CIS 520 Fall 2020, Project Checkpoint

Fake News Detection

Team Members:

- Haowei Liu; Email: `hwliu@seas.upenn.edu`
- Yuchen Zhang; Email: `zycalice@seas.upenn.edu`

Abstract

With the popularization of digital media, news has become readily available to human at a click of finger. This development drastically increased our access to information, information that are non-discriminatory in nature. The multitude of sources that are more often than not contradictory has created new challenges for human - what should we believe in. To tackle this problem, we explore a machine learning approach in this project. Using news that have been flagged by human as either fake or real as training dataset, we design classification models that can identify the legitimacy of a news article given its content. Specifically, we apply natural language processing techniques for text feature extraction, and consequently classification models for fake news detection. Our best-performing model reaches prediction accuracy of over 99%. We understand that our estimate for prediction performance might be limited to the scope of our data sources, and we welcome suggestions and applications of our models on different datasets.

1 Motivation

In the current era of information overflow, identifying information validity becomes a challenging task for news outlet, social media, and just the general public. This process becomes increasingly more relevant when platforms or individuals voice opinions incentivized by personal interests, and are not as much concerned about verifying these information. We see this phenomenon intensified in the midst of political turmoil. Therefore, we believe it's important to build a scalable model to preemptively identify fake news from more legitimate ones, which would otherwise be too inefficient to go through with human scrutiny.

TODO: elaborate on application For social media and news collection agencies, this would help them to flag unreliable news. For finance/investment companies, this could help them to have a better sense of the sentiment and to predict the financial markets more accurately.

2 Dataset

2.1 Dataset Introduction

With this goal in mind, we have located a dataset on Kaggle titled, **Fake and real news dataset**.[1] Formally, the dataset is called "ISOT Fake News Dataset".[2] It consists of a total of about 40,000 news articles, and manually annotated label of either real or fake news. This is a fairly balanced dataset, around 50% of the news article classified as fake news. This dataset was collected from real-word sources dated 2015 to 2017 - the real news had been obtained by crawling Reuters.com, whereas the fake news are collected from unreliable sources as flagged by PolitiFact and Wikipedia.[1].

News	Number of Articles
Real news	21,417
Fake news	23,481

Table 1: Data Summary

In addition to the labels, the provided features includes:

1. **title**: Title of the article.
2. **article**: The full article itself, including punctuation.
3. **subject**: For fake news, the subjects include "GovernmentNews", "Middle-east", "US News", "left-news", "politics", "News". For true news, the subjects include "World-News" and "PoliticsNews".
4. **date**: publication date of news.

2.2 Variables Exploration

2.2.1 Labels - Target variable

It is crucial for us to interpret how labels are created, which allows us to better understand our models and their applicability to new data.

According to the collection methodology section on Kaggle, the "true" articles are all from reuters.com.

[1]Kaggle dataset source, `https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset`.

[2]Kaggle dataset metadata - sources, `https://www.uvic.ca/engineering/ece/isot/assets/docs/ISOT_Fake_News_Dataset_ReadMe.pdf`

For "fake articles", they were collected from unreliable websites that were flagged by PolitiFact and Wikipedia. According to PolitiFact, below are the questions they consider **when choosing which claims to fact-check:**

- Is the statement rooted in a fact that is verifiable? (the team does not check opinions)
- Does the statement seem misleading or sound wrong?
- Is the statement significant?
- Is the statement likely to be passed on and repeated by others?
- Would a typical person hear or read the statement and wonder: Is that true?

Politifact also listed **how the team determines their ratings**. As opposed to binary "true" and "false", the team actually had six ratings, ranging from true, mostly true, half true, mostly false, false, and pants on fire. Kaggle turned this into binary categories, and we currently did not find resources to validate Kaggle's methodology.

The editors and reporters at Politifact create the ratings by discussing the following questions:

- Is the statement literally true?
- Is there another way to read the statement? Is the statement open to interpretation?
- Did the speaker provide evidence? Did the speaker prove the statement to be true?
- How have we (PolitiFact's) handled similar statements in the past? What is PolitiFact's jurisprudence?

Here are some excerpt of the examples of "true" news (each paragraph is a separate piece of news):

```
'WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the
U.S. military starting on Monday as ordered by federal courts, the Pentagon said on Friday, afte
r President Donald Trump's administration decided not to appeal rulings that blocked his transge
nder ban. Two federal appeals courts, one in Washington and one in Virginia, last week rejected
the administration's request to put on hold orders by lower court judges requiring the military
to begin accepting transgender recruits on Jan. 1. A Justice Department official said the admini
stration will not challenge those rulings. "The Department of Defense has announced that it will
be releasing an independent study of these issues in the coming weeks. So rather than litigate t
```

```
'The following statements\xa0were posted to the verified Twitter accounts of U.S. President Dona
ld Trump, @realDonaldTrump and @POTUS.  The opinions expressed are his own.\xa0Reuters has not e
dited the statements or confirmed their accuracy.  @realDonaldTrump : - The Massive Tax Cuts, wh
ich the Fake News Media is desperate to write badly about so as to please their Democrat bosses,
will soon be kicking in and will speak for themselves. Companies are already making big payments
to workers. Dems want to raise taxes, hate these big Cuts! [0724 EST] - Was @foxandfriends just
named the most influential show in news? You deserve it - three great people! The many Fake News
Hate Shows should study your formula for success! [0745 EST] - Home Sales hit BEST numbers in 10
years! MAKE AMERICA GREAT AGAIN [0856 EST] - House Democrats want a SHUTDOWN for the holidays in
order to distract from the very popular, just passed, Tax Cuts. House Republicans, don't let thi
s happen. Pass the C.R. TODAY and keep our Government OPEN! [0952 EST] -- Source link: (bit.ly/2
jBh4LU) (bit.ly/2jpEXYR) '
```

Figure 1: True news examples

Note that the latter example centers around Tweets, which is not the norm among "true" news. As specified in data collection method, true news are scraped from Reuters news and as a result, the texts contain its source tag, **"(Reuters)"**. This piece of rather identifying text will be removed in the proceeding analysis to ensure robustness of model when applied to different information sources.

Here are also some excerpt of the examples for fake news (each paragraph is a separate piece of news):

```
'A Bernie Sanders supporter who slammed Hillary Clinton at her own rally in Iowa was kicked off
stage after urging attendees not to vote for Clinton because you can t trust her.Kaleb Vanfosso
n, president of Iowa State University s Students for Bernie chapter, trashed Clinton s ties to W
all Street and lack of empathy for the average American. She is so trapped in the world of the e
lite that she has completely lost grip on what it s like to be an average person,  Vanfosson sai
d, as cheers erupted from the audience.  She doesn t care.  He s no Trump fan but this was suppo
sed to be a glowing endorsement of Hillary Clinton  Read more: BPR'


'We ve written about the amazing filmmaker, playwright and journalist, Phelim MacAleer and his f
earless wife Ann McElhinney several times over the past couple of years. They ve brilliantly exp
osed the Left  and their outrageous lies countless times. They are currently making a movie call
ed  Gosnell  that will tell the truth the media hid about the crimes committed by serial killer,
Dr.  Kermit Gosnell. This husband and wife team are doing the work honest journalists, if they w
ere being true to their profession, SHOULD be doing. Phelim s play  Ferguson,  that is now a mov
ie, (see full-length version below) has completely dismantled the  hands up don t shoot  lie tha
t was coordinated by the White House and professional race agitators.  This lie of course, was p
erpetrated by a willing media. MacAleer does a brilliant job of dispelling these lies by simply
using the actual testimony from the case.From Phelim MacAleer:These are professional actors read


'You re gonna love this!The left has been screaming for decades about the government interfering
in their right to kill their babies. Because after all, even though it s a human life in its ear
liest stages of development; it s growing inside their womb, so the government has no business p
rotecting the innocent life of that baby in THEIR bodies.Strangely enough, the left is angry abo
ut Trump s decision to let the states decide if they will or will not force its citizens to shar
e a bathroom or public shower with a person who claims to be a certain gender contrary to what t
heir genitals say they are. Clearly the hypocritical left can t make up their mind when it s oka
y, or when it s not okay for the government to tell citizens what they can or cannot do '
```

Figure 2: Fake news examples

As shown in the examples above, the contents of fake news vary from politics news with strong partisan sentiment to rather non-sensible passages. We observe that the syntax of fake news, in comparison to real news, seems to be more conversational.

We have several observations. First of all, as indicated by the "subject" variable in the original dataset, the range of topics is broader for fake news. In our initial version, we did not include subject as a feature nor as a filter. In our second version, we attempt to include only domestic politics related news, to create models that are not as affected by the structure of the inputs.

2.2.2 Features/predictors

We have shown the significance of our analysis target in the previous section, we will now provide an overview of our predictors, namely title, text, news subject and publication date. [3]

- **Publication date**
  Given publication date of each news article, we first saw that all news were published during 2015 to 2017, as shown below. Although there seems to more real news collected in 2017, we don't think there exists reasonable explanation to this phenomenon except for data collection method. And we are not going to discriminate between articles publication year.

[3]There are 626 fake news observations and 1 true news entry with empty text fields. We have removed these incomplete observations prior to generating the following graphs and statistics.

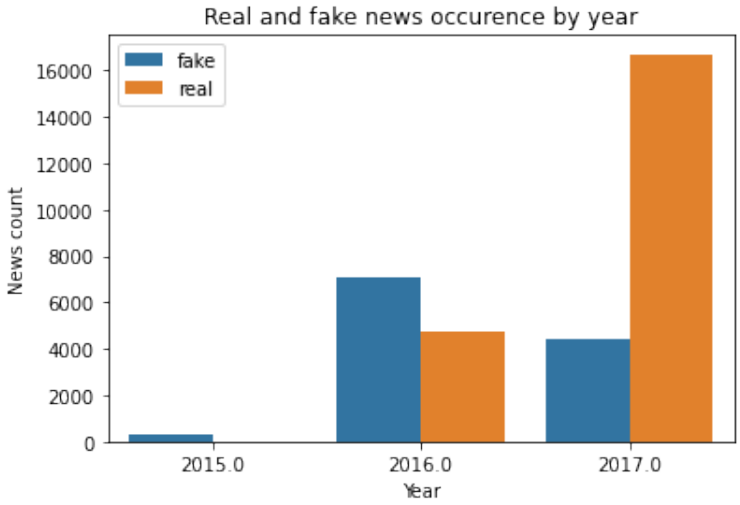


Figure 3: News publication - Year

On the other hand, we found some interesting pattern in the month and the day of week when the news is published. **Figure 4** shows that there are a lot more real news publication during the last four months of the year. We should be cautious of this trend since we only have three years' data, this, again, could be sensitive to how researchers collected these data. We also found that real news are more likely to have been published during weekdays. We suspect this might correspond to the majority of employees' schedules at more traditional publication organizations.

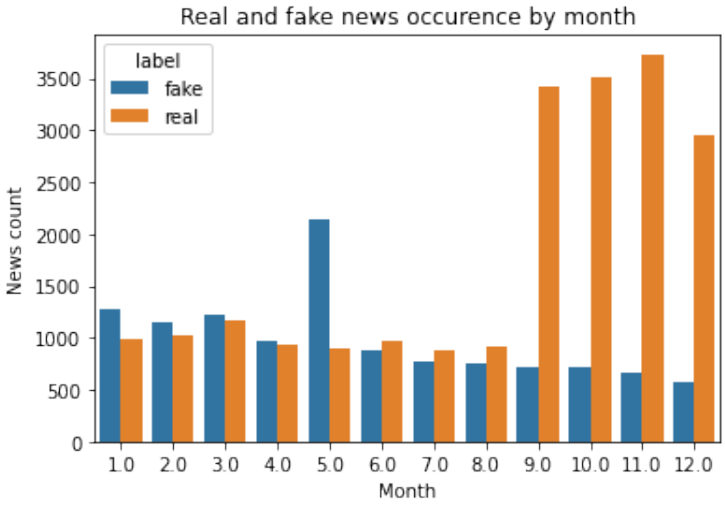


Figure 4: News publication - Month

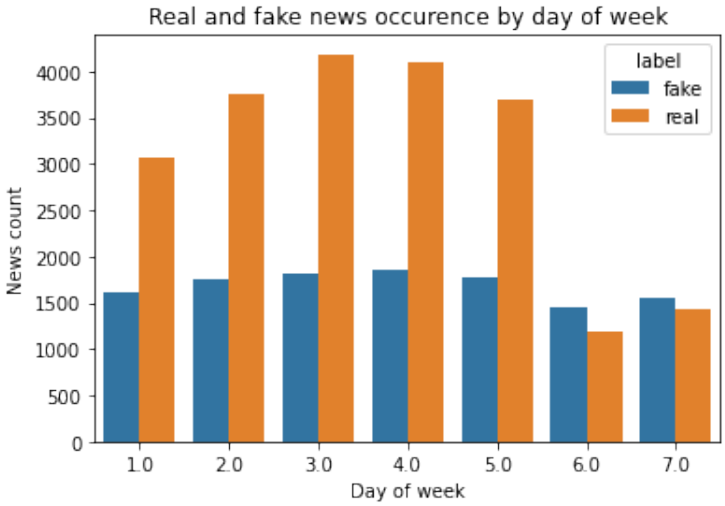


Figure 5: News publication - Day of week

- **Title**
  Given our inspection of news title, there doesn't seem to much interesting pattern. We did notice that only fake news titles tend to include an indicator of whether the news content includes video by adding **(VIDEO)** at the end of sentence. This practice results in a dominant occurrence of the word "VIDEO". We think this could be a strong but undesirable indicator of fake news since it's not related to news content, therefore we removed the trailing (VIDEO) in fake news titles.

  Furthermore, to visualize the vocabulary occurrences in fake and real news titles, we created the following word clouds graphs separately for real and fake news title.[4]

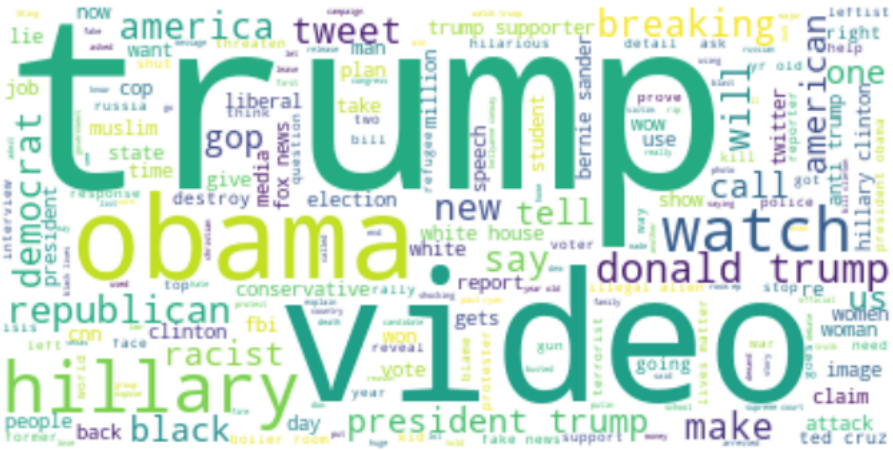


Figure 6: Fake news title - word cloud

[4]Words/phrases were converted to lowercase prior to generating graphs. This applies to the title wordcloud graphs as well.

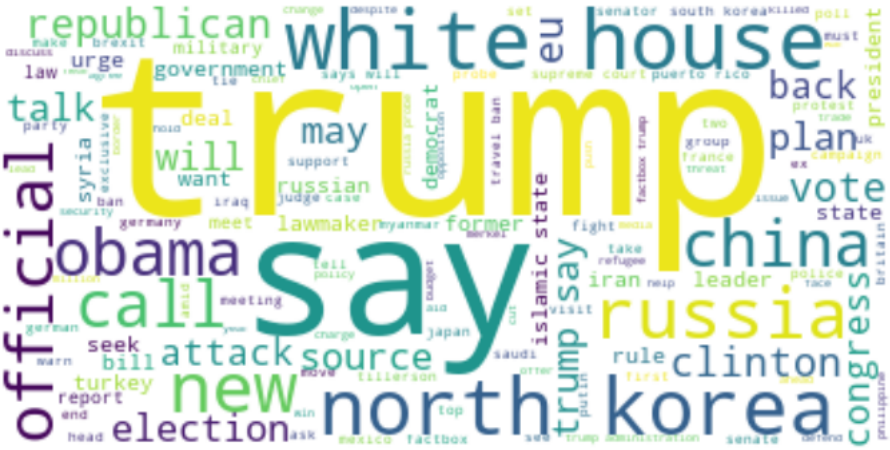


Figure 7: True news title - word cloud

- **Text**
  We have shown some examples of news text in the previous section, see Figure 1 and Figure 2. We think it's not exactly challenging for an educated adult to tell apart most real and fake news, since fake news tend to have more inciting sentiments and less rigorous sentences. The question remains whether machine learning models can learn to do the same thing by studying term frequencies and sentence embeddings. We demonstrate here the most frequent words/phrases in fake and real news texts.

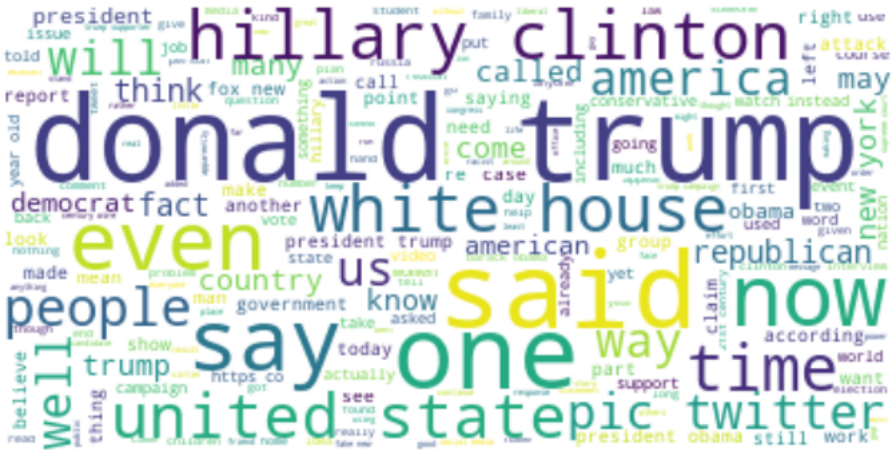


Figure 8: Fake news text - word cloud

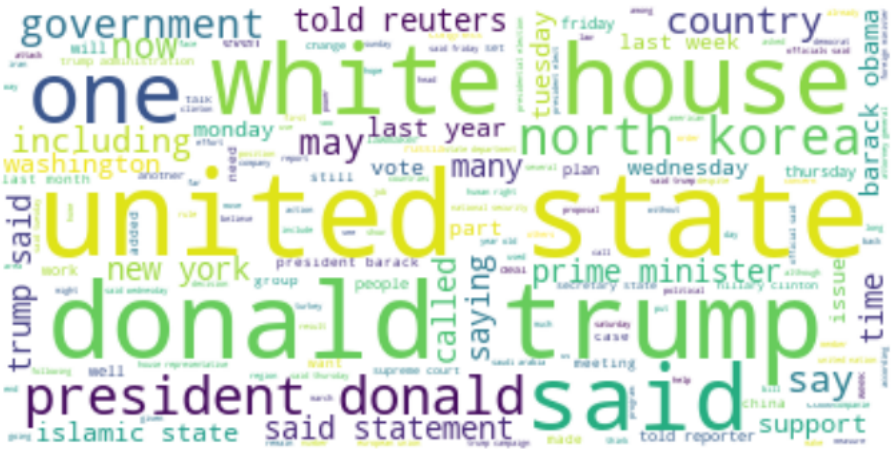


Figure 9: True news text - word cloud

From the 4 figures above, we see that phrases such as "donald trump", "white house", "united states", and other country names are frequent in both fake and news contents. This is to be expected since our data was collected from a pool of politics-centered news between 2015 and 2017. At the same time, there are words such as "twitter" and "breaking" that are more prominent in fake news, while "official" and "government" are highly used in real news. We expect our text feature extraction mechanisms to pick up these differences.

- **Subjects**
  Finally, we examined the news subjects reported in the dataset. We noticed that different categorization rules were applied to fake and real news, as illustrated in Figure 10.

News	9050
politics	6435
left-news	4310
Government News	1499
US_News	783
Middle-east	778

(a) Fake news subjects and occurrence counts

politicsNews	11271
worldnews	10145

(b) True news subjects and occurrence counts

Figure 10: News subject comparison

As demonstrated above in the label section, these subjects are different due to the different data collection processes for fake news and true news. We understand these categorization are not exactly informative, and the categories are not mutually exclusive among fake news. Thus, we decided not to include subjects as a feature. However, by investigation, we notice that the contents of *non-politics* fake and real news vary quite a lot, and we suspect that our model may learn to differentiate between real and fake news using content-specific information. This may be useful in some cases, but it would be also desirable to a classifier that performs well with more homogeneous contents, for instance identifying *politics-related* fake news. Therefore, we plan to experiment with subsets of data corresponding to politics subjects.[5]

2.2.3 Data Clean Summary

- Remove "Reuters" source from the article text (present in all true news)

[5]We deem fake news of subject *politics* and *Government News*, and real news of subject *politicsNews* to be politics relevant.

- Remove photo credits from the article text (prevalent in fake news)
- Remove "VIDEO" or related phrases in () in the title (prevalent in fake news)

3 Related Work

3.1 Prior Work Using the Same Dataset

Previously, Ahmed H, Traore I, and Saad S. utilized this dataset for the second part of their paper, Detecting Opinion Spams and Fake News Using Text Classification[1]. Similar to our initial finding (which will be explained further in the experiment section), the paper achieved a 98% accuracy initially using "a combination of news articles from different years with a broad variety of political topics." Therefore, the paper decided to only examine a subset of articles around the 2016 US election. The authors picked 2000 real articles and 2000 fake articles. Using TF-IDF, and TF with a combination of different feature size and n-grams, the paper represented results for SVM, LSVM (Linear Support Vector Machine), Logistic Regression, KNN, and Decision Trees. The best model is LSVM, using TF-IDF (as opposed to TF) and Bigram.

Due to time limitations, we will explore a subset of these models, but using Bert as a context-based embedding for comparison as opposed to TF (Term Frequency). At the time of this previous paper, which uses TF-IDF with n-gram, Bert has not been introduced yet.

3.2 Blogs and Tutorials on TF-IDF and Bert

One major challenge with our project is how to sensibly translate text information to numeric representation and use this information to predict the validity of a news article. This task falls under the widely discussed area of text classification. A variety of methodology and application have been proposed by previous research, and we found the following sources especially relevant to our project.

In a medium blog post regarding text classificaiton, the author detailed a complete workflow of text classification task from preprocessing the raw text to running machine learning model.[2] The author implemented two context-free feature extraction method, one based on term frequency, TF-IDF and the other using deep learning, word2vec. With these extracted features, the author proceeded to perform logistic regression and naive bayes for classification. The models were evaluated using accuracy score, and TF-IDF combined with logistic regression produced the best performance. Our takeaway from this article is that we also intend to perform feature extraction using term-frequency and pre-trained deep learning models. Building upon the two classification models mentioned so far, we are also planning to experiment with other classification method, such as random forest and SVM.

To use deep learning techniques to create word embeddings, we will reply on huggingface documentations and blog posts like `http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/`. This post provided a detailed tutorial and visual representation of the pre-trained DistilBERT model.

In sum, we intend to apply these aforementioned text embedding and deep learning methodologies to identify whether a news article is fake or real given its content.

3.3 Research and Academic Articles

We further explored some other research papers on the topic of Fake News detection to help us understand what are some commonly used models and methods in this domain.

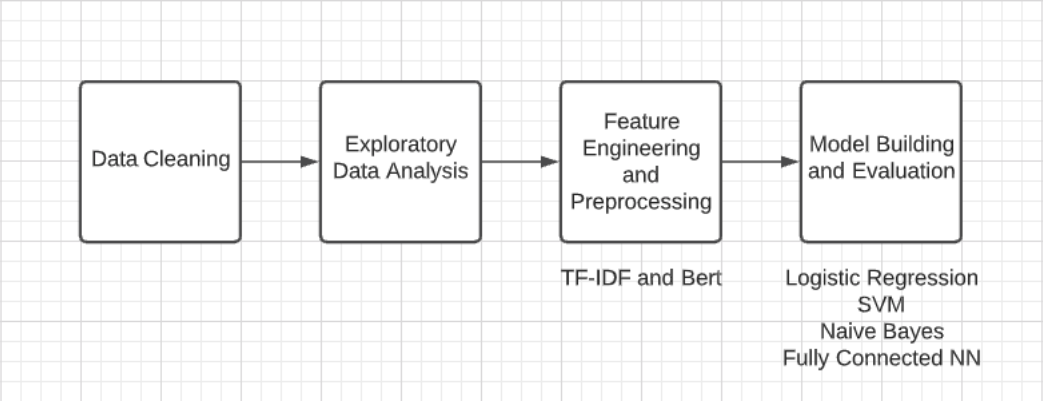
In paper Automatic Detection of Fake News [6], the authors used a linear SVM classifer with a 5-fold cross validation method. The paper used accuracy, precision, recall, and F-score as evaluation metrics. The paper grouped accuracy by feature types, finding that the best performing classifers are the ones "that rely on stylistic features", such as punctuation and readability. In addition, it is interesting that the paper tested

domain-specific news with different types of features. For technology, features that represent readability are the most important predictors.

Another paper The Language of Fake News: Opening the Black-Box of Deep Learning Based Detectors[3] used CNN to detect fake news. In addition, the model aimed to predict novel topics correctly. Although it is interesting, we will not do CNN or novel topics prediction centered training here due to resource and time constraints. We will use a fully-connected neural network model.

We also found a paper Text-mining-based Fake News Detection Using Ensemble Methods[4] which uses an Ensemble method. Since this is not a free resource, we only examined the abstract. The paper found that using a combination of features and word-vector representations, an accuracy of 95.4% is achieved using **ensemble** method.

4 Problem Formulation



Given human-annotated news data, we can solve our problems, fake news detection, as a supervised learning one. We would like to train a classifier that can accurately identify a news article as fake, given its title and/or text, as well as other properties associated with its publication, such as publication date information. In this specific project, we are only interested in classifying news either as real or fake, so our classification problem is further limited to a binary one. For this binary classification problem, the inputs are words in the news articles, titles and other potential features.

We starts the project by observing the data and performing some data cleaning as mentioned above. Using the cleaned data, we perform EDA. Next we conduct feature engineering.

The uniqueness of this classification problem lies in the nature of our input data - natural languages. We understand that for machines to understand these text information, we would need to find sensible ways to convert text into numeric representation in the feature engineering process. In designing this transformation mechanism, it's important to compare it to how human make the conscientious choice of marking a news article as illegitimate. For instance, we may choose to inspect the vocabularies used in certain text; we might also peruse the article and understand its meaning and sentiment. At a high level, we think that there are discernible differences between real and fake news characteristics; in other words, we believe that real news and fake news are similar to those in their own groups on certain aspects. This property should remain in the numeric representations - similar news title and articles should also be in close proximity to each other in the newly found vector spaces. We will experiment with both context-agnostic and context-sensitive text feature extraction methodologies. This correspond to our understanding of human decision making process. It should be noted that our main objective is not to develop new or improve upon current natural language processors, but rather employ existing NLP techniques for our ultimate goal of classification. More details will be mentioned in the Methods section.

Finally we perform several classification models and evaluate its performs, mostly using accuracy given

balanced data.

We'll discuss below the specific methods we use regarding (1) text feature extraction and (2) binary classification.

5 Methods

5.1 Feature Engineering

1. **TF-IDF vectorization with Principal Component Analysis**
   We first calculate TF-IDF for each data point/article.

   - Term Frequency ("TF"): TF of word "cat" for each document = number of "cat" count/ document word counts
   - Inverse Document Frequency ("IDF"): IDF of word "cat" = 1 + ln(Total Number Of Documents / Number Of Documents with term "cat" in it)
   - TF-IDF: TF * IDF

   Next, we performed PCA to reduce the dimension of the input. PCA will help us to retain the top directions (linear combinations of feautres) that can capture the most variances in the data.

2. **BERT/DistillBERT**
   Bert stands for Bidirectional Encoder Representation from Transformers. It is a power tool introduced in 2018, creating context-based word embeddings trained through bidirectional transformers. One could fine-tune the model parameters, or just take it as pre-trained. In addition, Bert can take punctuation, which could be important in the classification of fake vs true news.

   In this project, we use pre-trained DistilBERT[5], which is designed to be "smaller, faster, cheaper and lighter". This model effectively combined the step of PCA on Bert. This model works well with our dataset, given article length and sample size. The specific pre-tained model we used is "distilbert-base-uncased". This accepts punctuation but would require lower case letters. Finally, the max number of words + punctuations is 512, and the dimension of embedding is 768.

   We created distilBERT embeddings for two sets of features: 1) article titles, and 2) article text/bodies. There are two steps in getting the embeddings: 1) tokenize the cleaned text input and make them encoded inputs, and 2) obtain the embeddings from the model. Due to how long it takes to obtain the embeddings, we saved the outputs of the distilBert embedding in numpy formats for further use as inputs in models.

5.2 Modeling

The second step is to use various classification models to identify fake news. We intend to use the following models:

1. Baseline models: Logistic Regression, SVM
   We decided to use Logistic Regresion and SVM because they are classic models for classifications. In addition, the previous paper working on the same dataset has linear SVM as the best performing model.
2. Statistical Learning: Native Bayes, Random Forest [conditional on time]
3. Deep Learning: Fully-connected Neural Network

6 Experiments and Results

Our preliminary accuracy results shows that: TF-IDF achieves around 98%, Bert-title-LR achieves 96%, and Bert-text-LR achieves 99%.

Due to this observation, we decided to down scope the topic, like the previous paper. Specifically, we will keep news that are related to politics, but does not restrict it to a certain year.

7 Conclusion and Discussion

Acknowledgments

References

[1] Saad S Ahmed H, Traore I. Detecting opinion spams and fake news using text classification. *Journalof Security and Privacy*, 1(1), January/February 2018.

[2] Ishaan Arora. Document feature extraction and classification.

[3] N. OâBrien, Sophia Latessa, Georgios Evangelopoulos, and X. Boix. The language of fake news: Opening the black-box of deep learning based detectors. 2018.

[4] Raj N. Gala M. et al. Reddy, H. Text-mining-based fake news detection using ensemble methods. *Int. J. Autom. Comput.*, 2020.

[5] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[6] Alexandra Lefevre Rada Mihalcea Veronica Pérez-Rosas, Bennett Kleinberg. Automatic detection of fake news. *COLING*, 2018.