Identifying Influential Users' Professions via the Microblogs They Forward

Yuan Wang, Hangyu Mao, Zhen Xiao
Peking University, China
SocInf2017, 19 Aug 2017

Outline

- Background
- Data
- Our Method
- Evaluation
- Conclusion

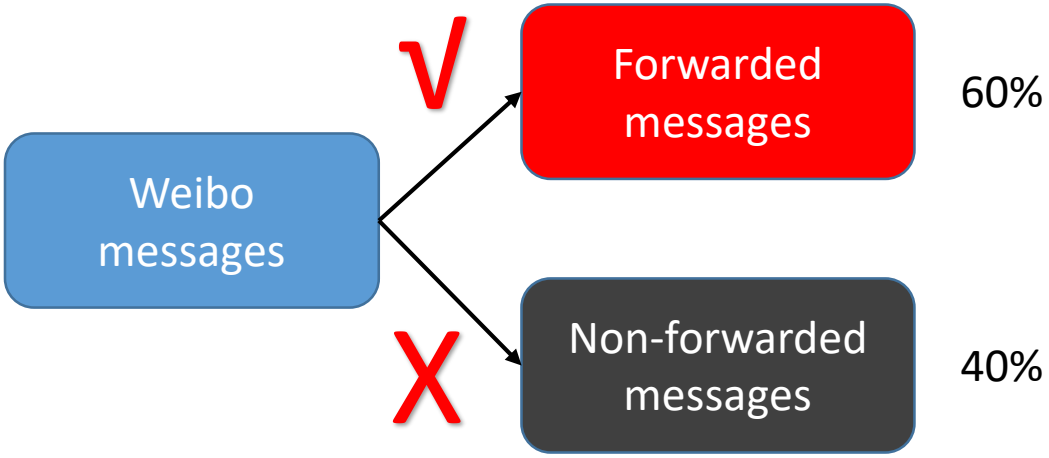
Introduction

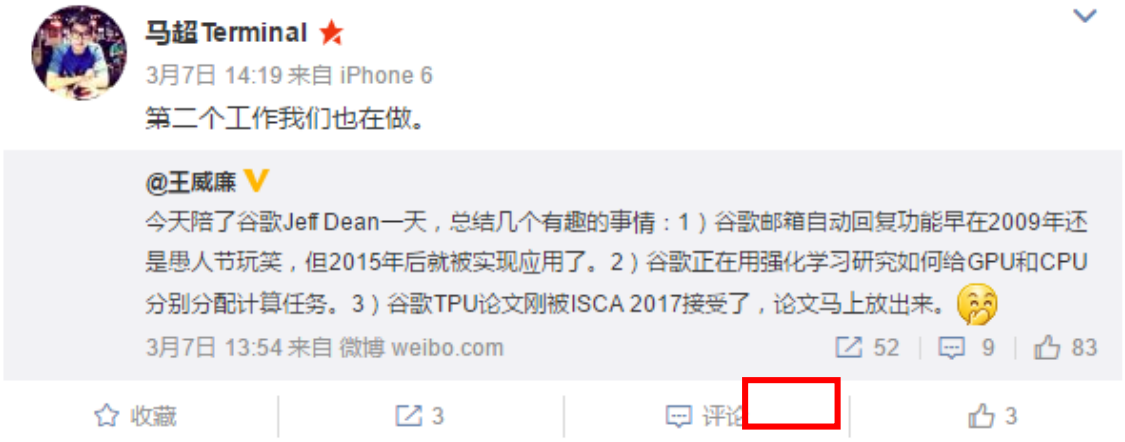
Profession



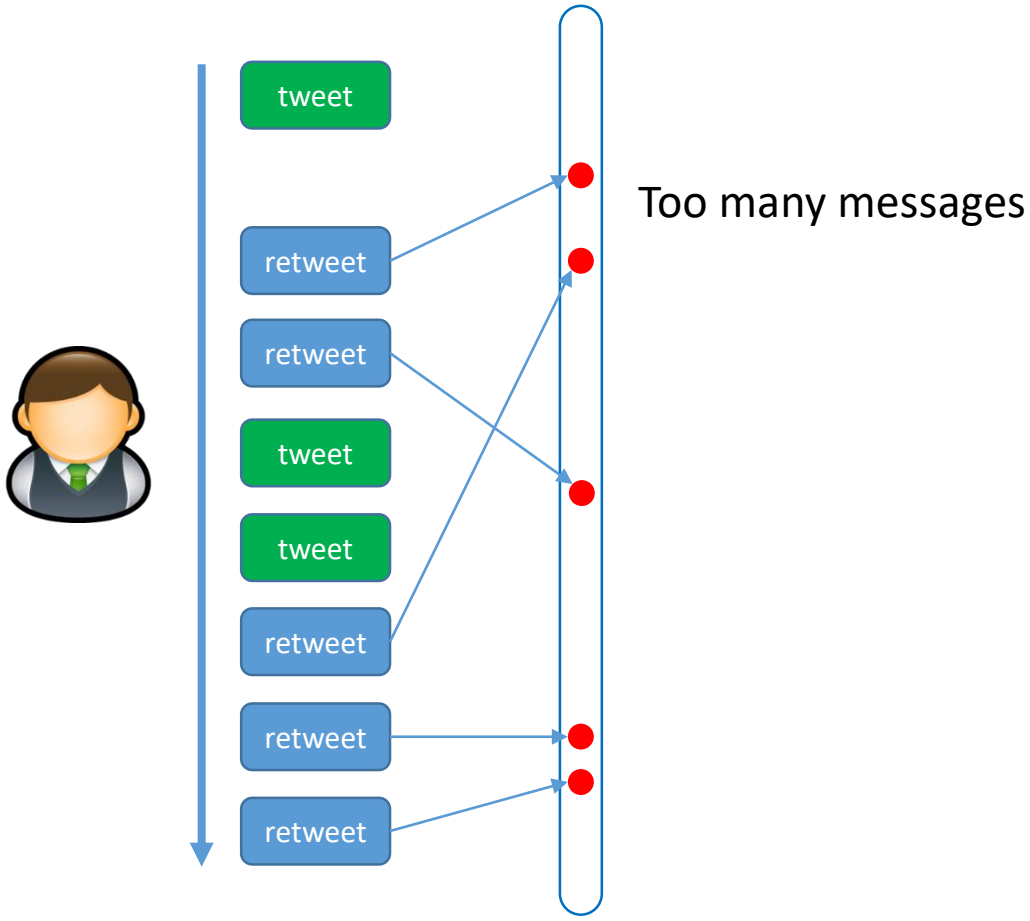

users are mainly organized by their professions

Message forwarding

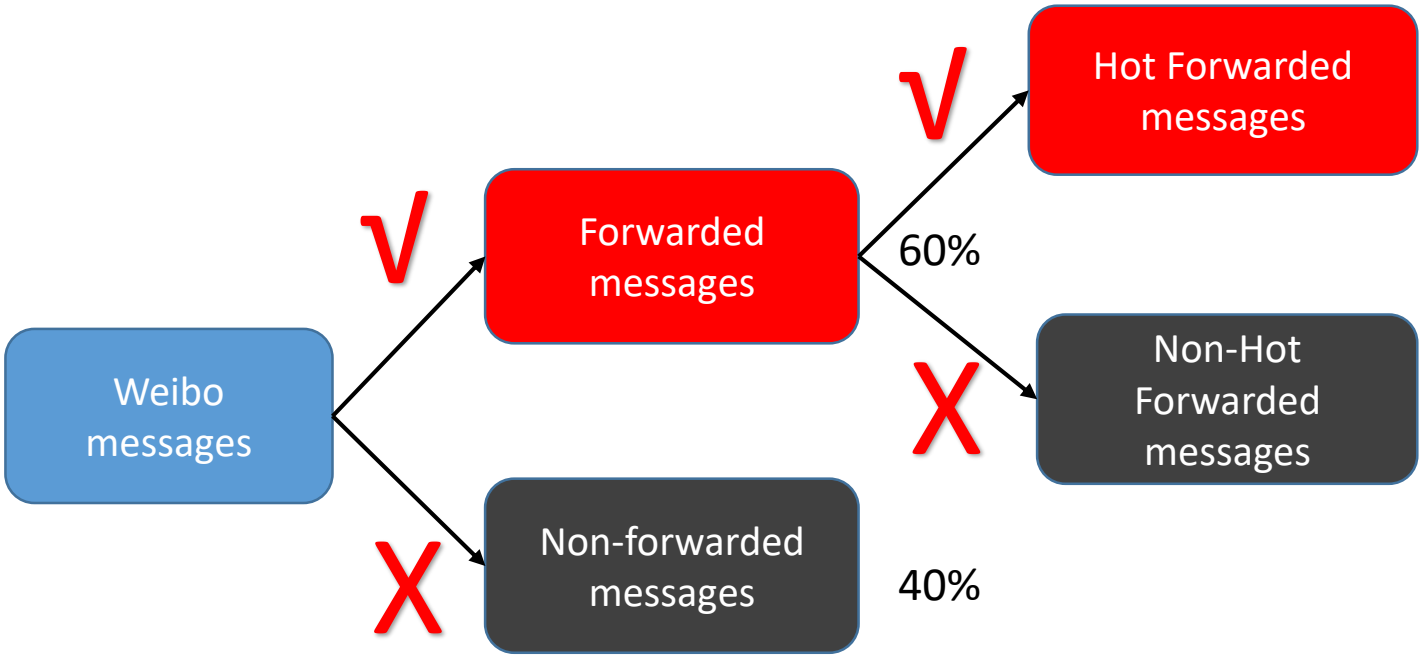




Challenge


tweet
retweet
retweet
tweet
tweet
retweet
retweet
retweet
Too many messages

Challenge


Weibo messages
√
Forwarded messages
X
Non-forwarded messages
40%
√
Hot Forwarded messages
60%
X
Non-Hot Forwarded messages

Our Framework

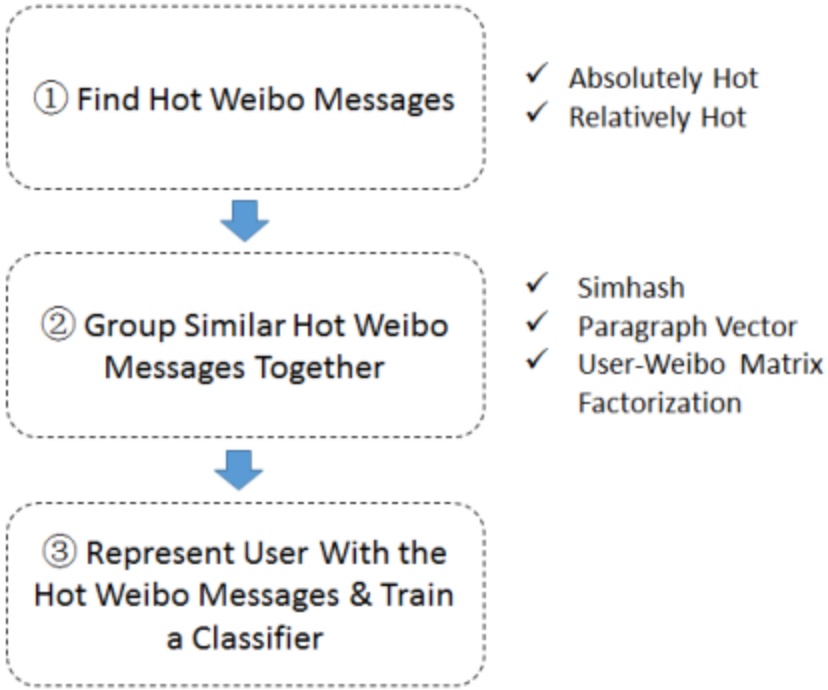


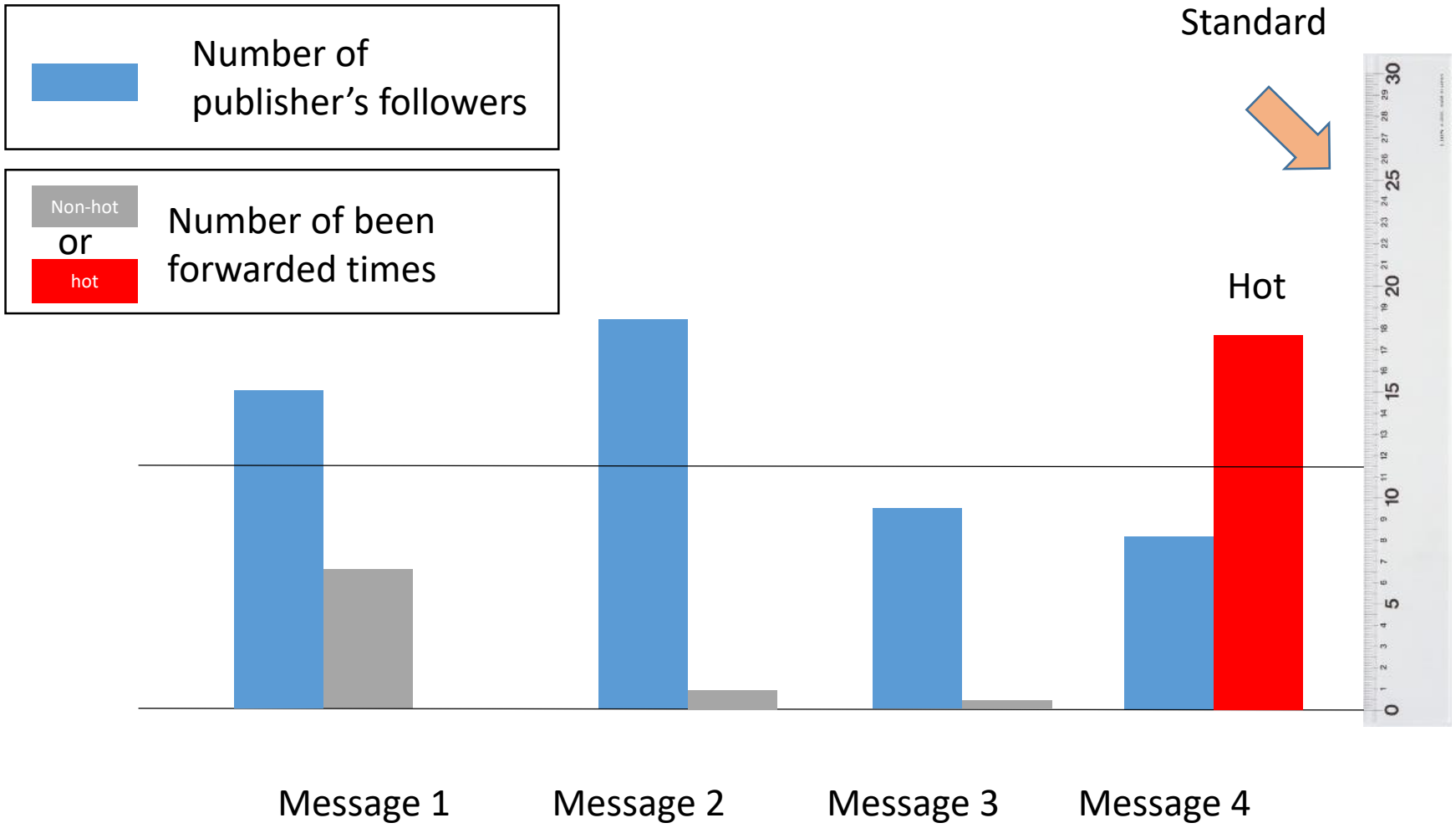
Fig. 1. The framework of PIFB

Our Data

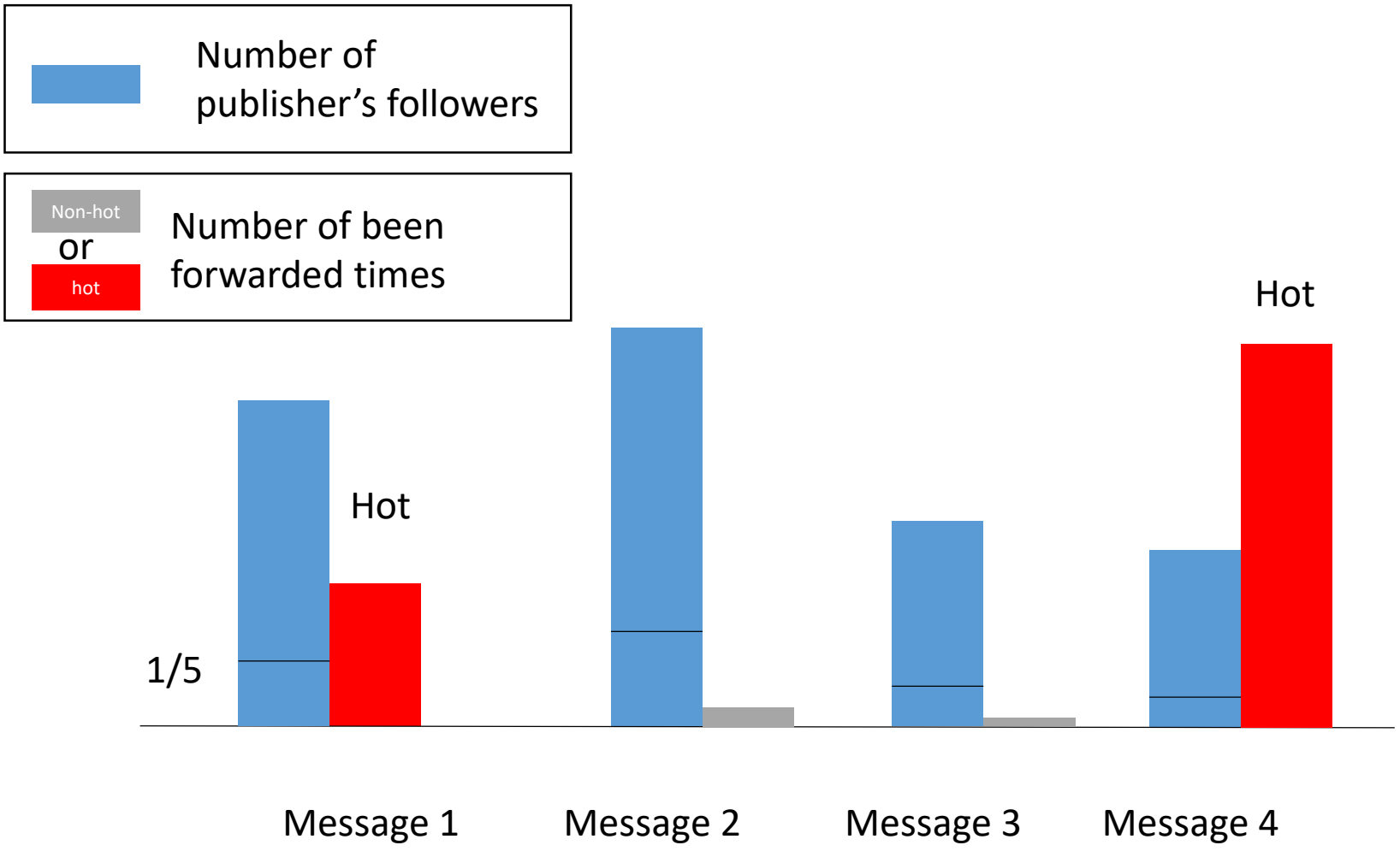
No.	Category	Ratio (%)
1	Media	26.3
2	Entertainment	10.1
3	Estate	9.1
4	Finance	8.6
5	Government	8.5
6	IT	8.4
7	Sports	6.4
8	Fashion	6.2
9	Education	5.9
10	Literature	5.4
11	Game	5.1

41,531 manually annotated influential users

Absolutely Hot Message



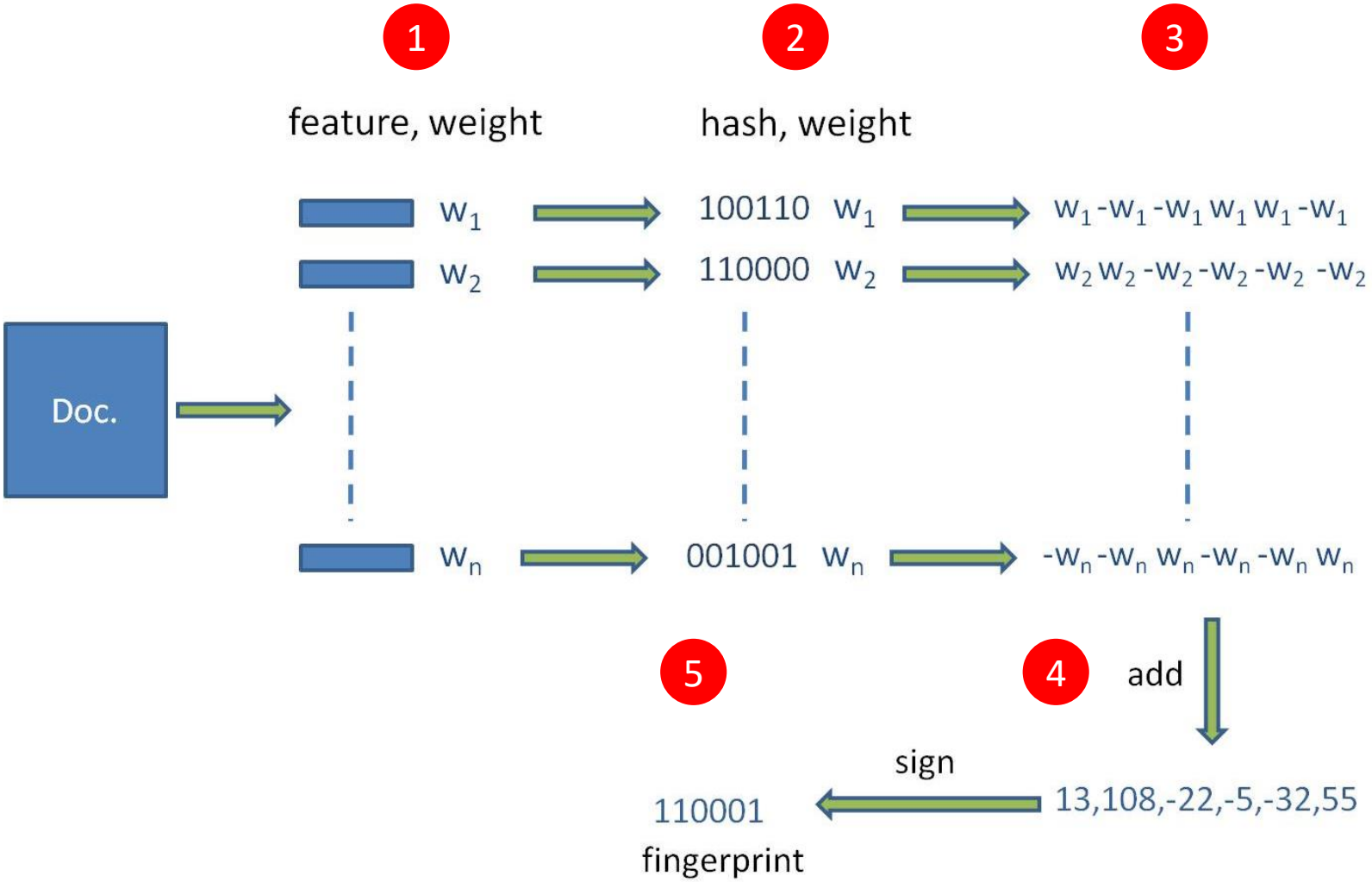
Relatively Hot Message



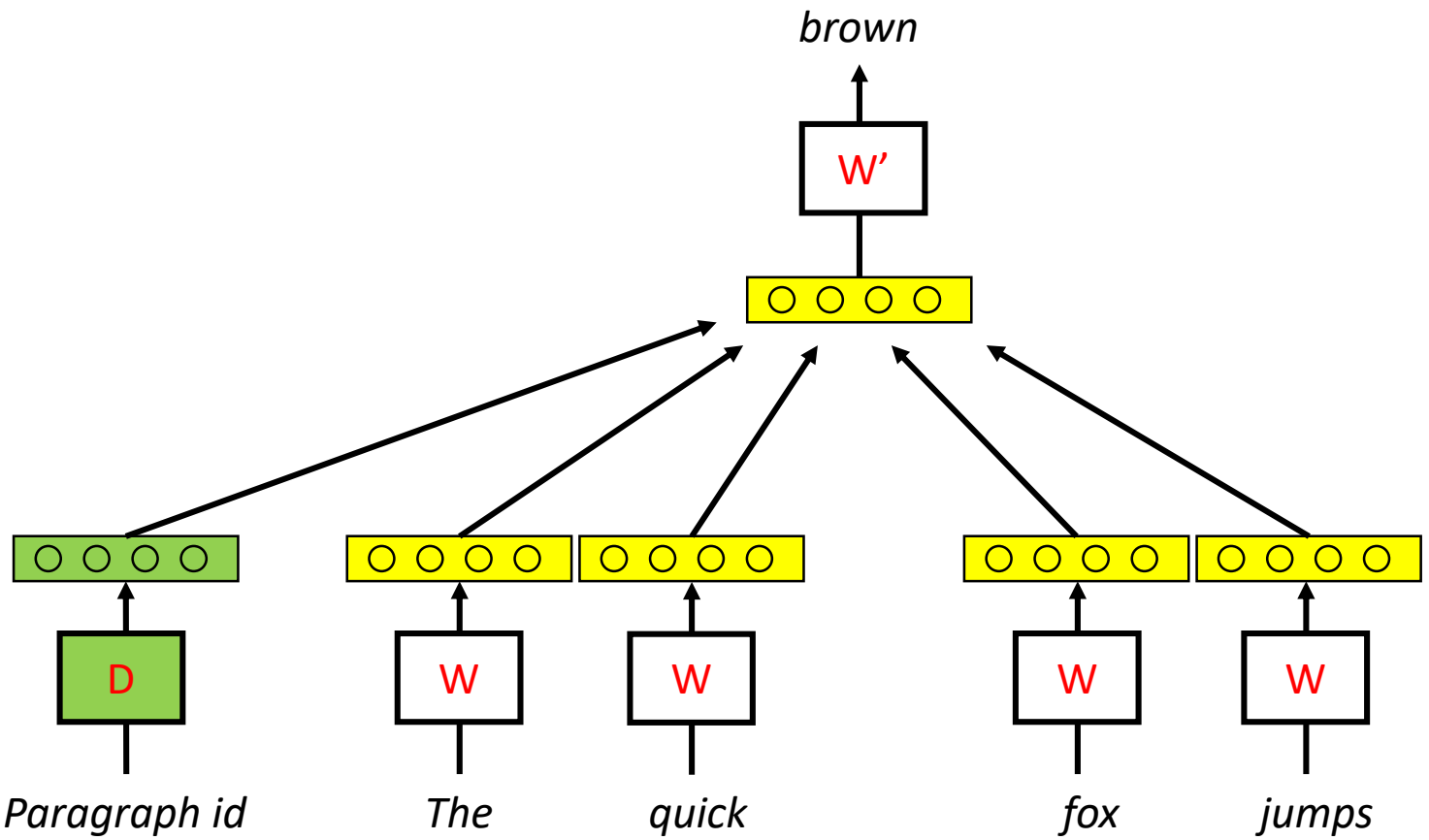
Group Similar Hot Messages Together

- Simhash
- Paragraph2vec
- Matrix factorization

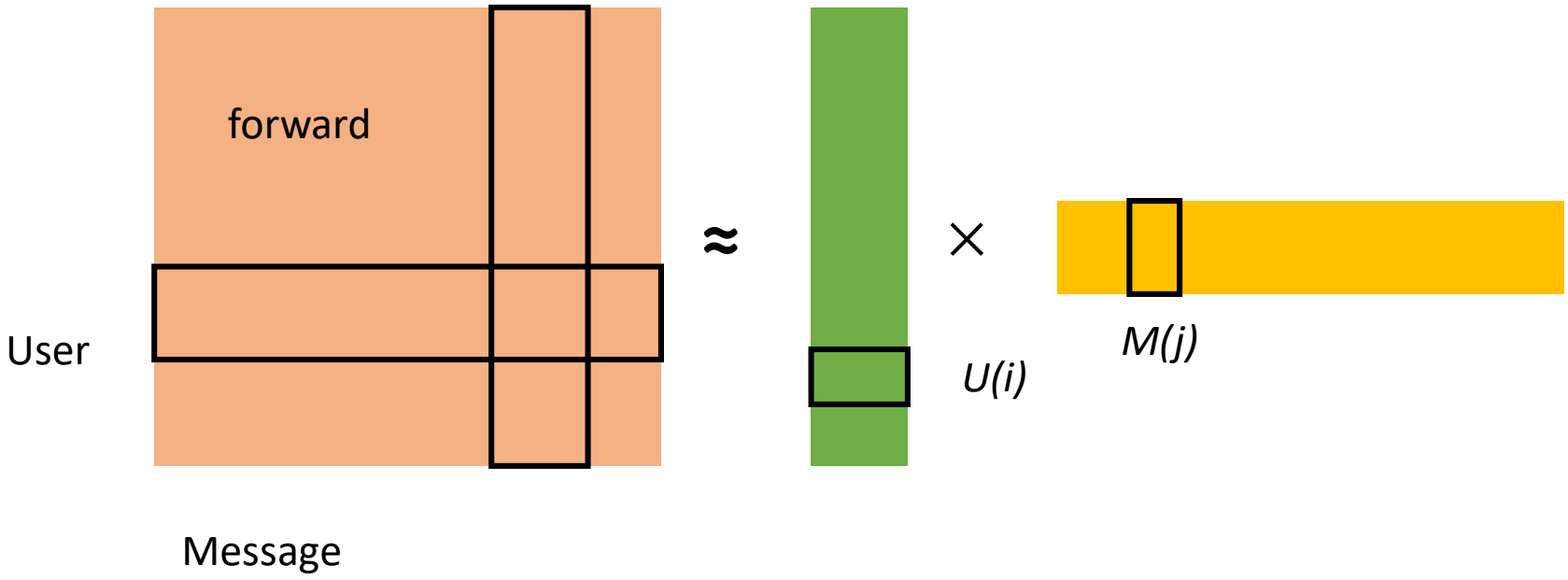
Simhash


1
2
3
feature, weight
hash, weight
Doc.
w1
w2
wn
100110 w1
110000 w2
001001 wn
w1 -w1 -w1 w1 w1 -w1
w2 w2 -w2 -w2 -w2 -w2
-wn -wn wn -wn -wn wn
4
add
5
sign
13,108,-22,-5,-32,55
110001
fingerprint

Paragraph2Vec

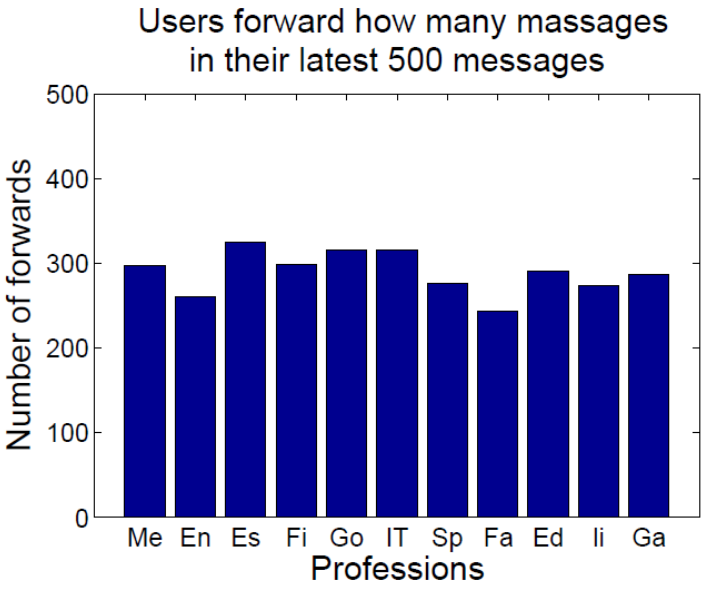


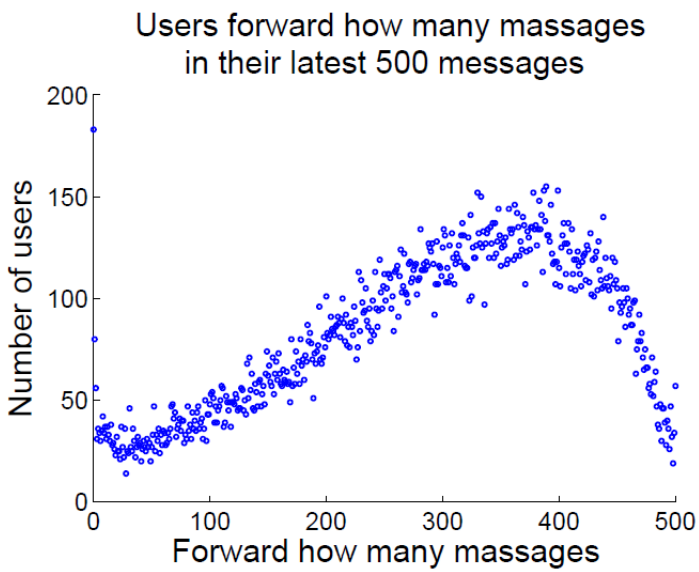
User-Weibo Matrix Factorization



User-weibo relationship matrix M

Evaluation


Users forward how many massages
in their latest 500 messages
Number of forwards
500
400
300
200
100
0
Me
En
Es
Fi
Go
IT
Sp
Fa
Ed
Ii
Ga
Professions


Users forward how many massages
in their latest 500 messages
Number of users
200
150
100
50
0
0
100
200
300
400
500
Forward how many massages

Evaluation

- Absolute hot
- Relative hot

Table 2. Number of absolutely hot messages

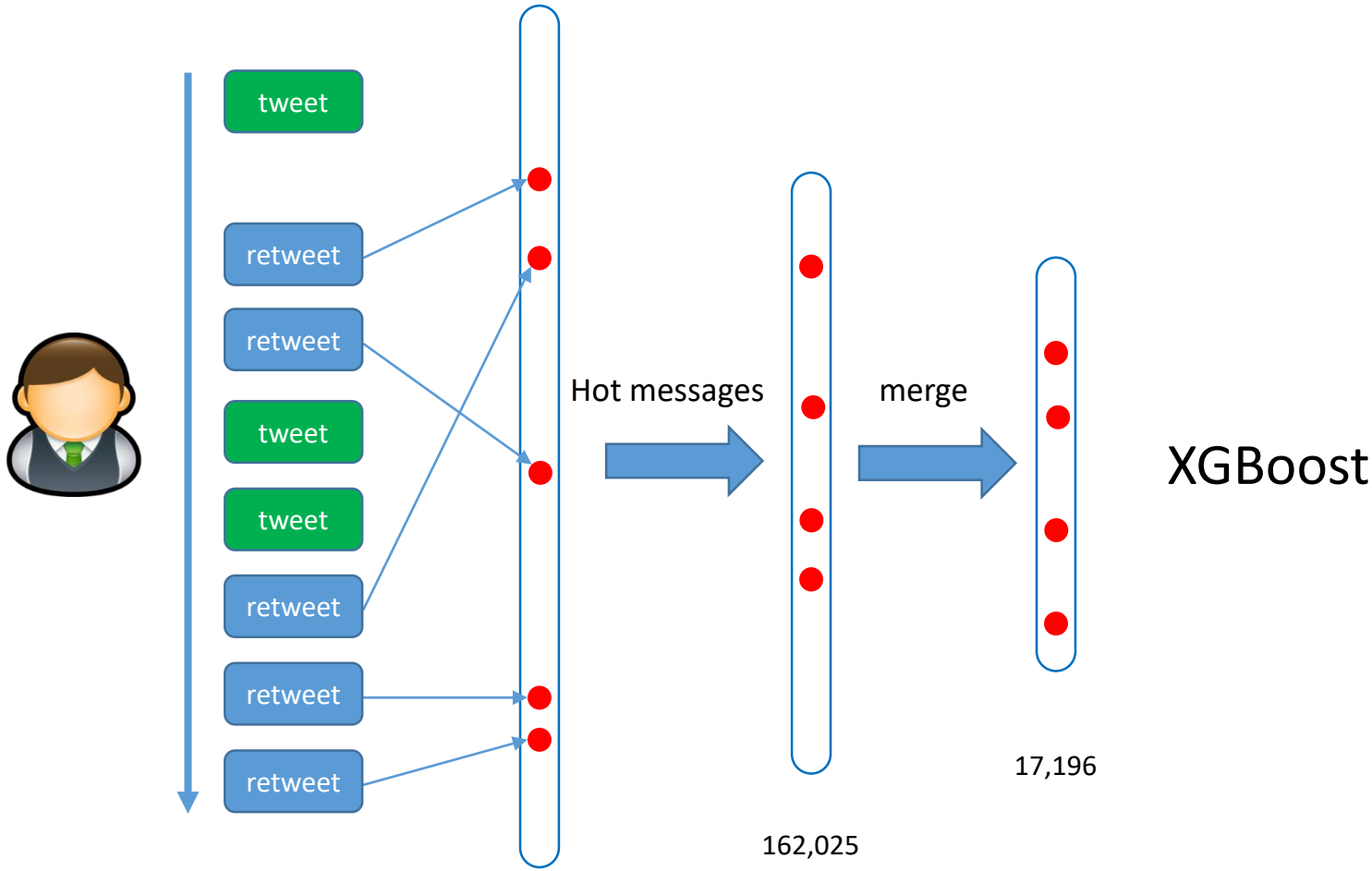
No.	Threshold	# before filter	# after filter
1	500	731,150	100,219
2	2000	426,019	82,339
3	10000	74,308	32,955

Evaluation

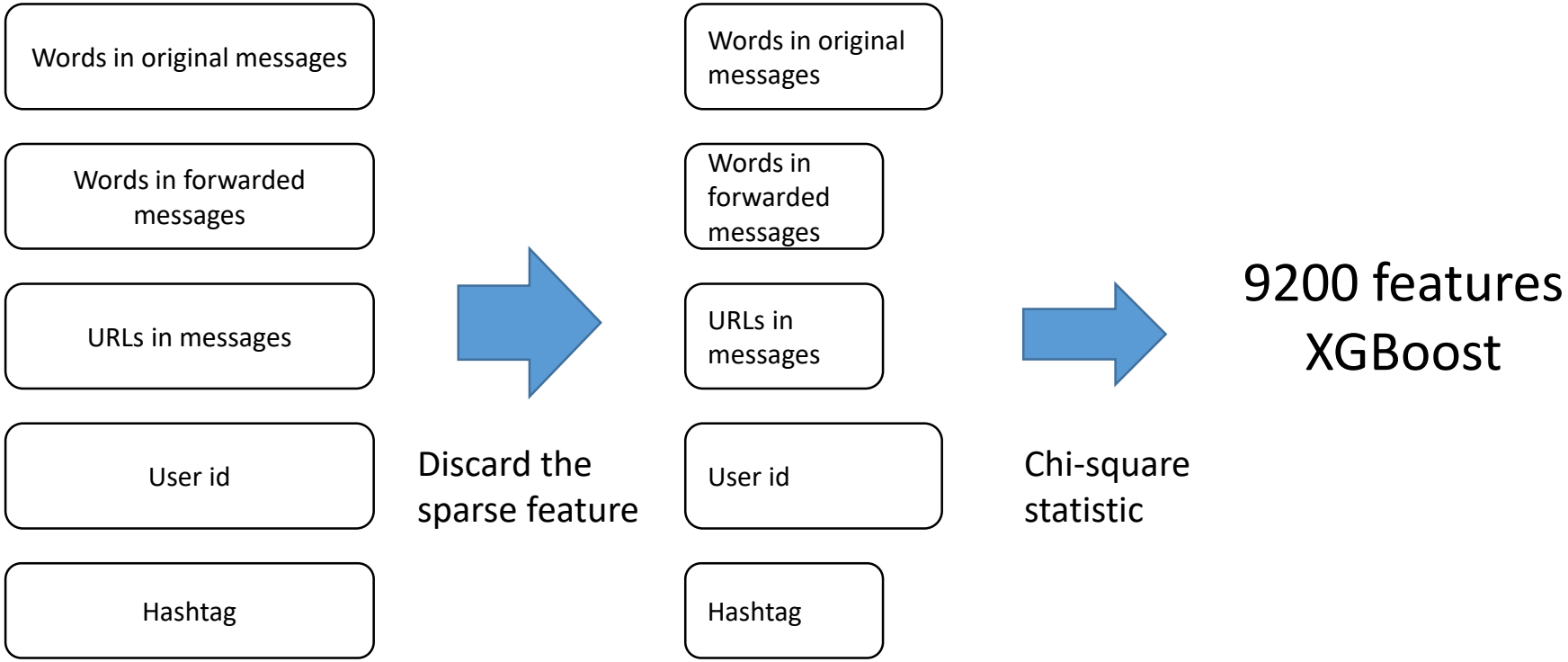
Table 3. Number of messages under different merging strategies

No.	Merging Strategy	# before	# after
1	Simhash	162,025	57,624
2	P2V	162,025	32,118
3	UWMF	162,025	27,129
4	Simhash+P2V+UWMF	162,025	17,196

Evaluation


tweet
retweet
retweet
tweet
tweet
retweet
retweet
retweet
Hot messages
merge
XGBoost
17,196
162,025

Baseline


Words in original messages
Words in forwarded messages
URLs in messages
User id
Hashtag
Discard the sparse feature
Words in original messages
Words in forwarded messages
URLs in messages
User id
Hashtag
Chi-square statistic
9200 features
XGBoost

Evaluation

No.	Method	Accuracy	Precision	Recall	F1
1	Baseline	62.38	64.03	60.29	62.10
2	Simhash	69.24 ↑ 6.86%	70.88	67.61	69.21 ↑ 7.11%
3	Simhash+P2V	73.79 ↑ 11.41%	73.90	71.28	72.57 ↑ 10.47%
4	Simhash+P2V+UWMF	73.98 ↑ 11.60%	74.81	72.95	73.87 ↑ 11.77%

Table 4: Evaluation results for various features and combinations. (%)

Evaluation

	Me	En	Es	Fi	Go	IT	Sp	Fa	Ed	Li	Ga
Me	76.7	5.6	2.7	3.5	4.1	3.3	2.2	0.9	0.6	0.2	0.2
En	7.2	74.5	0.2	3.3	0.7	1.4	4.4	5.1	0.2	1.3	1.7
Es	7.4	2.0	72.9	8.5	5.3	2.2	0.4	0.9	0.1	0.0	0.3
Fi	8.4	0.1	6.4	70.2	5.3	6.2	0.2	1.3	1.7	0.1	0.1
Go	4.9	2.2	0.4	4.2	78.2	2.9	4.1	0.4	2.5	0.2	0.0
IT	6.1	0.7	3.9	4.3	1.3	76.3	0.2	0.1	2.6	0.7	3.8
Sp	5.1	2.9	0.0	0.3	0.3	1.0	86.2	2.2	0.7	0.0	1.3
Fa	9.7	14.9	1.0	6.2	0.2	0.0	3.3	61.5	0.9	1.2	1.1
Ed	5.2	3.9	3.3	4.6	2.0	3.2	0.7	1.8	68.4	4.2	2.7
Li	13.7	7.2	0.7	1.3	0.6	1.4	0.4	3.3	9.8	60.9	0.7
Ga	5.2	3.9	0.8	0.0	0.4	7.1	1.2	4.3	0.1	0.3	76.7

Table 5: Distribution of identified professions in each profession.

Conclusion

- find the hot weibo messages
- group similar hot messages together
- design a classifier to predict users' professions

Questions