Identifying Fake News

Team Fake News

Research Question

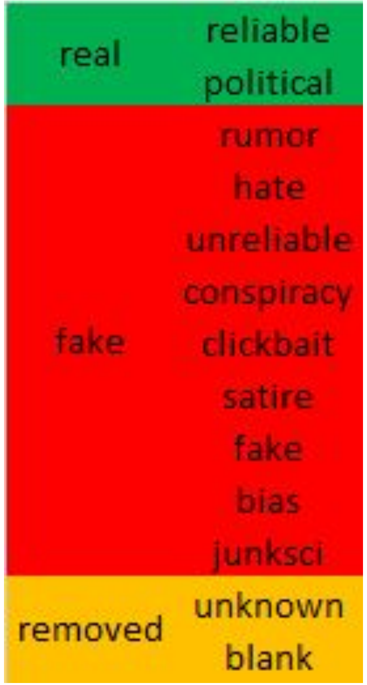
By looking at the text of an article/tweet/blog post, can you determine a statistical method that will determine whether or not it is fake news?

Background

- Information age
- The internet is ever-growing
  - 3.4 billion users
  - 1.6 billion domains
- Everyone is a news source
  - 6,000 tweets per second
- Everyone is looking for information
  - 40,000 google queries per second

Dataset

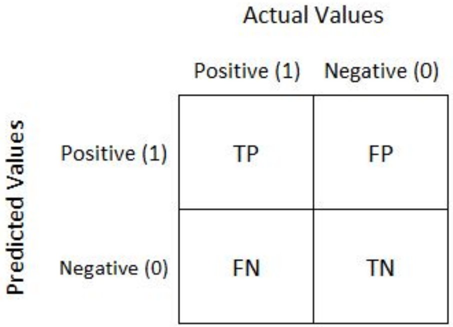
- Fake News Corpus
  - 29.3 GB
  - More than 9 million rows
  - Key variables: content, type
    - Content: text content of the article
    - Type: rumor, hate, unreliable, conspiracy, clickbait, satire, fake, reliable, bias, political, junksci, unknown, blank
    - Categorized pre-existing types into “real” and “fake”



real	reliable
	political
fake	rumor
	hate
	unreliable
	conspiracy
	clickbait
	satire
	fake
	bias
	junksci
removed	unknown
	blank

Methodology

- Group different types of news into fake and real news
- Use classification model to predict
- Represent results as confusion matrix
- Evaluation statistic: accuracy %



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Methodology

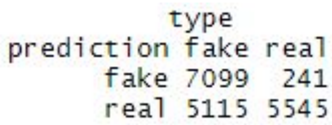
- Naive Bayes Model
  - Probabilistic classifier
  - Assigns every word a probability (real or fake)
  - Looks at combinations of words
- SVM
  - Non-probabilistic
  - Envisions articles as points in n-dimensional space
  - Splits points between real and fake

Methodology

- Technical Limitations
  - Space
  - 29.3 GB dataset
  - SVM is computationally expensive
  - Solution:
    - Load first n lines of dataset
    - Subsampling
      - Repeatedly sample dataset
      - Train model on sample
      - Aggregate results of cross-validation

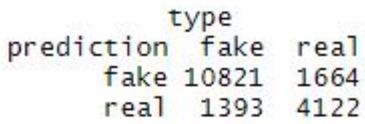
Results

- Naive Bayes Model
  - Accuracy: ~70%
  - Better at isolating real news



```
           type
prediction fake real
      fake 7099  241
      real 5115 5545
```

- SVM Model
  - Accuracy: ~83%
  - Better at isolating fake news



```
          type
prediction  fake  real
      fake 10821  1664
      real  1393  4122
```

Influencing Factors

- Space Limitations
  - Limited amount of data to load
  - SVM computationally expensive, limits sample size
- Time Limitations
  - How many subsets to take?
  - Diminishing returns
- Distribution of real and fake news in dataset
  - More fake news than real news
- Article categorization

Further Improvements

- Eliminate limitations
  - More computationally able computer
  - Load entire dataset
  - Larger sample sizes
  - Experiment with changing influencing factors
- Other classification methods
  - Linear/Quadratic Discriminant Analysis
  - Random Forest Classification
  - Logistic Regression: domain
- Semantic Analysis

Conclusion

- SVM best classification model
  - Computationally expensive, but more accurate
  - Not as good at determining real news, but low error % for fake news predictions
  - 83% accuracy rating
- Naive Bayes model deficient
  - Computationally efficient
  - “Baseline” model
  - Inaccurate in comparison to SVM