Utilizing Weak Supervision to Create S3D: A Sarcasm Annotated Dataset

Jordan Painter, Diptesh Kanojia, Helen Treharne

UNIVERSITY OF SURREY

Roadmap

- Motivations - ‘Why?’
- Contributions - ‘What?’
- Methodology - ‘How?’
- Results
- Conclusions

Motivation

Sarcasm Detection: Task of identifying if a given extract of data is sarcastic

- Use of sarcasm on social media can have diminishing effects on other NLP tasks (Sentiment Analysis).
- Small number of publicly available sarcasm detection datasets, many have decreased in size over time as tweets get deleted by users.
- All current methods of dataset annotation rely on trusting a user’s own judgement (self-annotation), or the slow process of manually annotating data.
- Lack of comparative analysis between state-of-the-art language models for the task of sarcasm detection.

“Oh yeah this phone is fantastic, I just love how the battery dies 3 hours after charging”

Contributions

Datasets:

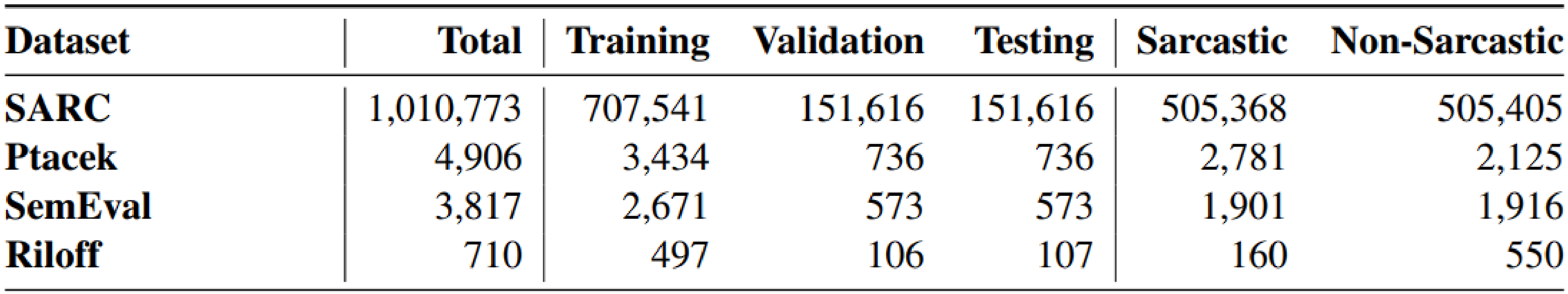
SAD - A dataset of 2,340 tweets, scraped by observing a #sarcasm hashtag, then manually annotated by 3 annotators.

S3D - A dataset of 100,000 tweets, annotated using our novel approach of weak supervision.

A performance **evaluation** of existing language models and datasets for the binary classification task of sarcasm detection.

Release of our code, data and models on both GitHub and HuggingFace publicly for the research community.

Existing Datasets



Dataset	Total	Training	Validation	Testing	Sarcastic	Non-Sarcastic
SARC	1,010,773	707,541	151,616	151,616	505,368	505,405
Ptacek	4,906	3,434	736	736	2,781	2,125
SemEval	3,817	2,671	573	573	1,901	1,916
Riloff	710	497	106	107	160	550

Ptacek - 4,096 self-annotated tweets - #sarcasm

SemEval 2018 - 3,817 manually annotated tweets

Riloff - 710 manually annotated tweets

SARC - Over 1,000,000 Reddit self-annotated Reddit comments – ‘/s’

SAD

- Using TWINT, we collected tweets containing a #sarcasm hashtag.

- Every sarcastic tweet would then become a tweet pair, by searching for a recent tweet by the same user that didn’t contain #sarcasm

- Tweet pairs were then manually annotated by three annotators

- A total of 2340 tweets annotated for sarcasm

Methodology

- Six datasets were used for training: four pre-existing, our new SAD dataset and a final ‘combined’ dataset.

- Every text extract was pre-processed to remove punctuation and capitalisation. Usernames were replaced with the generic ‘@user’.

- All examples of ‘#sarcasm’ were removed from relevant datasets.

- This pre-processed data was used to fine-tune five language models.

Language Models

- **BERT**

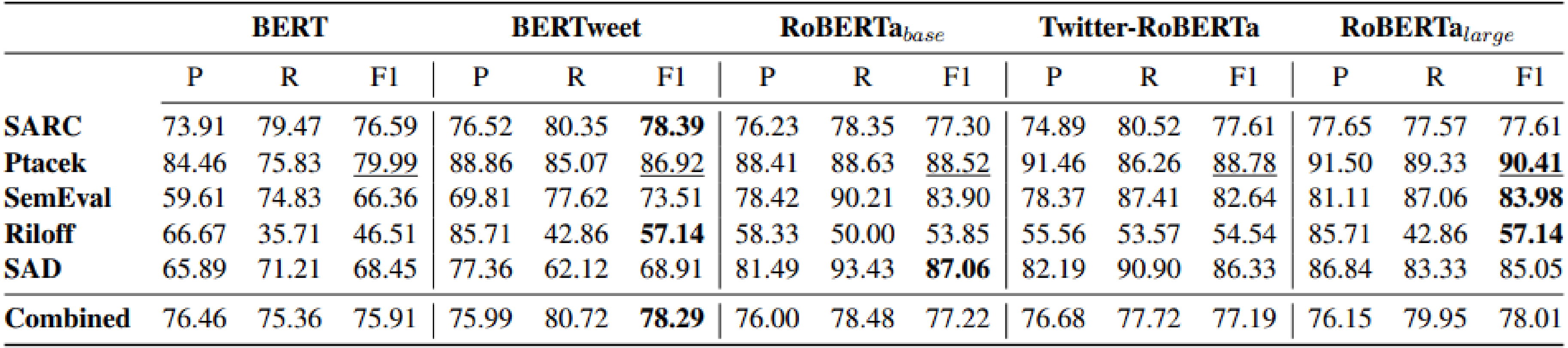
- **RoBERTa$_{\text{base}}$**

- **RoBERTa$_{\text{large}}$**

- **BERTweet** - A BERT model pre-trained using the RoBERTa pre-training procedure on a corpus of 850M tweets.

- **Twitter-RoBERTa** - A RoBERTa$_{\text{base}}$ model pre-trained on ~58M tweets.

Results



	BERT			BERTweet			RoBERTa$_{base}$			Twitter-RoBERTa			RoBERTa$_{large}$		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SARC	73.91	79.47	76.59	76.52	80.35	**78.39**	76.23	78.35	77.30	74.89	80.52	77.61	77.65	77.57	77.61
Ptacek	84.46	75.83	79.99	88.86	85.07	86.92	88.41	88.63	88.52	91.46	86.26	88.78	91.50	89.33	**90.41**
SemEval	59.61	74.83	66.36	69.81	77.62	73.51	78.42	90.21	83.90	78.37	87.41	82.64	81.11	87.06	**83.98**
Riloff	66.67	35.71	46.51	85.71	42.86	**57.14**	58.33	50.00	53.85	55.56	53.57	54.54	85.71	42.86	**57.14**
SAD	65.89	71.21	68.45	77.36	62.12	68.91	81.49	93.43	**87.06**	82.19	90.90	86.33	86.84	83.33	85.05
Combined	76.46	75.36	75.91	75.99	80.72	**78.29**	76.00	78.48	77.22	76.68	77.72	77.19	76.15	79.95	78.01

Table 3 in the paper shows the results of our deep-learning experiments, where P, R and F1 denote Precision, Recall and F1-score respectively.

We observed BERTweet achieved the highest F1-score on our largest dataset, ‘Combined’, of 78.29.

What is Weak Supervision?

- An approach to machine learning which allows for the creation of much larger datasets, at the expense of them being noisier.

- Use a pretrained model to label data.

- Removes the tediousness of manually annotating data.

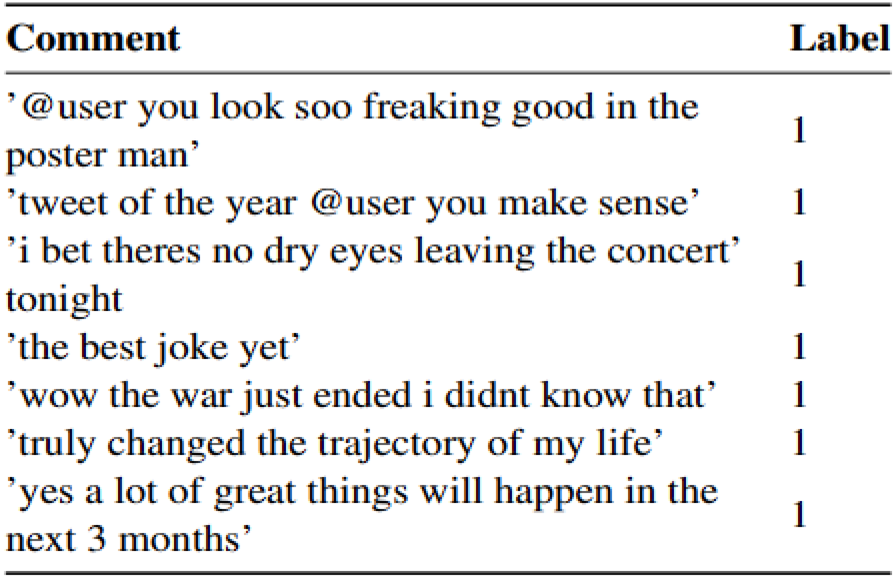
- Model has one unified idea of what ‘makes something’ sarcastic.

S3D – Our Weakly Supervised Dataset

S3D is a dataset of 100,000 tweets, making it the largest sarcasm annotated dataset of tweets, all labelled by our pre-trained BERTweet model.

Each tweet was pre-processed before being annotated.

The dataset contains 38,879 sarcastic tweets and 61,121 non-sarcastic tweets.



Comment	Label
’@user you look soo freaking good in the poster man’	1
’tweet of the year @user you make sense’	1
’i bet theres no dry eyes leaving the concert’ tonight	1
’the best joke yet’	1
’wow the war just ended i didnt know that’	1
’truly changed the trajectory of my life’	1
’yes a lot of great things will happen in the next 3 months’	1

Conclusion and Future Work

- A contribution of a small gold-standard and large silver-standard sarcasm detection dataset.

- An evaluation of multiple datasets and language models for sarcasm detection.

- Perform a more fine-grained annotation for sarcasm with subcategories

- Perform similar experiments for multimodal sarcasm detection

Thank You!