COMP4420 Project Proposal: Sarcasm Detection

Bui, Nam (#01963609), Conners, Riley (#01943861), Zuk, Sam (#01642608)

1 Introduction

Sarcasm is a feature of natural language that is notoriously difficult to define and identify in both the spoken and written word. The assumption that a statement will be recognized as sarcastic is typically contingent upon the listener/reader knowing some outside piece of contextual information beforehand. However, this external information isn't always known, and even when it is, the relationship between it and the statement at hand may not always be clear. When this happens, the meaning can be obscured as a result, often leading to avoidable scenarios involving miscommunication.

Recognizing sarcasm typically involves picking up on subtle cues and nuance that can be difficult to identify. This can often pose a challenge for populations who encounter greater difficulty when processing certain aspects of a language. For example, someone trying to interpret a language they don't speak natively will likely have to expend more mental effort to parse out meaning from words, which in turn makes it more difficult to pick up on nuance, including sarcasm. Being unfamiliar with the cultural norms, idioms, etc. that inform the established meaning of the locally spoken language can also be a source of confusion. In addition, many neurodivergent people, in particular those with autism, can struggle to recognize and/or communicate certain social cues in conversation due to differences between their cognitive experience of language and what is expected of them.

Finally, there are unique challenges faced in detecting sarcasm in the written word. It is often possible in practice to infer a statement is sarcastic, even without necessarily having the context to understand *why* by listening to changes in the tone of the speaker. However, when translated into the written word, some or all of this information is lost, making sarcasm even more difficult to detect when only text is given. With the Internet now being extremely important to modern infrastructure, and with text being the predominant medium for online communication, this problem has become increasingly apparent over the years. This project shall explore and contrast different approaches to disambiguating sarcasm by applying concepts from the fields of computational linguistics and machine learning.

2 Dataset

The dataset we plan on using for this project is a collection of 28,619 tagged newspaper headlines– of which 13,635 are from the satirical publication *TheOnion*, the other 14,984 being from the non-satirical publication *The Huffington Post* (*HuffPost*). The data was collected from TheOnion's "News in Brief" and "News in Photos" sections and HuffPost's news archive page in 2019 [1].

For each headline, the dataset contains a JSON object with three attributes:

- `is_sarcastic` (integer): the headline's label– 1 if sarcastic, 0 if not.
- `headline` (string): the text of the headline, case-converted to be all lowercase.
- `article_link` (string): the URL of the referenced article.

This project will attempt to use the contents of `headline` to predict the value of `is_sarcastic`. Values of `article_link` will not be used directly for the purposes of modeling sarcasm, but may be useful when attempting to decipher why certain models made certain predictions.

This dataset has advantages over text that could be found on social media platforms because news text is formal in nature. This means there are less words outside of the word2vec vocabulary, less spelling

mistakes, and little to no slang usage. Also, because The Onion is openly sarcastic by design, there is no ambiguity regarding if labels are correct.

However, there are downsides to news headline data. In this case, there are only two news sources being used, and the model could pick up on writing styles or other details instead of sarcasm. There is another potential issue that stems from The Onion's obvious use of sarcasm. In more nuanced cases where sarcasm is more subtle, a model could do poorly.

3 Evaluation Method

Since the dataset was assembled following a heated U.S. presidential election in 2016, there may be some bias in the data. It will be interesting to see which words correlate most strongly with sarcastic headlines.

Additionally, news headlines usually have a lot of proper nouns, so it may help to use named entity recognition when encoding the headlines.

Sentiment analysis is a core natural language processing task, so there is a lot of data available on what types of models are effective. We plan on using several for this project. Naive Bayes classifiers are lightweight models that have traditionally been used in sentiment analysis. Deep averaging networks are able to leverage the universal approximation properties of neural networks, but are lightweight since they don't capture context. In recent years, recurrent neural networks have gained popularity due to their ability to capture context with the attention mechanism. Since news headlines are often one or two sentences, there is not much need to capture long distance dependencies.

Since the task at hand is binary classification, precision, recall, and F1 are good metrics to use. Accuracy will also be used to compare findings to results from Misra et al [1].

4 Timeline

First, the dataset will be split into training, development, and test sets with a 70/10/20 split.

For the baseline model, we will use a Naive-Bayes with one-hot encoding.

After training and evaluating the baseline model, we would like to fine-tune word2vec to the headline-specific words.

Once the word embeddings have been fine tuned, we would like to train and evaluate either a DAN or LSTM model.

5 References

[1] Rishabh Misra and Prahal Arora. Sarcasm detection using news headlines dataset. *AI Open*, 4:13–18, 2023.