Artificial Intelligence

Masking and Generation: An Unsupervised Method for Sarcasm Detection

Rui Wang
Joint Lab of CMS-HITSZ,
Harbin Institute of Technology
Shenzhen, China
ruiwangnlp@outlook.com

Qianlong Wang
Joint Lab of CMS-HITSZ,
Harbin Institute of Technology
Shenzhen, China
qlwang15@outlook.com

Bin Liang
Joint Lab of CMS-HITSZ,
Harbin Institute of Technology
Shenzhen, China
bin.liang@stu.hit.edu.cn

Yi Chen
Joint Lab of CMS-HITSZ,
Harbin Institute of Technology
Shenzhen, China
yichennlp@gmail.com

Zhiyuan Wen
Joint Lab of CMS-HITSZ,
Harbin Institute of Technology
Shenzhen, China
wenzhiyuan2012@gmail.com

Bing Qin
Harbin Institute of Technology
Harbin, China
qinb@ir.hit.edu.cn

Ruifeng Xu*
Harbin Institute of Technology
Shenzhen, China
Peng Cheng Laboratory
Shenzhen, China
xuruifeng@hit.edu.cn

gesis
Leibniz-Institut
für Sozialwissenschaften

Reported by Zhaoze Gao

Artificial

1. Introduction
2. Approach
3. Experiments

Introduction

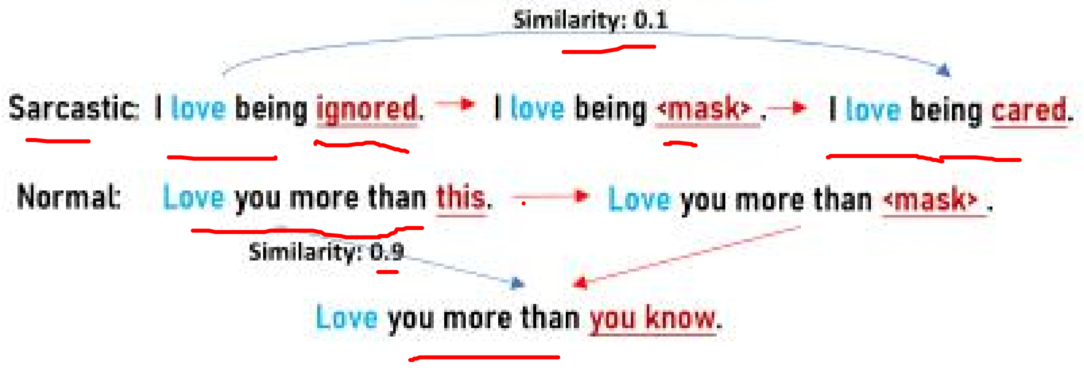
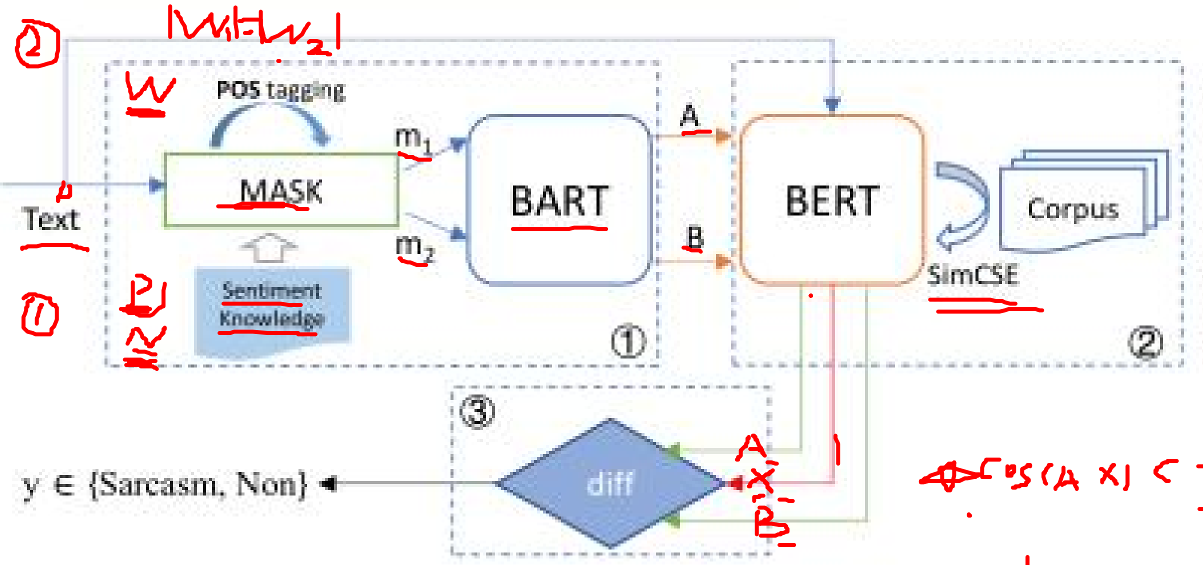


Figure 1: Red denotes the masked places and blue means decisive sentiment words. During the mask and generation procedure, sarcastic texts suffer more changes than normal texts. Hence, for sarcasm sentences, the similarity between original and reborn texts will be relatively lower.

Since the pre-trained generation model is pre-trained on general corpora where sarcastic texts are scarce, we assume that given a masked text,it can generate a relatively normal one according to the remaining logic information.

Approach



$\cos(A, X) < T \;||\; \cos(B, X) < T$

Figure 2: Architecture of our proposed method.

Approach

$$x = \{x_1, x_2, ..., x_n\}$$

$$PW = \{pw_1, pw_2, ..., pw_h\}$$

$$NW = \{nw_1, nw_2, ..., nw_k\}, h + k \leq n$$

$$SW = \{sw_1, sw_2, ..., sw_m, m \leq n\}$$

$$SW_1 \cup SW_2 = SW, \ |SW_1| = |SW_2|$$

Here, $PW \cup SW_1$ and $NW \cup SW_2$ are used to mask original sentence respectively. So we will obtain two masked sentences $x_{m1} = \{[m]_1, x_2, ..., [m]_n\}$ and $x_{m2} = \{x_1, [m]_2, ..., x_n\}$. These two masked sentences are fed into the pre-trained generation model to fulfill the generation procedure.

Approach

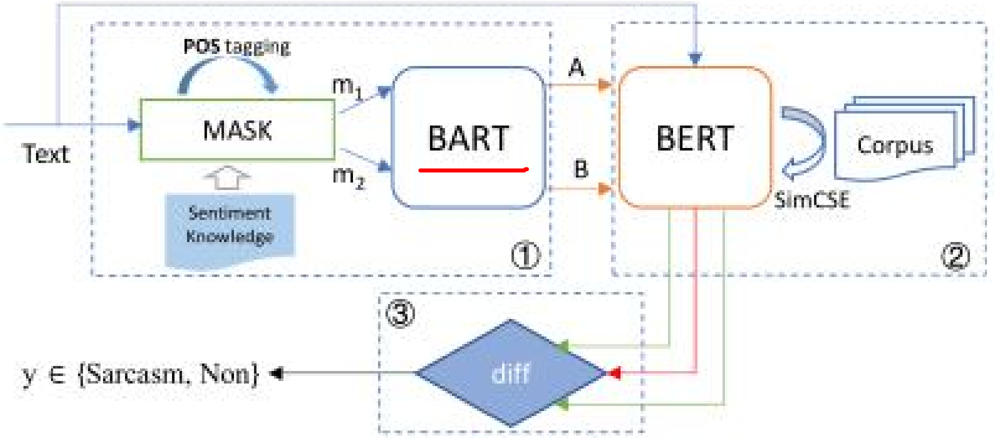


Figure 2: Architecture of our proposed method.

$$A\{a_1, ..., x_2, ..., x_{n-1}, ..., a_o\} = BART([m]_1, x_2, ..., x_{n-1}, [m]_n) \quad (1)$$

B

$$\boldsymbol{H_x, H_A, H_B} = BERT(x), BERT(A), BERT(B) \quad (2)$$

$$\mathbf{diff} = \mathbf{sim}(H_x, H_A) < threshold \,||\, \mathbf{sim}(H_x, H_B) < threshold \quad (3)$$

$$y = \mathbb{I}(\mathbf{diff}) \quad (4)$$

$$\begin{aligned} &\mathbf{rule\text{-}2}: \; y = \mathbb{I}(|\,\mathbf{sim}(H_x, H_A) - \mathbf{sim}(H_x, H_B)\,| < threshold) \\ &\mathbf{rule\text{-}3}: \; y = \mathbb{I}(\mathbf{sim}(H_A, H_B) < threshold) \end{aligned} \quad (5)$$

Experiments

Table 1: Statistics of training and test datasets.

Dataset	Train		Test	
	Sarcasm	Non	Sarcasm	Non
IAC-V1	862	859	97	94
IAC-V2	2, 947	2, 921	313	339
Tweet-1	23, 456	24, 387	2, 569	2, 634
Tweet-2	282	1, 051	35	113
Reddit-1	5, 521	5, 607	1, 389	1, 393
Reddit-2	6, 419	6, 393	1, 596	1, 607
iSarcasm	476	2, 346	124	582

Experiments

Table 2: Ablation study (F1). $\mathcal{S}$ denotes the split of affective words. $\mathcal{M}$+$\mathcal{G}$ denotes the Masking and Generation procedure.

MODEL	IAC1	IAC2	Tweet-1	Tweet-2	Reddit-1	Reddit-2
Our	55.44	63.17	58.76	58.31	54.91	55.80
w/o $\mathcal{S}$	54.32	60.26	56.23	57.75	52.52	54.58
w/o $\mathcal{M}+\mathcal{G}$	33.68	54.22	33.36	52.53	43.29	53.09

Experiments

Table 3: Main experimental results on different datasets. Average scores over five runs are reported. Best scores are in bold. Second best scores are underlined.

MODEL	IAC1		IAC2		Tweet-1		Tweet-2		Reddit-1		Reddit-2	
	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)	Acc.(%)	F1.(%)
Lexicon	47.64	40.29	44.01	39.12	59.00	55.86	57.43	51.7	43.06	42.71	42.77	41.47
TF-IDF-LDA	53.40	53.22	54.61	52.44	54.52	54.36	50.68	48.15	52.51	50.81	51.72	47.65
TF-IDF-Kmeans	49.73	49.35	51.68	47.52	52.27	44.1	**72.97**	51.86	49.68	46.74	52.58	43.29
BERT+word-Mask [11]	51.39	36.35	48.00	35.72	**59.46**	56.54	41.22	41.21	47.19	39.47	46.91	37.63
Ours	52.35	53.75	62.06	56.75	50.21	52.35	67.57	55.24	52.62	52.60	51.92	49.89
Ours+SimCSE	**57.59**	**55.44**	**64.30**	**64.27**	58.91	**58.76**	56.76	**58.31**	**53.30**	**54.91**	**56.16**	**56.14**

Experiments

Table 4: Experiment results on iSarcasm Dataset. Best Scores are in bold. Second best scores are underlined.

MODEL	Precision.(%)	Recall.(%)	F1.(%)
Lexicon	49.2	48.7	40.5
TF-IDF-LDA	15.7	49.0	42.6
TF-IDF-Kmeans	18.8	32.5	32.4
BERT+word-Mask [11]	16.7	**88.5**	24.0
LSTM	21.7	74.7	33.6
CNN	26.1	56.3	35.6
SIARN [25]	21.9	78.2	34.2
MIARN [25]	23.6	<u>79.3</u>	36.4
3CNN [10]	25.0	33.3	28.6
Dense-LSTM [27]	<u>37.5</u>	27.6	31.8
Ours	**50.7**	50.5	<u>50.1</u>
Ours+SimCSE	20.5	72.7	**52.1**

Experiments

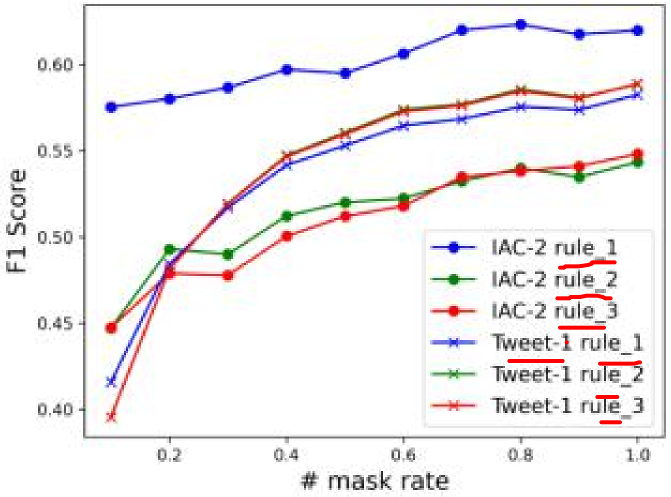


Figure 3: The performance of different prediction rules and mask rates. The average lengths of texts are 270 and 80 for IAC-2 and Tweet-1 respectively.

Experiments

Table 5: Different numbers of labeled data (development set) and corresponding thresholds. Average scores over 100 runs are reported. *T*- and *R*- represent Tweet and Reddit respectively.

number	10	50	200	10%	best-F1
T-threshold	0.9920	0.9839	0.9759	0.9759	0.9820
R-threshold	0.9076	0.9437	0.9317	0.9317	0.9260
Tweet-1	52.91	55.85	56.67	57.08	58.76
Reddit-2	50.35	53.76	55.02	55.03	56.14

Experiments

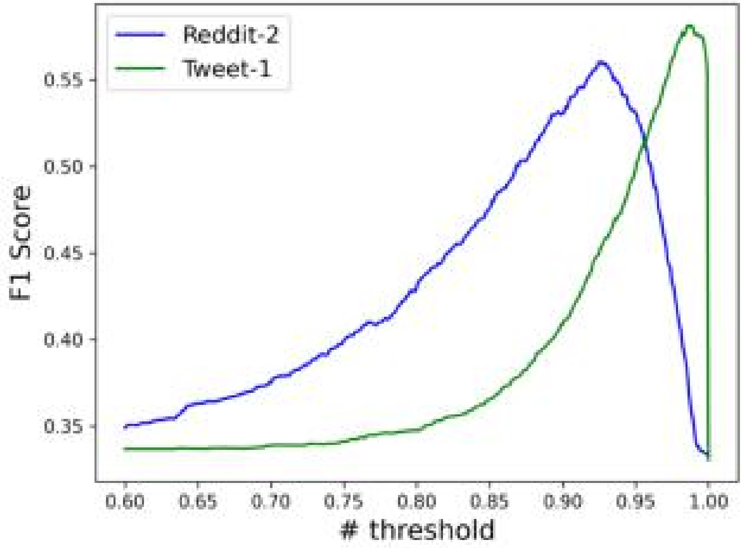


Figure 4: Thresholds and corresponding F1 scores on two datasets.

Experiments

	Lexicon	BERT+word -mask	Ours
1. yes cause a stupid looking Duck on a hat is pretty awesome. *Sarcastic*	√	√	√
2. I just love getting calls from restricted numbers. *Sarcastic* -> I just keep getting calls from restricted numbers. Retrieved word by BERT+word-mask	×	√	√
3. So the Romans nailed anyone up that organized the community! Did you get that from the film? *Sarcastic* -> So the Romans nailed anyone who did that to the community! Did you get that from the film? Generated text by Ours -> So the Romans didn't set anyone up that organized the event. Did you know that from the book? Generated text by Ours	×	×	√

Figure 5: Three examples for case study. Red denotes negative word. Blue denotes positive word. Purple represents masked tokens. Green represents corresponding generated tokens.

Thank you！