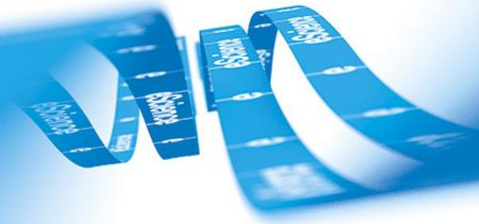
netherlands eScience center

CWI Centrum Wiskunde & Informatica

rijksuniversiteit groningen



Towards Transparent Linguistic Analysis of Dutch Newspaper Article Genres using Machine Learning

Erik Tjong Kim Sang, **Kim Smeenk**, Aysenur Bilgin, Tom Klaver, Laura Hollink, Jacco van Ossenbruggen, Frank Harbers and Marcel Broersma

CLIN29, Groningen, 31/01/2019

Task

Task: automatically predict genres of Dutch newspaper articles

Data: 2,930 Dutch newspaper articles with 16 different genre labels

Examples of genre labels: news, column, editorial, interview

Results

Previous work: Harbers and Lonij (2017) obtained 65% accuracy on this task

Our method: machine learning (MLP, NB, RF, SVM)

Result: 70% accuracy with SVM (interannotator agreement: 77%)

Application

We want to use the distribution of genres over time (1955-1995) to study the effects of depillarization of Dutch newspapers

The quality of the proposed genre labels should be very good, in particular: their predicted distributions should be excellent

Question

Can you convince us that the genre prediction system works well enough to base our future studies on?

1. Open the genre classification system
2. Look for components that could introduce **bias**
3. Improve the **transparency** of the system with data visualizations

We have built a platform supporting step 3

Dealing with OCR errors

VOOR AAN DE RADIO

EREDIVISIE

DOS—Feijenoord
Fortuna—PSV
Enschede—GVAV
Go Ahead—Ajax
ADO—Telstar
DWS—Sittardia
MVV—Heracles
Sparta—NAC

EERSTE DIVISIE

Eindhoven—NEC
RBC—Elinkwijk
Willem II—Velox
VVV—Blauw Wit
Alkmaar—Holland Sport
Volewijckers—Volendam
Veendam—Enschedese Boys
Excelsior—DHC

TWEEDE DIVISIE A

Hilversum—EDO
RCH—Graafschap
ZFC—Zwolse Boys
Heerenveen—Wageningen
Zwartemeer—AGOVV
Vitesse—Spel. Cambuur
PEC—FC Zaanstreek
Haarlem—Tubantia

TWEEDE DIVISIE B

Fortuna Vl.—Xerxes
Baronie—'t Gooi
HVC—BVV
SVV—Hermes DVS
DFC—Longa
Wilhelmina—Noord
Helmondia—Limburgia

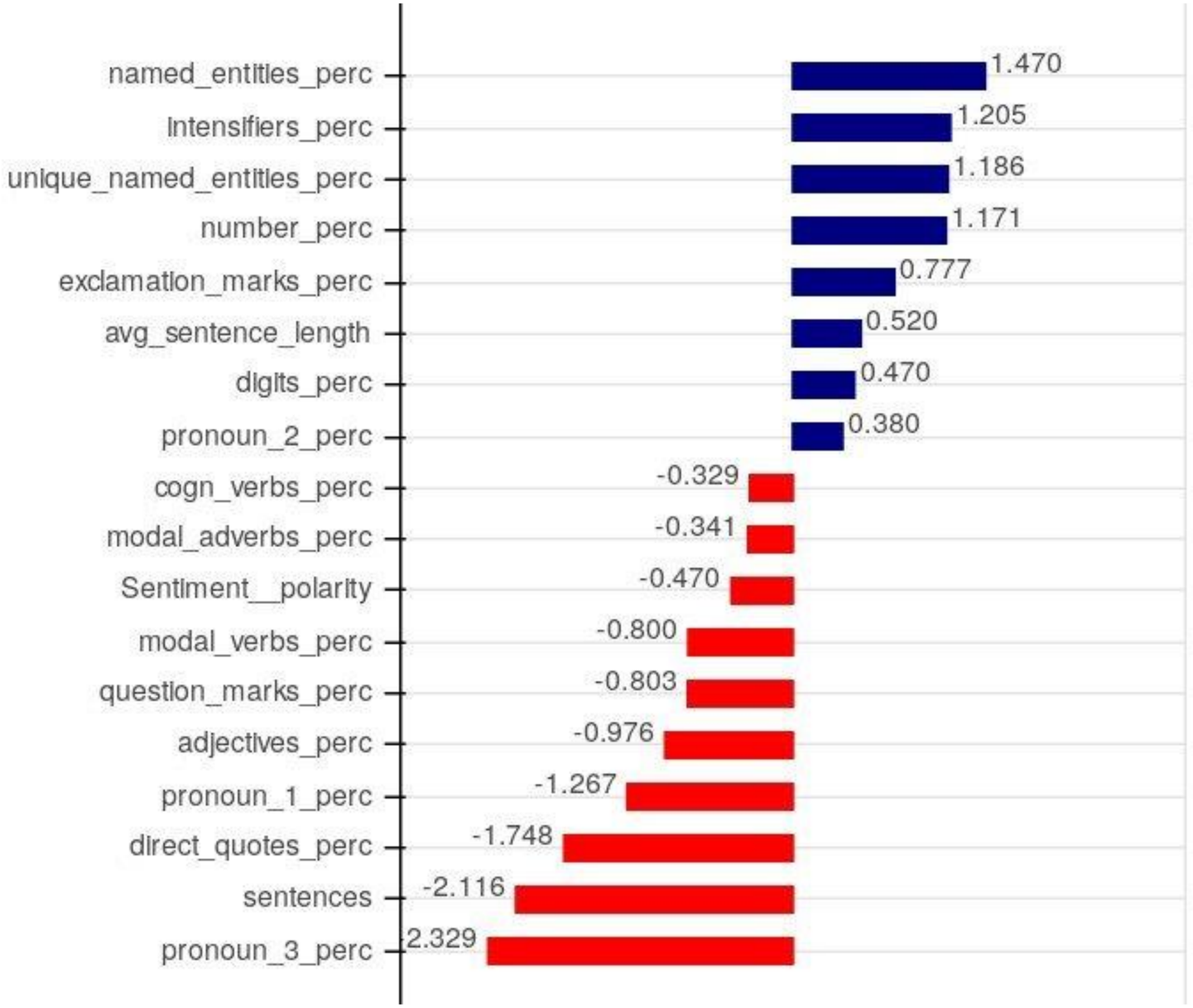
De vet gedrukte wedstrijden zijn opgenomen in de sport-toto. De cursief gedrukte zijn reserve-wedstrijden.

Paper version

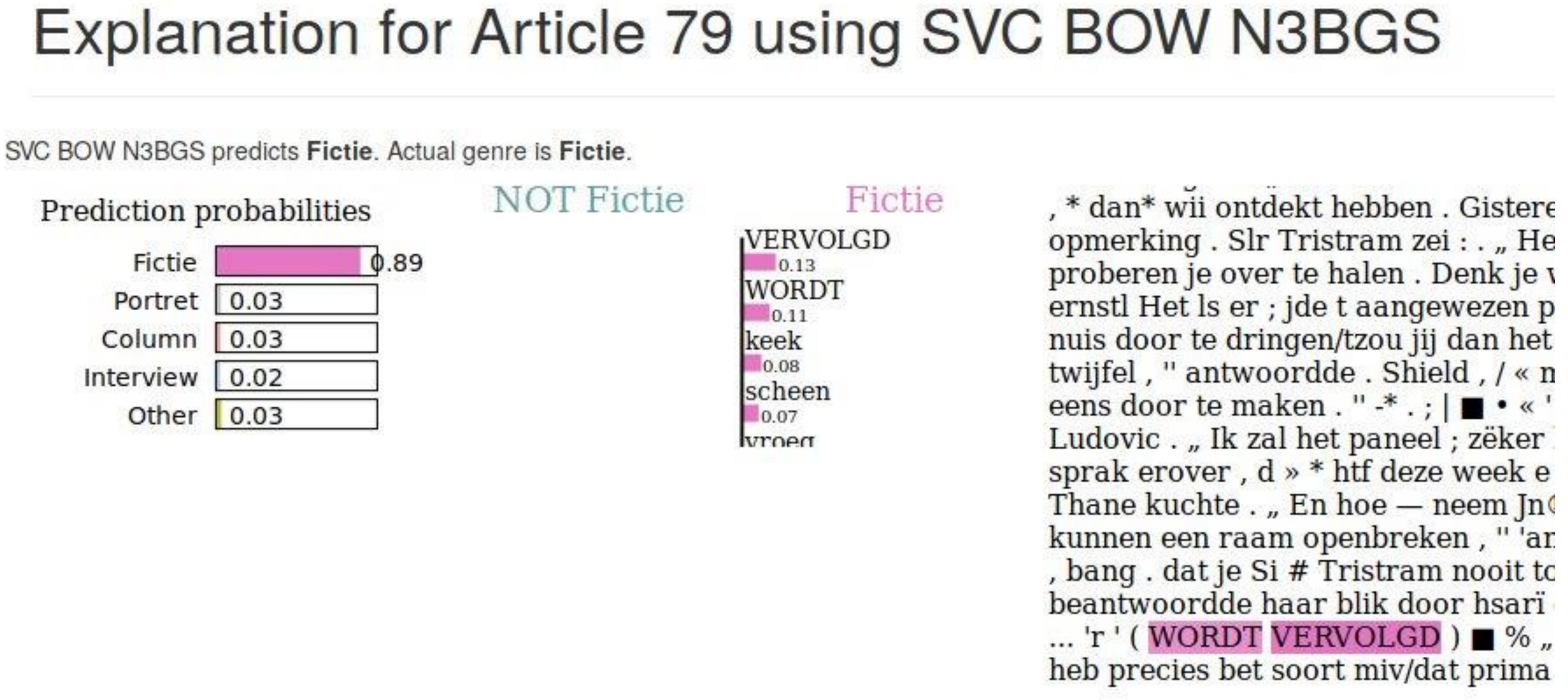
VOOR AAN DE RADIO
t TWEEDE DIVISIE A i Portu„a__Psv Hilversum-EDO _ Enschede rviv RCH—Graafschap 5 Go ZFC—Zwolse Boys f ADO—Telstii. Heerenveen—Wageningen .. . ï DWS—Sitter__, Zwartemeer—AGOVV i VlVV—HerVclés Vitesse—Spel. Cambuur 5 Sparta—Nac PEC-FC Zaanstreek EERSTE- DIVISIE ' ' Haarlem~Tubantia i SS-ar.™ TWEEDE DIVISIE B 'f Willen, H-lve'lov Fortuna Vl.-Xerxes •' VW—Blanw » Baronie—'t Gooi tSSB-S&» gfcfZe.DvS ■:.:::::. i 'SS:3S""U"■'■>'* '"'■'■"■ &£-__e*i_-■::::::: !• Helmondia—Limburgia «t zijn opgenomen in de sport-toto. De curfl.--.j_. ' '" drukte z'.l') reserve-wedstrijden. j'"'"v «__. ' *A- - - -v"-'^"-"JV-_-_-__r_-__^-».---I^v-"--__nj_-

Digital version

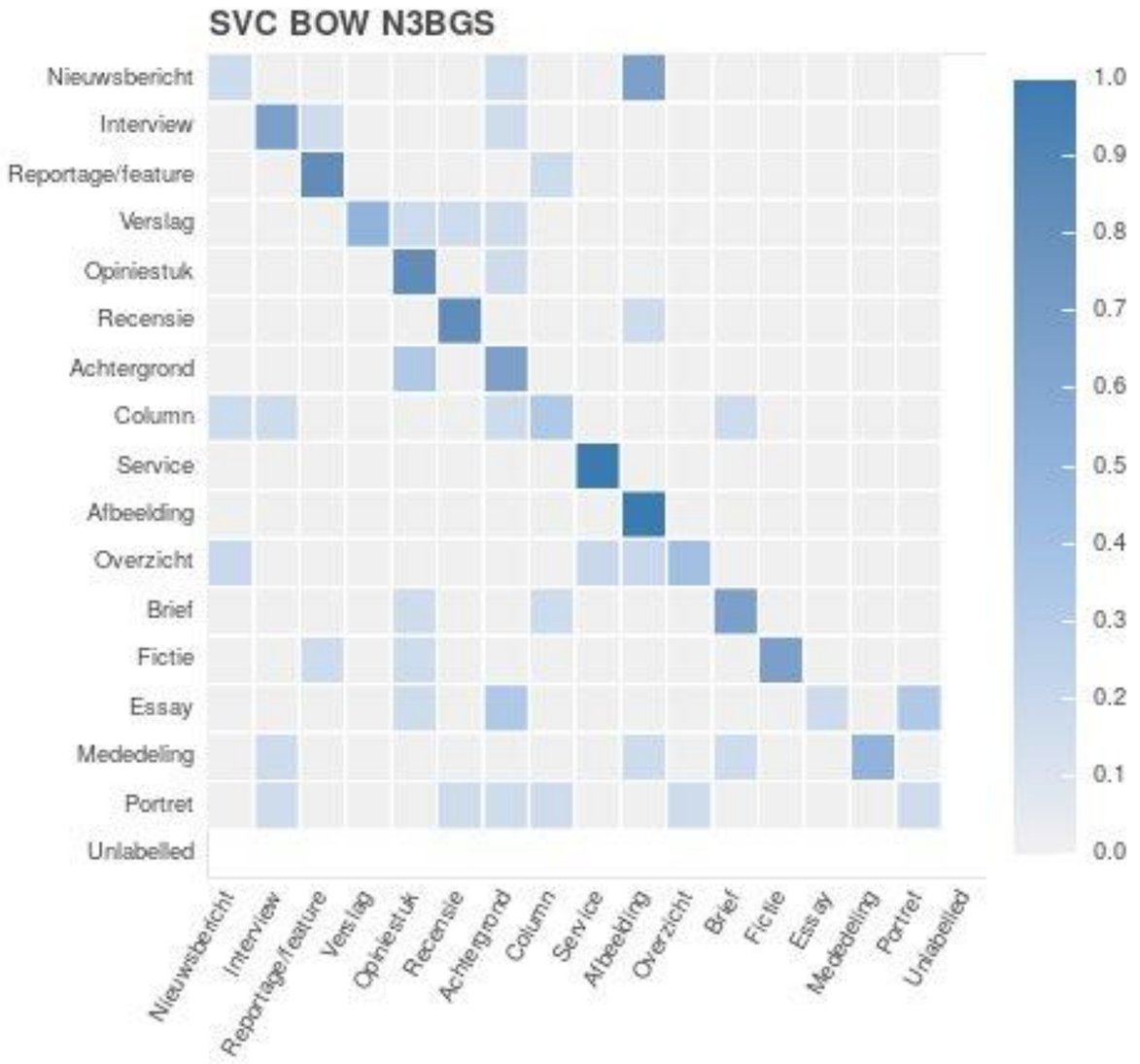
Example of important features for genre class comparisons: Interview (blue) vs Reportage (red)

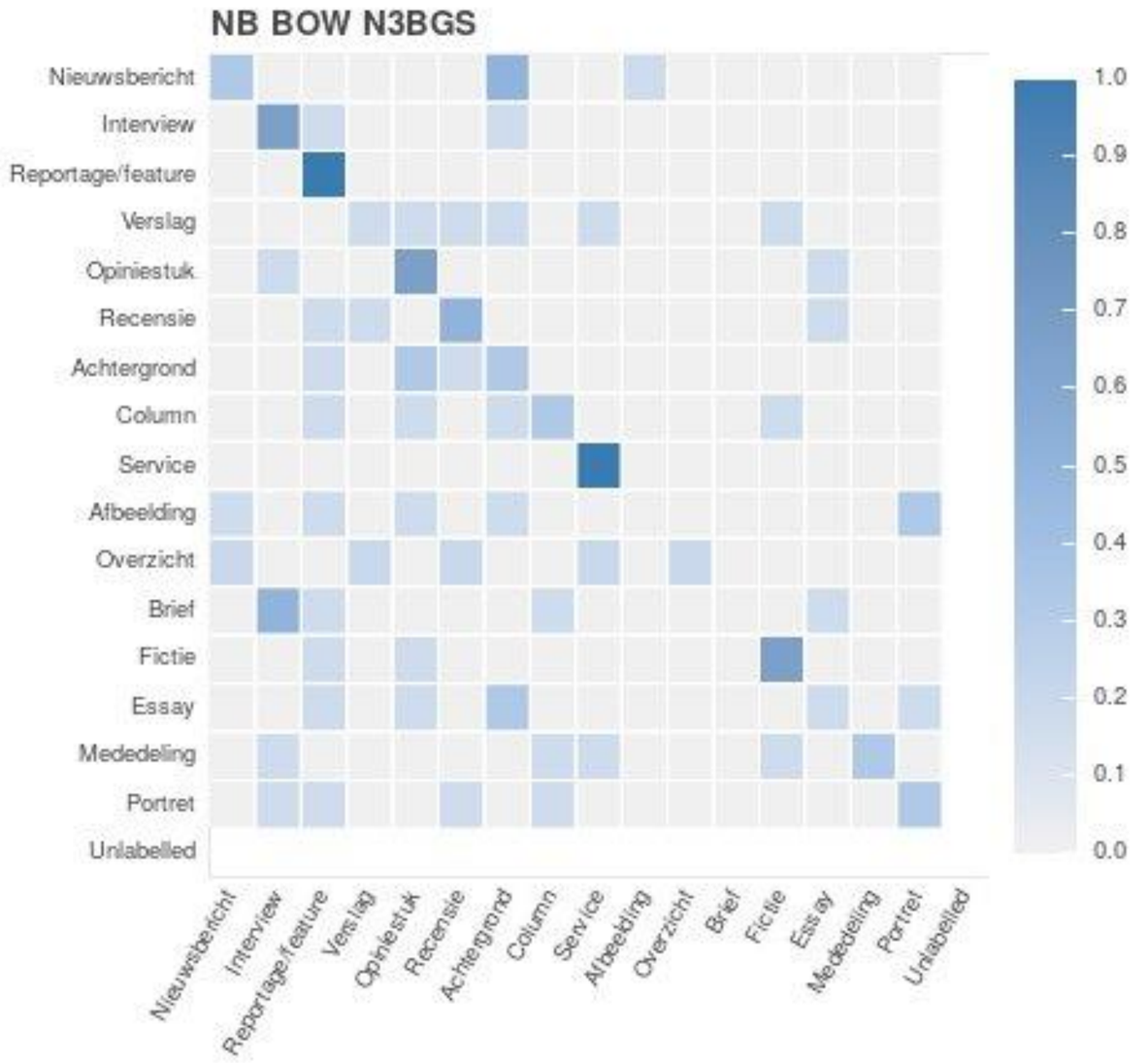


Visual explanation of genre class choice based on feature values

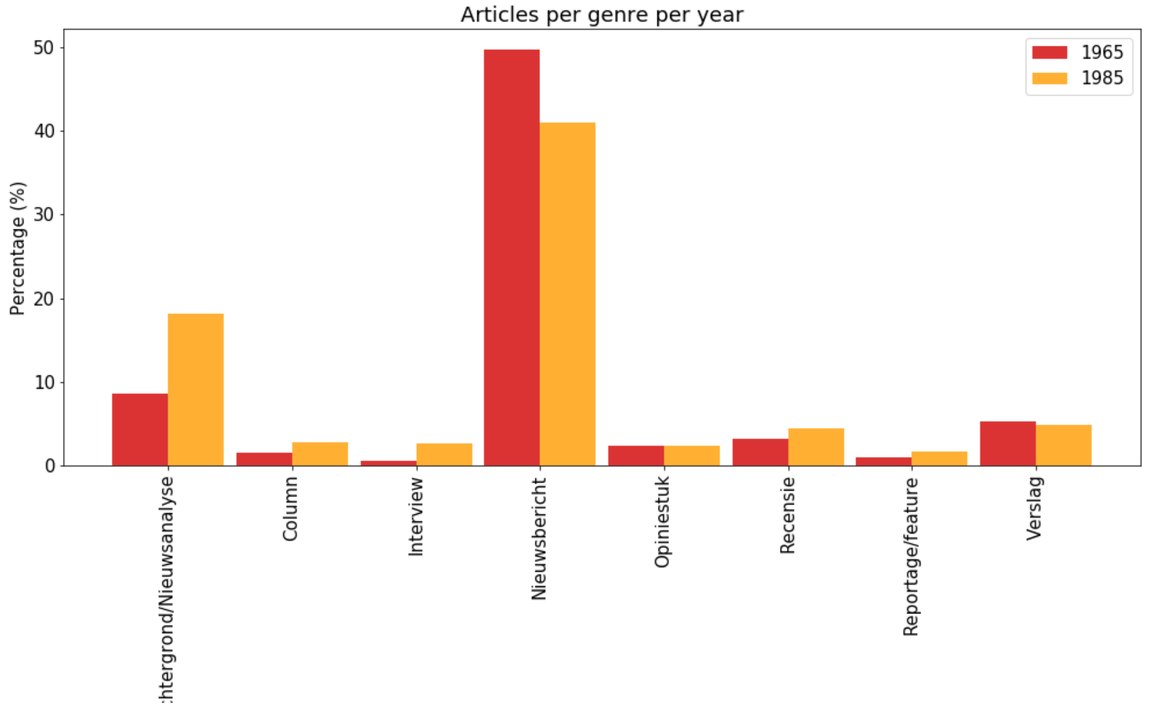


Visual explanation of genre class accuracies and genre class confusion

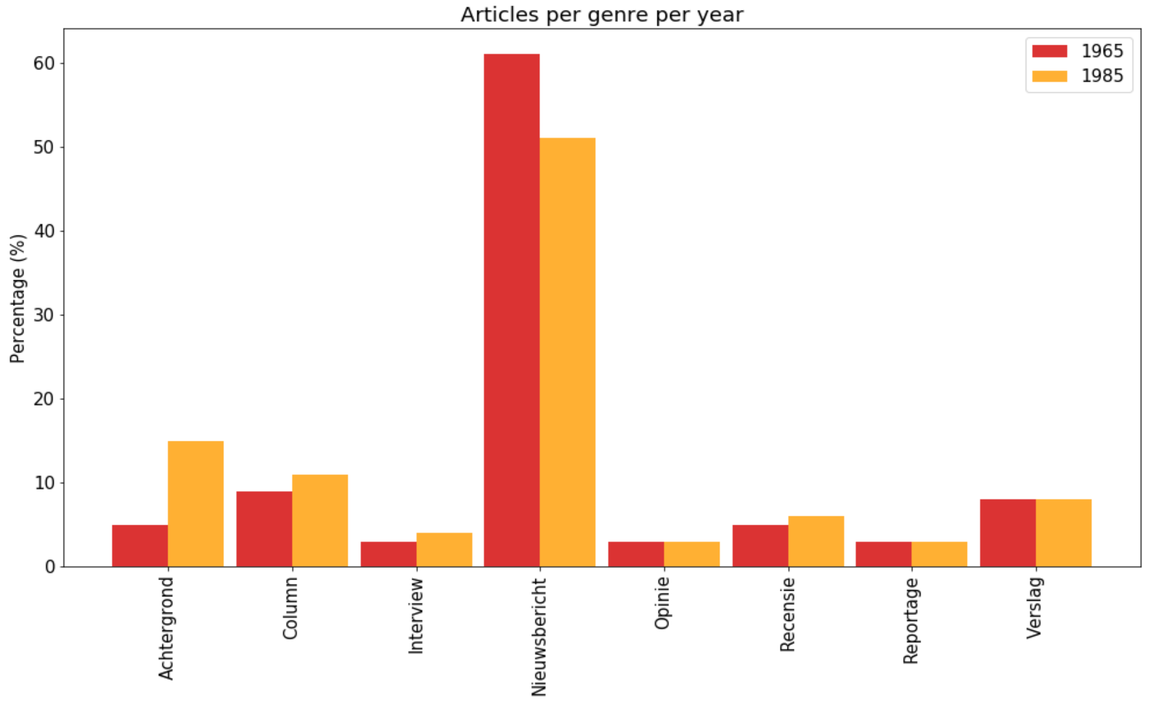




Gold standard data



Machine labeled data



The domain scientists regard the current quality of the predicted genre labels as too low to be used as a basis for further study

This involves both the label accuracy and the provided explanations for the labels

1. Collect more training data to improve model accuracy

2. Employ word vectors to overcome lack of training data

3. Look for better features, to generate better explanations

4. Evaluate alternative more advanced machine learners

Concluding remark

Improving the transparency of our classifier has improved the insights in the classification task, both for domain scientists and computer scientists