Sarcasm Detection in News Headlines - LSTM

Context:

Past studies in Sarcasm Detection mostly make use of Twitter datasets collected using hashtag based supervision but such datasets are noisy in terms of labels and language. Furthermore, many tweets are replies to other tweets and detecting sarcasm in these requires the availability of contextual tweets.

To overcome the limitations related to noise in Twitter datasets, this Headlines dataset for Sarcasm Detection is collected from two news website. TheOnion aims at producing sarcastic versions of current events and we collected all the headlines from News in Brief and News in Photos categories (which are sarcastic). We collect real (and non-sarcastic) news headlines from HuffPost.

Since news headlines are written by professionals in a formal manner, there are no spelling mistakes and informal usage. This reduces the sparsity and also increases the chance of finding pre-trained embeddings.

Furthermore, since the sole purpose of The Onion is to publish sarcastic news, we get high-quality labels with much less noise as compared to Twitter datasets.

Unlike tweets which are replies to other tweets, the news headlines we obtained are self-contained. This would help us in teasing apart the real sarcastic elements.

Attributes:

Each record consists of three attributes:

- `is_sarcastic`: 1 if the record is sarcastic otherwise 0
- `headline`: the headline of the news article
- `article_link`: link to the original news article. Useful in collecting supplementary data

Reading the data

In python, data can be read using the following function:

```
def parseJson(fname):
    for line in open(fname, 'r'):
        yield eval(line)
```

Example usecase: `data = list(parseJson('./Sarcasm_Headlines_Dataset.json'))`

Data reference: https://github.com/rishabhmisra/News-Headlines-Dataset-For-Sarcasm-Detection

Citation:

```
@article{misra2019sarcasm,
  title={Sarcasm Detection using Hybrid Neural Network},
```

```
  author={Misra, Rishabh and Arora, Prahal},
  journal={arXiv preprint arXiv:1908.07414},
  year={2019}
}

@book{book,
author = {Misra, Rishabh and Grover, Jigyasa},
year = {2021},
month = {01},
pages = {},
title = {Sculpting Data for ML: The first act of Machine Learning},
isbn = {978-0-578-83125-1}
}
```

Key asks:

- Predict the number of positive and negative reviews based on sentiments by using different classification models.