UTILIZING A TRANSPARENCY-DRIVEN ENVIRONMENT TOWARD TRUSTED AUTOMATIC GENRE CLASSIFICATION

A Case Study in Journalism History

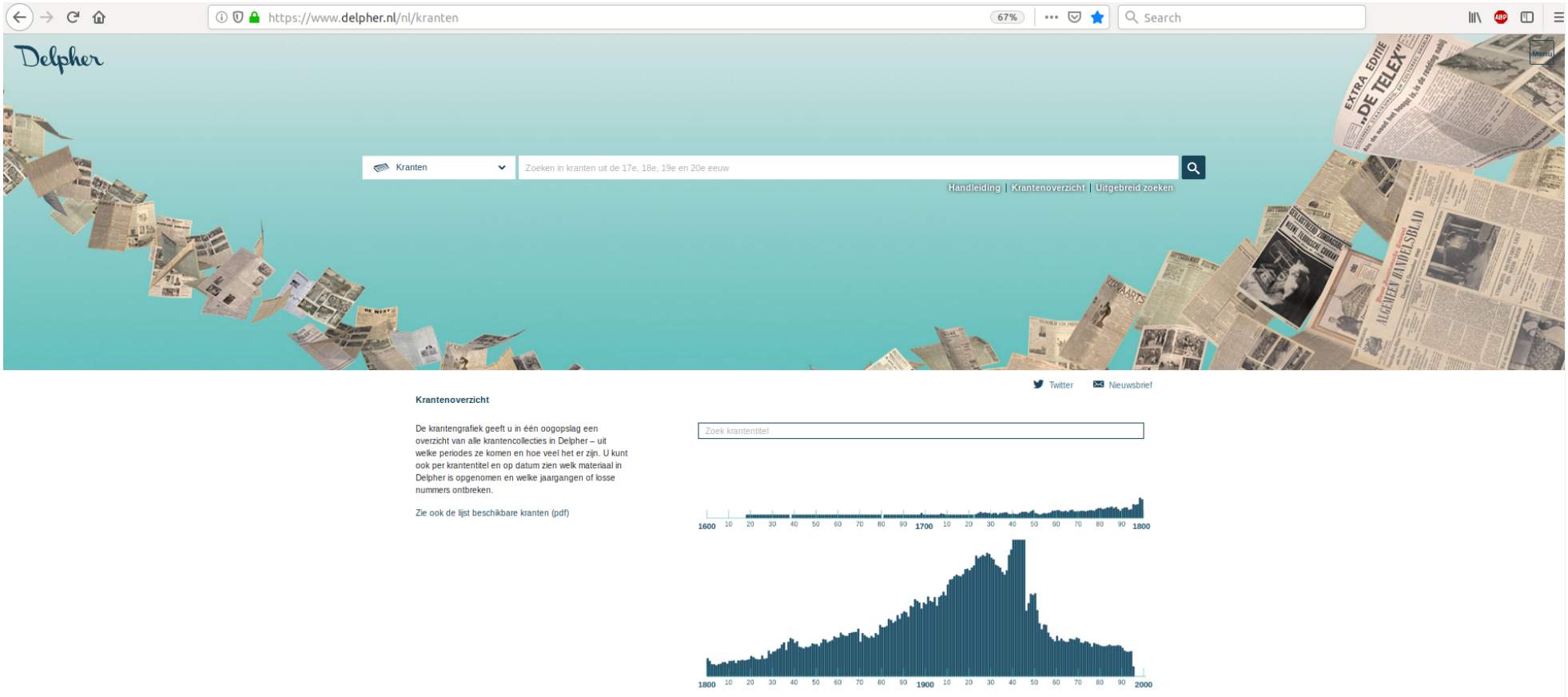
Aysenur Bilgin [1], **Erik Tjong Kim Sang**[2], Kim Smeenk[3], Laura Hollink[1], Jacco van Ossenbrugen[1], Frank Harbers[3], Marcel Broersma[3]

November 1, 2018

[1]CWI, [2]Netherlands eScience Center, [3]University of Groningen

aysenur.bilgin@cwi.nl

Newspaper archives



Newspaper archive Delpher.nl offers millions of newspaper pages

Motivation

- Study genre distribution in Dutch newspaper text
- However, digital archives do not contain genre labelling
- Due to big corpus size, we need machine learning

BUT... Can we trust the results of machine learning?

Transparency

- We need to be able to show what machine learning is doing
- Both for specialists and non-specialists (domain experts)
- We want transparency in all aspects of the labelling process

Transparent Workflow

1

Data pre-processing

Upload formatted data, choose representation type and pre-processing steps.

2

Optimal model selection

Optional step to detect the optimal machine learning algorithm and its hyper-parameters.

3

Configuration and training

Use the optimal choices from the previous step if applied or configure the algorithm and train.

4

Individual pipeline analysis

Using several metrics, receive insights about the performance of the pipeline (e.g. accuracy, precision, etc.) and its global interpretations if available.

5

Pipeline comparison

Based on test data, compare collective predictions of the pipelines in tabular views and get insights about the pipelines using local explanations.

6

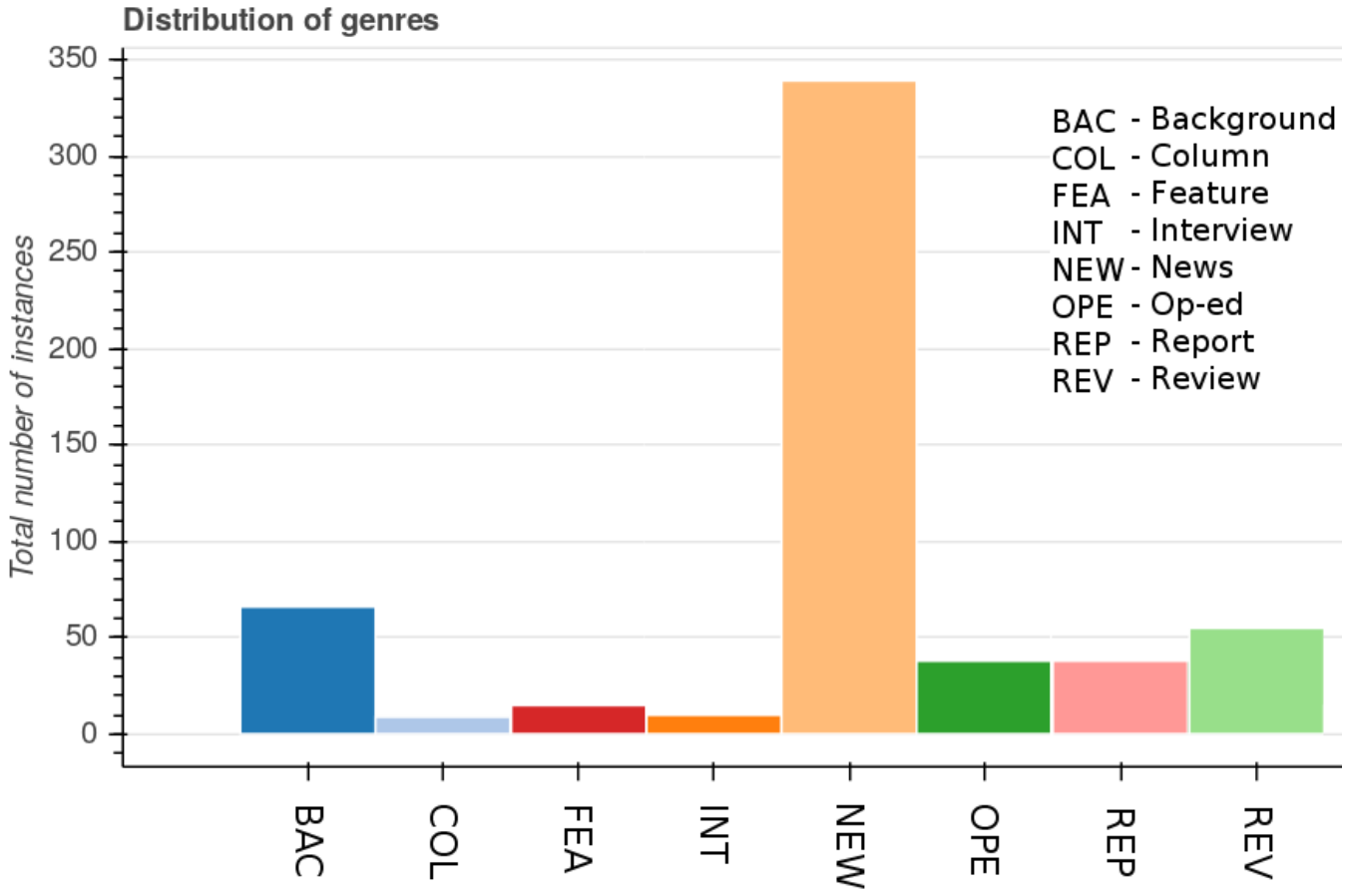
Domain hypothesis testing

Choose the best suited pipeline for the research hypothesis and apply it to large-scale real-world data.

Case Study

- We study distribution of genre (e.g. column, interview, review, etc.) over time
- We use supervised machine learning algorithms for genre labelling
- We can analyze the impact of different variants for:
  - Data preparation
  - Data pre-processing
  - Machine learning algorithms
  - Algorithm configurations
- We examine visualizations for fair comparison

Data preparation



We need a balanced dataset to ensure a fair model

Data pre-processing

We consider two types of document representations:

- Bag-of-words (i.e. words and their frequencies)
- Linguistic features (e.g. document length, number of pronouns, etc.)

Machine learning algorithms and configurations

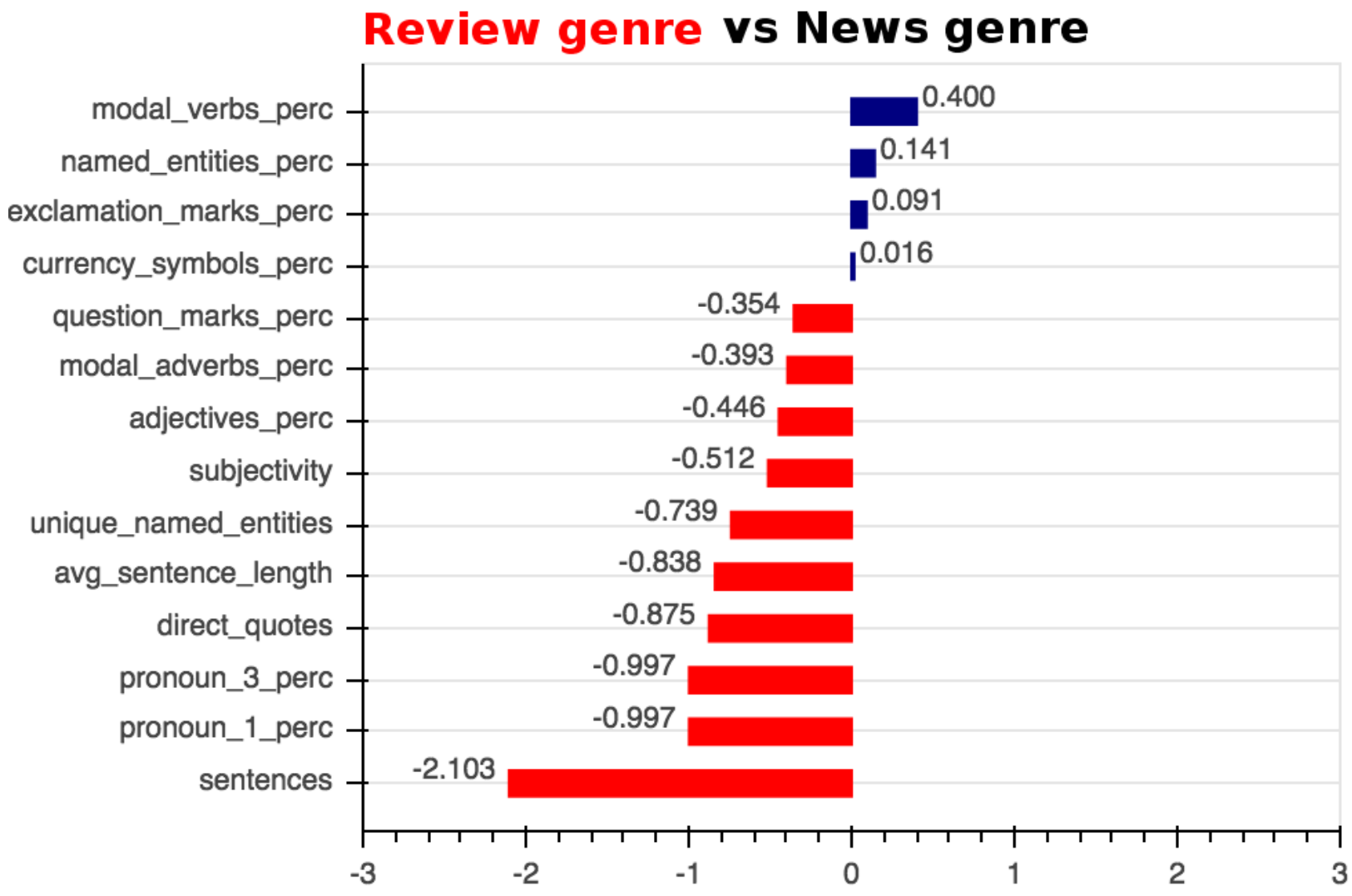
We evaluate three machine learning algorithms:

- Naive Bayes
- Support Vector Classifier
- Random Forest

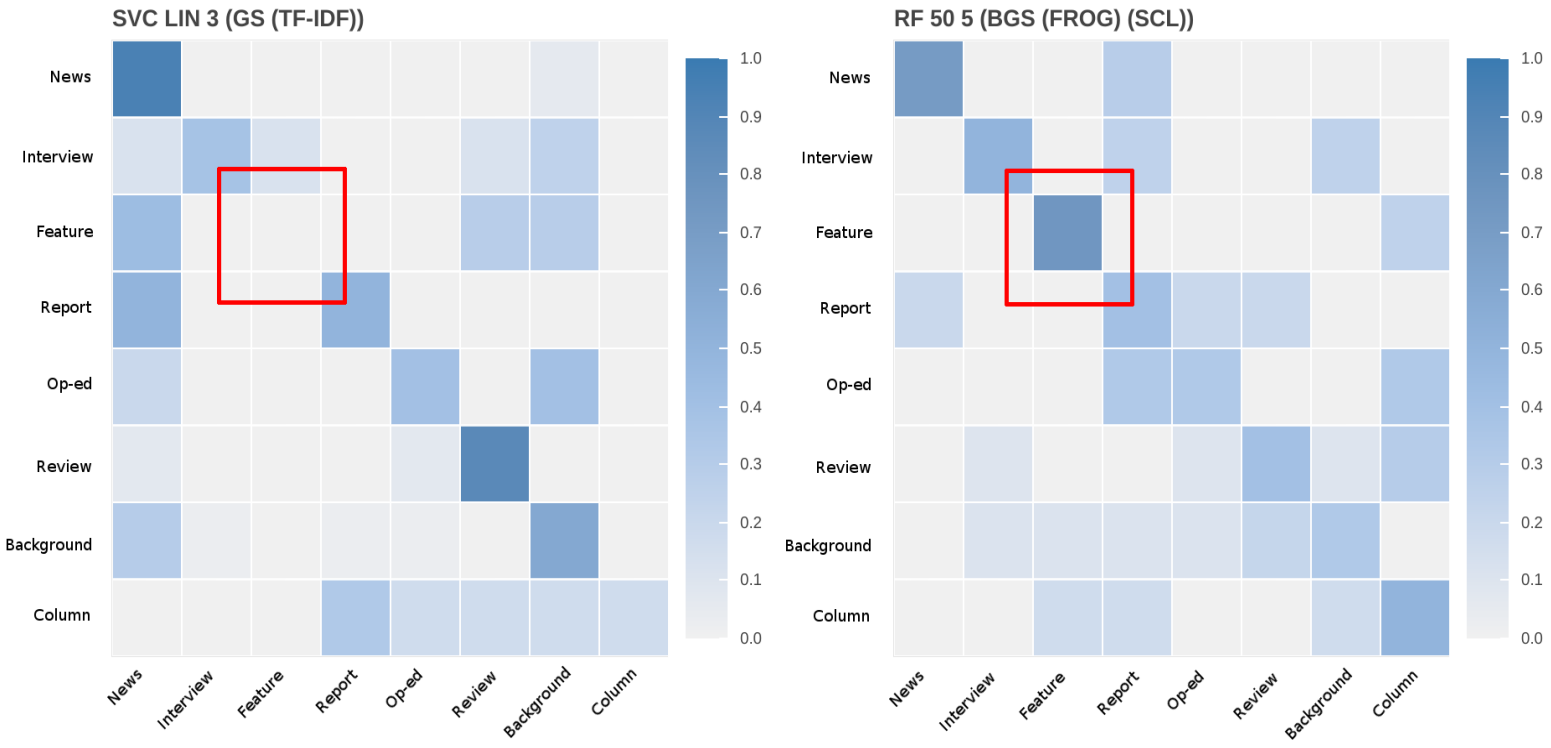
The algorithms have several configurable parameters that have impact on their behaviour.

We use automated techniques to select optimal values of these parameters.

Pipeline analysis - Feature importance ranking



Pipeline comparison using confusion matrices



Confusion matrices for two machine learning pipelines:

SVC accuracy is 70% (left) and RF accuracy is 41% (right)

Pipeline comparison using article-based view

Article #	Article text	Mutually Agreeing Pipelines	Prevailing Genre (prediction)	True Genre
1800	C.P.N . jaren van geharrewar is het er he ff 6 `` S '° tt : e van gekomen…	RF 1000 5 (GS (FROG) (SCL)) RF 50 5 (BGS (FROG) (SCL)) RF 1000 2 (CGS (FROG) (SCL)) SVC LIN 3 (CGS (TF-IDF) (SWR))	News	Op-ed
55	Produktschap gaat praten over vleesprijzen (Van onze Haagse redacteur)…	RF 1000 5 (GS (FROG) (SCL)) RF 50 5 (BGS (FROG) (SCL)) RF 1000 2 (CGS (FROG) (SCL)) SVC LIN 3 (CGS (TF-IDF) (SWR)) SVC LIN 3 (CGS (TF-IDF))	News	News
77	Artis schenkt Blijdorp orang oetang De directie van diergaarde Blijdorp…	RF 1000 5 (GS (FROG) (SCL)) RF 50 5 (BGS (FROG) (SCL)) RF 1000 2 (CGS (FROG) (SCL)) SVC LIN 3 (CGS (TF-IDF) (SWR)) SVC LIN 3 (CGS (TF-IDF))	News	News
111	Ingestudeerde spelpatronen komen niet uit de verf tegen Oostenrijk Door…	RF 1000 5 (GS (FROG) (SCL)) RF 50 5 (BGS (FROG) (SCL)) RF 1000 2 (CGS (FROG) (SCL)) SVC LIN 3 (CGS (TF-IDF) (SWR)) SVC LIN 3 (CGS (TF-IDF))	Background	Background
117	Simpele zege volleyballers op Portugal SOLBJERG , 2 april — Het…	RF 1000 5 (GS (FROG) (SCL)) RF 50 5 (BGS (FROG) (SCL)) RF 1000 2 (CGS (FROG) (SCL)) SVC LIN 3 (CGS (TF-IDF) (SWR)) SVC LIN 3 (CGS (TF-IDF))	News	News

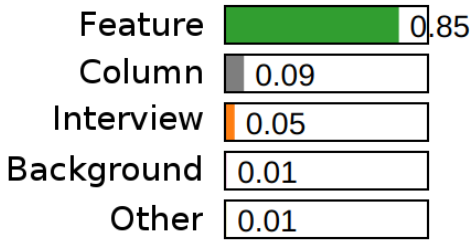
Pipeline comparison using set-based view

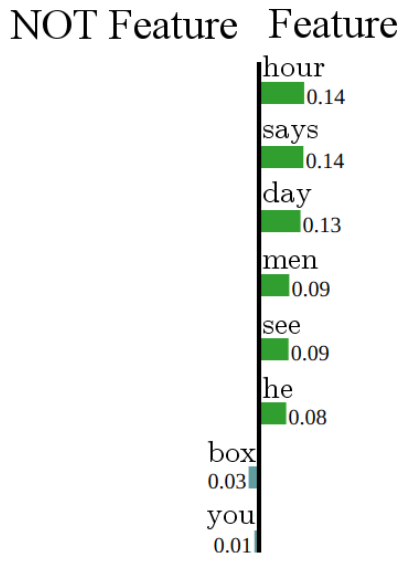
Mutually Agreeing Pipelines	Predicted Genre	Number of Articles	Articles
SVC LIN 2 (BGS (TF-IDF) (SWR)) SVC RBF 10 (BGS (TF-IDF))	Report	208	[Article 13] [Article 17] [Article 26] [Article 46] [Article 49] [Article 56] [Article 58] [Article 63] [Article 85] [Article 105] [Article 115] [Article 128] [Article 142] [Article 149] [Article 150] [Article 152] [Article 164] [Article 173] [Article 175] [Article 194] [Article 214] [Article 225] [Article 227] [Article 246] [Article 269] [Article 295] [Article 323] [Article 334] [Article 350] [Article 366] [Article 371] [Article 372] [Article 386] [Article 391] [Article 414] [Article 426] [Article 444] [Article 457] [Article 484] [Article 485] [Article 490] [Article 519] [Article 524] [Article 531] [Article 557] [Article 559] [Article 560] [Article 562] [Article 568] [Article 595] [Article 611] [Article 617] [Article 619] [Article 628] [Article 634] [Article 642] [Article 658] [Article 669] [Article 675] [Article 681] [Article 699] [Article 727] [Article 735] [Article 746] [Article 751] [Article 764] [Article 769] [Article 774] [Article 775] [Article 787] [Article 792] [Article 793] [Article 794] [Article 805] [Article 807] [Article 816] [Article 827] [Article 852] [Article 877] [Article 880] [Article 882] [Article 888] [Article 889] [Article 895] [Article 896] [Article 904] [Article 911] [Article 940] [Article 954] [Article 957] [Article 961] [Article 965] [Article 975] [Article 1006] [Article 1010] [Article 1011] [Article 1019] [Article 1042] [Article 1047] [Article 1079] [Article 1085] [Article 1121] [Article 1124] [Article 1134] [Article 1148] [Article 1152] [Article 1160] [Article 1161] [Article 1168] [Article 1172] [Article 1179] [Article 1184] [Article 1215] [Article 1217] [Article 1218] [Article 1266] [Article 1285] [Article 1288] [Article 1314] [Article 1339] [Article 1353] [Article 1395] [Article 1404] [Article 1407] [Article 1416] [Article 1437] [Article 1447] [Article 1484] [Article 1513] [Article 1514] [Article 1515] [Article 1526] [Article 1536] [Article 1545] [Article 1558] [Article 1564] [Article 1581] [Article 1619] [Article 1679] [Article 1686] [Article 1692] [Article 1723] [Article 1747] [Article 1754] [Article 1755] [Article 1757] [Article 1783] [Article 1794] [Article 1802] [Article 1808] [Article 1821] [Article 1825] [Article 1836] [Article 1838] [Article 1861] [Article 1862] [Article 1863] [Article 1869] [Article 1871] [Article 1884] [Article 1887] [Article 1891] [Article 1896] [Article 1905] [Article 1910] [Article 1915] [Article 1974] [Article 1986] [Article 2000] [Article 2008] [Article 2010] [Article 2027] [Article 2042] [Article 2046] [Article 2056] [Article 2096] [Article 2100] [Article 2107] [Article 2117] [Article 2148] [Article 2165] [Article 2183] [Article 2200] [Article 2213] [Article 2231] [Article 2232] [Article 2263] [Article 2268] [Article 2300] [Article 2311] [Article 2317] [Article 2318] [Article 2343] [Article 2351] [Article 2353] [Article 2360] [Article 2367] [Article 2386] [Article 2454] [Article 2456] [Article 2458] [Article 2467] [Article 2482] [Article 2489] [Article 2491] [Article 2493] [Article 2496] [Article 2502]
RF 1000 5 (GS (FROG) (SCL)) RF 1000 2 (CGS (FROG) (SCL)) RF 50 5 (BGS (FROG) (SCL)) SVC LIN 3 (CGS (TF-IDF)) SVC LIN 3 (GS (TF-IDF) (SWR)) SVC LIN 3 (GS (TF-IDF)) SVC LIN 3 (CGS (TF-IDF) (SWR))	News	142	[Article 32] [Article 46] [Article 49] [Article 51] [Article 115] [Article 128] [Article 142] [Article 149] [Article 150] [Article 152] [Article 194] [Article 214] [Article 224] [Article 225] [Article 235] [Article 245] [Article 254] [Article 266] [Article 286] [Article 310] [Article 323] [Article 350] [Article 366] [Article 371] [Article 372] [Article 375] [Article 391] [Article 414] [Article 426] [Article 444] [Article 457] [Article 484] [Article 485] [Article 490] [Article 557] [Article 559] [Article 560] [Article 562] [Article 611] [Article 636] [Article 642] [Article 658] [Article 706] [Article 714] [Article 727] [Article 735] [Article 772] [Article 775] [Article 793] [Article 807] [Article 837] [Article 852] [Article 877] [Article 888] [Article 895] [Article 896] [Article 904] [Article 911] [Article 940] [Article 947] [Article 954] [Article 957] [Article 961] [Article 965] [Article 1006] [Article 1019] [Article 1047] [Article 1071] [Article 1121] [Article 1134] [Article 1152] [Article 1160] [Article 1172] [Article 1179] [Article 1215] [Article 1218] [Article 1258] [Article 1285] [Article 1337] [Article 1353] [Article 1404] [Article 1416] [Article 1447] [Article 1484] [Article 1499] [Article 1507] [Article 1513] [Article 1580] [Article 1584] [Article 1617] [Article 1619] [Article 1679] [Article 1686] [Article 1723] [Article 1736] [Article 1747] [Article 1757] [Article 1794] [Article 1802] [Article 1817] [Article 1818] [Article 1821] [Article 1825] [Article 1836] [Article 1849] [Article 1861] [Article 1862] [Article 1869] [Article 1871] [Article 1873] [Article 1877] [Article 1884] [Article 1887] [Article 1891] [Article 1896] [Article 1910] [Article 1915] [Article 1974] [Article 2000] [Article 2008] [Article 2042] [Article 2046] [Article 2056] [Article 2100] [Article 2107] [Article 2148] [Article 2162] [Article 2165] [Article 2174] [Article 2183] [Article 2263] [Article 2266] [Article 2268] [Article 2300] [Article 2311] [Article 2317] [Article 2351] [Article 2360] [Article 2404] [Article 2456] [Article 2482] [Article 2491]

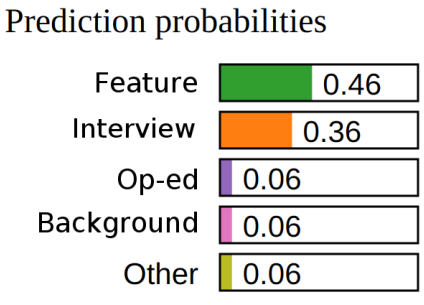
Pipeline comparison using explanation-based view

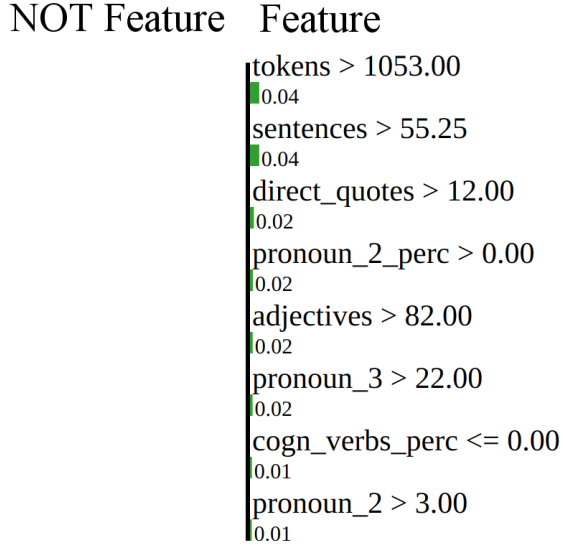
Article #	Pipeline	Predicted Genre	Text Explanation	Feature Explanation
1	RF 1000 2 (CGS (FROG) (SCL))	Feature	N/A	Click here for feature explanation
59	RF 1000 5 (GS (FROG) (SCL))	Feature	N/A	Click here for feature explanation
59	RF 1000 2 (CGS (FROG) (SCL))	Feature	N/A	Click here for feature explanation
59	RF 50 5 (BGS (FROG) (SCL))	Feature	N/A	Click here for feature explanation
66	RF 50 5 (BGS (FROG) (SCL))	Feature	N/A	Click here for feature explanation
80	RF 1000 5 (GS (FROG) (SCL))	Feature	N/A	Click here for feature explanation
80	RF 1000 2 (CGS (FROG) (SCL))	Feature	N/A	Click here for feature explanation
87	RF 1000 5 (GS (FROG) (SCL))	Feature	N/A	Click here for feature explanation
94	RF 1000 5 (GS (FROG) (SCL))	Feature	N/A	Click here for feature explanation

Local explanations



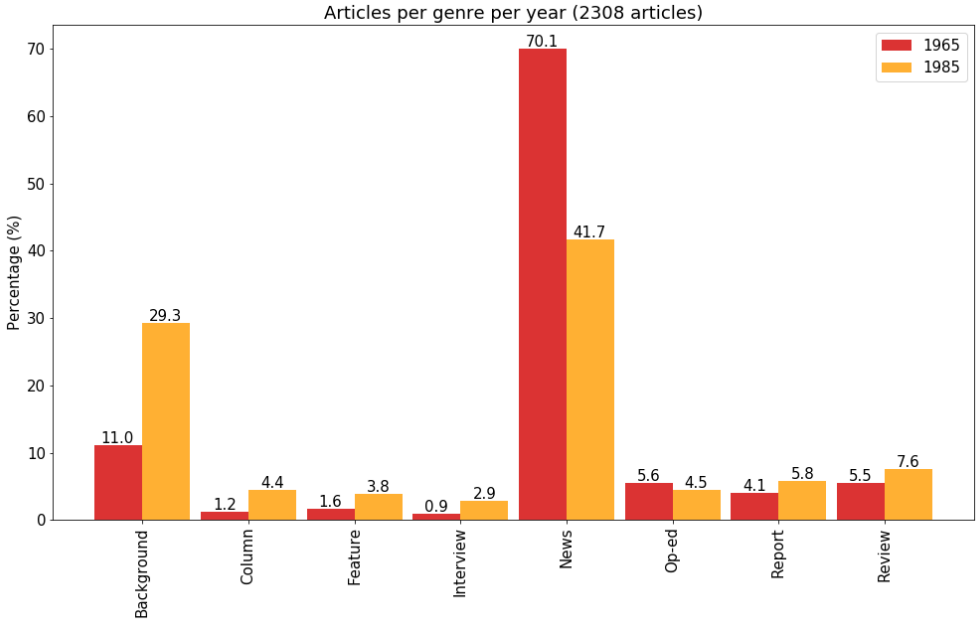




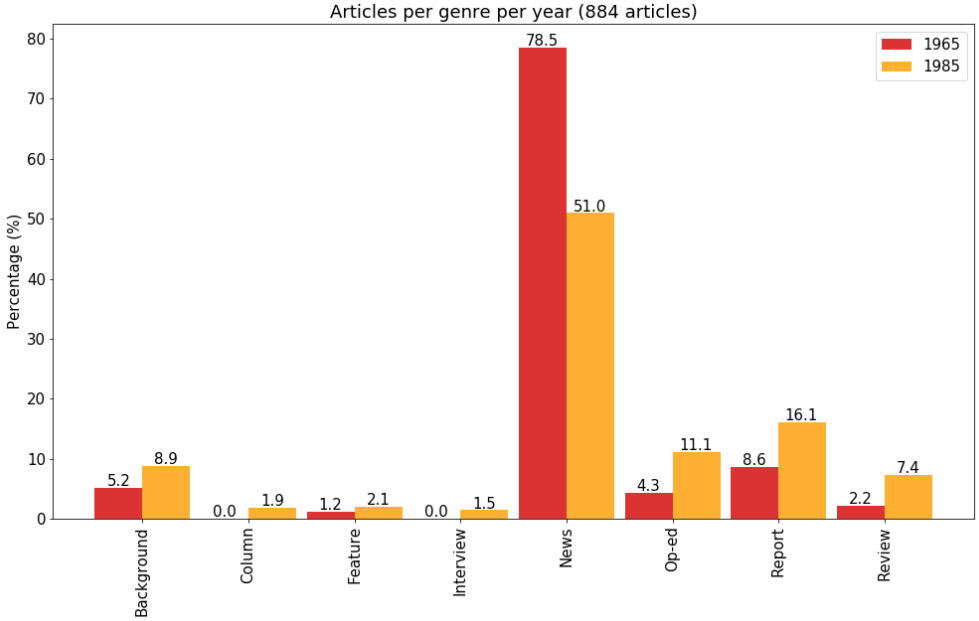


Domain hypothesis: Genre distribution over time

GOLD STANDARD



MACHINE-GENERATED



Concluding remarks

- With accuracy alone, we cannot select the best pipeline for the domain scientist
- Transparency promises in-depth insights and discussions for both the domain scientist and the computer scientist
- We can claim transparency increases trust of the domain scientist in the application of machine learning

UTILIZING A TRANSPARENCY-DRIVEN ENVIRONMENT TOWARD TRUSTED AUTOMATIC GENRE CLASSIFICATION

A Case Study in Journalism History

Aysenur Bilgin [1], **Erik Tjong Kim Sang**[2], Kim Smeenk[3], Laura Hollink[1], Jacco van Ossenbrugen[1], Frank Harbers[3], Marcel Broersma[3]

November 1, 2018

[1]CWI, [2]Netherlands eScience Center, [3]University of Groningen

aysenur.bilgin@cwi.nl