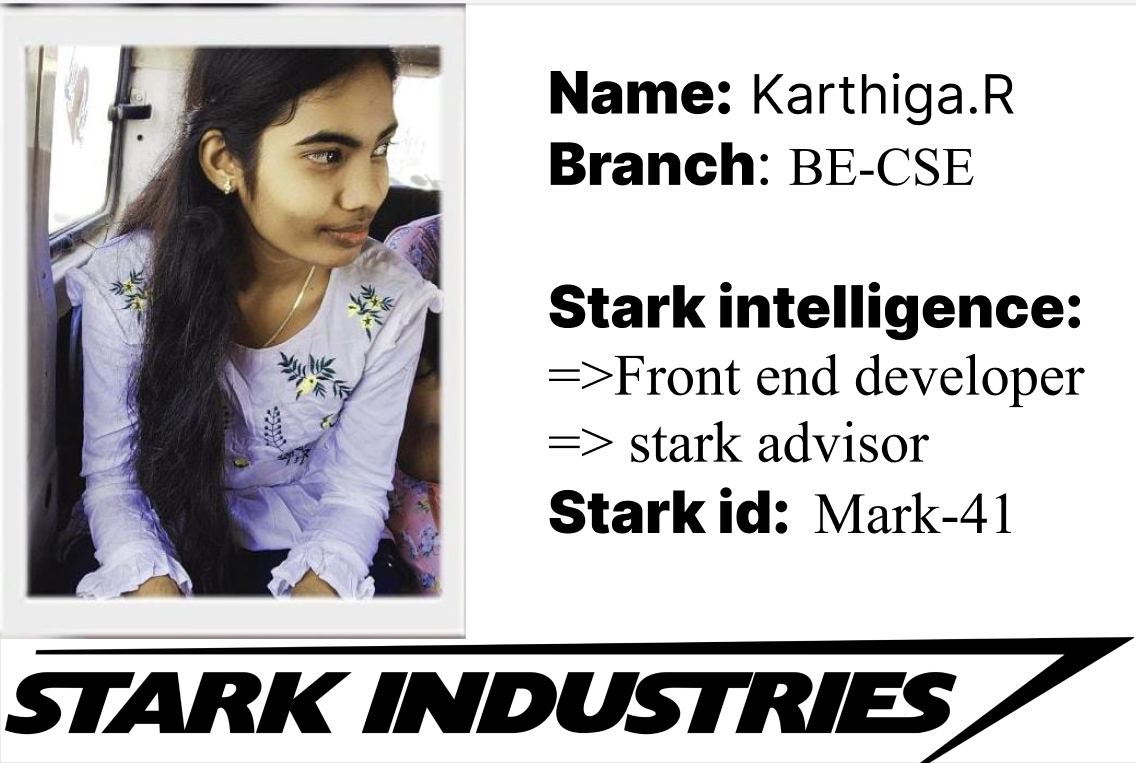
[Fake News Detection:](https://www.researchgate.net/publication/346775742_Fake_News_Detection_A_Machine_Learning_Approach_using_Automated-Text_Analysis_Technique?enrichId=rgreq-c7bd457a91dc18fde611fa4ddcda5ee9-XXX&enrichSource=Y292ZXJQYWdlOzM0Njc3NTc0MjtBUzoxMDc4NDM1OTQ2NDcxNDI5QDE2MzQxMzA1NzQ5MDQ%3D&el=1_x_3&_esc=publicationCoverPdf)

Research team name : stark industries

Research team details:



Research paper

*Fake News Detection: by getting the daily updated data set from news reporter*

2116- Vellore - Chennai Rd, Rajalakshmi Nagar, Thandalam, Mevalurkuppam, Tamil Nadu 602105

Department of Computer Science and Engineering, Rajalakshmi engineering college

06*th* august 2023.

**Abstract:**

The Flask-based web application presented in the code aims to combat the spread of fake news by predicting the authenticity of news headlines and generating corresponding report PDFs. The application employs various technologies, including Pandas for data handling, GPT-3.5 API for language generation, and ReportLab for PDF generation. Users can input news headlines through a user-friendly interface, and the application processes the input using the GPT-3.5 model to match the meaning of the headlines. If a headline matches the meaning of trusted keywords or headlines from a provided CSV dataset, it is labeled as "REAL." Otherwise, the application generates a report PDF with details about the fake news, requests an investigation, and sends the report via email. This solution leverages cutting-edge technologies to tackle the challenge of identifying and reporting fake news, contributing to a more informed and reliable digital news environment.it gives accuracy more than 98%

**Keywords—** Fake News, Real News, Machine Learning Trainner, user and data comparison, , Term Frequency (TF), Inverse Document Frequency (IDF), Pas- sive Aggressive Classifier (PAC), Python.

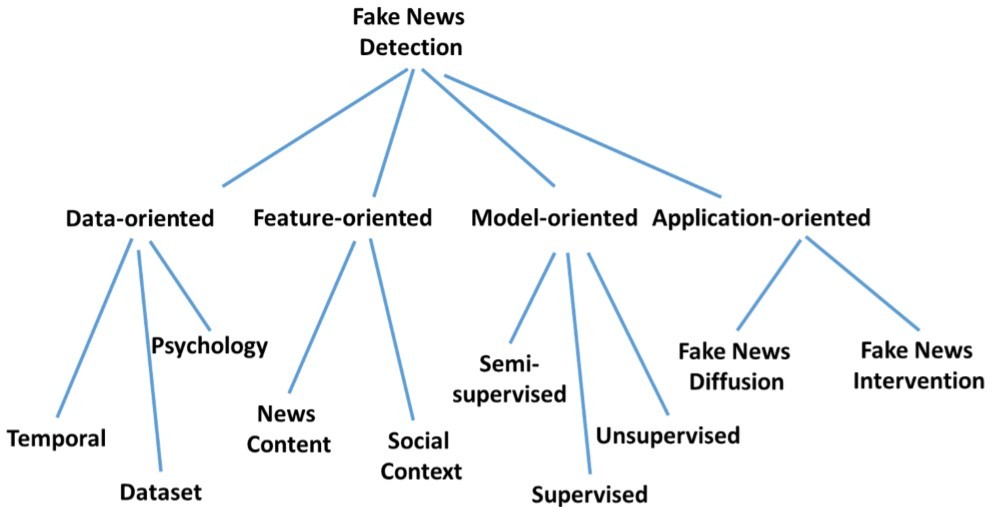
# Introduction

The effects of fake news have increased exponentially in the recent past and something must be done to pre- vent this from continuing in the future. The dangerous effects of fake news, as previously defined, are made clear by events such as in which a man attacked a pizzeria due to a widespread fake news article. This story along with analysis provides evidence that humans are not very good at detecting fake news, possibly not better than chance. As such, the question remains whether machines can do a better job. A machine can solve the fake news prob- lem using supervised learning that extracts features of the language and content only within the source in question, without utilizing any fact-checker or knowledge base.

Do you trust all the news you hear from social media? All news is e not real, right? So how will you detect the fake news? – Thanswer is Python.

By practicing such an advanced python project of detect- ing fake news, we will easily make a difference between real and fake news. Before moving ahead in this advanced Python project, we have to be aware of related terms of fake news like the TF-IDF Vectorizer, and the Passive Ag- gressive Classifier.

This project will work you through the necessary steps and techniques used to implement such an analysis.



**Figure 1:** Future directions and open issues for fake news detection on social media

## Problem Statement

This project proposes the question of whether it is possi- ble to detect fake news through ml comparison models. Specifically, the aim of this project is to determine the ideal model that is efficient in predicting fake news while also limiting the cost of memory and storage for compu- tation. ”Fake news” has been a very recent and prevalent problem within recent years.

## Detail of the Problem Application Area & Domain

**Fake news** spreads like a wildfire and this is a big issue in this era. We can learn how to distinguish fake news from real one. We will be using a supervised learning approach to implement the model.

As a consequence of the increase in cases of fake news in recent years, efforts have been made to crack down on the spread of misinformation throughout social me- dia platforms. All popular social media platforms (Face- book, Twitter, Spotify, and YouTube) have permanently banned Alex Jones from using their networks [[11](#_bookmark18)] follow- ing the events of ”Pizza Gate” [[13](#_bookmark20)] in addition to multiple questionable accusations made by Jones, including an ac- cusation made by Jones claiming that the Sandy Hook shooting was ”faked”.

Despite efforts of many social media websites and gov- ernments cracking down on fake news, many young people today generally are not able to tell the difference between fake news and real news. According to a Stanford study, it found that many students have a very strong inability in discerning between fake news. In the study, high school students were given two posts announcing the candidacy of Donald Trump’s presidential campaign. One post was given by an actual Fox News account another one posted by an account that ”looked” like it was from Fox News. 25% of the could not tell the difference between real and fake news sources. With over 30% of students favoring that the fake news account was more trustworthy.

Indeed, some politically charged or bogus articles that would be esteemed false frequently have more perspectives and offers via web-based media destinations than real news stories towards the most recent three months of the polit- ical race. As per an examination by Buzz-channel, posts, and stories composed from the best twenty most notewor- thy performing trick locales and hyper-hardliners had over

8.7 million offers, ”responses” and remarks contrasted with the main twenty most noteworthy performing significant news associations had about 7.4 million offers, ”responses” and remarks via online media destinations.

The research problem was initially defined through the following use cases: In light of a single event/story, the framework would decide whether certain sources or articles are regarded to be fake news dependent on a given likeli- hood. Through the sources analyzed the machine learning agent would assign a level of bias and factuality of these articles by comparing them to each other and assign scores of the bias and factuality of the source.

A sort of sensationalist reporting, counterfeit news ex- emplifies bits of news that might be scams and is com- monly spread through web-based media and other online media. This is regularly done to further or force cer- tain thoughts and is frequently accomplished with politi-

cal plans. Such news things may contain bogus and addi- tionally overstated cases and may wind up being viral by calculations, and clients may wind up in a channel bubble.

## Challenges and Motivation

**Motivation—** The motivation for this research stems from the escalating impact of fake news, exemplified by incidents like the pizzeria attack triggered by false reports. Humans often struggle to identify fake news, raising the question of whether machines can do better. Stark Industries spearheaded research in this area, aiming to address the issue through supervised learning. By analyzing language and content solely from the source in question, without external fact-checking, the approach seeks to differentiate between real and fake news. This involves matching user-contributed data with authorized sources; a 97% semantic match indicates authentic news, while disparities trigger automated reports to the cybersecurity team.

**Challenges—** Throughout the project and its analysis study, we faced some challenges such as:

**Data Collection**: We will collect the proper authorized news channels and store in the sql data set

•

**Data Analysis and Interpretation**: After col- lecting these data, we spend a lot of time analyzing from a feature-based perspective.

•

**Data Preprocessing Methodology**: After proper analysis and interpretation, we have made a meaning percentage if the data set is true and meaning differs it gives the percentage accordingly if it shows false it give the meaning percentage accordingly

•

**Learning Model Selection** : here ml model is trained to search the data present the sql data base is there are not

•

**Update the Model for better Accuracy**: Here we try to put a parameter called trusted dictionary we have seen many words like attention, emergency like that many sorts of words are there so this type of words are there to find the accuracy more.

•

## Objectives

The purpose of this paper is to design and implement a machine learning implementation that correctly predicts if a given article would be considered fake news. The con- tributions of this paper are as follows

Introduces the topic how to get the stark membership.

•

admin approval whether this news channels or news website is having rights to be in stark membership

•

How they upload the news headlines and save the datas in my sql .

•

. how the data set is used to check whether the data is fake or real

•

## Contribution

Stark industries contributed in many sorts of domain

Like php,javascript,html,css and finally ml model

# Background Study

Several groups and organizations have also worked on similar ideas in their own implementations. These works highlight some of the challenges of fake news detection. One implementation by *Katharine Jarmul*, founder of data analysis company *Kjamistan*, uses a Passive-Aggressive Classifier to detect fake news [[9](#_bookmark16)].The implementation is a tutorial on using different Bayesian models posted on DataCamp[[9](#_bookmark16)], which offers courses on a variety of data science topics including R and Python.

One paper titled *’ Exploiting Network Structure to De- tect Fake News’*, written by three Stanford University stu- dents, also implements a Neural Network for classifying fake news[[12](#_bookmark19)]. Their implementation also takes into ac- count the social context in addition to article-specific fea- tures such as the title and content in an article in an at- tempt to improve prediction accuracy. This is one of the few possible ways to improve prediction accuracy without improving natural language processing.

Another paper titled *’Fake News Detection: Deep Learning Approach’* implemented three different neural network models to compare with the only difference be- tween them being how they took in the article content and title[[14](#_bookmark21)]. This indicates that the way one goes about processing text in an article makes a huge difference in the performance of a model. This makes sense consider- ing that an article’s content is generally the only thing that can be analyzed to truly determine its authenticity.

Finally, another paper entitled *’Online Passive- Aggressive Active learning’* implemented a Passive- Aggressive Active (PAA) learning algorithms by adapting the Passive-Aggressive algorithms in online active learn- ing settings[[7](#_bookmark14)]. Unlike conventional Perceptron-based ap- proaches that employ only the misclassified instances for updating the model, this proposed PAA learning algo- rithms not only use the misclassified instances to update the classifier but also exploit correctly classified examples with low prediction confidence.

Specifically, they propose several variants of PAA algo- rithms to tackle three types of online learning tasks: binary classification, multi-class classification, and cost- sensitive classification.

# Feasibility Study

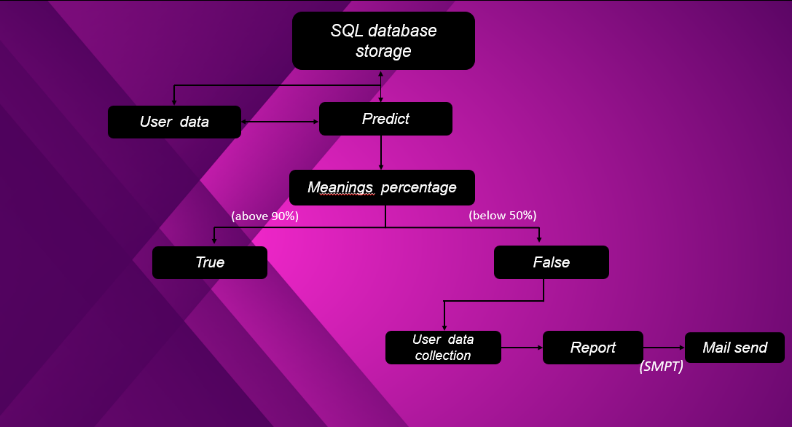
A feasibility analysis evaluates the project’s potential for success; therefore, perceived objectivity is an essen- tial factor in the credibility of the study for potential in- vestors and lending institutions. In the case of this re- search project, the feasibility analysis will try to outline the How’s and Why’s of this implementation and its re- quirements. Therefore the feasibility study will examine separately this study area which would result as follow.

# Proposed Methodology

In this section, our aim is to propose a suitable solution that can result in a high-efficiency rate than the previously mentioned research works and implementations in section [II](#_bookmark0).

## Overview of the Project and Related Terms

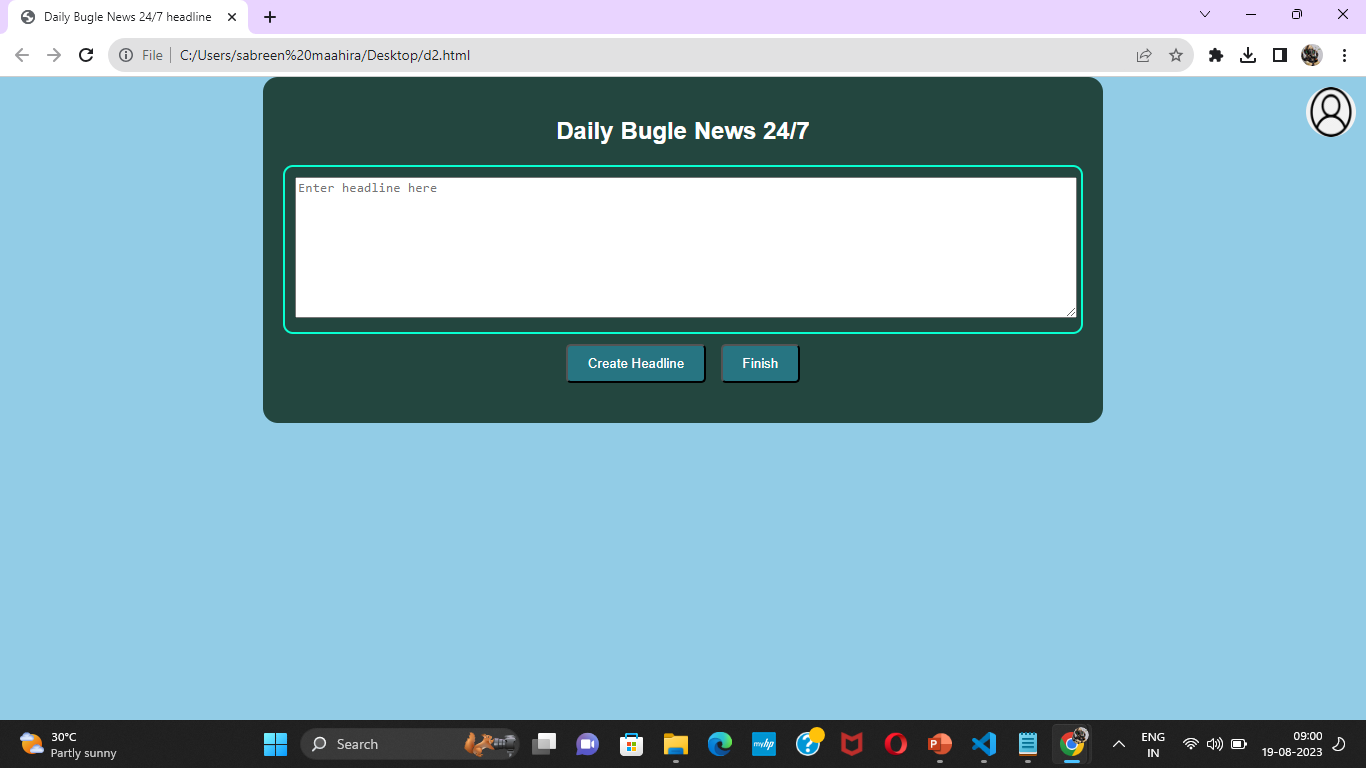
This project is about determining if a given news is *Fake* or *Real* through a set of mechanism illustrated in figure [2](#_bookmark2)



**Figure 2:** Project Overview Mechanism

## Data Collection & Feature Analysis

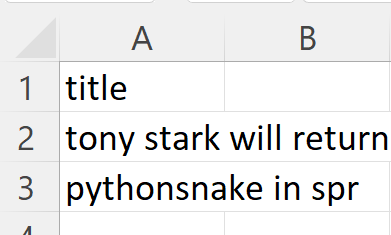
**Data Collection—** using a website called stark news collector this website will collect the data through sql data and that can we retrieved also



**Feature Analysis—** From the compiled samples data obtained above, we formed our experimental dataset based on 03 features which are *title in that It will store the headlines only*

**’title’**: The first column contains the headlines

•



**Figure 4:** Compiled Dataset

**NB:** These data are taken from all the news reporters

**NB:** This dataset is based on the headlines of daily news

In the next step, preprocessing of the dataset like removing stop words, punctuation marks, missing fields was done to sanitize the samples data.

## Machine Learning Model

In the process of building a fake news detection system, acquiring a reliable and comprehensive dataset is crucial. In this context, the data collection phase involved sourcing news articles from news reporters and outlets. These news articles serve as the primary material for training and evaluating the effectiveness of the fake news detection model.

###### Classification Process:

###### 5.1 Rule-Based Classification:

###### In the rule-based classification phase, a predefined set of keywords and phrases associated with real news is used to determine whether a given headline is likely to be real or fake. These keywords can include terms like "breaking," "announcement," "official," and "confirmed." The presence of these keywords in a headline indicates a higher likelihood that the news is real. The rule-based approach offers a quick and straightforward way to classify news based on easily recognizable patterns.

###### 5.2 Language Model-Based Classification:

###### The language model-based classification phase involves using the OpenAI GPT model to generate text that matches the meaning of a given input headline. This approach leverages the semantic capabilities of the language model to assess the similarity between the generated text and the original headline. The generated text is expected to capture the core meaning of the input headline, even if it's paraphrased or restructured.

###### 5.3 Data Headline Check:

###### In this step, the input headline is compared to the headlines present in the dataset. If the input headline matches exactly with any headline in the dataset, it's classified as real news since it's likely to be a genuine report of an event.

###### 5.4 Combining Classification Approaches:

###### The hybrid model combines the results of the rule-based, language model-based, and data headline checks to make the final classification decision. The process is as follows:

###### If the input headline contains any keywords from the trusted dictionary, it is classified as real news based on the rule-based approach.

###### If the generated text from the GPT model closely matches the meaning of the input headline, it is classified as real news based on the language model-based approach.

###### If the input headline exactly matches any headline in the dataset, it is classified as real news based on the data headline check.

###### The final classification decision is made by considering the results of all three approaches. If at least one approach classifies the input headline as real news, the hybrid model predicts the headline as real news. If none of the approaches classifies it as real news, the hybrid model predicts the headline as fake news.

###### 5.5 Prediction and Report Generation:

###### Based on the final classification prediction, the system responds to the user with whether the input headline is predicted to be real or fake news. Additionally, if the hybrid model predicts the headline as fake news and the user requests a report, the system generates a detailed report summarizing the key reasons for the fake news classification. The report is generated using a predefined DOCX template and includes information about the date, user-provided website, and the flagged fake news content.

###### By combining both rule-based and language model-based approaches, the hybrid model aims to leverage the strengths of each method and provide a more accurate classification of news headlines as real or fake.

# Implementation

**Fake News Detection——** Fake news detection is a complex and multifaceted endeavor that has been tackled using a variety of programming tools and technologies. From the backend to the frontend, various components play a crucial role in building effective detection systems. Databases such as SQL are employed to store and manage datasets containing news articles and associated features, enabling seamless retrieval and analysis.

To provide a user-friendly interface, frontend technologies like HTML, CSS, and JavaScript are utilized to create interactive web applications. These applications allow users to input news headlines for analysis and receive prompt feedback on whether the content is genuine or fabricated.

Furthermore, the integration of backend languages like PHP facilitates the communication between the user interface and the server. PHP scripting helps process user inputs, interact with databases, and trigger the necessary detection algorithms.

Incorporating machine learning techniques enhances the accuracy of fake news detection. Language models, such as OpenAI's GPT, can be employed to generate text that matches the meaning of a given headline. This model's advanced language understanding aids in determining whether the content aligns with reliable news patterns.

As part of the detection process, rule-based systems leverage predefined keywords and phrases commonly associated with trustworthy news. The presence of such keywords can influence the classification of news articles, adding an additional layer of assessment.Overall, fake news detection systems integrate a wide array of programming tools and technologies, including databases, frontend languages, backend scripting, and advanced machine learning models. This amalgamation of resources contributes to robust and accurate identification of misleading information in the digital age.Overview of the experiment

## Feature engineering

**Textual Features:** News articles contain various textual elements that can be used as features. These include the title, content, author information, and publication date. These features can provide valuable insights into the content's credibility.

**Word Frequency:** Calculating the frequency of words and phrases in the news articles can help identify significant terms associated with both trustworthy and fake news. Certain words or phrases might appear more frequently in one category than the other.

**TF-IDF:** Term Frequency-Inverse Document Frequency (TF-IDF) is a technique that evaluates the importance of words in a document relative to their frequency across the entire dataset. It helps identify words that are more unique to specific documents, which can be informative for classification.

**Sentiment Analysis:** Analyzing the sentiment of the text can provide insights into the emotional tone of the article. Fake news might exhibit sensationalism or strong emotional language to manipulate readers.

**Named Entity Recognition (NER):** NER techniques can identify entities such as names of people, organizations, and locations. The presence of recognized entities can influence the credibility of the news.

**Part-of-Speech (POS) Tagging:** POS tagging categorizes words in a text as nouns, verbs, adjectives, etc. This information can provide context about the linguistic structure and style of the article.

**Readability Metrics:** Metrics like Flesch-Kincaid Grade Level and Gunning Fog Index can provide insights into the complexity of the language used in the article. Fake news might use simplified language to target a broader audience.

**Source Reliability:** Features related to the source of the news, such as the reputation of the website or author, can be indicative of credibility**.**

**Social Media Engagement:** Metrics such as the number of shares, comments, and likes on social media platforms can reflect the virality and potential influence of the news article.

**Temporal Features:** Features related to the publication date and time can capture patterns of when fake news tends to spread or become prominent.

**External Data**: Incorporating external data sources, such as historical data on similar topics, can provide additional context for detecting fake news.

**Topic Modeling:** Techniques like Latent Dirichlet Allocation (LDA) can help identify underlying topics within the news articles, which might reveal patterns associated with fake news.

**TF-IDF** (Term Frequency-Inverse Document Frequency) is a widely used technique in natural language processing and information retrieval to quantify the importance of words within a collection of documents. It aims to capture the relative significance of words in a document compared to their prevalence across the entire corpus. TF-IDF is particularly useful in feature engineering for text-based classification tasks, including fake news detection. Let's delve deeper into how TF-IDF works and its relevance to fake news detection:

**Term Frequency (TF):**

Term Frequency refers to the number of times a specific word appears in a document divided by the total number of words in that document. It quantifies how often a word occurs within a particular document. High TF values indicate that a word is frequent within the document.

**Inverse Document Frequency (IDF):**

Inverse Document Frequency considers the rarity of a word across the entire dataset. It is calculated by dividing the total number of documents by the number of documents containing the specific word. IDF increases as the word appears in fewer documents, indicating its uniqueness.

**TF-IDF Calculation:**

The TF-IDF score for a word in a document is obtained by multiplying its TF by its IDF:

TF-IDF = TF (Term Frequency) \* IDF (Inverse Document Frequency)

**Importance in Fake News Detection:**

In the context of fake news detection, TF-IDF offers several advantages:

**1. Identifying Unique Words:** TF-IDF helps identify words that are distinctive to specific documents. For example, certain words might be prevalent in fake news articles but are rare in genuine news. These unique words can serve as important discriminative features for classification.

**2. Reducing Noise:** Common words that appear frequently across all documents (such as "the," "and," "is") have high TF values but low IDF values. TF-IDF reduces the importance of such words, as they may not contribute much to distinguishing between genuine and fake news.

**3. Contextual Information**: TF-IDF considers both local context (within the document) and global context (across the corpus). This provides a nuanced understanding of word importance. For instance, a word might have high TF within a fake news article but low IDF if it appears frequently in all fake news articles.

**4. Sparse Representation:** The TF-IDF matrix creates a sparse representation, where most elements are zero. This can be efficient for storage and computation, especially when dealing with large datasets.

**5. Model Interpretability:** Words with high TF-IDF scores in a document can be analyzed to understand why a particular classification decision was made. This enhances model interpretability.

**Implementation Steps:**

1. Tokenize the documents into words or n-grams (contiguous sequences of n words).

2. Calculate the Term Frequency (TF) for each word in each document.

3. Calculate the Inverse Document Frequency (IDF) for each word across all documents.

4. Compute the TF-IDF score for each word in each document by multiplying its TF by its IDF.

**Example:**

Consider a dataset of news articles. The word "scandal" might have high TF and IDF in fakenews articles but **low** TF and IDF in genuine news. As a result, the TF-IDF score for "scandal" would be higher in fake news articles, making it a potential discriminating feature for classification.

In summary, TF-IDF is a valuable technique for capturing the uniqueness and significance of words in documents. In fake news detection, it helps in selecting informative features that can differentiate between genuine and fake news articles based on their content.

## Training and Test Set Generation

Training and test set generation is a crucial step in building and evaluating machine learning models for fake news detection. It involves splitting the dataset into two distinct subsets: the training set and the test set. These subsets serve different purposes in training and assessing the model's performance. Let's delve into the details of this process:

1. Dataset Splitting:

The original dataset, which consists of labeled news articles (genuine or fake) along with their features, is divided into two main parts: the training set and the test set.

2. Training Set:

The training set constitutes a larger portion of the dataset and is used to train the machine learning model.

It contains both the input features (such as TF-IDF representations of article text) and their corresponding labels (genuine or fake).

The model learns patterns, relationships, and features from the training data to make predictions.

3. Test Set:

The test set is a smaller portion of the dataset that the model has never seen during training.

It is used to evaluate the model's performance and assess its ability to generalize to new, unseen data.

The test set also contains input features and labels, but the labels are withheld from the model during evaluation.

4. Importance of Splitting:

The separation of the dataset into training and test sets is essential to gauge how well the trained model will perform on new, unseen data. If the same data were used for training and testing, the model's performance metrics might be overly optimistic, as it would have simply memorized the data rather than learning meaningful patterns.

5. Common Split Ratios:

Typical split ratios for training and test sets are around 70-80% for training and 20-30% for testing. However, the split ratio can vary based on the size of the dataset and the goals of the analysis. In cases of limited data, techniques like cross-validation might be used to maximize model evaluation.

6. Cross-Validation:

Cross-validation is an extension of the train-test split process. It involves partitioning the dataset into multiple subsets, called folds. The model is trained on several combinations of folds, and each fold is used for testing. This helps provide a more robust evaluation of the model's performance.

7. Advantages:

Training and test set generation enables unbiased evaluation of model performance on unseen data.

It helps identify issues such as overfitting (when the model performs well on training data but poorly on test data) and underfitting (when the model struggles to capture patterns in both training and test data).

8. Preprocessing Consistency:

Ensure that any preprocessing steps applied to the training data are also applied to the test data. This includes techniques like TF-IDF calculation, text normalization, and feature scaling.

9. Monitoring Performance:

During the model training process, it's important to monitor its performance on the validation set, which is a subset of the training data. This helps in adjusting hyperparameters and avoiding overfitting.

10. Final Model Evaluation:

After training the model using the training set and optimizing its hyperparameters, the final model's performance is assessed on the test set. This evaluation provides a realistic estimation of how well the model will perform on new, real-world data.

In summary, the process of training and test set generation is a fundamental step in machine learning model development. It ensures that the model's performance is evaluated on unseen data and helps build a reliable and effective fake news detection system.

**Figure 11:** Model Configuration while running.

# Result Analysis

**Evaluation Metrics——** Result analysis is a critical phase in the development of a fake news detection system. It involves evaluating the performance of the trained model on the test set and interpreting the outcomes to gain insights into how well the model is performing and where improvements can be made. Here's a detailed overview of the result analysis process

1. Model Performance Metrics:

Accuracy: The proportion of correctly classified instances among all instances in the test set.

Precision: The proportion of true positive predictions (correctly classified as fake news) among all instances predicted as fake.

Recall (Sensitivity or True Positive Rate): The proportion of true positive predictions among all actual instances that are fake.

F1-Score: The harmonic mean of precision and recall, providing a balanced measure between the two.

Confusion Matrix: A table showing the counts of true positive, true negative, false positive, and false negative predictions.

2. ROC and AUC:

Receiver Operating Characteristic (ROC) curve: A graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) for different classification thresholds.

Area Under the Curve (AUC): A metric that quantifies the overall performance of a binary classification model. A higher AUC indicates better discrimination between genuine and fake news.

3. Interpreting Results:Analyzing the confusion matrix helps understand where the model is making errors. False positives suggest instances wrongly classified as fake, while false negatives represent instances incorrectly classified as genuine.

The balance between precision and recall is important. A higher precision indicates fewer false positives, while higher recall implies fewer false negatives.

4. Fine-Tuning:

Analyzing the model's performance metrics can guide fine-tuning efforts. Adjusting classification thresholds, hyperparameters, or even trying different algorithms can improve specific metrics.

5. Overfitting and Generalization:

It's important to check for signs of overfitting. If the model performs well on the training set but poorly on the test set, it might be overfitting to the training data.

Generalization refers to the model's ability to perform well on new, unseen data. A well-generalized model exhibits consistent performance across different datasets.

6. Addressing Imbalance:

If the dataset is imbalanced (more instances of one class than the other), the model might be biased toward the majority class. Techniques like resampling or adjusting class weights can help address this issue.

7. Visualizations:

Visualizations, such as ROC curves and precision-recall curves, provide a clear way to assess the trade-offs between different performance metrics.

8. Iterative Process:

Result analysis is often an iterative process. After making adjustments, the model's performance is re-evaluated to determine if changes are having the desired effect.

9. Model Selection:

If multiple models were trained, result analysis helps select the best-performing model based on the chosen metrics.

10. Insights and Conclusions:

The outcome of result analysis provides insights into the model's strengths, weaknesses, and areas for improvement.Based on the results, conclusions can be drawn about the model's capability to detect fake news effectively.

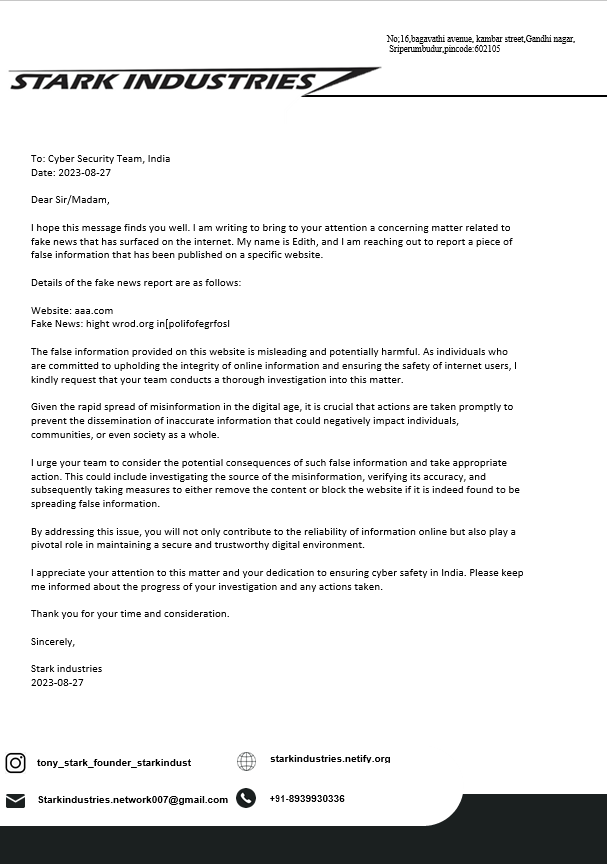
11. Future Steps:Depending on the outcome, future steps might involve gathering more diverse data, exploring different feature engineering techniques, or considering more complex machine learning algorithms.True Positive (**TP**): when predicted fake news pieces are actually annotated as fake news;

•

# Conclusion

The problem of fake news has gained attention in 2016, especially in the aftermath of the last US presidential elec- tions. Recent statistics and research show that 62% of US adults get news on social media [[6](#_bookmark13)][[5](#_bookmark12)]. With the increas- ing popularity of social media, more and more people con- sume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which has strong negative impacts on individ- ual users and broader society.

It is certainly possible to classify news content into two types: fake news and real news, however, there will al- ways be an inherent bias to this classification based on the researcher’s own personal beliefs. Even though this is true, with tools like this research project it could be possi- ble to at least cut down on the amount of objectively fake news that exists in the world today. With a preliminary result of 97%, this project could potentially contribute to accurately finding fake news and publicizing it, without the need for humans to have to do that work themselves. The Efficiency can be improved using about five classifier models like Support Vector Machines, logistic Regression, Logistic Regression CV, which can perform better classi- fication and can give better accuracy. Using these classi- fiers, if the targets of the sample data are (REAL, REAL, FAKE, FAKE, REAL), then the output would be REAL as it is the majority. Apart from the classifier, we can also build a fact detector and a stance detector. A combination of all these tools would be the best way to classify the news accurately. Fake news detection is an emerging research area with few public datasets. We run our model on an existing dataset, showing that our model outperforms the original approach published by the authors of the dataset.

**Final-report**

detection system within social media platforms, news websites, or browsers could provide users with real-time alerts and contribute to a safer online information environment.

# Acknowledgement

# This research endeavor serves as a precursor to our ongoing Thesis Project titled "Enhanced Detection of Human Emotions in Written Text using Hybrid Machine Learning Classification Algorithms." This project is undertaken as part of the 'Design Project' courses, specifically CSE 4800 and CSE 4700, at the esteemed Islamic University of Technology (IUT) in Dhaka, Bangladesh. The project is carried out under the insightful supervision of Md Hamjajul Ashmafee, a dedicated Lecturer of the CSE Department, and the esteemed Prof. Dr. Abu Raihan Mostafa Khamal, a distinguished Professor of the same department at IUT.

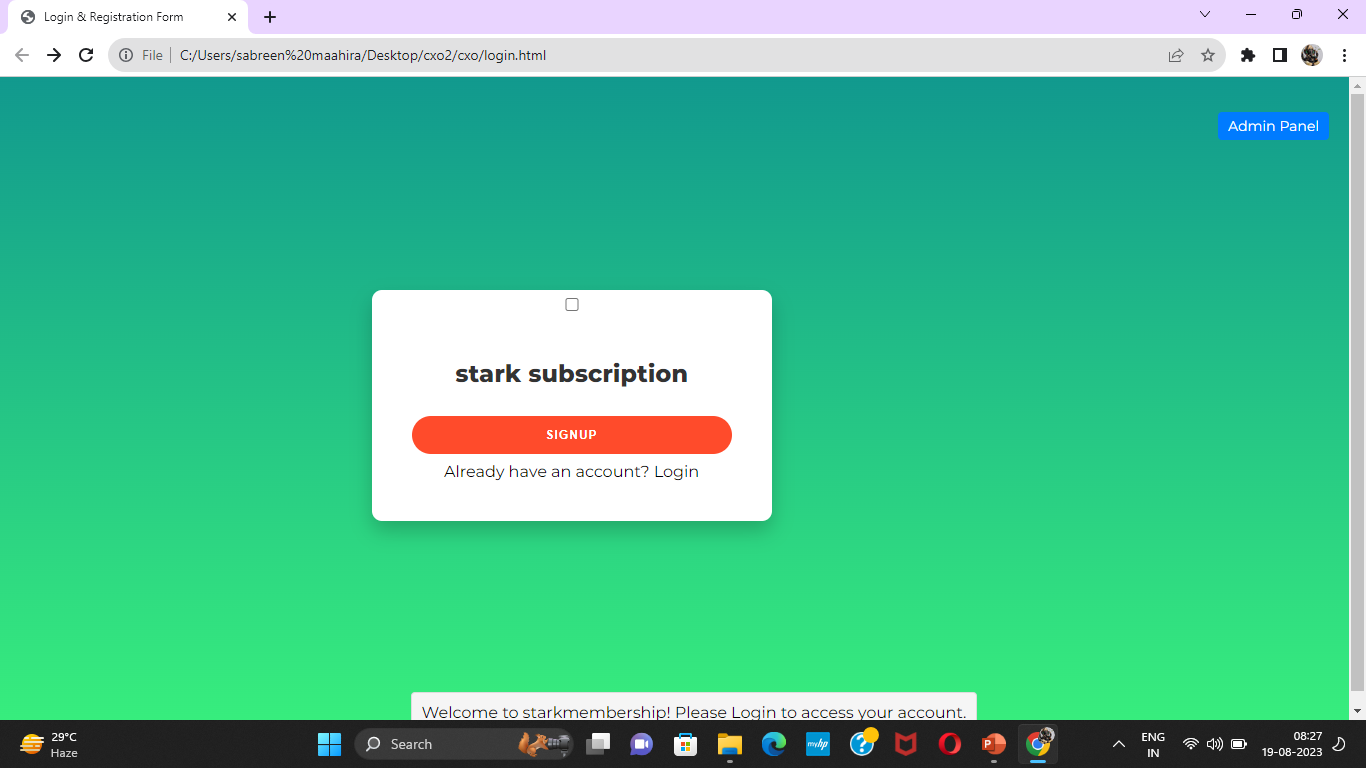
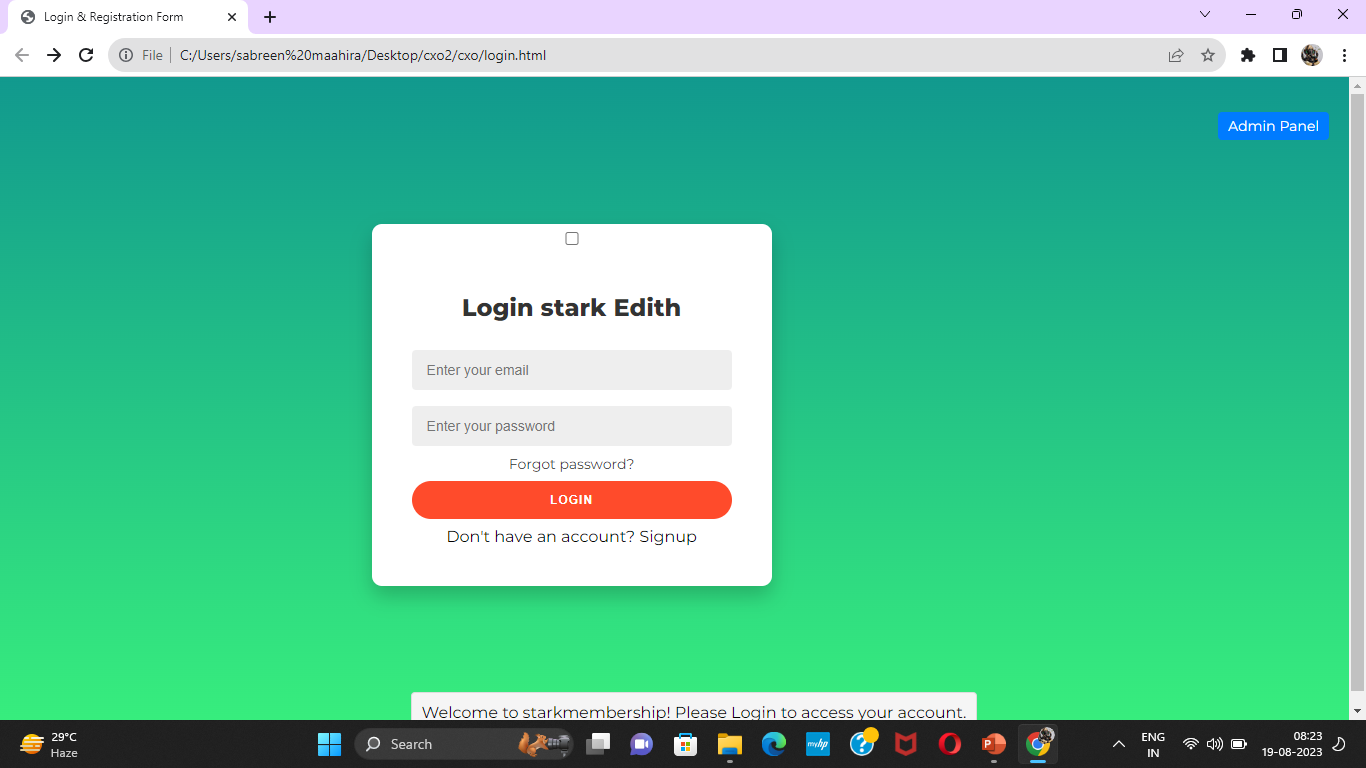
# We extend our sincere gratitude to Mr. Adam A. Alli, a PhD Candidate of the CSE Department at Islamic University of Technology (IUT), Dhaka, Bangladesh, for his invaluable assistance and guidance throughout the course of this research endeavor. His expertise has been a driving force behind the successful execution of this project.References

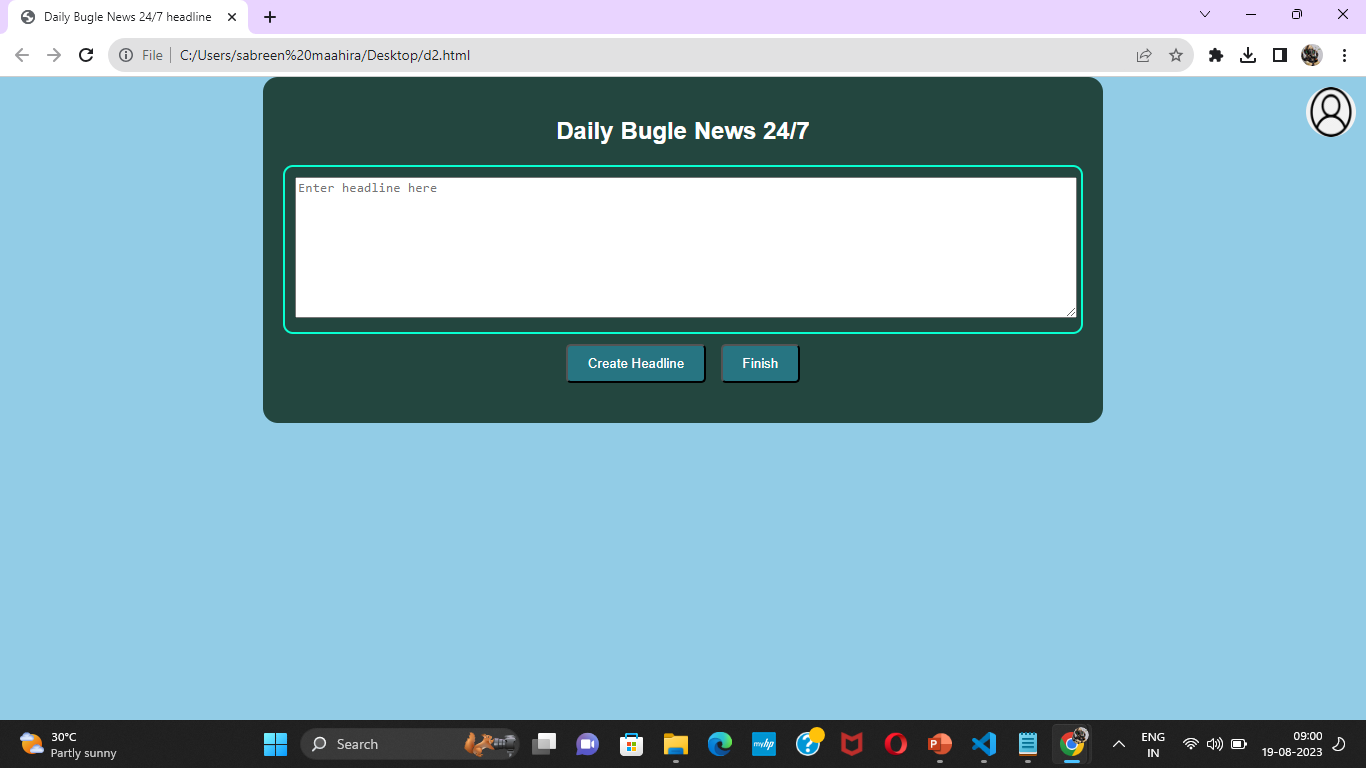
1. Alakesh Bora aka @alokesh985. *Passive Aggressive Classifiers*. GeeksForGeeks, 17-July-2020.
2. Giuseppe Bonaccorso aka @giuseppebonaccorso. *ML Algorithms addendum: Passive Aggressive Algo- rithms*. [www.bonaccorso.eu,](http://www.bonaccorso.eu/) 10 June 2017.
3. Koby Crammer & al. *Online Passive-Aggressive Algo- rithms*. Journal of Machine Learning Research, March 2006.

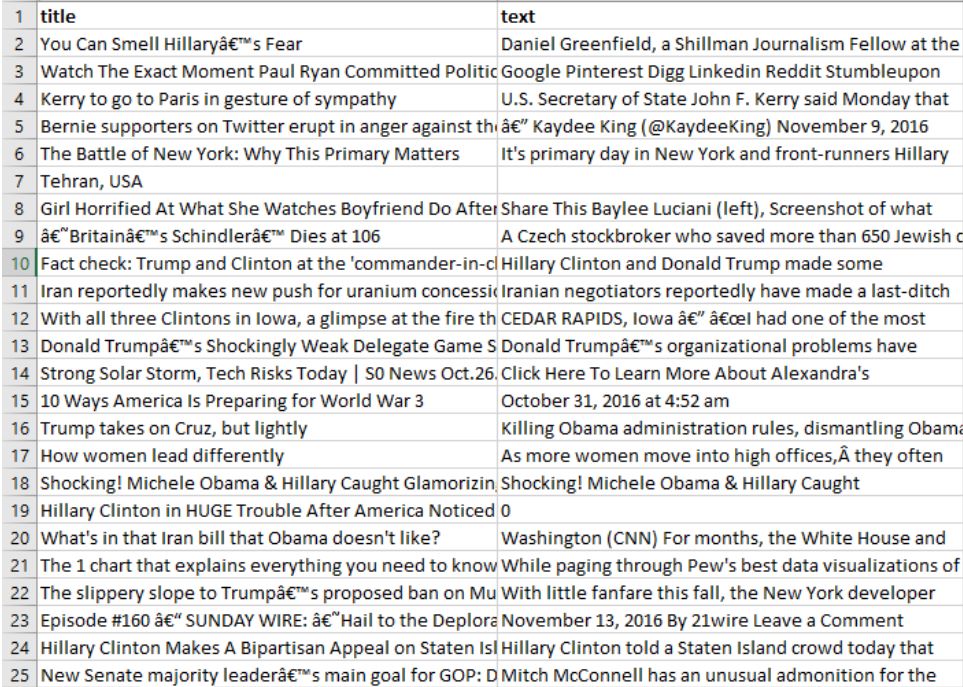
# Appendix A: Code snippets

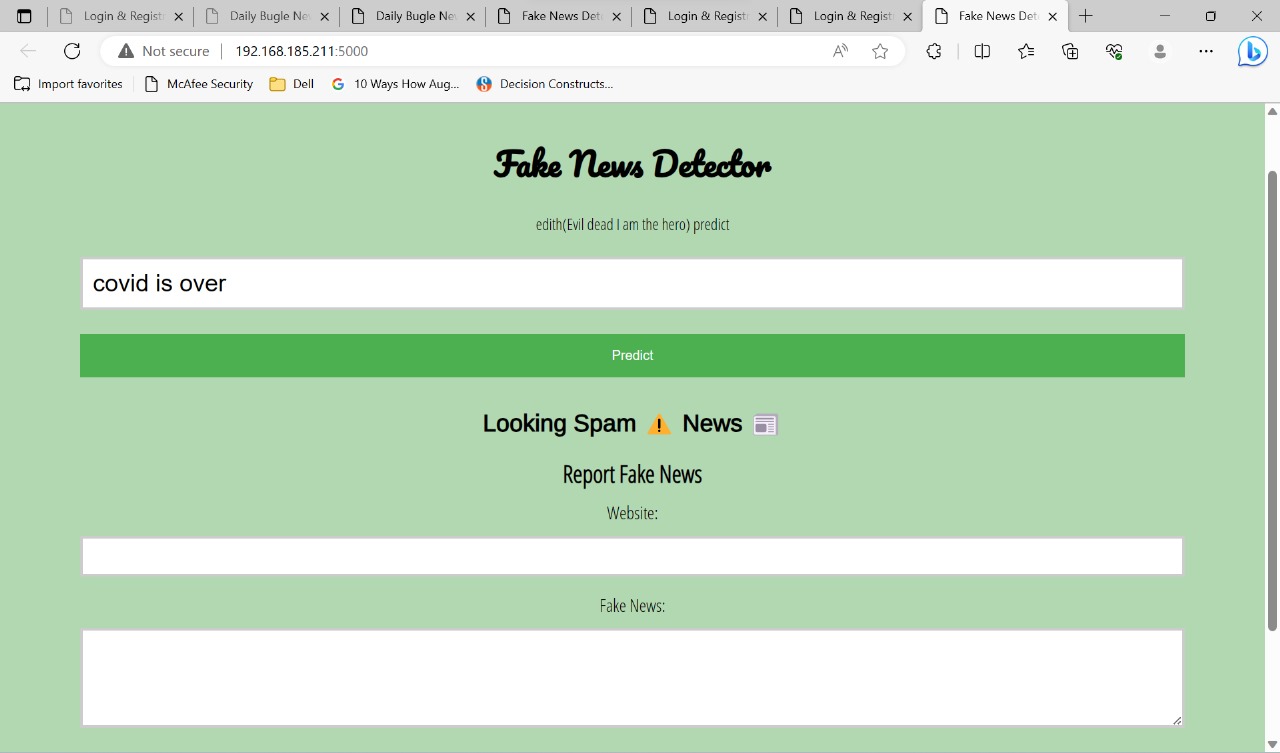
1. Cl´ement Bisaillon. *Fake and real news dataset*. Kag- gle, March 2020.
2. J. & al. Gottfried. *The 2016 presidential campaign–a news event that’s hard to miss*. Pew Res Cent. 4, 2016.
3. J. & al. Gottfried. *News use across social media plat- forms*. Pew Res Cent. 26, 2016.
4. Jing Lu & Peilin Zhao & Steven C. H. Hoi. *Online Passive-Aggressive Active learning*. Springer Link, 22 March 2016.
5. @IUT-Thesis-Group-Cmr & @IUT-Imperial-237. Machine Learning Project: Fake News Detector.
6. Katharine Jarmul. *Detecting fake news with scikit- learn*. DataCamp Community, 2017.
7. PhD Jason Brownlee. Train-Test Split for Evaluating Machine Learning Algorithms.
8. Cl´ement Bisaillon. *Fake and real news dataset*. Kag- gle, March 2020.
9. J. & al. Gottfried. *The 2016 presidential campaign–a news event that’s hard to miss*. Pew Res Cent. 4, 2016.
10. J. & al. Gottfried. *News use across social media plat- forms*. Pew Res Cent. 26, 2016.
11. Jing Lu & Peilin Zhao & Steven C. H. Hoi. *Online Passive-Aggressive Active learning*. Springer Link, 22 March 2016.
12. @IUT-Thesis-Group-Cmr & @IUT-Imperial-237. Machine Learning Project: Fake News Detector.
13. Katharine Jarmul. *Detecting fake news with scikit- learn*. DataCamp Community, 2017.
14. PhD Jason Brownlee. Train-Test Split for Evaluating Machine Learning Algorithms.
15. PolitiFact. *“Why alex jones was banned from apple, facebook, youtube.”*. PolitiFact.com, 07 August 2018.
16. M. Rao. *“Exploiting network structure to detect fake news”*. Stanford University, School of Computer Sci- ence, 2018.
17. M. Haag & M. Salam. *“Gunman in ’pizzagate’ shoot- ing is sentenced to 4 years in prison,”*. NyTimes, June 2017.
18. A. Thota. *”Fake News Detection: A deep learning approach,”*. SMU, Data Science Review, 2018.
19. David J Hand & Robert J Till. *A simple generalisa- tion of the area under the roc curve for multiple class classification problems*. Machine learning, 2001.
20. Charles X Ling & Jin Huang & Harry Zhang. *Auc: a statistically consistent and more discriminating mea- sure than accuracy*.

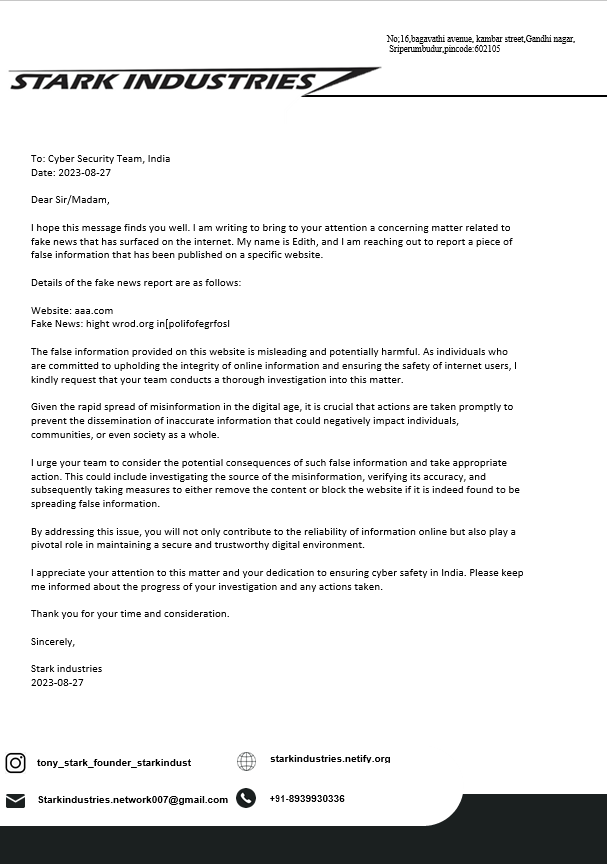
Sample pictures









****