Identifying Fake News

Project Option 1

By: Jamie Mandolini

01

Pre-Processing

Data Cleaning and Formatting

Cleaning Methods

- Change uppercase to lowercase
- Remove punctuation, digits, and stopwords
- Tokenize and lemmatize

Formatting

- Create corpus and convert tokens to numerical representation using TFIDF
- Create new dataframe with features:
  - Cosine similarity
  - Polarity for each title
  - Subjectivity for each title

*Note: EDA showed a very uneven dataset

02

Model Training

Model Training: Choosing a Classifier

Classifiers Used:

- Unweighted and balanced Random Forest
- Gaussian Naive Bayes
- Logistic Regression (one vs rest model)
- Unweighted, balanced, and weighted Multinomial Logistic Regression

03

ANALYSIS

Performance Metrics

Multinomial Logistic Regression Performance

	precision	recall	f1-score	support
agreed	0.66	0.46	0.54	22176
disagreed	0.00	0.00	0.00	2012
unrelated	0.78	0.91	0.84	52745
accuracy			0.76	76933
macro avg	0.48	0.46	0.46	76933
weighted avg	0.72	0.76	0.73	76933

Balanced Multinomial Logistic Regression Performance

	precision	recall	f1-score	support
agreed	0.61	0.55	0.58	22176
disagreed	0.03	0.22	0.06	2012
unrelated	0.86	0.73	0.79	52745
accuracy			0.66	76933
macro avg	0.50	0.50	0.48	76933
weighted avg	0.77	0.66	0.71	76933


Thanks!