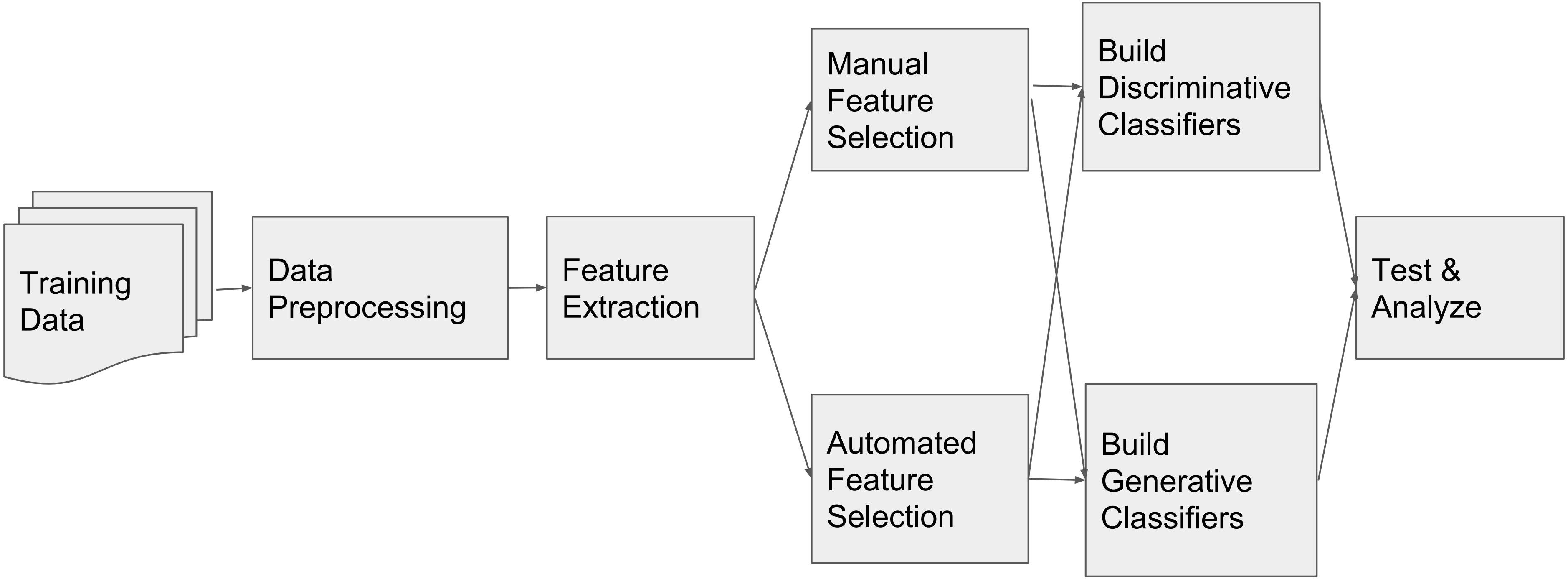
Authorship Classification

Joshua Shapiro | Ishan Sharma | Shuqing Zhang

Problem Definition

- Authorship classification has been a long studied problem in the domain of linguistics
- Existed long before computers
  - 1400s *Donation of Constantine*
  - Federalist Papers
  - Shakespeare’s Plays
- Current computational approaches exist, but vary tremendously on feature selection, accuracy, and performance.
- Goal: To discover subset of features that yields best accuracy with lowest performance cost.

Solution Overview



Training Data

Data Preprocessing

Feature Extraction

Manual Feature Selection

Automated Feature Selection

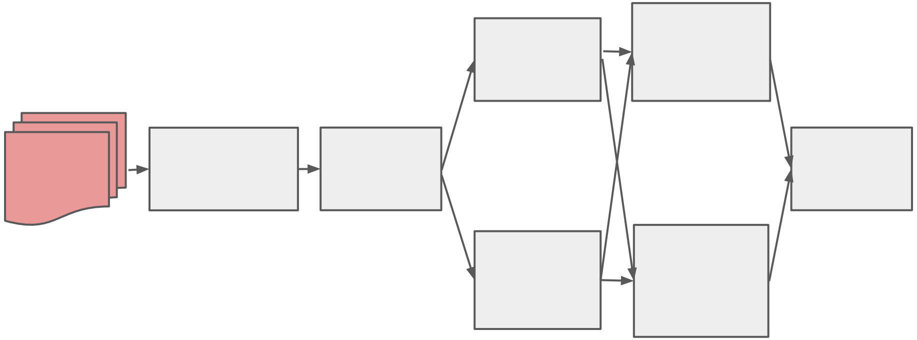
Build Discriminative Classifiers

Build Generative Classifiers

Test & Analyze

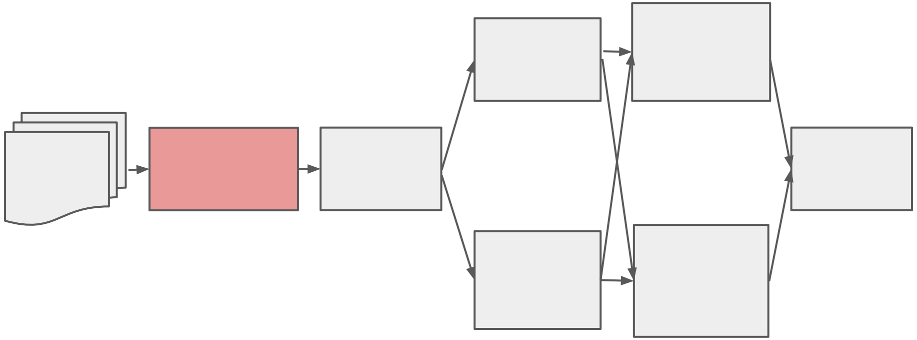
Training / Test Data

- Reuters 50-50 Dataset
- 2500 training documents, 50 authors
- 2500 test documents, 50 authors
- All documents are news articles



Data Preprocessing

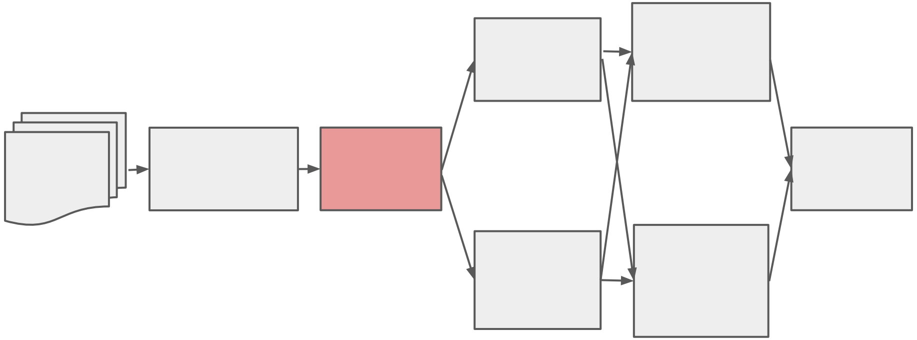
- Separate labels from data and parse articles into proper format
- Group articles by author
- Tokenize, lemmatize, and stem each article
- Generate POS tags for data



Feature Extraction

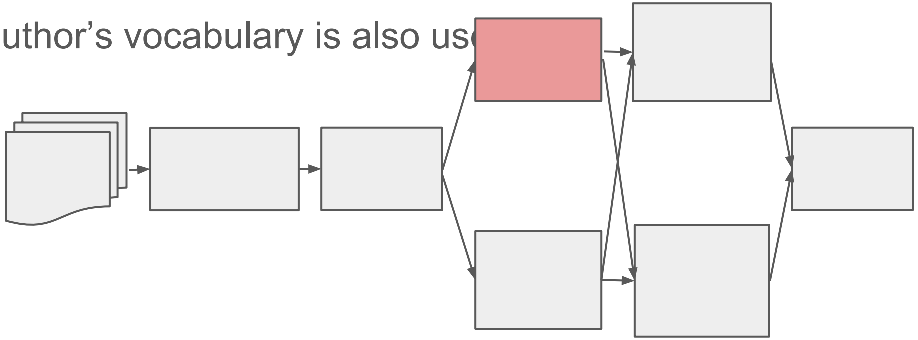
- Frequency of conjunctions
- Frequency of modals
- Frequency of determiners
- Frequency of quantifiers
- Frequency of pronouns
- Number of sentences per article
- Number of unique words per article
- Average length of sentence
- Average word length
- Number of total words per article
- Number of periods
- Number of commas
- Number of colons
- Number of semicolons
- Number of exclamation marks
- Number of question marks
- N-grams of text
  - 1,2,3
  - surface,tokens,lemmas

- Bag of words(TF-IDF)
- N-grams of parts of speech
  - 1,2,3
- POS dependencies



Manual Feature Selection

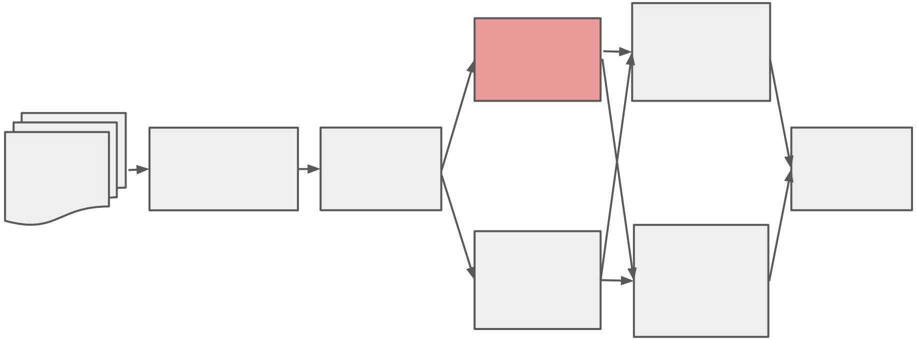
- Stylometry
  - The statistical analysis of style, stylometry, is based on the assumption that every author's style has certain features being accessible to conscious manipulation. Therefore they are considered to provide a reliable basis for the identification of an author.
- Bag of Words
  - The bag-of-words representation, where the document is represented with a vector of the word counts that appear in it. Depending on the classification method, the bag-of-words vector can be normalized to unity and scaled so that common words are less important than rare words
  - Measuring the “richness” or “diversity” of an author’s vocabulary is also us discriminating feature.
  - Use TF-IDF



Manual Feature Selection

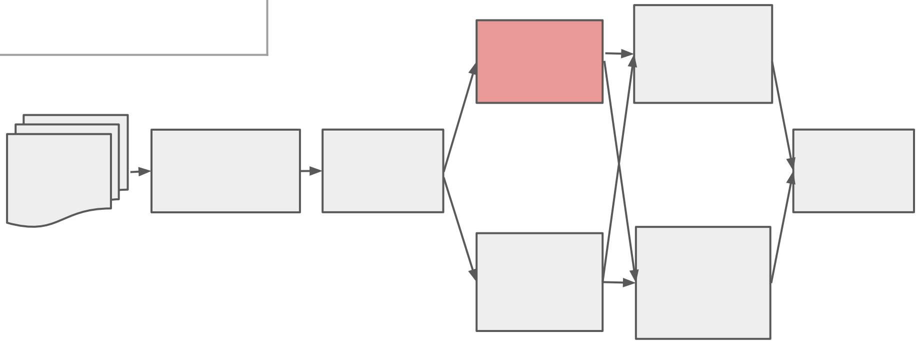
- Frequency of function words
  - The function words (modal, pronoun, conjunction) are used as a discriminating feature of author

- All features
  - Self explanatory



Stylometric Features

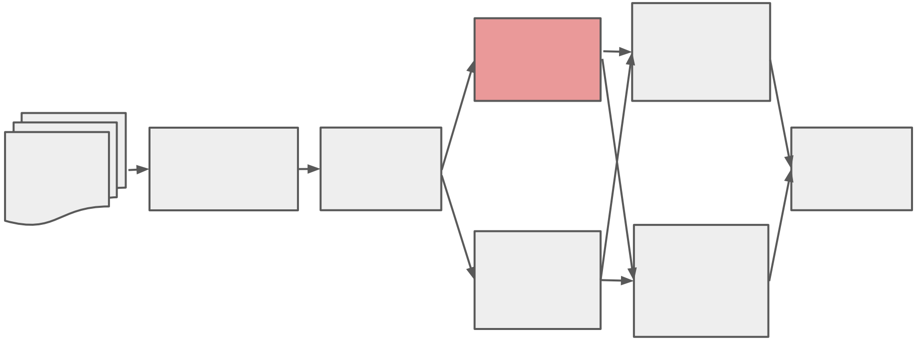
Number of Sentences	Number of words	Average Sentence Length
Average Word Length	Number of Different Words	Number of Periods
Number of Commas	Number of Colons	Number of Semi colons
Number of Exclamation Marks	Number of Question Marks	



Features

- Frequency of Function Words

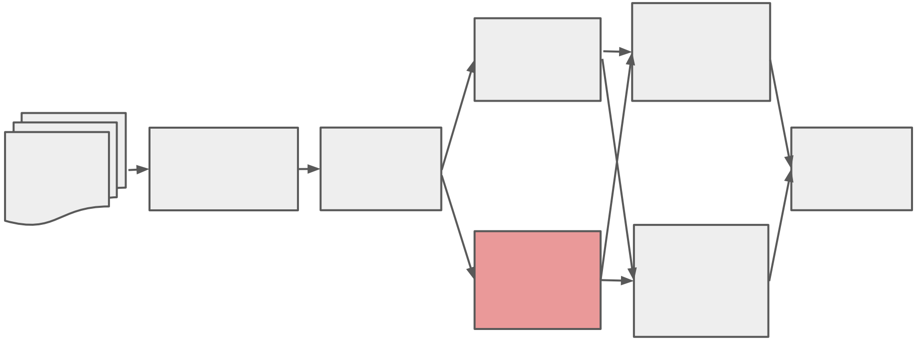
Conjunction	Modals	Determiners
Quantifier	Pronoun	



Automated Feature Selection

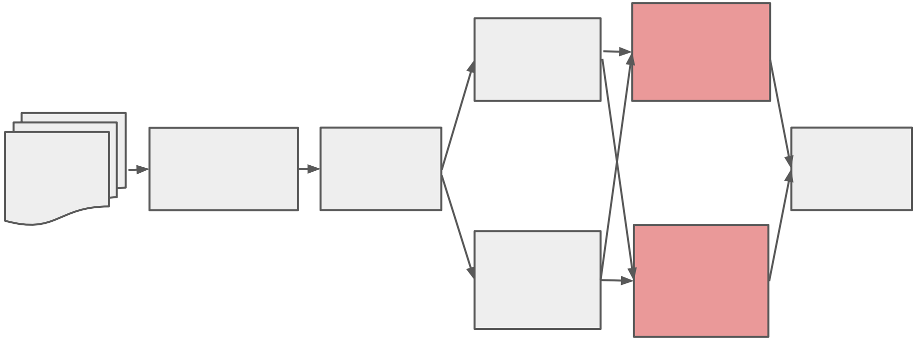
- Univariate Feature Selection -- SelectKBest, preprocessing
- Recursive Feature Elimination -- prune lowest weighted features
- L1-based feature selection -- SelectFromModel & LassoCV

- All part of scikit



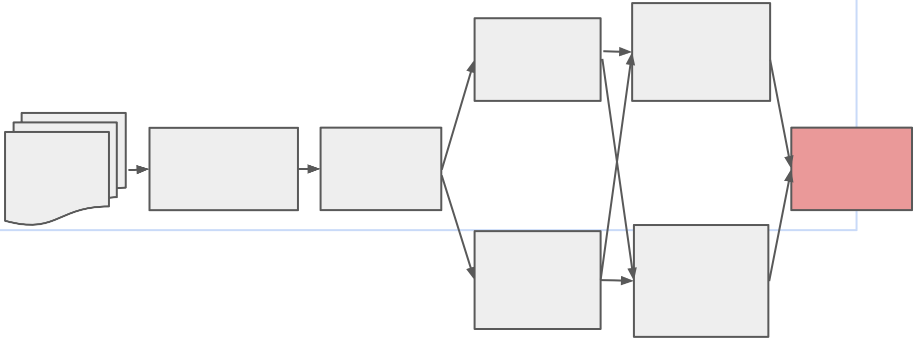
Build Classifiers

- Discriminative Classifier
  - SVM -- Linear, Polynomial, Gaussian Radial Basis Function
- Generative Classifier
  - Naive Bayes



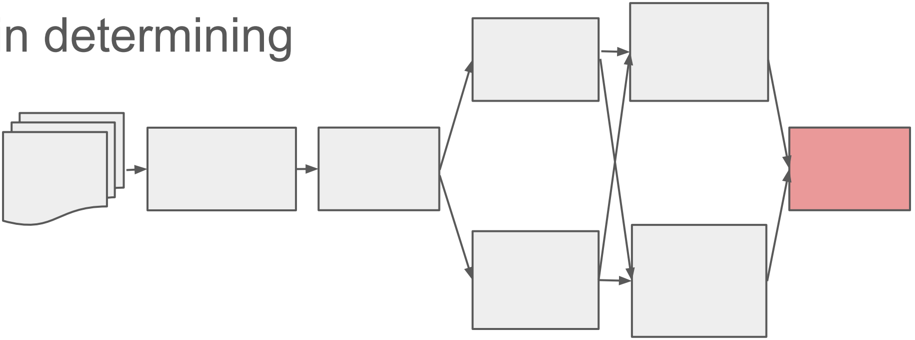
Test & Analyze

- Baseline
  - Trigram language models with smoothing
    - Provides baseline on actual data
    - Surface, token, lemma



Test & Analyze

- Given:
  - Feature subset
  - Classifier
- Output:
  - Accuracy
  - Performance
- By the end we will be able to determine the best feature subset and classifier pair that optimizes both accuracy and performance as well as see which individual features are most important in determining authorship classitionton



Reference

- Peng, F.; Schuurmans, D.; Keselj, V; Wang, S. (2003). Language Independent Authorship Attribution using Character Level Language Models. Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, 267--274
- A. (Ed.). (2015). Authorship Attribution with Python. Retrieved November 15, 2016, from Piazza.com/class/isg9se99b5s6nw?cid=1
- Nirkhi, S. (2015, May). Stylometric Approach For Author Identification of Online Messages. 1-2. Retrieved from http://ijcsit.com/docs/Volume 5/vol5issue05/ijcsit2014050536.pdf
- K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott,"Stylometry for E-mail Author Identification and Authentication",Seidenberg School of CSIS, Pace University, New York
- ZhiLiu, Reuter_50_50 Data Set. Retrieved November 15, 2016, from http://archive.ics.uci.edu/ml/datasets/Reuter_50_50
- Shihara, S. (n.d.). A Forensic Authorship Classification in SMS Messages: A Likelihood Ratio Based Approach Using N-gram. *Likelihood Ratio Calculation,* 51-53. Retrieved from http://www.alta.asn.au/events/alta2011/proceedings/pdf/U11-1008.pdf
- Castro, A., & Lindauer, B. (n.d.). Author Identification on Twitter. Retrieved November 15, 2016, from http://cs229.stanford.edu/proj2012/CastroLindauer-AuthorIdentificationOnTwitter.pdf
- Elayidom, S., Jose, C., Puthussery, A., & Sasi, N. K. (2013, September 5). TEXT CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION ANALYSIS. 1-8. Retrieved November 14, 2016, from http://airccse.org/journal/acij/papers/4513acij01.pdf
- Nirkhi, S., Dharaskar, R., & Thakare, V. (n.d.). Authorship Identification in Digital Forensics using Machine Learning Approach. Retrieved from http://www.ijltet.org/wp-content/uploads/2015/01/54.pdf
- AICBT, Authorship Attribution with Python. Retrieved November 8, 2016, from http://www.aicbt.com/authorship-attribution/
- Nazar, R; Marta, S.P. (2006) An Extremely Simple Authorship Attribution System. Proceedings of the Second European IAFL Conference on Forensic Linguistics / Language and the Law, Barcelona 2006.