Sarcasm Detection in News Headlines

Md. Mahmudul Haq
2016-2-60-031
rajon789@gmail.com

Moni Kishore Dhar
2016-2-60-099
kishor.hridoy1996@gmail.com

Reaz Ahmed
2016-1-63-029
reaz63029@gmail.com

Ahmed Omar
2016-2-60-093
omeruday@gmail.com

April 3, 2020

Contents

1 Introduction

Sarcasm is defined by the use of remarks that clearly mean, the opposite of what they say, made to hurt someone's feelings or to criticize something humorously. In our project, sarcasm detection refers to the task of accurately labeling a headline as sarcastic or non-sarcastic. A machine can't comprehend the motive of a headline as it only contains some words in a sentence or sentences with no intonation and facial expressions even if it contains so. However, a person can spot a sarcastic sentiment in a text and reason about what makes it so. It is also one of the many critical tasks of NLP to analyze sarcasm in a text to avoid misinterpretation of sarcastic remarks as literal statements. Accuracy and robustness of NLP pave the way to build such models.

2 Motivation

Let us start with an example, a sentence like "So thrilled to be on call for work the entire weekend!" could be classified merely a sentence with a positive sentiment. However, it is actually the opposite, that is cleverly implied through sarcasm. The use of sarcasm is prevalent across all social media, micro-blogging and e-commerce platforms. Sarcasm detection is imperative for accurate sentiment analysis and opinion mining. It can contribute toward enhancing automated feedback systems in the context of customer-based sites.

3 Problem

We have the News Headlines Dataset which contains three columns respectively article_link, headline and is_sarcastic with 26709 rows in a Json file format. This News Headlines Dataset for Sarcasm Detection is collected from The Onion website which aims at producing sarcastic versions of current events and HuffPost website which collects real news headlines.

article_link: contains links to the news articles.
headline: contains headlines of the news articles.
is_sarcastic: contains 0(for no sarcastic text) and 1(for sarcastic text).

4 Objective

Main goal of our project is to predict whether a news headline is sarcastic or not.

5 Model

To accomplish the goal, so far, we have used a machine learning model called naive_bayes from sklearn library. We have also used MultinomialNB from the same library. From the dataset, 80 percent of the data is used to train the model, and the rest 20 percent is for testing the model.

5.1 Pseudo code

Algorithm 1: Pseudo-code

Result: Write here the result

1. Read the dataset
2. import stopwords from nltk.corpus
3. import PorterStemmer from nltk.stem
4. Set a list variable corpus = []
5. **while** *headline != EOF* **do**

```
    headline = re.sub('[^a-zA-Z]', ' ',dataset['headline'][i])
    headline = headline.lower()
    headline = headline.split()
    ps = PorterStemmer()
    headline = [ps.stem(word) for word in headline if not word in set(stopwords.words('english')) ]
    headline = ' '.join(headline)
    corpus.append(headline)
```

end

6. Vectorize the words
7. fit corpus using CountVectorizer from sklearn.feature_extraction.text
8. Select training and testing set
9. Train the model using Naive Bayes
10. Show confusing matrix
11. Print the accuracy

6 Contribution

All of our team member has contributed throughout the project despite the existing situation of COVID-19.

6.1 Report Writing

- Introduction: Md. Mahmudul Haq, Ahmed Omar, Reaz Ahmed
- Md. Mahmudul Haq, Reaz Ahmed, Moni Kishore Dhar
- Problem: Moni Kishore Dhar
- Objective: Moni Kishore Dhar
- Model and pseudocode: Md. Mahmudul Haq (pseudocode), Moni Kishore Dhar (Model)

6.2 Source code

Contributed by all of us.