Wan ZHOU, Yong WANG, Cuiyun GAO, Fei YANG, 2022. Emerging topic identification from app reviews via adaptive online biterm topic modeling. *Frontiers of Information Technology & Electronic Engineering*, 23(5):678-691. https://doi.org/10.1631/FITEE.2100465

Emerging topic identification from app reviews via adaptive online biterm topic modeling

Key words: App reviews; Emerging topic identification; Topic model; Natural language processing

Corresponding author: Yong WANG
E-mail: yongwang@ahpu.edu.cn
ORCID: https://orcid.org/0000-0002-2719-1017

Motivation

1. Emerging topics in app reviews highlight the topics (e.g., software bugs) with which users are concerned during certain periods. Identifying emerging topics accurately, and in a timely manner, could help developers more effectively update apps.

2. The accuracy of emerging topic identification is reduced because reviews are short in length and offer limited information.

Main idea

1. The pre-training model (such as BERT) can be applied to text error correction by fine-tuning.

2. Adaptive online biterm topic model (AOBTM) can effectively capture topic distributions in short texts of different time slices, and outlier detection can identify outliers (i.e., emerging topics) from the topic distribution of texts.

3. Emerging topics are labeled by the related phrases and sentences, and the effectiveness of emerging topics can be evaluated by the similarity between official app changelogs and labels.

Method

An improved emerging topic identification (IETI) approach is proposed which consists mainly of three phases:

1. Reduce noisy data in app reviews through a series of natural language preprocessing methods.

2. Identify emerging topics in app reviews by AOBTM and outlier detection methods.

3. Use the relevant phrases and sentences to label emerging topics, and evaluate the effectiveness of emerging topics by official app changelogs.

Method (Cont'd)

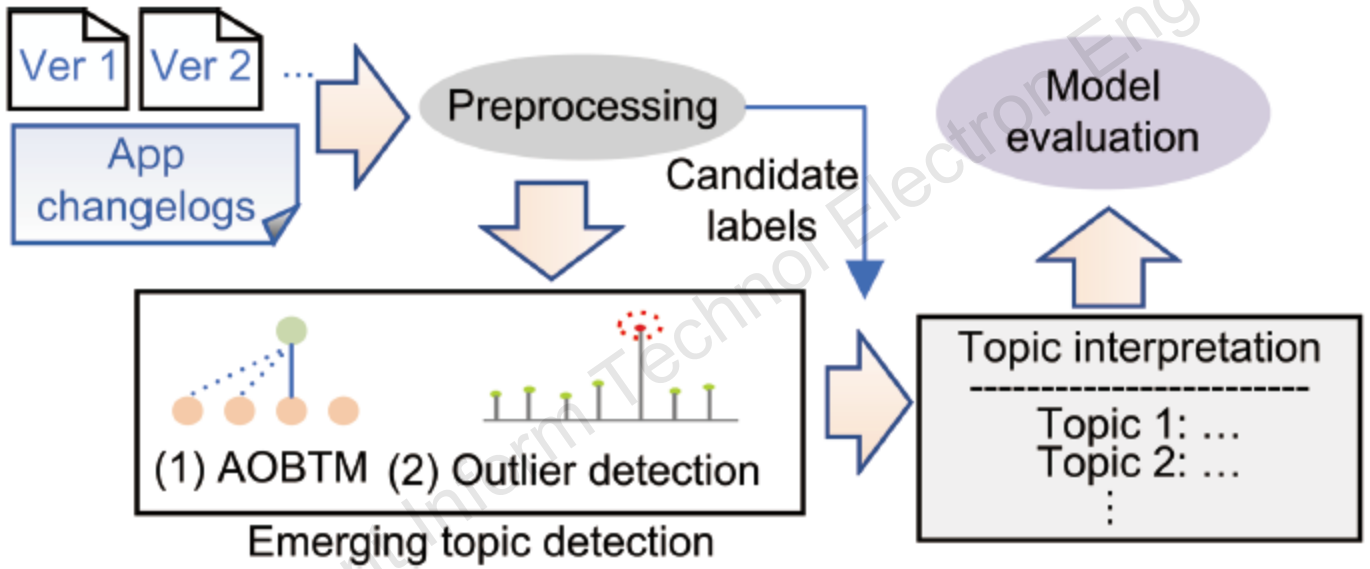

Fig. 4 Overview of the improved emerging topic identification (IETI)

Major results

Table 3 Comparison results of emerging topics identified by different models

App name	Method	Precision		Recall		F1 score	
		Phrase	Sentence	Phrase	Sentence	Phrase	Sentence
YouTube	OLDA	0.441	0.578	0.462	0.664	0.451	0.597
	IDEA	0.592	0.628	0.472	**0.666**	0.523	0.636
	IETI	**0.674**	**0.719**	**0.527**	0.625	**0.592**	**0.669**
Viber	OLDA	0.157	0.313	0.305	0.550	0.166	0.375
	IDEA	0.625	0.625	0.340	**0.651**	0.440	0.638
	IETI	**0.658**	**0.694**	**0.475**	0.642	**0.552**	**0.667**
SwiftKey	OLDA	0.100	0.367	0.567	0.617	0.148	0.458
	IDEA	0.517	0.583	**0.653**	**0.700**	0.523	0.587
	IETI	**0.565**	**0.622**	0.592	0.693	**0.578**	**0.656**
Clean master	OLDA	0.300	0.200	0.269	0.421	0.160	0.129
	IDEA	0.667	0.667	0.318	0.434	0.431	0.526
	IETI	**0.683**	**0.695**	**0.562**	**0.582**	**0.617**	**0.634**
Ebay	OLDA	0.167	0.500	0.238	0.488	0.196	0.494
	IDEA	0.229	0.646	0.251	0.527	0.227	0.580
	IETI	**0.574**	**0.693**	**0.434**	**0.572**	**0.494**	**0.627**
NOAA radar	OLDA	0.468	0.482	0.528	0.622	0.473	0.534
	IDEA	0.571	0.476	0.497	0.639	0.531	0.546
	IETI	**0.612**	**0.607**	**0.582**	**0.651**	**0.597**	**0.628**

Best results are in bold

Major results (Cont'd)

Table 4 Ablation study results of different models

App name	Method	Precision		Recall		F1 score	
		Phrase	Sentence	Phrase	Sentence	Phrase	Sentence
YouTube	IDEA	0.592	0.628	0.472	**0.666**	0.523	0.636
	IDEA$^+$	0.621	0.684	0.481	0.634	0.542	0.658
	IETI$^-$	0.643	0.652	0.501	0.639	0.563	0.645
	IETI	**0.674**	**0.719**	**0.527**	0.625	**0.592**	**0.669**
Viber	IDEA	0.625	0.625	0.340	0.651	0.440	0.638
	IDEA$^+$	0.641	0.637	0.396	0.644	0.490	0.641
	IETI$^-$	0.637	0.652	0.425	**0.667**	0.501	0.659
	IETI	**0.658**	**0.694**	**0.475**	0.642	**0.552**	**0.667**
SwiftKey	IDEA	0.517	0.583	**0.653**	**0.700**	0.523	0.587
	IDEA$^+$	0.531	0.572	0.567	0.684	0.545	0.623
	IETI$^-$	0.557	0.617	0.574	0.677	0.566	0.646
	IETI	**0.565**	**0.622**	0.592	0.693	**0.578**	**0.656**
Clean master	IDEA	0.667	0.667	0.318	0.434	0.431	0.526
	IDEA$^+$	0.643	0.681	0.487	0.526	0.554	0.594
	IETI$^-$	0.671	0.680	0.523	0.547	0.588	0.601
	IETI	**0.683**	**0.695**	**0.562**	**0.582**	**0.617**	**0.634**
Ebay	IDEA	0.229	0.646	0.251	0.527	0.227	0.580
	IDEA$^+$	0.471	0.659	0.341	0.543	0.401	0.595
	IETI$^-$	0.525	0.668	0.382	0.552	0.442	0.604
	IETI	**0.574**	**0.693**	**0.434**	**0.572**	**0.494**	**0.627**
NOAA radar	IDEA	0.571	0.476	0.497	0.639	0.531	0.546
	IDEA$^+$	0.585	0.523	0.526	0.640	0.552	0.578
	IETI$^-$	0.571	0.574	0.554	0.642	0.562	0.606
	IETI	**0.612**	**0.607**	**0.582**	**0.651**	**0.597**	**0.628**

Best results are in bold

Major results (Cont'd)

Table 5 Emerging topic detection results of NOAA radar in version 1.7

IDEA	IETI
Topic 7	Topic 1
Lightning strike, 0.545	Extend forecast, 0.435
Extend forecast, 0.447	Loop speed, 0.419
Waste money, 0.353	Pay money, 0.365
Topic 9	Topic 8
Loop speed, 0.481	Extend forecast, 0.422
Extend forecast, 0.417	Loop speed, 0.412
Show nothing, 0.4123	Cloud cover, 0.385

The numbers represent the similarity scores of phrase labels

Conclusions

1. An improved emerging topic identification (IETI) approach was proposed.

2. IETI reduces noisy data in app reviews through a series of preprocessing methods, including correcting misspelled words and restoring abbreviations.

3. IETI uses the topic model AOBTM to capture the topic distribution of app reviews, which overcomes the influence of the sparse feature of short text on topic identification.

Wan ZHOU received his BS degree in Computer Science and Technology from Huaibei Normal University, in 2019. He is currently pursuing the MS degree in Software Engineering with Anhui Polytechnic University. His research interests include app store analysis, natural language processing, and program debugging.

Yong WANG received his BS and MS degrees in Computer Science from Anhui Polytechnic University and his PhD degree in Computer Science and Technology from Nanjing University of Aeronautics and Astronautics. His current research interests include data mining, natural language analysis of software, and program debugging.

Cuiyun GAO received her BE degree from the Department of Communication Engineering, Shanghai University, in 2014, and her PhD degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China, in 2018. She is currently an assistant professor with Harbin Institute of Technology (Shenzhen). Her research interests include user experience study, natural language processing, and mobile app analysis.

Fei YANG received his BS and MS degrees in Computer Science from Shanghai Jiao Tong University, and his PhD degree in Computer Science from Eindhoven University of Technology, the Netherlands. His current research interests include deep learning, machine learning systems, and concurrency theory.