Building a Machine Learning-Enabled Fake News Detection Model

Presented By -
Vikas Jain
M.E. Software Engineering 2024-26

Motivation

1. The rise of fake news: The proliferation of social media and the internet has led to an increase in the spread of fake news and its potential consequences.

2. The need for solutions: The growing need for tools that can help detect and combat the spread of fake news, and the role that machine learning can play in addressing this issue.

3. The power of machine learning: Machine learning algorithms can be trained to accurately classify news articles as real or fake, and the benefits of using these algorithms over traditional rule-based approaches.

4. The importance of data: The challenges of collecting and preprocessing large datasets of news articles, and the importance of data quality in achieving accurate results.

Introduction

- Fake news is a growing issue, amplified by social media and the internet, with serious consequences on public opinion and politics.

- Machine learning can help detect and combat fake news, and this presentation outlines the key steps in developing a detection system: dataset collection, feature extraction, model training, and performance evaluation.

- Challenges include maintaining unbiased datasets and managing risks like false positives/negatives.

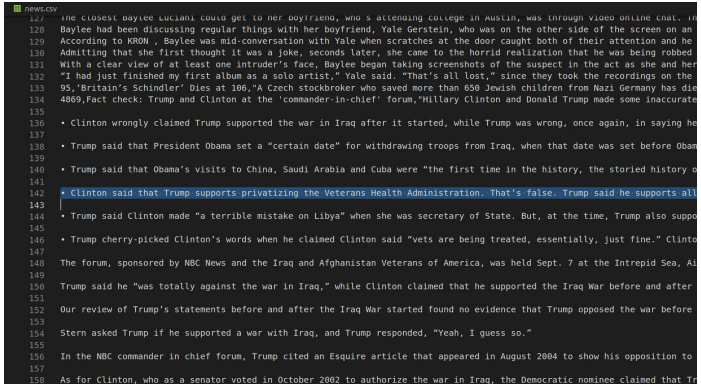
- The goal is to highlight effective techniques for fake news detection and its importance in combating misinformation.

Literature Review and Related Studies

- Various techniques have been proposed to detect fake news, including manual fact-checking[1], crowd-sourcing, and machine learning-based approaches.
- Machine learning-based approaches have shown promise in detecting fake news, as they can automatically learn patterns and features from large datasets of news articles.
- Commonly used machine learning algorithms for fake news detection include Support Vector Machines[2], Random Forests[3], and Naive Bayes.
- Feature extraction is a crucial step in fake news detection, as it involves identifying the most relevant features for classification, such as word frequencies[4] and n-grams.
- Feature selection techniques, such as chi-squared and mutual information, can be employed to identify the most discriminative features and reduce the dimensionality of the feature space.
- Several studies have evaluated the performance of machine learning-based fake news detection models using various metrics, such as accuracy, precision[5] and found that they can achieve high levels of accuracy.
- There is still room for improvement in fake news detection models, particularly in terms of their generalizability to new and unseen datasets, and their ability to detect more sophisticated forms of fake news, such as deep fakes.

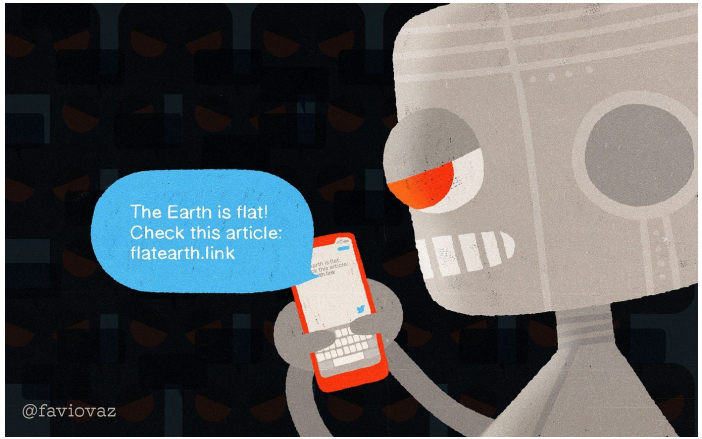
The Dataset

- Our dataset consist of news and facts more than 1,60,000+ News
- A large and diverse dataset is crucial for training and testing the fake news detection system.
- The dataset should consist of news articles labeled as either real or fake.
- The dataset should cover a variety of topics, including politics, science, entertainment, and sports.
- The dataset should include articles from reputable sources, as well as articles from known fake news sites.
- The dataset should be balanced, meaning there should be an equal number of real and fake articles.
- Data preprocessing techniques should be employed to clean and normalize the dataset.
- Feature extraction techniques should be used to identify relevant information within the articles.
- The dataset should be split into training and testing sets to evaluate the performance of the machine learning model.
- The dataset should be continuously updated and expanded to ensure the system's accuracy and relevance.
- Ethical considerations should be taken into account when collecting and using the dataset, including issues of privacy and bias.



Issues In Dataset

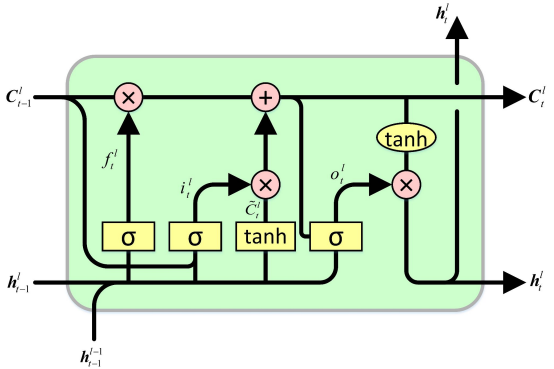
- Limited or Biased Sample
- Inconsistent Labeling
- Irrelevant Information
- Poor Quality Data
- Outdated Information
- Lack of Context
- Limited Coverage



Algorithms Proposed In The Machine Learning Model

1. LSTM

   Long short-term memory (LSTM) belongs to the complex areas of Deep Learning. It is not an easy task to get your head around LSTM. It deals with algorithms that try to mimic the human brain the way it operates and to uncover the underlying relationships in the given sequential data.



2. TF-IDF Vectorizer

Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF).

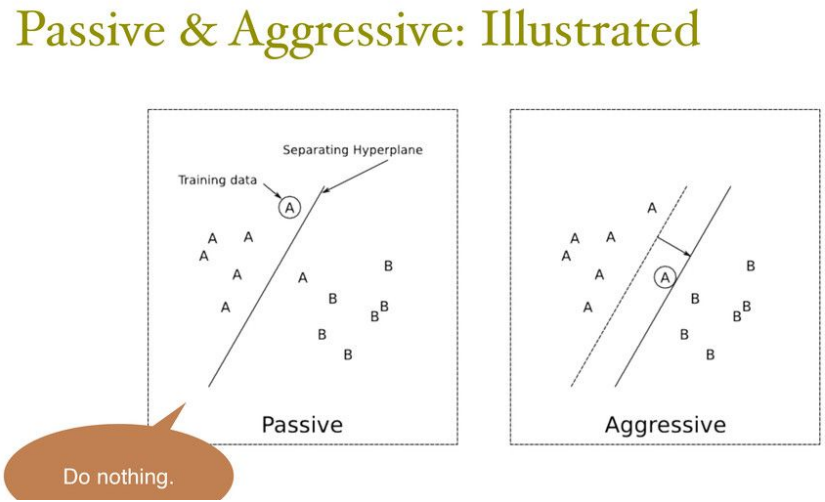
The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document. Term frequency represents every text from the data as a matrix whose rows are the number of documents and columns are the number of distinct terms throughout all documents.

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

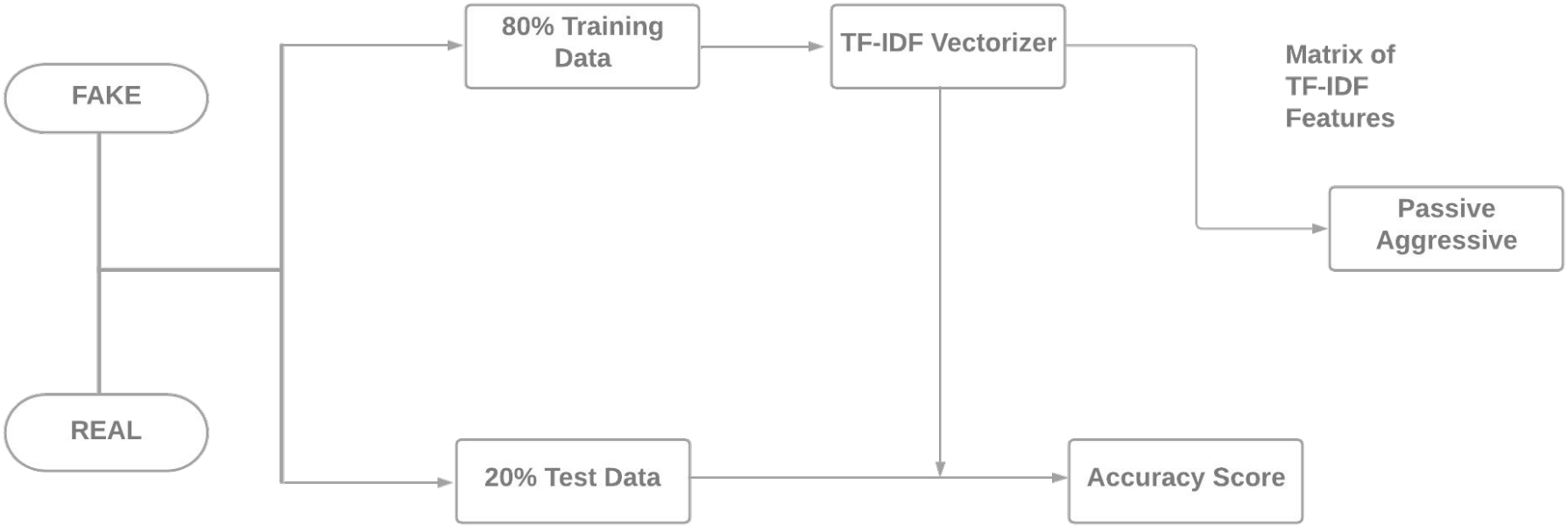
$$w_{i,j} = tf_{i,j} \times idf_i$$

3. 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.

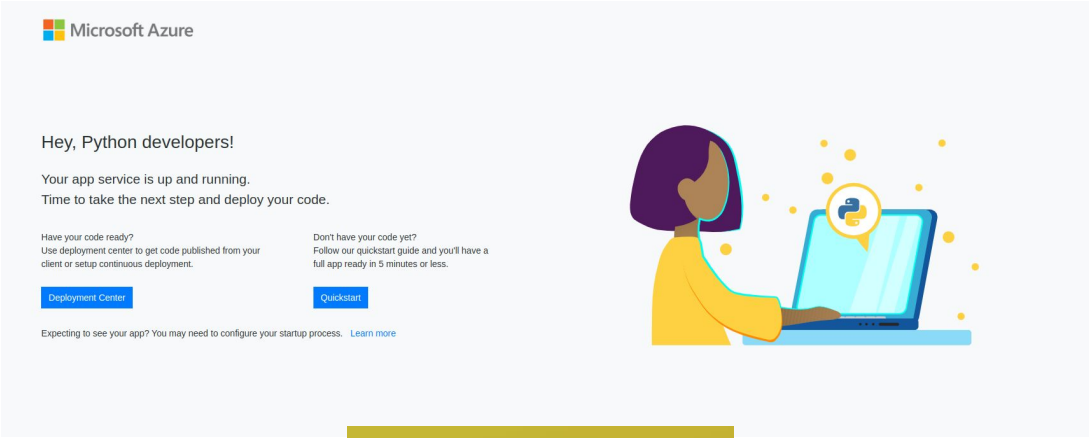


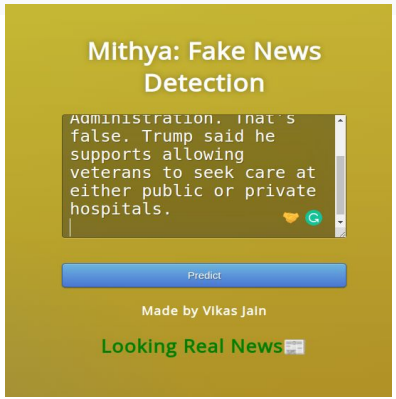
Solution Diagram



Results of Deployment

Cloud Deployment of Mithya



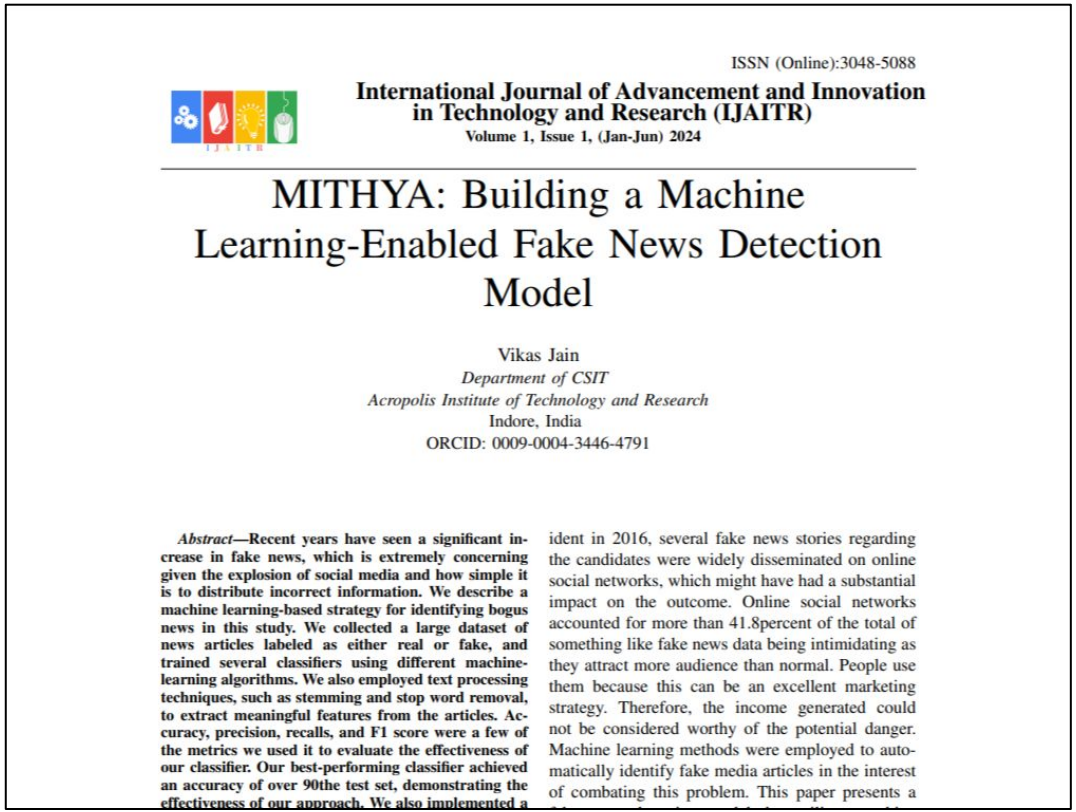


The Mithya app is also deployed over Microsoft Azure Cloud.

The Services used in the deployment of the Mithya model are:

Azure VMs
Azure App Services
Deployment center
Azure Data Lake services
Azure ML studio

Research Work



ISSN (Online):3048-5088

International Journal of Advancement and Innovation in Technology and Research (IJAITR)

Volume 1, Issue 1, (Jan-Jun) 2024

MITHYA: Building a Machine Learning-Enabled Fake News Detection Model

Vikas Jain

Department of CSIT

Acropolis Institute of Technology and Research

Indore, India

ORCID: 0009-0004-3446-4791

***Abstract*—Recent years have seen a significant increase in fake news, which is extremely concerning given the explosion of social media and how simple it is to distribute incorrect information. We describe a machine learning-based strategy for identifying bogus news in this study. We collected a large dataset of news articles labeled as either real or fake, and trained several classifiers using different machine-learning algorithms. We also employed text processing techniques, such as stemming and stop word removal, to extract meaningful features from the articles. Accuracy, precision, recalls, and F1 score were a few of the metrics we used it to evaluate the effectiveness of our classifier. Our best-performing classifier achieved an accuracy of over 90the test set, demonstrating the effectiveness of our approach. We also implemented a**

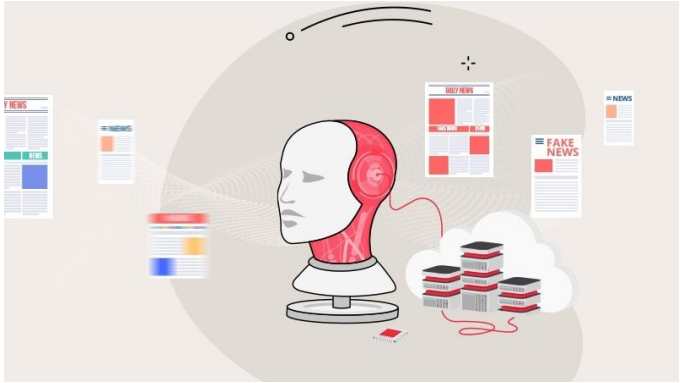
ident in 2016, several fake news stories regarding the candidates were widely disseminated on online social networks, which might have had a substantial impact on the outcome. Online social networks accounted for more than 41.8percent of the total of something like fake news data being intimidating as they attract more audience than normal. People use them because this can be an excellent marketing strategy. Therefore, the income generated could not be considered worthy of the potential danger. Machine learning methods were employed to automatically identify fake media articles in the interest of combating this problem. This paper presents a

Future Work

1. Text Classification Based Model such as Passive Aggressive Classifier are based on supervised Learning, which means these algorithms cannot be able to detect fake news beyond the dataset and also will not be able to learn by self.
2. RNN Models including GRU and Bi-GRU should be used to build the model with higher efficiency and accuracy for huge datasets.
3. The Model also needs continuous data to detect the real world news if it is fake or real.

Conclusion

Mithya can ring the initial alert for fake news. The Model produces worse results if the article is written cleverly, without any sensationalization. This is a very Complex problem but we tried to address it as much as we could. We believe the interface provides an easier way for the average person to check the authenticity of the news. Projects like this one with more advanced features should be integrated into social media to prevent the spread of fake news.



References

[1] c. Castillo, m. Mendoza, and b. Poblete. Predicting information credibility in time-sensitive social media. Internet research, 23(5):560–588, 2013.

[2] t. Chen, l. Wu, x. Li, j. Zhang, h. Yin, and y. Wang. Call attention to rumors: deep attention-based recurrent neural networks for early rumor detection. Arxiv preprint arxiv:1704.05973, 2017.

[3] y.-c. Chen, z.-y. Liu, and h.-y. Kao. Ikm at several-2017 task 8: convolutional neural networks for stance detection and rumor verification. Proceedings of removal. Acl, 2017.

[4] i. Augenstein, a. Vlachos, and k. Bontcheva. Usfd at several-2016 task 6: any-target stance detection on twitter with autoencoders. In semeval@naacl-hlt, pages 389–393, 2016.

[5]s. B. Yuxi pan, doug sibley. Talos. Http://blog.Talosintelligence. Com/2017/06/, 2017.

[6] b. S. Andreas hanselowski, avinesh pvs and f. Caspelherr. Team athene on the fake news challenge. 2017.

[7] g. Bonaccorso, "artificial intelligence – machine learning – data science," 10 06 2017.

[8] EANN: Event Adversarial Neural Networks for Multi-Modal

[9] Fake News Detection on Social Media: A Data Mining Perspective Kai Shuy, Amy Slivaz, Suhang Wangy, Jiliang Tang \, and Huan Liuy

[9] Automatic Deception Detection: Methods for Finding Fake News. Niall J.Conroy, Victoria L. Rubin, and Yimin Chen


Thank
YOU