news pieces are actually classified as fake news;
False Positive (FP): when predicted fake news pieces are actually classified as true news.

Confusion matrix: Basically this metrics how many results are correctly predicted and how many results are not correctly predicted.

Total	Class 1 (Predicted)	Class 2 (Predicated)
Class 1 (Actual)	TP	FN
Class 2 (Actual)	FP	TN

Implementation steps:

- Libraries required :

  1. NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
  2. Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.
  3. sk-learn: Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.
  4. nltk: The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

- Create features using TF-IDF Vectorizer.
- Split the data in training and testing.
- Then transfer the training data to the Passive Aggressive Classifier algorithm. Passive Aggressive Classifier is part of the online learning algorithms for machine learning. It works by responding as idle in the appropriate categories and responding as aggressive to any incorrect calculation.
- Then we will check the accuracy of the model and apply it to the web framework of the flask.
- Our web application will take news as installed and classify it as Fake or Real by model.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005065|

VI. CONCLUSION

However, social media has also been used to spread false stories, which have a powerful negative impact on individual users and the wider community. The aim of this project is to completely review, summarize, compare and evaluate current research on false news. This paper concludes that using passive aggressive and TF-IDF vectorizer is effective as we have achieved 94% accuracy in this model.

REFERENCES

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, ―Fake News Detection on Social Media: A Data Mining Perspective‖ arXiv:1708.01967v3 [cs.SI], 3 Sep 2017

[2] Uma Sharma, Siddarth Saran, Shankar M. Patil, 2021, Fake News Detection using Machine Learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU – 2020 (Volume 09 – Issue 03)

[3] Uma Sharma, Siddarth Saran, Shankar M. Patil, 2021, Fake News Detection using Machine Learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU – 2020 (Volume 09 – Issue 03)

[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] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 436- 441, DOI: 10.1109/MIPR.2018.00093.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details