A Deep Learning Method Study of User Interest Classification

Alexey Malafeev and Kirill Nikolaev
National Research University Higher School of Economics
Nizhny Novgorod, Russia

Task and Simplifications

- The task of text classification;
- Each text corresponds to a single interest:
  - Real-life texts often correspond to none, or to more than one;
  - A finite set of interests: user forum activity empirical analysis

Corpus

- **89 844** web-forum (kinopoisk, livelib) documents;
- Not less than 150 characters;
- Unbalanced dataset;
- Test + validation sets: 1000 each, 100 per class

Class Distribution

Interest	All texts		Over 150 characters	
Anime	7663	3,66%	3213	3,58%
Food	11751	5,61%	5866	6,53%
Art	2216	1,06%	1175	1,31%
Games	67282	32,10%	29550	32,89%
Books	18008	8,59%	9999	11,13%
Music	21637	10,32%	7974	8,88%
Nature	2578	1,23%	1057	1,18%
Travel	3137	1,50%	1914	2,13%
Films	12961	6,18%	5862	6,52%
Football	62397	29,77%	23234	25,86%
Total	209630		89844	

Preprocessing and Representations

Stop-words removal

Non-Cyrillic tokens removal;

PPMI: most informative ***words*** per class;

Doc2vec model (training set)

PPMI: most informative ***trigrams*** per class;

Class-specific element proportions as features: 10+10

Classic ML algorithms;

Different feature combinations;

Deep learning algorithms

Optimum search: **micro accuracy** as a primary metric

Statistics

Classic Machine learning vs Deep learning algorithms:

Model	Classification accuracy
Random Forest Classifier – 100 e.	0.595
Gaussian Naïve Bayes	0.627
1D CNN (26)	0.761
Feedforward: 32 – 32	0.771
Bidirectional LSTM (100)	0.785
LSTM (100)	**0.786**

Class weighting:

Model	Weighted	Unweighted
LSTM	**0.786**	0.743
Feedforward	**0.771**	0.723

Different representation results:

Model	D2v	D2v + W	D2v + Ch3	D2v + W + Ch3
LSTM	0.657	0.747	0.717	**0.786**
Feedforward	0.640	0.740	0.735	**0.771**

Word count for PPMI:

Model	100	200	300
LSTM	**0.786**	0.757	0.749
Feedforward	**0.771**	0.741	0.738

Summary

- **Most misclassified:** music (confused with all), nature (confused with travel);
- **Highest accuracy:** LSTM, D2v + words + char trigrams;
- **Future work:** further data, rebalanced datasets, SotA representations (ELMO, BERT), multiclass classification.

Dataset and source code:
https://bit.ly/32tHhT9

amalafeev@yandex.ru
kir.nikolaev.7@gmail.com