Survey of the State of the Art on Sarcasm Detection

Eddie Guo
University of California, Berkeley
eddieguo@berkeley.edu

Abstract

As a subfield of sentiment analysis, automatic sarcasm detection has gained increasing attention from the Natural Language Processing Community. In this survey, I will present an overview of the task, identify key datasets, and introduce state-of-the-art techniques to detect sarcasm.

1 Introduction

Sarcasm, according to Merriam-Webster, refers to the use of words that mean the opposite of what the person really wants to say, with the intention of insulting someone, or to show irritation, or just being funny. Rather than a direct insult, sarcasm is usually under the disguise of the use of positive words, which could only cause more harm to the audience. For example, saying "it is a perfect movie for people who can't fall asleep" to describe a movie as very boring is using sarcasm. Nowadays, irony is prevalent on online platforms such as Twitter and Reddit, and detecting sarcasm automatically is helpful for opinion mining, online harassment detection, and customer service.

While sarcasm has received growing attention from the natural language processing community, it is an exceptionally challenging task. As mentioned by Prof. Bamman, "Sarcasm detection is a good example of what we call AI-complete." It require[1]s knowledge of the entire world, including the speaker's past experiences in social interactions, the context of the current conversation, the general attitude of the speaker, and the relationship between the speaker and listener. The need for so much information makes it difficult to "code" sarcasm.

In recent years, researchers have proposed several novel approaches to automatically detect sarcasm, including deep neural networks and multi-modal NLP. In the rest of this survey, I will discuss datasets created or available for the task of sarcasm detection and state-of-the-art approaches for this task.

2 Dataset

The fundamental step for every NLP task is always data acquisition. For the task of sarcasm detection, there are mainly two modes for collecting data: API (Application Programming Interface) and accessible datasets like MUStARD (Multi-Modal Sarcasm Detection Dataset), SARC (Self-Annotated Reddit Corpus), SemEval (Semantic Evaluation), Internet Argument Corpus (IAC) (Verma et al., 2021).

Twitter and Reddit are two major data sources for sarcasm detection. For Twitter, tweets with "#sarcasm" are self-labeled as sarcasm by the author so researchers could conveniently obtain a large corpus of sarcastic texts (Tay et al., 2018). Similar to Tweets, sarcastic posts on Reddit are extracted via '/s' which are marked by the

[1] Word Count: 1927

author as sarcasm. Another major dataset is political debates from Internet Argument Corpus (IAC), which has been human-annotated for sarcasm (Tay et al., 2018). This data contains mainly long text, which is designed for research in political debate in online forums.

To enable the fast-growing field of multi-modal NLP, a new dataset (MUStARD) consisting of manually annotated short videos is introduced for sarcasm detection (Castro et al., 2019). They collected videos from TV shows such as Friends, The Big Bang Theory, The Golden Girls, and Sarcasmaholics Anonymous and humanly annotated the videos for sarcasm. The final dataset consists of 690 videos with an even number of sarcastic and non-sarcastic labels (Castro et al., 2019).

3 Approaches

In this section, I investigate several approaches for automatic sarcasm detection. In general, these approaches can be classified into rule-based, statistical, deep learning based, and multi-modal approaches. I will put more emphasis on findings in most recent deep neural network and multi-modal architectures.

3.1 Rule-Based Approaches

Rule-Based approaches attempt to detect sarcasm through direct evidence or indicators in the text without the need for training or using machine learning models. For example, (Maynard & Greenwood, LREC 2014) found that hashtags in tweets are particularly indicative of sarcasm. Thus they developed a hashtag tokenizer to split hashtags made of concatenated words and compiled a set of rules to determine sarcasm. One example of such rules is if the underlying sentiment of the hashtag is inconsistent with the rest of the tweet, then it will be labeled as sarcasm (Joshi et al., 2017).

3.2 Feature Sets for Statistical Sarcasm Detection

As discussed in Bamman and Smith (2015), past machine learning approaches treat the task as a text categorization problem and rely solely on the lexical indicators and linguistic features of the text itself. Features used in these models are mostly generated from the immediate text without any contextual information. However, as suggested by Bamman and Smith (2015), including extra-linguistic features from the context of the text could increase model accuracy while providing insight into sarcasm in conversation and more complex social interactions. Salient features used in their findings include the author and audience features, which encode their profile information and historical data on topics and sentiment. They also include environment features, which model the interaction between target tweets and their responses. Their result shows the gain in accuracy when adding more contextual features and their top-performing model uses a combination of tweet, author, audience, and response features (Bamman & Smith, 2015).

The importance of extralinguistic features is also noticed by Kolchinski and Potts (2018), who focus specifically on author features. They explored two methods to represent authors: a Bayesian approach that directly represents authors' tendency to be sarcastic and a dense embedding approach that can learn interactions between the author and the text (Kolchinski & Potts, 2018).

Another set of interesting features is proposed by Bouazizi and Otsuki Ohtsuki (2016) in a novel pattern-based approach. The author further breaks down sarcasm into 3 categories based on the purpose of sarcasm: Sarcasm as wit (used sarcasm with the purpose of being funny), Sarcasm as whimper (showing how annoyed or angry the person is), and Sarcasm as evasion (used when the person wants to avoid giving an answer).

3.3 Deep Learning Based Approaches

Deep neural nets are gaining increasing attention in the NLP community for sentiment analysis and sarcasm detection. As demonstrated by Zhang et al. (2016), there are two main advantages of neural models. First, neural networks are capable of extracting features automatically from the input text without the need for feature engineering in discrete models. Moreover, neural features are able to capture long-range and more subtle semantic patterns in the text, which are difficult to encode using discrete feature templates (Zhang et al., 2016)**.** Also, neural models use word embeddings that are pre-trained from large-scale raw texts and thus are capable of avoiding the feature sparsity problem in statistical machine learning models.

As introduced in Jaiswal (2020), there are a couple of Pre-trained Language Representation Models (PLRMs) being created, which are trained on large corpora and can be fine-tuned for specific tasks. Four commonly used PLRMs are Embeddings from Language Models (ELMo), Universal Sentence Encoder (USE), Bidirectional Encoder Representations from Transformers (BERT), and Robustly Optimized BERT Approach (RoBERTa). Among these contextual embeddings, RoBERTa turns out to yield the best performance across different tasks (Jaiswal, 2020). RoBERTa retrains BERT with improved training methodology. Specifically, RoBERTa removes the Next Sentence Prediction (NSP) task from BERT's pre-training and introduces dynamic masking so that the masked token changes during the training epochs. The model is also trained on more data (160 GB vs. 16 GB) (Jaiswal, 2020).

To improve the performance of deep neural networks, researchers have investigated linguistic phenomena related to sarcasm. One finding shows that sarcasm relies heavily on the contrast between words and phrases in a sentence (Tay et al., 2018). For example, in a sarcastic sentence "perfect movie for people who can't fall asleep, the authors identify word pairs {movie, asleep} that characterize sarcasm. Thus, they propose a multiple-dimensional intra-attention recurrent neural network that attempts to capture "contrast" or "incongruity" in the text. The concept of "incongruity" is also explored by Joshi et al. (2015), where the author introduces inter-sentential incongruity for sarcasm detection, and Riloff et al. (2013) also identify contrasting contexts using phrases learned through bootstrapping methods.

3.4 Multi-Modal Sarcasm Detection

Most of the previous work in sarcasm detection focuses only on textual data. However, there has been a growing interest in the NLP community to incorporate data from other modalities to build multi-modal classifiers for sarcasm detection. To see how images or video data might be useful to detect irony, consider the following scenario (Pan et al., 2020). The text "such a packed game. It is amazing we even got a seat" does not seem to imply sarcasm. However, the image followed by this message shows an almost stadium. The contrast between the image and text immediately points out the sarcasm in the context.

As a first step toward multi-modal sarcasm detection, Castro et al. (2019) proposed a new sarcasm dataset Multimodal Sarcasm Detection Dataset (MUStARD), compiled from popular TV shows. As introduced in Section 2, MUStARD consists of audiovisual utterances annotated with sarcasm labels (Castro et al., 2019). They then extract features from the dataset and perform several experiments. The result shows that multimodal variants significantly outperform their unimodal counterparts, with relative error rate reductions of up to 12.9% (Castro et al., 2019).

Cai et al. (2019) further refine multi-modal sarcasm detection by deeply fusing the three modalities of image, attribute, and text rather than naive concatenation. Text modality first undergoes early fusion where the attribute features are used to initialize a bi-directional LSTM network (Cai et al., 2019). Then all three modalities go through representation fusion, where they are transformed from raw vectors into reconstructed feature vectors (Cai et al., 2019). Finally, a modality fusion layer takes the weighted average of the vectors and output fused vector for classification (Cai et al., 2019).

A concept that is often ignored in past work is incongruity, the distinction between reality and expectation (Pan et al., 2020). Inspired by the incongruity manifested within and between modalities, Pan et al. (2020) proposed a BERT architecture-based model focusing on both intra and inter-modality incongruity for multi-modal sarcasm detection. To be specific, they designed the inter-modality attention to capture the incongruity between modalities and apply the co-attention mechanism to model the incongruity between text and hashtags as the intra-modality incongruity (Pan et al., 2020).

Building upon the achievements in multi-modality sarcasm detection so far, Chauhan et al. (2020) further explore the correlation between sarcasm and sentiment and emotion. To illustrate the motivation of their research, they showed an example of a man having dessert. Purely from the utterance of the speaker, he seems to enjoy the dessert. However, closely examining his facial expressions, the underlying sentiment of the speaker is negative, thus his utterance is showing sarcasm. Therefore, multi-modal data with a combination of video and text helps us understand the sentiment and emotion of the speaker, providing more context in addition to the utterance. To implement this approach, they two attention mechanisms: Inter-segment Inter-modal Attention and Intra-segment Inter-modal Attention to better combine the information across modalities and classify sarcasm, sentiment, and emotion more accurately (Chauhan et al., 2020).

4 Conclusion

Sarcasm detection as a subfield of NLP has grown significantly in recent years. Different from typical sentiment analysis tasks, sarcasm is much harder to detect due to its contextual nature. It also requires extensive social interaction experiences to perceive sarcasm, which makes it exceptionally challenging for machines to understand. Bearing the difficulty of the task, there have been several approaches to automatic sarcasm detection, including rule-based, statistical, deep learning based, and most state-of-the-art multi-modal techniques. The current trend of research in the field point to the incorporation of new extralinguistic and contextual features, and even data from different modalities such as image and video, into deep neural network architecture. Moreover, topics like irony and sarcasm detection, types of negations, and multipolarity that are mentioned in Verma et al. (2021) are also potential topics that could provide insight into refining sarcasm detection models.

5 References

1. Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1383–1392, Online. Association for Computational Linguistics.

2. Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis. In Proceedings of the 58th Annual Meeting of the Association for

Computational Linguistics, pages 4351–4360, Online. Association for Computational Linguistics.

3. Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

4. Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

5. Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with Sarcasm by Reading In-Between. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

6. Y. Alex Kolchinski and Christopher Potts. 2018. Representing Social Media Users for Sarcasm Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1115–1121, Brussels, Belgium. Association for Computational Linguistics.

7. Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained Attention Network for Aspect-Level Sentiment Classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3433–3442, Brussels, Belgium. Association for Computational Linguistics.

8. Aniruddha Ghosh and Tony Veale. 2017. Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 482–491, Copenhagen, Denmark. Association for Computational Linguistics.

9. Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

10. 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), pages 757–762, Beijing, China. Association for Computational Linguistics.

11. Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection. ACM Computing Surveys, 50(5), 1–22. https://doi.org/10.1145/3124420

12. Verma, P., Shukla, N., & Shukla, A. P. (2021). Techniques of SARCASM DETECTION: A Review. 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE).

https://doi.org/10.1109/icacite51222.2021.9404585

13. Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are Word Embedding-based Features Useful for Sarcasm Detection?. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1006–1011, Austin, Texas. Association for Computational Linguistics.

14. Bamman, D., & Smith, N. (2021). Contextualized Sarcasm Detection on Twitter. Proceedings of the International AAAI Conference on Web and Social Media, 9(1), 574-577. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14655

15. 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 (WSDM '15). Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/2684822.2685316

16. M. Bouazizi and T. Otsuki Ohtsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter," in IEEE Access, vol. 4, pp. 5477-5488, 2016, doi: 10.1109/ACCESS.2016.2594194.

17. P. Verma, N. Shukla and A. P. Shukla, "Techniques of Sarcasm Detection: A Review," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 968-972, doi: 10.1109/ICACITE51222.2021.9404585.

18. Diana Maynard and Mark Greenwood. 2014. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis.. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 4238–4243, Reykjavik, Iceland. European Language Resources Association (ELRA).

19. 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, Seattle, Washington, USA. Association for Computational Linguistics.

20. Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet Sarcasm Detection Using Deep Neural Network. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.

21. Nikhil Jaiswal. 2020. Neural Sarcasm Detection using Conversation Context. In Proceedings of the Second Workshop on Figurative Language Processing, pages 77–82, Online. Association for Computational Linguistics.

22. Yaghoobian, Hamed & Arabnia, Hamid & Rasheed, Khaled. (2021). Sarcasm Detection: A Comparative Study.

23. Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying Sarcasm in Twitter: A Closer Look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 581–586,

Portland, Oregon, USA. Association for Computational Linguistics.

24. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

25. Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 606–615, Austin, Texas. Association for Computational Linguistics.

26. *David Bamman explains why robots* [2]*don't understand sarcasm*. UC Berkeley School of Information. (n.d.). Retrieved May 10, 2022, from https://www.ischool.berkeley.edu/news/2016/david-bamman-explains-why-robots-dont-understand-sarcasm