Classification of User Interests by Text Messages: Method Survey

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

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

- Difference from existing datasets (e.g. Taiga: https://tatianashavrina.github.io/taiga_site/)
  - Pure social media (forums);
  - Short texts
  - Taiga: 2% social media vs 77% literary tests

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)

Dataset stats

Class	Pre-deletion texts	Post-deletion texts	Sentence tokens	Word tokens	Avg text length	Avg sentence length
Anime	7663	3213	13945	188424	59	14
Art	2216	1175	7270	98215	84	14
Books	18008	9999	60237	863698	86	14
Films	12961	5862	28679	417130	71	15
Food	11751	5866	30786	386307	66	13
Football	62397	23234	111951	1400314	60	13
Games	67282	29550	151834	2001324	68	13
Music	21637	7974	37931	497097	62	13
Nature	2578	1057	4465	60754	57	14
Travel	3137	1914	15212	193070	101	13

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, RNNMorph 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 сливка пить чай пить горячий пара минута кипение пройти 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 (10+10)

Model	Accuracy	Precision	Recall	F1
Random Forest Classifier – 100 e.	0.56	0.56	**0.72**	0.55
Multinomial Naïve Bayes	0.64	0.64	0.66	0.65
LinearSVC	0.65	0.65	0.65	0.64
Voting Classifier (the 3 above)	**0.66**	**0.66**	0.66	**0.66**
Random Forest Classifier – 100 e.	0.56	0.56	**0.72**	0.55

Classical Methods (BoW, 100)

Model	Accuracy	Precision	Recall	F1
ComplementNB	**0.71**	**0.71**	0.75	0.70
LinearSVC	**0.71**	**0.71**	**0.80**	**0.73**
Random Forest Classifier	0.61	0.61	0.74	0.62
Voting Classifier (gauss – linsvc – rfc 100)	0.70	0.70	0.76	0.71
ComplementNB	**0.71**	**0.71**	0.75	0.70

Deep Learning: Feature Combinations

Model	Bi-LSTM (LSTM200-D200-D100, d/o 0.2)	Feedforward (32-32)
300: Doc2vec	0.700	0.669
310, d2v-words	0.755	0.760
310, d2v-trigrams	0.681	0.704
320, d2v-words-trigrams	**0.781**	**0.775**

Deep Learning: Architectures

Model	Accuracy	Precision	Recall	F1
CNN: Conv1D – 26 filters, kernel size 10, Leaky ReLU rectifier	0.771	0.77	0.78	0.77
Feedforward: Dense 32 units – Dense 32 units	0.775	**0.78**	0.78	**0.78**
LSTM 100 – Dense 200 units – Dense 100 units – Dropout 0.2	0.777	**0.78**	0.78	**0.78**
Bidirectional LSTM 200 – Dense 200 units - Dense 100 units – Dropout 0.2	**0.781**	**0.78**	0.78	**0.78**
CNN: Conv1D – 26 filters, kernel size 10, Leaky ReLU rectifier	0.771	0.77	0.78	0.77

Best Results

Model	Accuracy	Precision	Recall	F1
LinearSVC (BoW 100)	0.71	0.71	0.80	0.73
Bidirectional LSTM 200 – Dense 200 units - Dense 100 units – Dropout 0.2	0.781	0.78	0.78	0.78
Fasttext (calibrated; 20000 seconds calibrating, training 34 minutes (after cal)	**0.85**	**0.85**	**0.85**	**0.84**

Studied dependencies

- Weighted vs Unweighted classes : 0.6-2.7% overall acc. growth;
  - Undersampling – smallest classes are too small;
  - Oversampling – require synonimical words / more texts (can't tweak vectors directly)
- Top-300, Top-200, Top-100 PPMI words / trigrams: the less, the better (~1-1.5% 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:
  - 0.85 acc.: FastText
- Means of improvement:
  - Additional data;
  - Data rebalancing;
  - Different representations: ELMO, BERT;
  - Reformulate task as topic modelling (detection instead of classification)

Classification of User Interests by Text Messages: Method Survey

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