Imbalanced Sentiment Classification Enhanced with Discourse Marker

ICANN19
28th International Conference on Artificial Neural Networks

Tao Zhang[1,2], Xing Wu[1,2,3], Meng Lin[1*], Jizhong Han[1] and Songlin Hu[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

[3] Baidu, Inc., Beijing, China

{zhangtao, linmeng, hanjizhong, husonglin}@iie.ac.cn, wuxing03@baidu.com

Introduction

Problem & Goal

Imbalance data exists in plenty of scenarios, making it hard to be utilized directly for supervised model training. This phenomenon emerges frequently in sentiment-related area, where individuals tend to select and share content based on what the majority agrees with.

We aim to use data augmentation methods to decrease imbalanced-ratio, boosting the performance of sentence level imbalanced text classification on sentiment datasets.

An important observation

Humans often express transitional emotion between two adjacent discourses with discourse markers like “but”, “though”, “while”, etc, and the head discourse and the tail discourse usually indicate opposite emotional tendencies. We simply use the inherent antonymy relation between discourses to generated new samples.

Raw:	The actress is beautiful, but the plot is terrible.	Negative
S1:	The actress is beautiful.	Positive
S2:	The plot is terrible.	Negative
S3:	The plot is terrible, but The actress is beautiful.	Positive

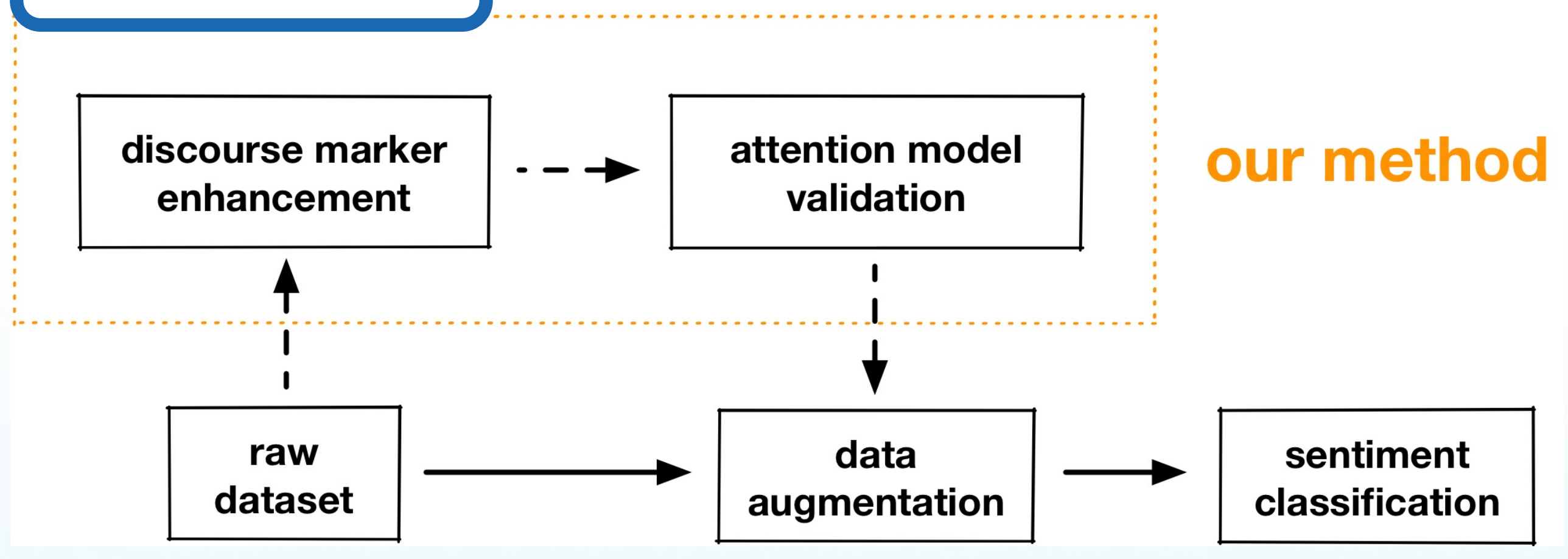
Existing works

a) Re-sampling methods, causing either information loss when under-sampling the majority class or overfitting when over-sampling the minority class.

b) Using generative models to generate a whole bunch of sentences as new samples, often leading to low quality generation.

c) Replacement-based methods find replaceable words or phrases and substituting them with synonyms, limited by the candidate vocabulary.

Method



① Crop & Swap

$$[S_1, dm, S_2], p/n \rightarrow \begin{cases} [S_1], n/p \\ [S_2], p/n \\ [S_2, dm, S_1], n/p \end{cases}$$

② Attention-based validation

$$h_i = Bi-LSTM(x_i)$$
$$f(h_i, v) = \tanh(W_h h_i + W_v v + b)$$
$$\alpha_i = \mathrm{softmax}(f(h_i, v))$$
$$c = \sum_i \alpha_i h_i$$
$$y = \mathrm{softmax}(W_s c + b_s)$$

Our method

We propose to use two simple operations named “crop & swap” to generate samples from original transitional sentences to augment sentiment datasets. We introduce a attention-based model to validate generated samples to avoid label inconsistency problem. The strength of our methods is as follows:

- Generated samples are from original dataset, introducing no noise and no change for data distribution.
- Breaking hard transitional sentences, into easily classified discourses.
- Serving as a pre-process step to decrease imbalanced-ratio, easily integrated with other data augmentation methods like re-sampling.

Experiment

Dataset

1. **MR Movie Review**, consisting of 5,331 positive and 5,331 negative reviews.
2. **SST2 Stanford Sentiment Treebank** (binary version).
3. **CR Customer**, reviews of various products with 2406 positive and 1367 negative samples.

Baselines & Algorithms

Datasets balanced with oversampling methods only, while ours is first decreasing imbalanced-ratio with our method and then using oversampling to balance datasets.

Machine learning models: naïve Bayes (NB), logistic regression (LR), support vector machine (SVM). Deep learning models: TextCNN, TextRNN, both with regular settings.

Effectiveness of our method

Method	Setting	MR	SST2	CR	Avg improvement
NB	w/os	72.79	73.31	74.19	-
	w/our + os	71.90	76.99	76.20	1.60
LR	w/os	68.88	69.02	74.60	-
	w/our + os	67.95	71.14	76.20	0.93
SVM	w/os	66.34	49.91	50.00	-
	w/our + os	50.00	69.41	50.00	1.05
CNN	w/os	71.33	75.28	77.82	-
	w/our + os	74.14	79.95	81.45	3.70
RNN	w/os	71.80	74.30	75.60	-
	w/our + os	75.34	79.39	77.02	3.35

Effectiveness of Validation

Method	IR	Setting	MR	SST2	CR	Avg improvement
CNN	5	wo/val	72.84	74.84	79.43	-
		full	74.14	79.96	81.45	2.81
	20	wo/val	66.18	70.12	62.70	-
		full	70.95	74.74	68.75	5.15
	100	wo/val	61.65	65.01	60.28	-
		full	67.95	69.08	61.69	3.93

Highly imbalanced datasets

IR	Method	Setting	MR	SST2	CR	Avg improvement
10	NB	w/os	68.15	68.86	69.75	-
		w/our + os	69.41	72.87	71.98	2.50
	LR	w/os	60.87	60.07	63.91	-
		w/our + os	63.74	65.24	70.96	5.33
	CNN	w/os	72.94	76.49	70.76	-
		w/our + os	74.51	79.74	75.40	3.15
	RNN	w/os	71.96	69.14	73.59	-
		w/our + os	71.90	77.05	76.01	3.42
20	NB	w/os	61.44	61.55	63.30	-
		w/our + os	63.21	67.55	68.75	4.41
	LR	w/os	53.85	54.31	54.43	-
		w/our + os	59.72	60.13	63.31	6.86
	CNN	w/os	68.26	67.87	53.42	-
		w/our + os	70.92	74.74	68.75	8.23
	RNN	w/os	60.93	70.84	60.69	-
		w/our + os	70.03	76.44	73.99	9.33
50	NB	w/os	55.10	56.12	55.84	-
		w/our + os	61.76	65.35	64.52	8.19
	LR	w/os	50.52	50.74	51.41	-
		w/our + os	56.24	56.01	57.46	5.61
	CNN	w/os	62.27	54.09	51.20	-
		w/our + os	67.69	68.97	67.14	12.08
	RNN	w/os	52.29	53.10	52.82	-
		w/our + os	67.48	71.06	63.91	14.75
100	NB	w/os	51.61	51.78	52.01	-
		w/our + os	61.76	65.34	64.51	12.07
	LR	w/os	50.15	49.97	50.40	-
		w/our + os	56.24	56.01	57.46	6.40
	CNN	w/os	53.74	50.74	50.60	-
		w/our + os	67.95	69.08	61.69	14.55
	RNN	w/os	50.62	50.85	50.60	-
		w/our + os	68.57	67.01	65.12	16.21

Conclusion

(1) We propose a novel two-step method, which first generates new samples according to transitional discourse markers and then validates polarity correctness with a pre-trained attention-based model.

(2) The experimental results proves that the semantics conveyed by transitional discourse marker can be utilized to generate sentimental discourses.

(3) Our method is simple and plug-and-play, serving as a upstream part in data augmentation.

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

网络空间技术实验室
Laboratory for Cyberspace Technologies