Context Based Sarcasm Detection

Project Report R1:

Team Members:

1. Arun Gaonkar (unity ID: agaonka)
2. Manasi Sanjay Ghosalkar (unity ID: mghosal)

1. Introduction

Sarcasm is a convoluted form of expression where meaning is conveyed implicitly. Recognizing sarcasm in a conversation is important for understanding the actual meaning and sentiment conveyed. The intended meaning is often different from what can be perceived by naive systems. This poses problems to many natural language systems, in particular, summarization and dialogue systems. It also has applications in understanding sentiment and opinions in modern communication channels such as tweets, comments and chatbots. Sarcasm detection remains a difficult task and this can be partly attributed to the fact that sarcasm relies heavily on the context of the dialogue taking place. For this project, we plan to implement sarcasm detection on news headlines by using text from the article it pertains to as context.

2. Literature Survey

Sarcasm Detection has been explored by Davidov et al. (2010) by applying semi-supervised techniques like SASI (Semi-supervised Sarcasm Identification Algorithm) by (Tsur et al., 2010) along with feature extraction on two different datasets, consisting of tweets and product reviews. Gonzalez-Ibanez et al. (2011) have also approached this task using supervised machine learning methods such as SVM and Logistic Regression. To address this problem in social-media domains, there have been works that deal with sarcasm in multi-modal settings such as texts and images (Schifanella et al. (2016) and Cai et al. (2019)). Recurrent Neural Networks like LSTM (Long Short Term Memory) networks with sentence-level attention have been used by Ghosh et al. (2018) by taking into consideration the sentences as well as the conversation context that the sentence responds to. This work shows that modeling the conversation context using a multiple-LSTM architecture yields better results in sarcasm detection as compared to just modeling the text in question. This argument is supported by other papers such as by Wallace et al. (2014) which shows that even human annotators have to rely on additional context in order to classify or deduce ironic content. This has also been applied by others such as Pant et al. (2020) using transformer model based approaches.

3. Dataset:

We plan to use the News Headlines for Sarcasm Detection dataset obtained from Kaggle. The dataset is a json file that contains 28619 records. Each of these records are in the following format:

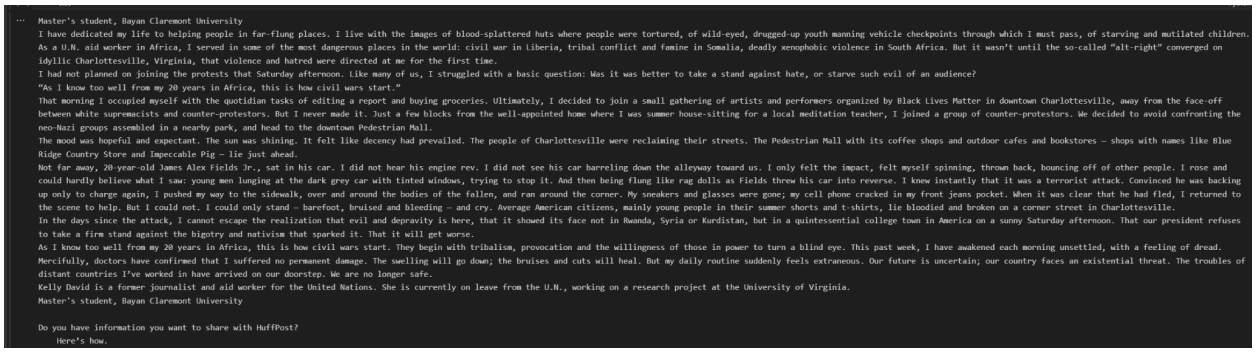
{
    “is_sarcastic”: 1,
    “headline” : “study: 83% of marathon spectators only attend for sick thrill of watching fellow man suffer”
    “article_link”:
"https://www.theonion.com/study-83-of-marathon-spectators-only-attend-for-sick-1828946111"
}

In each record,

1. The first item is the label. It is labeled 1 for sarcastic sentences and 0 for non-sarcastic sentences.
2. headline is the news article headline in the string format.
3. Last item is the article link, which hyperlinks to news articles in the form "https://…". To get the context of the news, the article link can be used.

From the dataset description it is verified that the 'headline' from the article link and the headline from the record are matching. Most of the news articles are from sources like "TheOnion.com", "huffingtonpost.com". The Huffington Post is an American news website and political blog. The Onion is an American digital media company and newspaper organization that publishes satirical and sarcastic articles. For example, in the above example the article_link theOnion.com navigates to this website. And another example from huffpost.com naviages to
https://www.huffpost.com/entry/charlottesville-car-attack_b_5995ddd3e4b01f6e801ce11a

Each article has a brief description. We are using BeautifulSoup and html parser to extract the paragraphs in the article. And these paragraphs are used as context for the headlines.



For example, the above image shows the parsed result of the hyperlink "https://www.huffpost.com/entry/charlottesville-car-attack_b_5995ddd3e4b01f6e801ce11a".
Some articles have multiple paragraphs, including non-related contents as well. To mitigate irrelevant details, we are considering only the first paragraph. (Extracting condition is that a paragraph will be considered only if it has more than 2 sentences in it.) The intuition behind this is that usually abstracts and overviews are mentioned in the first paragraph.
In the end, the dataset will consist of labels, article headline, and the article description. The task will be predicting the label (is_sarcastic or not) of the headline based on the article description.

4. Methods and Evaluation:

Initially embeddings (word2Vec,GloVe, paragraph2vec) are being used to get the vector representation.
Since the task involves extracting the context from a paragraph, sequence of sentences are important. Previous studies have shown that LSTM is best for sequential analysis. So initially LSTM will be tested for this application.
Another observation is that transformers in NLP are used to solve sequence to sequence tasks while incorporating long-range dependencies. This can be advantageous because some contexts have long paragraphs So we will be using a transformers based approach as well. For transformers, we are further planning to use pre-trained models. This is because the dataset might not be big enough to train all the parameters of models. Pre-trained models can be used to overcome this limitation. So will be implementing BERT and RoBERTa as well.

References

1. Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and Amazon. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics https://aclanthology.org/W10-2914/
2. Roberto Gonzalez-Ibanez, 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. https://aclanthology.org/P11-2102/
3. Debanjan Ghosh, Alexander Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. Computational Linguistics, 44:1–56. https://direct.mit.edu/coli/article/44/4/755/1620/Sarcasm-Analysis-Using-Conversation-Context
4. Rossano Schifanella, Paloma Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms.(https://dl.acm.org/doi/10.1145/2964284.2964321)
5. Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in twitter with hierarchical fusion model. pages 2506–2515. https://aclanthology.org/P19-1239/
6. Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In ACL.
7. Misra, Rishabh and Prahal Arora. "Sarcasm Detection using Hybrid Neural Network." arXiv preprint arXiv:1908.07414 (2019).
8. Misra, Rishabh and Jigyasa Grover. "Sculpting Data for ML: The first act of Machine Learning." ISBN 9798585463570 (2021).
9. Tanvi Dadu and Kartikey Pant. 2020. Sarcasm Detection using Context Separators in Online Discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 51–55, Online. Association for Computational Linguistics.