Fake News Detection

Team name: Cosmos

Nagavarun S.N(20BDA67), Ranjith Kumar K.N(20BDA56)

St. Joseph’s College Autonomous(Bangalore)

Abstract:

The term "fake news" has recently become popular. There was a time when everyone who required news had to wait for the next day's newspaper. However, with the rise of online newspapers that update news practically quickly, individuals have discovered a better and faster method to stay updated about topics of interest. Nowadays, social-networking systems, online news portals, and other online media are the primary sources of news, with fascinating and breaking news being disseminated at a quick rate. Many news portals, on the other hand, serve a specific purpose by disseminating distorted, partially true, and occasionally fictitious news that is intended to pique the interest of a specific set of individuals. Fake news has become a major source of worry due to its ability to cause misunderstanding and purposeful misdirection among the public.

Introduction:

Because of the simple accessibility and exponential expansion of information available on social media networks, distinguishing between false and real information has become difficult. The ease with which information may be shared has resulted in an exponential increase in the amount of information that can be falsified. Where the transmission of fraudulent material is rampant, the credibility of social media networks is also jeopardised. As a result, automatically verifying the information's source, substance, and publisher in order to categorise it as false or real has become a research problem. Machine learning has played an important role in data classification, but with significant drawbacks. This study examines a number of machine learning algorithms for detecting bogus and counterfeit news. The limitations of such approaches, as well as improvisational methods for applying machine learning and deep learning, are discussed.

Keywords:

Machine Learning, Deep Learning, Fake News Detection.

Literature Survey:

1)Fake News Detection using Bi-directional LSTM-Recurrent Neural Network(Pritika Bahada, *, Preeti Saxenaa ,Raj Kamalb)

This research provides a model for detecting false news that is based on a bi-directional LSTM-recurrent neural network. To evaluate the model's performance, two publicly available unstructured news article datasets are employed. The results reveal that the Bi-directional LSTM model outperforms other approaches for detecting false news, such as CNN, vanilla RNN, and unidirectional LSTM, in terms of accuracy.

2) Optimization and improvement of fake news detection using deep learning approaches for societal benefit(Hemant Palivela, Tavishee Chauhan)

In this paper to distinguish bogus news from authentic news, a deep learning-based technique was applied in this study. The suggested model was created using an LSTM neural network. A gloVe word embedding, in addition to the neural network, has been employed for vector representation of textual words. Tokenization has also been used for feature extraction or vectorization. The notion of N-grams is applied to improve the suggested model. A comparison of several false news detecting systems is examined.

3) Fake News Detection Using Machine Learning Algorithms (Uma Sharma, Sidharth Saran,, Shankar M. Patil)

The system is explained in three parts in this paper. The first component is static and relies on a machine learning classifier. They investigated and trained the model using four distinct classifiers before selecting the optimal classifier for final execution. The second portion is dynamic, taking the user's keyword/text and searching internet for the truth probability of the news. The third section contains the genuine.

4) Fake News detection Using Machine Learning (Nihal Fatima Baarir, Abdelhamid Djeffal)

In this research, they describe a unique approach and tool for identifying false news that employs the following techniques:

• Text pre-processing: heating and analysing the text by eliminating stop words and unusual characters.

• Text encoding: employing a bag of words and N-grams, followed by TF-IDF.

• Characteristic extraction: this enables for the exact detection of erroneous information. We consider the source of a news, its author, the date, and the sentiment conveyed by the text to be news features.

• Support vector machine: a supervised machine learning technique that can classify fresh data.

5) A systematic Review Fake News Detection Using Machine Learning approaches (Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita)

This research examines several Machine Learning methods for detecting bogus and fraudulent news. The limitations of such techniques and improvisation through the implementation of deep learning are also discussed.

Many automatic detection algorithms for false news and deception posts have been reported in the literature. Because there are many different sides to fake news detection, from employing chatbots to promote disinformation to utilising clickbait's to propagate rumours. There are several clickbait's accessible on social media networks such as Facebook that encourage sharing and like. A great deal of effort has gone into detecting fake information.

The authors of introduced many detecting strategies. The following Detection Methods have been proposed by the writers.

Aim of the Work

The goals of this project are to apply Machine Learning and Deep Learning to detect fake news based on the text content of articles. After that, construct an appropriate Machine/Deep learning model to recognise fake/true news.

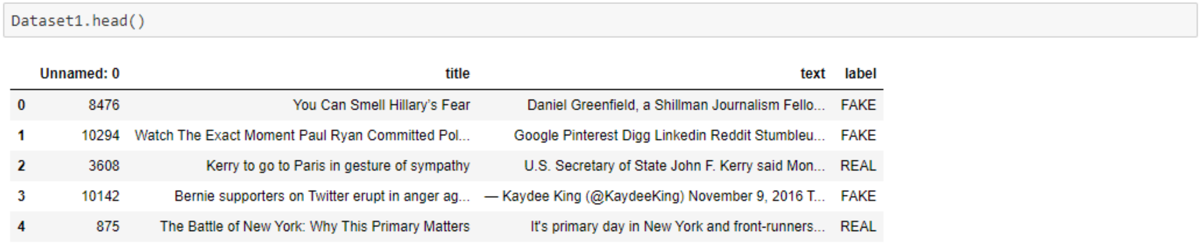
- Our sole objective is to classify the news from the dataset to fake or true news.
- Extensive EDA of news.
- Selecting and building a powerful model for classification.

Method and materials

Dataset:

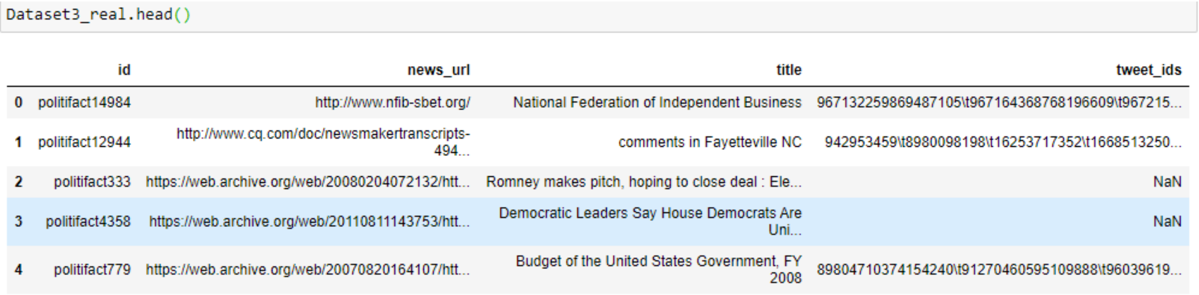
The datasets utilized in this study are all available in the public domain. The majority of the data came from Kaggle (https://www.kaggle.com/). Distinct datasets have different columns and data, such as [title, text, subject, news URL, author].

sample view of Dataset



```
Dataset1.head()
```

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL



```
Dataset3_real.head()
```

	id	news_url	title	tweet_ids
0	politifact14984	http://www.nfib-sbet.org/	National Federation of Independent Business	967132259869487105\t967164368768196609\t967215...
1	politifact12944	http://www.cq.com/doc/newsmakertranscripts-494...	comments in Fayetteville NC	942953459\t8980098198\t16253717352\t1668513250...
2	politifact333	https://web.archive.org/web/20080204072132/htt...	Romney makes pitch, hoping to close deal : Ele...	NaN
3	politifact4358	https://web.archive.org/web/20110811143753/htt...	Democratic Leaders Say House Democrats Are Uni...	NaN
4	politifact779	https://web.archive.org/web/20070820164107/htt...	Budget of the United States Government, FY 2008	898047103741542400\t912704605951098888\t96039619...

,

Packages: Sklearn (scikit-learn), NumPy, Pandas, matplotlib, seaborn, NLTK, Joblib, TensorFlow, Kera's, Seaborn

Algorithm's used: Logistic Regression, Decision Tree , Multinomial Naive Bayes', XGBoost, LSTM(Long-Short Term Memory)

Study Design:

Data Processing:

We used five different datasets in this research and combined them into a single data frame. Then we made a news column called 'Article' for the text, which will be the combination header and content. 1 was replaced to true and 0 was replaced to fake in the Label field.

Steps:

- Remove all unwanted columns.
- Remove All Missing Values Records.
- Removing all the extra information like brackets, any kind of punctuations - commas, apostrophes, quotes, question marks from Text.
- Remove all the numeric text, urls from Text.
- Processing of Text
  For every text analysis programme, this is a critical step. In the news, there will be a lot of useless information, which might be a problem when feeding a machine learning model. The machine learning model will not operate well until they are removed.
- Stop words in the news
  A stop word is a widely used word (such as "the," "a," "an," or "in") that a search engine has been configured to disregard while indexing and retrieving results as the result of a search query. We don't want these terms to eat up important processing time or take up space in our database.

Train/Test Split (75:25):

We split the dataset into 75:25 ratios for the train and test sets using the train test split function.

Model Building: Fake News Classifier:

It's not a typical machine learning challenge because we've successfully processed the text input. We classified the classes in the target feature using the sparse matrix.

Model Selection:

Using cross validation, first choose the best performing model. Let's have a look at all of the classification algorithms and go through the model selection procedure.

Models

Logistic Regression Classifier:

It is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

Decision Tree Classifier:

The decision tree classifier is a useful tool that operates on a flow chart-like structure and is mostly used to solve classification issues. Each decision tree internal node gives a condition or "test" on an attribute, and branching is based on the test conditions and results. Finally, the leaf node has a class label that is determined when all attributes have been computed. The categorization rule is represented by the distance between the root and the leaf. The fact that it can function with both a category and a dependent variable is incredible. They are effective in identifying the most significant variables and depicting the relationships between them. They are important in the creation of new variables and characteristics that are beneficial for data exploration and accurately forecast the target variable.

Extreme Gradient Boosting Classifier (XGBoost):

This is highly regarded by machine learning experts and is well-known among ML rivals. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. This system is available for all major data analysis languages and works well when dealing with multiclass classification situations. Because XGBoost is used to solve both regression and classification issues, it should do well in our classification challenge. Aggregates strong predictions from many weak learning decision trees.

Multinomial Naive Bayes Classifier:

Simple machine learning includes naive Bayes classifiers in machine learning. Using multinomial NB and pipelining ideas, Naive Bayes is a common technique for determining the accuracy of news, whether real or fraudulent. There are a variety of methods that focus on the same principle; thus, it isn't the only way to train such classifiers.

Long-Short Term Memory (LSTM):

In this section, we utilise a neural network to predict whether or not the supplied news is false. We'll start by developing and compiling the base model. The embedding layer will be the first, including inputs such as vocabulary size, vector characteristics, and sentence length. We next add a 30% dropout layer to prevent overfitting, as well as the LSTM layer, which includes 100 neurons. The sigmoid activation function is used in the last layer. Because we only have two outputs, we assemble the model using Adam optimizer and binary cross entropy as the loss function.

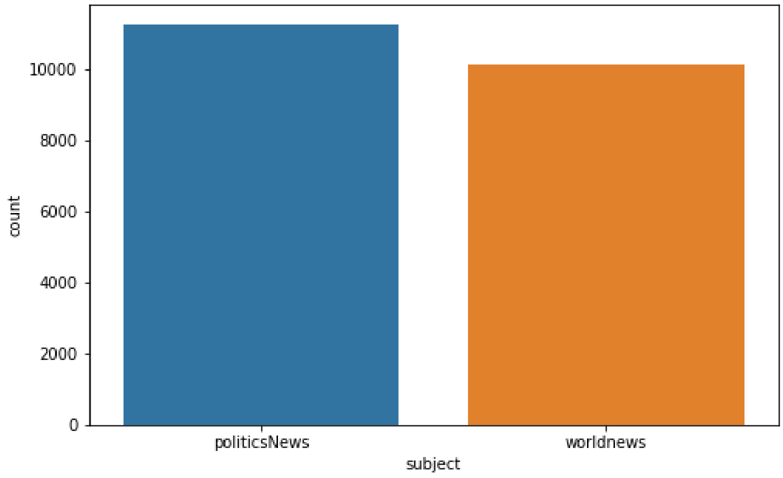
Evaluation of model:

In this we anticipate the output for our test data and use y test to assess the expected results.

Exploratory Data Analysis

1)Count of political news and world news

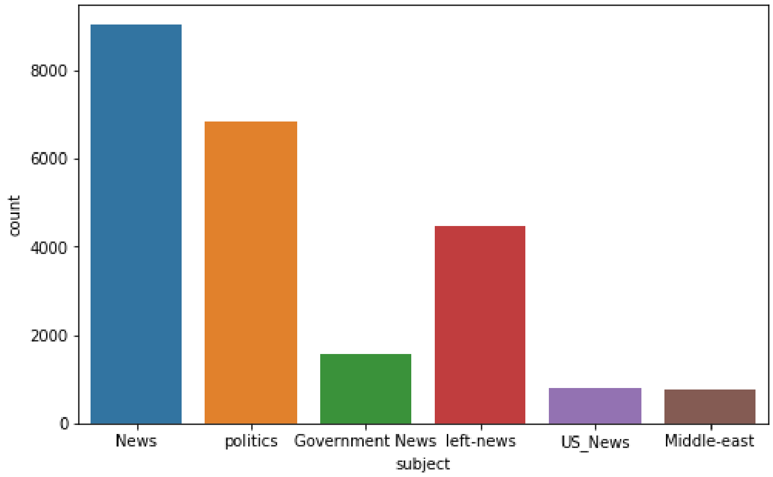
Let's check the count of political and world news and confirm whether our data is balanced or not



Insights:

- We have a pretty much balanced data
- But the count of Politics News is higher than the world news but not on a greater extent.

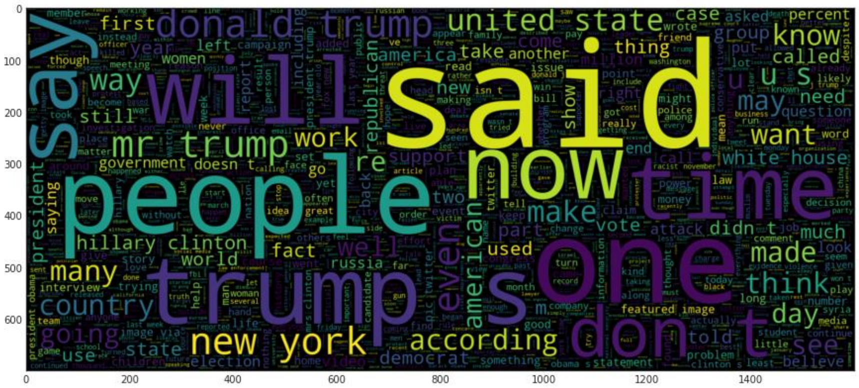
2)Plotting the Subjects in Fake news



Insights:

- Fake news is prevalent throughout the category, with the exception of politics and global news;
- True news is only found in politics and world news, and its prevalence is high.
- THIS IS A HIGHLY BIASED DATASET, so we can anticipate increased accuracy, but it doesn't mean it's a decent model given the dataset's poor quality.

3) Word Cloud of Fake and True News



Insights:

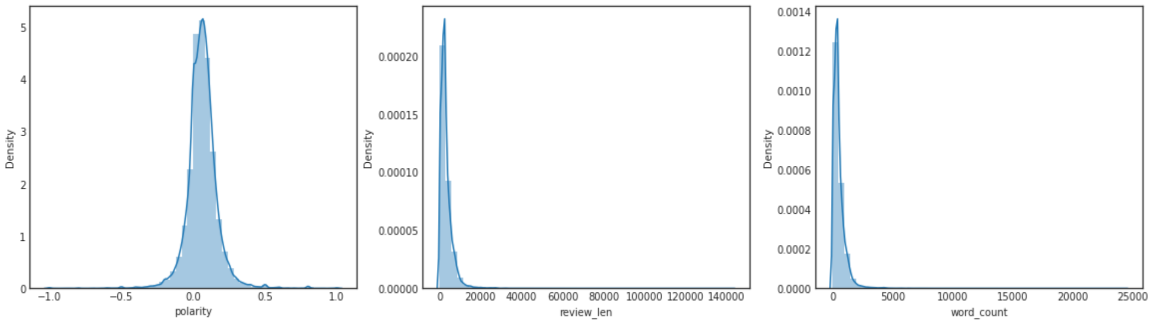
- Most of the fake news revolves around People and Trump, Said
- There are also fake news about privacy, internet etc.,

Deriving new features from the news

Let’s extract more features from the news feature such as

1. Polarity: The measure which signifies the sentiment of the news
2. Review length: Length of the news (number of letters and spaces)

3. Word Count: Number of words in the news



Insights:

- Most of the polarity are neutral, neither it shows some bad news nor much happy news
- The word count is between 0-1000 and the length of the news are between 0-5000 and few near 10000 words which could be an article.

Result and Discussion

Machine Learning Implementation: Let's take a look at all of the classification algorithms and run through the model selection process. The findings show that logistic regression outperformed the other algorithms, with Naive Bayes, XGBoost and Decision Tree coming in second and third, fourth, respectively. All of the classifiers for predicting the detection of fake news have been built here. Different classifiers are given the extracted characteristics. The classifiers such as Logistic Regression, XGBoost, Decision Tree, Multinomial Naive Baye’s, all of the classifiers utilised each of the retrieved characteristics. We compared the accuracy score and evaluated the confusion matrix after fitting the model.

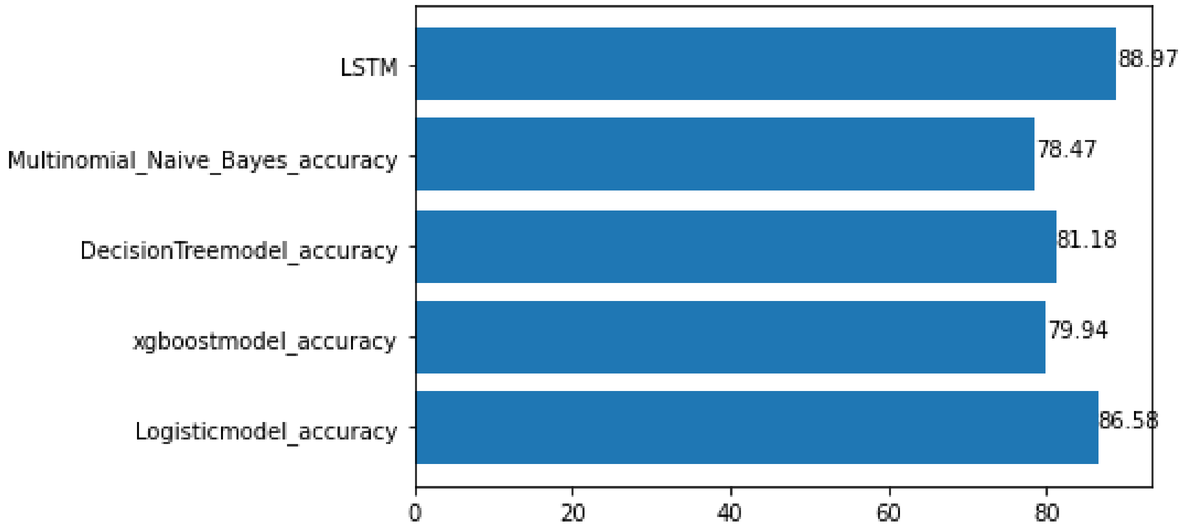
Deep Learning Implementation: In this section, we utilise a neural network to determine whether or not the supplied news is false. By employing LSTM (long short term memory), which aids in the storage of sequence data. The first layer of the LSTM is the embedding layer, which takes into account vocabulary size, vector characteristics, and sentence length. We next add a 30% dropout layer to prevent overfitting, as well as the LSTM layer, which includes 100 neurons. The sigmoid activation function is used in the last layer. Because we only have two outputs, we assemble the model using Adam optimizer and binary cross entropy as the loss function. Before fitting to the model, we consider the padded embedded object as X and y as y itself and convert them into an array. We have considered 10 epochs and 64 as batch size. It can be varied to get better results.

Model	Predict	Precision	Recall	F1-Score	Accuracy
Logistic Regression Classifier	0	0.85	0.90	0.87	0.87
	1	0.88	0.83	0.86	0.87
Decision Tree Classifier	0	0.76	0.93	0.84	0.81
	1	0.90	0.69	0.78	0.81
XGBoost Classifier	0	0.72	0.99	0.84	0.80
	1	0.98	0.60	0.74	0.80
Multinomial Naive Bayes Classifier	0	0.74	0.89	0.81	0.78
	1	0.85	0.67	0.75	0.78
LSTM	0	0.90	0.89	0.89	0.89
	1	0.88	0.89	0.88	0.89

From the classification report we can see the LSTM accuracy value is nearly around 89%. We have to concentrate on precision score and it is 89% which is small performance improvement from Logistic Regression.

Conclusion

We've completed a significant amount of work in terms of data processing and model construction. While vectorizing the text data, we may have experimented with changing the n-grams. The greatest accuracy score we have is 87.04, but don't worry, the model has been trained with over 61,000 records and will perform well. LSTM was the final and highest performing classifier we chose.



The majority of false news is found in the context of world news and politics. In light of the 2020 presidential elections in the United States. There is the potential for fake news to propagate, necessitating the use of these technologies. During this epidemic, fake news is rooted in order to play politics, scare people, and drive them to buy things. The majority of the information comes from Reuters. We don't know if this news organisation is swayed by politics. As a result, we must carefully analyse the source of news in order to determine if it is false or true.

References:

1)Fake News Detection using Bi-directional LSTM-Recurrent Neural Network

(Pritika Bahada, *, Preeti Saxenaa ,Raj Kamalb)

2) Optimization and improvement of fake news detection using deep learning approaches for societal benefit(Hemant Palivela, Tavishee Chauhan)

3) Fake News Detection Using Machine Learning Algorithms (Uma Sharma, Sidharth Saran,, Shankar M. Patil)

4) Fake News detection Using Machine Learning (Nihal Fatima Baarir, Abdelhamid Djeffal)

5) A systematic Review Fake News Detection Using Machine Learning approaches (Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita)

6) https://colah.github.io/posts/2015-08-Understanding-LSTMs/

7)https://towardsdatascience.com/covid-fake-news-detection-with-a-very-simple-logistic-regression-34c63502e33b

8) https://towardsdatascience.com/fake-news-detection-with-machine-learning-using-python-3347d9899ad1