Authorship Identification

Final presentation by Liming Gong, He Feng

04/10/2019

Baseline Model

- 80% training + 10% test + 10% other
- Features: lexical, syntactic, writing density, readability, POS trigram diversity, stopword frequency and average word frequency class
- Collective attribution per unknown author
- Use 1000 logistic regression as the classifier

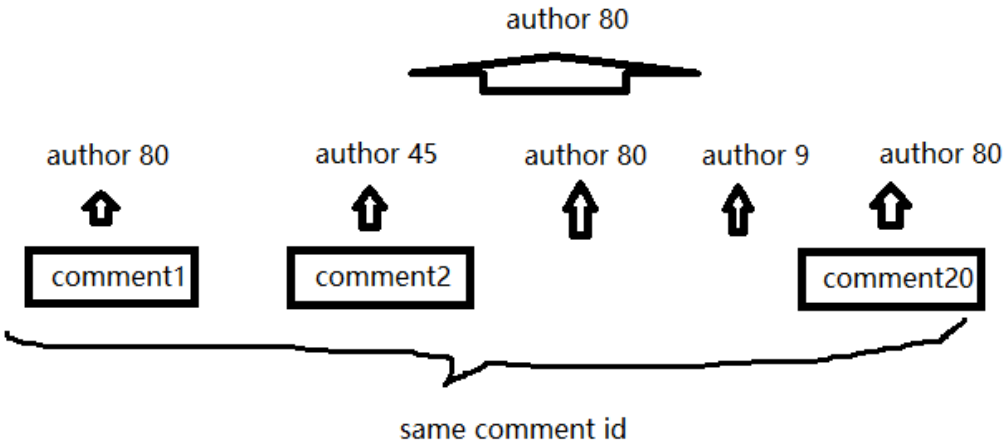
Table 2: Author verification results showing macro-averaged values

Dataset	Method	Positive Class			Negative Class			Accuracy
		Precision	Recall	F-score	Precision	Recall	F-score	
Amazon Reviews	NOS	0.8674	0.9165	0.8846	0.9193	0.8423	0.8696	87.94
Amazon Reviews	NRS	0.8600	0.9162	0.8806	0.9187	0.8331	0.8639	87.47
Yelp Hotel	NOS	0.8517	0.8921	0.8678	0.8915	0.8358	0.8579	86.39
Yelp Hotel	NRS	0.8636	0.8916	0.8732	0.8927	0.8495	0.8656	87.05
Yelp Restaurant	NOS	0.8595	0.8757	0.8617	0.8804	0.8449	0.8557	86.03
Yelp Restaurant	NRS	0.8567	0.8799	0.8628	0.8825	0.8401	0.854	86.00

The cumulative probability for rank 1 is 0.4245 for the Amazon dataset.

Our Result

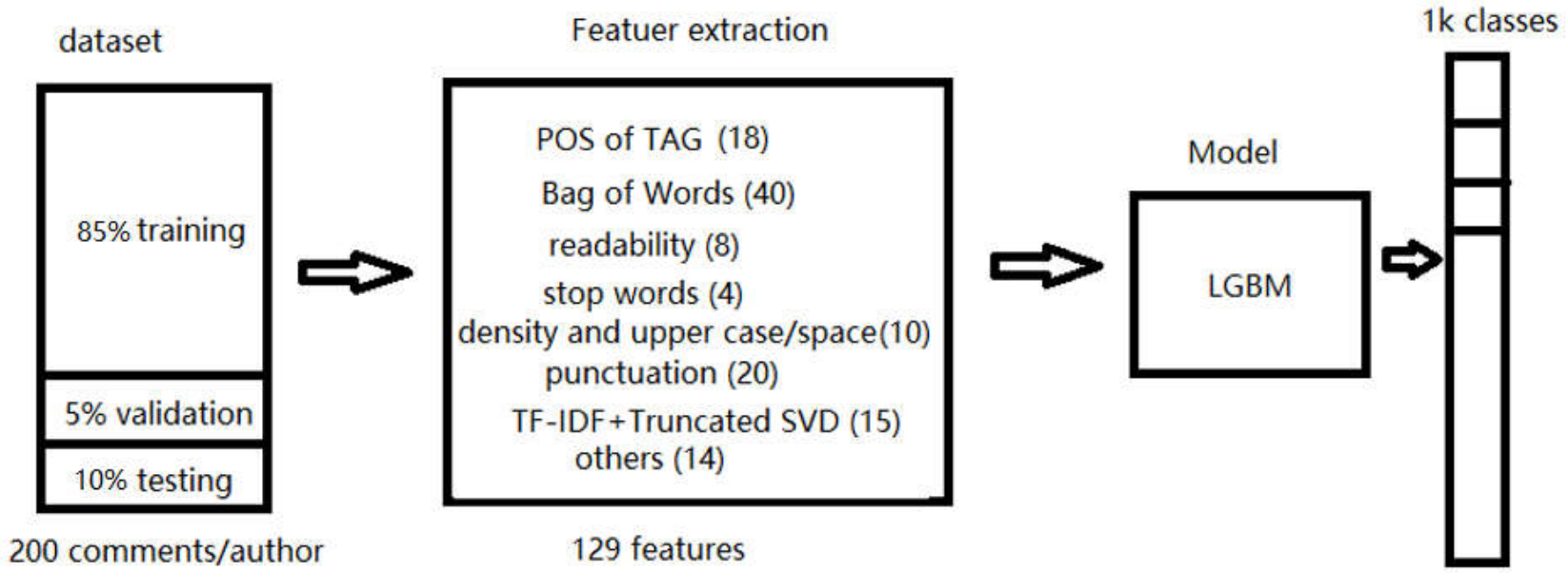
- A single combined multi-class classifier, directly classify between 1k authors.
- Rank1 accuracy: 43.82% > 42.45% baseline
- Collective voting(majority voting between 20 comments sharing the same comment id): 96.2%



Dependency

- Python3
- LGBM
- NLTK
- scikit-learn
- textstat
- Jupyter notebook
- pandas
- pymysql

Our architecture



Dataset Download

Large Scale Authorship Attribution of Online Reviews

CCLING 2016

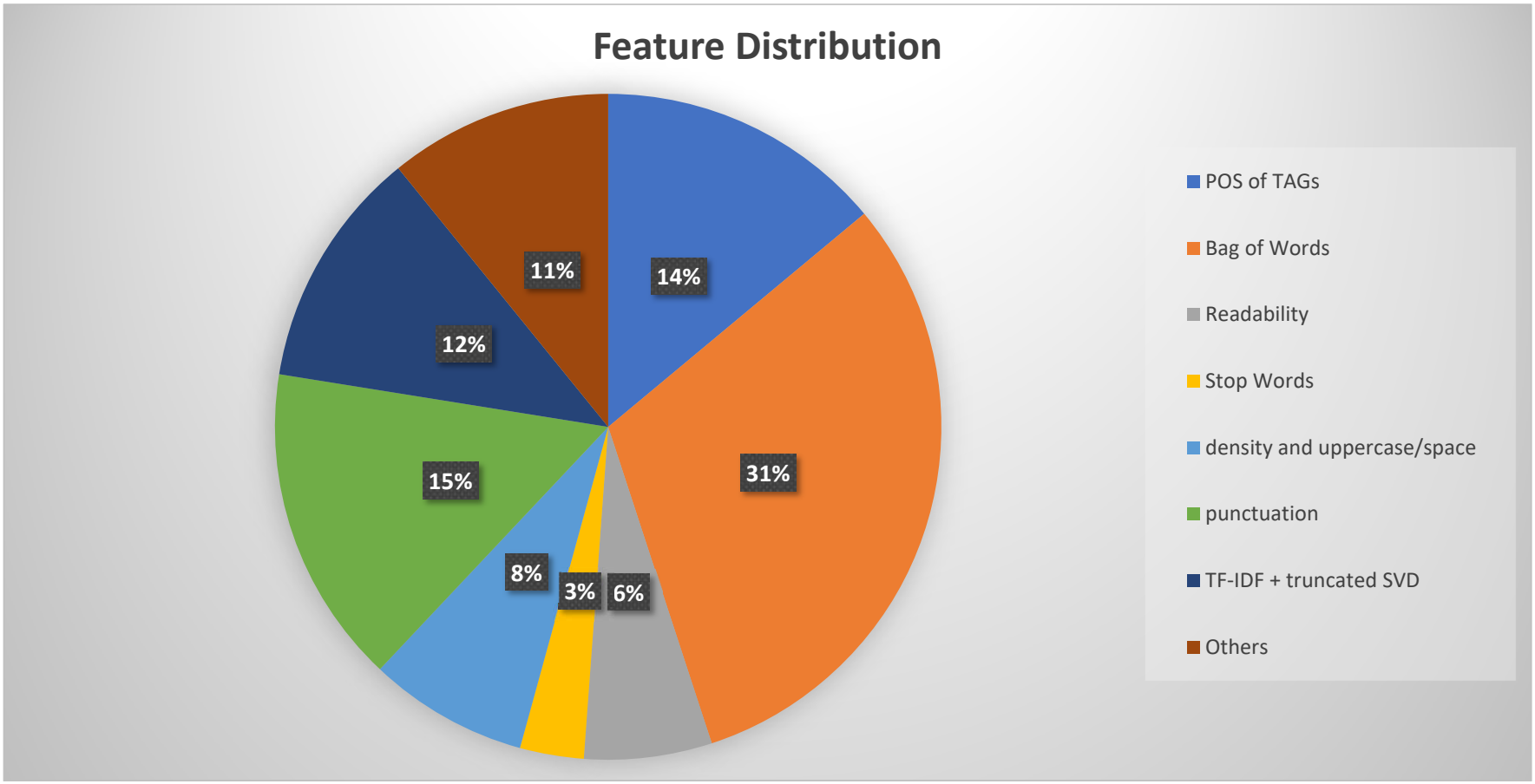
Authorship Attribution on Reviews (CICLING 2016) The dataset is presented as MySQL tables. You can get the data from the following links: Amazon Reviews Yelp Hotel Reviews Yelp Restaurant Reviews

Data preparation

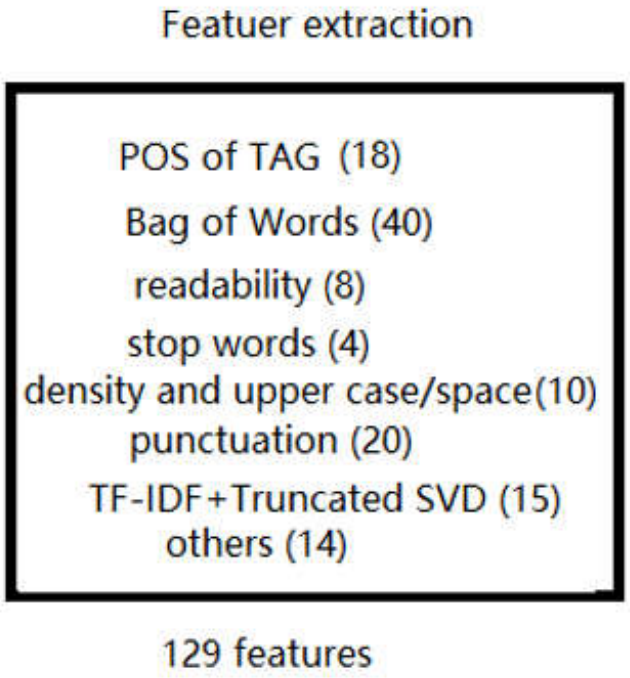
- Only amazon comments;
- 200 comments/author * 1000 author = 200k comments;
- 85% training + 5% validation + 10% testing;
- Use description + title

	description	title	custom_id
0	There is nothing special about this streamer. ...	Mediocre	0
1	For starters, I did not receive the keyboard t...	Can't Use	0
2	My home is located about half way between wher...	So Far, So Good	0
3	In the box is the quite attractive unit, a fil...	Almost	0
4	I have reviewed this previously but it shows a...	Repeat Review	0

Features, 129 in total



Features



- POS of TAGs: NN,NNP,DT,IN,JJ,NNS,UH,PDT,MD
- Bag of Words: 20 for title, 20 for description, only use 20 most frequent words.
- Punctuation: check frequency of ,;:!(?.-”&
- Readability, use below simultaneously:
  textstat.flesch_reading_ease
  textstat.flesch_kincaid_grade
  textstat.gunning_fog
  textstat.linsear_write_formula
- TF-IDF is huge, so I use truncated SVD to choose 10 most important features for description and 5 for title

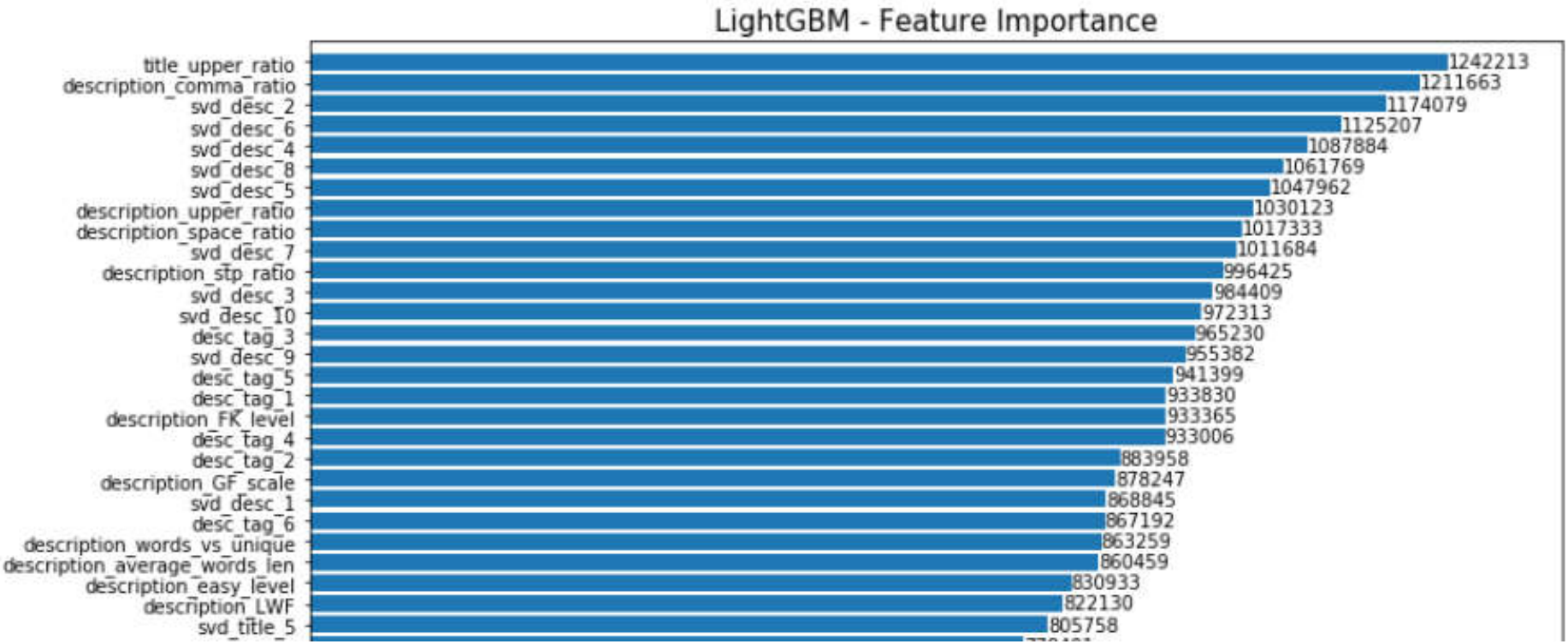
What is Light GBM

- Widely used in Kaggle competition as winner solution
- Based on gradient boosting tree method
- Support multi-class classification directly
- Can use GPU to accelerate
- Support early stopping
- Run very fast

Hyper parameters

```
params = {
        "objective" : "multiclass",
        "metric" : "multi_logloss",
        'boosting_type': 'gbdt',
        'num_class' : 1000,
        'max_bin' : 255,
        'metric_freq' : 5,
        "is_training_metric" : 'true',
        "learning_rate" : 0.01,
        "bagging_fraction" : 0.7,
        "feature_fraction" : 0.7,
        "bagging_freq" : 5,
        "bagging_seed" : 2018,
        "verbosity" : 1,
        'device': 'cpu',
        'gpu_platform_id': 0,
        'gpu_device_id': 0
    }
```

Feature importance



Summary

- Light GBM is strong
- Upper case ratio is a strong feature
- TF-IDF + truncated SVD choose very good and reliable features, even though not explainable
- Easy level is also important
- POS of TAGs are important

Complementary Materials

- My GitHub Code: https://github.com/stephenkung/authorship/
- Intel i5 + 24GB RAM, training 1530 epochs needs 8 hours.
- Light GBM: https://lightgbm.readthedocs.io/en/latest/