Twitter Feeds Profiling With TF-IDF

Juraj Petrik & Daniela Chuda

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF INFORMATICS
AND INFORMATION TECHNOLOGIES

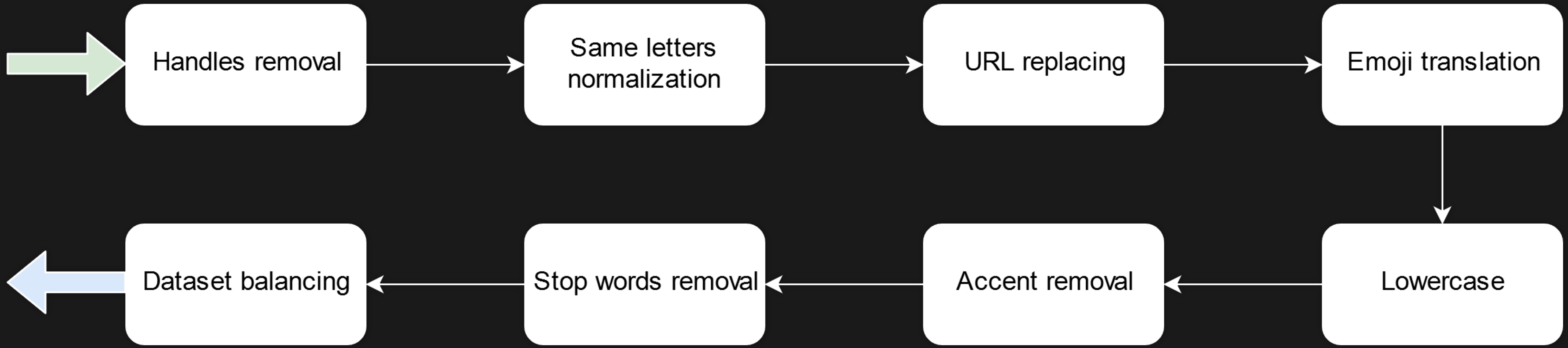
Task

- Given celebrity Twitter feed (English not guaranteed)
- Determine:
  - Fame level
  - Occupation
  - Age
  - Gender

Motivation

- Our background:
  - Source code authorship attribution – deep learning and frequency methods
  - Source code plagiarism detection – string similarity and character/word frequency methods
- Useful in plagiarism and also source code – comments for example

Preprocessing



First approach

- Convolutional hierarchical recurrent NN
- Class imbalance problem – trained network tends to prefer majority class
  - Oversampling, synthetic, random – better, but not enough
  - Undersampling - little to no effect
- Another problem – variable length feeds and pretty long
- Custom loss function to reflect f1 score
- ...also painfully slow
- Result from testing dataset 1 is from this approach

Preprocessing

Handles removal

- @superuser ->

Same letters normalization

- faaaaancy -> fancy

URL filtering

- https://t.co/adsadasd -> URL_TOKEN

Preprocessing

Emoji translation

- ☺ -> :smiling face:

Lowercase

- AaaaA -> aaaaa

Accent removal

- Čo sa deje -> Co sa deje

Stop words removal

- The, on, an, a… ->

Dataset balancing

- Random Oversampling
- SMOTE, TOMEK

Feature extraction

- N-gram based TF-IDF (1-3,5)
- Top 5000 features - grid search (matrix 5000x5000)

Classification

- One model per each “subtask”
- Random forest
- Extremely randomized trees
- Both have similar results, were more resistant to overfitting than our deep learning approaches
- Hyperparameter tuning – very similar results with 200+ trees

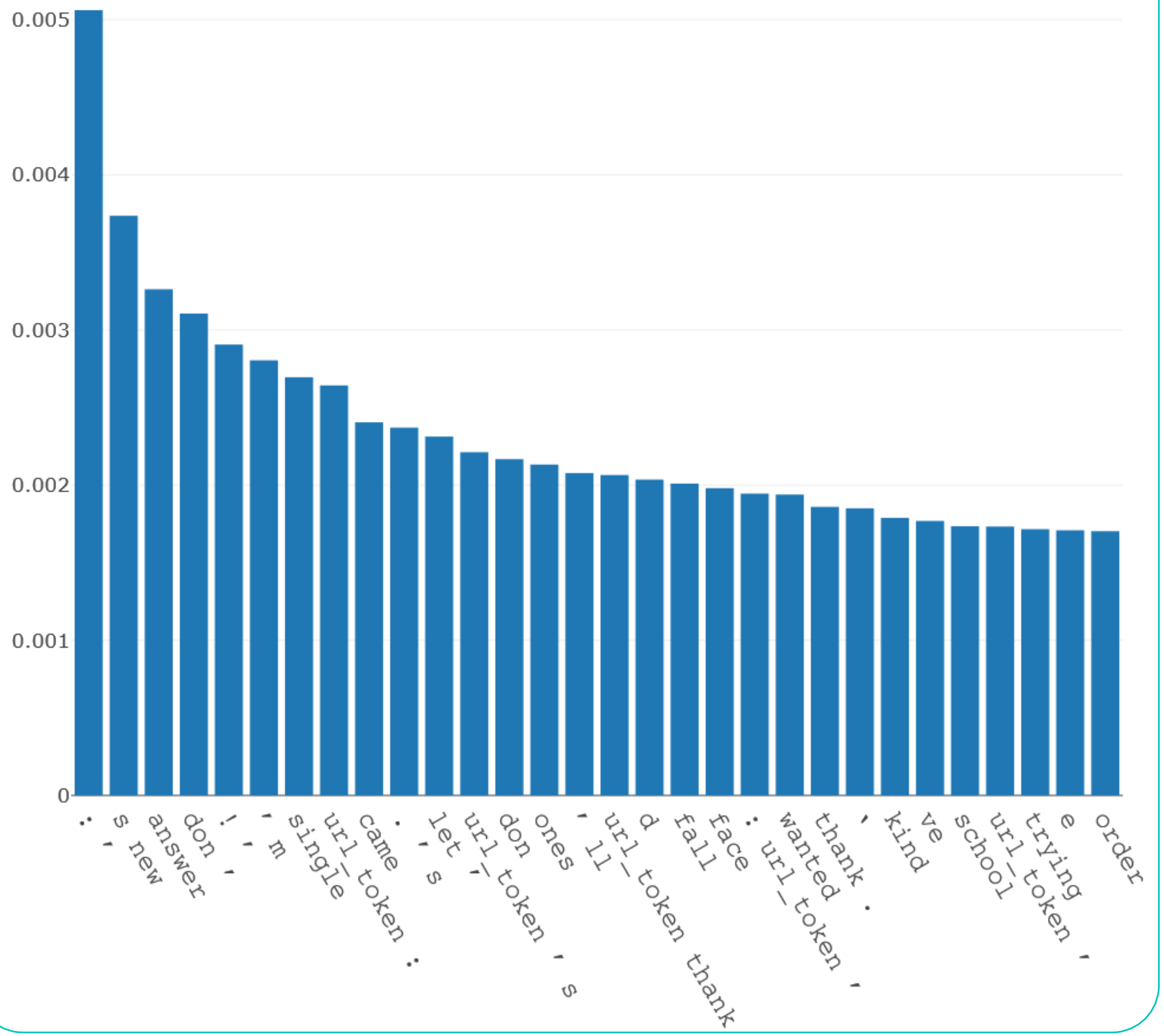
Regression

- Random forest regressor
- Used for birthyear trait
- Scaled to [0-1]
- Not so good in terms of the challenge as binning approaches

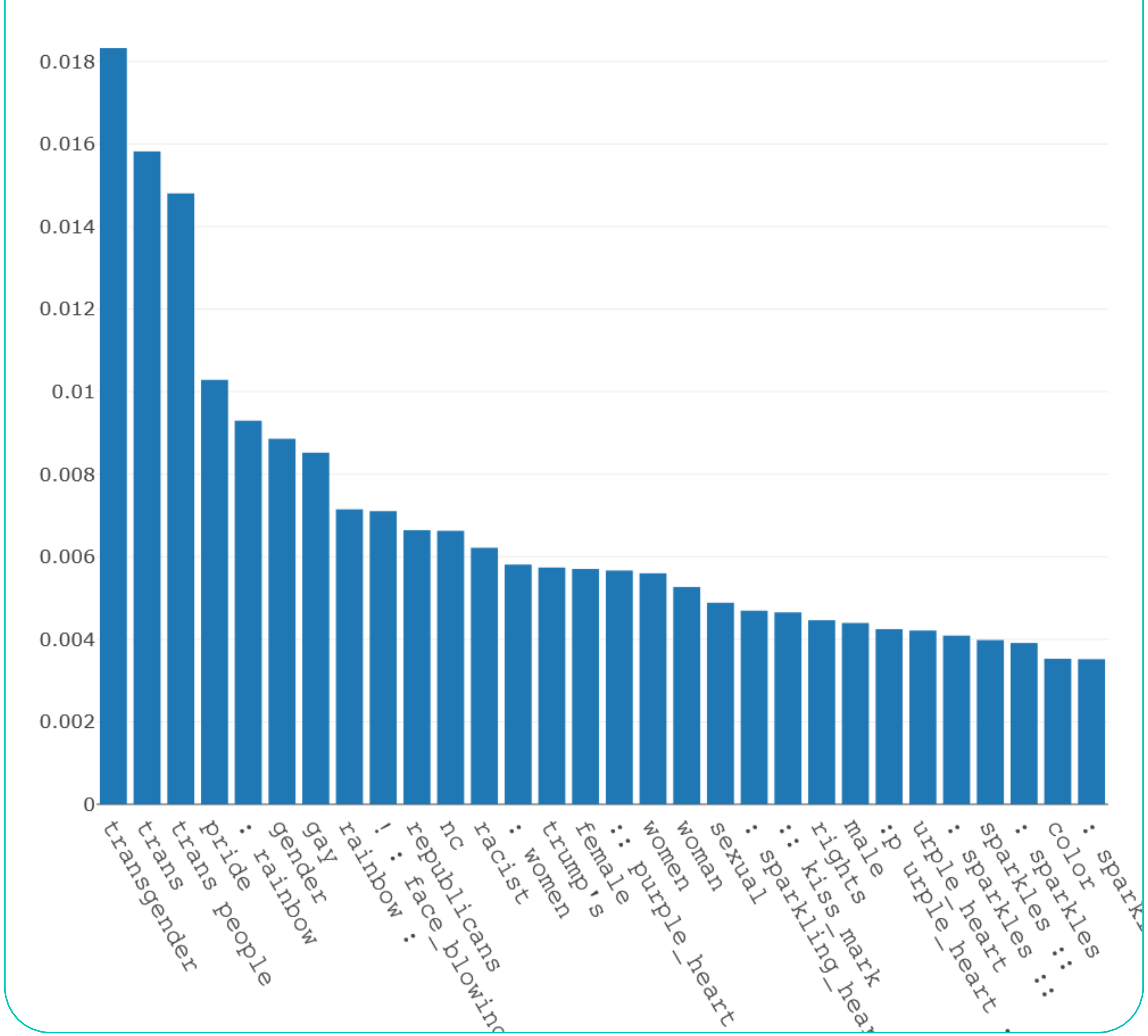
Name	cRank	F1				Accuracy				
		gender	occupation	fame	age	mean	gender	occupation	fame	age
radivchev19	0.558	0.608	0.461	0.547	0.657	0.743	0.930	0.757	0.770	0.517
morenosandoval19	0.497	0.560	0.418	0.517	0.515	0.627	0.861	0.722	0.547	0.376
martinc19	0.465	0.594	0.485	0.506	0.347	0.712	0.915	0.733	0.753	0.448
fernquist19	0.412	0.465	0.300	0.481	0.467	0.666	0.784	0.640	0.776	0.466
petrik19	**0.440**	**0.555**	**0.385**	**0.525**	**0.360**	**0.597**	**0.852**	**0.661**	**0.529**	**0.345**
asif19	0.401	0.587	0.427	0.504	0.254	0.696	0.905	0.758	0.776	0.346
bryan19	0.230	0.335	0.165	0.288	0.206	0.515	0.722	0.402	0.763	0.173

Name	female	male	nonbinary	star	superstar	rising	performer	creator	sports	manager	politics	science	professional	religious
	Classwise F1													
radivchev19	0.874	0.952	0	0.858	0.396	0.350	0.763	0.527	0.900	0.250	0.756	0.150	0.200	0
morenosandoval19	0.772	0.902	0	0.641	0.466	0.246	0.740	0.417	0.893	0.242	0.715	0.190	0.080	0
martinc19	0.835	0.943	0	0.848	0.383	0.178	0.730	0.470	0.869	0.300	0.736	0.142	0.200	0
fernquist19	0.449	0.866	0	0.869	0.258	0.111	0.617	0.362	0.785	0	0.632	0	0	0
petrik19	**0.759**	**0.894**	**0**	**0.620**	**0.434**	**0.292**	**0.708**	**0.344**	**0.854**	**0.086**	**0.700**	**0.142**	**0.160**	**0**
asif19	0.825	0.937	0	0.870	0.189	0.120	0.776	0.481	0.884	0	0.773	0.095	0	0
bryan19	0.014	0.838	0	0.865	0	0	0.318	0.108	0.550	0	0.218	0	0	0

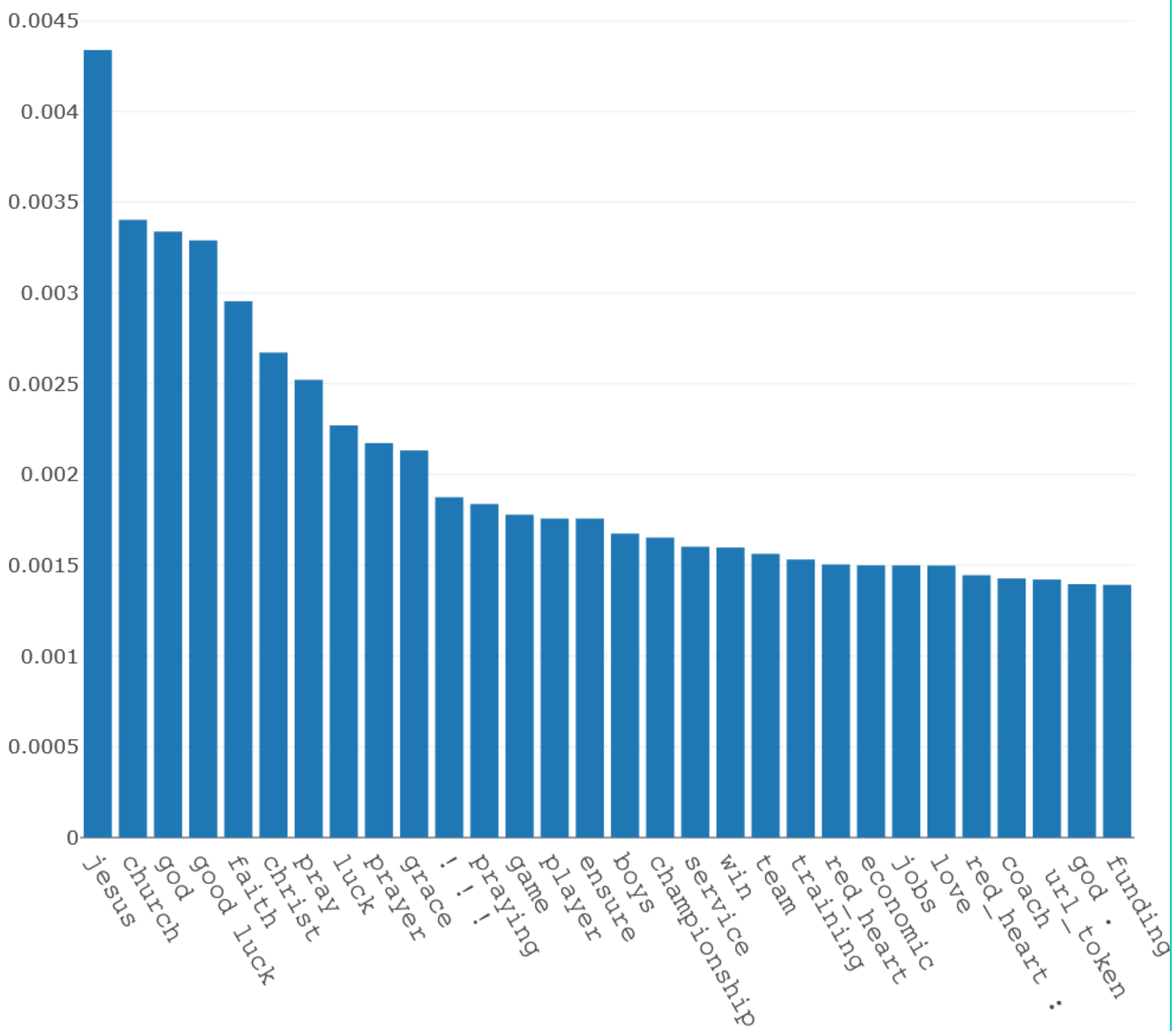
Feature importance - fame



Feature importance - gender



Feature importance - occupation



Possible improvements

- Oversampling – more sophisticated ones, focused on texts (synonyms, hypernyms from wordnet for example)
- Age prediction - regression vs bins (classification)
- Expand dataset – more data from Twitter (minority classes mainly)
- Language specific tuning