Predicting Paper Authorship

Springboard Capstone 1

The Problem

- Plagiarism in scientific writing
  - 30% of Scientists have witnessed it
- Outright fraud
  - Russian paper mill offered authorship on paper for $5000
- Can NLP help this problem?

Who Would be Interested?

+ Scientific Journals
+ Universities/University Libraries
+ Professors
+ The general public

Data Wrangling

Merging and Cleaning the Data

Data Wrangling

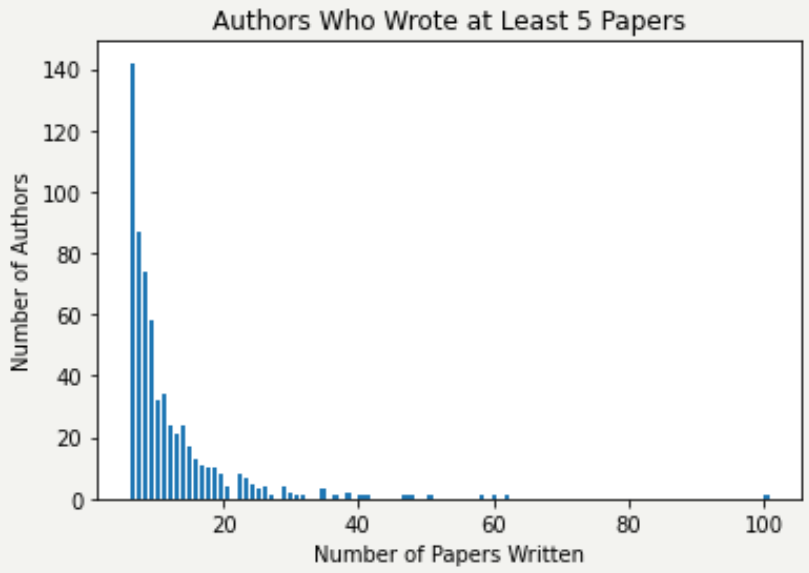
- Joined three separate .csv files such that each paper was duplicated once for each author
- Features included title, abstract, year published, event type, author, and paper text
- Four papers had duplicated paper texts
- Some identical author names had different author ids

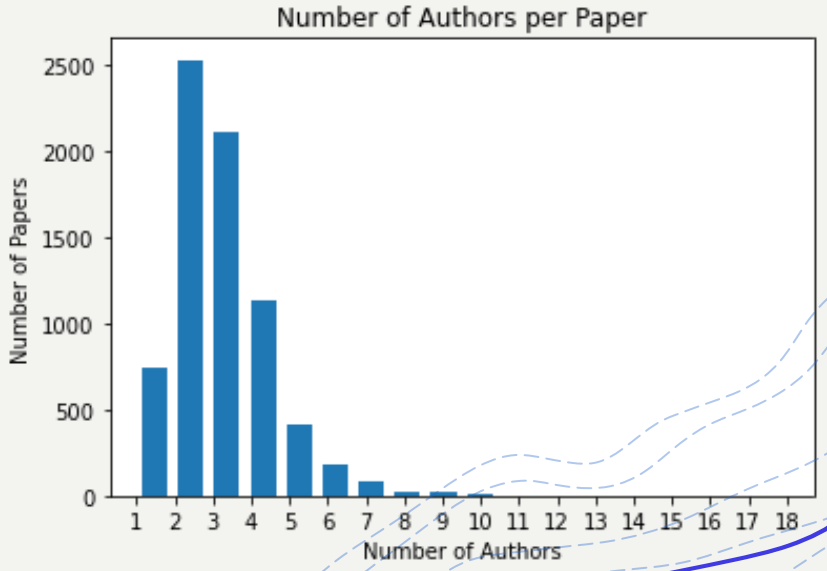
	year	title	event_type	pdf_name	abstract	paper_text	paper_id	author_id
0	2016	Only H is left: Near-tight Episodic PAC RL	Poster	6052-only-h-is-left-near-tight-episodic-pac-rl...	In many applications such as advertisement pla...	Launch and Iterate: Reducing Prediction Churn\...	6052	8474
1	2017	Deep Multi-task Gaussian Processes for Surviva...	Poster	6827-deep-multi-task-gaussian-processes-for-su...	Designing optimal treatment plans for patients...	Deep Multi-task Gaussian Processes for\nSurviv...	6827	9344
4	2017	Deep Multi-task Gaussian Processes for Surviva...	Poster	6827-deep-multi-task-gaussian-processes-for-su...	Designing optimal treatment plans for patients...	Deep Multi-task Gaussian Processes for\nSurviv...	6827	9351
2	2017	Bayesian Inference of Individualized Treatment...	Poster	6934-bayesian-inference-of-individualized-trea...	Predicated on the increasing abundance of elec...	Bayesian Inference of Individualized Treatment...	6934	9344
5	2017	Bayesian Inference of Individualized Treatment...	Poster	6934-bayesian-inference-of-individualized-trea...	Predicated on the increasing abundance of elec...	Bayesian Inference of Individualized Treatment...	6934	9351

Potential issues

+No record of who was the primary author!

+A very imbalanced dataset, most authors write only a few papers!

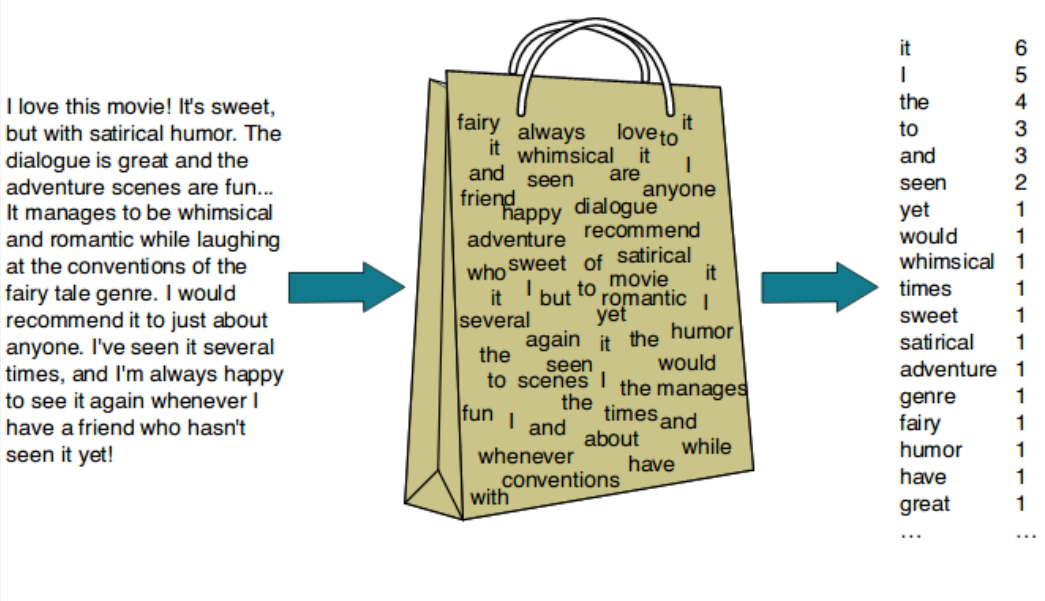




Initial Modeling

Predicting an author from just the title

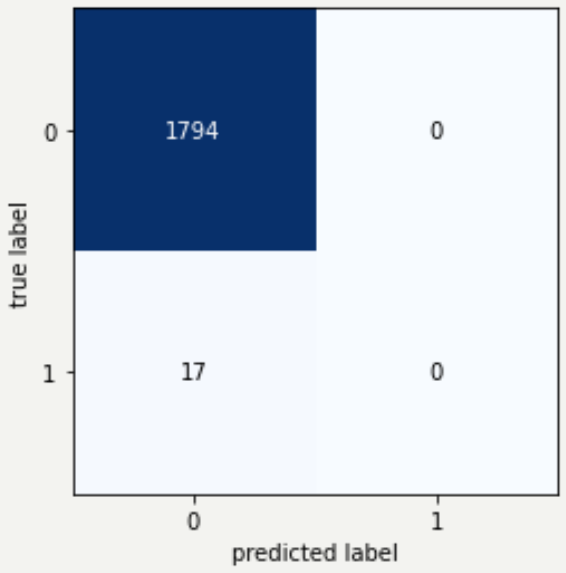
Initial Modeling

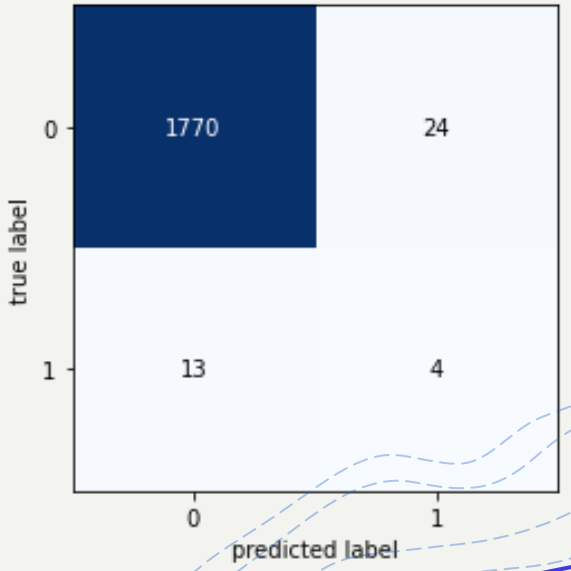


- Focused on predicting one author (Bernhard Schulkopf)
- Used only the title of the paper as a feature
- Scikit-learn's TFIDF vectorizer
- 4865 new features generated

Initial Modeling

- Very imbalanced, Bernhard is only 0.8% of the dataset!
- Used class weights, oversampling, and undersampling
- For this stage, only logistic regression was used

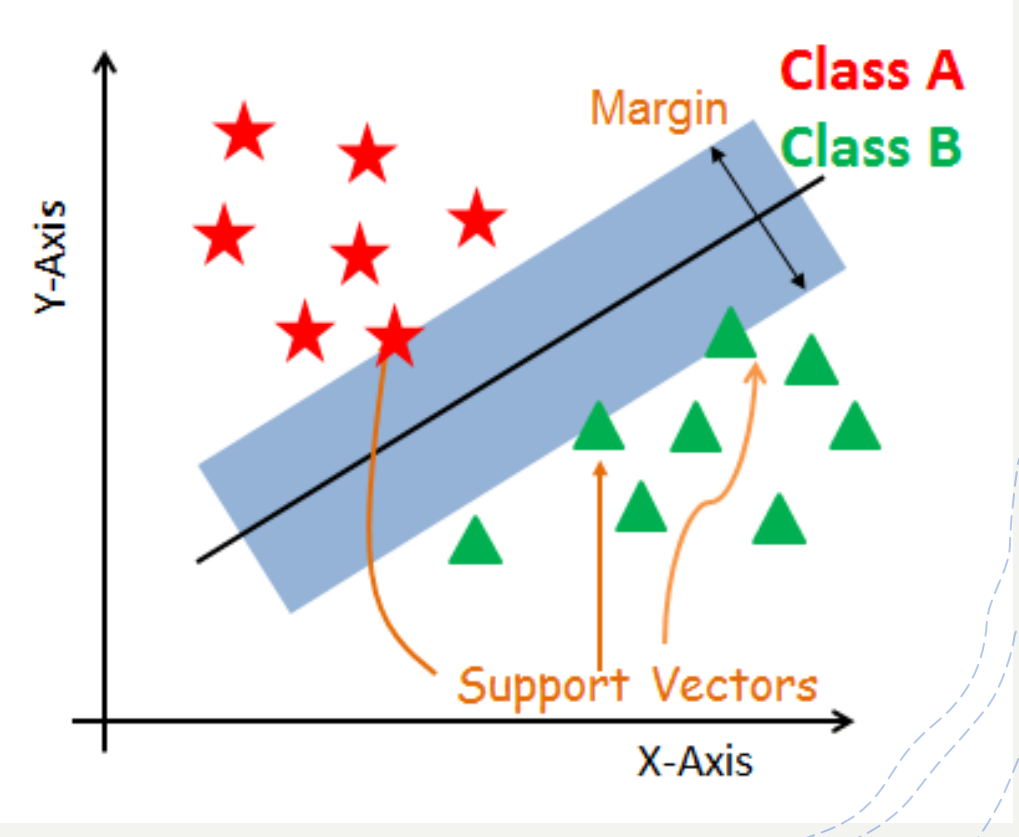




Predicting From the Paper Text

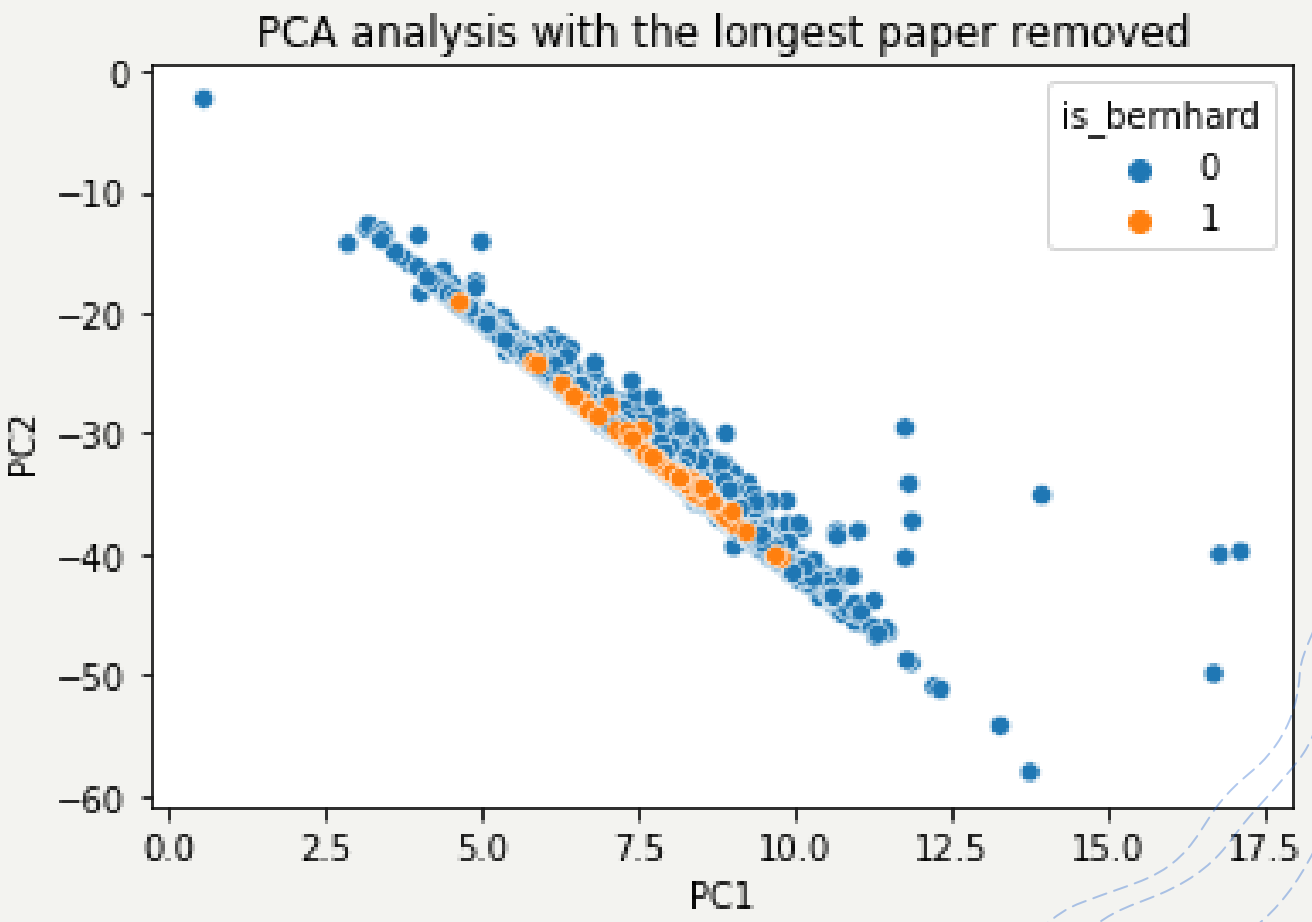
Predicting From the Paper Text

- Created new features (title length, paper length, avg word length)
- Models tested were logistic regression, LinearSVD, and random forest
- A “manual” grid-search was used to tune hyperparameters
- F2 score was used as the scoring metric
  - F2-Measure = (5 * Precision * Recall) / (4 * Precision + Recall)
  - More weight on minimizing false negatives than F1-score



PCA Analysis

- TFIDF vectorizer created almost 200,000 features
- A way to do topic modeling?
- Any more than 1000 components was impossible
- Ultimately did not help

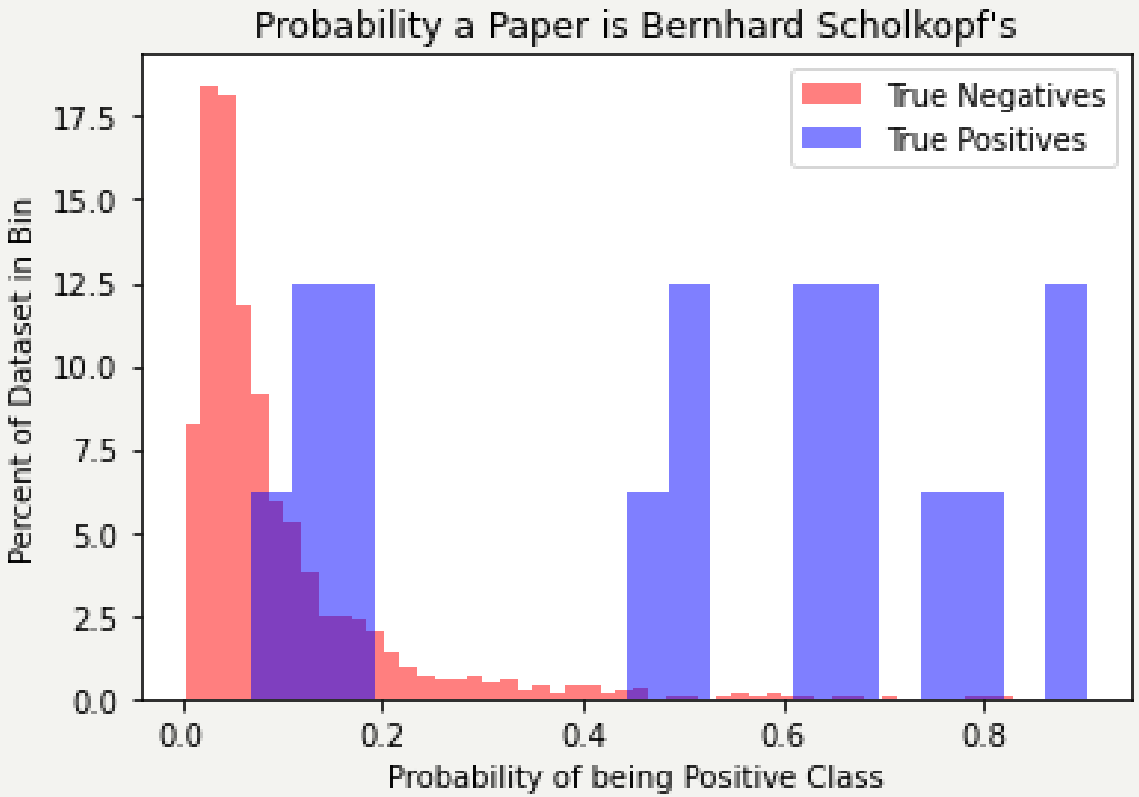


Predicting From the Paper Text

+ LinearSVD model gave best results
+ No inherent way to predict probabilities

	LogReg_C_value	SVC_C_value	Sampler	F2_score_train_avg	F2_score_test_avg	Model
4	NaN	0.1	None	0.839164	0.462798	SVC
7	NaN	0.05	ros	0.837961	0.461768	SVC
3	1.0	NaN	ros	0.874637	0.453942	LogisticRegression
5	NaN	0.1	ros	0.917847	0.445585	SVC
6	NaN	0.05	None	0.744874	0.434137	SVC
5	0.5	NaN	ros	0.801624	0.432396	LogisticRegression
2	1.0	NaN	None	0.804469	0.422821	LogisticRegression
1	5.0	NaN	ros	0.963199	0.419485	LogisticRegression

Predicting From the Paper Text

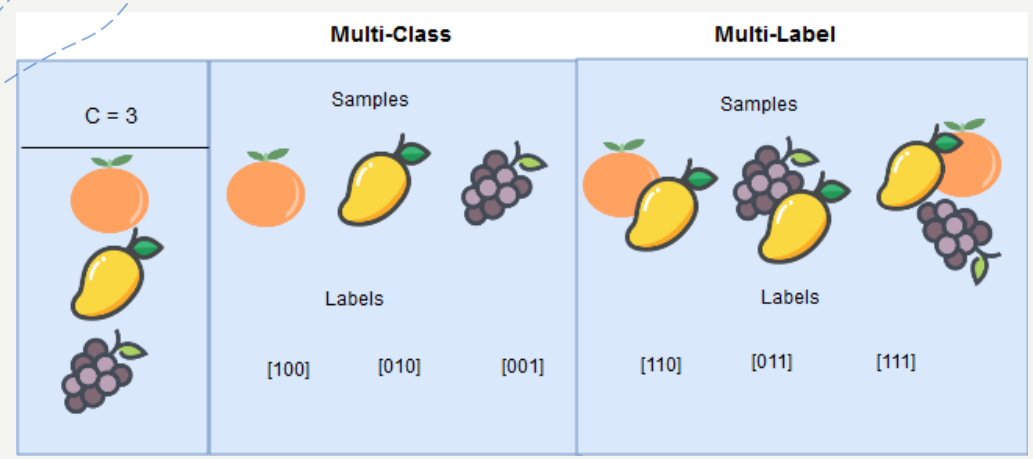


Multiple Authors

+

Multi-label Classification

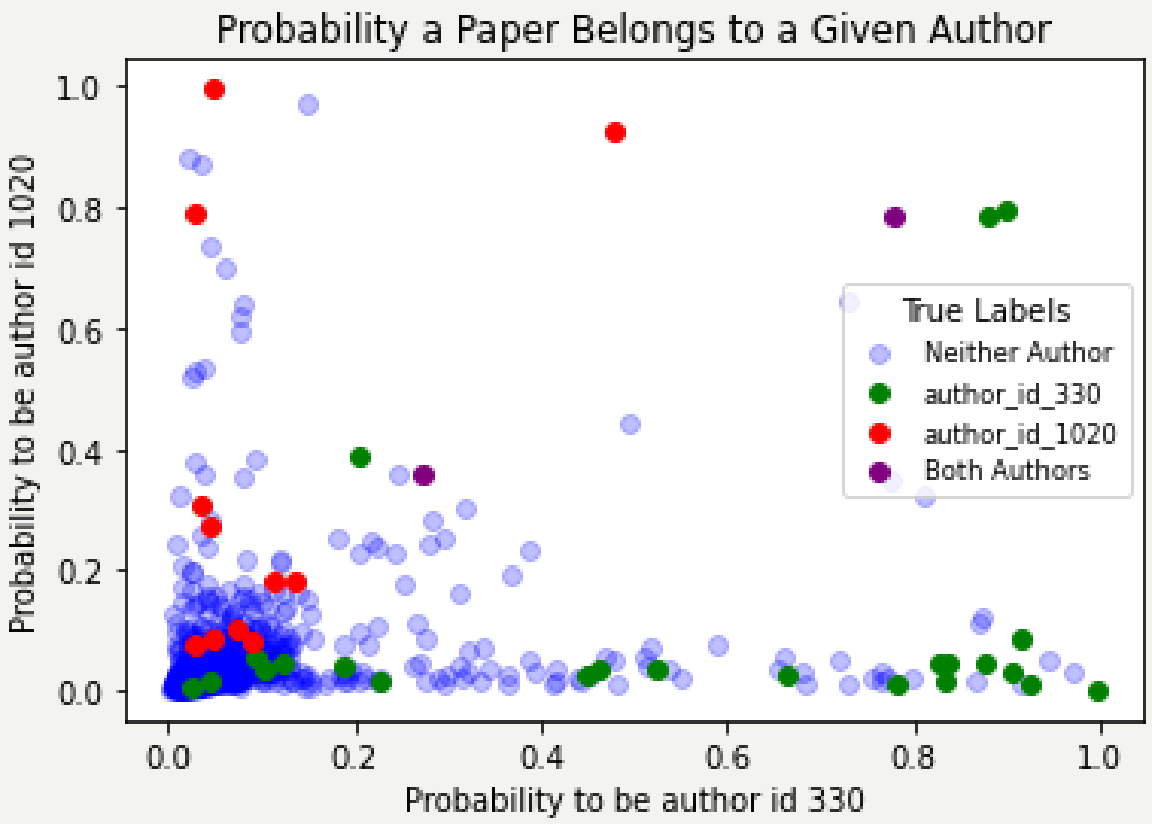
Multiple Authors



+ Papers can have multiple authors!
+ Used sklearn's MultiOutputClassifier
+ Had to use "next best" model because predicting probabilities was important!

Multiple Authors

- Initial test used two authors that co-wrote many papers
- Tested sklearn's MultiOutputClassifier and ClassifierChain
  - Both strategies were almost identical
- Why did ten authors provide the best results and 5 the worst?



	Num_authors_tested	F1_score_train_avg	F1_score_test_avg
1	10	0.790582	0.464970
2	15	0.799741	0.447966
0	5	0.775275	0.427062

Conclusion

- Final model not good enough to firmly predict, but much better than random
  - (F2 scores ~0.46)
- Largely achieved separation from the bulk of papers
- Could be useful with a human to help it
- Other vectorizers that retain word-order information could help
- Knowing primary authorship would help