Intended Sarcasm Detection in English

Sheikh Ayatur Rahman — Anika Tahsin — Ankita Roy — MD Ajmain Mahtab
School of Data and Sciences
Brac University

September 4, 2023

1 Abstract

Sarcasm on digital platforms has escalated, posing a unique challenge to online discourse and personal interactions. Automated sarcasm recognition systems have garnered significant attention due to their potential to enhance digital communication. This report provides a comprehensive analysis of a sarcasm recognition code, elucidating its core methodologies and evaluating its performance. Leveraging state-of-the-art natural language processing techniques, this paper aims to identify and categorize instances of sarcasm. We used the English parts of the competition dataset provided for iSarcasmEval: Intended Sarcasm Detection In English and Arabic to train our models. We framed the problem as a binary classification problem, where a model has to determine whether a piece of text is sarcastic or not, which is the same as Task A of the English section of iSarcasmEval: Intended Sarcasm Detection In English and Arabic. We trained a BERT-base model to perform this task and achieved a peak F1-score of 0.6529 on the sarcastic class, which outperforms the competition winning score of 0.6052 on the same task.

2 Introduction

Sarcasm, a nuanced form of verbal irony, emerges when there's a stark contrast between the literal words spoken and the underlying intended message. It often serves as a vehicle for expressing disdain or mockery towards a previously stated notion. In today's digital age, sarcasm finds its way onto social media platforms, and its subtle complexity poses a significant challenge to computational systems. These systems, vital for tasks such as sentiment analysis, opinion mining, author profiling, and harassment detection, are thrown off balance when confronted with sarcasm.

Notably, during SemEval competitions, it became evident that including sarcastic tweets led to a notable decrease in sentiment polarity classification accuracy, compared to non-sarcastic ones. These systems wield considerable influence in industries, shaping marketing strategies, administrative decisions, and investment choices.

In the realm of NLP, models for sarcasm detection are indispensable. These models usually

operate within a supervised learning framework, relying on datasets that categorize texts as either sarcastic or non-sarcastic. The conventional labeling methods involve distant supervision, where texts are tagged as sarcastic based on predefined criteria like the presence of #sarcasm, or manual labeling, where human annotators classify texts. However, both methods can introduce noise, leading to false positives and negatives due to subjective differences in annotator perceptions of sarcasm.

In response to these challenges, the iSarcasmEval project presents an innovative approach. We label texts for sarcasm by leveraging self-reporting from the authors themselves, eliminating the subjectivity associated with traditional labeling methods.

This project aims to advance the field of sarcasm detection by offering a fresh perspective on dataset labeling and exploring the complexities of sarcasm in different languages. Further details on dataset creation and SemEval tasks are covered in subsequent sections.

3 Data Analysis

The dataset was collected from the iSarcasmEval GitHub repository. Instead of relying on predefined tags or third-party annotators, they have gathered an English dataset for each sarcastic text, requesting its author to provide a non-sarcastic rephrased version conveying the same message.

In the dataset, they've gone a step further by having trained annotators categorize each text into one of the ironic speech categories. These categories include sarcasm, irony, satire, understatement, overstatement, and rhetorical questions. They collected additional information for each text. Additionally, they asked participants to explain why their texts were sarcastic and to provide non-sarcastic rephrases.

The data was split into three parts. In the first part, there was a list of sarcastic tweets and their corresponding non-sarcastic counterparts as rephrased by the annotators. In the second part, there was a list of texts and for each text, 5 annotators had to vote whether the text was sarcastic or not sarcastic, with the winning label being used as the ground label. In the third part, each row had a pair of texts, one sarcastic and one non-sarcastic - 5 annotators then voted on which one of the texts was sarcastic and the winning label was chosen as the ground label.

Since our task was a binary classification problem, we had to convert each dataset so that each entry consisted of one text, which was a label signifying whether it is sarcastic or not. For the first part of the dataset, we took the sarcastic text and labeled them as "sarcastic" while for the rephrased counterparts, we labeled them as "non-sarcastic". For the second part, we kept the text as is and used the winning label as the label for the text. For the third part, we took the text that had been labeled as sarcastic by the annotators and labeled them as "sarcastic" and we additionally took the other text in the pair and labeled them as "non-sarcastic" and added them to our dataset. This gave us a sufficiently large and varied

dataset to train our models.

The training dataset had 6,134 training examples while the testing dataset had 1,400 testing examples.

One example from the dataset:

- Language: English
- Sarcastic Text: "Gotta love people who follow you and unfollow because you don't follow them within an hour or 2. Sorry I don't stay on Twitter 24/7."
- Unsarcastic Rephrase: "I dislike people who follow me, only to unfollow me when I don't follow back right away. I'm not on Twitter that much to follow right away."

4 Model

In recent years, BERT has revolutionized the natural language processing world. Using transfer learning, BERT can be fine-tuned to achieve good results on a wide array of natural language processing tasks.

BERT is a large language model trained on two tasks - next sentence prediction (NSP) and masked language modeling (MLM). In NSP, BERT had to predict whether a sentence logically followed another sentence, and in MLM, some words of a sentence were masked out, and BERT had to predict the masked words. BERT was trained on a huge corpus of English text and can produce contextual word embeddings that can capture the semantic meaning of a word based on its context.

For our task, we fine-tuned BERT for our binary classification task.

During the fine-tuning process, we used the following hyperparameters.

Parameter	**Value**
Batch size	4
Epochs	10
Learning Rate	2e-5
Weight Decay	0.01

Table 1: Hyperparameters

5 Results

While training, we used the test data for evaluation with the maximum F1 score being reached on the 5th epoch: 0.652908.

Epoch	Training Loss	Validation Loss	F1 Score
1	0.538200	0.289030	0.590000
2	0.520300	0.486802	0.621723
3	0.347300	0.587522	0.649903
4	0.209100	0.792615	0.645283
5	0.092500	0.924092	0.652908
6	0.071800	0.868732	0.635161
7	0.071400	1.098429	0.629981
8	0.038900	1.232798	0.635161
9	0.024100	1.264945	0.640301
10	0.014200	1.327932	0.637736

Table 2: Training and Test results

6 Limitations

One notable limitation of our data is the source is primarily from Twitter. Although Twitter serves as a rich platform for real-world language usage, it may not be entirely generalizable. Different social media platforms and contexts may exhibit various degrees of sarcasm and linguistic differences, which might affect the model's performance.

7 Conclusion

In conclusion, this paper presents a Python-based framework for sarcasm detection using natural language processing techniques and the transformers DistilBERT model. The code, which leverages advanced transformer models and data preprocessing strategies, demonstrates its capability to categorize tweets as sarcastic or non-sarcastic effectively. Through extensive experimentation and evaluation, the model achieves promising accuracy while maintaining efficiency. The comprehensive analysis and evaluation presented in this paper offer valuable insights into the performance and potential applications of automated sarcasm detection systems, contributing to the ongoing efforts to mitigate the spread of harmful and offensive content in online discourse.