A deep-learning framework to detect sarcasm targets

[1]**Jasabanta Patro,** [2]**Srijan Bansal,** [3]**Animesh Mukherjee,**
Indian Institute of Technology Kharagpur, India – 721302
{[1]jasabantapatro, [2]srijanbansal97}@iitkgp.ac.in , [3]animeshm@cse.iitkgp.ac.in

Abstract

In this paper we propose a deep learning framework for sarcasm target detection in predefined sarcastic texts. Identification of sarcasm targets can help in many core natural language processing tasks such as aspect based sentiment analysis, opinion mining etc. To begin with, we perform an empirical study of the socio-linguistic features and identify those that are statistically significant in indicating sarcasm targets (p-values in the range $(0.05, 0.001)$). Finally, we present a deep-learning framework augmented with socio-linguistic features to detect sarcasm targets in sarcastic book-snippets and tweets. We achieve a huge improvement in the performance in terms of exact match and dice score as compared to the current state-of-the-art baseline.

1 Introduction

Computational sarcasm is a very well studied research area in computational linguistics (Joshi et al., 2017). Sentiment analysis and opinion mining of sarcastic texts are known to be difficult problems (Pang et al., 2008). For instance, in aspect based sentiment analysis, which deals with the identification of sentiment expressed toward different aspects or dimensions of the entities present in the text, it is very important to identify the sarcasm targets and sentiments toward them in the texts. Thus, if a user expresses a sarcastic utterance such as *"My laptop has an awesome battery life that lasts for 15 minutes"*, the tool should recognize that the speaker is expressing a negative sentiment toward the battery life of the laptop, even though, it has a positive sentiment word 'awesome' in it. Similarly the opinion mining tool should identify the negative opinion of the user expressed toward the entity "battery". Sarcasm target identification can also benefit natural language generation; for example, after detection of entity toward which a negative sentiment is expressed in a sarcastic text, a natural language generation system will have more context to generate a response. Similarly, a sentiment analysis tool will flag the sentiment in a sarcastic text toward the correct aspect of a product or the entity which can help to build a more accurate product review. In this paper we present a novel method for sarcasm target identification with the help of deep learning techniques in addition to a set of socio-linguistic features.

There is a lot of literature that deal with the sarcasm detection in text (Joshi et al., 2017), but only Joshi et al. (2018) have addressed the problem of sarcasm target identification. The sarcasm target is defined as the entity or situation that is being mocked or ridiculed at in the sarcastic text. Formally, the sarcasm target identification is defined as the task of building a system that takes a sarcastic text (book snippets, tweets etc.) as input, and either identifies a subset of words as sarcasm targets or outputs a fall-back label 'outside' if the target is not present in the text. For example in the sarcastic text *"I love to be ignored"*, the target is *"I"*. We consider two assumptions in this work as the same has been done in the baseline, – (a) every sarcastic text has at least one sarcasm target as this holds true by the definition of sarcasm, and, (b) the notion of sarcasm target is applicable for sarcastic texts only.

Sarcasm target identification is a difficult task, the primary reasons being,

- **Multiple candidate phrases**: There can be multiple target candidate phrases present in the sarcastic text. For example, in the sarcastic text, "The laptop heats up so much that I strongly recommend chefs to use it as a cook-top", the target candidates could be 'chefs', 'cook-top' and 'laptop'; however, only 'laptop' is ridiculed in this sentence.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6336–6342,
Hong Kong, China, November 3–7, 2019.

- **Multiple sarcasm targets**: There can be multiple sarcasm target phrases present in the sentence. For example, in the sarcastic text, "I used to be a middle-of-the-road kid, but now with my freaky looks I'm definitely an outsider. Hooray.", have two sarcasm targets, i.e., 'my freaky looks' and 'I'.
- **Absence of any target**: It is also possible that no sarcasm target is present at all in the sarcastic text. For example in the sarcastic text "Oh, and I suppose the apples ate the cheese." the sarcasm target has to be labelled as 'outside'.

The main contributions and results of this paper can be summarized as,

- An empirical study of the socio-linguistic features that are highly significant in identifying sarcasm targets.
- A novel deep learning framework augmented with socio-linguistic features to detect sarcasm targets in sarcastic texts. We achieve a huge improvement over Joshi et al. (2018) in sarcasm target detection in terms of the evaluation metrics – exact match and dice score.

In this paper our main motive was to establish that deep neural machinery can be effectively married with socio-linguistic features to detect sarcasm targets. This exercise was a proof of concept to show that this marriage is indeed useful. The code we developed for this work is made freely available[1].

2 Related works

Most of the papers in the area of computational sarcasm address the problem of sarcasm detection, i.e., classification of a text as sarcastic or non-sarcastic. Joshi et al. (2017) present a compilation of past works including the datasets, approaches, issues and trends in automatic sarcasm detection. They observe mainly three approaches to the sarcasm detection problem – semi-supervised extraction of sarcastic patterns (Tsur et al., 2010; Ptáček et al., 2014; Bouazizi and Ohtsuki, 2015; Riloff et al., 2013; Joshi et al., 2015), use of hashtag based supervision (Davidov et al., 2010; Abercrombie and Hovy, 2016), and use of contextual information for sarcasm detection (Hazarika et al., 2018; Wallace et al., 2014; Rajadesingan et al., 2015). Recently, Tay et al. (2018) presented an attention-based neural model to explicitly model contrast and incongruity. Kolchinski and Potts (2018) presented two methods for representing authors in the context of textual sarcasm detection; they show that augmenting a bidirectional RNN with these representations improves performance in sarcasm detection. Ghosh and Muresan (2018) did a thorough analysis of sarcasm markers in social media platforms like Twitter and Reddit; in their study they found that in Twitter while emoticons or emojies are the most discriminative markers to recognize sarcastic/ironic utterances, for Reddit the morphological markers (e.g., interjections, tag questions) are the most discriminative. In socio-linguistic literature even though there are many studies that observe propagation of hate speech (Ribeiro et al., 2018; Salminen et al., 2018) and abusive behaviour (Founta et al., 2018; Maity et al., 2018; Mathew et al., 2019a,b) in social media an in-depth analysis of how sarcastic message travels in social networks and how tweets around the targets behave is an area which social scientists need to investigate. To the best of our knowledge, only Joshi et al. (2018) addresses the problem of sarcasm target identification. This problem attempts to identify the entity toward which sentiment is expressed in a sentence which in turn can have a lot of applications. Our objective here is to leverage recent deep learning methods to escalate the overall performance on this task.

3 Dataset

We consider the dataset released by Joshi et al. (2018) for our experiments. The dataset has two types of sarcastic text - - book snippets and tweets. There are 224 book snippets and 506 tweets present in the data. The sarcasm targets present in these book snippets and tweets are manually annotated by three well experienced linguists who have at least five years of linguistic annotation experience for tasks such as sentiment analysis, word sense disambiguation and other related works. For the book snippets the average length of the sarcasm target is 1.6 words while it is 2.08 words for the tweets. The average length of the whole snippets is 27.74 words whereas for the tweets this is 12.97 words. For annotation, the annotators are given a bunch of sarcastic texts and asked to identify which words represent the target that the author is mocking? In case the annotators do not find specific words in the text that corre-

[1]Code:https://github.com/Srijanb97/Sarcasm_Target_Detection-EMNLP-

spond to a target, they label it as 'outside'.

4 Socio-linguistic features

In this section, we present various socio-linguistic features that show statistically significant differences between the words corresponding to the sarcasm targets and the rest of the words in the sarcastic text. The results are shown in Table 1. Some of the observations are,

- The distribution of location (LOC) and organisation (ORG) named entities are significantly different for the sarcasm target words compared to the other words ($p < 0.001$).
- The distribution of some of the POS tags (nouns, verbs, adjectives and modifiers) are significantly different for the target words compared to the other words.
- We calculate the LIWC[2] and Empath (Fast et al., 2016) category fractional distributions across the target and the other words in the snippets and tweets. Certain categories as noted in the table are significantly different. The LIWC and Empath dictionary has many pre-defined categories (e.g., 'social', 'family' etc.). Analysis using these dictionaries has been done on different collection of tweets in many past research (Fink et al., 2012; Schwartz et al., 2013; Maity et al., 2016) which forms our primary motivation for this study.

5 Methodology

The architecture of our proposed system is shown in Figure 1. The input to our system is a sarcastic text concatenated with a dummy word at the end of the sentence. We proceed with the hypothesis that each word is a potential candidate to be a sarcasm target. Thus for each word in the sentence we create three components, (i) left context, (ii) right context, and (iii) a word representation for itself. Suppose the input sarcastic sentence is represented by a sequence of words $w_1, w_2 ... w_N$, w_{N+1}, where w_{N+1} is a dummy word. We append a start token and an end token respectively at the beginning and the end of this sentence. These two tokens are never be considered as center word, but act as the left context for the first (w_1) word and the right context for last dummy word (w_{N+1}) respectively. Thus, for a word w_K, where $1 <= K <= N + 1$, the left context is defined as $[< start > w_1 : w_{K-1}]$ while the right context is defined as $[w_{K+1} : w_{N+1} < end >]$. Each word in the left context, right context and the central word are passed through an embedding layer to initialize them through pre-trained embeddings. We experiment with various pre-trained word embeddings like Glove, fast-text, elmo, BERT etc. The word representations are then passed to a LSTM or bidirectional LSTM (Bi-LSTM) layer or a target dependent LSTM (TD-LSTM) layer. In case of unidirectional LSTM (simple LSTM) layer, we keep the flow of hidden vectors in left context and right context as toward the center. Next we concatenate the hidden vectors of rightmost LSTM cell in left context, the central word LSTM cell hidden vector and the hidden vector of leftmost LSTM cell in right context, and pass them to a dense layer. In case of Bi-LSTM we concatenate both the forward and backward hidden vectors at each component before concatenating them again across the components. The dense representation is then concatenated with socio-linguistic features as we have obtained for the word w_k, and passed to a linear layer with sigmoid activation function, for the classification of the center word as sarcasm target or not.

[2]http://www.liwc.net/comparison.php

6 Experiments and results

6.1 Evaluation metrics

We consider two evaluation metrics – (i) exact match accuracy, and (ii) dice score, as has been also used in the baseline method (see (Joshi et al., 2018) for definitions).

6.2 Baselines

The baseline as described in Joshi et al. (2018) consists of two extractors joined by an integrator. The two extractors are (i) rule based, and (ii) statistics based. While the rule based extractor extracts candidate words for sarcasm target based on nine syntactic rules, the statistical extractor takes features such as lexical, POS tag, polarity, pragmatic features etc., and passes them to a classifier for the candidate word selection. The selected candidate words are then given as input to the integrator module, which is a hybrid 'AND' or 'OR' module, to select the final set of words as sarcasm targets.

6.3 Model setup and results

The setup: All the results reported are on 3-fold cross validation. Models were trained with Adam Optimizer having a learning rate $1e-5$ and a batch

Catrgory	Features for snippets	Features for tweets
NER	LOC(***), ORG(***)	LOC(***), ORG(***)
POS	CC(***), IN(***), MD(***), NNP(***), PRP(***), PRP$(***), RB(***), VB(***), VBD(***), VBP(***), VBZ(***), JJ(**), NN(**), NNS(**), TO(**), UH(**), WP(*)	IN(***), MD(***), NN(***), NNP(***), NNS(***), PRP$(***), RB(***), UH(***), VB(***), VBG(***), VBP(***), VBZ(***), CC(**), TO(**), VBD(**), JJ(*), JJR(*), JJS(*), RP(*)
Empath	Appearance(***), Feminine(***), White_Collar_Job(***), Beauty(*), Cleaning(*), Exotic(*), Farming(*), Occupation(*), Violence(*)	Affection(***), Ancient(***), Beach(***), Car(***), Cleaning(***), College(***), Domestic_Work(***), Driving(***), Eating(***), Economics(***), Family(***), Government(***), Health(***), Home(***), Love(***), Medical_Emergency(***), Meeting(***), Morning(***), Occupation(***), Ocean(***), Office(***), Optimism (***), Poor(***), Positive_Emotion(***), Reading(***), Sailing(***), School(***), Science(***), Sports(***), Swimming(***), Traveling(***), Water(***), White_Collor_Job(***), Work(***), Vehicle(***), Art(**), Blue_Collar_Job(**), Business(**), Clothing(**), Cooking(**), Dance(**), Exotic(**), Fabric(**), Feminine(**), Friends(**), Hygiene(**), Liquid(**), Musical(**), Law(**), Plant(**), Restaurant(**), Ship(**), Toy(**),Achievement(*), Air_Travel(*), Animal(*), Attractive(*), Childish(*), Children(*), Contentment(*), Exercise(*), Fashion(*), Furniture(*), Help(*), Leader(*), Leisure(*), Messaging(*), Music(*), Nervousness(*), Politics(*), Shopping(*), Sleep(*), Technology(*), Violence(*), Warmth(*), Weather(*), Wedding(*)
LIWC	Adverbs(***), Affect(***), AuxVb(***), CogMech(***), Conj(***), Excl(***), Humans(***), Negate(***), Past(***), Posemo(***), Ppron(***), Prep(***), Present(***), Pronoun(***), Relativ(***), SheHe(***), Social(***), Space(***), Tentat(***), They(***), Time(***), Verbs(***), Cause(**), Discrep(**), Funct(**), Future(**), Hear(**), Incl(**), Motion(**), Achiev(*), Article(*), Body(*), Death(*), Family(*), Insight(*), Percept(*), Quant(*)	Adverbs(***), Affect(***), AuxVb(***), Body(***), CogMech(***), Conj(***), Discrep(***), Excl(***), Family(***), Funct(***), Health(***), Home(***), Humans(***), Insight(***), Leisure(***), Past(***), Posemo(***), Ppron(***), Prep(***), Present(***), Pronoun(***), Sad(***), Social(***), Verbs(***), Work(***), Assent(**), Bio(**), Cause(**), Future(**), Ingest(**), Negate(**), Quant(**), Sexual(**), Certain(*), Friends(*), I(*), Incl(*), SheHe(*)

Table 1: Features that are significantly different for the target words compared to the other words. Features for which $p < 0.05$ ($p < 0.01$ and $p < 0.001$) is represented by (*) {(**), (***)} respectively.

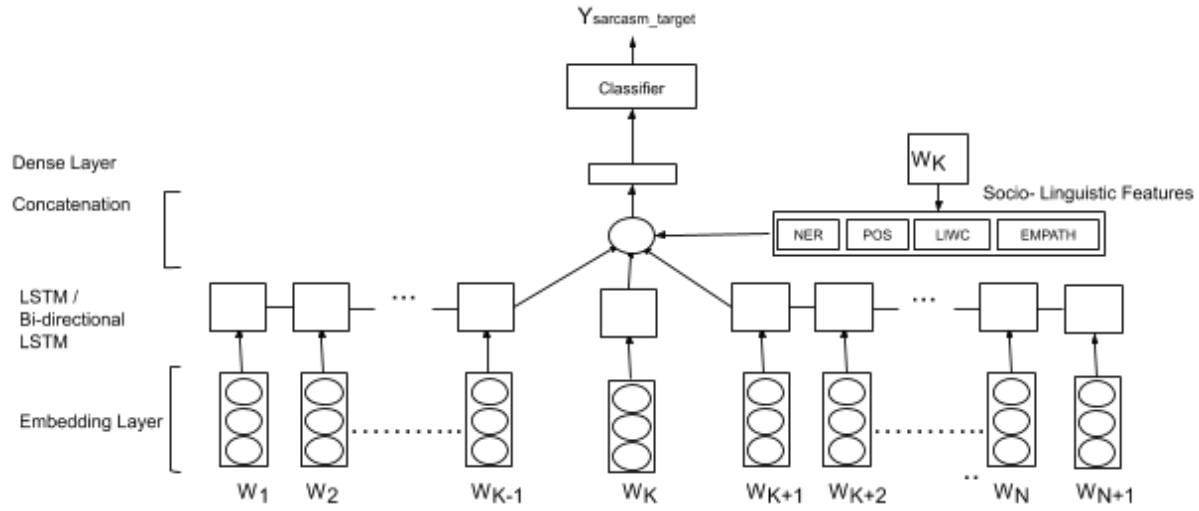

Figure 1: Architecture of the proposed system. w_K is the center word to be classified as target or not, $[< start > w_1 : w_{K-1}]$ is the left context and $[w_{K+1} : w_{N+1} < end >]$ is the right context.

size of 64. Best results were obtained with the Elmo embedding as initialization.

Results: We report the exact match and the dice score obtained from different variants of our model and compare them with the baseline in Table 2. We note that all the variants of our model outperforms the baseline approach by a large margin. Among non-augmented models, the variant with Bi-LSTM layer performs the best in most of the metrics. The dice score for the book snippets data is best when TD-LSTM is used. The augmentation of socio-linguistic features (BiLSTM layer + slf) the performance further leading to the establishment of new state-of-the-art in sarcasm target detection. In addition, for our model variants we also report the macro and micro F1-scores in Table 3. Once again the Bi-LSTM is indicative of the best performance in majority of cases. The macro-F1 and micro-F1 increases further for book snippets when augmented with socio-linguistic features.

Model	EM_T	DS_T	EM_S	DS_S
Baseline: AND	13.45	20.82	16.51	21.28
Baseline: OR	9.09	39.63	7.01	32.68
LSTM layer	26.01	82.84	23.37	87.57
Bi-LSTM layer	29.48	84.04	30.14	87.66
TD-LSTM layer	26.35	82.27	25.97	87.71
Bi-LSTM layer + slf	**30.12**	**84.11**	**31.17**	**88.16**

Table 2: Comparison our models with the baseline. T: tweets, S: snippets, sl: socio-linguistic features.

Model	$F1_T^\mu$	$F1_T^M$	$F1_S^\mu$	$F1_S^M$
LSTM layey	46.79	51.04	39.31	43.72
Bi-LSTM layer	**54.74**	**59.19**	48.27	45.36
TD-LSTM layer	41.59	49.28	47.23	46.42
Bi-LSTM layer + slf	50.30	55.76	**48.30**	**48.54**

Table 3: Macro (M) and micro (μ) F1 scores for the different models.

7 Conclusion

In this work, we have presented a deep learning model for sarcasm target identification. We outperform the only available baseline by a large margin. We identify various socio-linguistic features that differentiate the target text from the rest of the snippet/tweet. When these additional socio-linguistic features are fused into our deep learning framework they seem to improve performance for both snippets and tweets establishing new state-of-the-art for this problem.

References

Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113.

Mondher Bouazizi and Tomoaki Ohtsuki. 2015. Sarcasm detection in twitter:" all your products are incredibly amazing!!!"-are they really? In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM.

Clay Fink, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring gender from the content of tweets: A region specific example. In *ICWSM '12*, pages 459–462.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *ICWSM*.

Debanjan Ghosh and Smaranda Muresan. 2018. "with 1 follower i must be awesome : P". exploring the role of irony markers in irony recognition. In *ICWSM*.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. In *COLING*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50:73:1–73:22.

Aditya Joshi, Pranav Goel, Pushpak Bhattacharyya, and Mark Carman. 2018. Sarcasm target identification: Dataset and an introductory approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 757–762.

Y. Alex Kolchinski and Christopher Potts. 2018. Representing social media users for sarcasm detection. In *EMNLP*.

Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2018. Opinion conflicts: An effective route to detect incivility in twitter. In *ACM CSCW*.

Suman Kalyan Maity, Ritvik Saraf, and Animesh Mukherjee. 2016. #bieber + #blast = #bieberblast: Early prediction of popular hashtag compounds. In *CSCW '16*, pages 50–63.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019a. Spread of hate speech in online social media. In *ACM WebSci*.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019b. Thou shalt not hate: Countering online hatespeech. In *ICWSM*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 97–106. ACM.

Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira. 2018. Characterizing and detecting hateful users on twitter. In *ICWSM*.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714.

Joni O. Salminen, Hind Almerekhi, Milica Milenkovic, Soon-Gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *ICWSM*.

Hansen Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Richard E. Lucas, Megha Agrawal, Gregory J. Park,

Shrinidhi K. Lakshmikanth, Sneha Jha, Martin E. P. Seligman, and Lyle Ungar. 2013. Characterizing geographic variation in well-being using tweets. In *ICWSM '13*, pages 583–591.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *ACL*.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsma great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 512–516.