News Muse

Introduction:

- Fake News
    - Falls within the broader category of misinformation
    - Intended to mislead audience for gaining benefits
    - We consider only textual fake news

- Online News Popularity
    - Popularity - number of audiences, activities around some news
    - Metadata plays a key role
        - News construct
        - When was the news published?
        - Did the news have any media content?
        - Etc.

Datasets:

Kaggle Fake News Dataset:

- Source: https://www.kaggle.com/c/fake-news/data
- Attributes (total 5):
    - ID
    - Author
    - Title
    - Text
    - Label (fake/real)

UCI Online News Popularity Dataset

- Source: https://archive.ics.uci.edu/ml/datasets/online+news+popularity
- Attributes (total 61):
    - URL (1)
    - Timedelta (1)
    - Syntactical (6)
    - Topical (11)
    - Sentiment (16)
    - Keyword (10)
    - Media Frequency (4)
    - One-hot encoded publication day (8)
    - Self reference (3)
    - Target variable - number of social media shares (1)

Connecting the Datasets:

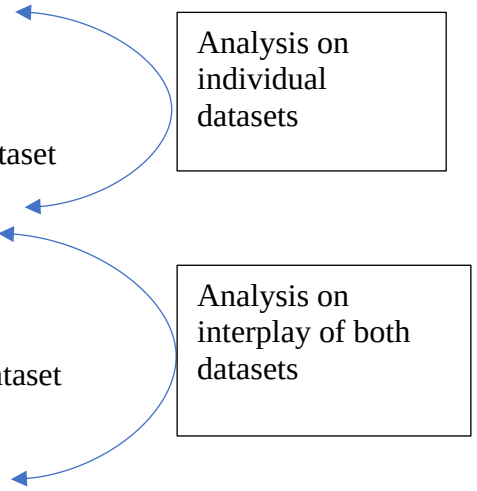
- The UCI dataset contains only metadata about news text (e.g. number of words in news article and title, subjectivity and sentiment of news article and title etc.)

- The Kaggle dataset contains only news title and text (no metadata)

- For combined analysis, we select the common metadata features from UCI dataset and extract the same features for the Kaggle dataset (following the feature guideline in UCI Online News Popularity dataset's original paper [1])

Goals:

- Detection of fake vs. real news
  - Classifier modeled on Kaggle Fake News Dataset
- Identification of popular vs. unpopular news
  - Classifier modeled on UCI Online News Popularity Dataset
- (Real) News Recommender
  - Clustering applied to the Kaggle Fake News Dataset
- Kaggle news impact analysis
  - Regressor modeled on UCI Online News Popularity Dataset
  - Applied to Kaggle Fake News Dataset
  - Reliant on domain adaptation techniques



Individual Dataset Analysis (1):

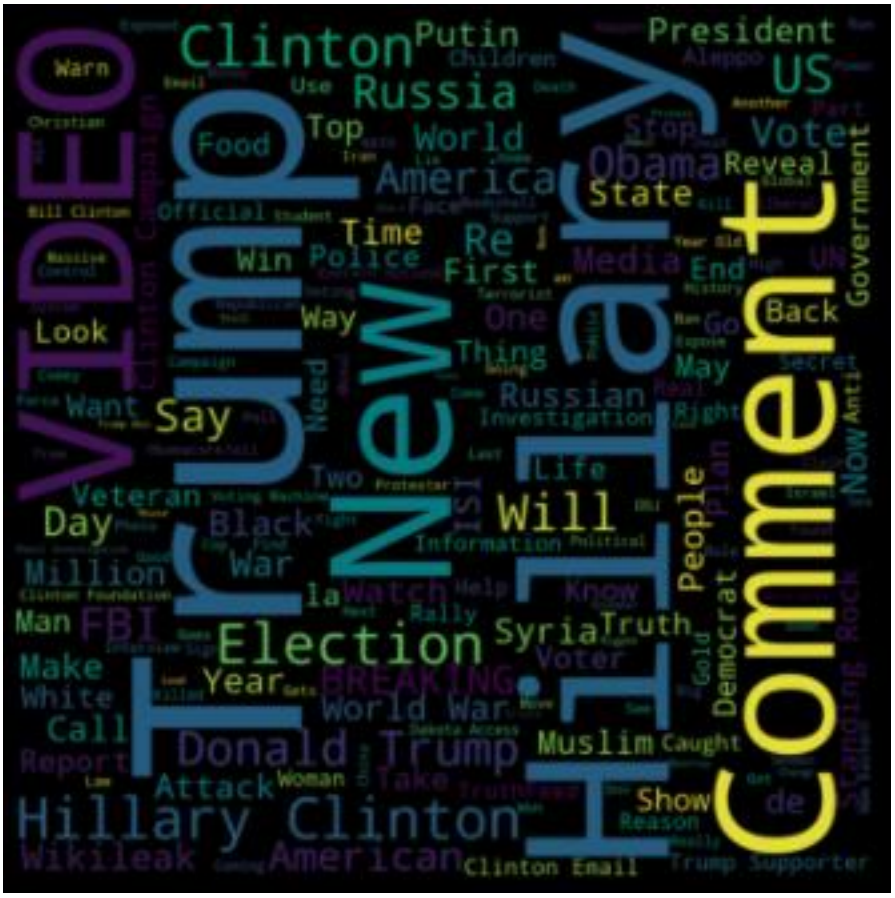
Fake News Classifier on Kaggle Fake News Dataset:

- Classify news articles
- Fake or Real
- Training Model - Labeled Kaggle Data source
- Data cleaning and Preprocessing
  - Remove NAN records and reindex
  - Drop insignificant columns
- Feature Engineering
  - Stemming and Lemmatization
  - Remove special characters
  - Change to lowercase
  - Remove stop words
  - Bag of words using Count Vectorizer
  - Term frequency and Inverse document frequency vectorizer.
- Classification Algorithms
  - Multinomial Naïve Bayesian
  - Passive Aggressive classifier
- Observations
  - Best accuracy: 91% (TF-IDF with PAC)
  - Confusion Matrix:
  - Passive Aggressive Classifier better than Multinomial Naive Bayesian

- Countvectorizer and TF-IDF have similar accuracy patterns across both algorithms.
- “Title” over “Text” trade off

- **Terms impacting Fake news**

```
[(-10.79985853025502, 'said mr'),
 (-10.394393422146857, 'breitbart news'),
 (-9.63670772044934, 'mr obama'),
 (-9.183776075102251, 'ms'),
 (-9.071157196994104, 'presid trump'),
 (-9.008099061026966, 'session'),
 (-8.980700086838851, 'breitbart'),
 (-8.954031839756691, 'sport'),
 (-8.802893210399425, 'islam state'),
 (-8.802893210399425, 'saturday')]
```

- 

- **Terms impacting Real news**

```
[(-4.661627643430595, 'trump'),
 (-4.673426122736554, 'clinton'),
 (-4.855485590680891, 'one'),
 (-4.856272063391502, 'peopl'),
 (-4.913569330764339, 'state'),
 (-4.99061554494066, 'would'),
 (-4.99402099694972, 'us'),
 (-5.1119796274340334, 'hillari'),
 (-5.113674447341442, 'like'),
 (-5.141305898573496, 'time')]
```

-

Individual Dataset Analysis (2):

Popularity Classifier on UCI Online News Popularity Dataset.

Features:

- Using the raw content of a new article, 60 features are derived and one target feature having the number of shares of that news article.

- We have 40k instances with 59 features (excluding URL and target) to be modeled for our data analysis.

- This set of features would cover below three main aspects of the raw content
  - Text aspect
  - Statistical aspect
  - General aspect

- There are both dependent and independent features in our dataset.

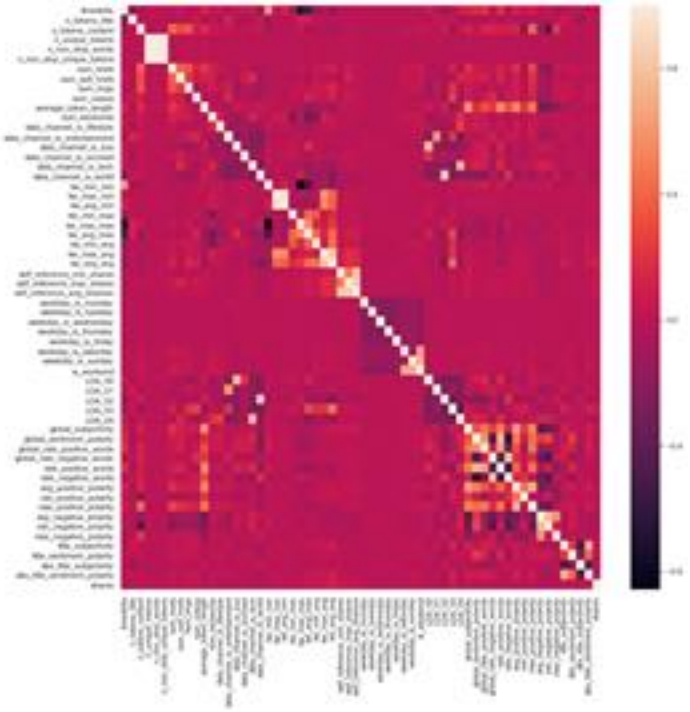
Exploratory Analysis

- We have two kinds of features – Binary and continuous
  - Binary – Features that are categorical and created by one hot encoding.
  - Continuous – Features that are derived from statistics (min, avg and max) of language processing aspects

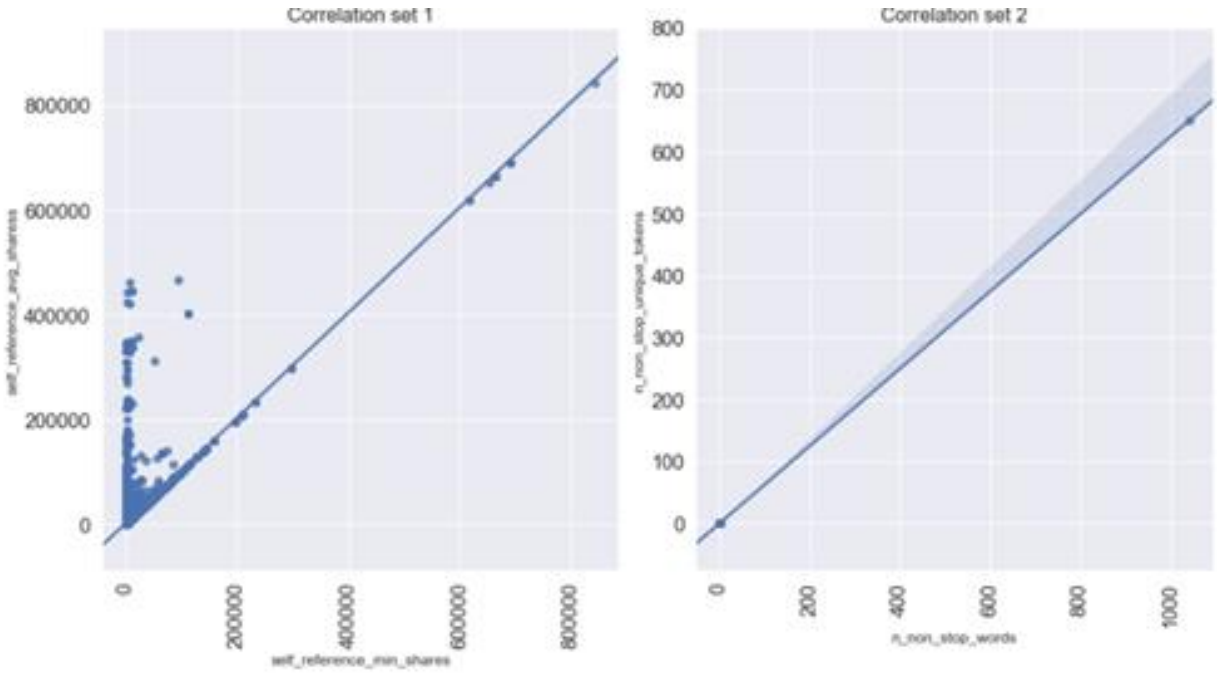
- We have features that are dependent and independent. Our features form different sets. Each set of these features corresponds to an aspect of the news article.

- Features from the same set are dependent and features across the sets are independent.

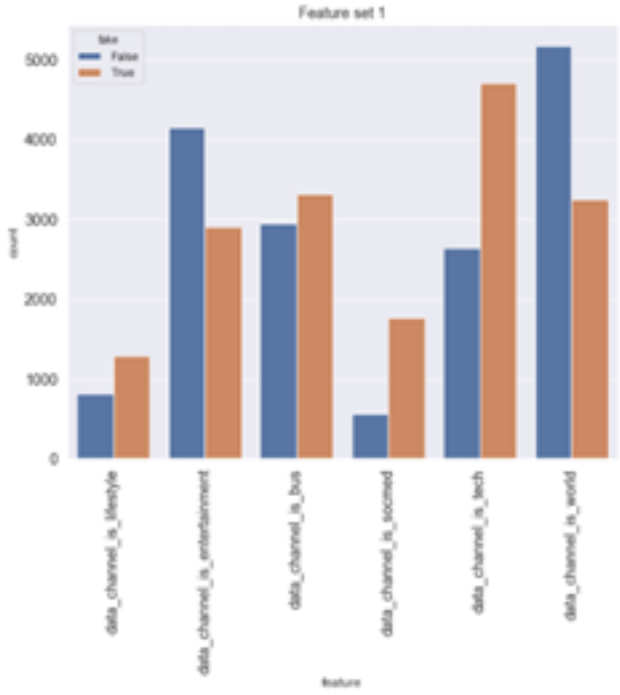
- We also have the features that have negative values and skewed features.



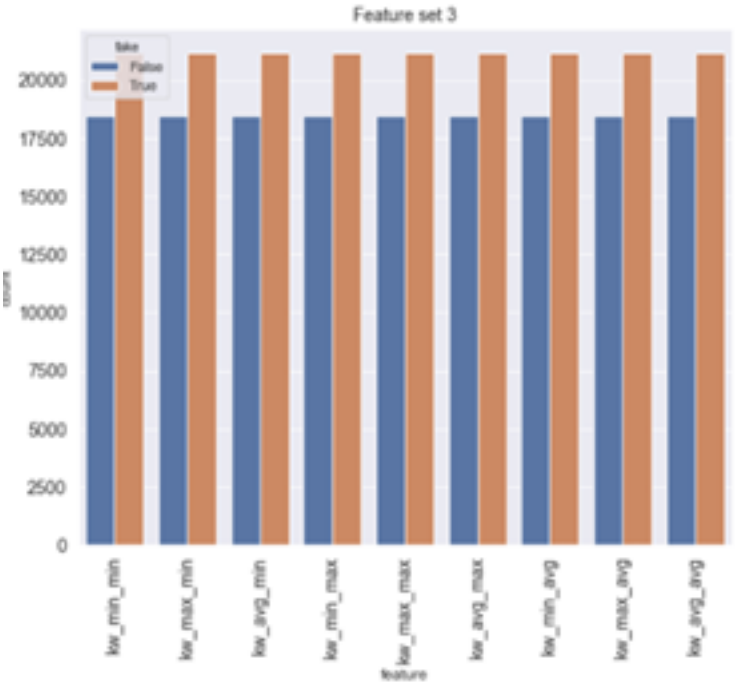
Correlation Heat map



- Regression plots between sample features



Binary Feature Count plot



Continuous feature count plot