Fake News Detection Using Machine Learning

Valluri Akash

Date: 06-05-2023

Abstract:

This paper presents a business idea for a machine learning-based product that can detect fake articles and spam content on the internet. The product will utilize state-of-the-art natural language processing techniques to analyse text data and identify patterns and features that are indicative of spam or fake content. The product will be aimed at businesses and individuals who need to protect their online reputation and prevent the spread of misinformation.

The proposed product will be built using advanced machine learning algorithms, including deep learning neural networks, to achieve high accuracy in detecting fake articles and spam content. The paper outlines the potential market size for this product and the potential revenue streams, including subscription-based models and partnerships with social media platforms and online news outlets. Overall, this business idea has significant potential to meet the growing demand for effective spam and fake article detection in the digital age.

1.Problem statement:

The increasing volume of online content has led to a rise in the spread of fake news and spam, which can harm businesses and individuals by promoting fraudulent or low-quality products or services. Spam can harm the reputation of a company or product, reduce productivity, and waste resources. Therefore, detecting and removing spam is critical for maintaining a healthy and efficient system.

Using a business idea for a fake article or spam detection using a machine learning model can solve a variety of problems, including:

1. Using machine learning can ***prevent fraud*** by detecting and filtering out fake articles and spam that promote fraudulent products or services.

2. Filtering out fake articles and spam *can* ***improve the customer experience*** by reducing irrelevant and misleading content.

3. Businesses can ***enhance their brand reputation*** by taking steps to filter out fake articles and spam, which can damage their reputation and erode customer trust.

4. Machine learning can help businesses ***identify trends and opportunities*** by analyzing data from articles and other sources, which can inform strategic decision-making.

5. Automating the filtering process using machine learning can ***save time and resources*** that would otherwise be spent manually filtering out fake articles and spam.

2. Market/Customer/Business Need Assessment:

Market Assessment:

- Research the current market for fake article/spam detection products/services and identify any gaps in the market.
- Identify potential competitors and their strengths and weaknesses.
- Consider the size of the market and potential growth opportunities.
- Determine if there is a need for a more advanced or user-friendly solution for fake article/spam detection.

Customer Assessment:

- Identify the target audience for the product/service and their specific needs and pain points when it comes to detecting fake articles/spam.
- Consider how the product/service can meet the needs of different types of customers (e.g. individual users, small businesses, large corporations).
- Research customer preferences when it comes to features, pricing, and customer support.

Business Assessment:

- Consider the potential revenue streams for the business (e.g. subscription fees, one-time purchase fees, advertising).
- Determine the cost structure for developing and maintaining the machine learning model and associated infrastructure.
- Identify potential risks and challenges associated with the business model (e.g. regulatory compliance, data privacy concerns).

3. Target Specifications and Characterization:

Target Customer Characteristics:

- Small to large businesses and organizations that rely heavily on email and messaging communication for their day-to-day operations.
- Individuals who receive a high volume of emails and messages and want to prioritize important ones over spam or promotional messages.
- People who want to protect themselves from phishing scams, malware, and other types of cyber threats through email and messaging platforms.

Target Specifications and Characterization:

- High accuracy rate of at least 95%
- Fast scanning speed
- Scalability to handle large volumes of messages without degradation in performance
- Customizable settings for spam filtering
- Integration with popular email and messaging platforms
- User-friendly interface
- Ensuring privacy and security of user data

4. External Search(information sources)

The dataset can be found on the Kaggle .The dataset consists about the URLs , Headlines , body of the articles .The sources of subsequent information is given below as references.

```
[9] df.head(10)
```

	URLs	Headline	Body	Label
0	http://www.bbc.com/news/world-us-canada-414191...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	1
1	https://www.reuters.com/article/us-filmfestiva...	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - "Last Flag Flying", a comed...	1
2	https://www.nytimes.com/2017/10/09/us/politics...	Trump's Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	1
3	https://www.reuters.com/article/us-mexico-oil-...	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...	1
4	http://www.cnn.com/videos/cnnmoney/2017/10/08/...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1
5	http://beforeitsnews.com/sports/2017/09/jetnat...	JetNation FanDuel League; Week 4	JetNation FanDuel League; Week 4\n% of readers...	0
6	https://www.nytimes.com/2017/10/10/us/politics...	Kansas Tried a Tax Plan Similar to Trump's. It...	In 2012, Kansas lawmakers, led by Gov. Sam Bro...	1
7	https://www.reuters.com/article/us-india-cenba...	India RBI chief: growth important, but not at ...	The Reserve Bank of India (RBI) Governor Urjit...	1
8	https://www.reuters.com/article/us-climatechan...	EPA chief to sign rule on Clean Power Plan exi...	Scott Pruitt, Administrator of the U.S. Enviro...	1
9	https://www.reuters.com/article/us-air-berlin-...	Talks on sale of Air Berlin planes to easyJet ...	FILE PHOTO - An Air Berlin sign is seen at an ...	1

Dataset after resetting the index after eliminating nan values

	index	URLs	Headline	Body	Label
0	0	http://www.bbc.com/news/world-us-canada-414191...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	1
1	1	https://www.reuters.com/article/us-filmfestiva...	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - "Last Flag Flying", a comed...	1
2	2	https://www.nytimes.com/2017/10/09/us/politics...	Trump's Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	1
3	3	https://www.reuters.com/article/us-mexico-oil-...	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...	1
4	4	http://www.cnn.com/videos/cnnmoney/2017/10/08/...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1
5	5	http://beforeitsnews.com/sports/2017/09/jetnat...	JetNation FanDuel League; Week 4	JetNation FanDuel League; Week 4\n% of readers...	0
6	6	https://www.nytimes.com/2017/10/10/us/politics...	Kansas Tried a Tax Plan Similar to Trump's. It...	In 2012, Kansas lawmakers, led by Gov. Sam Bro...	1
7	7	https://www.reuters.com/article/us-india-cenba...	India RBI chief: growth important, but not at ...	The Reserve Bank of India (RBI) Governor Urjit...	1
8	8	https://www.reuters.com/article/us-climatechan...	EPA chief to sign rule on Clean Power Plan exi...	Scott Pruitt, Administrator of the U.S. Enviro...	1
9	9	https://www.reuters.com/article/us-air-berlin-...	Talks on sale of Air Berlin planes to easyJet ...	FILE PHOTO - An Air Berlin sign is seen at an ...	1

Reference Links:

https://ijcrt.org/papers/IJCRT2006189.pdf

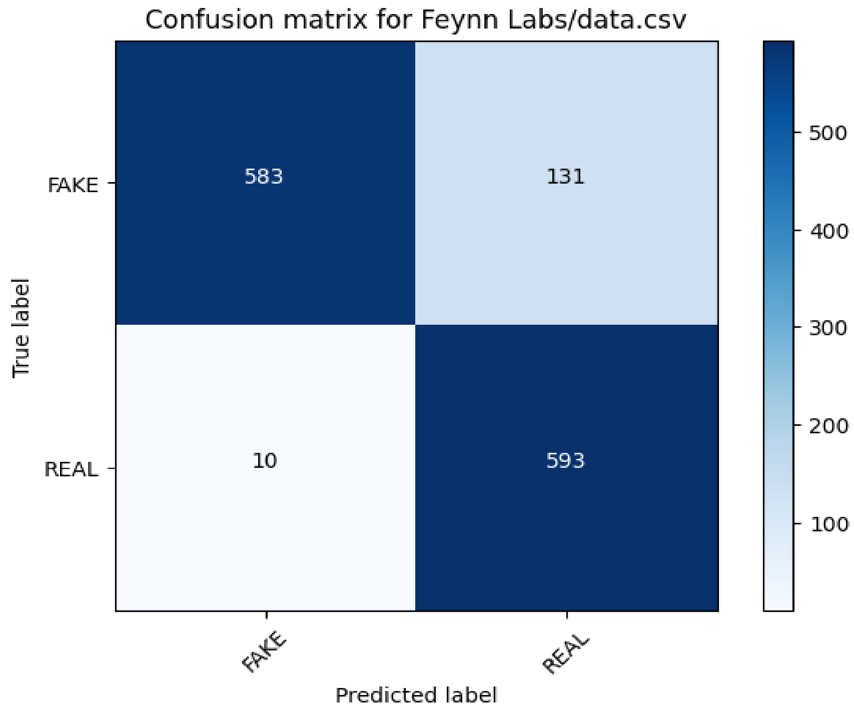
https://www.kaggle.com/code/ashishkumarbehera/fake-news-classifier

https://www.kaggle.com/datasets/jruvika/fake-news-detection?resource=download

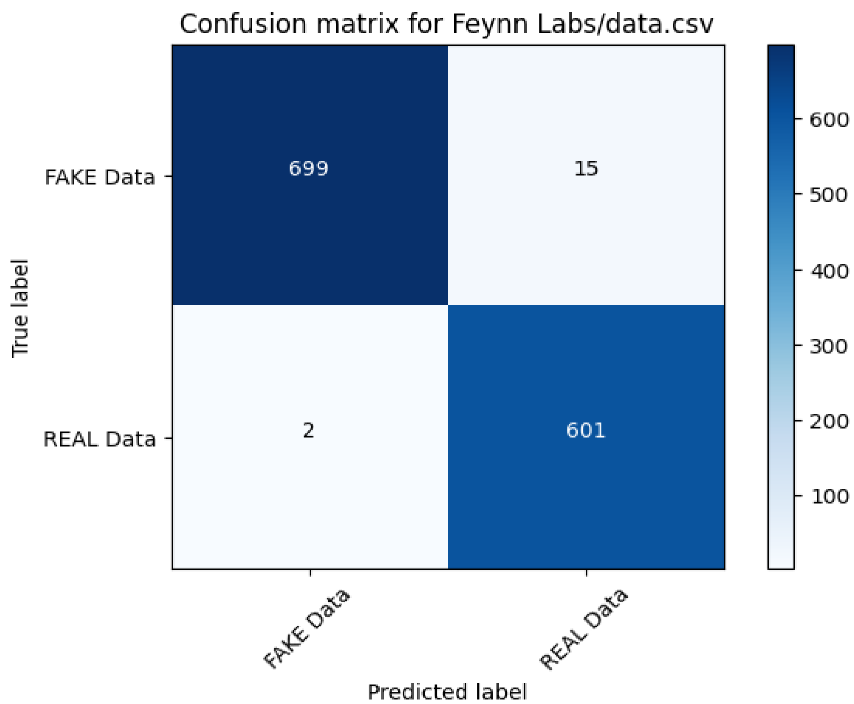
5.Bench marking

Confusion matrices:

(i)Using multinomial NB Classifier:



(ii)Passive Aggressive Classifier:



6.Applicable Patents:

According to the International Journal of Creative Research Thoughts (IJCRT) -Volume 8, Issue 6 June 2020, ISSN: 2320-2882; The current journal may incorporate this patent for the inspiration of the methodology used as well as EDA analysis to some extent.

7.Applicable Regulations:

These regulations vary depending on your location and the type of data you are processing, but here are some general guidelines to consider:

1. **Privacy regulations**: When collecting and processing personal data, you must comply with privacy regulations such as the General Data Protection Regulation (GDPR) in the European Union or the California Consumer Privacy Act (CCPA) in the United States.
2. **Intellectual property laws**: You should be careful not to infringe on any intellectual property rights when using machine learning algorithms to detect fake articles/spam. This includes copyright, trademark, and patent laws.
3. **Anti-spam laws**: Depending on where you operate, there may be anti-spam laws that you need to comply with. For example, in the US, the CAN-SPAM Act regulates commercial email messages.
4. **Data protection laws**: Depending on the data you are processing, you may need to comply with additional data protection laws. For example, if you are processing health data, you may need to comply with the Health Insurance Portability and Accountability Act (HIPAA).
5. **Ethical considerations**: As with any business that involves processing data, you should be mindful of the ethical implications of your product or service. This includes being transparent about how you collect and use data, being mindful of bias in your algorithms, and ensuring that your algorithms are not used to discriminate against any individuals or groups.

8.Applicable Constraints:

1. Machine learning models for fake article/spam detection require significant ***storage space*** and may impact hardware requirements and costs.
2. ***Budget constraints*** are a significant factor in developing machine learning models, including the resources required to build and maintain them.
3. ***Expertise*** in machine learning, natural language processing, and data analytics is necessary to build and maintain a machine learning model for fake article/spam detection.
4. ***High-quality data*** is required to train a machine learning model for fake article/spam detection, and obtaining this data can be challenging.
5. ***Ethical*** considerations, including fairness, transparency, and accountability, must be taken into account when developing and deploying machine learning models for fake article/spam detection.

9. Business Model:

One potential business opportunity for a machine learning model that detects fake articles or spam is to create a tool or service for media companies, news outlets, and social media platforms to automatically filter out fake news, spam, and other low-quality content.

i. The tool or service could be sold as a subscription-based service, with different pricing tiers based on the number of articles or posts analysed per month or the complexity of the machine learning models used. This could be particularly valuable for social media platforms, which have struggled to effectively police their platforms for fake news and spam.

ii. Another potential business opportunity would be to license the machine learning model to companies that are looking to build their own spam or fake news detection systems. This could include companies in the media, advertising, or e-commerce industries, as well as government agencies that are looking to monitor online content for disinformation or propaganda.

iii. Recently Telecom Regulatory Authority of India (TRAI) to implement new rules from Monday (May 1) to combat fake and promotional calls and messages that customers often receive. To achieve this, TRAI has mandated that all telecom companies must incorporate Artificial Intelligence (AI) spam filters into their call and SMS services. Which becomes huge business opportunity for this model.

10.Concept Generation:

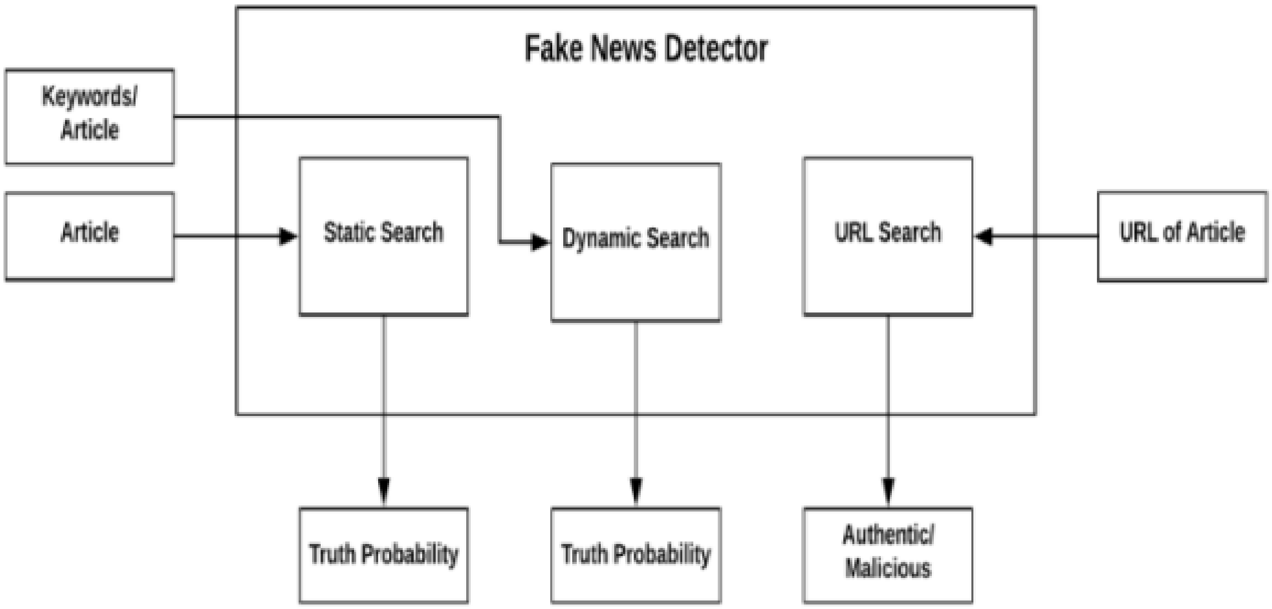
Machine learning models can be trained to analyze patterns and characteristics of language, such as syntax, grammar, and vocabulary, to detect fake articles or spam. By analyzing large amounts of text data, machine learning models can identify anomalies that are common in fake articles or spam, such as unusual language patterns or repeated phrases. Once trained, these models can be integrated into business or product idea development processes to automatically detect and remove fake articles or spam, improving the accuracy and reliability of the data used for decision-making.

Furthermore, businesses can also use machine learning models to continuously monitor for new patterns and characteristics of fake articles or spam. As new techniques for generating fake articles or spam emerge, machine learning models can be updated to detect these new patterns and improve their accuracy over time. This can help businesses stay ahead of emerging threats and maintain the integrity of their brand and product offerings. Overall, machine learning models offer a powerful and scalable solution for detecting and removing fake articles or spam in business and product idea development.

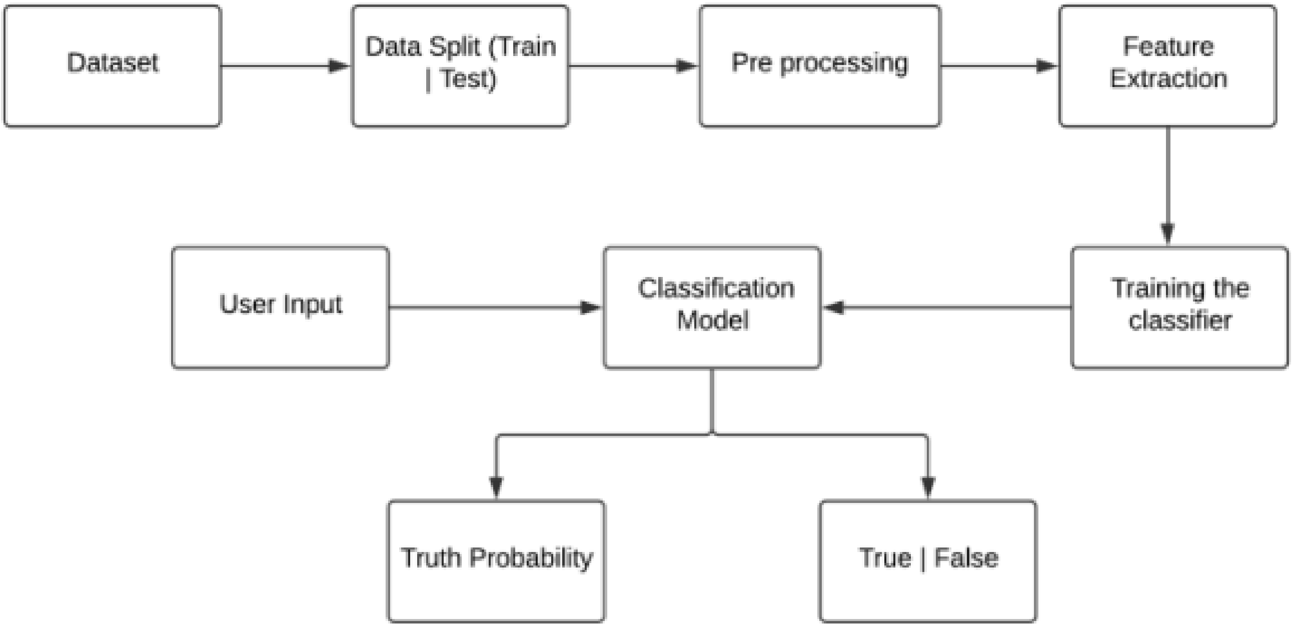
11. Concept Development:

We have used Python and its Sci-kit libraries. Python has a huge set of libraries and extensions, which can be easily used in Machine Learning. Sci-Kit Learn library is the best source for machine learning algorithms where nearly all types of machine learning algorithms are readily available for Python, thus easy and quick evaluation of ML algorithms is possible. We have used Django for the web based deployment of the model, provides client side implementation using HTML, CSS and JavaScript. We have also used Beautiful Soup (bs4), requests for online scrapping.

System Design:



System Architecture:



12.Final Product Prototype

12.1 Abstract:

Our implementation contains 3 search fields which are

1) Search by article content.

2) Search using key terms.

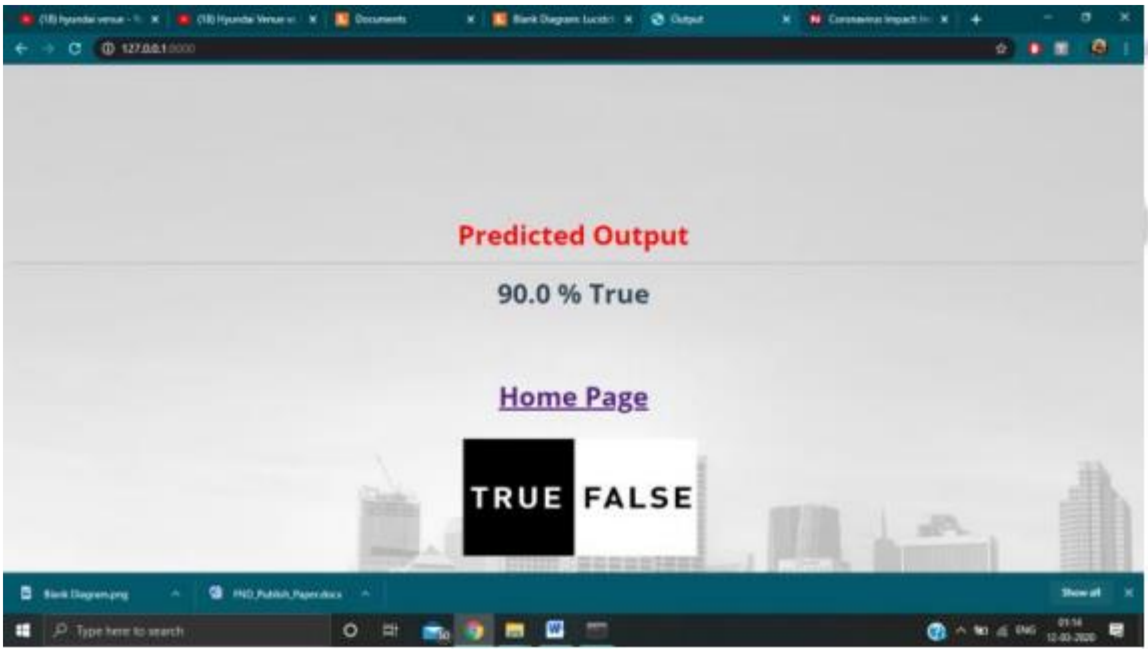
3) Search for website in database.

In the first search field we have used Natural Language Processing for the first search field to come up with a proper solution for the problem, and hence we have attempted to create a model which can classify fake news according to the terms used in the newspaper articles. Our application uses NLP techniques like Count Vectorization and TF-IDF Vectorization before passing it through a Passive Aggressive Classifier to output the authenticity as a percentage probability of an article.

The second search field of the site asks for specific keywords to be searched on the net upon which it provides a suitable output for the percentage probability of that term actually being present in an article or a similar article with those keyword references in it. The third search field of the site accepts a specific website domain name upon which the implementation looks for the site in our true sites database or the blacklisted sites database. The true sites database holds the domain names which regularly provide proper and authentic news and vice versa. If the site isn't found in either of the databases then the implementation doesn't classify the domain it simply states that the news aggregator does not exist.

Expected Software Overview

Predicted Output



13.Product details:

13.1 Working:

i. **Data pre-processing:**

- Collect a large dataset of articles, including both genuine and fake/spam articles.
- Clean the data by removing any HTML tags, URLs, or other non-textual elements.
- Tokenize the text into individual words or subworlds.
- Normalize the text by converting everything to lowercase and removing punctuation.
- Remove stop words, which are common words that don't carry much meaning (e.g., "the", "and", "a").

ii. **Text pre-processing:**

- Convert the text data into numerical representations, such as word embeddings or bag-of-words representations.
- Split the data into training and testing sets.

iii. **Model selection:**

- Choose a machine learning model that is suitable for text classification, such as Naive Bayes, logistic regression, or a neural network.We used Multinomial NB classifier and Passive Aggressive classifier.
- Tune the hyperparameters of the model using techniques such as grid search or random search.

iv. **Model building:**

- Train the machine learning model on the training set.
- Evaluate the model's performance on the testing set using metrics such as accuracy, precision, recall, and F1 score.

- If the model's performance is not satisfactory, try tweaking the model architecture or hyperparameters and retrain the model.

v. **Deployment:**

Deploy the machine learning model to a production environment, where it can be used to classify new articles as genuine or fake/spam.

13.2 Algorithm:

The problem can be broken down into 3 statements

1) Use NLP to check the authenticity of a news article.

2) If the user has a query about the authenticity of a search query, then we he/she can directly search on our platform and using our custom algorithm we output a confidence score.

3) Check the authenticity of a news source. These sections have been produced as search fields to take inputs in 3 different forms in our implementation of the problem statement.

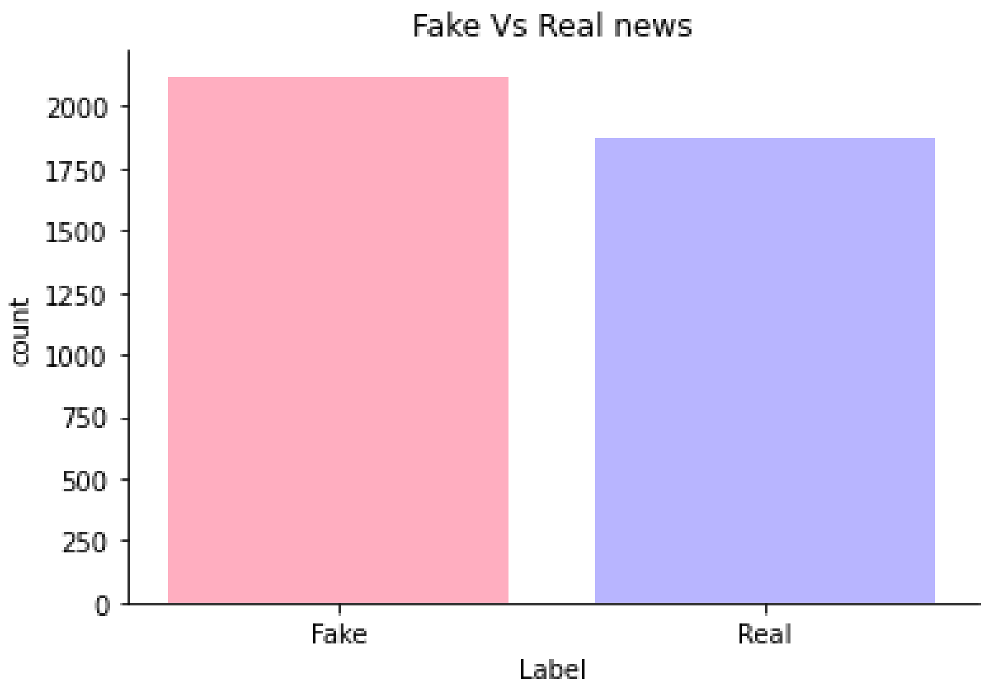
13.3 Data Sources:

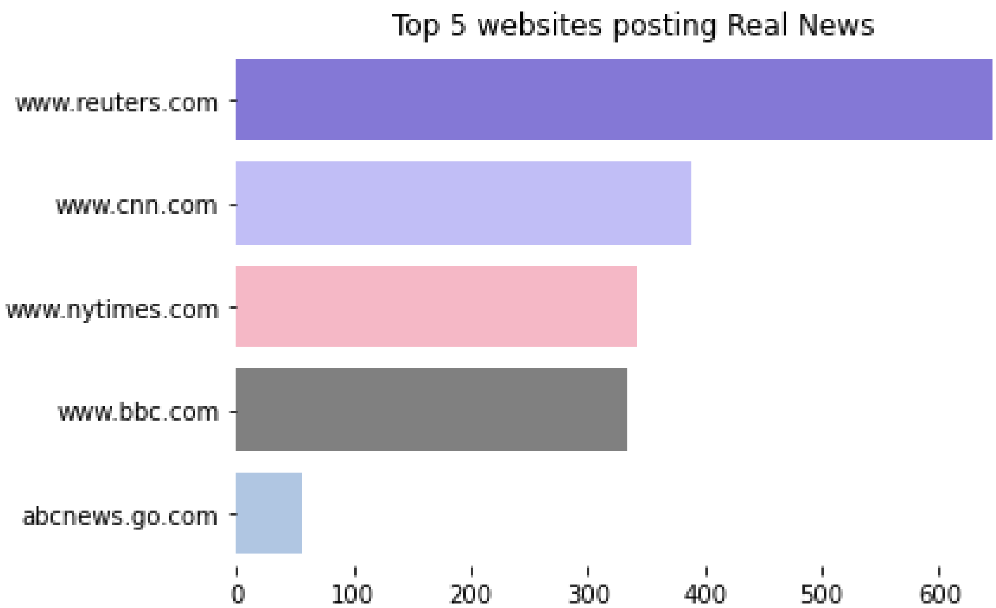
We can get online news from different sources like social media websites, search engine, homepage of news agency websites or the factchecking websites. On the Internet, there are a few publicly available datasets for Fake news classification like Buzzfeed News, LIAR , BS Detector etc. These datasets have been widely used in different research papers for determining the veracity of news. In the following sections, I have discussed in brief about the sources of the dataset used in this work. Online news can be collected from different sources, such as news agency homepages, search engines, and social media websites.

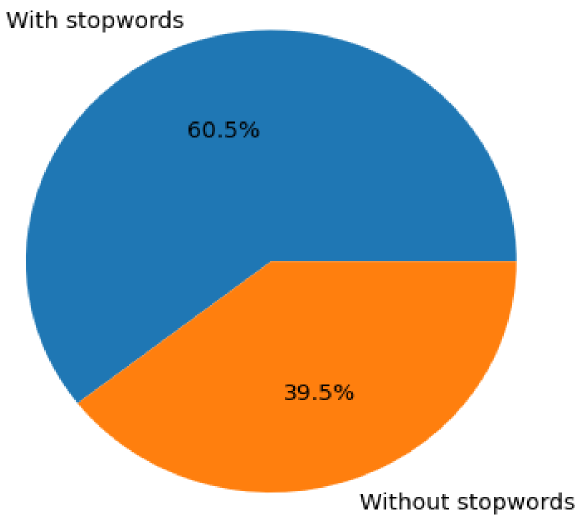
However, manually determining the veracity of news is a challenging task, usually requiring annotators with domain expertise who performs careful analysis of claims and additional evidence, context, and reports from authoritative sources. Generally, news data with annotations can be gathered in the following ways: Expert journalists, Fact-checking websites, Industry detectors, and Crowd sourced workers. However, there are no agreed upon benchmark datasets for the fake news detection problem. Data gathered must be pre-processed- that is, cleaned, transformed and integrated before it can undergo training process.

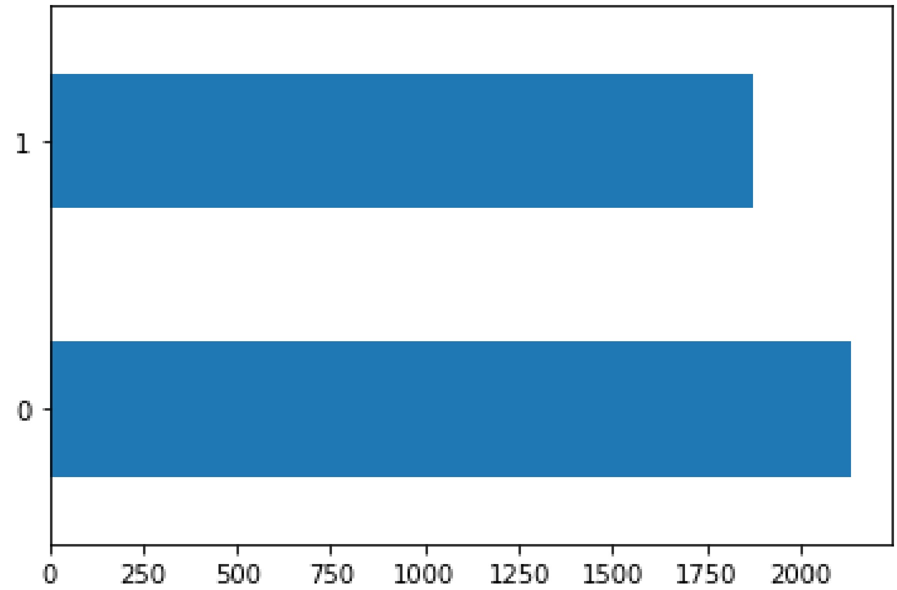
14. Code Implementation/Validation

14.1 Some Basic Visualizations on Real World or Augmented Data:









14.2 Code

```
import pandas as pd
df=pd.read_csv('data.csv')
```

```
[10] df.head()
```

	URLs	Headline	Body	Label
0	http://www.bbc.com/news/world-us-canada-414191...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	1
1	https://www.reuters.com/article/us-filmfestiva...	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - “Last Flag Flying”, a comed...	1
2	https://www.nytimes.com/2017/10/09/us/politics...	Trump’s Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	1
3	https://www.reuters.com/article/us-mexico-oil-...	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt’s Cheiron Holdin...	1
4	http://www.cnn.com/videos/cnnmoney/2017/10/08/...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1

14.2.1 Data Pre-processing:

```
[11] ## Get the Independent Features
     X=df.drop('Label',axis=1)
```

```
[12] X.head()
```

	URLs	Headline	Body
0	http://www.bbc.com/news/world-us-canada-414191...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...
1	https://www.reuters.com/article/us-filmfestiva...	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - "Last Flag Flying", a comed...
2	https://www.nytimes.com/2017/10/09/us/politics...	Trump's Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...
3	https://www.reuters.com/article/us-mexico-oil-...	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...
4	http://www.cnn.com/videos/cnnmoney/2017/10/08/...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...

```
[13] ## Get the Dependent features
     y=df['Label']
```

```
y.head()
```

```
0    1
1    1
2    1
3    1
4    1
Name: Label, dtype: int64
```

```
[15] df.shape
```

```
(4009, 4)
```

```
[16] from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
```

```
[17] df=df.dropna()#dropping the Nan values
```

```
[18] df.shape
```

```
(3988, 4)
```

```
messages=df.copy()
```

```
[20] messages.reset_index(inplace=True)## reseting the index after eliminating nan values
```

```
messages.head()
```

	index	URLs	Headline	Body	Label
0	0	http://www.bbc.com/news/world-us-canada-414191...	Four ways Bob Corker skewered Donald Trump	Image copyright Getty Images\nOn Sunday mornin...	1
1	1	https://www.reuters.com/article/us-filmfestiva...	Linklater's war veteran comedy speaks to moder...	LONDON (Reuters) - "Last Flag Flying", a comed...	1
2	2	https://www.nytimes.com/2017/10/09/us/politics...	Trump's Fight With Corker Jeopardizes His Legi...	The feud broke into public view last week when...	1
3	3	https://www.reuters.com/article/us-mexico-oil-...	Egypt's Cheiron wins tie-up with Pemex for Mex...	MEXICO CITY (Reuters) - Egypt's Cheiron Holdin...	1
4	4	http://www.cnn.com/videos/cnnmoney/2017/10/08/...	Jason Aldean opens 'SNL' with Vegas tribute	Country singer Jason Aldean, who was performin...	1

```
[22] messages['Body'][6]
```

```
'In 2012, Kansas lawmakers, led by Gov. Sam Brownback, a Republican, enacted a tax cut that eliminated state income taxes entirely for pass-through entities — such as sole proprieto
rships and limited liability partnerships — which are taxed at the owner's individual income tax rate. The law also lowered individual income tax rates, cutting the top rate to 4.9
percent from 6.4 percent.\nThe tax package reduced state revenue by nearly $700 million a year, a drop of about 8 percent, from 2013 through 2016, according to the Kansas Legislativ
e Research Department, forcing officials to shorten school calendars, delay highway repairs and reduce aid to the poor. Research suggests the package did not stimulate the economy,
certainly not enough to pay for the tax cut. This year, legislators passed a bill to largely rescind the law, saying it had not worked as intended.\n"It caused a lot of budget insta
bility," said State Senator Jim Denning, a Republican who led the effort to repeal the pass-thro...'
```

14.2.2 Text Pre-processing:

```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
[24] from nltk.corpus import stopwords
     from nltk.stem.porter import PorterStemmer
     import re
     ps = PorterStemmer()
     corpus = []
     for i in range(0, len(messages)):
         review = re.sub('[^a-zA-Z]', ' ', messages['Body'][i])
         review = review.lower()
         review = review.split()

         review = [ps.stem(word) for word in review if not word in stopwords.words('english')]
         review = ' '.join(review)
         corpus.append(review)
```

```
[25] corpus[3]
```

```
'mexico citi reuter egypt cheiron hold limit right partner mexican nation oil compani pemex onshor cardena mora project industri regul said wednesday tie mark second joint ventur pe
mex equiti partner sinc energi open final end compani decad long monopoli allow develop project privat foreign oil compani cardena mora squar mile sq km field locat tabasco state be
liev contain million barrel oil equival boe proven probabl possibl reserv'
```

```
[26] ## Applying TFidf Vectorizer
     from sklearn.feature_extraction.text import TfidfVectorizer
     tfidf_v=TfidfVectorizer(max_features=5000,ngram_range=(1,3))
     X=tfidf_v.fit_transform(corpus).toarray()
```

```
[29] ## Divide the dataset into Train and Test
     from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=0)
```

```
[30] tfidf_v.get_feature_names_out()[:20]
```

```
array(['aaron', 'aaron judg', 'aaron kesel', 'aaron rodger', 'ab',
       'abandon', 'abc', 'abe', 'aberti', 'abil', 'abl', 'abort',
       'abroad', 'absenc', 'absolut', 'abus', 'ac', 'academ', 'academi',
       'acc'], dtype=object)
```

```
tfidf_v.get_params()# things what you have applied in tfidf
```

```
{'analyzer': 'word',
 'binary': False,
 'decode_error': 'strict',
 'dtype': numpy.float64,
 'encoding': 'utf-8',
 'input': 'content',
 'lowercase': True,
 'max_df': 1.0,
 'max_features': 5000,
 'min_df': 1,
 'ngram_range': (1, 3),
 'norm': 'l2',
 'preprocessor': None,
 'smooth_idf': True,
 'stop_words': None,
 'strip_accents': None,
 'sublinear_tf': False,
 'token_pattern': '(?u)\\b\\w\\w+\\b',
 'tokenizer': None,
 'use_idf': True,
 'vocabulary': None}
```

```
[32] count_df = pd.DataFrame(X_train, columns=tfidf_v.get_feature_names_out())
```

```
{'analyzer': 'word',
 'binary': False,
 'decode_error': 'strict',
 'dtype': numpy.float64,
 'encoding': 'utf-8',
 'input': 'content',
 'lowercase': True,
 'max_df': 1.0,
 'max_features': 5000,
 'min_df': 1,
 'ngram_range': (1, 3),
 'norm': 'l2',
 'preprocessor': None,
 'smooth_idf': True,
 'stop_words': None,
 'strip_accents': None,
 'sublinear_tf': False,
 'token_pattern': '(?u)\\b\\w\\w+\\b',
 'tokenizer': None,
 'use_idf': True,
 'vocabulary': None}
```

```
[32] count_df = pd.DataFrame(X_train, columns=tfidf_v.get_feature_names_out())
```

```
[33] count_df.head()
```

	aaron	aaron judg	aaron kesel	aaron rodger	ab	abandon	abc	abe	aberti	abil	...	yorker	young	younger	youth	youtub	zach	zapatista	zealand	zero	zone
0	0.073735	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 5000 columns

14.2.3 Model selection and model building

Multinomial NB Classifier:

```
[36] from sklearn.naive_bayes import MultinomialNB
     classifier=MultinomialNB()
```

```
[37] from sklearn import metrics
     import numpy as np
     import itertools
```

```
classifier.fit(X_train, y_train)
pred = classifier.predict(X_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred)
plot_confusion_matrix(cm, classes=['FAKE', 'REAL'])
```

```
accuracy:   0.893
Confusion matrix, without normalization
```

Passive Aggressive Classifier:

```
[41] from sklearn.linear_model import PassiveAggressiveClassifier
     linear_clf = PassiveAggressiveClassifier(n_iter_no_change=50)
```

```
linear_clf.fit(X_train, y_train)
pred = linear_clf.predict(X_test)
score = metrics.accuracy_score(y_test, pred)
print("accuracy:   %0.3f" % score)
cm = metrics.confusion_matrix(y_test, pred)
plot_confusion_matrix(cm, classes=['FAKE Data', 'REAL Data'])
```

```
accuracy:   0.983
Confusion matrix, without normalization
```

- Accuracy obtained for the Passive Aggressive Clasifier is higher than compared to the other

Colab link for code implementation:

https://colab.research.google.com/drive/1o8kAvzA9hSGnQ3H15Vyf0iPvcrEdMCbb#scrollTo=ytDsYv-KsNHL

15. Conclusion:

In the 21st century, the majority of the tasks are done online. Newspapers that were earlier preferred as hard-copies are now being substituted by applications like Facebook, Twitter, and news articles to be read online. WhatsApp's forwards are also a major source. The growing problem of fake news only makes things more complicated and tries to change or hamper the opinion and attitude of people towards use of digital technology. When a person is deceived by the real news two possible things happen- People start believing that their perceptions about a particular topic are true as assumed. Thus, in order to curb the phenomenon, we have developed our Fake news Detection system that takes input from the user and classify it to be true or fake. To implement this, various NLP and Machine Learning Techniques have to be used.