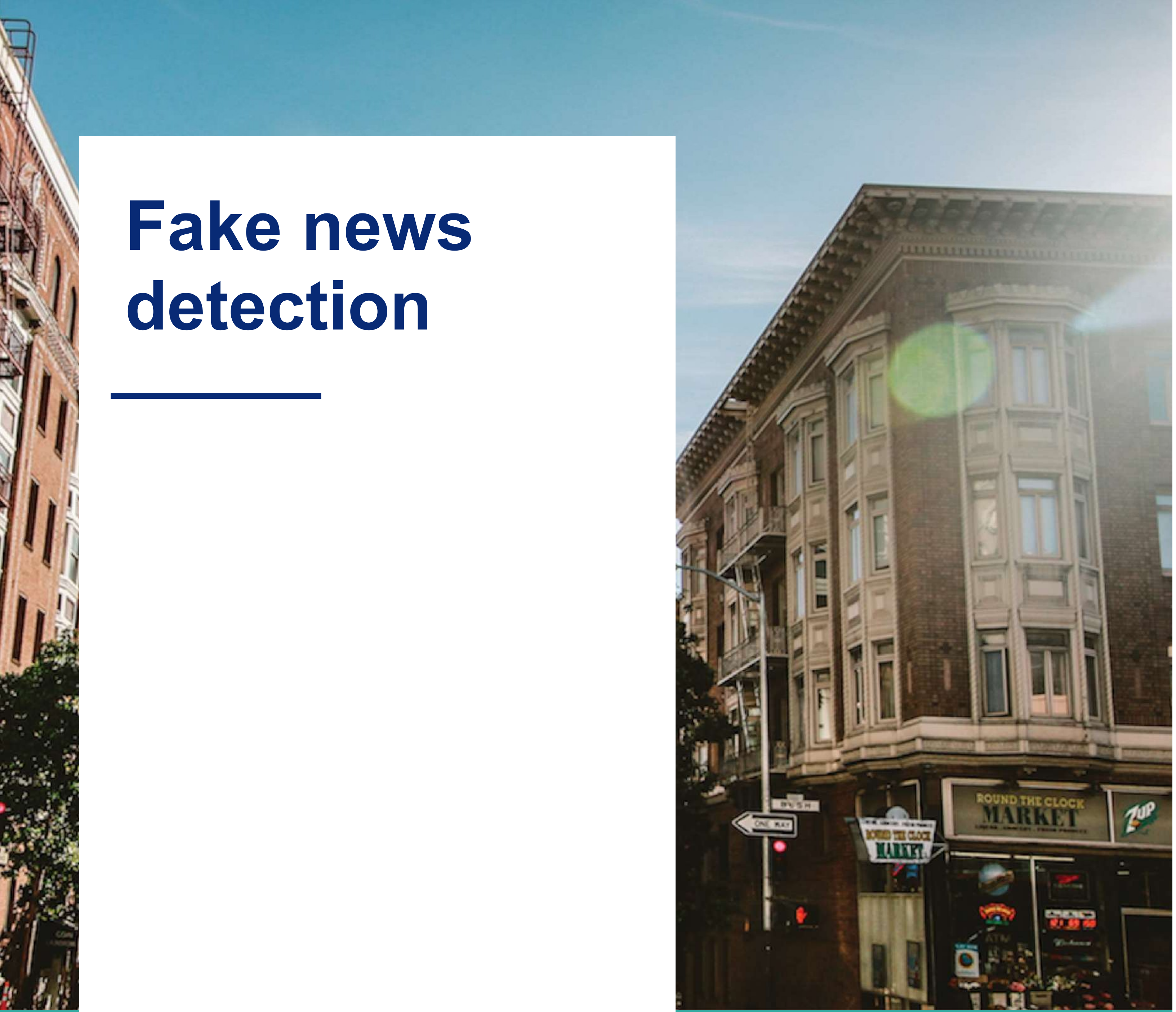


Fake news detection

Group Assignment

Arnab Biswas
Maddi Pranav Reddy
Mohan Nishantam

Introduction

The bane of fake news

Digital media today has extended its reach over print media because of low cost of production and distribution and interactivity. According to americanpressinstitute.org around 42% of adults aged 18 – 34 pay for digital media subscriptions compared to 14% adults aged 65+. While conversely, 72% of adults aged 65+ prefer print media subscriptions over 42% adults aged 18 – 34. With such an impressionable age bracket digital media has a disproportionate effect on general worldview. The paradox lies in the fact that it is not easy to distinguish between true news and fake news.

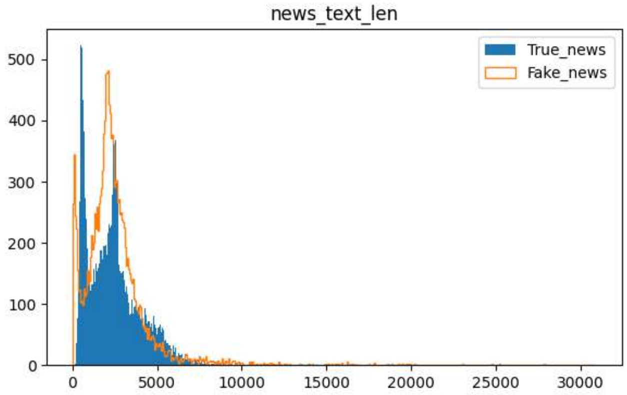
Problem statement

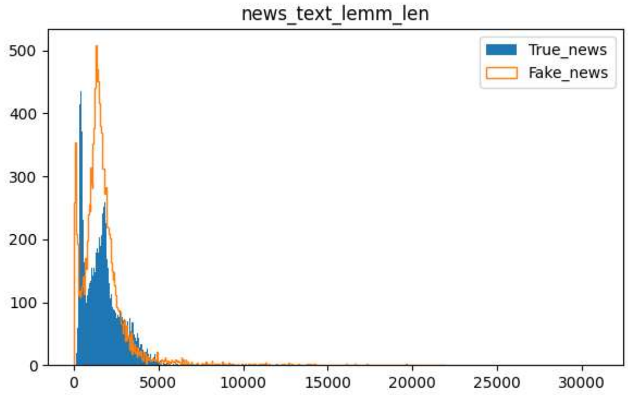
To classify news text as true or fake using supervised machine learning techniques based on a corpus of labelled news texts using ‘word2vec-google-news-300’ pre-trained word2vec model.

Exploratory data analysis

Text length

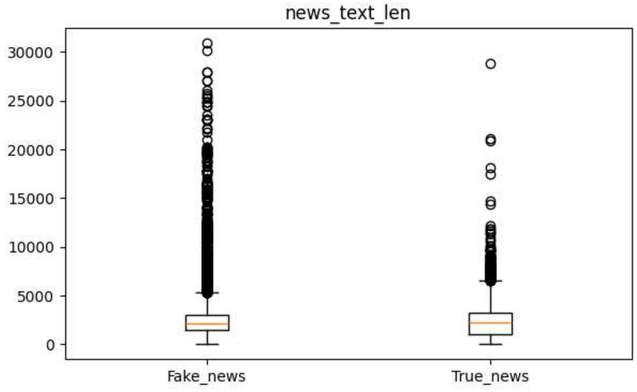
A comparison of distribution is shown below:

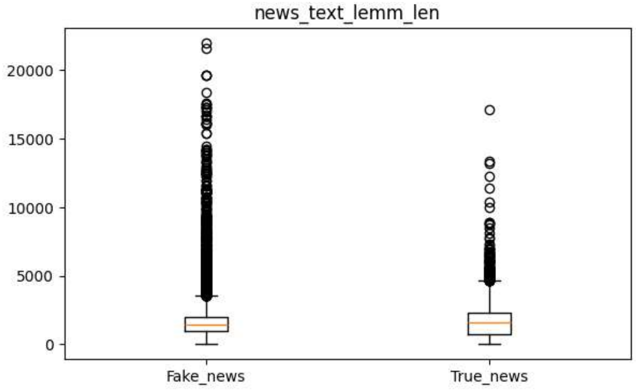




- Both fake and true news show a double peak and have a more or less similar distribution.
- But the magnitude of the peaks is reversed in both cases which means chances of fake news with long text is more than otherwise and vice versa for true news.
- Thus, length of text can be used as an input feature for news classification

A comparison of spread is shown below:

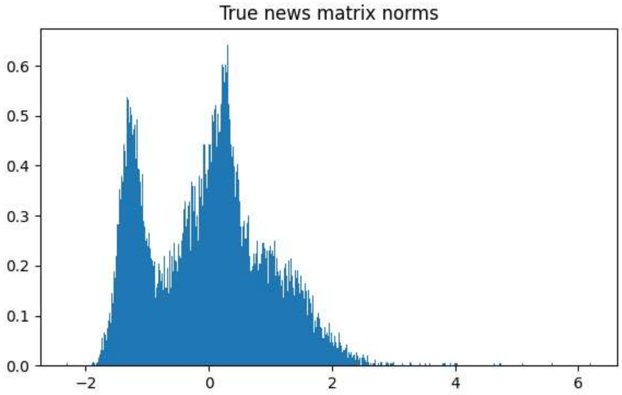


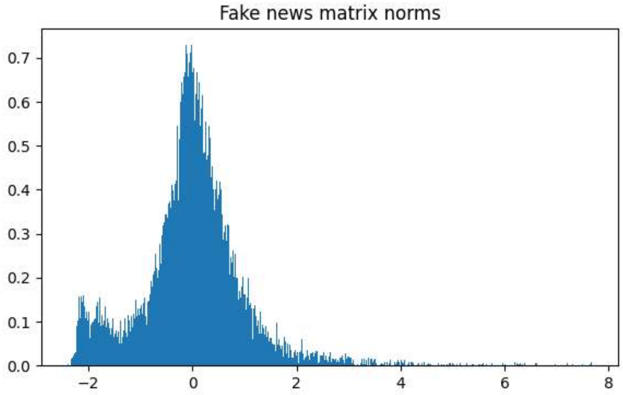


Text length of fake news is more spread out with more outliers than true news highlighting difference in editing standards.

Text hyperspace volume

A comparison of text matrix norm is shown below:

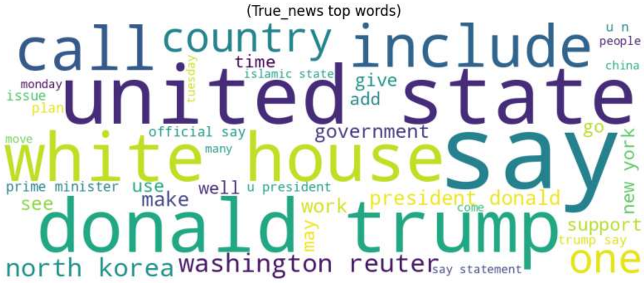

True news matrix norms

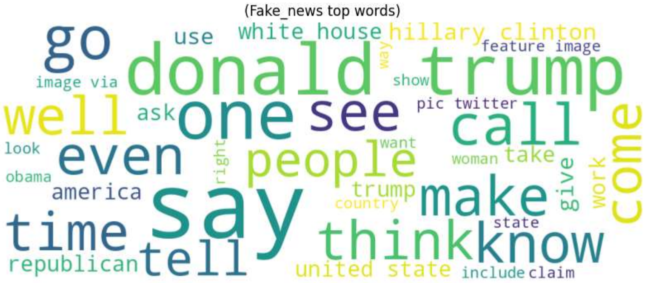

Fake news matrix norms

- Text matrix is composed of all the vectors of the contained words.
- Frobenius norm is used as the measure which is an indicator of the depth of the meanings of the words the text is comprised of.
- True news shows double peak which means more depth and variety in content
- Fake news shows a single peak which means less depth and variety in content
- This can be used as a feature for news classification

Semantic analysis

A comparison of top 40 most used words is presented below:


(True_news top words)

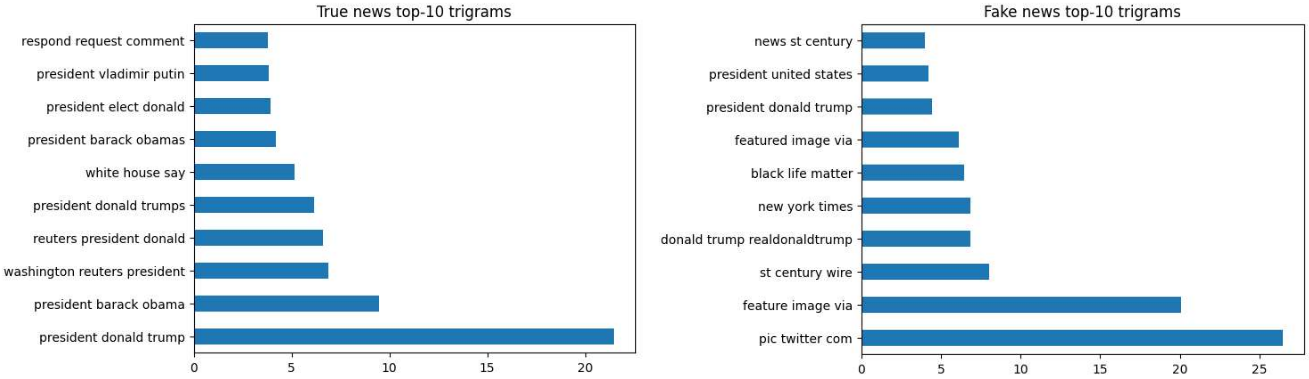

(Fake_news top words)

- Fake news texts have more references to people like ‘Donald Trump’, ‘Hillary Clinton’, ‘Obama’. True news has more references to geopolitical entities like ‘White-house’, ‘Washington’, ‘Islamic State’, ‘China’, ‘North Korea’.
- True news has more references to official sources like Reuters while fake news has more references to social media like Twitter.
- Words like ‘claim’, ‘include’, ‘want’, ‘even’ indicate a tendency of over generalization in fake news. While words like ‘plan’, ‘statement’, ‘issue’ ‘official’

indicate definitiveness in true news. True news also uses words like ‘may’ in order to not overgeneralize facts or claims.

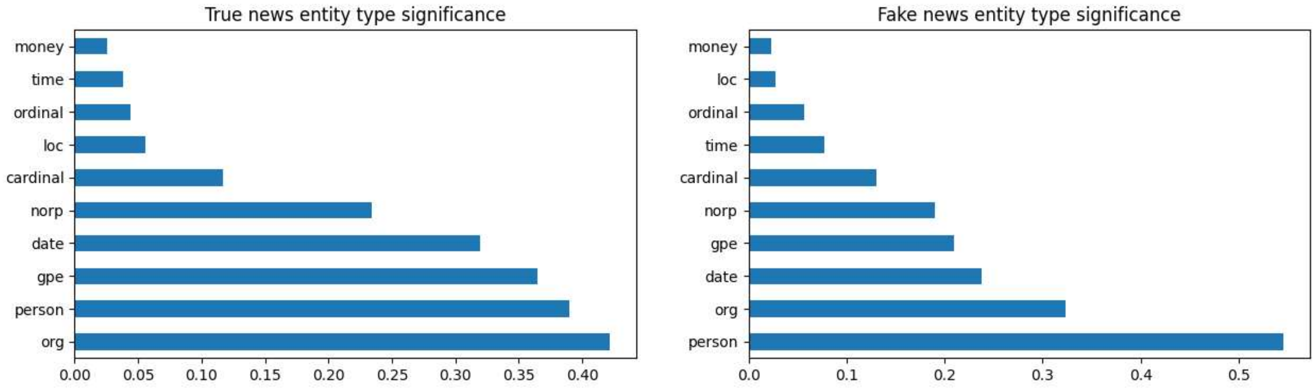
- Words like ‘see’ and ‘look’ in fake news indicates more use of images or image sources as proofs or justifications. Word ‘see’ is not used much in true news.
- Fake news contains many action words like ‘go’, ‘tell’, ‘come’, ‘think’, ‘know’ indicate effort on the creator to engage the reader into believing something and propagating the same. True news does not contain such references.
- Fake news also references gender issues more than true news.
- Thus, word vectors can be used as input features for classifying news.

A comparison of top 10 trigrams is presented below:



- Top trigrams indicate more reference to persons in true news than fake news.
- Fake news trigrams indicate references to images, social media and news sources like NY Times.
- True news also uses political designations of persons referenced more than fake news.
- Fake news also contains more references to social issues more than political news.

A comparison of Tf-Idf entity type significance is shown below:



- Tf-Idf gives importance to words depending on the usage and rarity. Above comparison shows differences of context between true and fake news.
- True news texts focus on organizations, persons, geo-political entities, dates and groups proportionately.
- Fake news focuses more on persons disproportionately.
- This characteristic can be used as an input feature.

Approach to model training

- Cleaning text by removing punctuations, number, special characters and converting to lower case.
- Lemmatizing text by keeping only the NN, NNS tags because nouns carry the most of the semantic sense of a text (many verbs are generated from nouns as well like, I googled the meaning of the term).
- Creating word vectors by using word2vec-google-news-300 pre-trained model. It is a 300-dimension vector-model of 3 million words.
- Extracting sentence vector by taking the unit of the sum of all its word-vectors. Vector sum is the resultant of all the components of the vectors and averaging it will scale the resultant down to one word. This encapsulates all the semantic sense of the all the words in text into a single word.
- Extracting the Frobenius norm of the sentence matrix composed of vectors of all its words and standardizing across all documents. This signifies the distance of the sentence matrix from origin.

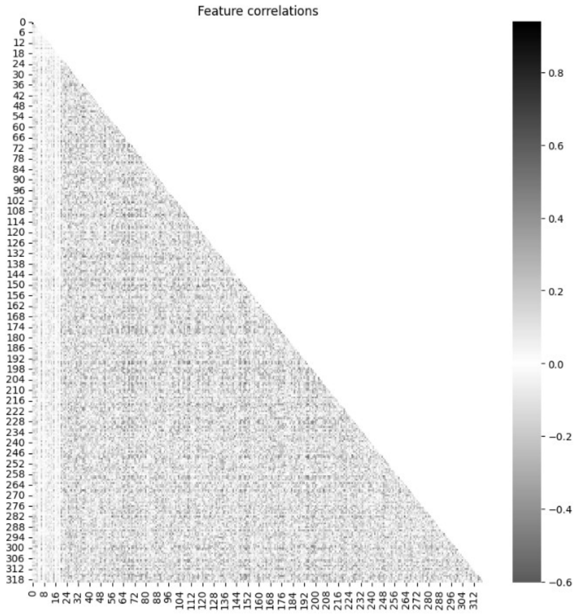
- Extracting the lemmatized text length and standardizing across all documents. Since length is a determining characteristic between true and fake news as per EDA.
- Extracting the most common Named_Entity_Type from text and standardizing across all documents. Most common entity carries the main semantic theme of a text.
- Hyper-parameters of the models are tuned using GridSearchCV

Evaluation metric

- F1 score is chosen as the best metric because it balances both precision and recall.
- Since both target classes are more or less balanced, both precision and recall can be prioritised instead of only recall.
- Mis-identifying true news as fake and fake news as true can be equally damaging to the reader.

Best model

- Logistic Regression (regularization: l1, solver: liblinear).
- **Accuracy: 96%, Precision: 95%, Recall: 96%, F1 score: 96%**
- Low variance model and in this instance, it has the highest scores amongst all three models of DecisionTree, RandomForest.
- Simple model can suffice for binary classification.
- Design matrix is not a sparse matrix.
- Very low correlations between feature variables.
- Takes less time to train.


Feature correlations