Sarcasm Detection with Neural Networks

In this work, a binary classification model was designed that processes individual word with natural language processing methodologies to predict the presence of sarcasm in News Headlines.

NLP Terminologies

- *Tokenization* is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens.
- When we pre-process the text data, not all the input sentences are of same length, some of the sentences are longer or shorter. The solution to this problem of different lengths of inputs is *padding* with zero to a fixed length.

  padding='post': add the zeros at the end of the sequence to make the samples of the same size.
- *Word embedding* is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. A word embedding is a learned representation for text where words that have the same meaning have a similar representation.

DATASET

News Headlines Dataset for Sarcasm Detection

Data Source: https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection?select=Sarcasm_Headlines_Dataset.json

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 News 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*.

This new dataset has following advantages over the existing Twitter datasets:

- 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 *TheOnion* 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.

Content

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