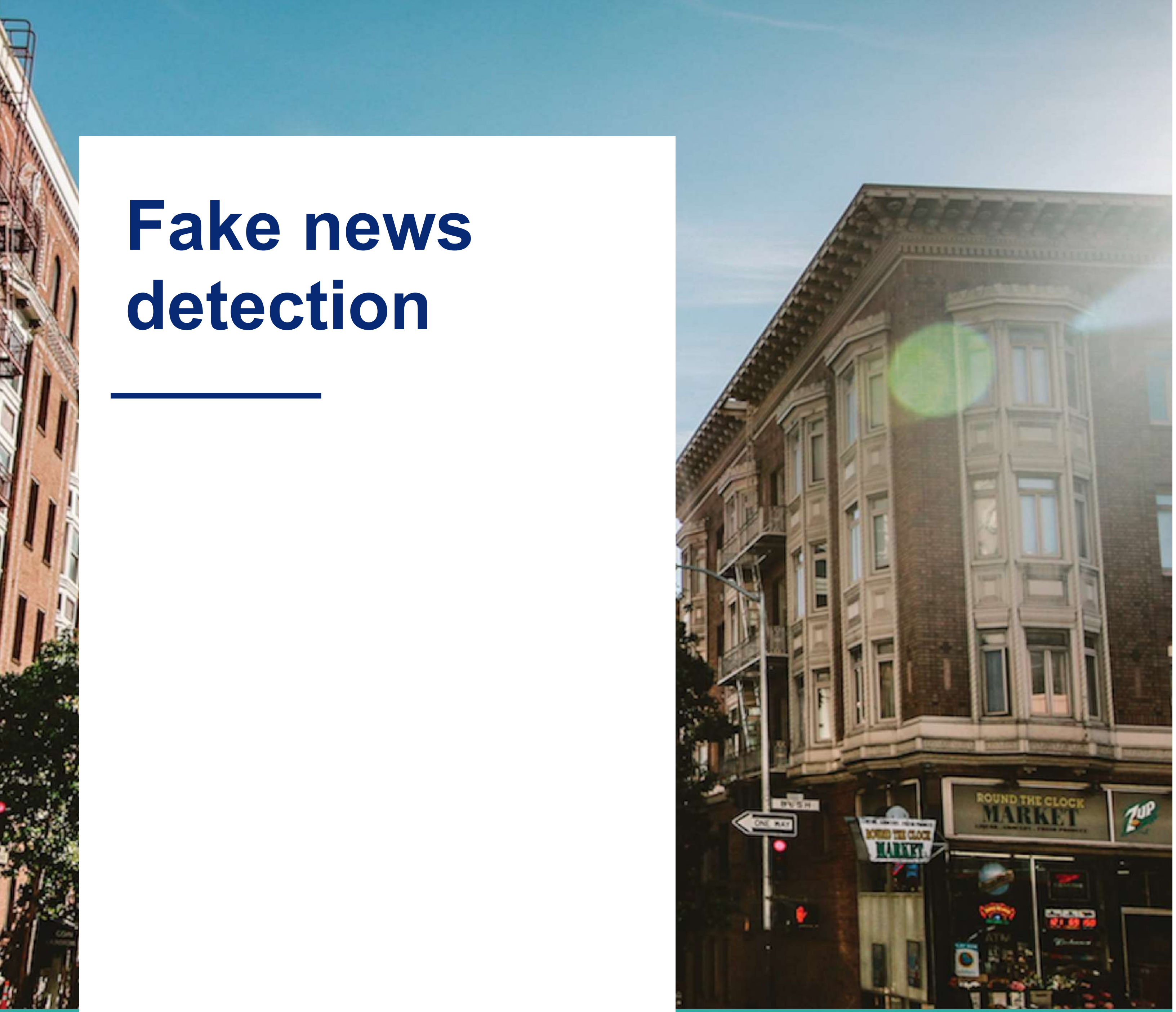
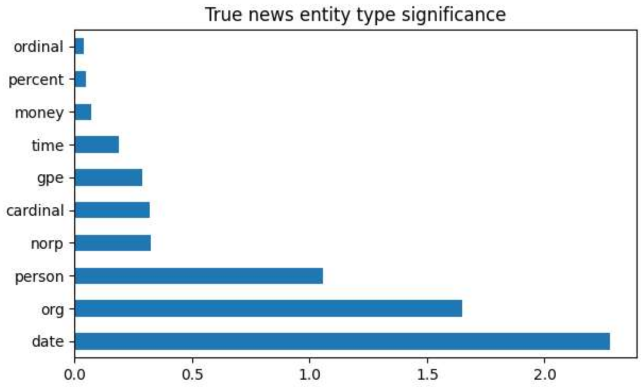


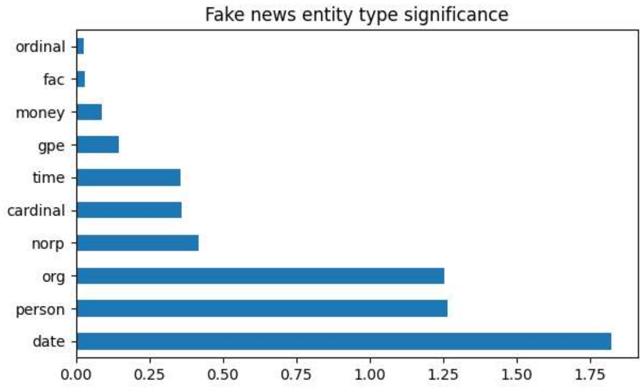
Fake news detection

Group Assignment

Arnab Biswas
Maddi Pranav Reddy
Mohan Nishantam

Entity types

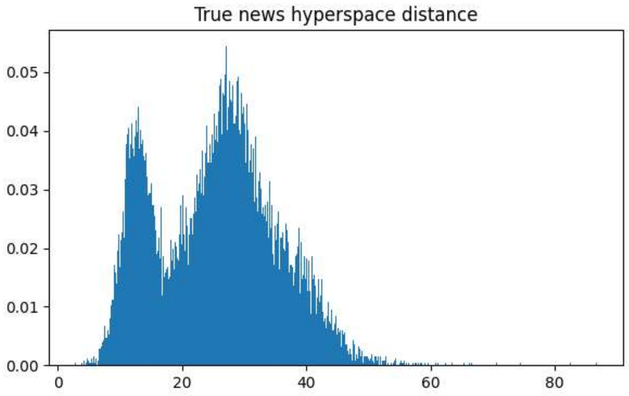


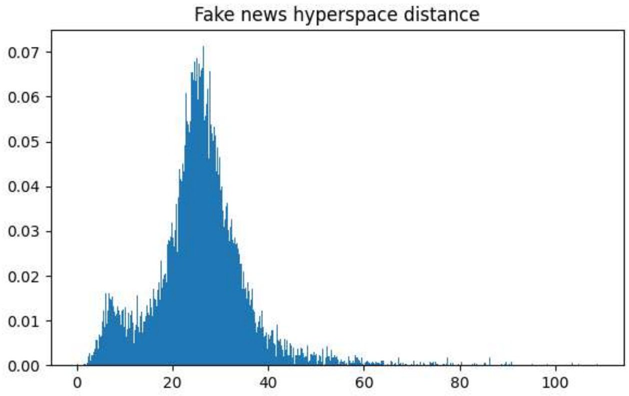


- True news texts focus on organizations, persons, geo-political entities, dates and groups proportionately.
- Fake news focuses more on persons disproportionately.
- This characteristic can be used as an input feature.

Text hyperspace distance

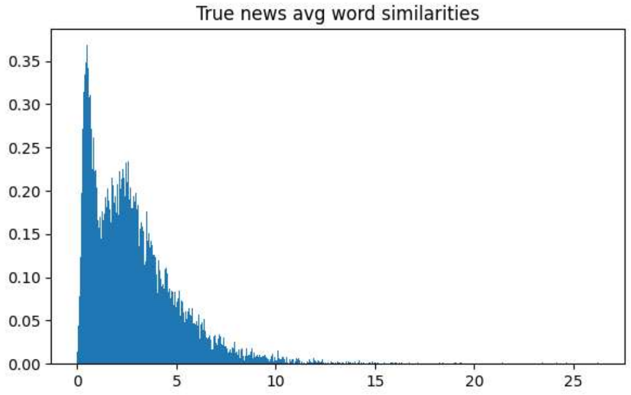
A comparison of text matrix norm is shown below:

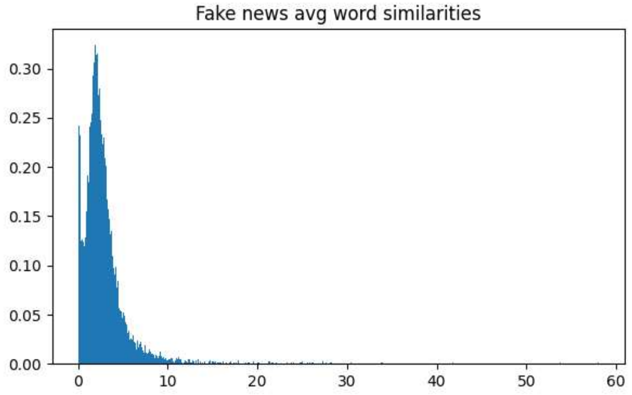




- Text hyperspace distance is the matrix norm of all word vectors in the text
- True news shows double peak which means more depth and variety in content
- Fake news shows a single peak which means less depth and variety in content
- This can be used as a feature for news classification

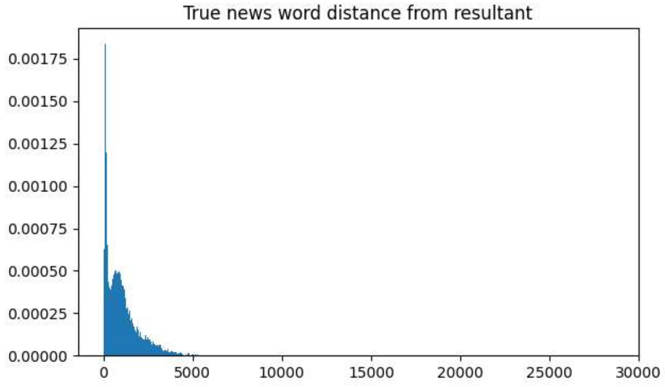
Average word similarities

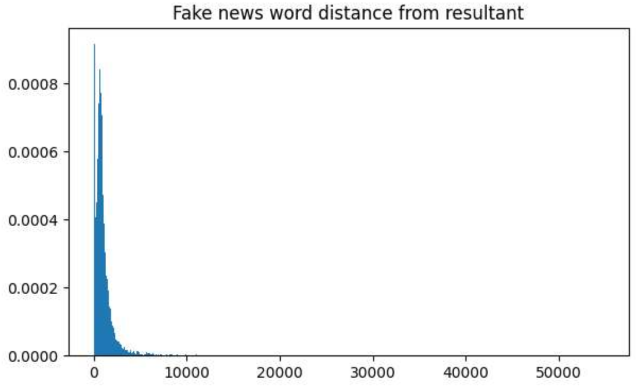




- Average word similarities is the mean of the dot product of each word vector with its resultant taken across each dimension.
- Peak average word similarity for fake news occurs at a higher value than true news which signifies a more purposeful and directed context.

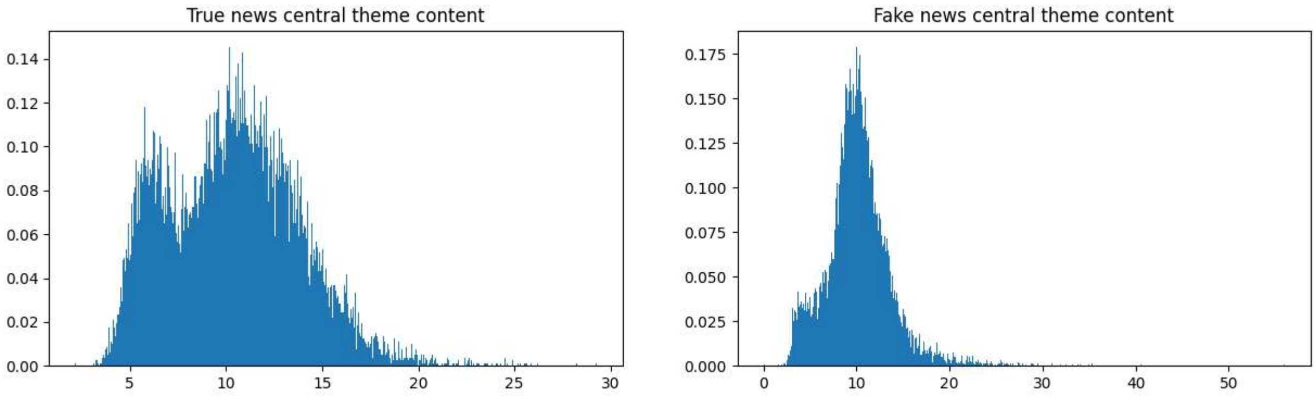
Average word distance from resultant





- 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.

Central theme content



- Theme content is the norm of PCA decomposed word-vectors reduced to 1x300 dimension.
- True news has a more varied theme than fake news which indicates its connection with reality and use of a variety of words.
- Fake news is written with a specific purpose keeping in mind a target audience hence limited theme content.

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

- Non usage of standardized features causes slower convergence but improves prediction power.
- Adding more features like word similarities, word distance, theme vector, NER counts improve prediction power but increases training time.
- Adding more derived features increase multi-collinearity which adversely impacts logistic regression.

Evaluation metric

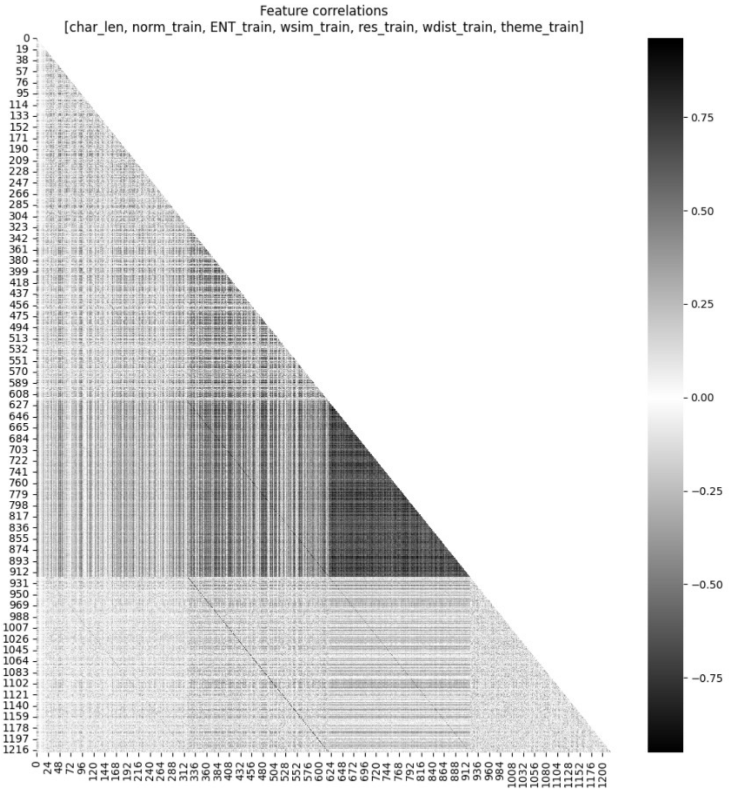
- F1 score is chosen as the best metric because it balances both precision and recall.
- Since both target classes are more or less balanced, both precision and recall can be prioritized instead of only recall.
- Mis-identifying true news as fake and fake news as true can be equally damaging to the reader.

Best model:

Logistic Regression

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

- **92.77%, Precision: 92.26%, Recall: 92.61%, F1 score: 92.43%**
- 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.