哈爾濱工業大學深圳研究生院
Harbin Institute of Technology Shenzhen Graduate School

A Joint Model for Chinese Microblog Sentiment Analysis

Yuhui Cao, Zhao Chen, Ruifeng Xu, Tao Chen

Harbin Institute of Technology, Shenzhen Graduate School

Content

Task: Topic-Based Chinese Message Polarity Classification

Task Description:

- Classify the message into positive, negative, or neutral sentiment towards the given topic.
- For messages conveying both a positive and negative sentiment towards the topic, whichever is the stronger sentiment should be chosen.

Introduction

Task Characteristics:

- Real and noise data
- Imbalance data between classes
- Short but meaningful message

Examples:

- 好看？吗？//【Galaxy S6：三星证明自己能做出好看的手机】

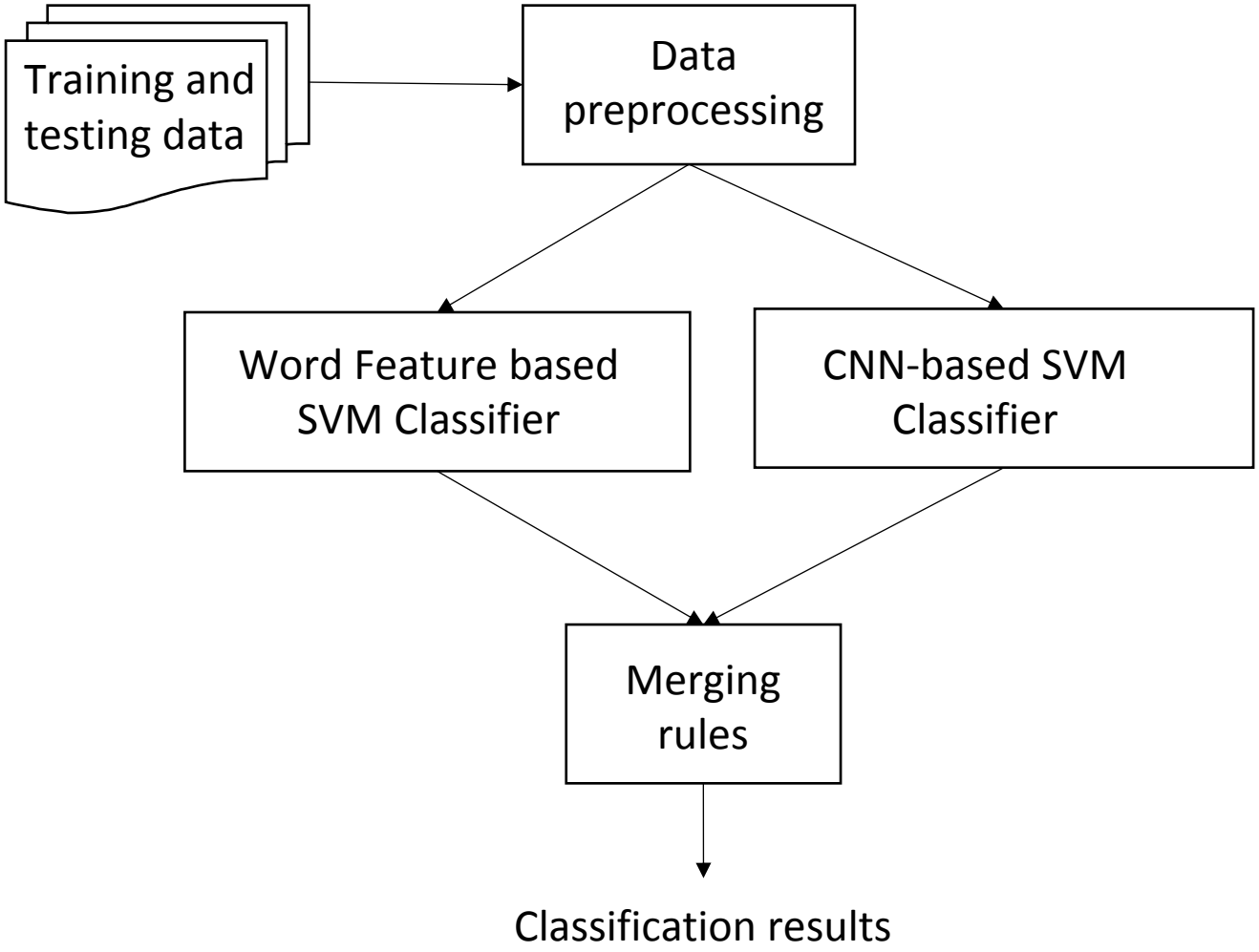
http://t.cn/RwHRsIb(分享自 @ 今日头条)

- # 三星 Galaxy S6# 三星 GALAXY S6 三星，挺中意 [酷][酷] [位置] 芒砀路
- 雾霾是什么？面对纯蓝的天，相机失焦了。 [位置]北门街

Framework of our model

- Data preprocessing: rule-based process
- Word feature based SVM classifier: unigram + bigram + sentiment words
- CNN-based SVM classifier: word embedding + convolutional neural network
- Integrated strategy: multi-classifier results fusion

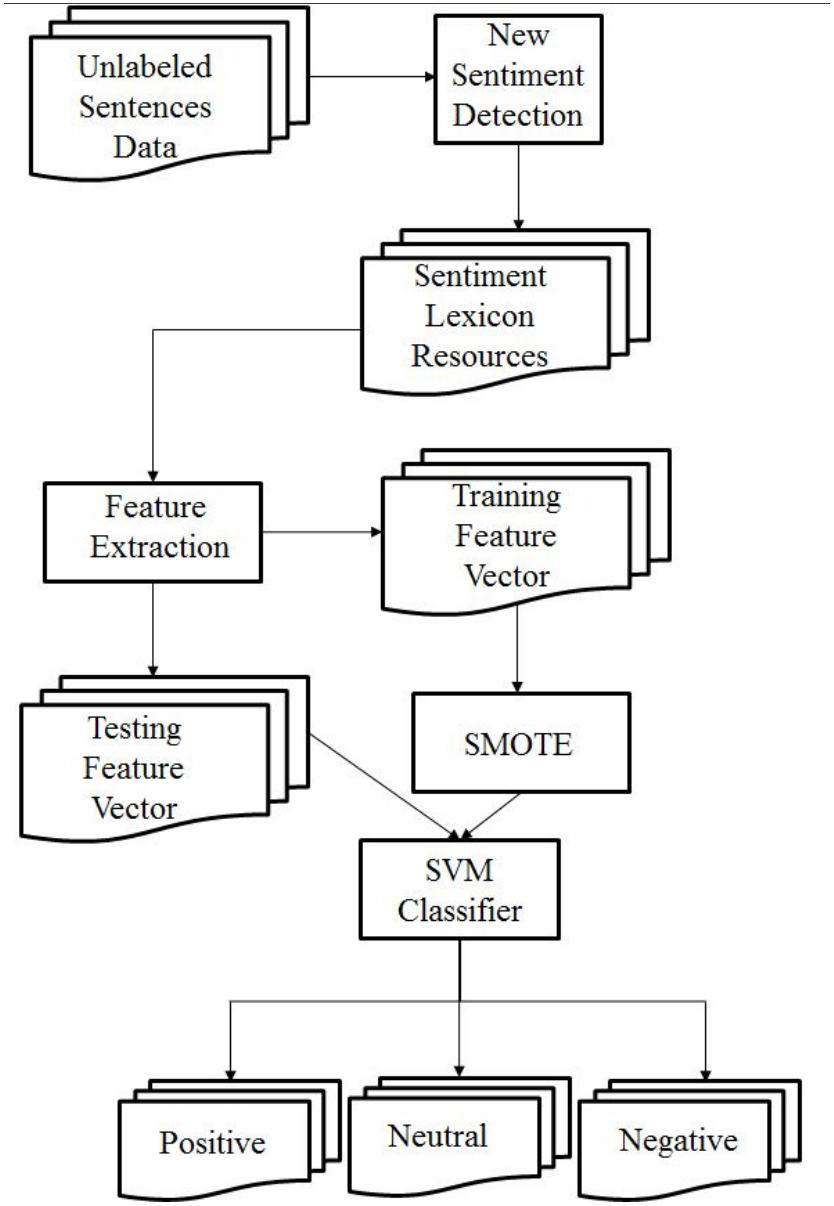
Framework of our model



Data preprocessing rules with illustrations

Rules	Raw Text	Processed Text
Sharing news with personal comments	好看？吗？ //【Galaxy S6：三星证明自己能做出好看的手机】http://t.cn/RwHRsIb (分享自 @今日头条)	好看？吗？
Removing HashTag	#三星 Galaxy S6# 三星GALAXY S6，挺中意[酷][酷] [位置]芒砀路	三星 GALAXY S6，挺中意[酷][酷]
Removing URL	699欧元起 传三星Galaxy S6/S6 Edge售价获证实 （分享自 @新浪科技）http://t.cn/RwTo3on	699 欧元起 传三星Galaxy S6/S6 Edge 售价获证实（分享自 @新浪科技）
Removing nickname	玻璃取代塑料，更美 Galaxy S6 的 5 大妥协 http://t.cn/RwHY6Az罗永浩 我去小米和三星这是要闹哪样，，，老罗。。不能忍啊，，，，，@锤子科技营销帐号 @罗永浩	http://t.cn/RwHY6Az 罗永浩 我去小米和三星这是要闹哪样，，，老罗。。不能忍啊，，，，，
Removing information sources	【视频：三星S6 对比 苹果iPhone6 MWC2015 @youtube 科技 ~】http://t.cn/RwHQzJ8（来自于优酷安卓客户端）	【视频：三星S6 对比 苹果iPhone6 MWC2015 @youtube 科技 ~】http://t.cn/RwHQzJ8

Framework


Unlabeled Sentences Data
New Sentiment Detection
Sentiment Lexicon Resources
Feature Extraction
Training Feature Vector
Testing Feature Vector
SMOTE
SVM Classifier
Positive
Neutral
Negative

Word Feature based Classifier

Sentiment Lexicon expansion: To expand existing sentiment lexicon, POS tags, word frequency, mutual information and context entropy are used to mine the new sentiment word from twenty million microblog text.

Positive Words	Negative Words
人气王，亮骚，人气爆棚	人渣，吐槽，坑爹，仆街
卖萌，傲娇，傲娇，共赢	伤退，伪娘，作孽，做空
典藏版，劲爆，劲歌热舞	偷腥，偷食，傻冒，傻叉
力挺，牛逼，完爆，给力	傻帽，傻缺，利空，劳神
炫酷，靠谱，重磅，利好	卖腐，厚黑，脑残，无语

Word Feature based Classifier

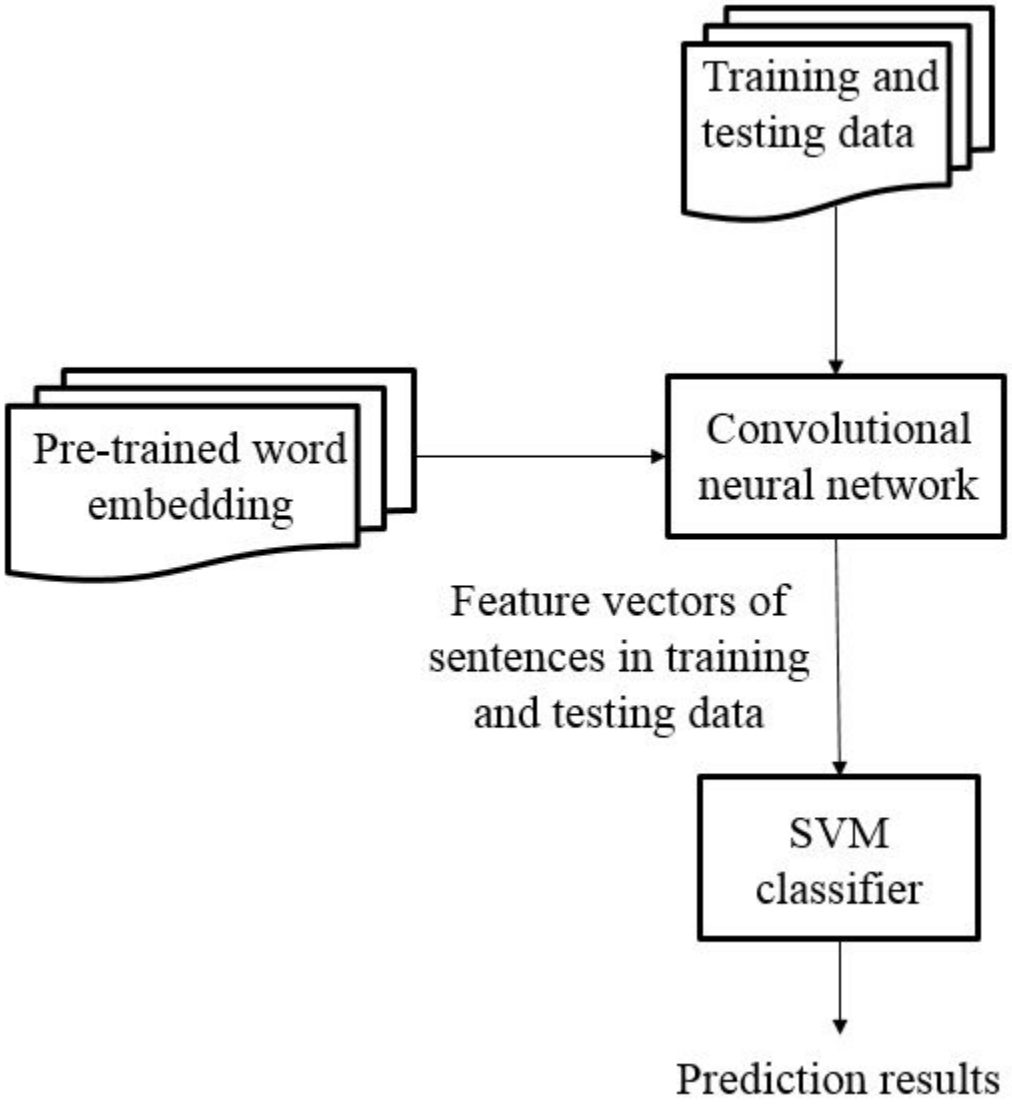
Word features: unigram, bigram, uni-part-of-speech, bi-part-of-speech, sentiment lexicons

Features Selection Methods: CHI-test, TF-IDF

Imbalance Data Problem: use SMOTE algorithm to undersampling the major class and oversampling the minor classes.

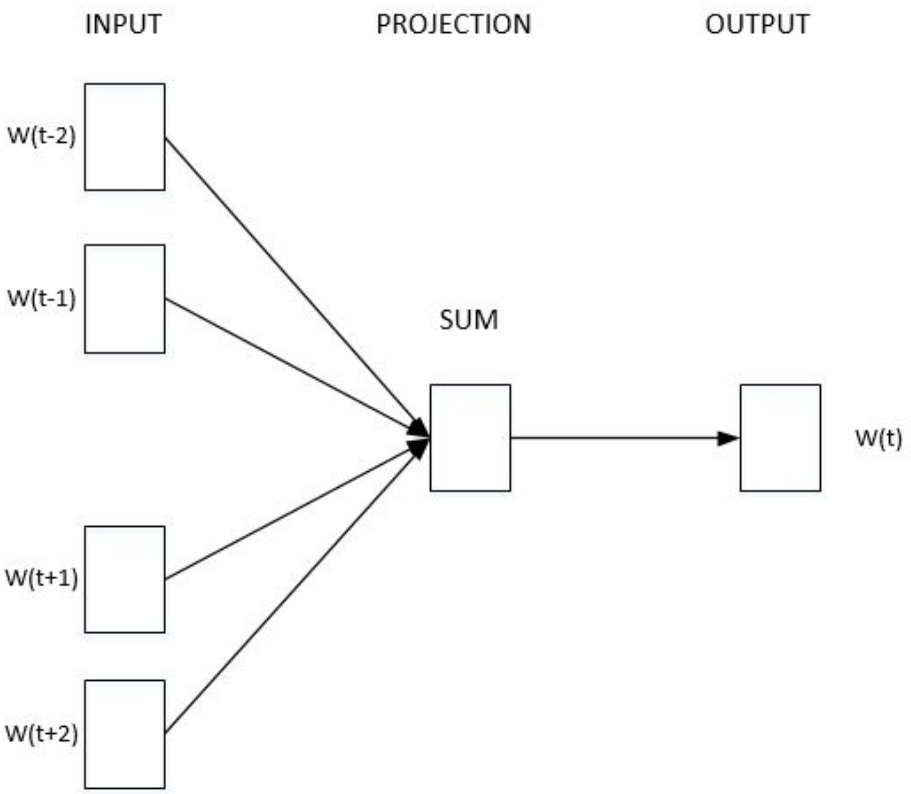
Classifier: SVM with linear kernel

CNN-based SVM Classifier



1. Word embedding

- Train the CBOW model using 16GB Chinese microblog text
- Obtain 200-dimension word embeddings for Chinese microblog text



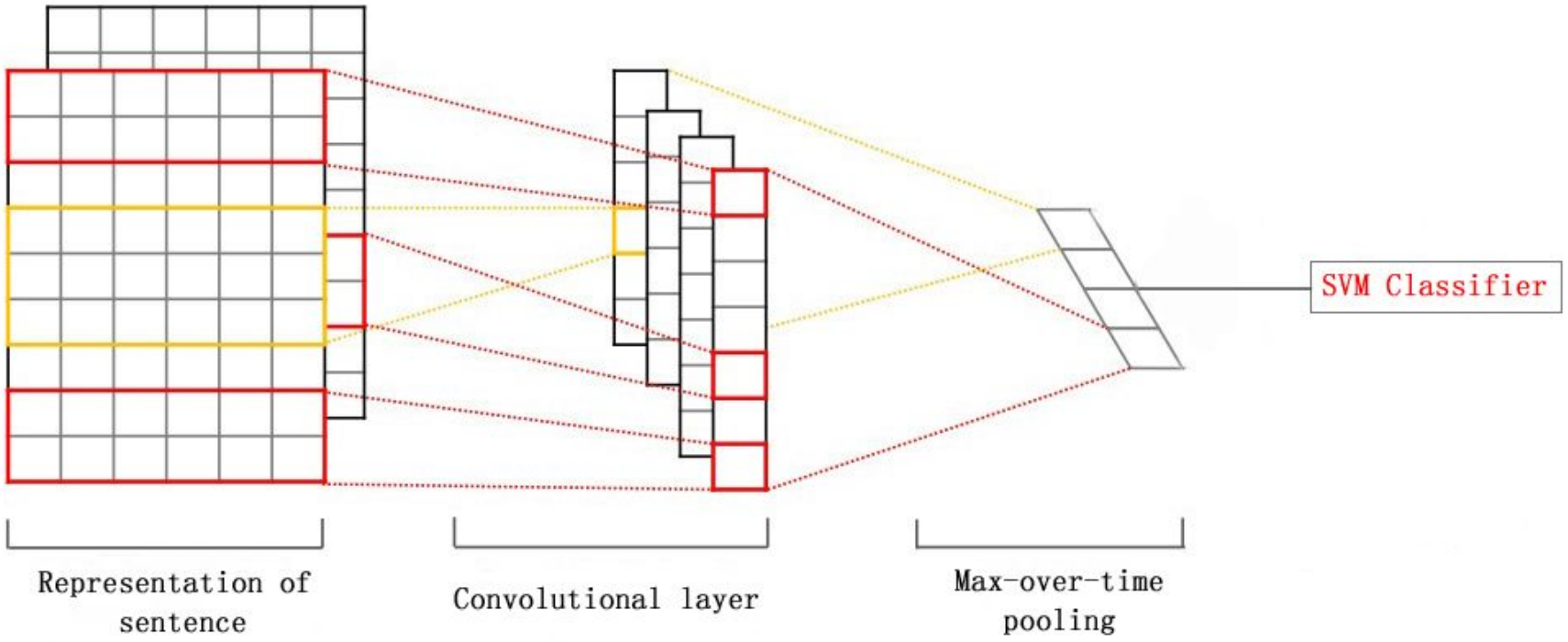
2. CNN-based SVM classifier

Input: a matrix which is composed of the word embeddings of microblogs

Features: use CNN to constitute the distributed paragraph feature representation

Classifier: SVM with linear kernel

2. CNN-based SVM classifier



Outputs merging

- Two classification outputs are the same
  =>The final output is the same
- Two classification outputs are different
  =>The final result is determined from the merge rules
  These rules are based on the statistical analysis on the individual classifier performances on training dataset.

Final result	Classifier 1	Classifier 2
neutral	positive	neutral
neutral	negative	neutral
neutral	neutral	positive
neutral	neutral	negative
negative	positive	negative
positive	negative	positive

Experiments

- Data set

  Training data: 4905 microblogs (394 positive, 538 negative and 3973 neutral), 5 topics

  Testing data: 19469 microblogs, 20 topics

- Metrics

$$Precision = \frac{System.Correct}{System.Output}$$

$$Recall = \frac{System.Correct}{Human.Labeled}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Performances in restricted resource subtask

	All			Positive			Negative		
Team Name	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
TICS-dm	0.83	0.83	0.83	0.62	0.51	0.56	0.82	0.46	0.59
NEUDM2	0.74	0.74	0.74	0.31	0.08	0.13	0.44	0.08	0.13
LCYS_TEAM	0.72	0.64	0.68	0.26	0.05	0.09	0.40	0.10	0.16
HLT_HITSZ	**0.68**	**0.68**	**0.68**	**0.21**	**0.40**	**0.28**	**0.45**	**0.60**	**0.52**

Performances in unrestricted resource subtask

	All			Positive			Negative		
Team Name	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
TICS-dm	0.85	0.85	0.85	0.58	0.62	0.60	0.79	0.61	0.69
xk0	0.74	0.74	0.74	0.19	0.01	0.03	0.40	0.05	0.09
NEUDM1	0.74	0.74	0.74	0.26	0.11	0.16	0.46	0.33	0.38
HLT_HITSZ	**0.71**	**0.71**	**0.71**	**0.24**	**0.41**	**0.30**	**0.51**	**0.54**	**0.53**

Experiments

Performances by different classifiers in unrestricted resource subtask

	Neutral			Positive			Negative		
Approach	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Classifier 1	0.67	0.67	0.67	0.20	0.42	0.27	0.44	0.49	0.46
Classifier 2	0.60	0.60	0.60	0.18	0.61	0.28	0.42	0.67	0.52
Merging	0.71	0.71	0.71	0.24	0.41	0.30	0.51	0.54	0.53

Conclusion

- Data preprocessing
- Word feature based SVM classifier
- CNN-based SVM classifier
- Integrated strategy

- Second rank on micro average F1 value
- Fourth rank on macro average F1 value

Q&A

哈爾濱工業大學深圳研究生院
Harbin Institute of Technology Shenzhen Graduate School

A Joint Model for Chinese Microblog Sentiment Analysis

Thanks