Machine Learning and NLP-based Fake News Detection.

Ketan Pande[1],
Department of Computer Engineering,
Wagholi, Pune

Rahul Prajapati[1],
Department of Computer Engineering,
Wagholi, Pune

Sadik Pathan[1]
Department of Computer Engineering
Wagholi, Pune

Akshay Patil
Department of Computer Engineering
Wagholi, Pune

Abstract:

Fake news has been a problem ever since the internet boomed. The easier access and exponential growth of the knowledge offered on social media networks have created it knotty to largely differentiatebetween false and true information. Opposing such fake news is important because the world's view and mindset are shaped by information. People form their own opinions through the day-to-day news. If this information is false, it can have devastating consequences. The quality of social media networks is additionally at stake wherever the spreading of pretend data is prevailing. Machine learning and Natural Language Processing have competed for a significant role in the classification of the data though with some limitations. The need of an hour is to stop these types of fake news especially in developing countries like India and focus on the correct, proper news article which will not affect people's mentality negatively.

Keywords: Machine Learning, Scikit-learn, supervised learning, NlTK, tfidf, NLP, Flask, bagging, boosting, etc.

I- INTRODUCTION

In today's world, most of the information is hazy available on various platforms suchas Twitter, blogs, online e-paper, social media. Today's youngsters spent most of thetime on social media or the internet. News on social media is additional appealing and fewer expensive compared to the optional ancient news organization and it's simple to do that 3 magical things share, like and comment however despite giving the profit, this category of reports from social media is minor than another earliest news sources. In Today's world, anybody can post content over the internet and some individuals hence post some misleading information.

Falsified news is any textual or non-textual content that is fake News and is generated so the readers will start believing in something which is not true. Opposing this fake news is the prime thingbecause the world's view is shaped by information. Many Machine Learningalgorithms have been used on different types of datasets to classify the news is fake or real.

The main objective is to discover the faux news, which is a classic text classification drawback with an undemanding proposition. The projected system helps to seek out the authenticity of the news. If the news isn't real, then the user is suggested with the relevant article.

II- EXISTING SYSTEMS

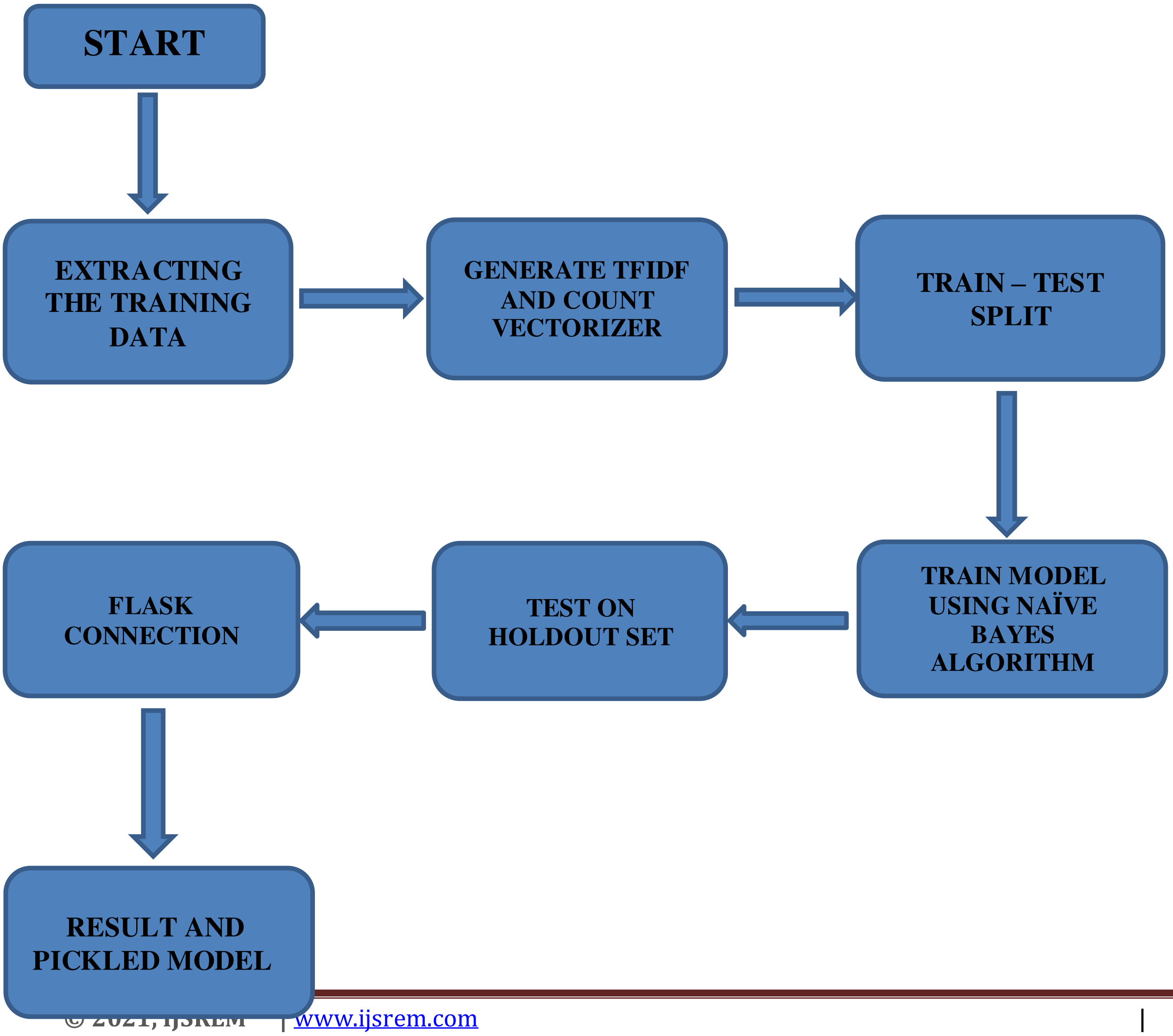
<u>Title: - Fake News Detection Using Machine Learning approaches: A systematic Review</u> Author: - Syed isfaq Manzoor, Jimmy Singla, Nikita

Description: - The easy access and exponential growth of the information available on social media networks has made it intricate to distinguish between false and true information. The easy dissemination of information by way of sharing has added to exponential growth of its falsification. The credibility of social media networks is also at stake where the spreading of fake information is prevalent. Thus, it has become a research challenge to automatically check the information viz a viz its source, content and publisher for categorizing it as false or true. Machine learning has played a vital role in classification of the information although with some limitations. This paper reviews various Machine learning approaches in detection of fake and fabricated news. The limitation of such and approaches and improvisation by way of implementing deep learning is also reviewed.

Title: - A smart System for Fake News Detection Using Machine Learning
Author: - Anjali Jain, Avinash Shakya, Harsh Khattar
Description: - Most of the smart phone users prefer to read the news via social media over internet. The news websites are publishing the news and provide the source of authentication. The question is how to authenticate the news and articles which are circulated among social media like WhatsApp groups, Facebook Pages, Twitter and other micro blogs & social networking sites. It is harmful for the society to believe on the rumors and pretend to be a news. The need of an hour is to stop the rumors especially in the developing countries like India, and focus on the correct, authenticated news articles. This paper demonstrates a model and the methodology for fake news detection. With the help of Machine learning and natural language processing, author tried to aggregate the news and later determine whether the news is real or fake using Support Vector Machine.

Title: Fake News Detection by Decision Tree
Author: - Shikun Lyu, Dan Chia-Tien Lo
Description: - Fake news detection research has appeared for a couple of years and is a relatively new and difficult research field. The difficulties come from the semantics of natural languages and manual identification via human beings, let along machines. In this project, we propose to analyze the performance of several machine learning algorithms integrating tools such as Fake News Tracker [1], doc2vec, Support Vector Machine (SVM), and decision trees. Our preliminary results indicate that the SVM and the decision trees are suitable to identify fake news with an acceptable accuracy of 95 percent. Typically, the decision trees method shows a better result than SVM. Future research directions will be addressed

III- PROPOSED SYSTEM


START
EXTRACTING THE TRAINING DATA
GENERATE TFIDF AND COUNT VECTORIZER
TRAIN – TEST SPLIT
TRAIN MODEL USING NAÏVE BAYES ALGORITHM
TEST ON HOLDOUT SET
FLASK CONNECTION
RESULT AND PICKLED MODEL

1] Extracting the training dataset:

We had 2 datasets. "True.csv" and "Fake.csv"."True.csv" contained only true news and Fake.csv contained only fake news. Then we combine these both files using "concat" function in pandas module and shuffled them together by "shuffle" function so that both of them get mixed together in a single file .Our dataset consisted of some columns which were of no use i.e it would not have affected the accuracy of our model so we drop them ,after dropping the unnecessary columns in the dataset we removed punctuation, stop words by using the necessary respective modules from python as such things can affect the accuracy to a certain extent.

After stop words removal, we performed Exploratory Data Analysis (EDA) on remaining columns to determine the relation between columns. We trained various models like decision tree classifier, logistic regressor, RandomForestClassifier
and Naive Bayes Classifier and we find the accuracy of all these Algorithms, Decision Tree had the highest Accuracy among all, but it was only when the dataset was large, for smaller datasets it was giving a less accuracy. On the Other hand, Naive Bayes was giving almost the same accuracy for larger and smaller datasets. So, we came to the decision that for our model Naive Bayes
Classifier would be the better choice.

We created the user interface where user can enter the url and our System will predict whether the news in that url is true or fake. As soon as the user clicks the predict button, web Scrapping is done on the news and the data in the news on that url is extracted and that news is converted into an article.

2.Generate TFIDF and Count vectorizer:

As soon as the news is converted into an article, we downloaded that article using download () function. After downloading article, we parsed that article using parse () method and then we applied Natural Language Processing (NLP) technique in that article. We also applied Term Frequency Identification(tfidf), Count Vectorizer on that article and we remove all the stop words inside that article and stored the whole updated article in one variable. [Stop words are those words in English which have no meaning. So, removing such words is good option for our Algorithm.]

3.Train Test Split:

From sklearn.model_selection we import train_test_split in order
to split the data inside the dataset. We trained our 75%-80% of the data and remaining 25%-20% of data is used to test the model.

4.Train model using Naive Bayes Classifier:

After splitting the data inside the dataset ,75%-80% of the split data is used to train the Naive Bayes classifier, Naïve Bayes will algorithm will get processed on the data and will generate some trained data and remaining 20%-25% data will be used for testing the data at later stage which will be used to generate the accuracy, f1-score, confusion matrix and other performance metrics.

5.Test on holdout set:

After the model is trained and validated with our algorithm, we get the final estimate of the machine learning model performance. After this we test our unseen data using test set to check the performance of our model.

6.Flask connection:

when we think about creating a reliable, scalable and maintainable web application in python, the first thing that comes into mind is Flask framework. The framework is the code library that makes the developers life easier when building the web application. In our project we had a necessity of creating a model where user can enter the url to find out whether the news on that url is fake or real. So that is the reason why we established flask connection.

7.Result and Pickled Model:

Python pickle module is used for serializing and de-serializing python object structure. The process to convert any kind of python object (list, dict, etc) into byte stream (0s and 1s) is called as pickling or serialization or flattening or marshalling.

We use dump () function in python pickle module for converting the objects into streams of 0s and 1s.

IV – ALGORITHM AND TERMS USED

1. Machine Learning:

Machine learning may be a growing technology that permits computers to find out automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it's getting used for recognition, speech- recognition, email filtering, Facebook- tagging, recommender system, and lots of more.

I. Supervised Learning:

Supervised learning is that the type of machine learning during which machines are trained using well "labeled" training data, and on basis of that data, machines predict the output. The labeled data means some input file is already tagged with the right output.

A supervised learning algorithm aims to seek out a mapping function to map the input variable(x) with the output variable(y).

1) Naïve Bayes:

Naïve Bayes is additionally a supervised Machine Learning algorithm, which is predicated on the Bayes theorem and is employed for solving classification problems.

It is mostly used in text classification which includes a high-dimensional training dataset. Naive Bayes classifier is one of the simple, most effective, and probabilistic classification algorithms which predict based on the probability of an object.

Ex: Spam Filtration, Sentiment analysis, etc.

2)Bayes Theorem:

Bayes Theorem provides a principled way for calculating a conditional probability.

It is a deceptively simple calculation, although it can be used to easily calculate the conditional probability of events where intuition often fails.

Although it is a powerful tool in the field of probability, Bayes Theorem is also widely used in the field of machine learning. Including its use in a probability framework for fitting a model to a training dataset, referred to as maximum a posteriori or MAP for short, and in developing models for classification predictive modelling problems such as the Bayes Optimal Classifier and Naive Bayes.

3) Natural Language Processing (NLP):

NLP stands for natural language Processing, which is an eternal part a part of computing, Human Language, and AI. It is the technology that's employed by machines to know, analyze, manipulate, and interpret human languages. It helps developers to arrange knowledge for performing tasks.

Ex. Translation, automatic summarization, Named Entity Recognition (NER), speech recognition, relationship extraction, and topic segmentation

4) TFIDF:

TF-IDF is an information retrieval and knowledge extraction subtask which aims to precise the importance of a word to a document which is a component of a set of documents that we usually name a corpus. It is usually used by some search engines to help them obtain better results that are more relevant to a specific query.

5) NLTK:

NLTK (Natural Language Toolkit) could also be a set that contains libraries and programs for statistical language processing. It is one of the foremost powerful NLP libraries, which contains packages to make machines understand human language and reply to them with an appropriate response.

6) Classification Report:

A Classification report is employed to live the standard of predictions from a classification algorithm.

TERMS OF CLASSIFICATION REPORT:

1. Accuracy:

It defines how often the model predicts the right output.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision:

It is often defined because the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, what percentage of them were true.
algorithm.

$$Precision = \frac{TP}{TP+FP}$$

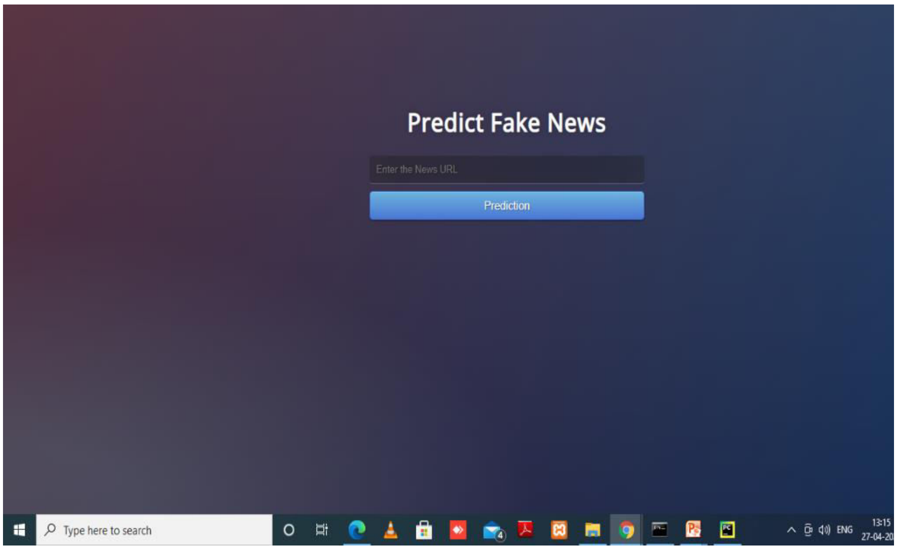
3. Recall:

It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.

$$Recall = \frac{TP}{TP + FN}$$

V – RESULT

This is the User Interface (UI) of our page where user will be prompted to check whether the news is Fake or Real.



User will have to enter the news URL in the box provided and then click on predict button. As soon as the user click the predict button, the result will be displayed under the button as “The news is Real” or “The news is Fake”.

VII- CONCLUSION

The concept of deception detection in social media is particularly new and there is ongoing research in hopes that scholars can find more accurate ways to detect false information in this booming, fake-news-infested domain. For this reason, this research could also be wont to help other researchers discover which combination of methods should be wont to accurately detect fake news in social media.

We must have some mechanism for detecting fake news, or at the very least, an awareness that not everything we read on social media may be true, so we always need to be thinking critically. This way we will help people make more informed decisions and that they won't be fooled into thinking what others want to control them into believing.

VII- BIBLIOGRAPHY

[1] A. N. K. Movanita, “BIN: 60 Persen Konten Media Sosial adalah Informasi Hoaks (BIN: 60 percent of social media content is hoax),” 2018. [Online]. Available: https://nasional.kompas.com/read/2018/03/15/ 06475551/bin-60- persen-konten-media- social-adalah-informasi-hoaks

http://www.sciencedirect.com/science/article/p ii/S0378437119317546
[2] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, and M. Akbar, “Fake news detection using deep learning models: A novel approach,” Transactions on Emerging Telecommunications Technologies, 2019.

detection within online social media using supervised artificial intelligence algorithms,”
Physica A: Statistical Mechanics and its
Applications, vol. 540, p. 123174, 2020. [Online]. Avail
[4] D. Pomerleau and D. Rao, “Fake News Challenge,” 2017. [Online]. Available: https://www.fakenewschallenge.org
[5] A. Thota, “Fake News Detection: Deep Learning approach,” SMU Data Science Review, vol. 1, no. 3, pp. 1–20, 2018.
[6] T. Shaikh, A. Anand, A. Ekbal, and P. Bhattacharyya, “A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11608 LNCS, pp. 345– 358, 2019.
[7] H. S. Nugraha and S. Suyanto, “Typographic-Based Data Augmentation to Improve a Question Retrieval in Short Dialogue System,” in 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISR), Dec 2019, pp. 44–49. [Online]. Available: https://ieeexplore.ieee.org/document/9034594
[8] T. Library, "Characteristics of Fake News & Media Bias," [Online]. Available: https://libguides.tru.ca/fakenews/characteristic s.
[9] N. S. SRIJAN KUMAR, "False Information on Web and Social Media: A Survey," p. 35, 2018.
[4] K. Shu, "Fake News Detection on Social Media: A Data Mining perspective," p. 15, 2017.
[10] The New York Times, "As Fake News Spreads Lies, More Readers Shrug at the Truth," [Online]. Available: https://www.nytimes.com/2016/12/06/us/fake news-partisan-republican-democrat.html. [Accessed 14 04 2018].