**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 less universal.

**Proposed Architecture**

*Model Selection*

Given the complexity of the task and prevalence of similar architectures identified in literature, neural networks will be the model class evaluated for the task. Selection Criteria will consider the architectural features with respect to the task in addition to observations from literature on their suitability.

*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 traditional 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.

**Baseline Model Evaluation**

*Neural Network Selection*

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.

Following hyperparameter tuning, three models with the greatest performance metrics were re-trained using 5-fold cross validation to evaluate their robustness to varied data. The results were used to identify the optimal model.

**Evaluation Metrics**

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.

Initial evaluation during the hyperparameter tuning process identified the GRU model as achieving maximum accuracy while minimizing loss. To balance computational cost with assessment of as many iterations of the model as possible during hyperparameter tuning cross validation was performed on the three models identified with the greatest performance and the final optimal results was selected from this pool.

**Performance Evaluation**

Initial evaluation during the hyperparameter tuning process identified the GRU models as the greatest performing, which is aligned with expectations with respect to their architectural features and literature assessment discussed previously.

*Figure X* Optimal GRU Model for Sarcasm Detection.

While the model yielded the greatest performance with respect to accuracy, f1 score and loss compared, the discrepancy between training metrics compared to test and validation subsets provide strong evidence of overfitting. Such an effect is consolidated by the inconsistencies in results from each fold during cross-validation, however this effect was present in all evaluated models.

*Table X* Performance Metrics of Optimal GRU Model for Sarcasm Detection.

Given this approach utilised the word vectors as the sole source of information to train the model, this result may be related to the over-reliance of the model on semantic information, where sentiment has also been evidenced to be relevant. Further adaptations to the model which enable the consideration of sentiment to a greater extent may be of value for the task. Section X discusses the implementation of a strategy to provide enhanced sentiment awareness to the model.

**Attention Mechanism**

**Architecture**

To integrate sentiment awareness into the model in the most computationally efficient manner, which minimising the addition of noise which may arise from vectors with overlapping information, an attention mechanism which weights the semantic and contextual information gleaned from the word vectors alongside sentiment embeddings.

Sentiment metrics selected for use in this layer were limited to degree of positivity and negativity. Such a selection directly addresses the limitations of adjacent semantics and opposing sentiment, in addition to providing means to learn patterns relating to sentiment incongruency, differing presentation of sarcasm in positive and negative contexts which have each been observed to contribute to what makes sarcasm. This work was limited to the evaluation of the effects of dimensional theory centric embeddings as these parameters were shown to have statistically significant differences in sarcastic text compared to non-sarcastic text. The evidence for parameters generated from basic theory data was more limited. Given this observation, the trade off between increasing computational complexity and likelihood of noise against the increased granularity of the data which enhance learning of subtle patterns indicated that the results were not likely to yield favourable results from their inclusion. The values previously generated in chapter X were limited to emojis only; no dataset could be identified which contains predictions for the up-to-date corpus therefore this resource is of significant value for such a task. Word values were obtained from the SentiWordNet dictionary which provides polar information for degree of positivity and negativity for a large set of words. The sentiment data was converted to embeddings and scaled to integer values for ease of interpretation:

Where and represent the positive and negative sentiment intensity of the words respectively where large values indicate greater intensity towards the given polarity and the converse for lesser values. Out of vocabulary words were represented by 0, indicating an objective word. Given an input string:

Sentiment embeddings (,) containing and respectively can be constructed for string *t*:

The embeddings are introduced into the attention layer of the optimised GRU model and the weighted contribution of the sentiment of each word is determined via a matrix multiplication between the sentiment values and the weight matrix:

Which is subsequently adjusted by a bias term:

Where:

Where *o* represents the output of the GRU layer and and represent positive and negative sentiment intensity embeddings respectively. Each respective transformation is calculated using independent weight and bias to enable varying importance to be assigned to each data type, providing a more flexible framework for optimisation during model training. The summation of these weights is obtained, and a hyperbolic tangent function is applied:

Where:

This introduces non-linearity to the model, increasing its capacity to learn complex patterns which cannot be entirely represented by linear relationships. Without the use of an activation *n* layers in a model can be reduced to a single linear layer, which mirrors linear regression thus is only capable of learning linear functions within the feature space.

Given a definition of linear regression as follows:

Two linear layers can be reduced to a single linear layer as follows:

Where and . Such a definition can be extended to *n* layers by induction. The hyperbolic tangent activation specifically was selected for the introduction of non-linearity due to its mapping of values to a range of -1 to 1 which mitigates instability and exploding gradients, which is not achieved by alternatives such as the sigmoid function. The function was additionally widely documented within literature for similar implementations within neural network and attention architectures.

A softmax function is applied to the output of the tanh function to convert unnormalized attention scores into interpretable probability distributions, which can be added to a total of 1 where greater probability is assigned to higher scores:

Where attention weights, represent the contribution of each token in the string to the final text representation. Finally, the attention weights are applied to the input word vectors:

This approach allows the model to learn the most relevant semantic information from the vectors in addition to considering the sentiment information for its prediction. The use of sentiment determined externally to the model enables the model to learn based off prior knowledge learned from word tokens previously, while also considering the most computationally efficient method to introduce sentiment information.

**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.