# Accurate Estimation of Class Distributions in Textual Data

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

CLIN30, Groningen, 30/01/2020

## Task

Task: automatically predict genres of Dutch newspaper articles

Data: 9,246 Dutch newspaper articles with 16 different genre labels

Examples of genre labels: News, Column, Editorial, Interview


Nieuwe Amsterdamsche Courant

ALGEMEEN HANDELSBLAD

Suezkanal was drie dagen geblokkeerd

President van PANAMA gisteren vermoord

IRAK verbreekt betrekkingen met Rusland

Oostelijke wind

Hammarskjoeld op weg naar Peking

Bosbranden in Zuid-Australië

REGEN BLUSTE VUUR

BIJ WEER IN EUROPA

Vredelievende woorden van premier Malenkow

Nog slechts twee infiltranten in N-Guinea vrij

GEJAAGD DOOR DE WIND

NIEUWJAARSWENSEN

VANDAAG

12 Pagina's

Minister Kernkamp luidde 1954 uit

Nieuwjaarsrede van minister dr Drees

## 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

We have built machine learning models for this task optimizing accuracy

But the research task is easier: we do not need to predict all labels correctly to obtain good distribution counts

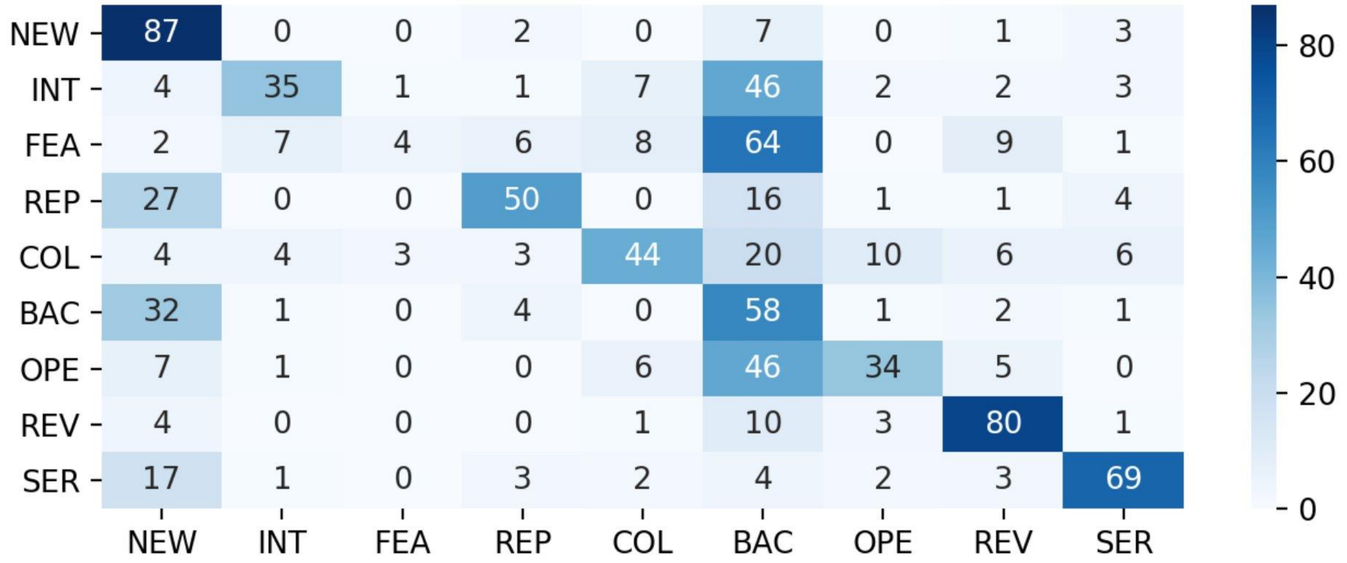
But the task is also more difficult: we also want to predict label distributions which are different from that of the training data

Results of cross-validation experiments with FastText on the train data

| Genre | Target count | Predicted count | Prediction error |
|---|---|---|---|
| News | 3697 | 4117 | +11% |
| Interview | 133 | 105 | -21% |
| Feature | 118 | 20 | -83% |
| Report | 482 | 426 | -12% |
| Column | 184 | 156 | -15% |
| Background | 1281 | 1466 | +14% |
| Op-ed | 215 | 170 | -21% |
| Review | 307 | 385 | +25% |
| Service | 1914 | 1486 | -22% |

1. Divide the predicted counts per genre in the test data by the predicted percentage of the class in the training data

2. As in point 1. but instead use the counts from a confusion matrix



## Results of re-estimating the test data counts

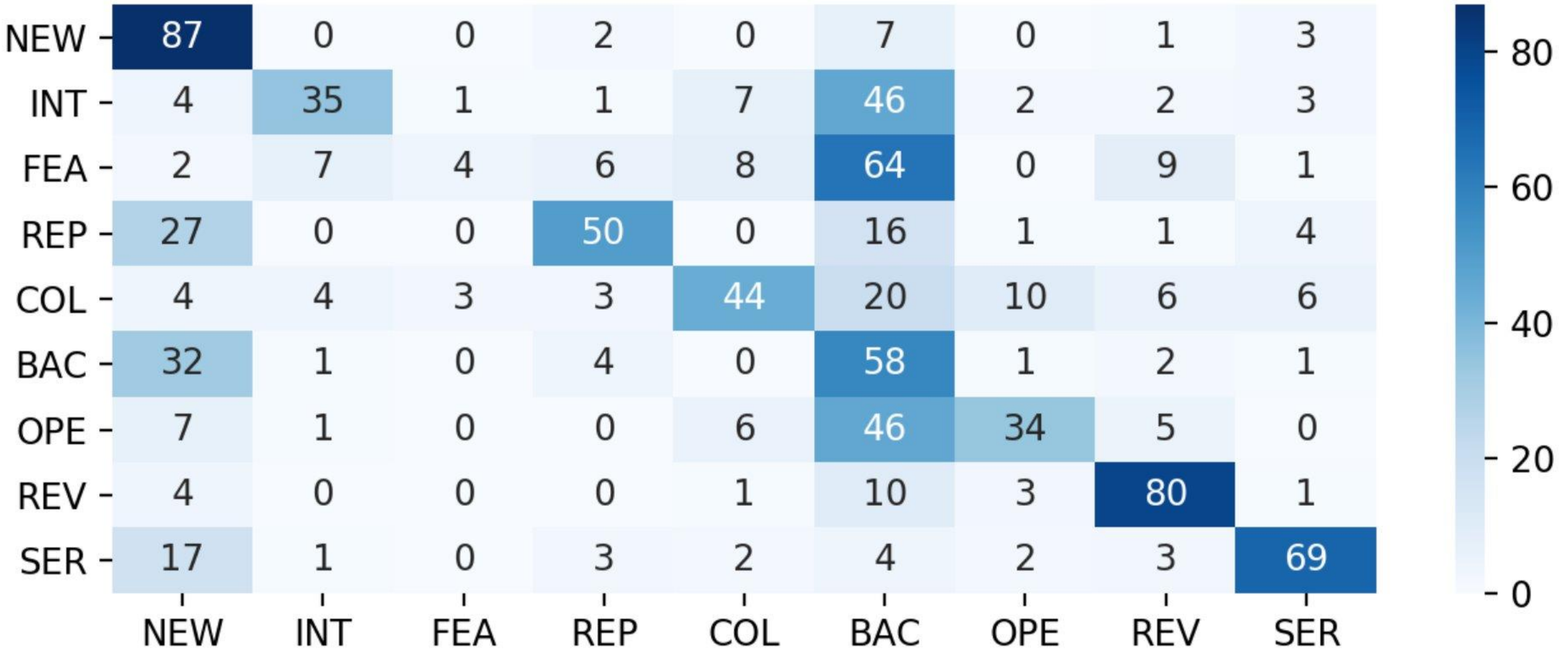
| Genre | Target | Predicted | Error | Re-estimated 1 | Re-estimated 2 |
|---|---|---|---|---|---|
| News | 345 | 371 | +8% | -3% | +8% |
| Interview | 37 | 12 | -68% | -59% | -56% |
| Feature | 19 | 3 | -84% | -7% | -22% |
| Report | 28 | 19 | -32% | -23% | +32% |
| Column | 39 | 23 | -41% | -30% | -33% |
| Background | 129 | 184 | +43% | +24% | +11% |
| Op-ed | 32 | 29 | -9% | +14% | -5% |
| Review | 73 | 64 | -12% | -30% | -33% |
| Service | 213 | 210 | -1% | +27% | +17% |
| Average | - | - | 33% | 24% | 24% |

## Results per newspaper on all data (train+test)

| Newspaper | Error | Re-estimated 2 | Newspaper-specific |
|---|---|---|---|
| NRC | 29% | 19% | 12% |
| Telegraaf | 24% | 21% | 10% |
| Volkskrant | 26% | 11% | 11% |

Newspaper-specifc involved the method Re-estimated 2 trained on only data of that particular newspaper rather than on data of all three newspapers.

Confusion matrix of the training data: cells contain percentages



George Forman 2005 presents:

- Adjusted Counts (similar to this paper) for binary identification
- Mixture Models

Related code: https://github.com/aesuli/semeval2016-task4/blob/master/semeval_quantification.py

Dallas Card & Noah A. Smith 2018 present:

- Calibrated Probabilistic Classify and Count

Code: https://github.com/dallascard/proportions

# Concluding remarks

We need accurate predictions of distribution for longitudional analysis of genres of Dutch newspapers

We presented methods for re-estimating distribution counts obtained from imperfect machine learning results

More work is needed forfurther improvements