Know your neighbours: Efficient neighbor Profiling via Follower Tweets

Boshko Koloski
Senja Pollak
Blaž Škrlj

EMBEDDIA

University of Ljubljana
Faculty of Computer and Information Science

Jožef Stefan Institute, Ljubljana, Slovenia

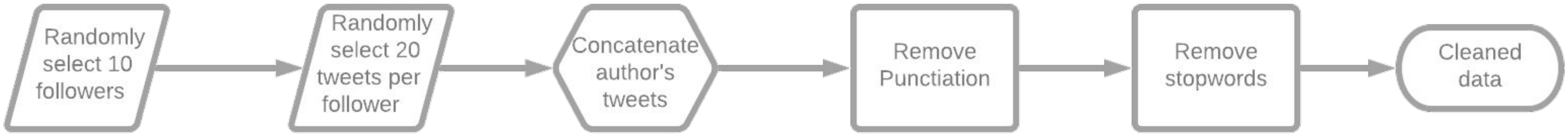
Task

- Given the tweets of the followers of a celebrity, determine celebrity’s:
    - Occupation
    - Gender
    - Birth Year
- Data balanced towards gender and occupation.
- Evaluate harmonic mean of F1-scores.

Our approach

- For each task (gender, age, birth year), we use a separate model.
- Employs lightweight method.
- Altered the birth year task.

Data preparation



Feature generation

Example tweet:

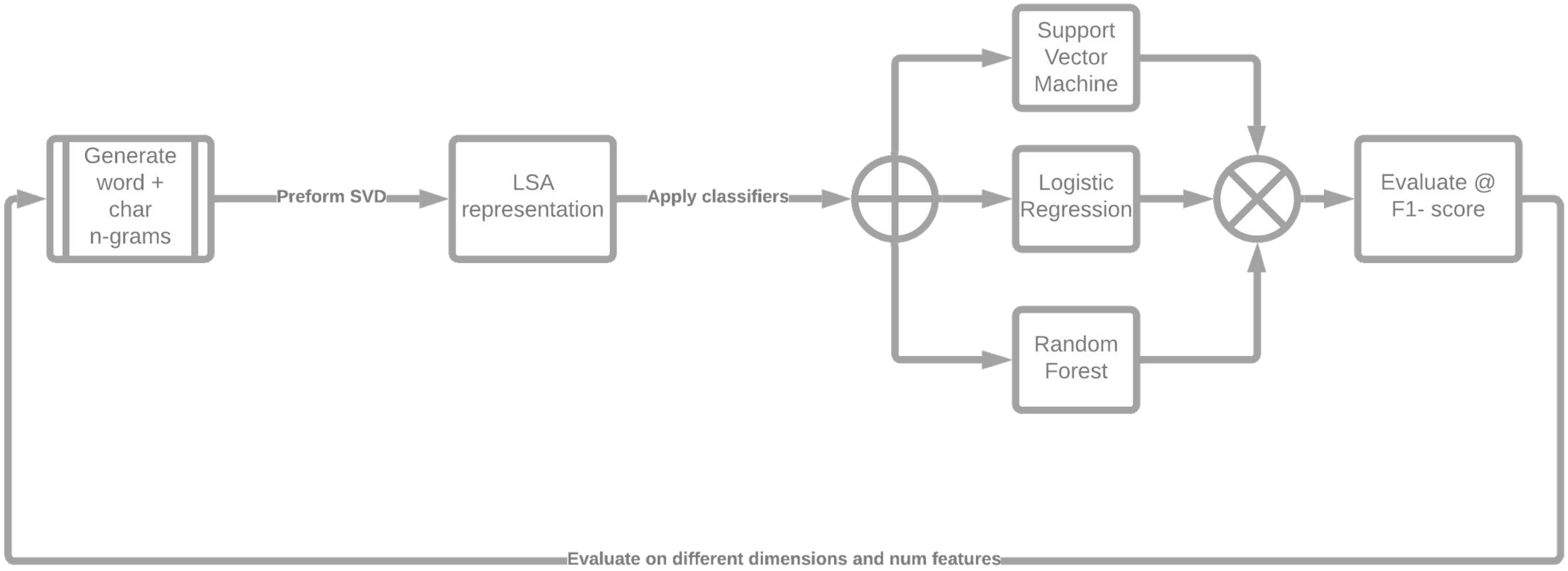
1) Character n-grams (1,2) :
   - 1-gram: d, o, n ; 2-gram: do, on, nt ;
2) Word n-grams (2,3) :
   - 2-grams: dont know; 3-gram: dont know where;
3) TF-IDF on generated features

Latent Semantic Analysis

Optimization

- Grid search on:
  - Number of generated features, n : [2500, 5000, 10000, 20000, 30000]
  - Number of dimensions in the SVD, d : [128, 256, 512, 640, 768, 1024]
- Model fine-tuning(regularization):
  - ElasticNet regularization
    - Lasso
    - Ridge

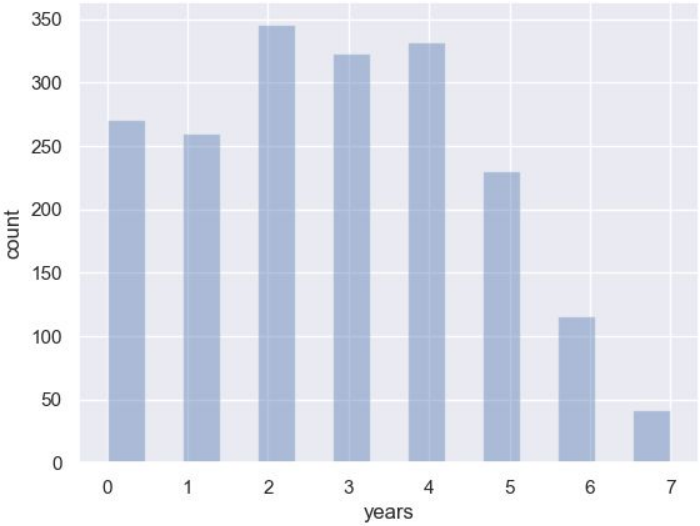
Learning pipeline



Birth year prediction

- [R] Regression
    - `birthyear = max(1949, min(predicted_year, 1999))`
- [FC] Full classification
    - `one class per year = 60 classes`
- [AC] Altered Classification
    - `custom intervals:`
      `[1949,1958];[1959,1966];[1967,1973];`
      `[1974,1980];[1981,1986];[1987 ,1991];`
      `[1992,1995];[1996,1999];`

Altered years



Results on training sets

Table 2: Final evaluation on training data on TIRA.

name	#features	#dimensions	f1 age	f1 gender	f1 occupation	crank
model-**AC**-2	20000	512	**0.358**	**0.665**	0.656	**0.516**
model-**AC**-1	20000	512	0.346	0.663	**0.669**	0.509
model-**FC**-2	10000	512	0.313	0.639	0.632	0.473
model-**FC**-1	10000	512	0.291	0.605	0.648	0.452
model-**R**	10000	512	0.298	0.612	0.613	0.453
baseline-ngrams	#	#	0.362	0.584	0.521	0.469

Test set evaluation

TEAM	TEST-DATASET			
	CRANK	AGE	GENDER	OCCUPATION
baseline-ngram-celebrity-tweets	0.631	0.500	0.753	0.700
hodge20	0.577	0.432	0.681	0.707
koloski20	0.521	0.407	0.616	0.597
tuksa20	0.477	0.315	0.696	0.598
baseline-ngram-follower-tweets	0.469	0.362	0.584	0.521
random	0.333	0.333	0.500	0.250

Conclusion

- Small sample based LSA gives competitive results for occupation and gender prediction
- Thresholding the years introduces significant improvement

Further work

- Development of improved strategies for thresholding
- Investigate performance in multilingual setting (see our paper on PAN Fake news detection Koloski et al. 2020)
- Adding background knowledge