**Proposed Novel Methodology**

Work carried out to date has identified several features which may serve to improve outcomes for sarcasm detection models. Chapter 6 has concluded that collection methodologies for annotated data for training has significant implications for the data obtained in terms of the structural and sentimental features. Based upon this assessment, expansions upon the current training dataset must be carried out mindfully to ensure that the characteristics of organic sarcastic data are retained. Additional observations regarding features which have a statistically significant difference in sarcastic and non-sarcastic text were identified; this will form the basis for feature extraction prior to model training. Additional observations which will be subject to evaluation for consideration in the proposed model include:

* Both semantics and sentiment have been identified as features with distinct markers in sarcasm. Word vectors are limited in their capacity to effectively convey both semantic relationships and sentiment concurrently; such an effect is fundamental to the nature of language; where sentiment of two words is opposing, the words may be semantically linked. Consider *good* and *bad* which are semantically adjacent however the sentiment is opposite.
* Subjectivity of sarcasm identification is increased where the topic relates to personal beliefs rather than more generalised humour. An observation relating to this effect is that the belief held by the majority seems to be classified as non-sarcastic at higher rates and the converse is true where the belief is regarded as more controversial. There were notable differences in the use of emojis in these cases- sentiment congruence was high between emojis and text in the non-sarcastic content and the opposite was true for the sarcastic content.
* The presentation of sarcasm is different in varying contexts. Negative sarcastic tweets were observed to use emojis disproportionately to reduce perceived negativity. The converse was true to some extent for positive sarcastic content; however, the effect was far less universal.

**Proposed Architecture**

*Sentiment-Aware Attention Mechanism*

This section aims to evaluate an approach to ensure both semantic and sentiment information can be considered for the purpose of sarcasm detection as previous evaluation has determined that both features are likely to play a role in understanding the underlying patterns in sarcastic content. The consideration of two sets of word vectors, optimised for semantic and sentiment information respectively is a possible solution for exploration, which may enable a more nuanced representation of the words by the model. This option may improve outcomes for highly specific tasks such as sarcasm detection. However, there are limitations to this approach; this would significantly increase complexity and increase data requirements for the task. Additionally, where the vectors contain overlapping information, there may be interference, reducing model performance necessitating increasingly robust measures against noise which would likely impede the learning of subtle nuances which characterise sarcasm. This evaluation leads to a conclusion that a better approach would consider these features in a manner which does not include additional word vectors.

The fundamental purpose of attention mechanisms is to mirror cognitive attention within text. This is computed though a process detailed within section X for transformer-based models. Attention mechanisms are not generally utilised for the purpose of enhancing the sentiment awareness of a model, however in this context this poses potential to ensure both sentiment and semantic information can be considered in the model architecture. To achieve this, the standard attention mechanism will be modified to weight word importance with regards to their polar sentiment.

**Data Preparation**

*Expansion of Current Training Dataset*

Evaluation to this point has established that sarcasm detection is a highly complex process, and thus is likely to require a large dataset for training. Previous work in chapter 6 has established that annotation strategies which rely on weak labelling using hashtags yield text which represents sarcasm significantly differently than human-annotation, however this is not to say that all data contained in these datasets are unrepresentative of organic sarcasm. This work aims to evaluate a series of datasets annotated for sarcasm detection collected in this manner and identify any text which is aligned with observations about the human-annotated data which was shown to be characteristic of sarcastic text previously. The annotation strategies for the evaluated datasets are as follows:

1. Annotation based on the presence of #sarcasm, #sarcastic or #irony in the text.
2. Annotation of based upon the presence of #sarcastictweet and #not in addition to a set of offensive vocabulary words.
3. Annotation strategy based on the presence of key words and hashtags in addition to semi-supervised learning techniques.
4. Annotation strategy based on the presence of key words and hashtags in addition to semi-supervised learning techniques.
5. Annotation based on the idea that sarcastic content consists of text containing both positive and negative verb phrases in the same document (tweet) in addition to a set of key words.

*Statistical Evaluation to Identify Aligned Features*

The goal of this work was to establish a method for removal of data which was unrepresentative of organic sarcasm present in the datasets due to the limitations of hashtag-based annotation strategies. Such a task was carried out by establishing a range for features previously observed to have a statistically significant difference between sarcastic and non-sarcastic text occur in the validated data at 95% confidence. Using this analysis, data can be selectively omitted from the dataset to leave only data which is the most aligned with the data found using validated methodologies. Standard calculations for a 95% confidence interval were implemented with the following adjustments to fit the context of the data:

* Where measuring instances of a feature, the minimum value for the lower limit of the range at 95% confidence was defined as zero to align with what is possible.
* Rounding of the results to up the nearest integer was utilised in cases where metrics represent the counting of features, as this must result in an integer value. Justification for this action is clear considering the example of the hashtags per tweet parameter; the range at 95% confidence is defined as 0.00 < x < 0.967 which results in the exclusion of all instances where a hashtag occurs, which is not necessarily valid. The adjustment of the obtained result in cases like this are reasonable as the dataset which establishes these boundaries is relatively small and the range must be viewed in the context of the limitations associated with small datasets.

*Table X* Baseline for Evaluation of Tweets for Organic Sarcasm Features.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | | **Standard Deviation** | | **Range at 95% confidence** | |
|  | **Positive Subset** | **Negative Subset** | **Positive Subset** | **Negative Subset** | **Positive Subset** | **Negative Subset** |
|  | **Emoji-Based Features** | | | | | |
| **Emojis Per Tweet** | 0.569 | 0.507 | 0.803 | 0.680 | UL: 2.17  LL: 0.00 | UL: 1.87  LL: 0.00 |
| **Sentiment Score** | 0.323 | -0.0514 | 0.412 | 0.413 | UL: 1.15  LL: -0.500 | UL: 0.775  LL: -0.878 |
| **Degree of Positivity** | 0.466 | 0.261 | 0.255 | 0.200 | UL: 0.977  LL: 0.00 | UL: 0.661  LL: 0.00 |
| **Degree of Negativity** | 0.143 | 0.312 | 0.178 | 0.229 | UL: 0.498  LL: 0.00 | UL: 0.770  LL: 0.00 |
| **Degree of Neutrality** | 0.391 | 0.427 | 0.155 | 0.118 | UL: 0.701  LL: 0.0800 | UL: 0.662  LL: 0.191 |
|  | **Text-Based Features** | | | | | |
| **Degree of Positivity** | 0.667 | 0.0707 | 0.282 | 0.0698 | UL: 1.23  LL: 0.102 | UL: 0.210  LL: 0.00 |
| **Degree of Negativity** | 0.0520 | 0.589 | 0.0695 | 0.246 | UL: 0.191  LL: 0.00 | UL: 1.08  LL: 0.00 |
| **Degree of Neutrality** | 0.281 | 0.340 | 0.247 | 0.205 | UL: 0.775  LL: 0.00 | UL: 0.751  LL: 0.00 |
| **Hashtags per Tweet** | 0.164 | 0.104 | 0.637 | 0.431 | UL: 1.44  LL 0.00 | UL: 0.967  LL: 0.00 |
| **Laughter Markers per Tweet** | 0.0336 | 0.0116 | 0.211 | 0.107 | UL: 0.455  LL: 0.00 | UL: 0.226  LL: 0.00 |
| **Affirmatives per Tweet** | 0.502 | 0.533 | 0.844 | 0.828 | UL: 2.19  LL: 0.00 | UL: 2.19  LL: 0.00 |
| **Negations per Tweet** | 0.379 | 0.651 | 0.779 | 1.06 | UL:1.94  LL: 0.00 | UL: 2.78  LL: 0.00 |
| **Intensifiers per Tweet** | 0.247 | 0.320 | 0.483 | 0.678 | UL: 1.21  LL: 0.00 | UL: 1.67  LL: 0.00 |
| **Interjections per Tweet** | 1.08 | 1.70 | 1.74 | 1.70 | UL: 4.57  LL: 0.00 | UL: 4.78  LL: 0.00 |
| **Relevant Punctuation per Tweet** | 0.423 | 0.528 | 1.03 | 1.29 | UL: 2.59  LL: 0.00 | UL: 3.11  LL: 0.00 |
| **User Mentions per Tweet** | 0.227 | 0.258 | 0.584 | 0.754 | UL: 1.40  LL: 0.00 | UL: 1.77  LL: 0.00 |
| **Capitalised Words per Tweet** | 2.00 | 2.13 | 2.19 | 2.28 | UL: 6.38  LL: 0.00 | UL: 6.69  LL: 0.00 |
| **Mid-Word Capitalisations per Tweet** | 0.551 | 0.603 | 1.35 | 1.69 | UL: 3.25  LL: 0.00 | UL: 3.94  LL: 0.00 |

Previous chapters establish that sarcasm presents differently where there is positive and negative sentiment in the text. For this reason, evaluation of each dataset has been broken down into these subsets for a more granular approach to the task. Figures X and X show results of the analysis for each dataset.

A group of blue and white bars

Description automatically generated

*Figure X* Atypical Feature Identification in Datasets Annotated using Hashtag Strategy by Feature.

A screenshot of a graph

Description automatically generated

*Figure X* Atypical Feature Identification in Datasets Annotated using Hashtag Strategy by Subset.

*Results*

In every case a significant proportion of text prompts were found to contain atypical presentations of sarcasm based on the established baseline from the results contained in the datasets which utilised human-annotation methods. Given the more aligned results observed for datasets 3 and 4, the addition of machine learning improves outcomes for training data collection. Details with regards to the techniques were not provided within the associated literature therefore greater depth of evaluation is difficult. Across the datasets available, 72% of the tweets were found to display incongruent characteristics compared to the present gold-standard collection strategy used in the iSarcasm dataset. Given the assessed datasets are the most prominently referenced across sarcasm detection literature, validity concerns are evident with regards to the true capabilities of resulting models to identify organic sarcasm online.

For the purposes of expansion upon the present training dataset, a cautious approach was implemented which includes only data which has no features which present atypically compared to the defined baseline. This yielded an additional 38593 sarcastic tweets for consideration during model training.

This strategy ensures that the bias associated with poor annotation strategy is mitigated, however the results are limited to align with the characteristics of the previously validated data. In any instance using this strategy, the scope of what is represented in the final dataset is limited to what aligns with the baseline. Where some valid instances of organic sarcasm are unrepresented in the baseline, they will be excluded from the collected data also. However, this serves to exclude data which is unrepresentative of how sarcasm usually presents; it is not typical for sarcastic content to contain hashtags like #sarcasm therefore the training data should reflect this. Expansion upon the availability of human-annotated data in future would serve to improve outcomes for representation of organic sarcasm within the dataset, leading to a more comprehensive understanding of patterns which are present in sarcastic content.

*Data preparation*

To ensure patterns learned for sarcastic content are as robust as possible, it is important to ensure that the data is processed appropriately prior to model training. The process for cleaning was dependent on the type of model to be trained. The processes have been detailed in section X and X. Following the data cleaning process, a control subset was generated which omitted all emojis from the text.

A diagram of a process

Description automatically generated with medium confidence

*Figure X* Data cleaning steps for neural network model training.

**Model Selection and Evaluation**

Neural networks are a primary area of potential identified in literature for sarcasm detection tasks, with the primary models used being CNN, LSTM and GRU models. Suitability based on their respective architectures are evaluated in section X. Each identified model was hyperparameter tuned using a similar methodology to that which was described previously in section X, with the goal of identifying the most optimal model with regards to accuracy and generalisation capabilities. The assessment metrics were adjusted to account for the nature of the task, binary classification.

During hyperparameter tuning accuracy was the primary metric used to evaluate model performance. The metric reports the proportion of correct predictions made by the model. This was selected based on the ease of interpretation and its suitability given the balanced nature of the dataset.

Following initial tuning, during the assessment of models with varying data input, F1 score was additionally considered which provides insight into the precision and recall of the model.

**Results**

\*\*Plan for this work:

Building the sentiment embedding

Start using dimensional data- more simple (can expand to include basic theory also if I have time)

I have positive and negative scores for all the emojis from previous work

Combine this with sentiwordnet and I will have a full set of info for emoji and words

(Check later if emojis generally are high or low attention for sentiment- maybe also in general if relevant)

Scale the sentiment scores from 1-10 and round to the nearest number-> granular but not extremely complicated.

If I take a sentence

How are you today?

Each word gets a positive and negative sentiment intensity.

Make a positive and a negative sentiment matrix

Final embedding combines results from the positive and the negative sentiment embedding.

Incorporating the embedding into the attention mechanism:

Alternative option-> emsemble methods?

For the controversial opinions mining?

* Maybe manually identify relevant topics (discuss why and limitations of the previous topic modelling here)
* Make embedding to incorporate this info into the model?
* Discuss that this is to some degree accounted for using the sentiment aware attention mechanism as incongruence/congruence between text and emoji sentiment can be identified this way- which was one marker

**Evaluation of Present State-of-the-art Sarcasm Detection Models**

The literature review highlighted several model architectures which are commonly used for sarcasm detection. The aim of this section is to evaluate the most frequently implemented models with regards to their strengths and limitations for sarcasm detection and their suitability to integrate architectural features to enhance their sarcasm detection capabilities.