Sarcasm Detection in News Headlines - Management Summary

Course: 194.093 Natural Language Processing and Information Extraction 24W

Group 16 - Snack Overflow: Terezia Olsiakova (12331438), Arian Etemadihaghighi (12242985), Viktoriia Ovsianik (12217985), Maximilian Scheiblauer (11776651)

January 26, 2025

1 Introduction & Task Overview

Our project aimed to develop an approach to classify news headlines as sarcastic or non-sarcastic. This classification focuses on distinguishing between humorous, sarcastic writing (e.g., from "The Onion") and straightforward news (e.g., from "The Huffington Post"). We also explored the patterns of sarcasm and performed semantic analysis to understand underlying headlines better.

2 Data

2.1 News Headlines for Sarcasm Detection

Our main dataset for this task was "News Headlines for Sarcasm Detection" [1]. The dataset has 28,600 annotated headlines from "The Onion" (sarcastic) and "The Huffington Post" (non-sarcastic), classes are balanced. There were no missing values and just a few duplicates that we deleted.

2.2 Headlines Generated by ChatGPT

Additionally, we generated 2,000 new headlines, evenly split between sarcastic and non-sarcastic. Half of these were crafted to mimic the style of the original dataset, while the other half followed a more generic, random style. This allowed us to evaluate whether the model could accurately classify sarcastic headlines that deviated from "The Onion's" distinctive tone.

2.3 Tweets Data

Additionally, we tested our models on a tweets dataset [2], where tweets were annotated as sarcastic or non-sarcastic by their original authors. This allowed us to evaluate the generalizability of the models to a different type of text, distinct from the news headlines. Tweets usually lacking contextual detail, tend to be more subjective, emotional, and written in informal manner.

3 Methods & Approach

3.1 Baselines

We developed and evaluated four models to classify sarcasm (2 deep learning and 2 non-deep learning): DistilBERT, Neural Network, Naive Bayes (Bag of Words), and Logistic Regression. Table 1 shows the measured performance metrics of the baseline models.

3.2 Error Analysis

We evaluated the performance of the baseline models not only quantitatively but also qualitatively to discover the main challenges of our task. For that, we analyzed the instances in which the models predicted incorrectly, focusing on the false negatives, so cases when a sarcastic headline was classified as non-sarcastic as we deem it more important to focus on not letting sarcasm in the headlines go undetected.

[1] Headlines Dataset: `https://github.com/rishabhmisra/News-Headlines-Dataset-For-Sarcasm-Detection`

[2] Tweets Dataset: `https://github.com/iabufarha/iSarcasmEval/blob/main/train/train.En.csv`

We observed, that for some of the headlines, it was difficult to detect sarcasm simply because real-world knowledge was required to understand the headline as sarcastic, for example, the headline "7.1 Billion Demonstrate in Favor of Global Warming." is sarcastic simply because of the large number which is not detectable by the design of our trained baselines. Moreover, the analysis revealed that some of the words show a bias towards a certain class, for example, the word "Trump" is most frequent in the non-sarcastic headlines.

3.3 Advanced Approaches

With that, we dug deeper into the task by exploring several approaches based on our insights from the error analysis. First, we asked ourselves whether we were detecting sarcasm or rather the specific writing style of "The Onion". Therefore, we tested our best baseline on a new unseen dataset consisting of tweets and we saw a significant decline in the performance, which could either be caused by the diversity of the headlines vs. tweets data but could also hint that there might be some underlying patterns in the data that come from the specific writing style.

Moreover, we performed a pattern analysis by also incorporating the headlines generated by ChatGPT. On this new data, our baseline performed even better than originally, as shown in Table 1, reinforcing the idea that the writing style does play a role in the models' classification. Namely, the identified common patterns among the sarcastic headlines include the words: "nation", "area man", "local", and "study".

Lastly, we explored the possibility of extracting syntactic features from the headlines to improve the performance over the baselines. From the POS tags and syntax trees, features were extracted using the methods of recursive feature elimination and random forest, then a logistic regression and random forest models were trained on the top 5 features, which together with the Naive Bayes baseline formed an ensemble model with majority voting. As shown in Table 1 this approach does not outperform the baseline model itself.

4 Results

To summarize, Table 1 shows the results of our experiments across various model architectures as well as different datasets which correspond to the ones introduced in Section 2. The best-performing baseline was the Naive Bayes and the only improvements of this performance were delivered by using the Naive Bayes or the Voting Classifier on the headlines generated by ChatGPt that mimic those from the original dataset which is a consequence of the patterns found in the data as discussed in Section3.3.

Model	Neural Network	Logistic Regression	DistilBERT	Naive Bayes				Voting Classifier			
Dataset	Headlines	Headlines	Headlines	Headlines	ChatGPT (mimicked)	ChatGPT (generic)	Tweets	Headlines	ChatGPT (mimicked)	ChatGPT (generic)	Tweets
Precision	0.84	0.83	0.82	**0.85**	0.95	0.71	0.52	0.74	0.88	0.68	0.62
Recall	0.83	0.83	0.75	**0.85**	0.89	0.74	0.15	0.71	0.87	0.68	0.59
F1-score	0.83	0.83	0.78	**0.85**	0.92	0.73	0.23	0.72	0.87	0.68	0.60

Table 1: Performance metrics per model and dataset.
The best baseline's metrics are in **bold**, and improvements over the baseline are underlined.

5 Challenges

The biggest challenge that we encountered when detecting sarcasm in news headlines was ultimately the need for external knowledge of the world to address more nuanced sarcasm in headlines. Moreover, we discovered that our main dataset displayed specific patterns not related to sarcasm itself but the writing styles of "The Onion" which we did not originally anticipate when choosing the data for the task.

We also found it challenging to improve upon the baselines as these yielded already good results regarding performance metrics, and in our experience, the errors that occurred could only be mitigated with the added external knowledge as we discussed. Regarding the different sarcasm datasets, we conclude that the models do not generalize well across them.

On a final note, we have found that sarcasm is not necessarily well-defined and it is debatable whether it comes down to detecting humor itself.

6 Conclusion & Further Steps

In conclusion, detecting sarcasm from news headlines proved to be a difficult classification task. Sarcasm is not exactly defined and often requires real-world knowledge and surrounding context to be detected by machines, possibly as well as humans. Our approach showed that even a simple baseline reaches a good enough performance, which is harder to improve on even with more advanced methods.