Mining for Credibility ⛏

we done boiz

Mahdokht Afravi
Jonathan Avila
Cristian Ayub
Gerardo Cervantes

Background

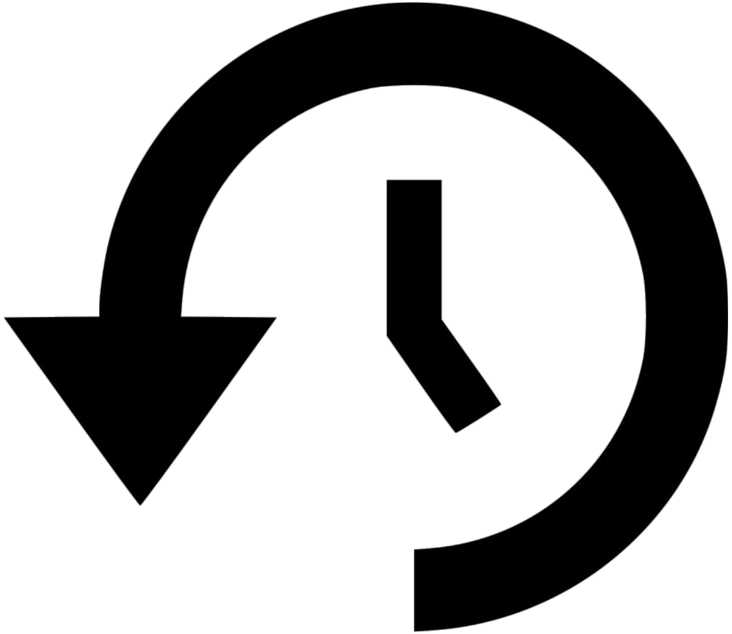
Fake News Corpus

- Labels
  - Fake News
  - Conspiracy Theory
  - Credible
  - Proceed with caution
- Relation between the articles contained
- Likelihood of an article label

Phase I

Recap

- Date clean up
- Preprocessing
- Created Vocabulary
- Algorithms
    - k-means
    - DBSCAN



Recap - Preprocessing

- Text to numerical representation
  - Stop word removal
  - Stemming - PorterStemmer
  - Lowercasing
  - Tokenization
  - Punctuation removal

Results

- Results
  - Sparse data
  - Negative Silhouette Coefficients
- Change of strategy
  - Choose articles
  - Trim subset
  - Change algorithm

Phase II

Clustering

- k-means
  - Average Silhouette
  - Using *k*=8 resulted in interesting set of clusters
- DBSCAN
  - Find ‘good’ parameters
  - Minimize noise
  - Distance measures

Results (k-means)

- 20k word vocab size
- 30k total articles
- Silhouette coefficient: 0.217
- 8 clusters

# Articles	Conspiracy	Fake	Reliable
19671	5499	6209	7885
7174	3168	2981	1025
1845	845	411	589
833	367	299	167
314	83	78	153
181	18	1	162
42	14	14	14
18	6	7	5

Table 1: Eight clusters, with number of articles for each label

Results (k-means)

2 biggest clusters

- Cluster with 19671 articles: 5529 conspiracy, 6249 Fake, 7893 Reliable
  - Words: 'one', 'state', 'peopl', 'us', 'time', 'would', 'said', 'year', 'like', 'also', 'trump', 'christian'
- Cluster with 7181 articles: 3178 conspiracy, 2969 Fake, 1034 Reliable
  - Words: 'one', 'new', 'state', 'american', 'peopl', 'us', 'time', 'would', 'year', 'like', 'govern'

2 smallest clusters

- Cluster with 18 articles: 6 conspiracy, 7 Fake, 5 Reliable
  - 'one', 'new', 'state', 'american', 'time', 'would', 'even', 'report', 'year', 'write', 'world', 'like', 'govern', 'iran', '2009'
- Cluster with 64 articles: 25 conspiracy, 21 Fake, 18 Reliable
  - 'one', 'new', 'state', 'say', 'peopl', 'time', 'would', 'even', 'report', 'year', 'get', 'go', 'like', 'also', 'govern'

Results (k-means)

Interesting clusters

- Cluster with 181 articles: 18 conspiracy, 1 Fake, 162 Reliable
    - 'se', 'de', 'la', 'lo', 'al', 'el', 'con', 'su', 'un', 'en', 'que', 'para', 'del', 'es'
- Cluster with 290 articles: 77 conspiracy, 72 Fake, 141 Reliable
    - ['war', '–', 'one', 'state', 'american', 'peopl', '’', 'us', 'time', 'would', '“', '”', 'presid', 'also', 'govern']

Results (DBSCAN)

- Better results with higher *eps*
  - Decreased amount of noise points
  - Distance measured from noise points to nearest cluster's core points
- Bad metrics
  - Silhouette coefficient
  - Homogeneity

Matrix	Clusters	Noise
15K x 20K	2	4.6%
30K x 20K	3	4.3%

Phase III

Model

- Motivation is to get a binary output on whether an article is fake or credible
- Features are our vocabulary with term frequency
- Training data has 10k reliable and 10k fake articles
- 75% train data, 25% test data

Linear Regression Results

- Test accuracy: **66.3%**
- Baseline: 50.0%
- Threshold 0.5

High coefficients	‘fortyeight', 'heartach', 'dissoci', 'compatriot', 'harbing', 'valerian', 'mosh', 'cha', '3800', 'turncoat', 'olympia', 'wacko'
Low coefficients	'transvers', 'claudia', 'manitoba', 'molecular', 'hydrolyz', 'landhold', 'noncompetit', 'gorg', 'conciliatori', 'nazca', 'generalpurpos', 'vino’

Naïve-Bayes

- Bad results because of features
- Articles used were *fake* and *reliable*

Conclusion

Conclusion

- Reduced amount of noise in all clusters
- Analysis of cluster data was easier
- Labeling prediction had an accuracy of 66% accuracy
- Deep learning techniques could have better accuracy