Classification of User Interests by Text Messages with Recurrent Neural Networks

Alexey Malafeev (amalafeev@yandex.ru),
Kirill Nikolaev (kinikolaev@edu.hse.ru),
National Research University Higher School of Economics

NET-2019

Acknowledgements

- This research is being conducted within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (grant No. 19-04-004) and of the Russian Academic Excellence Project "5-100".

User interest detection

- Social networks grow in popularity:
  - User data analysis: recommenders, targeted advertising;

- Mostly utilize personal data (age, nationality, etc.) or metadata (search history, user ratings);

- Interest: in a particular sphere (football, music, etc.); preference – positive / negative attitude towards certain objects (football teams, composers, etc.)

NLP application

- It seems appropriate to use NLP:
  - Formulate and solve automatic interest detection task by user's written texts;
  - Can be applied for recommender systems and / or targeted advertising, esp. when personal user data is unavailable.

- Interest detection only (multi-class classification)
  - A pre-requisite for preference detection in an unstructured text (object-oriented sentiment analysis)

The Task

- A set of documents $D = \{d_1, \ldots, d_n\}$;
- A finite set of classes $C = \{c_1, \ldots, c_m\}$ - interests expressed in these documents:
  - Each document d has only one corresponding class c;
- Find a classification function f:
  - For any given pair $<d, c>$, determine, whether the document corresponds to the interest: $f: D \times C \rightarrow \{0, 1\}$

Existing classification methods

- Naïve Bayes [McCallum, Nigam, 1998];
- Logistic regression [Genkin et al., 2007];
- SVM [Joachims, 1998];
- Random Forest Classifier [Svetnik et al., 2003];
- CNN [Zhang et al., 2015]
- RCNN [Lai et al., 2015]
- LSTM [Zhou Ch. et al., 2015]
- Attention LSTM [Zhou P. et al., 2016]

Exiting representation methods

- Classical methods:
  - Bag-of-words;
  - Word-level and character-level N-grams;
  - Specific binary features;
  - Regular expression-based features;
- Embeddings:
  - Word2vec, Doc2vec, FastText;
  - BERT, ELMO

Dataset

- 209 630 text documents:
  - Web-forum messages: forum.kinopoisk.ru, www.livelib.ru;
- Ten classes: anime, art, books, food, films, football, games, music, nature, travel;
- 43% > 150 characters: 89844 texts;
- Average text length post-filtering – 427 characters
- Very imbalanced;
- Validation, Test sets:
  - 1000 texts each (100 random per category)

Per-class distribution

Class	Pre-filtering		Post-filtering	
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	

Text Preprocessing

- Stop-words, Latin characters, Mystem lemmatization.
- Before:
  - **Я кофе только со сливками пью.А чай я пью то-же только горячий если даже пару минут после кипения прошло,я снова его включаю:Кстати врач сказал,что такой горячий нельзя пить,но это уже бесполезно я уже зависим**
- After:
  - **кофе сливки пить чай пить горячий пара минута кипение проходить снова включать кстати врач сказать горячий пить это бесполезный зависеть**

Text Representations

- Doc2Vec: 300d, 600d;
- 10 x2 complex features:
  - Top PPMI words;
  - Top PPMI character trigrams.
- Examples:
  - FOOD: *аппетит* (appetite), *кофе* (coffee), *блюдо* (dish), *гарнир* (garnish); *'ыр '*, *'кеф'*, *'ощ '*, *'сыр'* (parts of food product names)
  - BOOKS: *слог* (author style), *паланик* (palahniuk), *книжный* (book), *роман* (novel); *'афк'*, *'гюг'*, *'фка'*, *'руэ'*, *'дюм'*, *'элш'*, *'амю'*, *'юма'*. (parts of famous writers' names)

Positive Pointwise Mutual Information [Bouma, 2009]

- The most pertinent words and character trigrams to use in class prediction;
- Feature values: the proportion of class-specific elements among all words/trigrams in a given text;
- 20 values: 10 for per-class words, 10 for character trigrams.

Classical Methods

Model (10 + 10)	Validation set	Test set
Gaussian NB	**0.643**	**0.627**
Linear SVC	0.537	0.540
K-Nearest	0.547	0.563
Random Forest Classifier – 100 e.	0.609	0.595
Voting Classifier (Gaussian, Linear SVC, Random Forest – 100)	0.602	0.601

Deep Learning: Feature Combinations

Model	Validation Set	Test Set
300, d2v: Bi-LSTM 100 – Dense 200 – Dense 100 – d/o 0.2	0.67	0.669
310, d2v-words	0.767	0.745
310, d2v-trigrams	0.756	0.722
320, d2v-words-trigrams: Bi-LSTM 100 – Dense 100, d/o 0.2	**0.796**	**0.785**

Deep Learning: Architectures

Model	Classification Accuracy (micro)
Gaussian Naïve Bayes	0.627
CNN: Conv1D 26	0.761
Feedforward: Dense 32 – Dense 32	0.771
Bidirectional LSTM 100 – Dense 100 – Dropout 0.2	0.785
LSTM 100 – Dense 200 – Dense 100 – Dropout 0.2	0.786

Studied dependencies

- Weighted vs Unweighted classes : 0.5-2% overall acc. growth;
- Top-300, Top-200, Top-100 PPMI words / trigrams: the less, the better (0.5-2.5% acc. growth);
- 300d vs 600d D2V: 300d – 1% acc. Growth;
- Most problematic classes: Nature, Travel, Art.

Conclusions

- User interest text classification task formulated;
- Multi-class dataset collected;
- First results obtained:
  - 0.786 acc.: LSTM, 300 + 20;
- Means of improvement:
  - Additional data;
  - Data rebalancing;
  - Different representations: 3D, ELMO, BERT;
  - Error analysis

References

1. McCallum A., Nigam K. A comparison of event models for naive bayes text classification // AAAI-98 workshop on learning for text categorization, vol. 752, no. 1, 1998. P. 41-48.
2. Genkin A., Lewis D., Madigan D. Large-scale Bayesian logistic regression for text categorization // Technometrics 49, no. 3 (2007). P. 291-304.
3. Joachims Th. Text categorization with support vector machines: Learning with many relevant features // In European conference on machine learning, Springer, Berlin, Heidelberg, 1998. P. 137-142.
4. Svetnik V., Liaw A., Tong C., Culberson J., Sheridan R., Feuston B. Random forest: a classification and regression tool for compound classification and QSAR modeling // Journal of chemical information and computer sciences 43, no. 6 (2003). P. 1947-1958.

References

5. Zhang X., Zhao J., LeCun Y. Character-level convolutional networks for text classification // In Advances in neural information processing systems. 2015. P. 649-657.
6. Lai S., Xu L., Liu K., Zhao J. Recurrent convolutional neural networks for text classification // In Twenty-ninth AAAI conference on artificial intelligence. 2015.
7. Zhou Ch., Sun Ch., Liu Zh., Lau F. A C-LSTM neural network for text classification // arXiv preprint arXiv:1511.08630 (2015).
8. Zhou P., Shi W., Tian J., Qi Zh., Li B., Hao H., Xu B. Attention-based bidirectional long short-term memory networks for relation classification // In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, 2016. P. 207-212.
9. https://github.com/Pythonimous/forum-classifier (Last access 16.05.2019)

Classification of User Interests by Text Messages with Recurrent Neural Networks

Alexey Malafeev, Kirill Nikolaev,
National Research University Higher School of Economics,
Nizhny Novgorod, 2019