News aggregator web app with Fake news detection

Sachin[1], Tushar Khandelwal[2], Vikas Kumar[3]

[1]Department of Applied Mathematics Delhi Technological University, India
[2]Department of Applied Mathematics Delhi Technological University, India
[3]Department of Applied Mathematics Delhi Technological University, India

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: in recent times, due to the advent of technology, due to its ease of availability have made the passing of information very easy that leave little signs of authenticity. Such unvalidated news are termed as fake news. Manipulation of human mind has been illustrated for many decades via an adequate usage of news filled with hate and incorrect content. Recent progress in the field of smartphones, software has caused an abrupt increase in the realism of forged content and they can be made very conveniently. Dissemination of content with such absurdity is an easy task to do.. To accomplish this daunting task, we decided to use machine learning algorithm, some classifiers and Django. We begin by extracting features from passed on news. We will train a model using those extracted features. We will compare the performance of our model with some results from a standard data set. We will keep on improvising our model till it becomes good enough to work with real world data.

Keywords: BeautifulSoup, Dataset, Decision Tree, Fake News, News Aggregator

1. Introduction

1.1 Motivation

The initiate idea was to develop an app capable of questioning the news articles which are provided to us by the major news agencies worldwide for some time and translate all the content into reading points through which the user will have an proper understanding of news.We have seen evolution of decline in viewership of news papers and TV. Youth instead of using conventional means prefers to use social networking as source of information. Conventional means of information are on the verge of going extinct

1.2 Problem definition

This paper works on this paper works on the premise of finding if given is fake news or not. So, we will build our web application in Django server for news scraping and aggregation with the help of python. This app will follow two big News Companies Times of India and Google news . All the scraped content will be layed out on a single Website. Different machine learning models are utilized for prediction if given news is fake or genuine

1.3 What is news aggregator?

It is web application which collects information(news articles) from more than one like websites to be presented at single easily accessible link for user. News aggregator carrier is a totally critical begin of the day.You can easily find news sites and information hub on the web. They submit their content on more than one systems. It is not efficient if to gain information(news) about your surroundings have to open 10-20 websites everyday. It is time wasted to get information. It can give you leverage over those who don't have it. Now, is there a way we can make it easier? Yes!!Being informed is everything nowadays and that's where News Aggregator comes in. With it you can choose which websites to follow and which news articles should be collected for you. With pressing of a button all the news related to your interests are at your disposal from different websites otherwise it would take forever to get to desired newsThis task otherwise takes too much time on our scheduleA news aggregator is a system that takes news from several resources and puts them all together. A good example of news aggregator are **JioNews** and **Google News**.

1.4 Why to build a news aggregator?

There are hundreds of news websites, they do cover news on several broad topics, out of which only a few of them are of our interest. A news aggregator can be a tool to save a lot of time and with some modifications and filteration , we can fine tune it to show only news of our interest..\[2]

1.5 Why fake news is circulated?

In order to solve the issues discussed in several issues need to be solved. First the main goal of this graduation project is to find a solution for discrediting fake news. In order to do so a platform will be developed that aggregates news and personalises them to the user's interests. Hence researching existing methods of aggregating information from various websites and personalizing them according to a user's interests is necessary.

Due to the fact that the terms fake news and real news can be vague and situation depended, a definition inegard to how those terms will be handled throughout this report will be provided

Fake news: news that intentionally spreading false information.
Real news: News published by generally highly trusted publisher they have rigrous step to validate news.

1.6 What is Web Scraping and How to Use It?

Web Scripting is a programmed strategy to get a lot of information from sites. The majority of this information is unstructured information in a HTML design which is then changed over into organized structured form in a DB page or a data set so it tends to be utilized in different applications. There are various approaches to perform web scratching to get information from sites. these incorporate utilizing on the web administrations, specific API's or in any event, making your code for web scratching without any preparation. Numerous huge sites like Google, Twitter, Facebook, StackOverflow, and so forth have API's that permit you to get to their information in an organized arrangement. This is the most ideal alternative yet there are different locales that don't permit clients to get to a lot of information in an organized structure or they are essentially not excessively mechanically progressed. Around there, it's ideal to utilize Web Scraping to scratch the site for information. Web scratching requires two sections to be specific the crawler and the scrubber. The crawler is a man-made reasoning calculation that peruses the web to look through the specific information needed by following the connections across the web. The scrubber, then again, is a particular device made to extricate the information from the site. The plan of the scrubber can fluctuate enormously as indicated by the intricacy and extent of the task so it can rapidly and precisely extricate the information

2. Research work done

We started our learning through online courses on Machine Learning.Then we read some of research papers. There are several researches done in the same field here we display news which are relevant to user and which are more of genuine news.

3. Literature review

The amount of data generated in the world today is very huge. This data is generated not only by humans but also by smartphones, computers and other devices. Based on the kind of data available and a motive present, certainly, a programmer will choose how to train an algorithm using a specific learning model. Machine Learning is a part of Computer Science where the efficiency of a system improves itself by repeatedly performing the tasks by using data instead of explicitly programmed by programmers. Further let us understand the difference between techniques of Machine Learning- Supervised and Unsupervised learning

3.1 Supervised learning

In Supervised Learning , it works with functions of regression and classification. Supervised Learning maps labelled information to output and in this the output information styles are recognized by the system.

3.2 Unsupervised Learning

In Unsupervised Learning, there may be no entire and smooth labelled dataset in unsupervised learning . Unsupervised learning is self-prepared learning . It predicts the output by learning patterns and features from unlablled datsset

3.3 Natural language processing

Branch of artificial intelligence which falls under machine leaning that deals with interaction between human and computer using the natural language. The ultimate objective of this method to be used in this project is to read, decipher and validate a news.

4. Methodology

This project has been divided into two main firstly building a news web app aggregating news from sources using Django serve, second implementing a machine leaning model which will be detecting and validate a news is fake or genuine

4.1 Dataset

You can locate many datasets for fake information detection on Kaggle or many different sites. I have downloaded those datasets from Kaggle. There are datasets one for faux news and one for Real news. In real news, there may be 21417 news, and in Fake news, there may be 23481 news. Both datasets have a label column wherein 1 for Fake news and zero for Real news. We have blended each datasets the use of pandas integrated function.

4.1.1 Cleaning data: Splitting of data: Dataset may contains redundant value and duplicate values for which it has to be cleared. Cleaning of the data set so we will make one 'one_drop' function which cleans the data.

4.1.2 Splitting Data: Splitting the data is the most essential step in machine learning. We are splitting the data in two frames train dataset and test dataset so that we can train our model on train dataset and checking the efficiency of model developed on test dataset. We split the test data in 25% as the train dataset and 75% test data set We train our model on the trainset and will test data on testing dataset . We split our data in train and test with the use of train_test_split function from python library Scikit learn.

.

Till now we can see our news aggregator web application is prepared which contain news title ,we can download this information and convert into csv record yet performing fake news identification on exceptionally little dataset prompts extremely terrible AI model with extremely less exactness so we chose to utilize dataset from Kaggle site which is immense dataset .

4.2 Building the Web App

4.2. 1. Building News Aggregator

We'll build our news aggregator in 3 parts. These are following:

1. We'll research on html source code of news sites and build a website scrapper for each
2. Then, We'll setup our django server
3. Finally, we'll integrate everything altogether

Then we'll utilise BeautifulSoup python library for scraping of News content from Times of India and Google News. With the help of BeautifulSoup, we are able to interpret the HTML substance of the given URL and giving us access to its components by recognizing them with their labels and properties Therefore, we will utilize it to remove certain bits of text from the sites. It is an incredibly simple to-utilize yet amazing bundle. With right around 3–5 lines of code we will actually want to extricate any content we need from the web To give BeautifulSoup the HTML code of any page

.

4.3 Building fake news model

Jupyter notebook framework is used to implement. Jupyter notebook is notebook is an open-source, browser-based tool act as a virtual lab notebook to support workflows, code, data, and visualizations detailing the research process. It is machine and human-readable, which facilitates interoperability and scholarly communication. These notebooks can live in online repositories and make connections with datasets, code, methods documents, workflows, and publications easily. Jupyter notebook is the one which make science more open.

4.4. Feature Extraction

Machine learning works only with numerical data so first data is converted from textual content to numerical data. For that, we need to preprocess the converted data and that is known as Natural Language processing

In-textual content preprocess we're cleansing out textual content via way of means, taking away stopwords, unique symbols and numbers, etc. After cleansing the statistics we ought to feed this article statistics right into a vectorizer so that it will convert this article statistics into numerical features.

So , We have employed the use of Tfidf Vectorizer. It will transform each news content into a matrix. Matrix will contain tfdif features will helps us in recognize the significance of word in the corpus while analysing the article

Methods involved here

4.4.1. Logistic Regression

Logistic regression is tool used for classifying textual content . it is employed because of its presentation of intuitive equation to organize issues into binary or a couple of classes We accomplished hyperparameters tuning to get the nice end result for all independent datasets, whilst a couple of parameters are examined earlier than obtaining the most accuracies from LR version A Sigmoid function is used for converting the output to probability value.

4.4.2. Decision tree classification

In this Decision Tree, Models for classification or regression are build in the form of tree structure. This works by splitting of datasets into smaller subsets and hand in hand a related decision tree is build incrementally. Tree with a decisions and leaf nodes are resultant tree is the final output. **Decision node** (Result) contains two or more branches (e.g., WIN,TIE ,LOSE). **Leaf nodes** (Tournament Winner) shows us a decision or classification.

4.4.3. Gradient boosting classifiers

Gradient boosting classifiers are a set of Machine Learning algorithms that integrate many susceptible studying fashions collectively to create a firm predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting is in-demand classifier because of oits ability to handle the intricate datasets

4.44. Random Forest

The Random Forest(RF) is an arrangement calculation comprising of numerous decision trees. It utilizes packing and highlight unpredictability.Uncorrelation between models the key in this classifier . A G roup of different decision trees will give out a far more correct prediction than they would give out on their own. Reason behind this some trees might have slightly less accurate predecitions but as a whole group they give out accurate predictions because of very low correlation

5. Conclusion

This project could be practically used by media companies to automatically predict whether the circulating news is fake or not. The process could be done automatically without having humans manually review thousands of news-related articles.

Now, we can configure this to gather your favorite article websites. Many times, bots are not legally allowed to scrape content. So, web scraping comes at its own cost.But, for our purpose, we now know some very cool basics. We also have a very interesting project to showcase.

Graph theory and machine learning techniques can be employed to identify the key sources involved in spread of fake news. Likewise, real time fake news identification in videos can be another possible future direction.

References

1. N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.
2. S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 171–175.
3. ShlokGilda,Department of Computer Engineering, Evaluating Machine Learning Algorithms for Fake News Detection,2017 IEEE 15th Student Conference on Research and Development (SCOReD)
4. Douglas, "News consumption and the new electronic media," The International Journal of Press/Politics, vol. 11, no. 1, pp. 29–52, 2006.View at: Publisher Site | Google Scholar
5. J. Wong, "Almost all the traffic to fake news sites is from facebook, new data show," 2016.View at: Google Scholar
6. M. J. Lazer, M. A. Baum, Y. Benkler et al., "The science of fake news," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.View at: Publisher Site | Google Scholar
7. S. A. García, G. G. García, M. S. Prieto, A. J. M. Guerrero, and C. R. Jiménez, "The impact of term fake news on the scientific community scientific performance and mapping in web of science," Social Sciences, vol. 9, no. 5, 2020.View at: Google Scholar
8. D. Holan, 2016 Lie of the Year: Fake News, Politifact, Washington, DC, USA, 2016.
9. S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake News: Evidence from Financial Markets," 2019, https://ssrn.com/abstract=3237763.View at: Google Scholar
10. Robb, "Anatomy of a fake news scandal," Rolling Stone, vol. 1301, pp. 28–33, 2017.View at: Google Scholar
11. https://data-flair.training/blogs/django-project-news-aggregator-app/
12. https://www.hackersfriend.com/articles/building-news-aggregator-web-app-with-django-using-python-web-scraping
13. J. Soll, "The long and brutal history of fake news," Politico Magazine, vol. 18, no. 12, 2016.View at: Google Scholar