Sarcasm detection using nlp*

*Note: Sub-titles are not captured in Xplordocse and should not be used

1st Md. Tawsifur Rahman
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Md. Siam Azad Sadman
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Md Sabbir Hossain
CSE dept of Brac University(of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4thFarah Binta Haque
CSE dept of Brac University(of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5thAnnajiat Alim Rasel
CSE dept of Brac University(of Aff.)
name of organization (of Aff.)
City, Country
annajiat@gmail.com

6th Md Mahbub Alam Prithibi
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

7thSayma Akter Lubna
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—The emergence of Natural Language Processing (NLP) offers a new lens to decode the subtleties of sarcasm. This paper explores the complexities of sarcasm detection and the promise that NLP holds in unlocking its secrets. We delve into the linguistic, sentimental, and contextual intricacies that make sarcasm so elusive. We then showcase how NLP techniques, ranging from machine learning to deep learning, are being employed to identify sarcasm markers and bridge the gap between literal and intended meaning. In this study, we aim to minimise the amount of data required to achieve near-optimal performance in sarcasm detection. We employ two datasets composed of concise texts: iSarcasm- tweets and SARC- Reddit post comments. To reduce the size of the training dataset, we implemented a data pruning method. Additionally, we investigate various loss functions.**

***Index Terms*—nlp, sarcasm, sarcasm detection**

I. Introduction

Sarcasm, that sly, playful friend of irony, has long eluded the grasp of machines. But with the rise of Natural Language Processing (NLP), a new dawn is breaking. We're now equipped with tools to delve into the depths of human communication and finally understand the subtle nuances of sarcasm.

Imagine a world where machines can decipher the "not really" behind a "great weather, huh?". A world where robots don't take your "loving" punch in the arm literally. This is the promise of sarcasm detection, and NLP is the key.

II. Literature Review

We opted to fine-tune the sarcasm detection model based on RoBERTa as proposed by Abaskohi et al. [2022] [1]. In their study, the authors explored various models, including SVM-based, BERT-based, Attention-based, and LSTM-based, along with different data augmentation strategies such as generative-based and mutation-based approaches. They attempted to expand the training set by augmenting the iSarcasm dataset with Sentiment140 (Go et al. [2009]) [5] and Sarcasm Headlines Dataset (Misra and Arora [2019]), but notably did not augment with SARC, our chosen dataset.

The authors found that fine-tuning RoBERTa for sentiment classification, without data augmentation and further fine-tuning only on the iSarcasm dataset, achieved the best performance with an F1 score of 0.414. However, this fell short of the RoBERTa model fine-tuned for sarcasm detection by Hercog et al. [2022] [2], which utilized a training set comprising iSarcasm and a subset of SARC, achieving a higher F1 score of 0.526 on iSarcasm.

The iSarcasm dataset (Oprea and Magdy [2019]) [3] is unique as it consists of tweets labeled by the authors themselves as sarcastic or non-sarcastic, focusing on intended sarcasm. This labeling approach differs from datasets that label perceived sarcasm, and models trained on such datasets performed poorly on iSarcasm. The dataset comprises 777 sarcastic and 3,707 non-sarcastic tweets.

The Self-Annotated Reddit Corpus (Khodak et al. [2017]) [4] is a dataset of Reddit comments labeled for perceived sarcasm, using the "/s" token convention. However, SARC is noisy due to false negatives (sarcastic comments without "/s") and false positives (non-sarcastic comments with "/s"). We use a balanced subset of SARC as our starting point, attempting to prune it while maintaining similar performance.

Identify applicable funding agency here. If none, delete this.

The pruning process involves adopting the student-teacher setting for perceptron learning described by Sorscher et al. [2022] [7].

Sorscher et al. [2022] [7] explored ways to select training samples to achieve a reduction in test error better than scaling laws. They showed that optimal pruning strategies depend on the amount of initial data, and exponential scaling is possible with respect to pruned dataset size by choosing an increasing Pareto optimal pruning fraction based on the initial dataset size. These results apply not only to training from scratch but also to fine-tuning, which aligns with our setting.

We also experiment with different loss functions and their impact on task performance. Abaskohi et al. [2022] [1] used Cross Entropy Loss, and to address class imbalance in the iSarcasm dataset, we consider Weighted Cross Entropy Loss. Squared Hinge Loss, known for good performance in classification tasks, is also explored. Janocha and Czarnecki [2017] [8] compare various losses and recommend using squared hinge loss when operating in a data-scarce regime, as it converges faster and exhibits more robustness to noise in training set labeling and input space. This robustness is particularly beneficial when augmenting with SARC, a noisier dataset than iSarcasm labeled using distant supervision.

III. Methodology

We employ the student-teacher perceptron learning technique developed by Sorscher et al. [2022] [7]. This method utilizes a teacher model to categorize data points, retaining only those with high (low) misclassification confidence when the teacher model was trained on abundant (scarce) data. In our scenario, the initial baseline model trained on the iSarcasm dataset serves as the teacher model for classifying SARC training samples. We then arrange the teacher model's misclassifications based on confidence (logit score). A misclassified example with low confidence (logit score) is considered an easy example, while one with high confidence (logit score) is considered a hard example. We create three new datasets from these ordered data points: iSarcasm + 10,000 SARC training examples misclassified with the lowest confidence (Easy SARC Examples), iSarcasm + 10,000 SARC training examples misclassified with the highest confidence (Hard SARC Examples), and iSarcasm + 10,000 random SARC examples. The model is then fine-tuned on these new datasets, and the results are presented in Table 1. We observe that using the dataset with easy SARC examples outperforms the baseline, consistent with Sorscher et al.'s [2022] [7] findings that "The optimal pruning strategy changes depending on the amount of initial data; with abundant (scarce) initial data, one should retain only hard (easy) examples." Our case involves scarce initial data due to the small iSarcasm dataset, and we observe that adding easy examples to the dataset enhances model performance.

References

[1] Amirhossein Abaskohi, Arash Rasouli, Tanin Zeraati, and Behnam Bahrak. Utnlp at semeval-2022 task 6: A comparative analysis of sarcasm detection using generative-based and mutation-based data augmentation. arXiv preprint arXiv:2204.08198, 2022.

[2] Maciej Hercog, Piotr Jaroński, Jan Kolanowski, Paweł Mieczyński, Dawid Wiśniewski, and Jedrzej Potoniec. Sarcastic roberta: A roberta-based deep neural network detecting sarcasm on twit- ter. In International Conference on Big Data Analytics and Knowledge Discovery, pages 46–52. Springer, 2022.

[3] Silviu Oprea and Walid Magdy. isarcasm: A dataset of intended sarcasm. arXiv preprint arXiv:1911.03123, 2019.

[4] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. arXiv preprint arXiv:1704.05579, 2017.

[5] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12):2009, 2009.

[6] Rishabh Misra and Prahal Arora. Sarcasm detection using hybrid neural network. arXiv preprint arXiv:1908.07414, 2019.

[7] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. arXiv preprint arXiv:2206.14486, 2022.

[8] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. arXiv preprint arXiv:1702.05659, 2017.