31st International Conference on Tools with Artificial Intelligence

Design and Research of Composite Web Page Classification Network Based on Deep Learning

Zhao Qiuhan

School of Cyberspace Security

Beijing University of Posts and Telecommunications

Overview

- Demand Analysis
- Related work and Challenge
- Approach
- Experiments
- Conclusion

Overview

- Demand Analysis
- Related work and Challenge
- Approach
- Experiments
- Conclusion

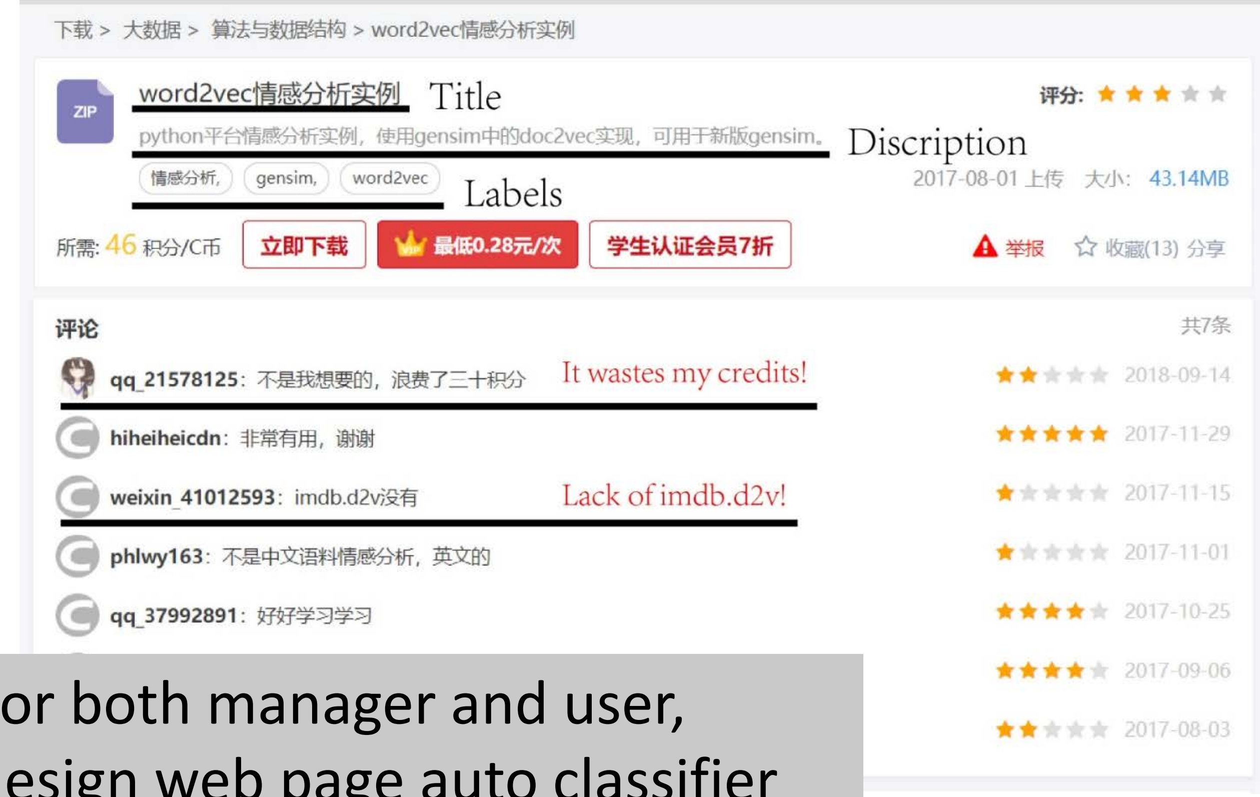
Demand Analysis

- The total number of domestic websites has reached 37.93 million. [*China Internet Network Information Center+ 2019*]

- For **website manager**
  - High labor costs
  - Rely on professional knowledge for classification

Demand Analysis

- For **user**
  - Unable to get the information you want quickly



下载 > 大数据 > 算法与数据结构 > word2vec情感分析实例

word2vec情感分析实例 Title

python平台情感分析实例，使用gensim中的doc2vec实现，可用于新版gensim。 Discription

情感分析, gensim, word2vec Labels

评分: ★★★☆☆

2017-08-01 上传 大小: 43.14MB

所需: 46 积分/C币

评论 共7条

qq_21578125：不是我想要的，浪费了三十积分 It wastes my credits! ★★☆☆☆ 2018-09-14

hiheiheicdn：非常有用，谢谢 ★★★★★ 2017-11-29

weixin_41012593：imdb.d2v没有 Lack of imdb.d2v! ★☆☆☆☆ 2017-11-15

phlwy163：不是中文语料情感分析，英文的 ★☆☆☆☆ 2017-11-01

qq_37992891：好好学习学习 ★★★★☆ 2017-10-25

★★★★☆ 2017-09-06

★★☆☆☆ 2017-08-03

For both manager and user,
design web page auto classifier

Overview

- Demand Analysis
- Related work and Challenge
- Approach
- Experiments
- Conclusion

Related work and Challenge

- Text-based:

  KNN [*Lin+ 2011*], SVM [*Xu+ 2011*], NN [*Kim+ 2014, Lai+ 2015, Yu+ 2018*]

The page contains ***irrelevant information*** increasingly, such as advertisement, Link and recommendation etc. These noises will greatly ***interfere with feature extraction and reduce the accuracy*** of classification.

- URL,HTML etc.-based:

  Neighboring webpage [*Qi+ 2006*], Url [*Yang+ 2016*]

Slower, ignoring the web ***body text***.

Overview

- Demand Analysis
- Related work and Challenge
- Approach
- Experiments
- Conclusion

Approach

- Idea

For complex networks

- ***Two different branches*** respectively target the short text information of the webpage and the long text information.

Fight against the Noisy

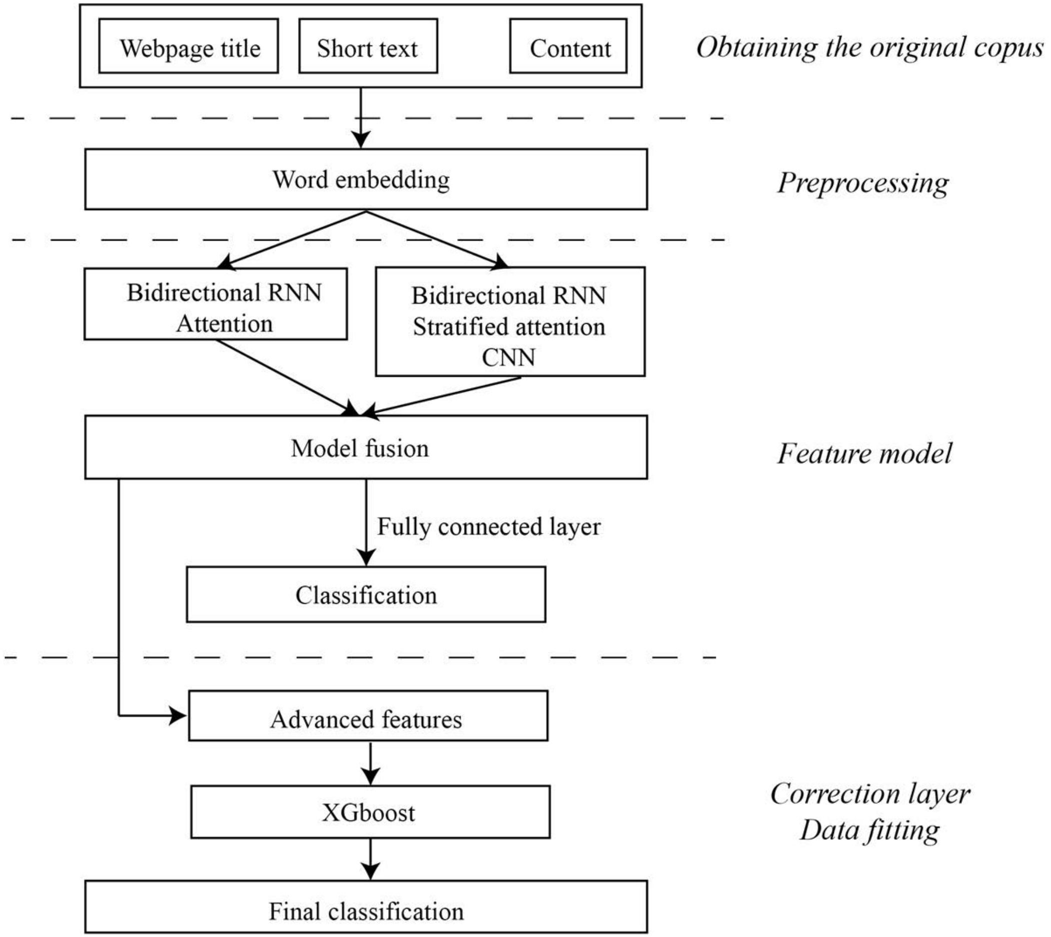
- Introducing Attention in short extracted network.
- Introducing Attention separately in word and paragraph level of long extracted network.

Integrated learning for better accuracy

- Using XGboost as the ***correct layer***.

Approach

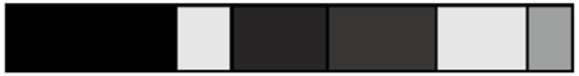
- Overview of our approach



Approach

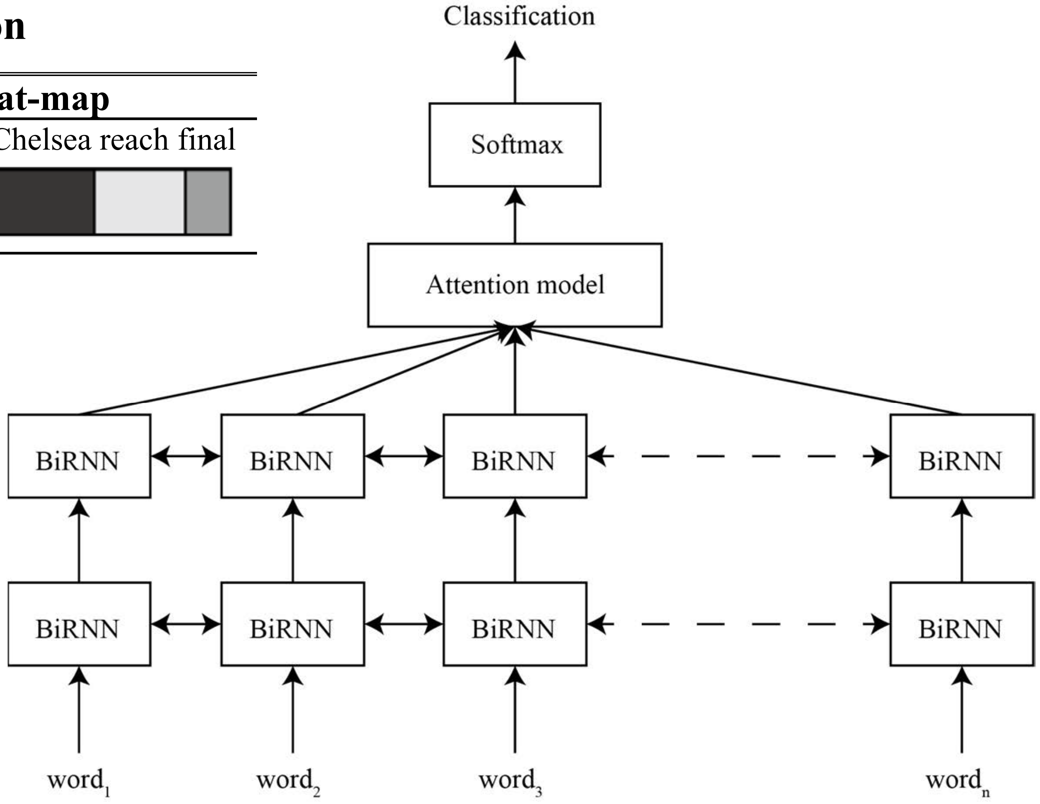
- Short Extracted Network

Table I. Heat Map for Attention

Original title	Attention heat-map
goalkeeper saved the penalty twice, Chelsea reached the final.	Goalkeeper save penalty Chelsea reach final 

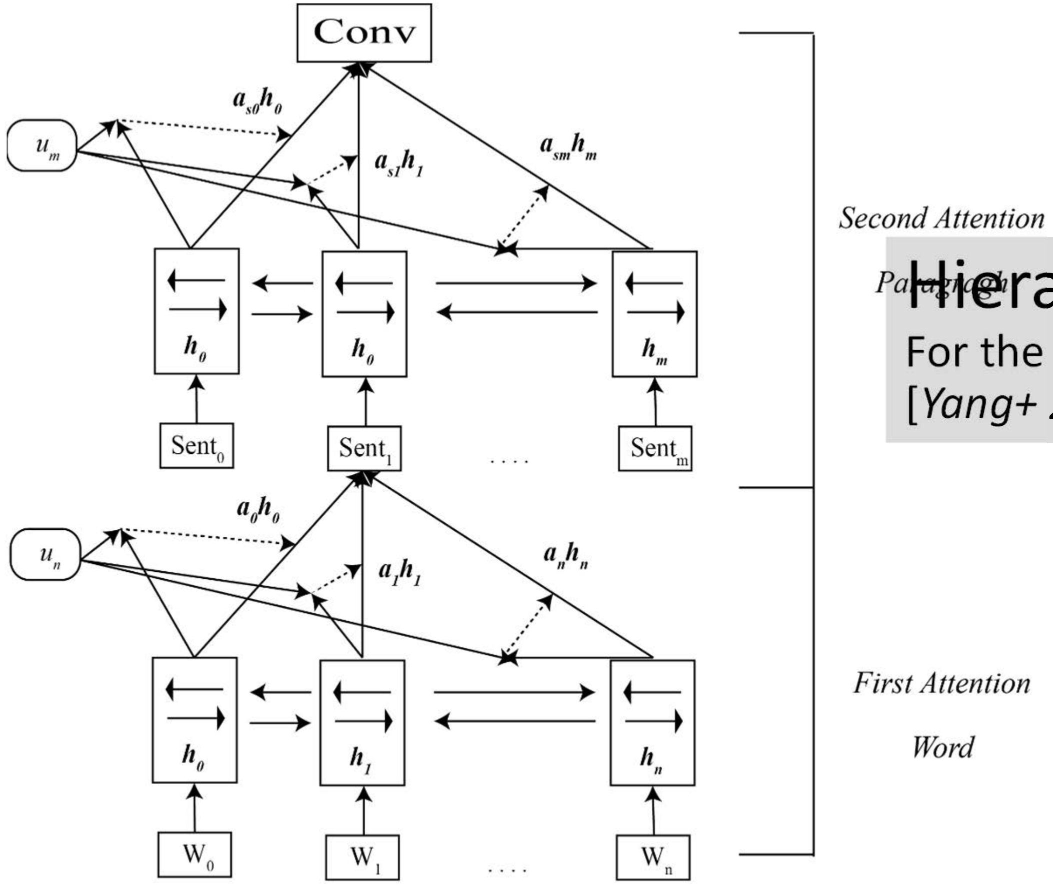
GRU + Attention

GRU is generally used instead of the traditional RNN structure to eliminate the gradient dispersion problem. [*Zhou+ 2016*]



Approach

- Long Extracted Network

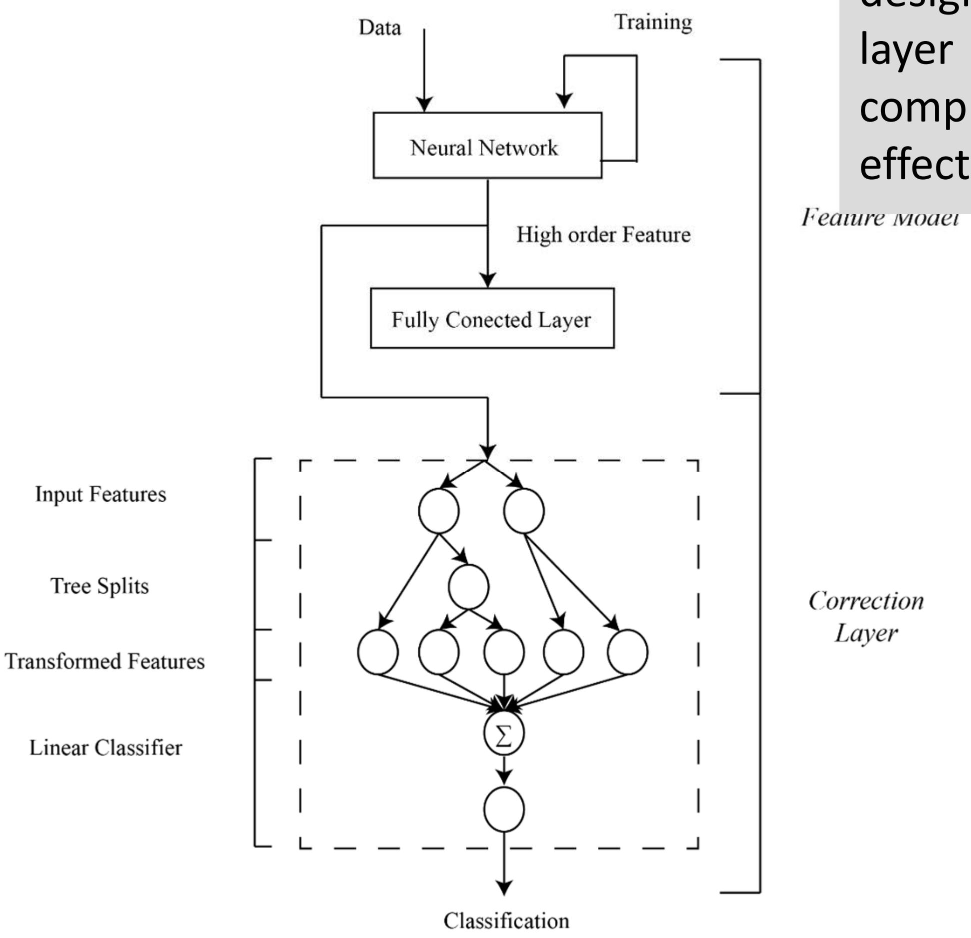


Hierarchical attention

For the key paragraph and key word.
[*Yang+ 2016*]

Approach

- Correct Layer



XGboost [*chen+ 2016*]

We improve the accuracy of the model by designing a correction layer. The correction layer uses the additive training principle to complete the correction of the classification effect of the individual neural network model.

Table II. Parameters of XGboost

Parameter	**Option/value**
objective	binary:logistic
booster	gbtree
eval_metric	logloss
eta	0.1
max_depth	9
subsample	0.9
min_child_weight	5
silent	1

Overview

- Demand Analysis
- Related work and Challenge
- Approach
- Experiments
- Conclusion

Experiments

Dataset*

- *Totally 270912 pieces from common Chinese Portals (Tencent, Sina, etc).*
- *Label: Entertainment, Games, Education, Arts, Finance, Technology, Cars, Sports , Fashion.*

Data overview

- *We choose the mode as the length of the vector expression to ensure* semantic integrity.

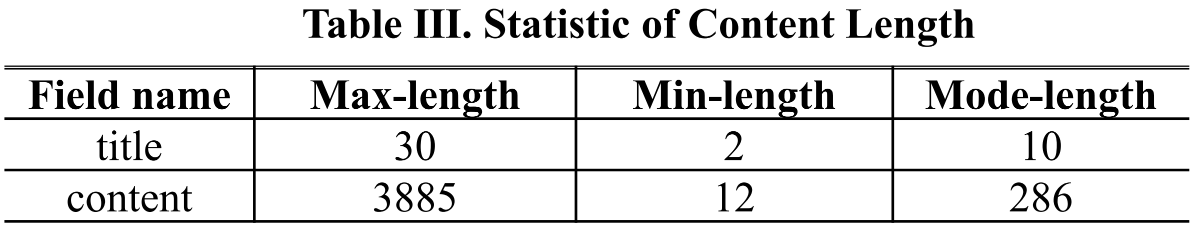


Table III. Statistic of Content Length

Field name	Max-length	Min-length	Mode-length
title	30	2	10
content	3885	12	286

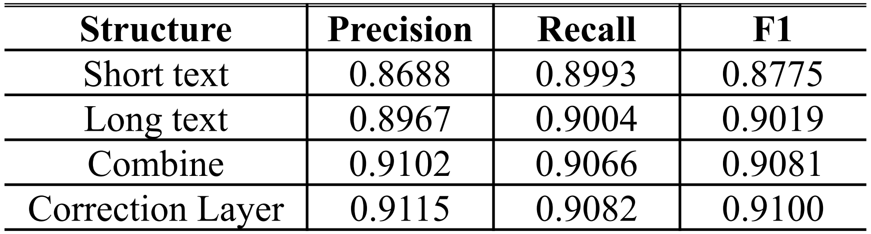
** This project is funded by Beijing Institute of Science and Technology Information.*

Experiments

Tuned

Comparison of Untuned Model

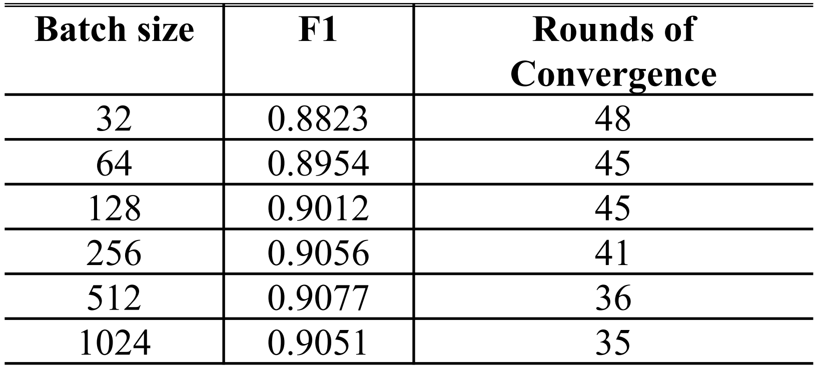
Table IV. Verification Model Design



Structure	Precision	Recall	F1
Short text	0.8688	0.8993	0.8775
Long text	0.8967	0.9004	0.9019
Combine	0.9102	0.9066	0.9081
Correction Layer	0.9115	0.9082	0.9100

Influence of Batch Size

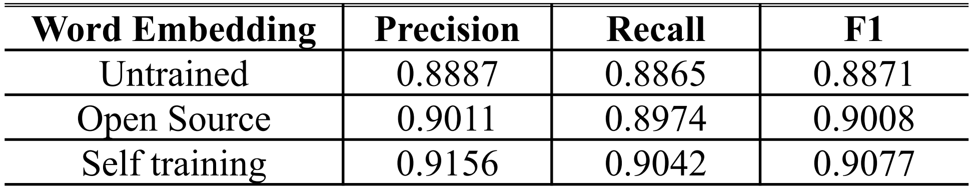
Table V. F1-Batch Size



Batch size	F1	Rounds of Convergence
32	0.8823	48
64	0.8954	45
128	0.9012	45
256	0.9056	41
512	0.9077	36
1024	0.9051	35

Different Embedding Matrix

Table VI. web-based corpus-based pre-training word vector successfully introduces external semantics



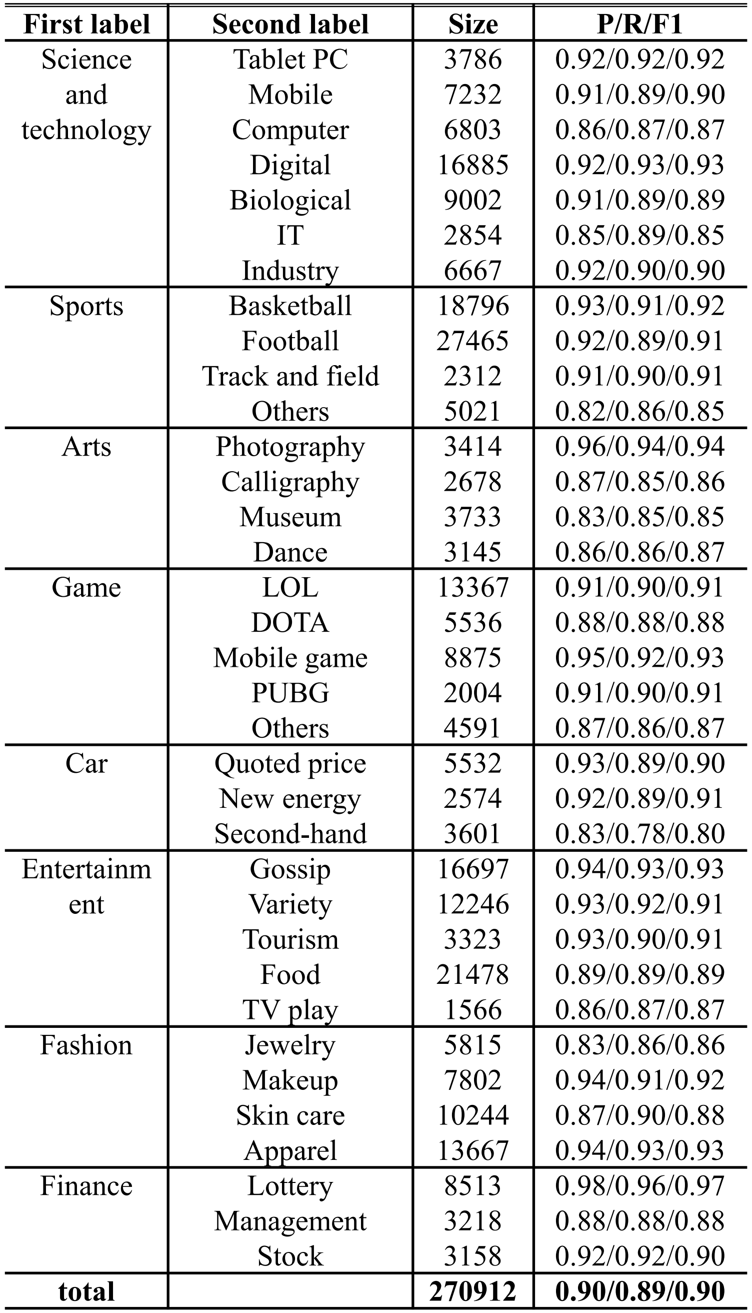
Word Embedding	Precision	Recall	F1
Untrained	0.8887	0.8865	0.8871
Open Source	0.9011	0.8974	0.9008
Self training	0.9156	0.9042	0.9077

Experiments

Result

The proposed algorithm achieves 0.9 under the second-level label, and further improves to at least 0.94 under the first-level label.

Table VII. Final Result On our Corpus

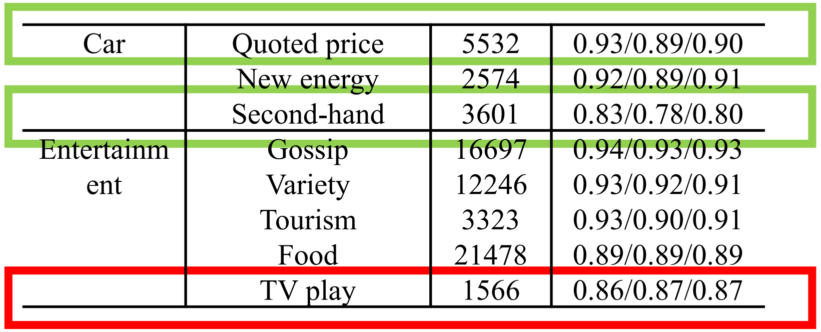


First label	Second label	Size	P/R/F1
Science and technology	Tablet PC	3786	0.92/0.92/0.92
	Mobile	7232	0.91/0.89/0.90
	Computer	6803	0.86/0.87/0.87
	Digital	16885	0.92/0.93/0.93
	Biological	9002	0.91/0.89/0.89
	IT	2854	0.85/0.89/0.85
	Industry	6667	0.92/0.90/0.90
Sports	Basketball	18796	0.93/0.91/0.92
	Football	27465	0.92/0.89/0.91
	Track and field	2312	0.91/0.90/0.91
	Others	5021	0.82/0.86/0.85
Arts	Photography	3414	0.96/0.94/0.94
	Calligraphy	2678	0.87/0.85/0.86
	Museum	3733	0.83/0.85/0.85
	Dance	3145	0.86/0.86/0.87
Game	LOL	13367	0.91/0.90/0.91
	DOTA	5536	0.88/0.88/0.88
	Mobile game	8875	0.95/0.92/0.93
	PUBG	2004	0.91/0.90/0.91
	Others	4591	0.87/0.86/0.87
Car	Quoted price	5532	0.93/0.89/0.90
	New energy	2574	0.92/0.89/0.91
	Second-hand	3601	0.83/0.78/0.80
Entertainment	Gossip	16697	0.94/0.93/0.93
	Variety	12246	0.93/0.92/0.91
	Tourism	3323	0.93/0.90/0.91
	Food	21478	0.89/0.89/0.89
	TV play	1566	0.86/0.87/0.87
Fashion	Jewelry	5815	0.83/0.86/0.86
	Makeup	7802	0.94/0.91/0.92
	Skin care	10244	0.87/0.90/0.88
	Apparel	13667	0.94/0.93/0.93
Finance	Lottery	8513	0.98/0.96/0.97
	Management	3218	0.88/0.88/0.88
	Stock	3158	0.92/0.92/0.90
total		**270912**	**0.90/0.89/0.90**

Overview

- Demand Analysis
- Related work and Challenge
- Approach
- Experiments
- Conclusion

Conclusion



Car	Quoted price	5532	0.93/0.89/0.90
	New energy	2574	0.92/0.89/0.91
	Second-hand	3601	0.83/0.78/0.80
Entertainment	Gossip	16697	0.94/0.93/0.93
	Variety	12246	0.93/0.92/0.91
	Tourism	3323	0.93/0.90/0.91
	Food	21478	0.89/0.89/0.89
	TV play	1566	0.86/0.87/0.87

- **Method**

  - Combine long and short extracted NN.

  - XGboost for the correct layer.

- **Further development**

  - ***The unbalanced corpus problem***. Initial quantity affects classification trend(shown as red block).

  - *Cross subclass categorization problem*(Shown as green blocks). Combination of the latest methods such as Bert [*Devlin+ 2018*] to further enrich the original semantics.

Thank you!