

Project

Natural Language Processing

Description

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.

In this hands-on project, the goal is to build a model to detect whether a sentence is sarcastic or not using Bidirectional LSTMs.

Dataset

News Headlines dataset for Sarcasm Detection

The dataset is collected from two news websites, theonion.com and huffingtonpost.com. This new dataset has the 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 that reply to other tweets, the news headlines 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

Reference

<https://github.com/rishabhmisra/News-Headlines-Dataset-For-Sarcasm-Detection>

Implementation steps

Instructions to perform all the below steps are mentioned in the question notebook with the respective marks.

1. Read and explore the data

2. Drop one column
3. Get length of each sentence
4. Define parameters
5. Get indices for words
6. Create features and labels
7. Get vocab size
8. Create a weight matrix using GloVe embeddings
9. Define and compile a Bidirectional LSTM model
10. Fit the model and check the validation accuracy

Project Support

You can clarify your queries related to the project by dropping a mail on Olympus

Happy Learning!