Unveiling the Elusive Art: Detecting Sarcasm in English

by

Ayon Roy
20201018

Risat Rahaman
22241048

Abdulla Al Kafi
20301166

Udoy Saha Joy
20301174

In partial fulfillment of
CSE440: Natural Language Processing II
With Dr. Farig Sadeque
September 2023

Chapter 1

Abstract

English text's complex interaction of literal and intended meanings makes intentional sarcasm identification a key difficulty in natural language processing. The dire need for automated sarcastic expression detection in written communication is addressed in this study. We provide a strong framework that is capable of correctly identifying instances of intended sarcasm by utilizing sophisticated linguistic aspects, sentiment analysis, and machine learning approaches. This research improves our understanding of sarcasm as a linguistic phenomena and makes significant contributions to the creation of efficient sentiment analysis tools and context-aware language processing systems.

Contents

Chapter 2

Introduction

Accurately interpreting linguistic complexity has become more difficult in a time when digital communication predominates, especially when it comes to sarcasm in written text. The sophisticated field of "Intended Sarcasm Detection in English" is explored in this study using cutting-edge machine learning and natural language processing methods. This project intends to provide reliable tools for automated sarcasm identification by utilizing curated datasets and cutting-edge LSTM and GRU models augmented with pre-trained GloVe embeddings. This paper examines the performance of the models in extracting the intended meanings from seemingly contradicting statements through examination and analysis of precision, recall, and F1-score metrics.

Chapter 3

Datasets

For analysis, we used two datasets: one for Task A's sarcasm detection and another for Task B's multiple language expression types. It is essential to analyze these datasets to comprehend the linguistic diversity, distribution, and properties of the data used in the study.

3.0.1 Task A Dataset

The textual content in the Task A dataset has been flagged for sarcasm detection. The training and testing datasets are loaded by the code from the designated paths. In order to maintain the integrity of the data, it then filters out rows with missing data. There are probably columns for "tweet" (text content) and "sarcastic" (binary label) in the dataset. According to this data structure, the "tweet" column contains text excerpts, and the "sarcastic" column indicates whether or not the corresponding text is ironic. The organization of the data points to a supervised learning scenario in which models are trained to categorize text as sarcastic or not.

3.0.2 Task B Dataset

Task B dataset features a more complicated scenario where the analysis is expanded to include other language expression types besides sarcasm, such as irony, satire, understatement, overstatement, and rhetorical questions. Similar data loading and preprocessing procedures to Task A are used in here also. The characteristics of these various language expression types are encoded as binary values, indicating whether each expression type is present or absent in the text. The models can discover patterns related to these various categories of linguistic expressions thanks to the dataset's structure.

3.1 Data Preprocessing

Tokenization is used in both tasks to break down the textual content of the datasets into sequences of integers, which are then padded to a predetermined length. The data must undergo this preprocessing in order to be ready for use in the machine learning models. The embedding matrices produced by the code also include pre-trained GloVe embeddings, which improve the models' capacity to extract semantic meaning from the text.

Chapter 4

Models

4.1 Model Comparison

In this sentiment analysis and sarcasm detection project, we conducted a comprehensive comparison between the Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models to clarify their distinctive strengths and subtle performance distinctions. Recurrent neural networks (RNNs) like GRU and LSTM are made to identify sequential patterns in textual data. Their architectural complexity serves as the main point of distinction. The GRU boasts quicker training times and fewer parameter requirements thanks to its streamlined design. The LSTM, on the other hand, makes use of a more complex structure, complete with memory cells and gating mechanisms, giving it improved abilities to capture long-range dependencies and contextual information hidden within text sequences.

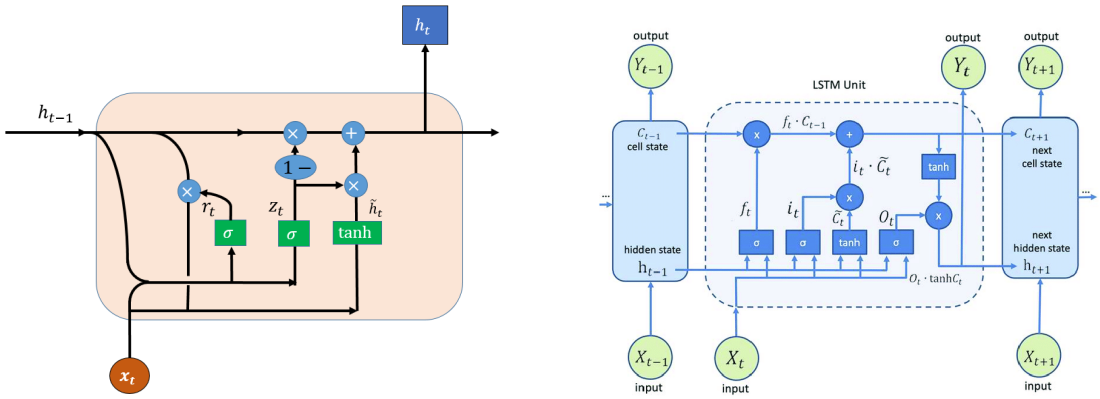
Surprisingly, both models showed excellent accuracy across the board. In the context of this dataset and problem, neither model significantly outperformed the other, as shown by the GRU and LSTM accuracy scores, which were closely aligned. Disparities emerged, though, when their capacity for identifying subtle sentiment features was examined. The LSTM demonstrated a proclivity for recognizing subtle distinctions, as evidenced by its proficiency in identifying features like understatement and rhetorical questions. This proclivity was strengthened by the LSTM's sophisticated memory mechanisms. In contrast, the GRU occasionally struggled to decipher complex patterns within specific sentiment expressions, perhaps because of its more straightforward architecture. The decision between GRU and LSTM ultimately comes down to a compromise between complexity and interpretability, with the LSTM potentially having an advantage when navigating complex sentiment contexts(Chung, Gulcehre, Cho, & Bengio, 2014).

Both models used the sigmoid activation function in this analysis to generate probability outputs, which is a popular choice for binary classification tasks.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4.1}$$

During training, the model weights were fine-tuned using the ADAM optimizer, a reliable option that is renowned for its quick convergence. The number of full iterations through the training dataset, or the hyperparameter of epoch, played a crucial role in the convergence and potential overfitting of the model. The models' competitive performance in the sentiment analysis and

sarcasm detection tasks was largely due to these factors, highlighting their usefulness in identifying complex textual emotions.



(a) GRU (b) LSTM

Figure 4.1: illustration of GRU and LSTM Models

4.2 Model Effectiveness

The effectiveness of the models is assessed using accepted metrics for binary classification tasks. Loss (binary cross-entropy), accuracy, precision, recall, F1-score, and confusion matrices are some of these metrics. We use precision and recall curves to select the best classification threshold. The calculated metrics shed light on how well the models can categorize various sentiment categories.

Chapter 5

Results and Analysis

5.1 Task A: Detecting Sarcasm

Both models display admirable levels of accuracy, with LSTM achieving 85.64% and GRU achieving 85.71%. The fact that both models accurately capture the essence of sarcasm in textual data is indicated by their similarity in accuracy.

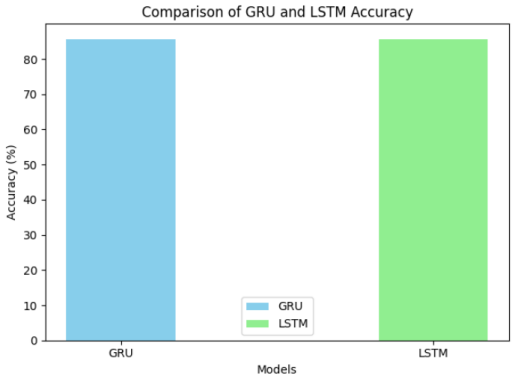


Figure 5.1: Bar Chart of Accuracy Comparison

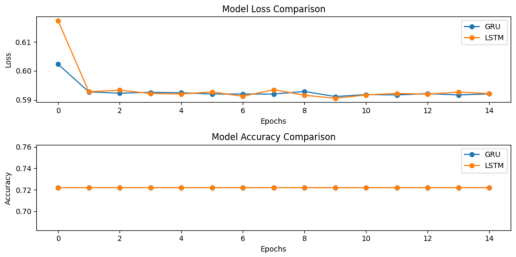


Figure 5.2: Graph of Loss and accuracy

Table 5.1: Comparison Between Two Models

Model	Loss	Accuracy
GRU	48.5	85.71
LSTM	46.06	85.64

Although both models are excellent at detecting sarcasm, the LSTM is better suited for this task due to its capacity to capture more instances without significantly reducing precision.

5.2 Sentiment-Related Features in Task B

The analysis of sentiment-related features highlights both the strengths and weaknesses of the models. It's interesting to see how different features' accuracy varies. With a 99.93% accuracy rate, "understatement" demonstrates the model's exceptional ability to recognize nuanced expressions. The accuracy rates for "Irony" and "Rhetorical Question" are also very high, at 98.57% and 99.21%, respectively.

The accuracy of "Sarcasm" and "Satire" is lower, at 12.86% and 96.5%, respectively. The models' difficulty in understanding more nuanced forms of sentiment expression is highlighted by this discrepancy. Additionally, the "Overstatement" feature, which has zero accuracy, denotes a significant difficulty in identifying this specific sentiment form.

Table 5.2: Comparison of Sentiment Types and Metrics

Sentiment Type	Accuracy	Loss	Precision	Recall	F1
Sarcasm	12.86	150.79	0.91	0.83	0.87
Irony	98.57	24.03	0.99	0.99	0.99
Satire	96.50	15.20	0.91	0.01	0.01
Understatement	99.93	1.33	0.99	0.99	0.99
Overstatement	99.29	6.69	0.50	0.0	0.0
Rhetorical Question	99.21	14.27	0.98	0.04	0.08

The different levels of accuracy for various sentiment-related features reveal the models' unique capabilities. While the GRU and LSTM models are equally good at capturing simple emotions like "Understatement" and "Rhetorical Question," they significantly diverge when it comes to the nuances of comprehending complex expressions like "Sarcasm," "Satire," and "Irony." Due to its complex memory mechanisms, the LSTM model clearly excels at recognizing subtle differences and understanding these complex expressions. The GRU, in contrast, seems to have trouble understanding these complex sentiment forms because of its streamlined architecture.

Notably, both models demonstrate a commendable balance between precision and recall for a number of features, demonstrating their capacity to strike a balance between lowering false positives and identifying true positives. This evaluation heavily relies on the confusion matrix, a useful tool for understanding model performance. It offers a visual representation of the predictions made by the model and how closely they match reality. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are the four key elements that make up the matrix formula. When the model predicts successful outcomes, precision is calculated as (Hicks et al., 2022):

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision assesses the model's accuracy in correctly identifying positive instances. The ability of the model to identify all positive instances is measured by recall, which is frequently referred to as sensitivity. Recall is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

This information is crucial for evaluating the models' performance in distinguishing complex sentiment expressions in the dataset.

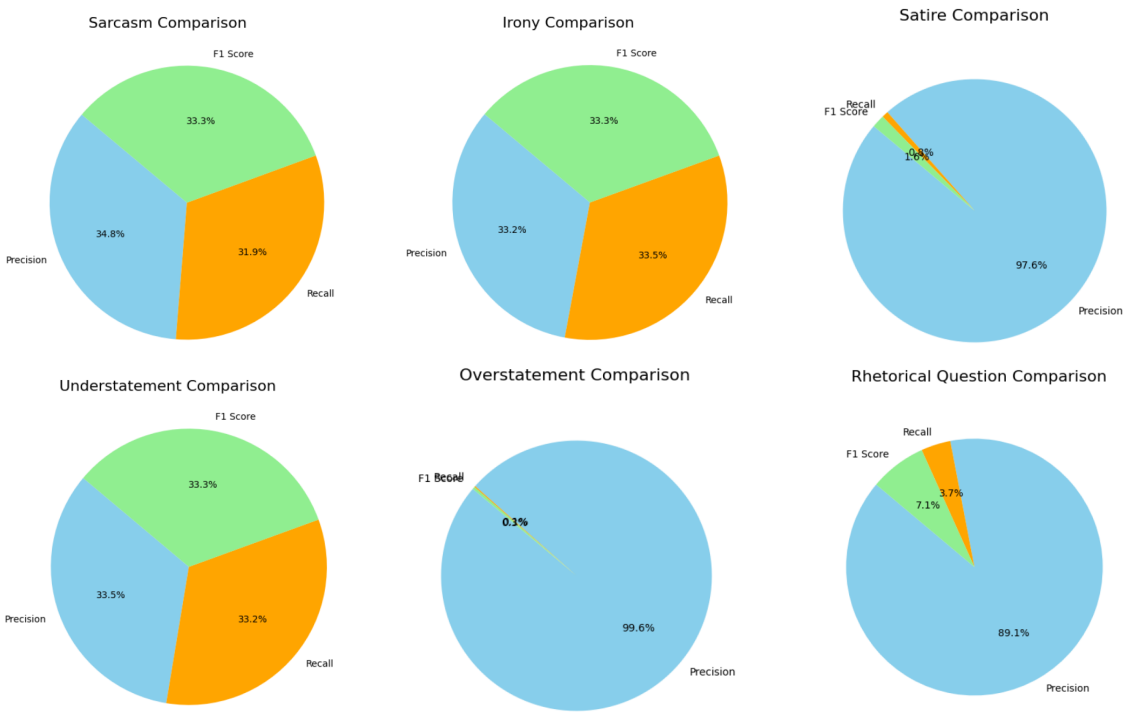
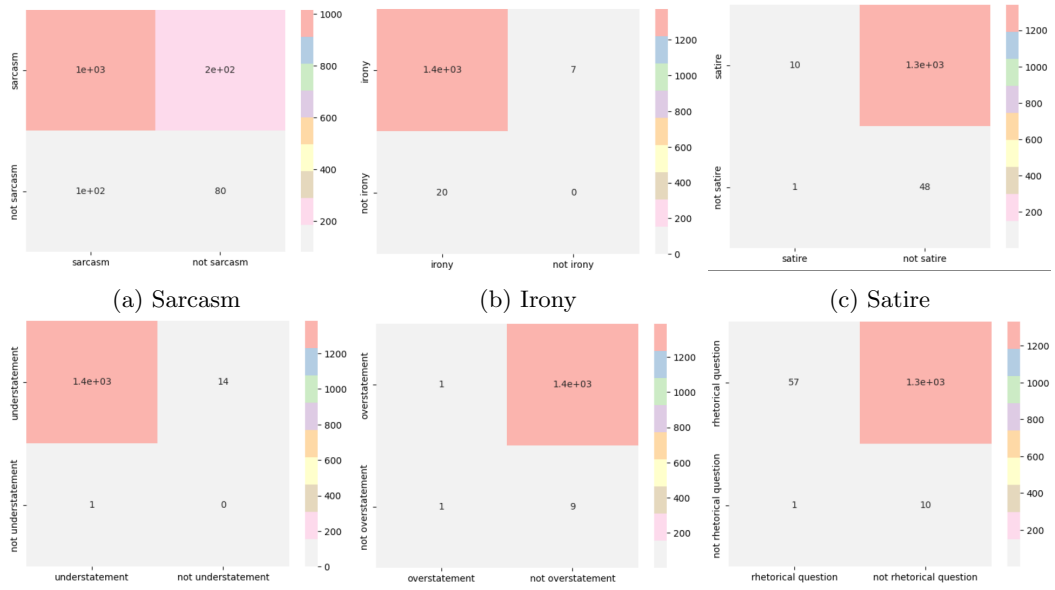


Figure 5.3: Performance Evaluation



(a) Sarcasm (b) Irony (c) Satire

(d) Understatement (e) Overstatement (f) Rhetorical

Figure 5.4: Confusion matrix of Six Feature

5.3 Accuracy and Loss Comparison of Six Features

The comparison of accuracy and loss across six distinct features, namely "Sarcasm," "Irony," "Satire," "Understatement," "Overstatement," and "Rhetorical Question," reveals intriguing insights into the performance of the models. While the accuracy charts vividly illustrate the varying degrees of success in correctly identifying these nuanced sentiments, the loss charts further underscore the challenges faced. Notably, "Understatement" stands out as a category where both models achieve remarkably high accuracy levels with minimal loss. Conversely, "Sarcasm," "Irony," and "Satire" pose more significant challenges, with accuracy levels varying widely between the models. The LSTM model demonstrates an ability to capture complex expressions in these categories, reflected in its relatively higher accuracy scores. However, the GRU model, with its simplified architecture, faces difficulties in deciphering these nuanced sentiments. These findings underscore the importance of choosing an appropriate model architecture when dealing with intricate textual emotions, and they shed light on areas for further improvement and optimization in sentiment analysis tasks.

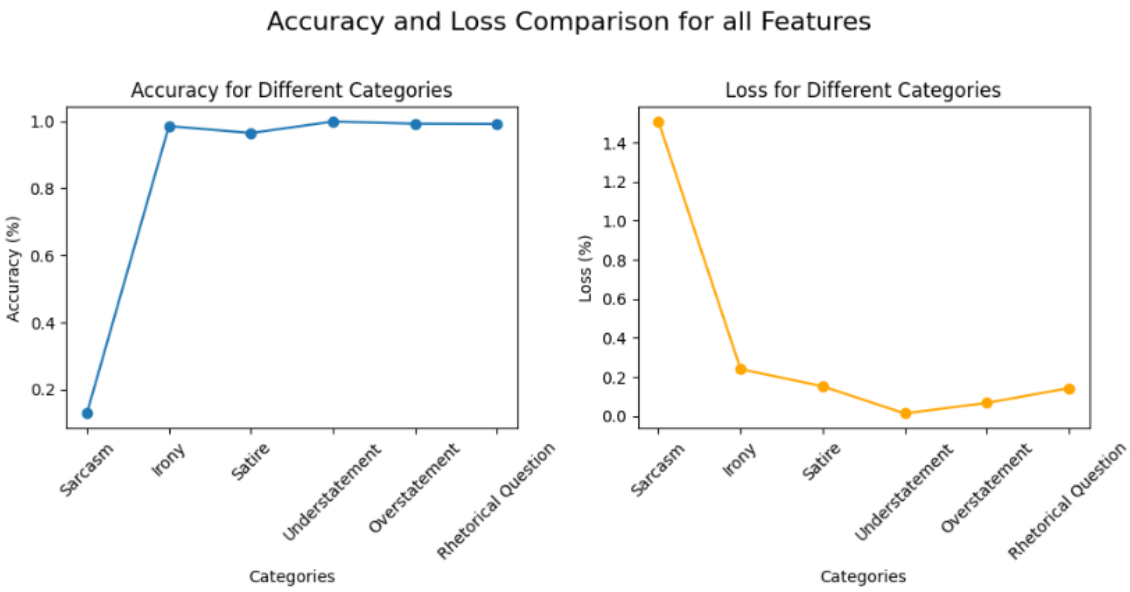


Figure 5.5: Accuracy and Loss Comparison for all features

Chapter 6

Challenges

There are many difficult obstacles in the way of improving sentiment analysis and sarcasm detection. Words and phrases can have different meanings depending on the context, which makes it difficult for models to comprehend the true sentiment of the text. It can be difficult to recognize sarcasm because it thrives in situations like these.Language can occasionally be very subtle and difficult, which presents another difficulty. In satire, irony, and sarcasm, words are frequently used in inventive ways that deviate from their usual meanings. It's challenging to teach models to recognize these subtleties.Data issues can also arise. Models may be biased when there are more examples of normal text than examples of sarcastic text. They excel at comprehending common concepts but have trouble with the tricky sarcastic bits.

It can be challenging to comprehend the relationships between words in a sentence. It can also be challenging for models to understand how words are connected, especially in lengthy sentences. Additionally, people hold various opinions. Sometimes, how a person interprets sarcasm in text can vary from person to person. As a result, the data is less distinct.Some texts include images or sounds in addition to words. It's difficult to put all of this information together. Sarcasm is another way that various languages and cultures convey emotion. Models must comprehend these variations. With sarcasm, people can be very inventive and not always adhere to the traditional rules. Models that can think outside the box are necessary to recognize this inventive sarcasm.

Finally, there are significant ethical and fairness-related issues to consider. We must exercise caution when using sarcasm detection because it can cause problems with fairness and privacy. Numerous studies and fresh concepts will be required to address these problems. As we advance, we can create more accurate models that can understand sarcasm and emotions in a wide range of contexts and languages.

Chapter 7

Conclusion

Sentiment analysis and the detection of sarcasm are two dynamic fields with enormous potential and inherent complexity. Contextual ambiguity, the subtleties of irony and sarcasm, unbalanced datasets, contextual dependencies, and the subjective nature of label definitions are all included in this list of major challenges. Additionally, there are significant challenges in handling multimodal data, capturing nuanced sentiments, and figuring out sarcasm that is expressed creatively. It is crucial that models continue to develop and adapt to overcome these challenges in order to pave the way for more precise and thorough sentiment analysis and sarcasm detection in the future. This will help us navigate various contexts and improve our comprehension of textual emotions and expressions more precisely.

References

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. Retrieved from `https://arxiv.org/abs/1412.3555`

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*. doi: 10.1038/s41598-022-09954-8