# COMP 550 Final project

**Anonymous ACL submission**

## Abstract

This paper aims to investigate the impact of contextual information and syntactic structure on sarcasm detection. In order to do so, we implemented the sarcasm detection model proposed by **?**, trained it on a dataset of Reddit posts, and conducted two experiments. The first one consisted of separating the dataset based on the number of contextual utterances, training the model for each subset, and assessing the variation in accuracy. The second experiment consisted of modifying the model to include input vectors representing the syntactic features of the response and assessing the accuracy. Our findings demonstrate that incorporating additional contextual utterances enhances sarcasm detection accuracy up to a certain threshold, beyond which accuracy declines. Furthermore, explicitly considering syntactic features had no added value to the model's performance. These results provide insights into the role of context and syntax in computational sarcasm detection.

## 1 Introduction

Sarcasm detection is a challenging task in Natural Language Processing (NLP) due to its reliance on implied meanings that often contradict literal wording. Sarcasm frequently depends on subtle linguistic cues and contextual nuances, which makes it difficult for machines to interpret. This project explores the role of contextual information and syntactic structure in improving sarcasm detection accuracy.

We hypothesize that the amount of contextual information related to a target response plays a significant role in detection performance. Specifically, we propose that incorporating additional contextual utterances enhances accuracy up to a certain threshold, beyond which excessive context may confuse the model and degrade performance by diluting its focus. Additionally, we investigate the influence of syntactic features, such as negations, modifiers, and the interrogative or exclamatory nature of a response, on sarcasm detection. These features, we suspect, are critical in capturing the nuanced nature of sarcastic expressions.

To test these hypotheses, we implement the method proposed by **?**, and evaluate its performance when trained on varying amounts of contextual information, as well as a fixed number of utterances. We also introduce modifications to the approach to validate our claims. For comparative analysis, we consider two baseline models: (1) a simple logistic regression model[1] of our implementation that uses only the response text without context, and (2) a publicly available GitHub implementation of **?** designed to only analyze responses without context (**?**).

## 2 Related work

Sarcasm detection has received significant attention within the natural language processing world, and as a result, many experiments and methodologies have been done and developed to accurately identify sarcastic expressions. We review some of the studies that have contributed to the field, highlighting their approaches and findings:

- **Transformer Based Approach**
  **?** developed a transformer-based model that combines pre-trained language representations and recurrent convolutional neural networks. This model was able to achieve state-of-art performance on different datasets including Twitter and Reddit datasets. Their combination allows the model to capture both contextual and sequential patterns in sarcastic expressions effectively. Also, it highlights the effectiveness of hybrid architectures, which enable the capture of contextual nuances in sarcastic expressions. While our model also leverages Transformers, it expands upon their approach by integrating syntactic features, and allows for more granular

[1] The implementation is inspired by PA1

analysis of sarcastic expressions. Potamias's model relies only on RCNNs which may lead to higher computational costs and it does not explicitly incorporate syntactic cues.

- **Linguistic and Syntactic Feature Integration** **?** proposed a framework that combines context, emotions and sentiment features with different pre-trained transformer models. They employed CNNs to analyze incompatibilities characteristics in sarcastic expressions, which allowed them to achieve improved performance on sarcasm detection tasks. However, this approach only focuses on the interplay between sentiment and sarcasm but it does not fully explore the sequential and hierarchical nature of conversational data, unlike our model, which models hierarchical dependencies in conversational data and builds upon that by using syntactic features.

- **Multimodal and Ensemble Methods** **?** introduced a parallel deep learning approach using multiple LSTM networks to process pop culture text and English humor literature. Their method was designed to learn from diverse datasets, which allowed the model to achieve high accuracy scores in detecting sarcasm. This article demonstrates the importance of ensemble and multimodal techniques. This model effectively learns patterns from diverse datasets, improving generalizability. However, the use of LSTMs can be computationally expensive and slower compared to transformer-based models. Whereas, our model takes advantage of using transformer-based architecture to improve efficiency and scalability, and it integrates syntactic and contextual information which are critical for handling complex sarcastic expressions.

# 3 Method

## 3.1 Dataset and Preprocessing

The dataset used in this experiment was the Reddit dataset from the ACL FigLang Workshop 2020 (**?**). Unlike traditional sentiment analysis datasets that focus solely on standalone responses, this dataset includes a response and its context, comprising a variable number of preceding utterances. This design enables the study of sarcasm as a contextual phenomenon, emphasizing the relationship between the response and its context (**?**). The dataset consisted of 4,400 data points in the training set and 1,800 data points in the test set. To create a validation set, 20% of the training data was set aside. Each data point contained a response string, a context list with two or more utterances, and a string label indicating whether the response was sarcastic.

To ensure uniformity, all context lists were padded to match the longest sequence in the dataset. Padding was performed using the special token `[PAD]` to ensure that models like BERT would not assign attention to the padded sections, thereby preserving the integrity of the contextual information. Additional preprocessing steps were applied to prepare the dataset for training. Sarcasm labels were converted to binary integers (0 for non-sarcastic and 1 for sarcastic) to simplify model learning. The BERT cased tokenizer was used to tokenize both the context and response texts, giving attention to capitalization, and employing subword tokenization for compatibility with transformer-based models.

These preprocessing steps ensured the dataset was optimized for capturing the intricate relationship between context and response.

## 3.2 Baseline Models

Two baseline models were used to benchmark the performance of our hierarchical sarcasm detection model:

1. **Logistic Regression Model:** This baseline, implemented with `scikit-learn`, leverages only the response text, disregarding contextual information. Features are extracted using a Bag-of-Words representation, making it a minimalist approach to sarcasm detection that relies purely on lexical patterns and word frequencies within the response.

2. **Simplified Hierarchical Sarcasm Detection Baseline** (**?**) The second baseline model was sourced from a publicly available GitHub repository. It is a simplified sentiment analysis model that was adapted from the hierarchical architecture proposed in the same article that inspired our main implementation. It implements sarcasm detection using only the response text. The original model was designed for a different simpler dataset, but it was adapted to work with our dataset for consistency and comparison. This model was chosen because it aligns closely with the first baseline (logistic regression with Bag-of-Words features) in that it does not utilize contextual information. By evaluating this model on our dataset, and adjusting the hyper-parameters accordingly with the main

model, we aim to ensure the robustness of our comparisons and to highlight the added value of incorporating context.

### 3.3 Main Model

The model used for our sarcasm detection experiment implements a sophisticated hierarchical deep learning approach. It is designed to capture the nuanced linguistics in sarcastic text, and is based on the hierarchical BERT architecture presented in the paper "A novel Hierarchical BERT Architecture for Sarcasm Detection" (**?**). This model has five main layers:

1. **Sentence Encoding Layer:** BERT is used to encode context utterances and responses into vector representations. A BiLSTM is applied to capture semantic relationships in the response.

2. **Context Summarization Layer:** A 2D convolution operation reduces the dimensionality of the context representation while maintaining essential features.

3. **Context Encoder Layer:** BiLSTM processes the sequential context to extract temporal features and create a concise context representation.

4. **CNN Layer:** The relationship between the response and the summarized context is modeled using shared convolutional layers with varying kernel sizes to capture N-gram features.

5. **Fully Connected Layer:** The extracted features are passed through a dense layer, followed by a sigmoid function to compute sarcasm probabilities.

The model was implemented using the PyTorch framework, with the HuggingFace Transformers library employed for the BERT tokenizer and encoder components. Training was conducted using the AdamW optimizer and binary cross-entropy loss function. The evaluation process utilized the provided dataset, which consists of sarcastic and non-sarcastic text samples along with their corresponding contextual information. This dataset enabled hierarchical processing, facilitating a thorough assessment of the model's accuracy and effectiveness. Furthermore, the study investigates the model's ability to capture and interpret both contextual information and syntactic structures, providing insights into its performance in understanding nuanced language patterns

### 3.4 Contextual information

To evaluate the impact of the number of utterances on sarcasm detection, the training, validation, and test sets were equally partitioned based on the number of contextual utterances included. Multiple versions of the model were then trained, each using a fixed number of utterances in the context, resulting in five distinct models corresponding to contexts containing 2 to 6 utterances. This range was constrained by the dataset's distribution, as the maximum number of utterances in any of the 4,400 training data points was 8. However, only two data points contained 7 utterances, and just one contained 8 utterances, making it scientifically unjustifiable to train a model on such a minimal subset of the data. Finally, each model was tested using its respective dataset to assess its accuracy in sarcasm detection, enabling a comparative analysis of the influence of context size.

### 3.5 Syntactic structure

The dataset analysis revealed that sarcastic sentences often feature specific syntactic structures, such as rhetorical questions, exclamatory statements, and excessive use of modifiers, which serve as key linguistic cues. To investigate the role of syntax in sarcasm detection, these patterns were analyzed alongside the hierarchical model to assess their contribution to distinguishing sarcastic from non-sarcastic text. It was hypothesized that when contextual information is insufficient, the model relies more on syntactic features.

SpaCy's dependency parser was used to extract syntactic features, including part-of-speech tags, dependency relations, and head-token relationships. These structures were converted into numerical vectors for model input, focusing on the most common patterns: negations, rhetorical questions, exclamations, and modifier frequency. The vectors were then passed through a fully connected layer to map them into a learned embedding space.

## 4 Results

### 4.1 Main Model

The main model was trained with early stopping criteria to prevent overfitting. After 8 epochs of training, early stopping was triggered, resulting in the best-performing model obtained at epoch 3, which achieved 64% accuracy[2] on the test set. To contextualize this result, the reference article

[2]Training/Validation accuracy and loss curves can be accessed via README for more details

that inspired our solution reported an accuracy of 66%. The slight difference in performance may be attributed to the hyperparameters used in the article, which were not discussed by the author. Due to limited computational resources, hyperparameter tuning was not conducted in this study.

Additionally, the main model outperformed the baseline models. Specifically, baseline models one and two achieved accuracies of 59% and 49%, respectively. All three models were trained on the same training set and tested on the same testing set. Notably, the baseline models did not account for context, whereas the main model explicitly incorporated contextual features.

This analysis highlights the critical role of context in classification tasks. The superior performance of the main model over the baselines suggests its effectiveness in capturing context as a significant feature for accurate classification.

### 4.2 Utterance specific models [3]

As previously stated, this study hypothesizes that while context enhances sarcasm detection, the relationship between the number of utterances and accuracy is not directly proportional and instead decays beyond a certain threshold. The results of this experiment are presented in Table 1.

| Number of utterances | Accuracy |
|---|---|
| 2 | 63.85% |
| 3 | 69.59% |
| 4 | 64.89% |
| 5 | 61.73% |
| 6 | 54.15% |

Table 1: Impact of context size on sarcasm detection accuracy. The results demonstrate diminishing returns beyond three utterances.

The results reveal varying performance depending on the number of context utterances. The highest accuracy is observed for data points with three utterances (69.59%), followed by four utterances (64.89%) and two utterances (63.85%). Interestingly, performance begins to decline with five utterances (61.73%) and decreases further for six utterances (54.15%).

This trend suggests that incorporating multiple context sentences is beneficial for sarcasm detection up to a certain threshold, likely due to the added information that helps disambiguate sarcasm. However, Beyond this threshold, additional context introduces noise, reducing accuracy.

These findings highlight the delicate balance required when using contextual information in sarcasm detection models. While context is undeniably valuable, too much of it may dilute or obscure the relevant features, underscoring the need for careful consideration when determining the optimal amount of context to include.

### 4.3 Syntax sensitive model

The second experiment aimed to explore the role of the syntactic features of the response itself, in the context. However, after training and evaluating the updated model with the added syntactic features, the same accuracy (64%) was observed compared to the main model. This result disproves our hypothesis and suggests that the addition of syntactic features neither improved nor distracted the model from focusing on contextual cues. Since BERT already encodes syntactic structures through its attention mechanism, explicitly adding syntax features may have introduced redundancy without providing additional benefits.

In a platform like Reddit, known for its casual communication style and user base predominantly aged 18–29, sarcasm often relies more on context than on rigid syntactic structures (**?**). As a result, emphasizing syntax does not suggest improvement in this setting.

However, it is worth to mention that while the accuracy metric is identical, this model has a different confusion matrix compared to the main model. Indeed, the main model favors precision over recall, while the syntax sensitive model favors recall over precision [4].

## 5 Discussion and Conclusion

This study explored the complexities of sarcasm detection in natural language processing, focusing on the role of context and syntactic features. By implementing a hierarchical BERT-based architecture inspired by Srivastava et al. (2020) **?**, we investigated how varying the amount of contextual information and incorporating syntactic features impacted the performance of sarcasm detection models.

[3] Confusion matrices can be accessed via the README for more details

[4] The confusion matrix can be access via the README for more details

## 5.1 Discussion

### 5.1.1 Role of Contextual Information

Our findings highlight the critical role of context in improving sarcasm detection accuracy. Models trained on three context utterances consistently outperformed those with fewer or excessive utterances, achieving a peak accuracy of 69.59%. This result aligns with our hypothesis that context enhances understanding up to a threshold, beyond which irrelevant or noisy information may degrade model performance. The diminishing returns observed with five or more utterances emphasize the importance of optimizing context size, or more specifically to which utterances should be payed more attention in relation to the response, to balance information richness and noise.

### 5.1.2 Impact of Syntactic Features

Contrary to our expectations, adding explicit syntactic features resulted in the same accuracy (64%) as the main model. This outcome suggests that BERT's self-attention mechanism already captures syntactic relationships effectively, making additional explicit inputs redundant. The team referenced to prior work by **?**, which has shown that BERT implicitly encodes syntactic structures, including dependency relations, within its learned representations. Adding explicit syntactic features introduces overlapping information, which does not improve performance and may even add unnecessary computational overhead. In addition, it suggests that context is the primary feature that the model is using for classification. This finding underscores the need for careful feature selection and highlights the limitations of syntax-driven approaches in informal communication platforms like Reddit, where conversational style often overrides formal syntactic patterns.

### 5.1.3 Comparison to Baselines

The hierarchical BERT model demonstrated superior performance compared to baseline models, achieving an accuracy of 64% on the test set. However, the Simplified Hierarchical Sarcasm Detection Baseline achieved a significantly lower accuracy of 49%, highlighting the importance of incorporating conversational context. The simplified model, while aligned with a straightforward classification task using only response text, may suffer from overfitting due to its complexity. Similarly, the logistic regression baseline achieved its peak accuracy of 59% on a trigram BoW. These lower accuracies illustrate the added value of incorporating hierarchical context encoding. These results validate the effectiveness of hierarchical models in capturing complex dependencies in conversational data; i.e, capturing contextual relationship in a conversation.

## 5.2 Conclusion

This work advances the field of sarcasm detection by providing empirical evidence for the optimal use of contextual information and by analyzing the limitations of syntax-driven approaches. The hierarchical BERT model's superior performance reinforces the importance of leveraging both semantic and contextual information for this challenging task.

While this study provided valuable insights, it is not without limitations. The lack of hyperparameter tuning, due to computational constraints, may have marginally hindered the model's performance. Future research could explore fine-tuning hyperparameters such as learning rate, batch size, and dropout rate to optimize convergence and generalization. Additionally, investigating adaptive context-selection techniques could further enhance accuracy and robustness. The findings support the hypothesis that context plays a critical role in sarcasm detection and that accuracy declines beyond a certain threshold of utterances. Therefore, future work could focus on developing strategies to effectively allocate attention across utterances. Furthermore, incorporating multimodal data, such as emojis and images, could pave the way for more sophisticated sarcasm detection in diverse communication contexts.

By addressing these challenges, future research can build on this foundation to develop more robust, efficient, and adaptable sarcasm detection systems for real-world applications.

## Statement of contributions

For this project, Simon led research on existing sarcasm detection models, and implemented, trained, and tested both baseline and hierarchical BERT models. On the other hand, Leen conducted the syntactic structure experiment while Idris conducted the contextual information experiment. All three members of the team contributed to the writing of the report.

## Limitations

This study has several notable limitations that impact the generalizability and robustness of its findings.

First, the dataset used for experimentation was sourced exclusively from Reddit. While Reddit users often communicate casually and employ unique forms of sarcasm, this style may not translate to other platforms, such as LinkedIn, where communication is more formal. This platform-specific focus restricts the applicability of our findings to other online contexts.

Second, the dataset was limited to English, which constrains the study's relevance to other languages. Sarcasm can manifest differently across languages and cultures, potentially requiring language-specific models or approaches to capture these nuances effectively.

Finally, the study faced significant computational resource limitations. Despite paying for Colab's premium services, the resources were insufficient to enable exhaustive experimentation, particularly for fine-tuning hyperparameters. This constraint may have hindered the optimization of the model's performance and limited the scope of experiments that could be conducted.

In conclusion, the findings of this study are constrained by the exclusive use of a single Reddit dataset, the restriction to English-language data, and limited computational resources. Future research should address these limitations by leveraging multilingual and cross-platform datasets and employing more extensive computational resources to enhance model performance and generalizability.

## Ethics Statement

The data used for this study is from Reddit, a public platform. Hence, it may include content related to controversial, political, or social topics. However, the team maintains neutral stance regarding the views expressed. Furthermore, all members of the team recognize that sarcasm detection systems could be used in ways that raise ethical concerns. To address this, the team emphasizes the importance of responsible use of NLP, and reaffirms that this work is exclusively intended for academic purposes.

## References