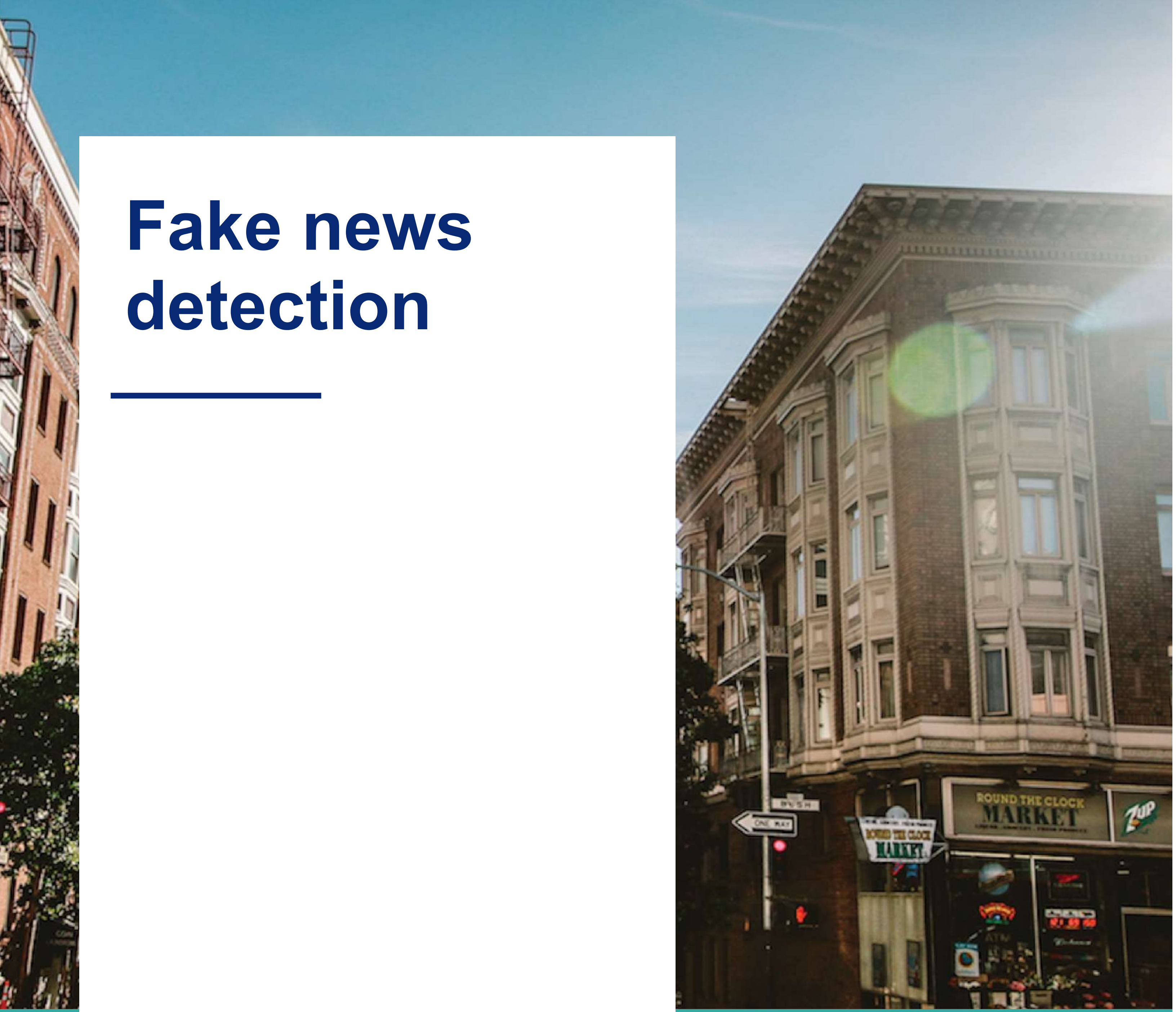
Fake news detection



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

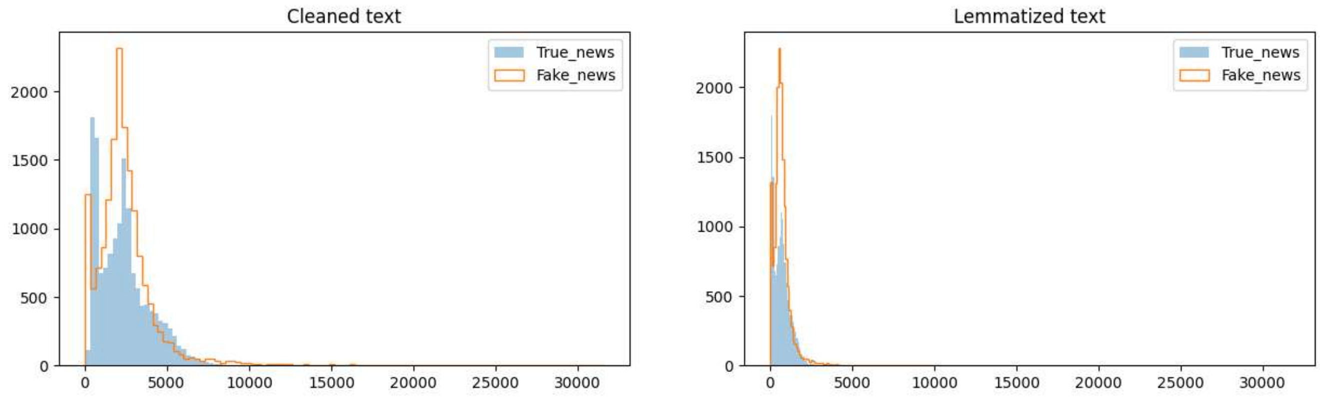
To classify news text as true or fake using supervised machine learning techniques based on a corpus of labelled news texts using ‘word2vec-google-news-300’ pre-trained word2vec word-vector model.

Assumptions

- Sentence vector is assumed to be sum of all its word vectors. Vector sum summarizes each component of its constituent vectors. In sematic perspective, it is a singular representation of all the meaningful dimensions of its words.
- Sentence matrix is composed of all its word vectors. Frobenius norm of this matrix represents the volume of the matrix hyperspace. Larger the magnitude of its word vectors larger the volume. Semantically, larger volume represents more varied usage and conceptual significance. True news may have a more conceptually significant context.
- Word similarity is the dot product of the resultant along each dimension with the vector sum of all the words, taken across each dimension.
  (resultant = word vectors (n_words x 300) sum axis = 1)
  (dot product = resultant.T , word vectors (n_words x 300))
  Semantically it represents how close each word is to the central meaning of the text. Thus, a high value of dot product indicates a more directed and purposeful context. Fake news may have a more purposeful context.
- Text lengths convey editing standards which is more significant in print media. Attempt has been made here to extend the same to digital media. High editing standards consider factors of readability like attention span, text volume first impression, distinction of key words etc. These influence text length. Fake news may lack such standards.
- Nouns in texts are typically used as subjects and objects of sentences and convey the semantic meaning of the context e.g. Elephant jungle may convey the context of elephant lives in a jungle. Thus, named entities attached to nouns also convey the theme of a text. True and fake news may display varying patterns and counts of named entities contained in them.

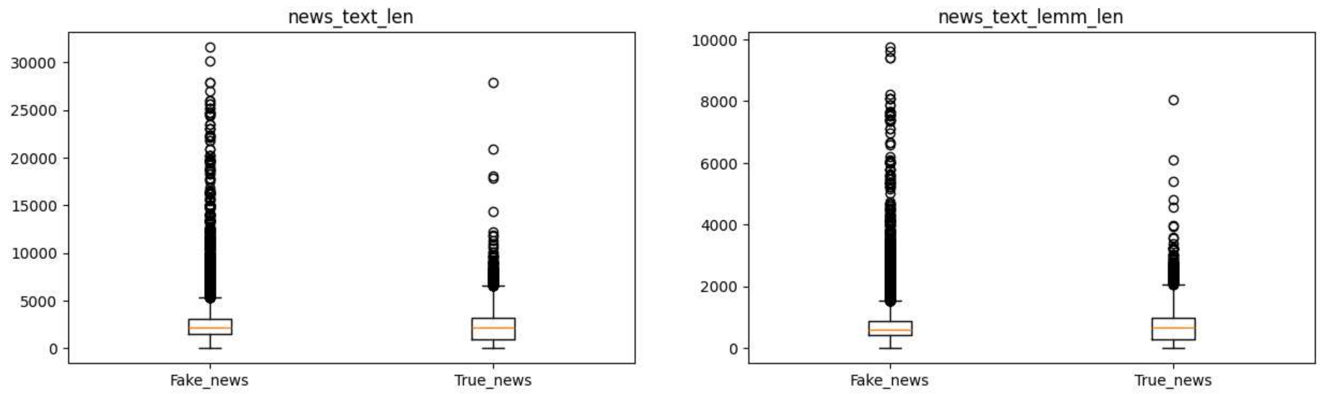
Exploratory data analysis

Text length distribution



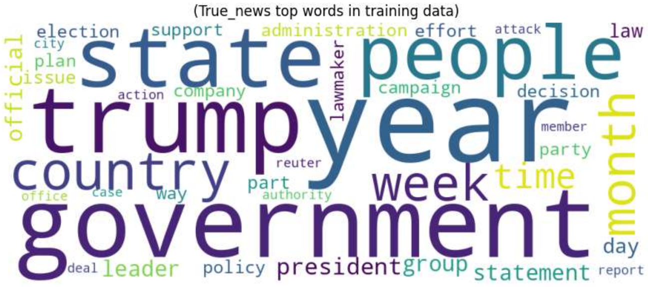
- Both fake and true news show a double peak and have a more or less similar distribution.
- But the magnitude of the peaks is reversed in both cases which means chances of fake news with long text is more than otherwise and vice versa for true news.
- Thus, length of text can be used as an input feature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

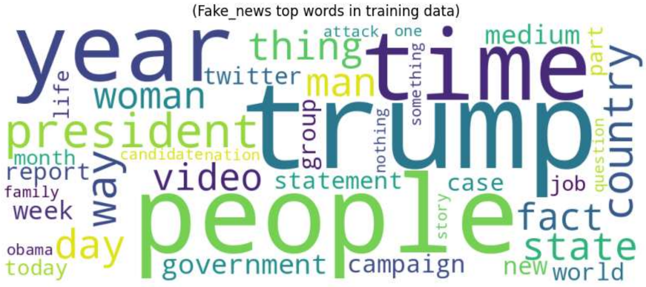
Text length spread



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

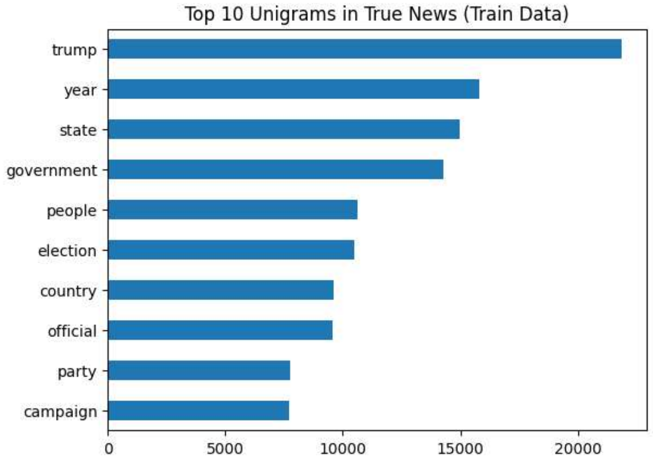
Word Clouds

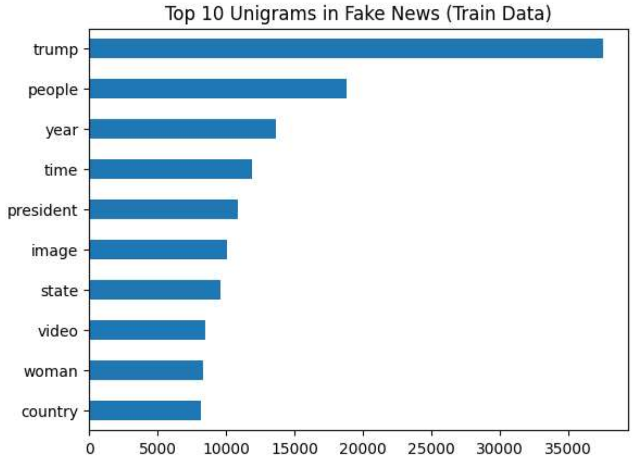


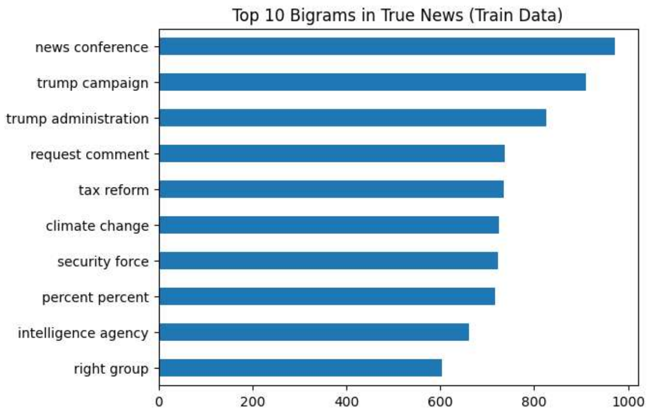


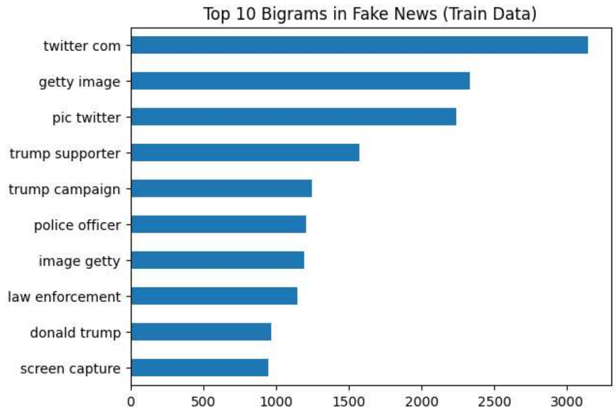
- Fake news texts have more references to people like ‘Donald Trump’, ‘Obama’. True news has more references to geopolitical entities like ‘government’, ‘state’, ‘country’ ‘administration’.
- True news has more references to official sources like Reuters while fake news has more references to social media like Twitter.
- Fake news references social issues (‘man’, ‘woman’, ‘family’). True news focuses on official issues (‘state’, ‘country’, ‘administration’)
- Fake news conveys a lack of definitiveness (‘story’, ‘question’, ‘nothing’, ‘thing’, ‘something’) while True news conveys definitiveness (‘policy’, ‘statement’, ‘rule’).
- True news focuses on results, like use of word ‘election’ while fake news focuses on targets like ‘campaign’.
- Use of word ‘fact’ in fake news as in ‘it is a fact that…’ convey a justification to a claim often used without evidence.
- Fake news uses visual context as proof.

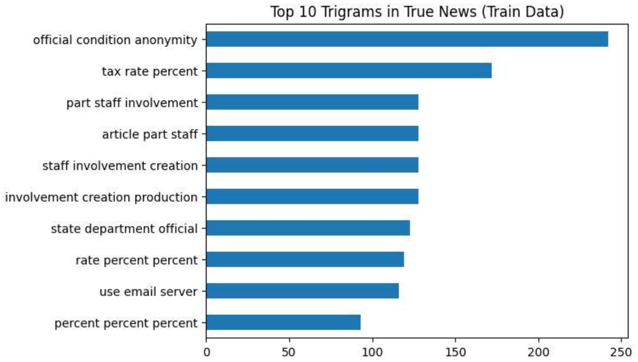
N-grams

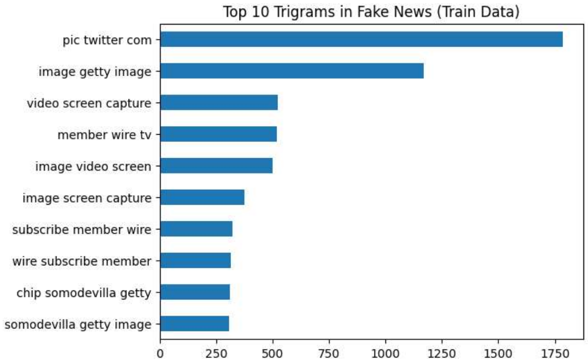






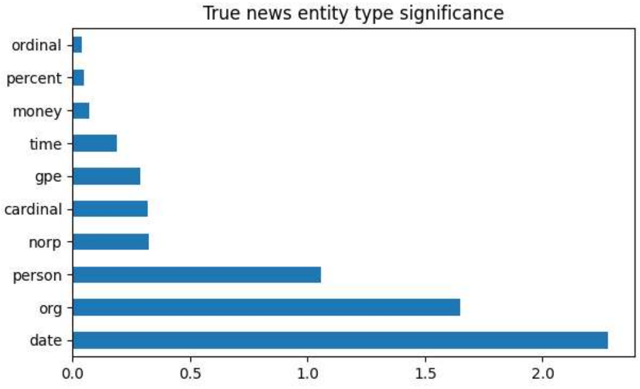


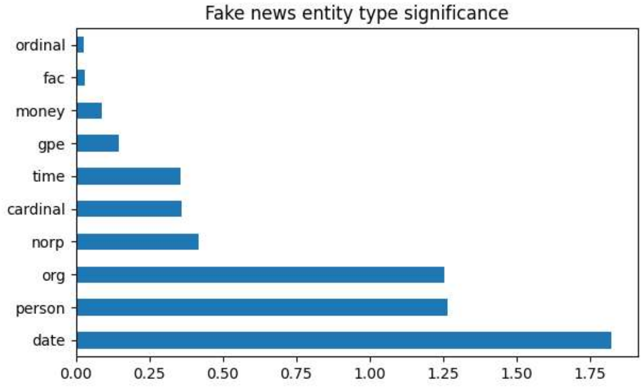




- Fake news contain references to images, videos and screen captures as proofs.
- True news admits unverifiability in sources unlike fake news ('official condition anonymity').
- Subject matter of true news is focused on real issues ('tax rate', 'climate change' 'court appeal', 'council resolution').

Entity types



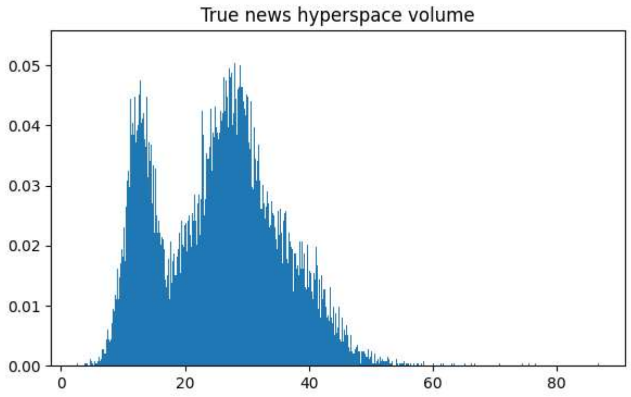


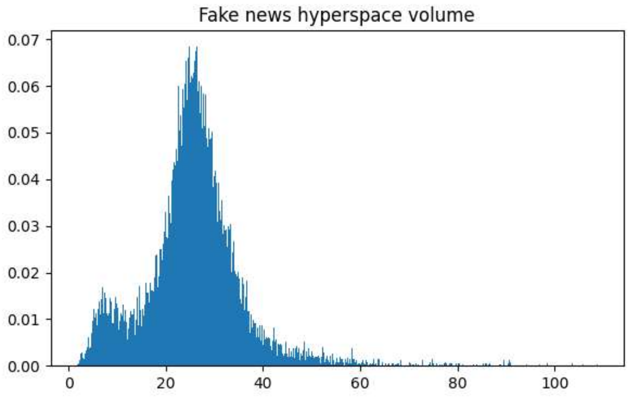
- True news texts focus on organizations, persons, geo-political entities, dates and groups proportionately.

- Fake news focuses more on persons disproportionately.
- This characteristic can be used as an input feature.

Text hyperspace volume

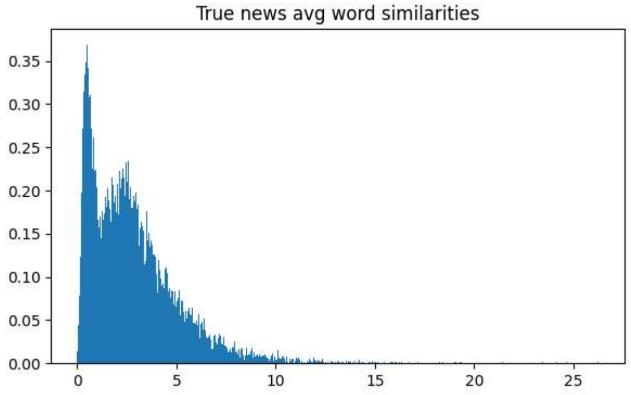
A comparison of text matrix norm is shown below:

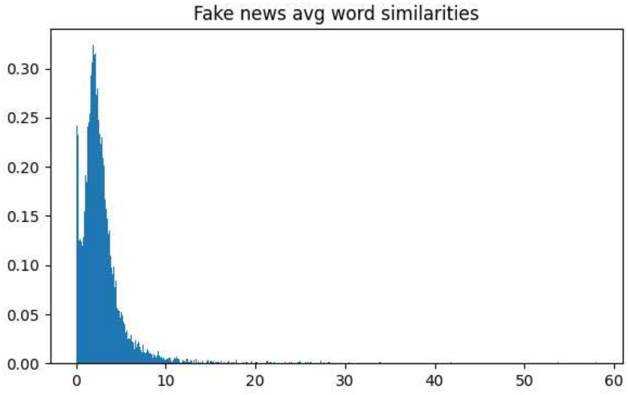

True news hyperspace volume


Fake news hyperspace volume

- 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

Average word similarities


True news avg word similarities


Fake news avg word similarities

- Peak average word similarity for fake news occurs at a higher value than true news which signifies a more purposeful and directed context.

Approach to model training

- Cleaning text by removing punctuations, numbers, special characters, quotation marks, texts in square brackets, NLTK stop-words 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 word vector-model of 3 million words.
- Extracting sentence vector by taking the sum of all its word-vectors.
- Extracting the Frobenius norm of the sentence matrix composed of vectors of all its words and standardizing across all documents.
- Extracting word similarities (dot product of each word vector and sum of all word vectors)
- Extracting the lemmatized text length.
- Extracting the counts of Named_Entity_Type from text.
- Hyper-parameters of the models are tuned using GridSearchCV

Impact

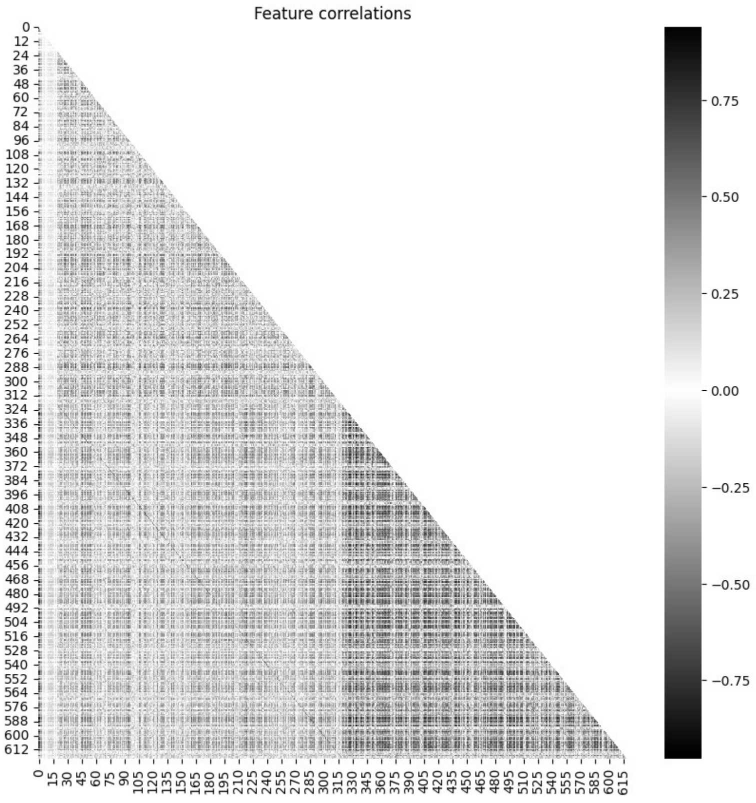
- Using standardized values (like standardized text lengths) causes faster convergence but reduces prediction power.
- Adding more features like word similarities, NER counts improve prediction power but increases training time.
- Adding more derived features increase multi-collinearity which adversely impacts logistic regression

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 prioritized instead of only recall.
- Mis-identifying true news as fake and fake news as true can be equally damaging to the reader.

Best model: Random Forest

- (n_estimators= 200, min_samples_leaf= 1, min_samples_split= 5).
- **Accuracy: 91.76%, Precision: 91.59%, Recall: 91.08%, F1 score: 91.34%**
- It has better performance than Decision Tree.
- It has marginally better performance than Logistic Regression which has a slightly better recall (92%) but lower accuracy (91.63%), lower precision (90.59%) and lower F1 score (91.29%).
- Higher precision means more confidence in the model.



- Unaffected by correlations between independent variables.
- Ensemble model and hence generalizes better to new input data.
- Flexible and adaptable to changing input data.
- But it is a high maintenance model which takes longer time to train and has more hyper-parameters to tune.