Big Data Analytics

Project 1 Report - Team 4

Fake News Detection with Machine Learning

Rushikesh Naidu, Anuj Vyas, Harsh Pathak, Nguyen Vo

Abstract :-

News can sometimes be highly influential medium to influence individual into believing a theory. In order to be fair to the context in the news, every news item must be free from bias and true. However over many years, there has always been the presence of fake news spreading around the world influencing and motivating several people to believe in misleading things. From hoax to conspiracies there has been an array of fake news on the internet. We plan to test using machine learning and deep learning models whether a given news is fake indeed. In order to do so we crawl our source website https://www.snopes.com/ which itself dynamically defines if a news is fake or true. We base our performance with the ground truth from Kaggle Datasets. We release our source code and dataset used in this project in [11]

---

1. Introduction

Fake news is a neologism often used to refer to fabricated news. This type of news, found in traditional news, social media or fake news websites, has no basis in fact, but is presented as being factually accurate. Fake news is written and published with the intent to mislead in order to damage an agency, entity, or person, and/or gain financially or politically, often using sensationalist, dishonest, or outright fabricated headlines to increase readership, online sharing, and Internet click revenue. In the latter case, it is similar to sensational online "clickbait" headlines and relies on advertising revenue generated from this activity, regardless of the veracity of the published stories.

Fake news detection on social media has unique characteristics and presents new challenges. First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult to detect based on news content. Thus, we need to include auxiliary information, such as user social engagements on social media, to help differentiate it from the true news. Second, exploiting this auxiliary information is nontrivial in and of itself as users' social engagements with fake news produce data that is big, incomplete, unstructured, and noisy.

We will crawl website snopes.com for articles which have been debunked as true or false using Scapy tool. After that, we do some processing to extract content of collected articles. Finally, we build machine learning to detect fake news. Our machine learning model is based on Deep Learning models proposed in paper [1] .

Different from [1], we use attention mechanism to learn better latent representation of each article for better classification result. We may use hierarchical LSTM to better classify articles. Furthermore, our work uses Snopes dataset which is different from dataset used in [1].

The anonymity and ambiguity of fake news allows us to define it in a broad spectrum. In order to understand and detect a fake news we need to analyze what type of issue does the fake news detection face. Mentioned below are a list of different types of fake news and inorder to successfully fulfil our project goals we need to tackle all of these issues.

Figure 2. Future directions and open issues for fake news detection on social media

- Data-oriented: it focuses on different aspects of fake news data, such as benchmark data collection, psychological validation of fake news, and early fake news detection.
- Feature-oriented: it aims to explore effective features for detecting fake news from multiple data sources, such as news content and social context.
- Model-oriented: it opens the door to build more practical and effective models for fake news detection, including supervised, semi-supervised and unsupervised models.
- Application-oriented: it encompasses research that goes beyond fake news detection, such as fake new diffusion and intervention.

3. Related Work

To achieve the goal of detecting fake news, many research work are put to build machine learning model to detect credibility of news. Two most earlier work on this domain is [5][6]. However, these work employed manual engineering features which is time consuming to derive. In this proposal, we aim to build deep learning models to detect fake news. In particular, we will learn latent features of articles collected on Snopes.com. There are several work about detecting fake news on using Deep learning. Typically, [7] proposed a hybrid model to detect fake news. We simply use models proposed in [1] with some modifications such as hierarchical model. etc.

2. Motivation

Fake news is prevalent on social media due to its negative impact on presidential election in 2016. It is important to detect whether a news is fake or not. The end users that benefit from this project is the whole society since people want to know what is fake and what is real automatically.

3. Method

3.1 Dataset Source :-

The dataset was obtained by crawling our source website which is https://www.snopes.com/. The website provides news along with checking facts to categorize a news into a particular category. There are various categories including History, Crime, Entertainment and Fake News. The news ahs various sources and these sources are operated on to check and categorize them into particular categories.

3.2 Data Gathering :-

We select dataset from Kaggle available at https://www.kaggle.com/arminehn/rumor-citation/data# due to the available ground truth. We further collect the content of webpages to build machine learning model. We chose Snopes website to crawl web pages' content since this website is more famous.

3.3 Data Description (Web Crawling) :-

We use requests package in Python to crawl totally 562 webpages. To avoid being blocked by Snopes website due to abusing behavior, we make our crawler sleep 30% after crawling 10 webpages. The crawling process was pretty fast. It took us around 1 day to get all html pages.

3.4 Text Scraping :-

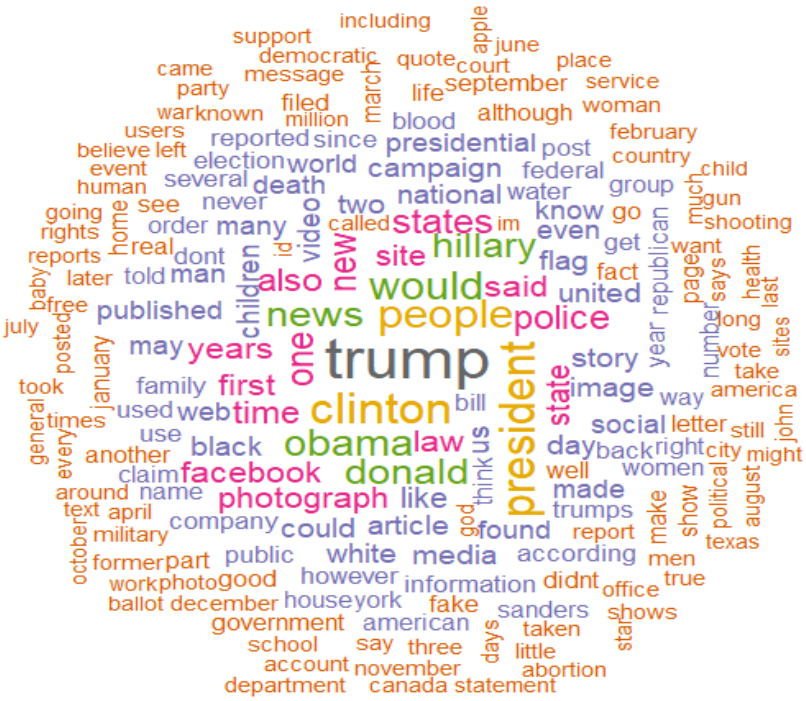
Once we have crawled the html files to our database from the website, the files contain all the information about the website informations as well as other information that we do not require. Thus it is important to remove the text from the paragraphs of the news body. In order to do this we used the

BeautifulSoup library in Python. Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

The result from performing operations using beautiful soup on the datasets, we obtain the text files containing the body of all 561 news crawled from our source website.

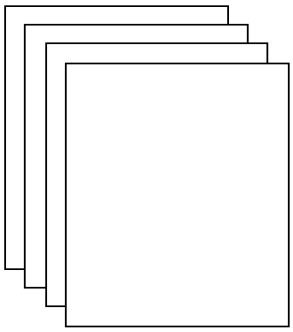
3.5 Data Analysis (Word Cloud) , (Topic Modelling) :-

In the figure, we plot top 200 words with highest tf-idf value. As we can see, the most popular words are about trump, clinton, obama. These words are mostly about presidential election in 2016.

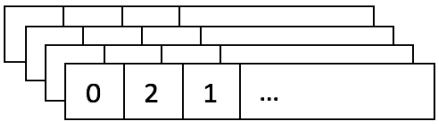


Topic Modelling:

Topic modelling is a method through which we can understand the pattern present in the text corpus. Thus help us in better decision making. It is a field of text mining. There are different methods of obtaining topics from a text but Latent Dirichlet Allocation (LDA) is the most popular method. LDA process the documents and then create a backtrack to figure out what topics would be created through those documents. LDA is matrix factorization method with Documents N as D1, D2….Dn and vocabulary size of V words as W1, W2….. Wn. LDA changes matrix into two lower dimensional networks – V1 and V2. V1 is a document-topics network and V2 is a subject – terms framework with measurements (N, K) and (K, V) separately, where N is the number of reports, K is the number of themes and M is the lexicon size.



N documents constructed from a vocabulary of V words



Represented as N V-dimensional vectors (bag-of-word)

Goal: Find the document and topic vectors which explain the observed data

https://www.datacamp.com/community/tutorials/lda2vec-topic-model

Results:

```
Topic 0:
trump donald people said news president new like social media
Topic 1:
article clinton united trump president obama web people new said
Topic 2:
news web published article trump said donald people president social
Topic 3:
obama president new article trump news web united facebook like
Topic 4:
president obama new said united news published states people time
Topic 5:
people new time states united said like media article social
```

Word cloud for the LDA2VEC Topic modelling:

4. Analytics

4.1 Machine Learning Models :-

4.1.1 Multinomial NB

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector $x_1$through x_n, Bayes' theorem states the following relationship:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Using the naive independence assumption that

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$ for all i,

This relationship is simplified to

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters. (For theoretical reasons why naive Bayes works well, and on which types of data it does, see the references below.)

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

4.1.2 Logistic Regression

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. Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest.

4.1.3 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

4.1.4 XG Boost

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solved our problem in a fast and accurate way. It iteratively combines a number of “weak” learners to create a strong classifier. The weak learner here is a decision tree. The learners are added based on a loss function which is to be minimized at each iteration. It regularizes model formalization to control over-fitting. It took two seconds to build a model on the training data.

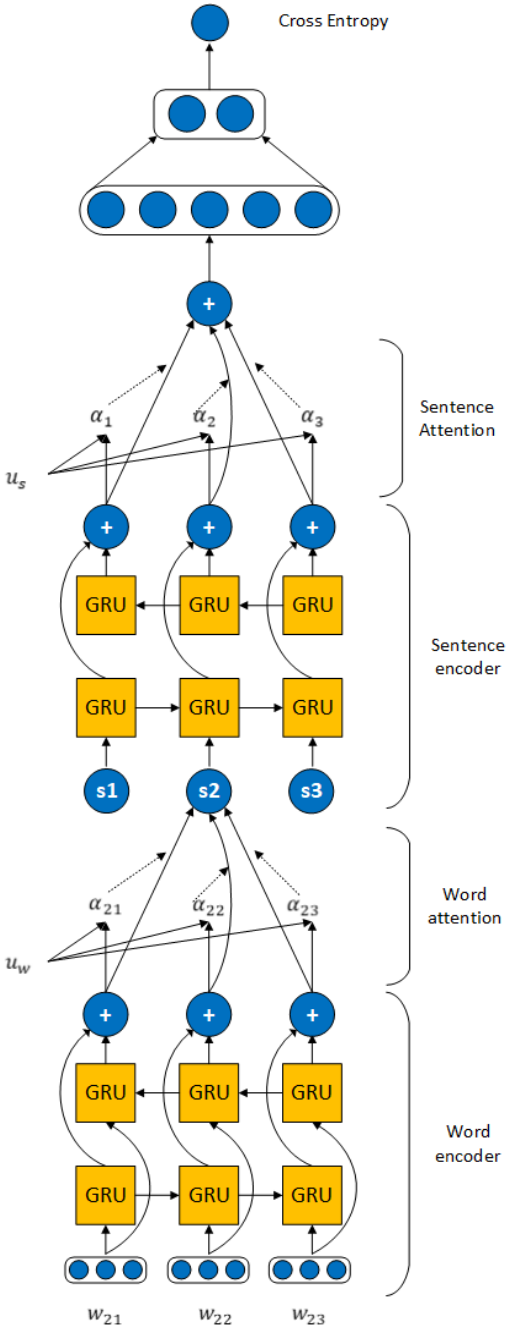
4.2 Deep Learning Model :-

In this subsection, we describe motivation and our Deep Learning model we use for fake news detection. Due to the rise of big data, manual engineering features for machine learning technique becomes an intensive task. Therefore, the need to automatically extract features to build machine

learning models attracts significant attention from community. Deep learning model with the ability to learn high quality latent representation of text, images, audio and videos are the best choice for this task. In this project, we aim to build several deep learning models and compare them against traditional machine learning models. We propose to use bi-directional GRU and attention model to classify if a news is fake or not. The architecture of our proposed is as follows. This architecture consists of 5 main components.

- The first layer is Word encoder in which we look up vector representation of words based on word2vec model. We use pre-trained word vectors at https://github.com/mmihaltz/word2vec-GoogleNews-vectors/blob/master/GoogleNews-vectors-negative300.bin.gz
- The second layer is Word Attention. The main intuition is that in each sentence, some words may play different roles in expressing the truthfulness of the sentence.
- The third layer is Sentence Encoder. From vector representation of words, we form sentence representation by summing all word vectors based on attention weight
- The fourth layer is Sentence Attention. With similar intuition with Word Attention, the truthfulness of a document depends on some key sentences only. Therefore, we downgrade sentences that are not discriminate in classifying fake news.
- The final layer is softmax layer, we sum up all sentence vectors in a news and forward it into a fully connected layer. In this layer, we used DropOut = 70% as a way to regularize the network. At the end of this layer, we use Cross Entropy as a loss function to optimize the network.

We train our model with mini-batch size = 50 documents. Each document, we limit at most 100 sentences. Each sentence we limit at most 100 words.


Cross Entropy
α_1
α_2
α_3
u_s
Sentence Attention
GRU
Sentence encoder
s1
s2
s3
α_{21}
α_{22}
α_{23}
u_w
Word attention
Word encoder
w_{21}
w_{22}
w_{23}

5. Results and Evaluation

5.1 Multinomial NB

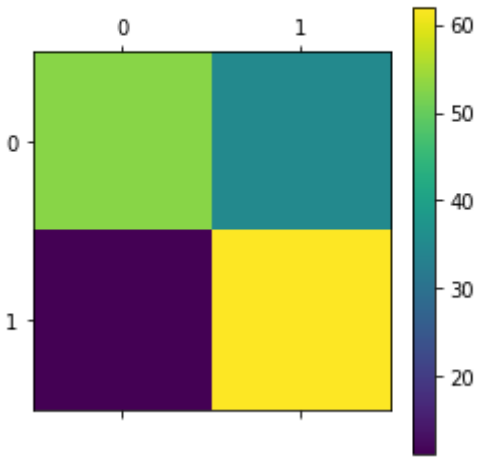
This method outperformed other machine learning models and was comparable to the deep learning model.

Test Accuracy : 71.5%

Confusion Matrix :

	precision	recall	f1-score	support
False	0.83	0.60	0.70	88
True	0.64	0.85	0.73	73
avg / total	0.74	0.71	0.71	161

```
[[53 35]
 [11 62]]
```

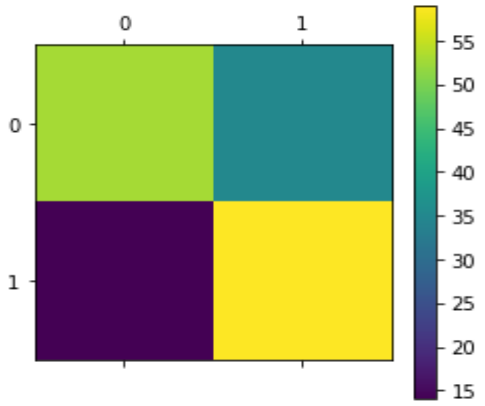


5.2 Logistic Regression

Test Accuracy : 69%

	precision	recall	f1-score	support
False	0.79	0.60	0.68	88
True	0.63	0.81	0.71	73
avg / total	0.72	0.70	0.69	161

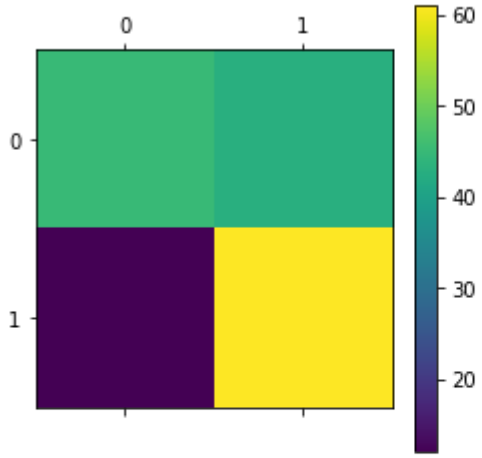
```
[[53 35]
 [14 59]]
```



5.3 Random Forest

Test Accuracy : 65%

```
[[45 43]
 [12 61]]
```

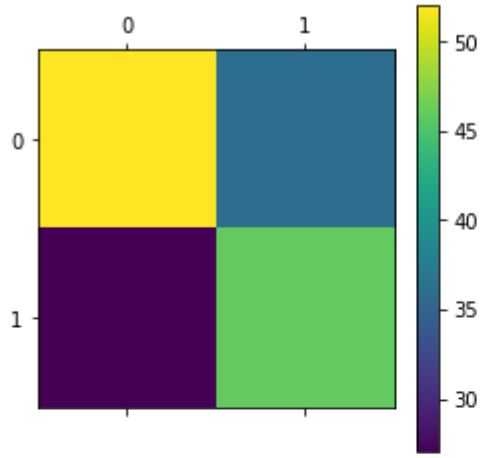



```
print(classification_report(test_labels, pred))
```

	precision	recall	f1-score	support
False	0.79	0.51	0.62	88
True	0.59	0.84	0.69	73
avg / total	0.70	0.66	0.65	161

5.4 XG Boost

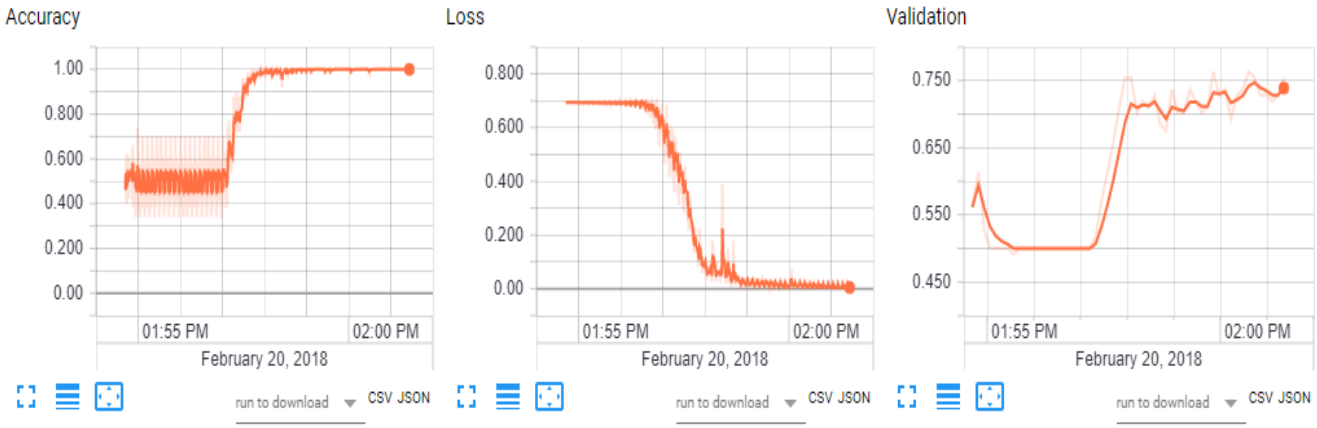
Test Accuracy : 60.5%



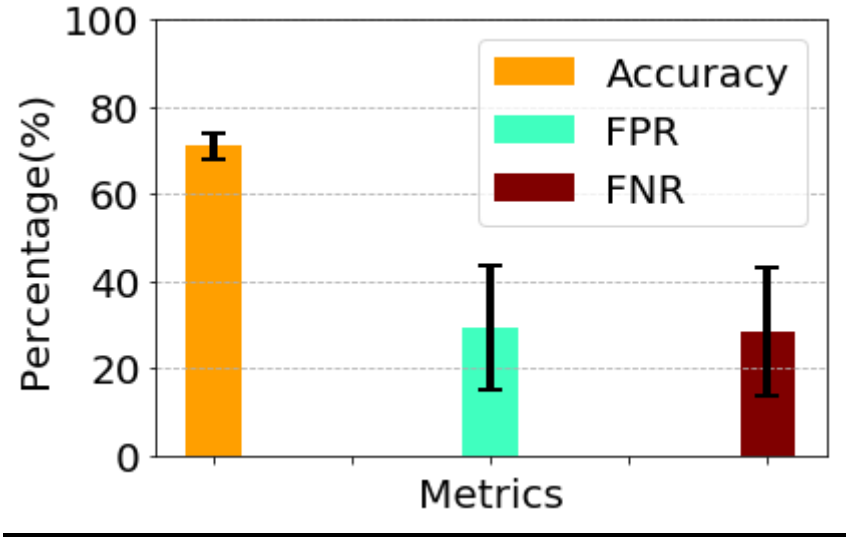

```
print(classification_report(test_labels, pred))
```

	precision	recall	f1-score	support
False	0.66	0.59	0.62	88
True	0.56	0.63	0.59	73
avg / total	0.61	0.61	0.61	161

5.5 Deep Learning Model



This figure shows the accuracy and loss in training set. As we can see, the accuracy in training set reached almost 100% and loss is almost zero. The validation accuracy increased over time.



Mean acc: 71.05%, Mean FP: 29.47% , Mean FN: 28.42%

As we can see that performance of the Bidirectional GRU with Attention Mechanism shows the best performance among all machine learning models even though the dataset is quite small.

To further understand the attention mechanism, we show the visualization in the following 6 figures. In particular, we visualize the weights of sentences based on their contribution in detecting fake news (i.e. attention weight of sentences) and we also visualize the weights of words in each sentence. The red color stands of importance of sentences and the green color stands for the weights of words.

As we can see, the attention mechanism is super useful in recognizing the most important sentences and the most important words inside each sentence. The observation is that if a document consists of many important sentences, our deep learning model can classify it correctly. If there is limited important words and sentences, it is likely that model will be likely misclassify it (See Document 6 as an example)

Document 0 - Label: Fake News - Predicted: Fake

musician hank williams endorsed hillary clinton .

june nevada county scooper published article reporting musician hank williams endorsed hillary clinton president united states act would quite given williams long politically involved republican country music artist hank williams son legend hank made startling .

williams withdrawn support republican candidate donald endorsed democrat nominee hillary .

move shocks williams fans .

article piece fabricated .

nevada county scooper fake news web disclaimer scooper satirical website scope .

sometimes often .

provide fake news social criticism satirical .

intention fool trick obviously firmly believe persons willingness listen injecting yes sometimes .

although hank williams endorsed hillary clinton writing country music star pledged support gop candidate donald trump telling rolling stone february .

Document 1 - Label: Fake News - Predicted: Fake

police found cow eyeballs rectum arresting drunk .

july web site crazed published article reporting police wyoming found cow eyeballs man named roy tilbotts rectum stopped suspicion police made routine traffic stop early thursday morning got bargained roy stepped el camino field sobriety test casper police noticed several eyeballs slide right pant leg onto .

feeling could potential murderer police quickly drew guns cuffed .

tilbott assured police eyeballs instead cow eyeballs johnson meats tilbott employed .

company wont let us take animal home instead toss tilbott said police .

theyre .

allowed take scrap meat parts .

company start green .

dont even recycling .

truth another piece satire fake news web .

Document 2 - Label: Fake News - Predicted: Fake

people die younger .

may scientific journal nature published brief report two psychologists titled live .

itpurported demonstrate statistically significant difference longevity right left based data collected professional baseball investigate relationship handedness age analysed baseball players listed baseball encyclopedia dates birth well throwing batting .

subject assigned handedness group throwing batting hand indicated change hand .

mean age death years .

years .

authors published another life equally prestigious new england journal medicine sampling death records year two california demonstrated argued even larger difference lifespan left right order test relation handedness life span general obtained death certificates two counties southern .

two thousand questionnaires concerning handedness deceased family member sent listed next resulted usable cases male subjects female .

subjects designated threw ball right .

subjects assigned .

Document 3 - Label: True News - Predicted: Legitimate

campaign ad president lyndon johnson featured purported republican voter expressing concerns eerily echoed threads debate gop march facebook page web site quartz shared video calledconfessions originally political advertisement clip rapidly gained along skepticism viewpoints expressed neatly echoed political schisms .

var fjs js confessions republican ad presidential election going thanks uncanny relevance presidential .

posted quartz march speaker bill discussed lifelong identity loyal republican voter expressing reservations candidacy barry unsuccessfully challenged president lyndon johnson concerns aired text bottom described tendency goldwater rapidly reverse position deny statements key well attitude duringa point american historywhen specterofnuclear destruction hardest thing whole campaign sort one goldwater statement .

reporter go senator goldwater hell blah blah blah whatever end .

goldwater wouldnt put .

cant follow .

serious put .

serious says wouldnt put .

dont get .

president ought mean .

Document 4 - Label: True News - Predicted: Legitimate

december amateur photographer snapped pictures snow ground algerian city borders sahara .

december amateur photographer posted album facebook images capturing snow algerian town located hills bordering sahara several online news outlets published stunning .

according last significant snowfall occurred area february last major snowfall called hit sefra february snowed whopping .

subsequent snow also appeared viewers expressed skepticism authenticity photographs likelihood town located next sahara would see snow first .

fact snowed sefra december via nasa satellite data shows pockets snow snowfall indeed rare occurrence temperatures led air cool enough produce snow .

historical data region show december january temperatures often dip .

Document 6 - Label: Fake News - Predicted: Legitimate

video clip shows meteorite striking pickup truck .
group recreationists videotaping desert outing suddenly spot something unusual sky .
flaming ball smashes directly ground around passing pickup blast sending cameraman .
even truck apparently unscathed promptly starting driving cheers .
astounding .
another case edited television commercial presented .
clip taken tv commercial advertising toyota produced automaker method .
full clip plainly displays toyota name logo .

Based on these visualizations, we conclude that attention mechanism is helpful in detecting fake news. Especially, it is valuable to visualize what was learned in deep learning model instead of considering it as black box. This paves the way for more interpretable deep learning models in the future.

5.6 Evaluation:-

Model	Testing Accuracy
XG Boost	60.5%
Random Forest	65%
Logistic Regression	69%
Deep Learning Model	71.05%
Multinomial NB	71.5%

6. Conclusion

In this project we build many machine learning models to detect fake news. We employ a bidirectional Hierarchical GRU with attention mechanism to detect fake news. The performance of Deep Learning is the best even though the training dataset size is small. In the future work, we will integrate more datasets to see if performance can be improved.

7. References:

[1] William Yang Wang "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection ACL 2017

[2] Andreas Vlachos, Sebastian Riedel, Fact Checking: Task definition and dataset construction ACL 14

[3] William Ferreira, Andreas Vlachos, Emergent: a novel data-set for stance classification NAACL-HLT 2016

[4] https://www.kaggle.com/mrisdal/fake-news

[5] Carlos Castillo, Marcelo Mendoza, Barbara Poblete, Information Credibility on Twitter, WWW 2011

[6] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, Qiaozhu Mei, Rumor has it: Identifying Misinformation in Microblogs EMNLP 2011

[7] Natali Ruchansky, Sungyong Seo, Yan Liu CSI: A Hybrid Deep Model for Fake News Detection CIKM 2017

[8] Zichao Yang, Diyi Yang , Chris Dyer , Xiaodong He , Alex Smola , Eduard Hovy Hierarchical Attention Networks for Document Classification NAACL 2016

[9] https://www.kdnuggets.com/2017/10/guide-fake-news-detection-social-media.html

[10]Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 2017. Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), pp.22-36.

[11] https://github.com/nguyenvo09/fake_news_detection_deep_learning

References for concepts/code/data sets:-

1. https://github.com/nguyenvo09/fake_news_detection_deep_learning
2. http://www.fakenewschallenge.org/
3. https://nycdatascience.com/blog/student-works/identifying-fake-news-nlp/
4. https://www.dataquest.io/blog/web-scraping-tutorial-python/