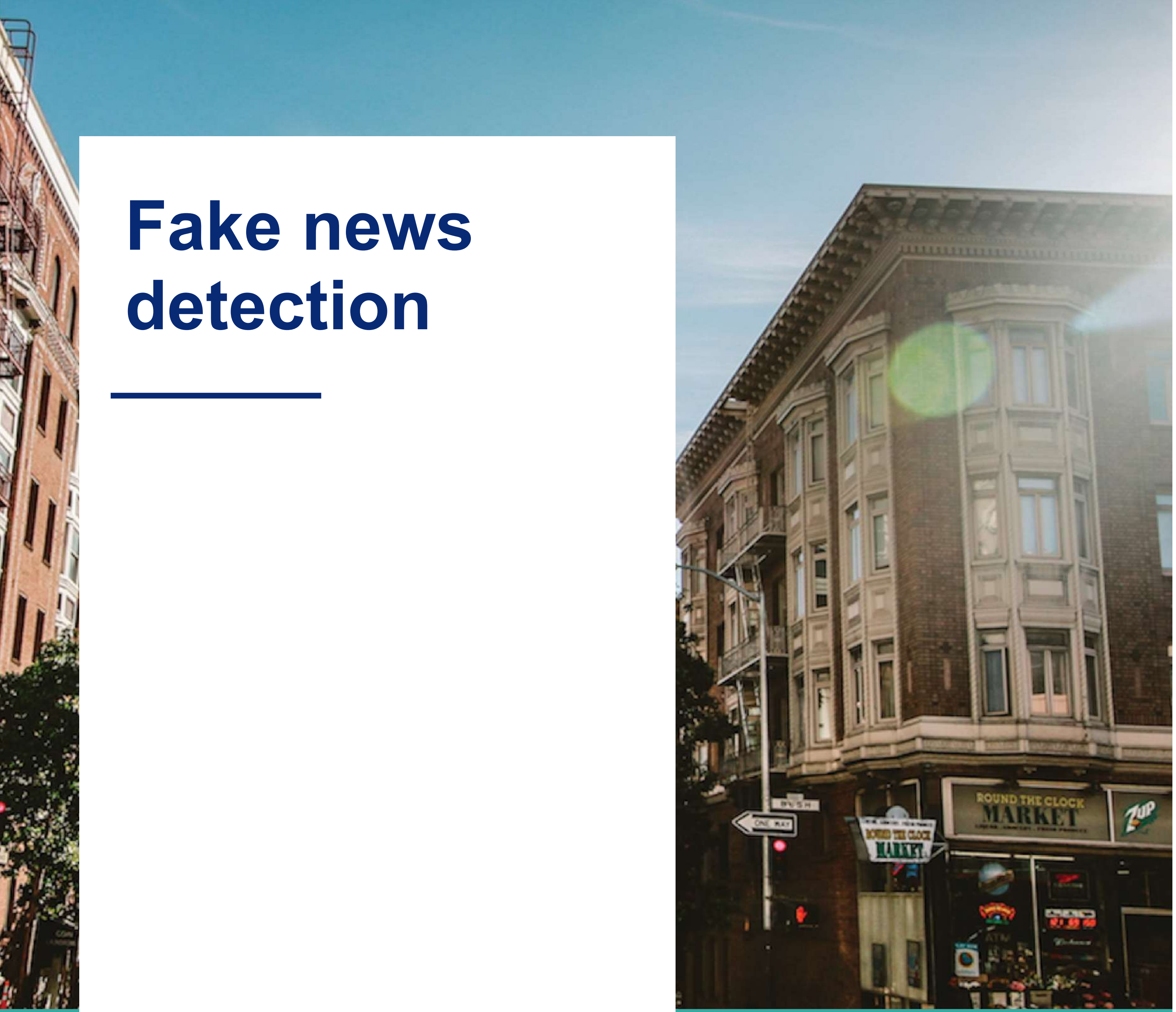
Fake news detection



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.

Assumptions

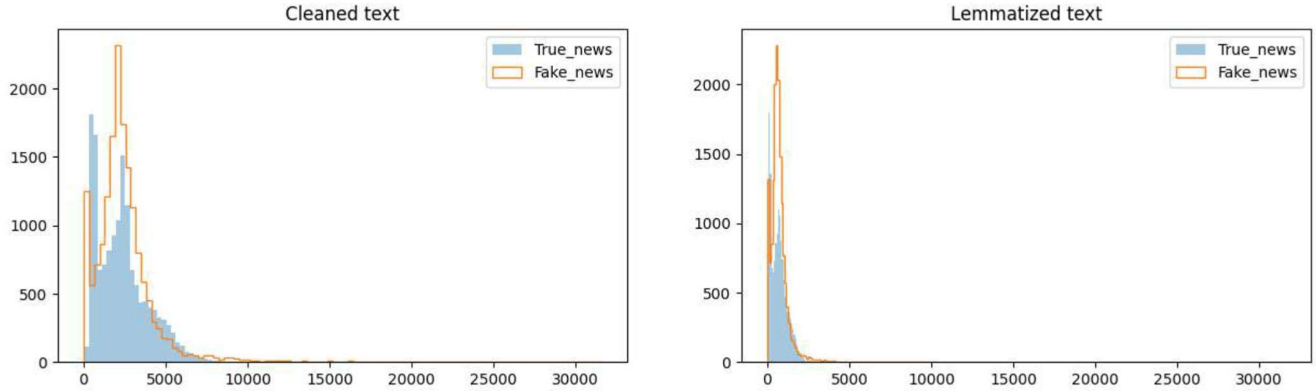
- **Text lengths (scalar)** convey editing standards which is more significant in print media.
  Attempt has been made here to extend the same to digital media. High editing standards consider factors of readability like attention span, text volume first impression, distinction of key words etc. These influence text length. Fake news may lack such standards.
- **Sentence vector (1x300)** is assumed to be sum of all its word vectors. Vector sum summarizes each component of its constituent vectors. In sematic perspective, it is the average representation of all the meaningful dimensions of its words.
- **Sentence matrix (scalar)** 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 (1x300)** 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.
- **Word distance (1x300)** is the length of each word vector from the resultant of all word vectors in the text. A higher distance means more varied word usage. It is assumed that true news with diverse content and high editing standards will have higher word distance than fake news.
- **Named entity (1x18)** are attached to nouns which are 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.

- **Most distant word pairs (1x300)** are word pairs that are semantically most distinct amongst words used. Euclidean distance is used as a measure of distance. Looking at such pairs in a text reveals a pattern that a few words are repeated in such pairs. Such words are not necessarily the most frequently used in a text but can be considered the most important and carry the main semantic theme of the text which can be used as a distinguishing feature.

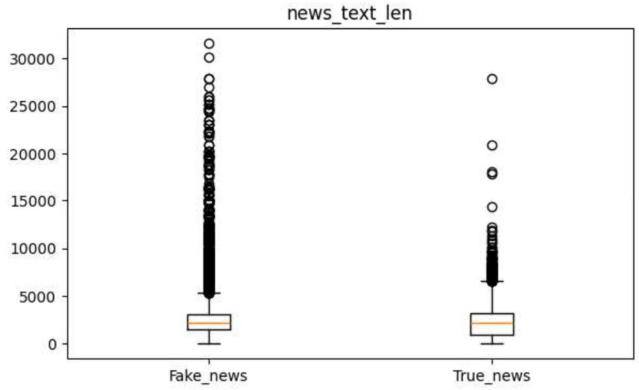
Exploratory data analysis

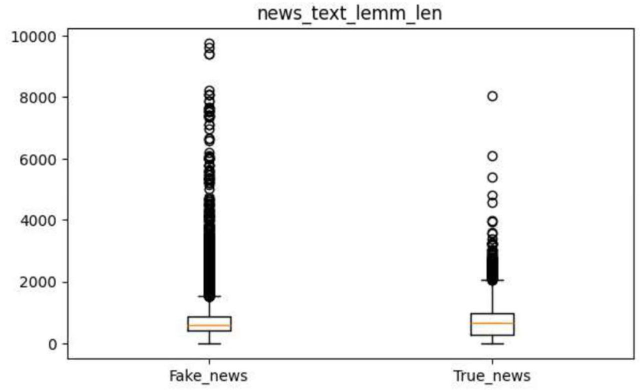
Text length distribution



- 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

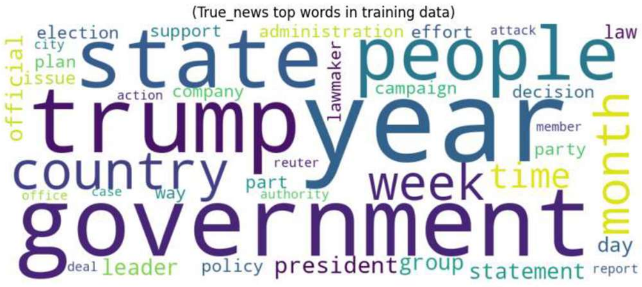
Text length spread

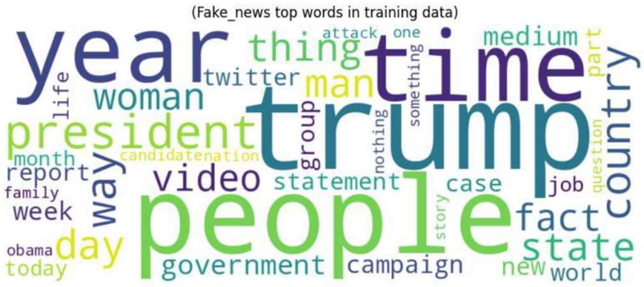




- Text length of fake news is more spread out with more outliers than true news
- highlighting difference in editing standards.

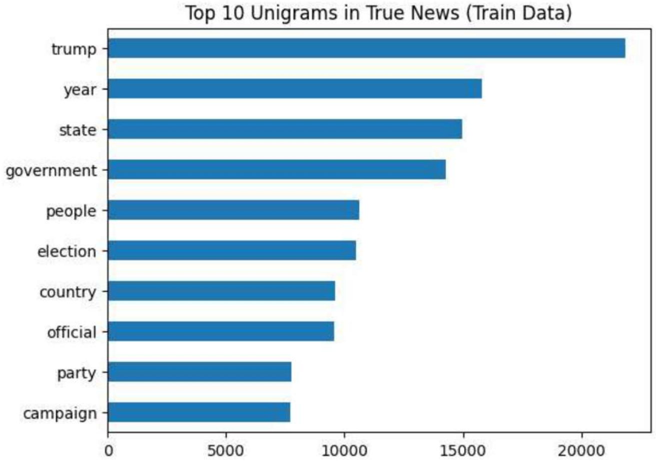
Word Clouds

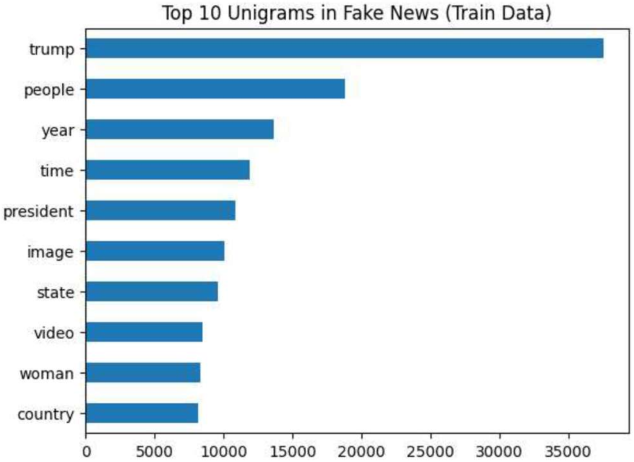


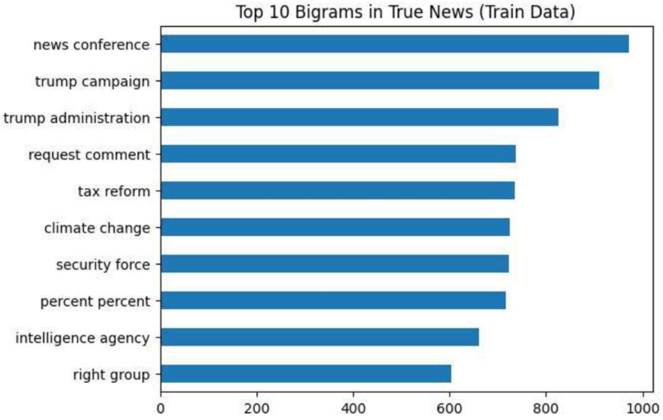


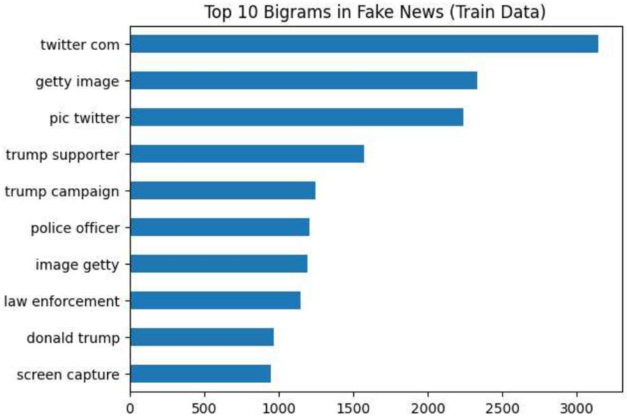
- Fake news texts have more references to people like ‘Donald Trump’, ‘Obama’. True news has more references to geopolitical entities like ‘government’, ‘state’, ‘country’ ‘administration’.
- True news has more references to official sources like Reuters while fake news has more references to social 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.

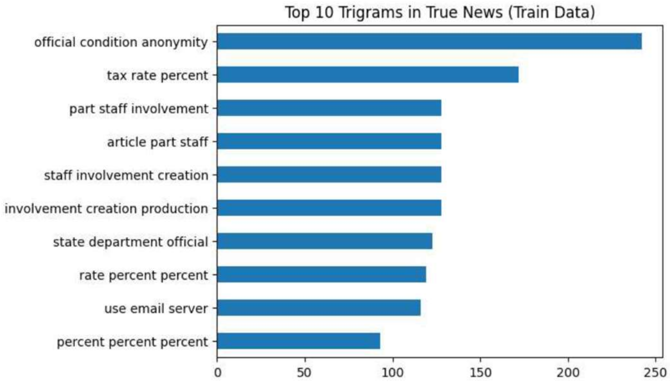
N-grams

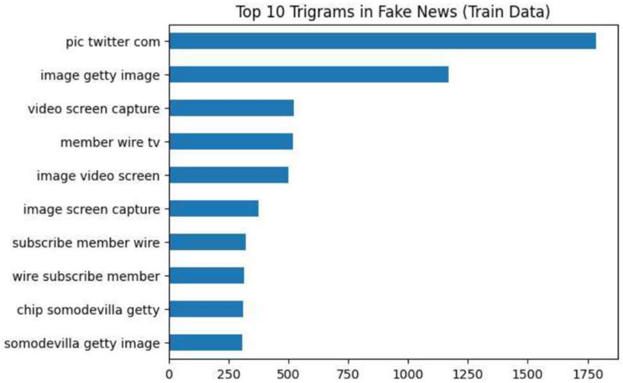

Top 10 Unigrams in True News (Train Data)


Top 10 Unigrams in Fake News (Train Data)


Top 10 Bigrams in True News (Train Data)

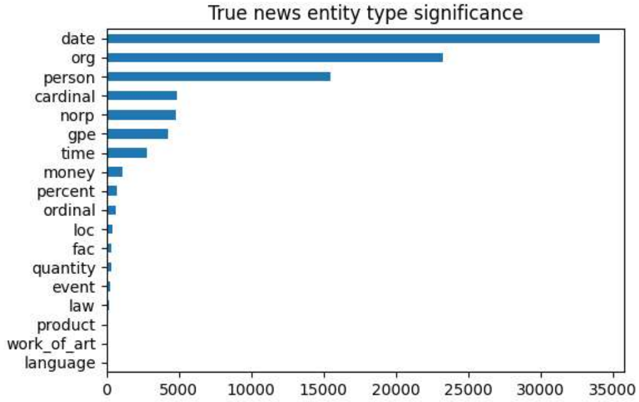

Top 10 Bigrams in Fake News (Train Data)

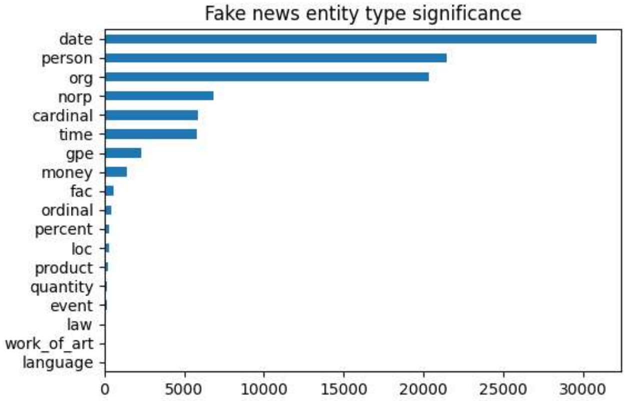

Top 10 Trigrams in True News (Train Data)


Top 10 Trigrams in Fake News (Train Data)

- 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’).

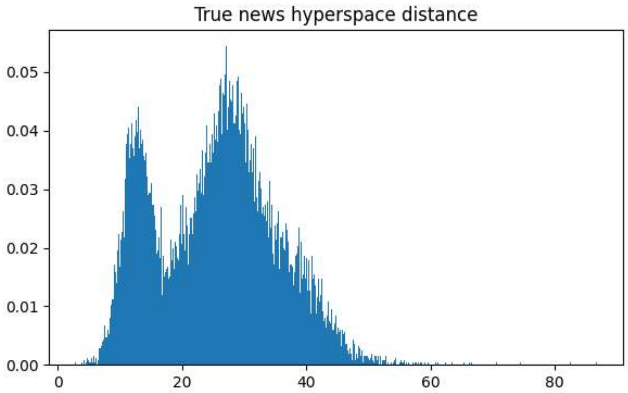
Entity types

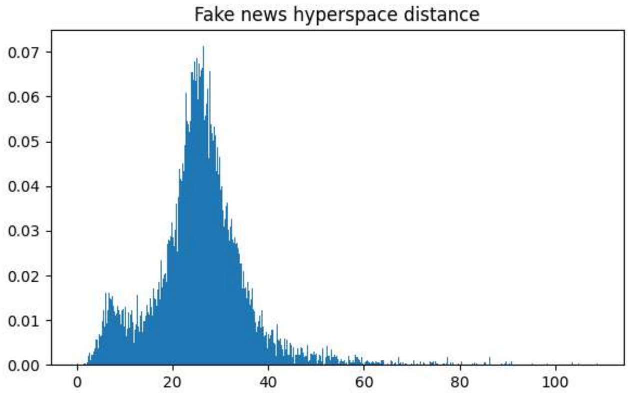




- True news texts focus on organizations, persons, geo-political entities, dates and groups proportionately.
- Fake news focuses more on persons disproportionately.
- The counts of entity types can be used as an input feature.

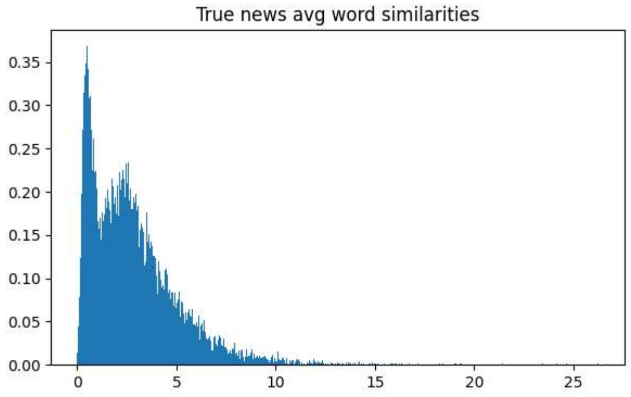
Text hyperspace distance

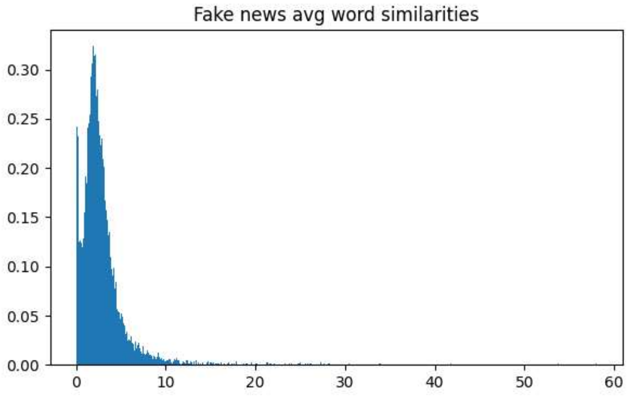




- Text hyperspace distance is the matrix norm of all word vectors in the text
- True news shows double peak which means more depth and variety in content
- Fake news shows a single peak which means less depth and variety in content
- This can be used as a feature for news classification

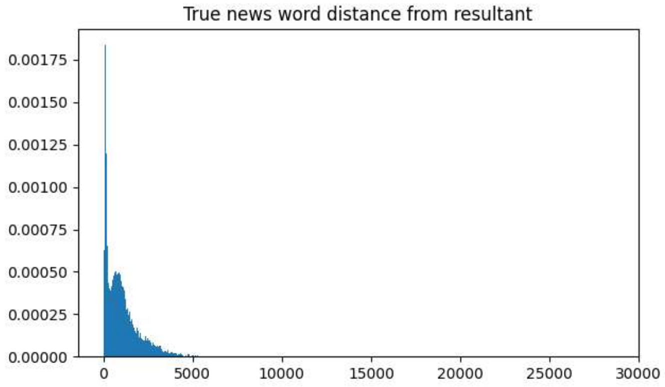
Average word similarities

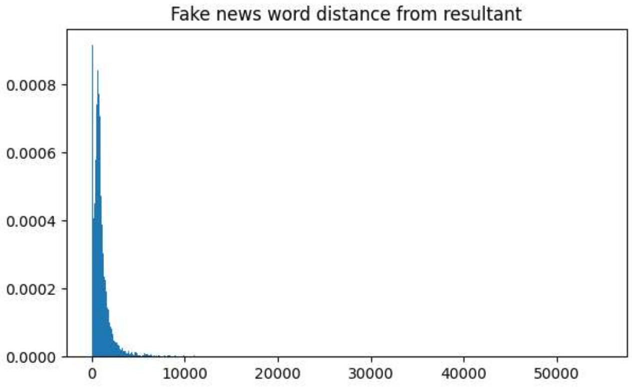




- Average word similarities is the mean of the dot product of each word vector with its resultant taken across each dimension.
- Peak average word similarity for fake news occurs at a higher value than true news which signifies a more purposeful and directed context.

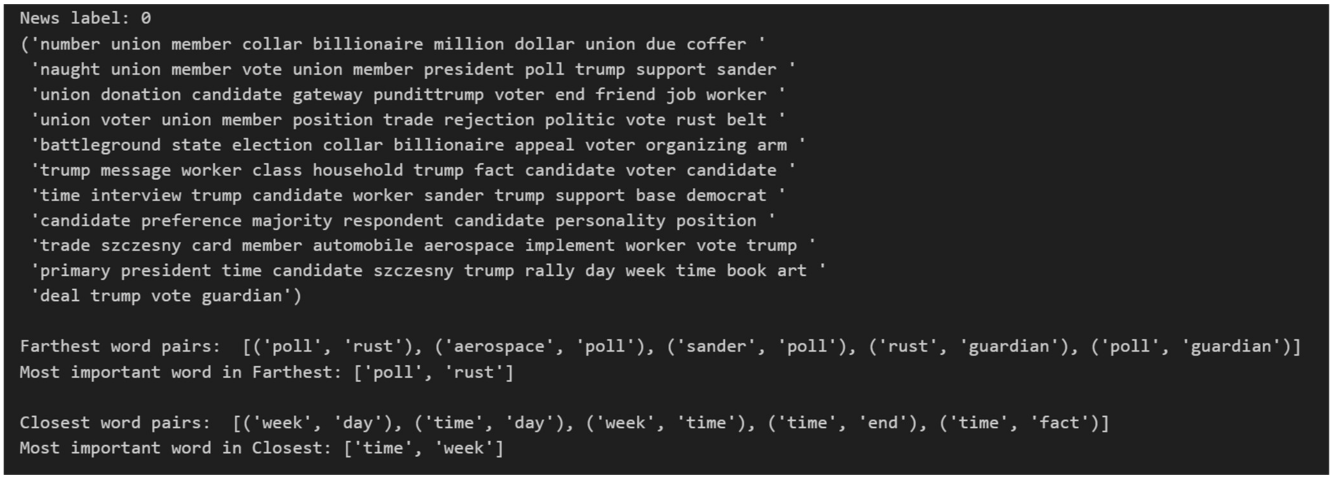
Average word distance from resultant





- Average word distance is the mean of the length of each word vector from the resultant of all word vectors in the text.
- Length of word distance vectors is more widely dispersed for true news than fake news which indicates variety of words is more for true news than fake news.

Most important words

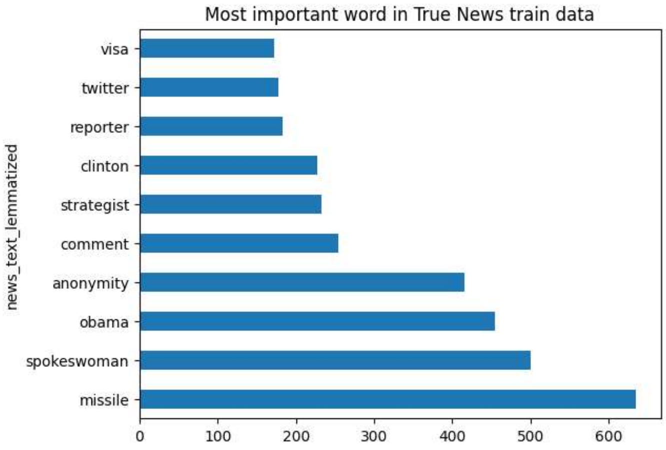


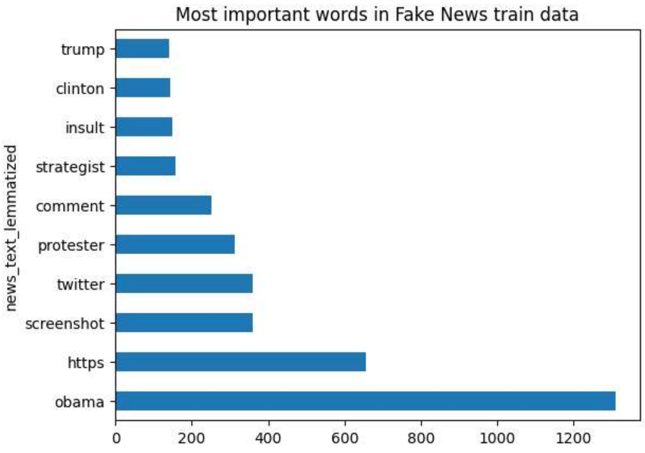
```
News label: 0
('number union member collar billionaire million dollar union due coffer '
 'naught union member vote union member president poll trump support sander '
 'union donation candidate gateway pundittrump voter end friend job worker '
 'union voter union member position trade rejection politic vote rust belt '
 'battleground state election collar billionaire appeal voter organizing arm '
 'trump message worker class household trump fact candidate voter candidate '
 'time interview trump candidate worker sander trump support base democrat '
 'candidate preference majority respondent candidate personality position '
 'trade szczesny card member automobile aerospace implement worker vote trump '
 'primary president time candidate szczesny trump rally day week time book art '
 'deal trump vote guardian')

Farthest word pairs:  [('poll', 'rust'), ('aerospace', 'poll'), ('sander', 'poll'), ('rust', 'guardian'), ('poll', 'guardian')]
Most important word in Farthest: ['poll', 'rust']

Closest word pairs:  [('week', 'day'), ('time', 'day'), ('week', 'time'), ('time', 'end'), ('time', 'fact')]
Most important word in Closest: ['time', 'week']
```

- In the above example, words like ‘vote’, ‘voter’, ‘union’, ‘member’ appear multiple times but are not the most important according to linear separation
- Both closest word pairs and farthest word pairs are extracted, but only the farthest word pairs convey the theme of the text.
- Thus, the words ‘poll’ and ‘rust’ capture the central theme of the text and can be considered as the most important words bearing the semantic theme.





- A summary of important theme words based on news classes are shown above.

Approach to model training

Pre-processing

- 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).

Feature extraction

- 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 number of words from text *(scalar feature)*
- Extracting counts of noun entity types from text *(count vector)*
- Extracting sentence vector as resultant of all its word-vectors *(math vector)*
- Extracting length of sentence vector *(scalar)*
- Extracting the matrix norm of the word-vector set of all words *(scalar)*
- Extracting dot product of each word vector with the sentence vector *(math vector)*
- Extracting length of vector difference between each word vector and sentence vector *(math vector)*
- Extracting word vector of top unigram of each news text *(frequency vector)*
- Extracting mean of word vectors of each word in top trigram of each news text *(frequency vector)*
- Extracting word vectors of the most important word of each text *(frequency vector)*

Model training

- Logistic Regression treated as the base simplest model with least hyper-parameters for tuning. F1 score is used for all models for best model tuning.
- Only scalar features (word counts, word repetitions, text matrix norm, and length of resultant vectors are passed as input features.
- Base model performed little better than a random guesser with an accuracy of 58%.
- Adding count vector of entity types increased base model accuracy to 62%.
- Adding resultant of word vectors for texts increased model accuracy significantly to 90%.
- Adding other math vectors (word-similarity, word-distance) increased accuracy of base model to 92%.

- Then frequency vectors (unigram, bigram, trigram, most important word) were passed individually, two at a time and all together along with other features like scalars, count vector and math vectors to check for performance improvement.
- Decision Tree and Random Forest models are tuned using the GridSearchCV on the same input features for better results.

Impact

- Amongst all the frequency vectors (unigram, bigram, trigram, most important word) that are passed individually along with the scalars, count vectors and math vectors:
  - unigram frequency vector increased the prediction power of the base model the highest,
  - next came the trigram frequency vector in performance improvement.
- While bigram and most important word frequency vector reduced prediction power.
- Using two frequency vectors also reduced the prediction power of the base model.
- Non usage of standardized features causes slower convergence but improves prediction power.
- Adding more features like word similarities, word distance, noun entity count vector improve prediction power but increases feature extraction 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:

Logistic Regression

(solver: liblinear, regularization: l1, penalty: 0.1).

- **93.06%, Precision: 92.47%, Recall: 93.03%, F1 score: 92.75%**
- It has better performance than Decision Tree and marginally better than Random Forest
- Takes lesser time to train.
- Simple model sufficient for binary classification.
- Has fewer hyper-parameters and hence low maintenance.
- Future data can be expected to have balanced classes because of plentiful true and fake news.
- But it is adversely affected by multi-collinearity, and inherent inflexibility due to lack of more hyper-parameters.