An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection

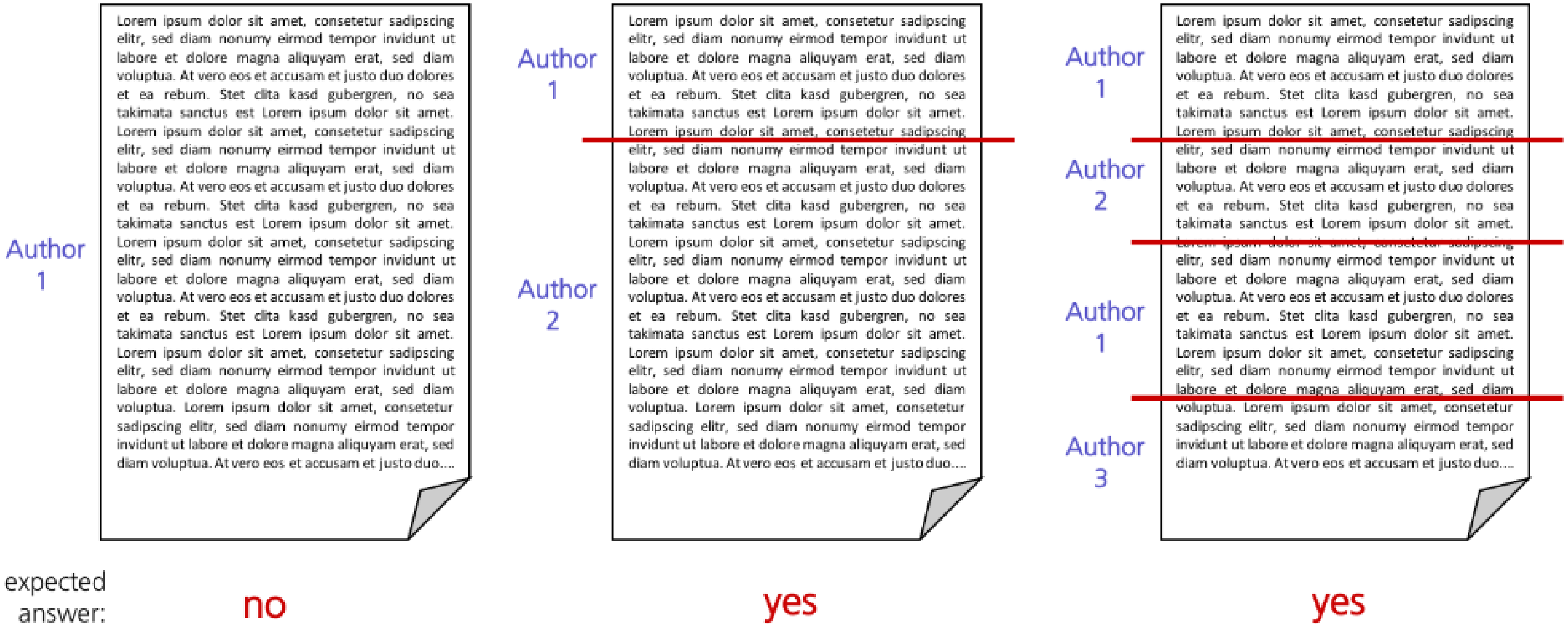
PAN at CLEF-2018

D. Zlatkova, D. Kopev, K. Mitov,
A. Atanasov, M. Hardalov, I. Koychev
Sofia University, Bulgaria

P. Nakov
Qatar Computing Research Institute, HBKU, Doha, Qatar

10-14 Sept. 2018, CLEF, Avignon

The Task


Author
1
Author
1
Author
2
Author
1
Author
2
Author
1
Author
3
expected
answer:
no
yes
yes

Related Work

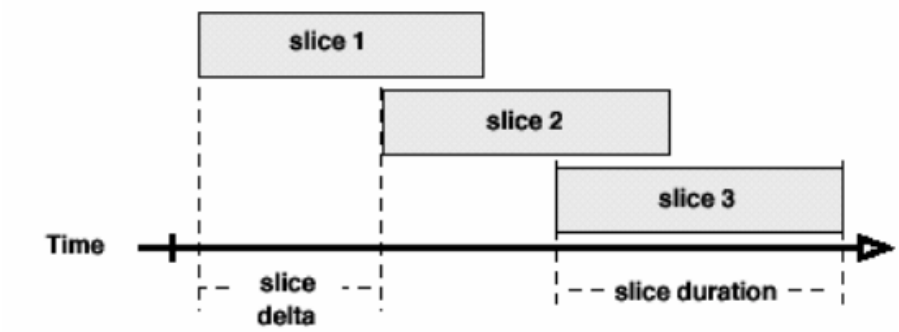
- General approaches for Style Breach Detection:
  - unsupervised methods
  - stylometry and TF-IDF features
- **Wilcoxon Signed Rank test** to check whether two segments are likely to come from the same distribution (Karas et al.)
- Outlier detection using **cosine-based distance** between sentence vectors using pre-trained skip-thought models (Safin and Kuznetsova)

Data Preprocessing

- Special tokens
  - http://www.java2s.com -> _URL_
  - 666573452995633321265321111111 -> _LONG_NUM_
  - /Users/Shared/Client/Blizzard -> _FILE_PATH
  - ========== -> _CHAR_SEQ
  - Taumatawhakatangihangakoauauo-> _LONG_WORD_
- Split hyphenated words
  - Pretends-To-Be-Scrum-But-Actually-Is-Not-Even-Agile

Text Segmentation

- Sliding Window
- 1/3 overlap
- Window size: 1/3 of doc length
- Max diff of feature vectors



Lexical Features

Characters:

- **spaces**
- digits
- commas
- (semi)colons
- apostrophes
- **quotes**
- **parenthesis**
- number of paragraphs

Words:

- **POS-tags**
- short (< 4 chars)
- **long (> 6 chars)**
- average length
- all-caps
- capitalized

Sentences:

- **question**
- **period**
- exclamation
- **short (<100chars)**
- long (>200 chars)

More Features

- Stop words: **you, the, is, of, ...**
- Function words: **least, well, etc, whether, ...**
- Readability, e.g Flesch reading ease:

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

- Vocabulary richness
  - Average word frequency class
    - frequency class of '*the' is 1*
    - frequency class of '*doppelganger' is 19*
  - Proportion of unknown words (not in corpus)

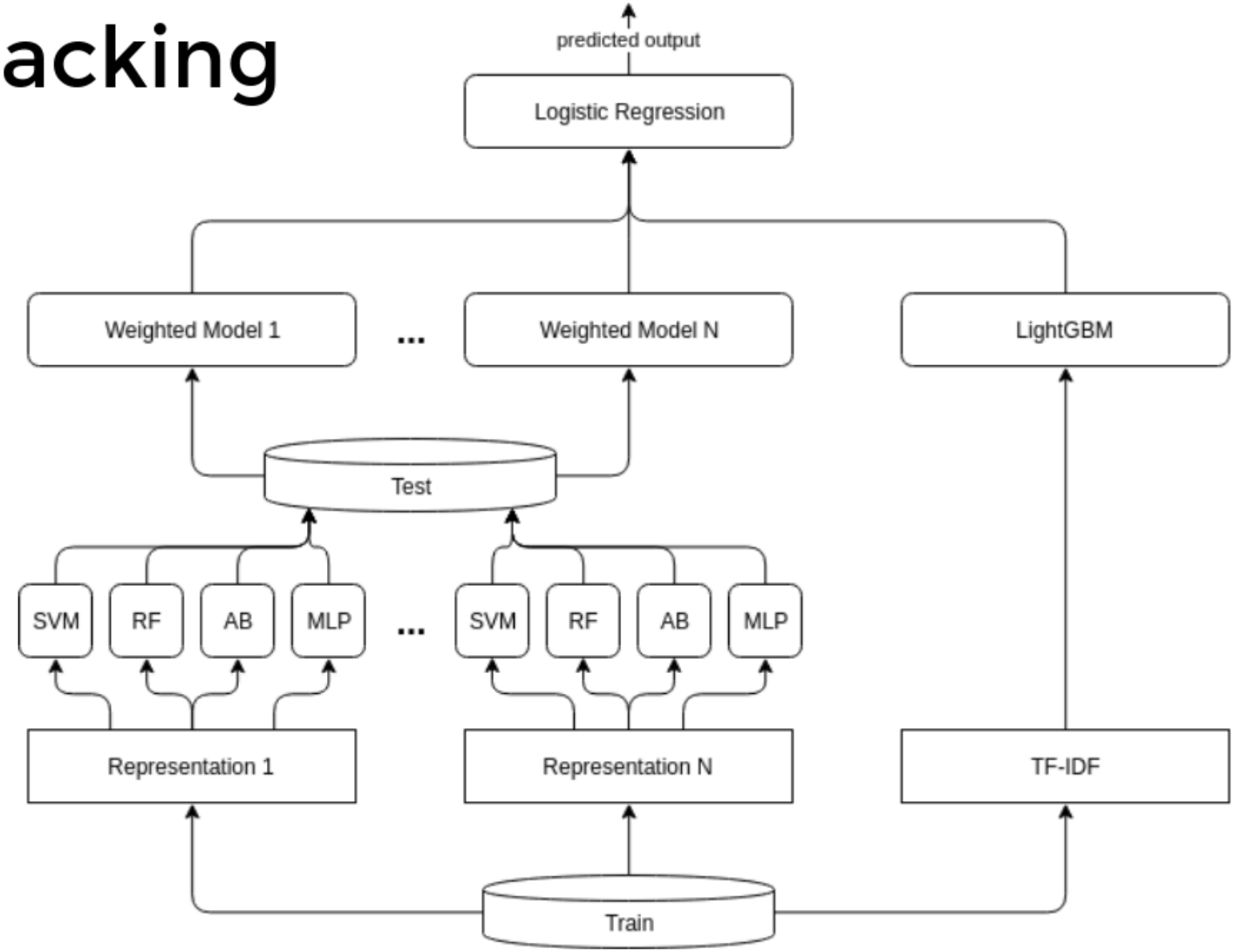
Even More Features

- Repetition
  - average number of occurrences of unigrams, bigrams, ..., 5-grams
- Grammar Contractions
  - *I will* vs. *I'll*
  - *are not* vs. *aren't*
- Quotation variation: ' vs. "

LightGBM + TF-IDF

- Character [2-6]-grams (up to 300k)
- Word [1-2]-grams (up to 300k)
- Logistic Regression for feature selection
- Parameter tuning to avoid overfitting
- Bagging
- Training TF-IDF on test documents

Stacking


predicted output
Logistic Regression
Weighted Model 1
...
Weighted Model N
LightGBM
Test
SVM
RF
AB
MLP
...
SVM
RF
AB
MLP
Representation 1
Representation N
TF-IDF
Train

Results

Classifier	Dataset	Accuracy
MLP w/ TF-IDF (Baseline)	validation	70.64
LightGBM w/ TF-IDF	validation	86.53
Stacking	validation	80.47
Stacking w/ LightGBM	validation	87.00
Stacking w/ LightGBM	test	89.35

Results

Table 10. Evaluation results of the style change detection task.

Submission	Accuracy	Runtime
Zlatkova et al.	**0.893**	01:35:25
Hosseinia and Mukherjee	0.825	10:12:28
Safin and Ogaltsov	0.803	00:05:15
Khan	0.643	00:01:10
Schaetti	0.621	00:03:36
C99-BASELINE	0.589	00:00:16
rnd2-BASELINE	0.560	–
rnd1-BASELINE	0.500	–

Style Breach Detection

- **PAN 2017** dataset
  - 134 training examples
  - 0 to 8 breaches
- use the developed **supervised** method
- search for breaches **recursively**
- outperforms **baseline** models

Conclusion

- High accuracy for **Style Change Detection** is achievable.
- **Ensembles** perform best.
- Using a supervised method to detect **exact breaches** is promising, but needs further work.

https://github.com/machinelearning-su/style-change-detection

References

1. *Karaś, D., Śpiewak, M., Sobecki, P.: OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection—Notebook for PAN at CLEF 2017.*
2. *Safin, K., Kuznetsova, R.: Style breach detection with neural sentence embeddings—notebook for PAN at CLEF 2017.*
3. *Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, Martin Potthast: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection.*