Fake News Detection using Machine Learning

Vikas Kumar *2K17/CO/370*
Tarunesh Kumar Gautam *2K17/EE/222*
Hemprakash Meena *2K17/EE/090*
Yashpal *2K17/CO/387*

Abstract

With the rise of social media and cyber crime, fake news has become a major social problem. Sometimes it spreads more and faster than the true information. Sometimes It can be very dangerous to a person or even for a whole community. So having a system to detect these fake news becomes the need of the hour. In this paper we are trying to detect fake news using machine learning. We will evaluate and visualise the dataset and then finally we will be using different algorithms to evaluate the news as the fake one or the real one. The results of this project demonstrate the ability for machine learning to be useful in this task. We have built a model that catches many intuitive indications of real and fake news as well as an application that aids in the visualization of the classification decision.

Introduction

Fake news is news designed to deliberately spread hoaxes, propaganda and disinformation. It denotes a type of yellow journalism which intentionally presents misinformation or hoaxes spreading through both traditional print news media and recent online social media. This is done to further impose certain ideas and is often achieved with political agendas. Often , fake news will mimic real headlines and twist the story. Fake news may be a relatively new term but it is not necessarily a new phenomenon. Fake news has technically been around at least since the appearance and popularity of one-sided, partisan newspapers in the 19th century. However, advances in technology and the spread of news through different types of media have increased the spread of fake news today. As such, the effects of fake news have increased exponentially in the recent past and something must be done to prevent this from continuing in the future.

The dangerous effects of fake news, as previously defined, are made clear by events such as [5] in which a man attacked a pizzeria due to a widespread fake news article. This story along with analysis from [6] provide evidence that humans are not very good at detecting fake news, possibly not better than chance . As such, the question remains whether or not machines can do a better job. There are two methods by which machines could attempt to solve the fake news problem better than humans. The first is that machines are better at detecting and keeping track of statistics than humans, for example it is easier for a machine to detect that the majority of verbs used are “suggests” and “implies” versus, “states” and “proves.” Additionally, machines may be more efficient in surveying a knowledge base to find all relevant articles and answering based on those many different sources. Either of these methods could prove useful in detecting fake news, but we decided to focus on how a machine can solve the fake news problem 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.

By collecting examples of both real and fake news and training a model, it should be possible to classify fake news articles with a certain degree of accuracy. The goal of this project is to find the effectiveness and limitations of language-based techniques for detection of fake news through the use of machine learning algorithms. The outcome of this project should determine how much can be achieved in this task by analyzing patterns contained in the text and blind to outside information about the world.

Fake news features

- They often have grammatical mistakes.
- They are often emotionally colored.
- They may try to affect the reader's opinion on some topics.
- They often use attention seeking words and news format and click bait.
- They are too good to be true.
- Their sources are not genuine most of the times

We are using the python programming language to implement this. We used numpy, pandas, and sklearn libraries. We also used seaborn for dataset visualization.

Dataset

We have taken the dataset from kaggle. There are two files, one for real news and one for fake news (both in English) with a total of 23481 fake news and 21417 real news.

Each file in the dataset has 4 columns.

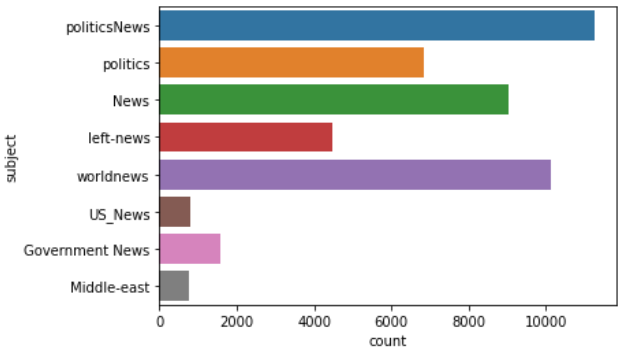
1. **Title** : News headline
2. **Text** : The actual text content of the news.
3. **Subject** : News type
4. **Date** : created date

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016

23481 rows × 4 columns

Features will be extracted from these four basic components. Fake news is used to influence the consumer, and in order to do that, they often use a specific language in order to attract the readers. On the other hand, non-fake news will mostly stick to a different language register, being more formal. This is linguistic-based features, to which can be added lexical features such as the total number of words, frequency of large words or unique words.

Visualisation




```
subject
Government News      1570
Middle-east           778
News                 9050
US_News               783
left-news            4459
politics             6841
politicsNews        11272
worldnews           10145
Name: text, dtype: int64
```

Before moving on first we made an extra column in both of our dataset files. We marked 0 for fake news and 1 for real news.

Fake news

	title	text	subject	date	flag
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
...	...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

23481 rows × 5 columns

Real news

	title	text	subject	date	flag
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1
...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

21417 rows × 5 columns

After this we merged both the files into a single file which now has both fake and real news data. Now we have 44898 news entries in the dataset.

	title	text	subject	date	flag
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
...	...	...	...	...	...
44893	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
44894	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
44895	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
44896	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
44897	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

44898 rows × 5 columns

Finally we also shuffled the data and removed and stored 20 news data to manually verify our ml model.

Data Cleaning

1. Removed duplicate data. After this only 44669 news entries remain.
2. We merged subject, title and text columns into a single news_data column
3. Now removed subject, title and text columns
4. Covert news text in lowercase and remove punctuation special chars., extra space, urls
5. Finally checked for null values in the dataset.

After this data cleaning we only have two columns in our dataset

Flag : fake or real

News_data : subject + title + text

	flag	news_data
26593	1	politicsNewsExclusive: Trump administration ey...
14664	0	politicsBREAKING: OBAMA JUST RELEASED GITMO Pr...
30531	1	politicsNewsU.S. to continue supporting engage...
30304	1	politicsNewsCommentary: Trump can't fight Isla...
11248	0	politicsVICE PRESIDENT PENCE BREAKS TIE In Bil...

Fake News Detection

First we define our dependent and independent variables

```
# independent variable
x = news["news_data"]
# dependent variable
y = news["flag"]
```

We splitted the dataset into 75 % training and 25% testing datasets.

Convert a collection of raw documents to a matrix of TF-IDF features. TF-IDF which stands for Term Frequency – Inverse Document Frequency. It is one of the most important techniques used for information retrieval to represent how important a specific word or phrase is to a given document**.**

Term Frequency

The number of times a word appears in a document divided by the total number of words in the document. Every document has its own term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Inverse Data Frequency

The log of the number of documents divided by the number of documents that contain the word w. Inverse data frequency determines the weight of rare words across all documents in the corpus.

$$idf(w) = log(\frac{N}{df_t})$$

Lastly, the TF-IDF is simply the TF multiplied by the IDF.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

So now we have converted our raw dataset into a matrix of TF-IDF features. We are ready for fake news detection using four ml algorithms

1. Logistic Regression
2. Decision Tree Classification
3. Gradient Boosting ClassifieR
4. Random Forest ClassifieR

Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

This is the final result of our logistic regression implementation

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5893
1	0.99	0.99	0.99	5275
accuracy			0.99	11168
macro avg	0.99	0.99	0.99	11168
weighted avg	0.99	0.99	0.99	11168

Decision Tree Classification

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

We have used gini classification criteria. This is the final result of our decision tree classification

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5893
1	1.00	1.00	1.00	5275
accuracy			1.00	11168
macro avg	1.00	1.00	1.00	11168
weighted avg	1.00	1.00	1.00	11168

Gradient Boosting Classifier

Gradient Tree Boosting or Gradient Boosted Decision Trees (GBDT) is a generalization of boosting to arbitrary differentiable loss functions. GBDT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems in a variety of areas including Web search ranking and ecology.

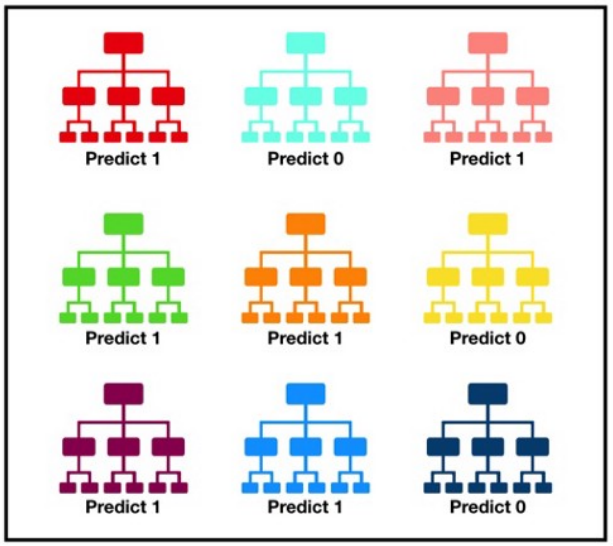
This is the final result of our gradient boosting classification

	precision	recall	f1-score	support
0	1.00	0.99	1.00	5893
1	0.99	1.00	1.00	5275
accuracy			1.00	11168
macro avg	1.00	1.00	1.00	11168
weighted avg	1.00	1.00	1.00	11168

Random Forest Classifier

In random forests each tree in the ensemble is built from a sample drawn with replacement from the training set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of size max_features.

The purpose of these two sources of randomness is to decrease the variance of the forest estimator. Indeed, individual decision trees typically exhibit high variance and tend to overfit. The injected randomness in forests yields decision trees with somewhat decoupled prediction errors. By taking an average of those predictions, some errors can cancel out. Random forests achieve a reduced variance by combining diverse trees, sometimes at the cost of a slight increase in bias. In practice the variance reduction is often significant hence yielding an overall better model.

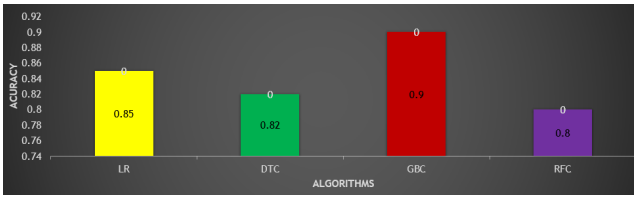


Tally: Six 1s and Three 0s
Prediction: 1

This is the final result of our Random forest classification

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5893
1	0.99	0.99	0.99	5275
accuracy			0.99	11168
macro avg	0.99	0.99	0.99	11168
weighted avg	0.99	0.99	0.99	11168

Results



References

1. Pallavi B. Petkar and S. S. Sonawane, “Fake News Detection: A Survey of Techniques,” International Journal of Innovative Technology and Exploring Engineering, pp. 383-386, Vol. 9, 2020.
2. Fake Data Analysis and Detection Using Ensembled Hybrid Algorithm “https://ieeexplore.ieee.org/document/8819741/references#references”
3. Fake News Detection by Decision Tree “https://ieeexplore.ieee.org/document/9249688”
4. Kaggle, Fake News Detection, Kaggle, San Francisco, CA, USA, 2019, https://www.kaggle.com/jruvika/fake-news-detection.
5. Kaggle, False News Detection, 2018, https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset
6. Scikit-learn ML in Python Userguide https://scikit-learn.org/stable/index.html