LREC-COLING 2024

SarcNet: A Multilingual Multimodal Sarcasm Detection Dataset

Tan Yue[1,2], Xuzhao Shi[1], Rui Mao[2], Zonghai Hu[1], Erik Cambria[2*]

[1]School of Electronic Engineering, Beijing University of Posts and Telecommunications

[2]School of Computer Science and Engineering, Nanyang Technological University

Speaker：Tan Yue

CONTENTS

Part 01

Research Challenge and Motivation

Research Challenge and Motivation

Sarcasm

Sarcasm is defined as *the activity of saying or writing the opposite of what you mean, or of speaking in a way intended to make someone else feel stupid or show them that you are angry*

What a wonderful weather!

Research Challenge and Motivation

Challenge

1. **Dataset limitation**: Limited dataset availability has led many studies to use a single dataset, raising concerns about inconclusive findings and model generalizability

2. **Multilingual nature**: The existing dataset only includes English text, which is inadequate for the multilingual nature of social networks

3. **Labeling method**: The current labeling method assigns one label to multimodal instances, which can lead to biased evaluations when used for both unimodal and multimodal models.

Research Challenge and Motivation

Label

New Label

Modalities		Label
Image		1
Text	What a wonderful weather!	1
Multimodal		1

Modalities		Label
Image		0
Text	What a wonderful weather!	0
Multimodal		1

Research Challenge and Motivation

Contribution

1. **Dataset**: A new dataset, SarcNet, for multilingual and multimodal sarcasm detection is introduced, consisting of 3,335 image-text pairs with 10,005 labels

2. **Labeling schema** : We modify the multimodal sarcasm dataset's labeling schema to address shortcomings and improve interpretability by labeling each sample distinctly for different modalities

3. **Benchmark**: We conducted extensive experiments and set benchmarks on the dataset using baseline models.

Part 02

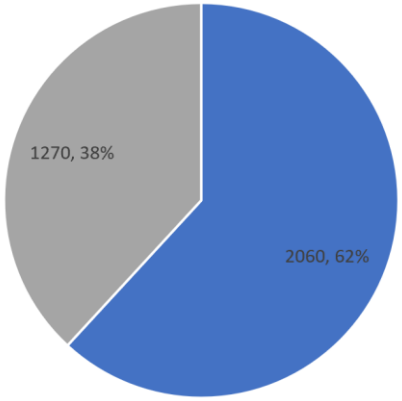
SarcNet Dataset

SarcNet Dataset

Collection

- **Topics:** #irony; #sarcasm; #讽刺(sarcasm in Chinese); ……
- **Key words:** 卷王(refers to someone who obsessively focuses on studying or working to the exclusion of all else, often in a competitive and intensive manner); ……
- Each sample contains an image and a piece of text.

Language	Train	Val.	Test	Total
Chinese	1242	424	399	2065
English	756	247	267	1270
Total	1998	671	666	3335



Chinese 62% English 38%

SarcNet Dataset

Annotation

We distinctly labeled the image data, text data, and multimodal data (image-text pairs), resulting in a total of 10,005 labels.

Mod.	Class	Train	Val.	Test	Total
Text	Sarcasm	864	274	308	1446
	Non-Sarcasm	1134	397	358	1889
	Total	1998	671	666	**3335**
Image	Sarcasm	623	224	221	1068
	Non-Sarcasm	1375	447	445	2267
	Total	1998	671	666	**3335**
M-mod.	Sarcasm	1116	372	387	1875
	Non-Sarcasm	882	299	279	1460
	Total	1998	671	666	**3335**
Total		**5994**	**2013**	**1998**	**10005**

SarcNet Dataset

Quality Control

Each sample was annotated by two independent annotators. If two annotators agree on a label, it becomes the ground truth. If they disagree, a third annotator is involved, and the majority-agreed label is selected as the ground truth.

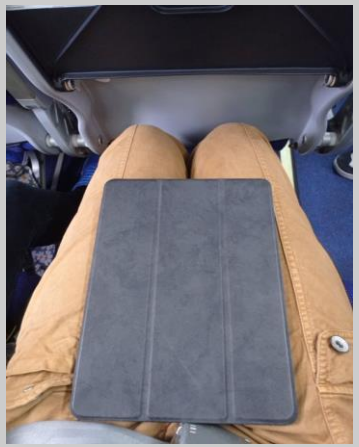
We achieve Cohen’ s Kappa by **0.9032**, **0.7129** and **0.9769**, for textual, visual and multimodal labels respectively,

$$\mathcal{K} = \frac{P_o - P_e}{1 - P_e}$$

Cohen's Kappa	Interpretation
$\mathcal{K} \leq 0$	No agreement
$0 < \mathcal{K} < 0.20$	Slight agreement
$0.20 \leq \mathcal{K} < 0.40$	Fair agreement
$0.40 \leq \mathcal{K} < 0.60$	Moderate agreement
$0.60 \leq \mathcal{K} < 0.80$	Substantial agreement
$0.80 \leq \mathcal{K} < 1$	Near perfect agreement
1	Perfect agreement

SarcNet Dataset

Dataset Analysis

		Label		Label
Image		0		1
Text	Thank God for Sooooooooo much Leg space I can't even cross my feet, let alone my legs	1	Lol	0
Multimodal		1		1

	{0,0,1}	{1,0,1}	{0,1,1}	{1,1,1}	{0,0,0}	{1,0,0}	{0,1,0}
Image_label	0	1	0	1	0	1	0
Text_label	0	0	1	1	0	0	1
Multi_label	1	1	1	1	0	0	0
Num	127	315	716	717	1411	36	13

Part 03

Experiment

Experiment

Modality	Method	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
Image	ResNet	67.32	50.16	62.58	55.69
	MobileNetV3	68.37	53.65	56.56	55.07
	ViT	68.62	53.40	60.28	56.63
Text	LSTM	68.47	66.55	63.96	65.23
	TextCNN	70.67	68.08	67.16	67.62
	BERT	73.57	73.74	66.56	69.97
	MultiL-BERT	74.92	76.01	66.88	71.16
	XLM-RoBERTa	76.88	78.52	68.83	73.36
Multimodal	Res-BERT	79.48	83.45	80.40	81.89
	KnowleNet	80.38	85.06	79.85	82.37
	DT4MID(BERT)	80.53	80.35	84.82	82.52
	DT4MID(XLM-R)	82.73	88.86	80.36	84.40

Experiment

Case Study

The figure is an example of a sarcastic text-image pair with the label of {0, 0, 1}. The literal meaning of **Jack Ma（“马云”）** in Chinese is “**horse”（“马”）** and “**cloud”（“云”）**.

Text content: 让 # 文心一言 # 画个马云(Let #ERNIE Bot# draw the Jack Ma). A difficult case for detection.

Part 04

Conclusion

Conclusion

Discussion

- We propose SarcNet, a novel multilingual and multimodal sarcasm detection dataset, comprising 3,335 image-text pair samples and yielding over 10,000 labels.
- The distinct image and text labels prove advantageous for more effectively testing unimodal models.

Future Work

Expand dataset:

- More data: From 10K to 100K
- More languages: German, French……
- More modalities: Text, Video, Audio……

LREC-COLING 2024

Thank you

E-mail：yuetan@bupt.edu.cn

Speaker：Tan Yue