Classifying Federal Register Documents by Type

Mark Febrizio

June 22, 2022

DATS 6103 – Summer 2022

What is the Federal Register?

- Daily journal of the U.S. government
- Print and online
- 4 main sections
- 4 document types:
  - Notices
  - Proposed Rules
  - Final Rules
  - Presidential Documents

NATIONAL ARCHIVES

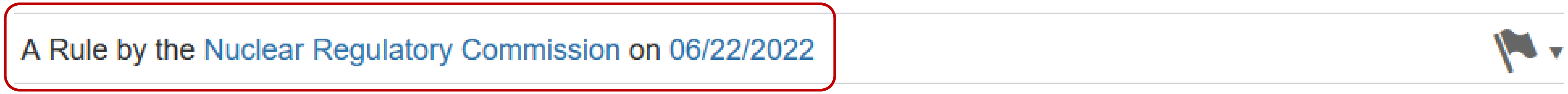
FEDERAL REGISTER

The Daily Journal of the United States Government

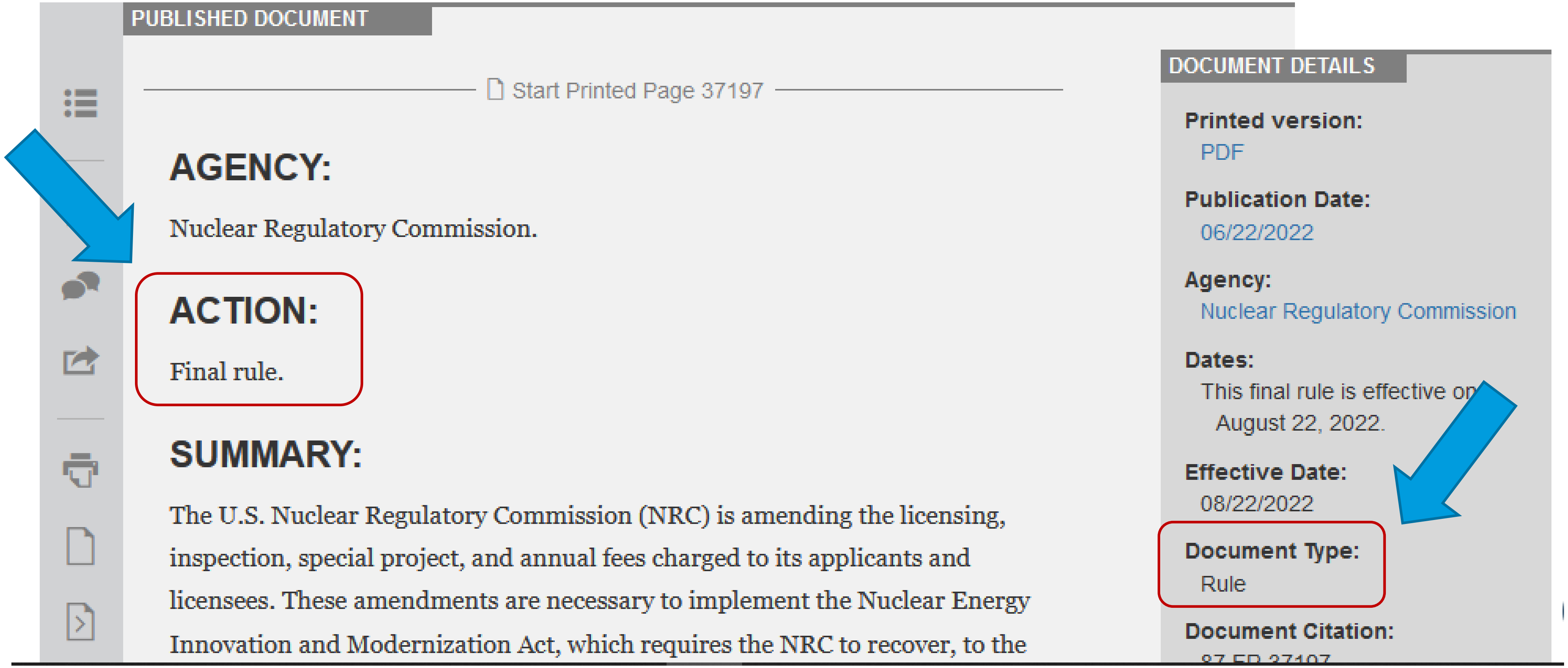
The Problem: Uncategorized Documents

- Particularly an issue in the 1990s data
- Available data from 1994–1999:
  - 201,591 total documents
  - 166,031 categorized as 1 of 4 main types
  - 32,468 lack “type” labels
- Problem: Severe undercounting for analysis
- Solution: use labeled documents to build classifier for document type

Revision of Fee Schedules; Fee Recovery for Fiscal Year 2022



A Rule by the Nuclear Regulatory Commission on 06/22/2022



PUBLISHED DOCUMENT

Start Printed Page 37197

AGENCY:

Nuclear Regulatory Commission.

ACTION:

Final rule.

SUMMARY:

The U.S. Nuclear Regulatory Commission (NRC) is amending the licensing, inspection, special project, and annual fees charged to its applicants and licensees. These amendments are necessary to implement the Nuclear Energy Innovation and Modernization Act, which requires the NRC to recover, to the

DOCUMENT DETAILS

Printed version:
PDF

Publication Date:
06/22/2022

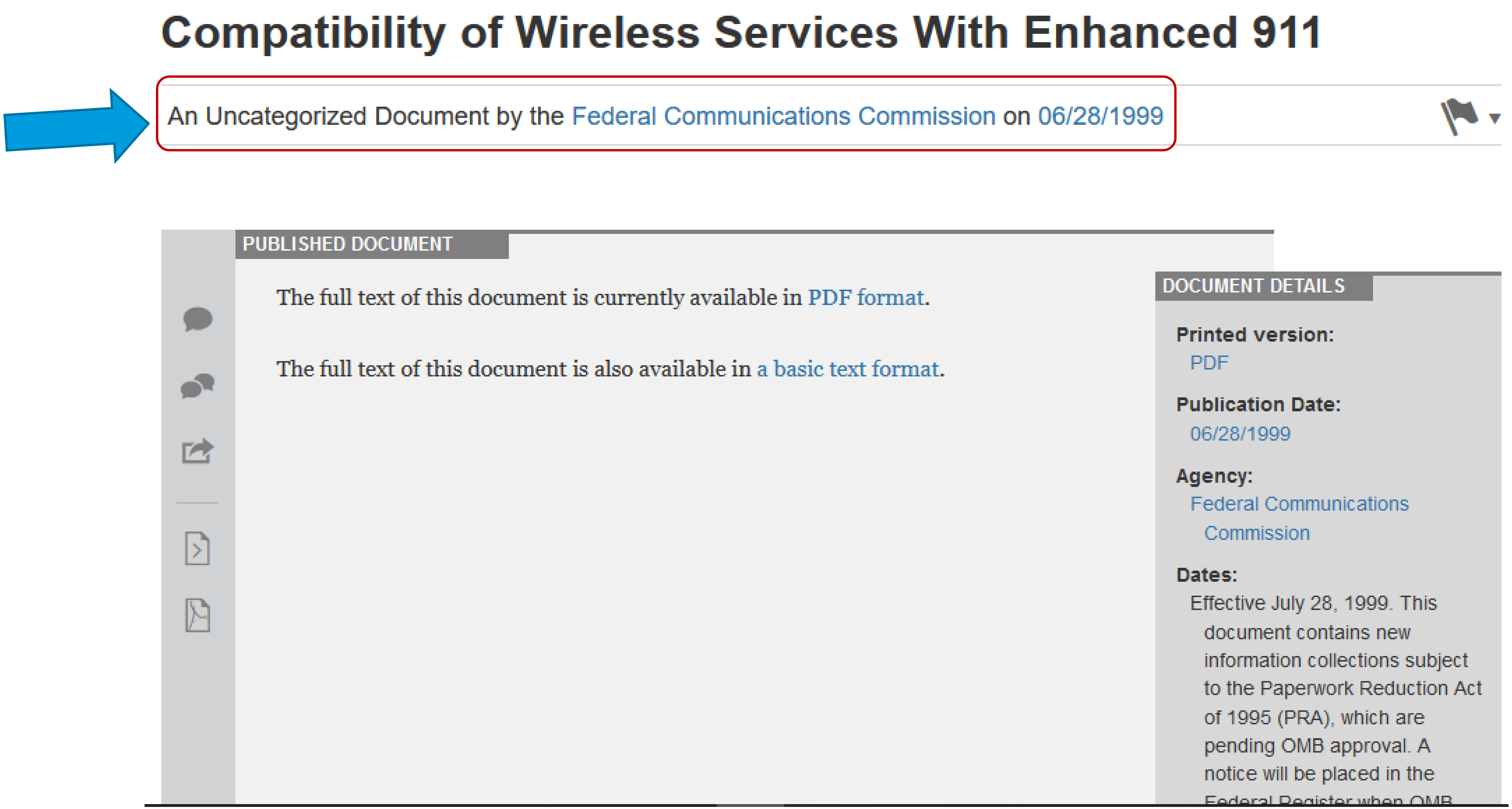
Agency:
Nuclear Regulatory Commission

Dates:
This final rule is effective on August 22, 2022.

Effective Date:
08/22/2022

Document Type:
Rule

Document Citation:
87 FR 37197



Compatibility of Wireless Services With Enhanced 911

An Uncategorized Document by the Federal Communications Commission on 06/28/1999

PUBLISHED DOCUMENT

The full text of this document is currently available in PDF format.

The full text of this document is also available in a basic text format.

DOCUMENT DETAILS

Printed version:
PDF

Publication Date:
06/28/1999

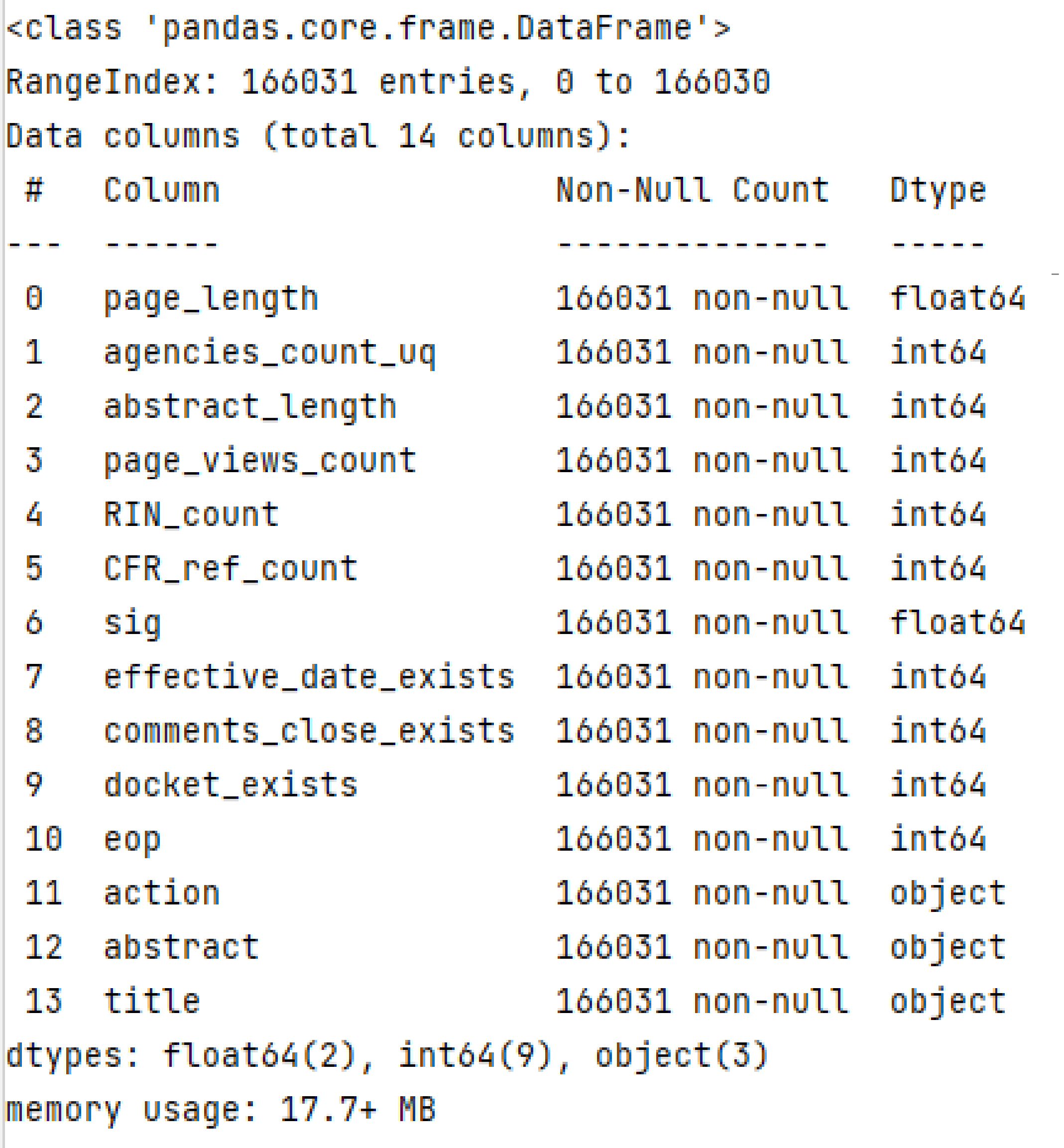
Agency:
Federal Communications Commission

Dates:
Effective July 28, 1999. This document contains new information collections subject to the Paperwork Reduction Act of 1995 (PRA), which are pending OMB approval. A notice will be placed in the Federal Register when OMB

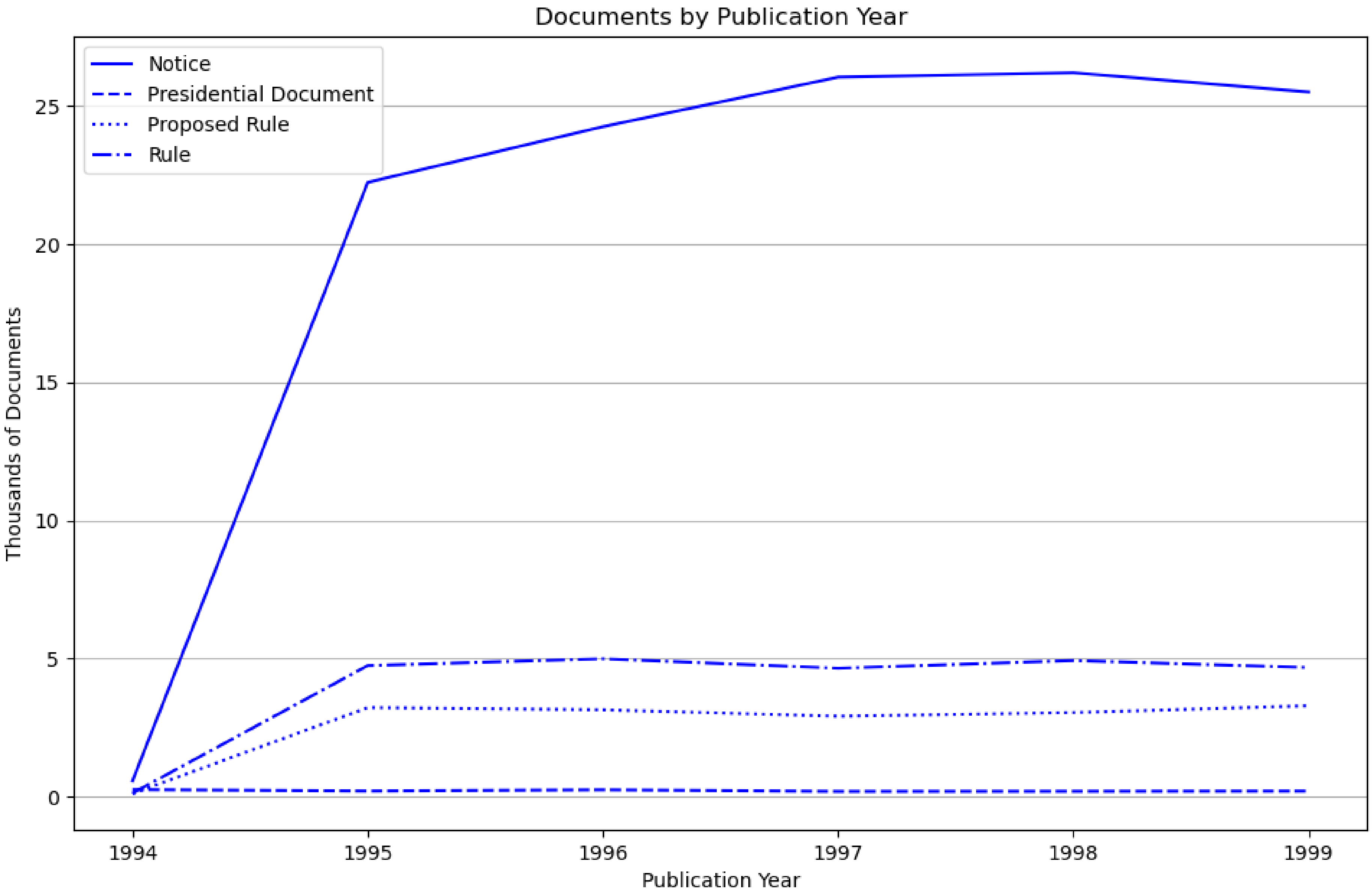
Data

Processed dataset:

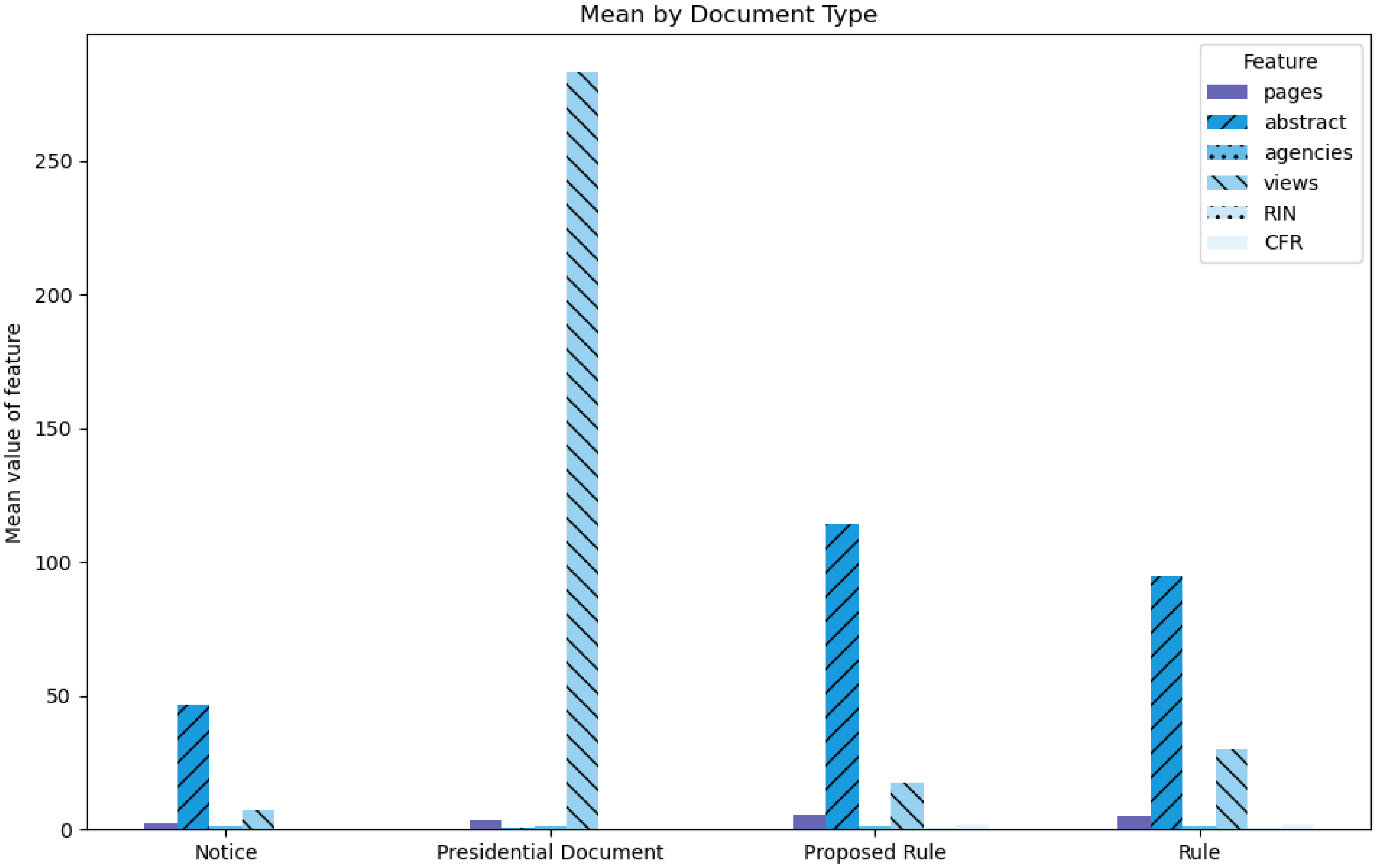
- 166,031 documents
- 6 numeric variables
- 5 categorical variables
- 3 text variables



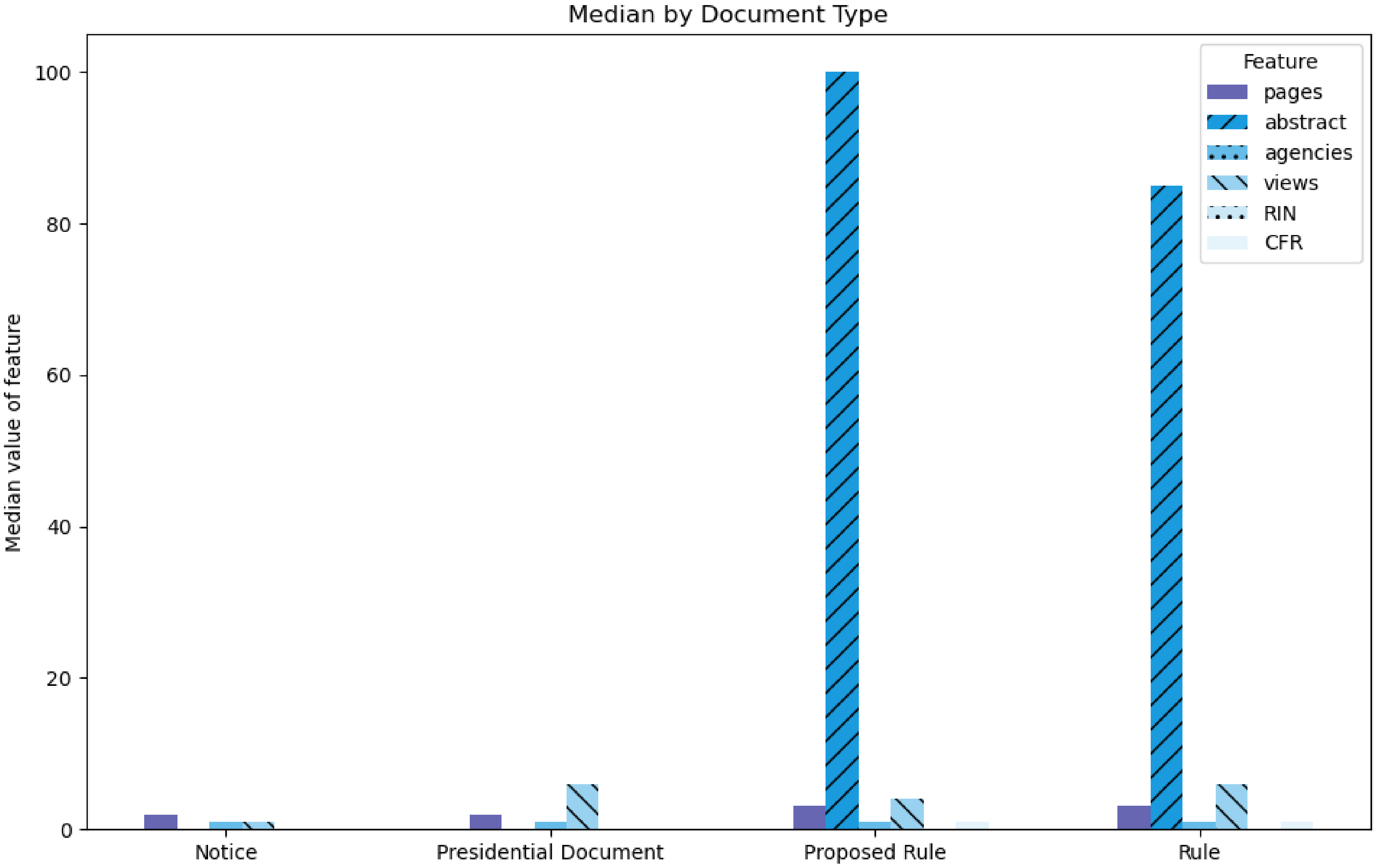
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 166031 entries, 0 to 166030
Data columns (total 14 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   page_length           166031 non-null  float64
 1   agencies_count_uq     166031 non-null  int64
 2   abstract_length       166031 non-null  int64
 3   page_views_count      166031 non-null  int64
 4   RIN_count             166031 non-null  int64
 5   CFR_ref_count         166031 non-null  int64
 6   sig                   166031 non-null  float64
 7   effective_date_exists 166031 non-null  int64
 8   comments_close_exists 166031 non-null  int64
 9   docket_exists         166031 non-null  int64
 10  eop                   166031 non-null  int64
 11  action                166031 non-null  object
 12  abstract              166031 non-null  object
 13  title                 166031 non-null  object
dtypes: float64(2), int64(9), object(3)
memory usage: 17.7+ MB
```



Source: Federal Register API and authors' calculations.



Source: Federal Register API and authors' calculations.



Source: Federal Register API and authors' calculations.

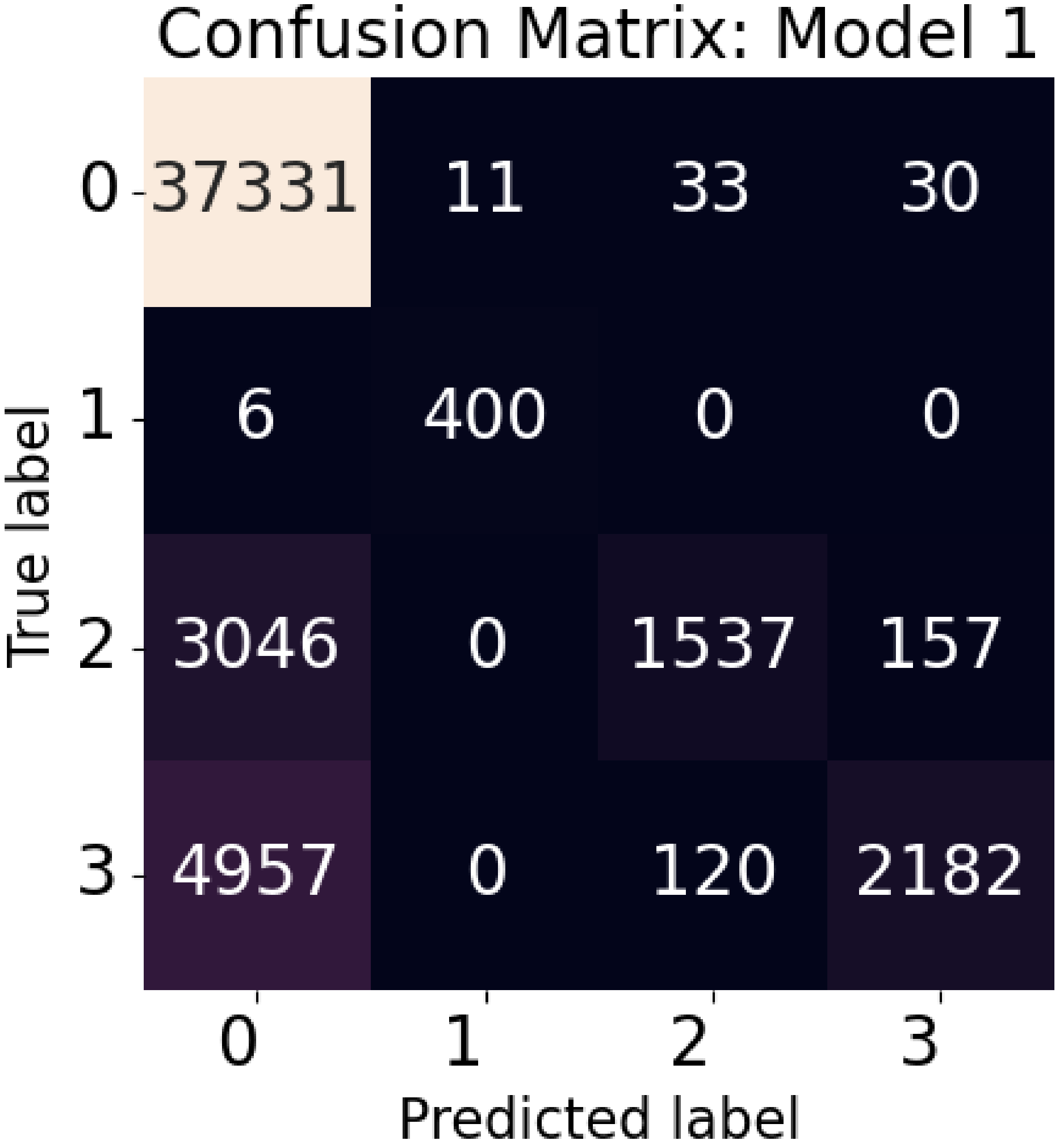
Preprocessing

- Data cleaning
- Imputing missing values
- Create/extract new variables from document metadata
- Numeric transformer (min-max scaler)
- Categorical transformer (one-hot encoder)
- Label encoder for target
- Text feature extraction (tf-idf vectorizer)
- Train-test split: 70%-30% (116,221 vs. 49,810)

Modeling

1. Complement Naïve Bayes (categorical and numeric)
2. AdaBoost (categorical and numeric)
   - ➢ 1000 Complement NB estimators
3. Voting Classifier (categorical and numeric)
   - ➢ Complement NB
   - ➢ Logistic Regression (balanced weights)
   - ➢ KNN → grid search for k == [5, 99, 341]
4. Complement NB (text)
   - ➢ tf–idf: term-frequency * inverse document-frequency

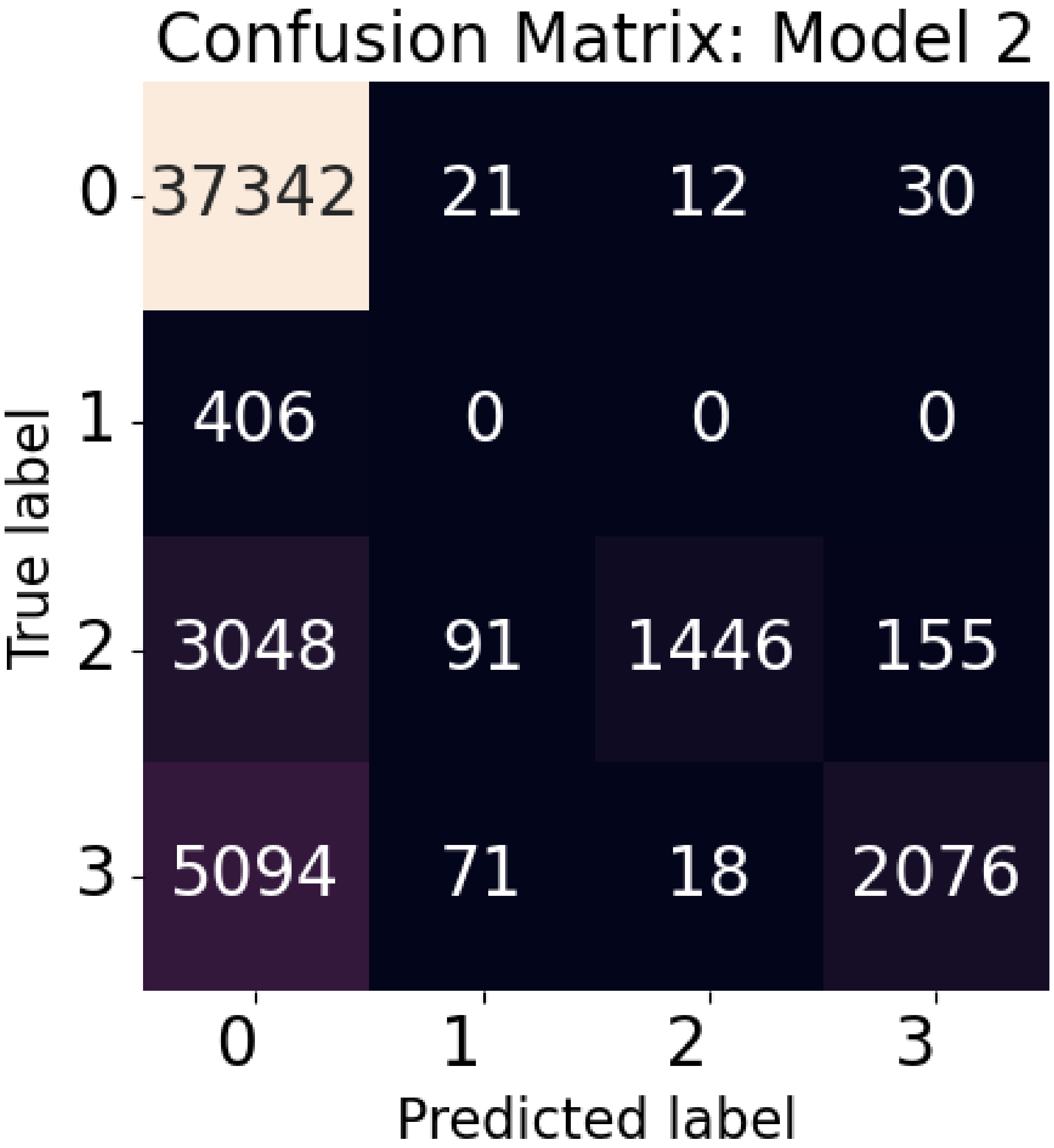
Complement NB (categorical/numeric)



Confusion Matrix: Model 1

True label \ Predicted label	0	1	2	3
0	37331	11	33	30
1	6	400	0	0
2	3046	0	1537	157
3	4957	0	120	2182

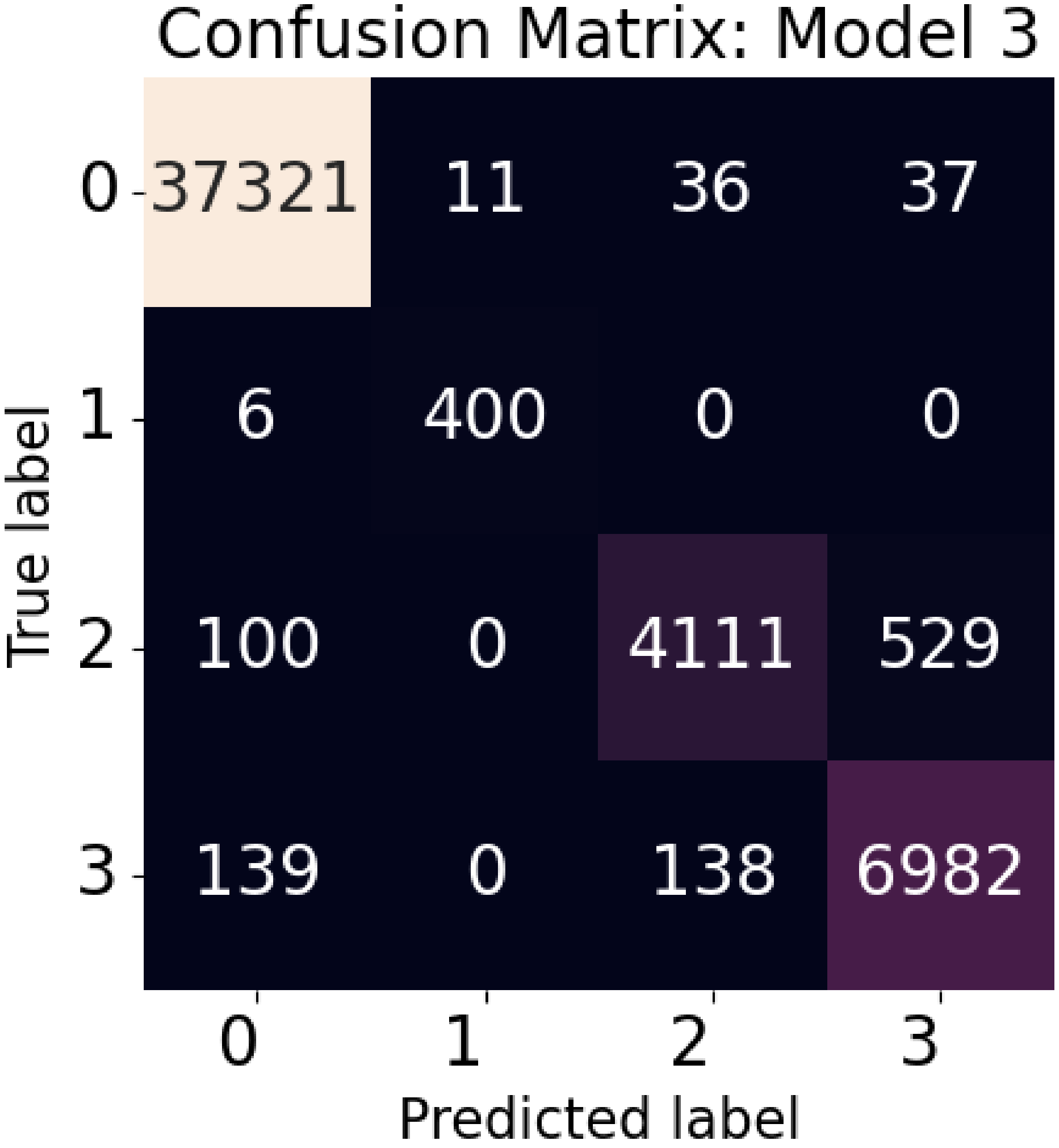
AdaBoost
(1000 NB estimators)



Confusion Matrix: Model 2

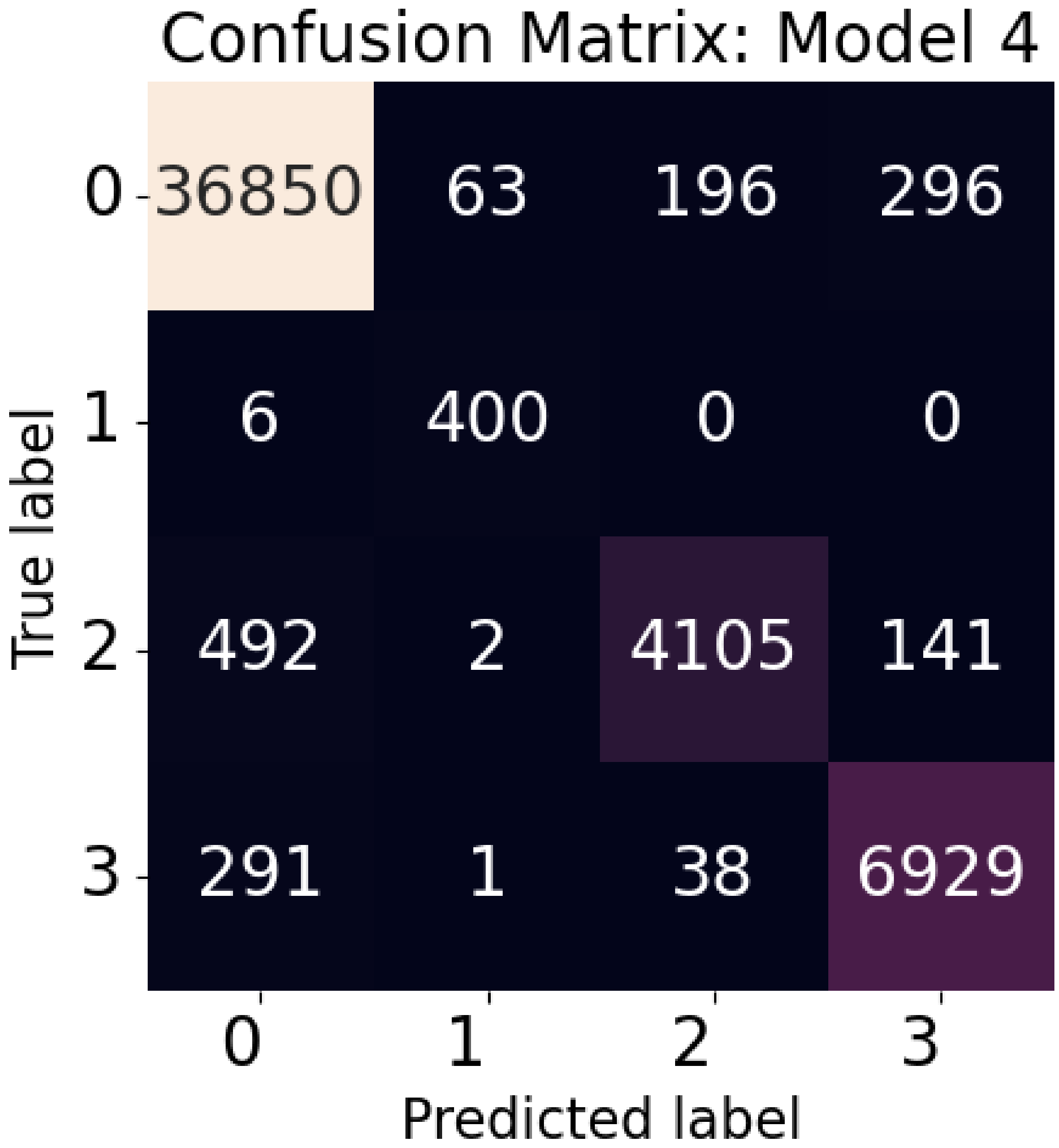
True label \ Predicted label	0	1	2	3
0	37342	21	12	30
1	406	0	0	0
2	3048	91	1446	155
3	5094	71	18	2076

Voting Classifier
(NB, Logit, KNN)



True label \ Predicted label	0	1	2	3
0	37321	11	36	37
1	6	400	0	0
2	100	0	4111	529
3	139	0	138	6982

Complement NB
(tf-idf vectors)



Confusion Matrix: Model 4

	0	1	2	3
0	36850	63	196	296
1	6	400	0	0
2	492	2	4105	141
3	291	1	38	6929

Model	*Classifiers*	*Features*	*Accuracy*	*F1-Score*
1	NB	5 categorical 6 numeric	0.832	0.797
2	Boost 1000 * NB	5 categorical 6 numeric	0.820	0.781
3	Hard voting NB, Logit, KNN	5 categorical 6 numeric	0.980	0.979
4	NB	1 text	0.969	0.969

Improvements

- Consider different hyper-parameters for AdaBoost (1000x too high?)
- Integrate text features with categorical/numeric features
- Analyze full text of documents → tf-idf vectorizer
- Clean up my code…
- Any other suggestions?

Thank you for listening!

Questions?