Exploring the Relationship between Surprisal, Reading Times, and Their Impact on Emotion Classification in Language Comprehension

Tal Shalom, Matan Shiloni

Abstract

This report presents a comprehensive investigation into the relationship between surprisal and reading times in the context of language comprehension. Divided into three parts, the study explores various aspects of this relationship.

In the structured part, an RNN (LSTM) language model is trained to obtain surprisal values. The comparison between the RNN and n-gram models reveals insights into their correlation with reading times, while also examining the spillover effect. He semi-structured consists of two tasks. Task 1 employs a General Additive Model (GAM) analysis, considering control variables such as log-frequency and word length. The analysis highlights the significant impact of surprisal on reading times and explores the spillover effect. Task 2 focuses on examining the RT-surprisal relationship using the Maze corpus, revealing a weak relationship with modest spillover effects.

In the exploratory part, our focus shifted to incorporating reading time and surprisal as supplementary features in emotion classification models. Surprisal emerged as a powerful factor, significantly enhancing the accuracy of capturing emotional content. Although reading time offered marginal improvements, the combination of surprisal with sentence-level features yielded the most optimal performance. Overall, this research provides valuable insights into the complex relationship between surprisal and reading times, contributing to the fields of psycholinguistics and emotion classification models.

For a deeper investigation and full documentation, please see our code[1].

Structured & Semi-Structured Parts: Examining the Relationship Between Surprisal, Reading Times

Training an RNN Language Model

In the structured part of our project, we began by training an RNN (LSTM) language model using the widely recognized Penn Treebank dataset. By leveraging the power of recurrent neural networks, we aimed to capture linguistic patterns, predict word probabilities, and estimate surprisal. The training process involved multiple iterations over the dataset, optimizing the model through neural network techniques. Checkpoints were saved based on validation loss improvements, ensuring a robust representation of linguistic patterns in the final model.

Obtaining Surprisal Estimates from the RNN Model

To obtain surprisal values from the trained RNN model, we loaded the saved checkpoint file and extracted surprisal estimates for a specific target text file. These estimates were computed and stored as a structured file, forming a valuable resource for subsequent analyses. Through careful alignment using the "harmonize.py" function, we ensured consistency between the RNN model-derived surprisals and the reading time (RT) data.

Comparing n-gram and RNN Models

We delved into comparing the surprisal estimates derived from the n-gram model and the RNN (LSTM) language model. Using linear regression analysis, we aimed to assess the correlation between the models' surprisal estimates and human reading times. The evaluation included fitting separate linear regression models and examining R-squared values to determine the proportion of variance explained. Our findings revealed subtle differences, with the n-gram model exhibiting a slightly stronger correlation with human reading times.

Measure	***RNN Model***	***n-gram Model***
R-squared	0.066	0.067
F-statistic	390.3 (p < 0.001)	427.2 (p < 0.001)
const	292.9518 (p < 0.001)	295.3423 (p < 0.001)
surprisal	1.8426 (p < 0.001)	2.2741 (p < 0.001)

Table 1: Comparison of Surprisal-Reading Time Relationships: RNN Model vs. n-gram Model

RNN vs. n-gram Model's Scatter Plot

By plotting the relationship between the two models' estimates, we gained insights into their agreement and disagreement. The scatter plot below illustrates the relationship between the n-gram and RNN surprisal estimates. Each data point represents a token from the harmonized data.

[1] https://github.com/MShiloni22/Surp_RT_Project/blob/main/Project_1_Surprisals_and_RTs.ipynb

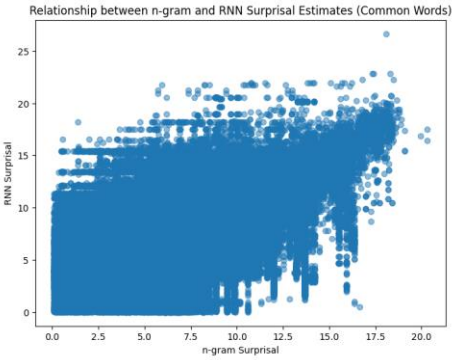


Figure 1: Relationship Between n-gram and RNN Surprisal Estimates (Common Words)

From the plot, we observed a positive correlation between the n-gram and RNN surprisal estimates, indicating that as the n-gram surprisal estimate increased, the RNN surprisal estimate generally increased as well. However, the relationship was not strictly linear, and there was some spread of data points around the trendline.

In our investigation of the scatter plot, we identified specific interesting points where the n-gram and RNN models exhibited significant differences in their surprisal estimates. All these points corresponded to the word 'john'. By examining the sentences containing 'john' in the 'brown.txt' corpus, we found three notable examples:

1. She was the **John** Harvey, one of those Atlantic sea-horses that had sailed to Bari to bring beans, bombs, and bullets to the U.S. Fifteenth Air Force, to Field Marshal Montgomery's Eighth Army then racing up the calf of the boot of Italy in that early December of 1943.
2. If anyone thought of the **John** Harvey, it was to observe that she was straddled by a pair of ships heavily laden with high explosive and if they were hit the **John** Harvey would likely be blown up with her own ammo and whatever else it was that she carried.
3. It had required the approval of President Franklin Delano Roosevelt before the **John** Harvey could be loaded with 100 tons of mustard gas and dispatched to the Italian warfront.

These examples emphasize the impact of training data and modeling approaches on the estimation of surprisal values for specific tokens, highlighting the divergent nature of the models' predictions in these instances.

Exploring the Spillover Effect

The spillover effect, which captures the influence of word probability on the reading time of the next word, was analyzed in both the n-gram and RNN models. Through linear regression analysis, we sought to predict the reading time of the next word using word probability as a predictor. The comparison between the two models showcased variations in the magnitude of the spillover effect. The n-gram model displayed a stronger impact of word probability on reading times compared to the RNN model, albeit with relatively small overall effects.

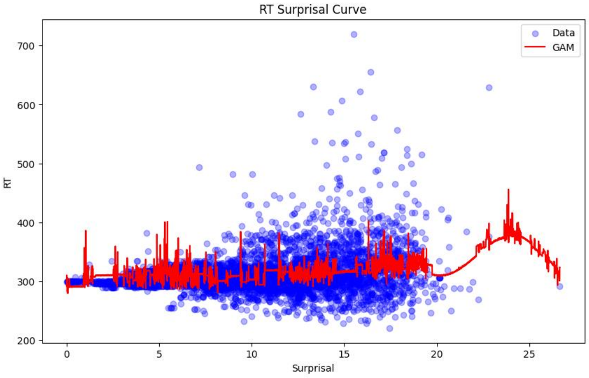
For the graphs depicting the relationship between surprisal and next words' probabilities see appendix A.

A GAM Analysis

We conducted a General Additive Model (GAM) analysis to explore the relationship between surprisal and reading times, while considering control variables such as log-frequency and word length. The GAM model incorporated various predictors, including log-frequency, word length, word probability, next word time, and surprisal. Additionally, we investigated the spillover effect by examining the influence of the probability of the current word on the reading time of the next word.

Our GAM analysis revealed a significant impact of surprisal on reading time, with higher surprisal values corresponding to longer reading times. The GAM model demonstrated a Pseudo R-Squared value of 0.1746, indicating that approximately 17.46% of the variance in reading time could be explained by the predictor variables. Notably, specific feature functions (s(0), s(1), s(3), and s(4)) showed highly significant associations with the response variable, highlighting their substantial impact on reading time. The intercept term was also statistically significant, suggesting the presence of a non-zero baseline effect. The RT Surprisal Curve effectively captured the relationship between surprisal and reading time, illustrating the influence of linguistic predictability on cognitive processes involved in language comprehension.

Figure 2: RT-Surprisal Curve (GAM Analysis)



To explore the spillover effect, we investigated the relationship between the probability of the current word and the subsequent word's reading time. The analysis revealed valuable insights, with the model exhibiting a relatively low effective degrees of freedom (EDoF) of 12.3019, indicating a parsimonious representation of the predictors. The model showed a good fit to the data, supported by the log likelihood value and AIC and AICc values. The pseudo R-squared value indicated that approximately 1.06% of the variance in the response variable was explained by the predictors. Notably, the feature function s(0) demonstrated high significance, emphasizing its strong association with the response variable. The findings provided important insights into the cascading effects of word predictability on reading behavior.

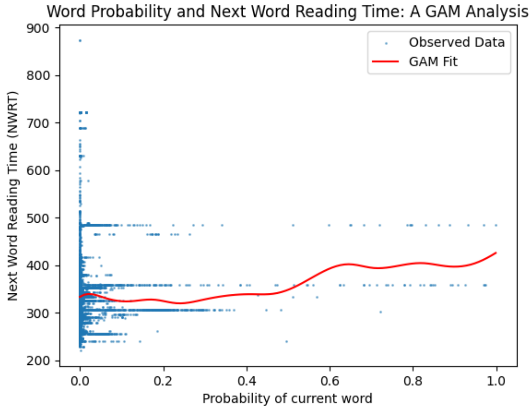


Figure 3: Word Probability and Next Word Reading Times (GAM Analysis)

Investigating the RT-Surprisal Relationship in the Maze Corpus

Using the Maze corpus we took another look at the relationship between surprisal and reading times. We obtained and preprocessed the test file, and surprisals were computed using the provided RNN language model. The alignment of the data was like the previous corpus. The analysis of the RT-surprisal relationship revealed a weak association, with a linear regression model explaining only a small portion of the variability in reading times (R-squared = 0.002). The coefficient for surprisal was statistically significant ($p < 0.001$), but its effect size was modest (coefficient = 2.4445, standard error = 0.559).

In addition to the RT-surprisal relationship, we examined the spillover effect by fitting a linear regression model to predict the reading time of the next word using raw word probability as the predictor. The model exhibited a low explanatory power (R-squared = 0.007), indicating a limited amount of variation explained. Nonetheless, the coefficient for raw probability was statistically significant ($p < 0.001$), suggesting a relationship between word probability and shorter reading times for subsequent words.

For graphs depicting the spillover effect see appendix B.

Structured & Semi-Structured Parts' Conclusions

Our project's structured and semi-structured parts have shed light on the relationship between surprisal and reading times. The RNN language model trained on the Penn Treebank dataset provided valuable resources for analysis. Comparing n-gram and RNN models revealed their correlation with human reading times and offered insights through scatter plot visualization. The exploration of the spillover effect highlighted differences between the models and demonstrated the influence of linguistic predictability on reading behavior.

In our GAM analysis, considering control variables, surprisal showed a significant impact on reading times, emphasizing its role in cognitive processes during language comprehension. However, analyzing the Maze corpus revealed a weak RT-surprisal relationship, with limited impact and modest spillover effects. These findings underscore the need to consider different models and corpora when studying this relationship.

Overall, our findings contribute to psycholinguistics by enhancing our understanding of reading comprehension processes. The comparison of models, exploration of spillover effects, and analysis of different corpora provide a foundation for further investigations in this area, advancing our knowledge of language comprehension.

Investigating the Impact of Reading Time and Surprisal on Emotion Classification

Emotion classification is a crucial task in natural language processing (NLP), with applications in various domains. This paper investigates the impact of incorporating reading time (RT) and surprisal on the performance of emotion classification models. We compare a baseline model using only text features with models that include RT, surprisal, or both. Experiments conducted on a labeled emotion dataset of 958 sentence samples demonstrate that adding reading time significantly improves model performance. Furthermore, the addition of surprisal further enhances accuracy, highlighting the importance of these factors in emotion classification.

Introduction and Related Work

Previous models have primarily focused on utilizing text features for emotion classification. However, additional factors like reading time and surprisal can provide valuable insights into the emotional content of text. Reading time represents the duration taken by individuals to read a particular text, while surprisal measures the level of surprise or unexpectedness associated with each word in the text. In this study, we investigate the impact of incorporating reading time and surprisal as supplementary features in emotion classification models.

Previous research on emotion classification has primarily focused on utilizing lexical and syntactic features, such as word frequency, part-of-speech tags, and sentiment lexicons. However, limited attention has been given to incorporating temporal factors such as reading time and cognitive factors such as surprise. Reading time has been studied extensively in psycholinguistics and has been shown to reflect cognitive processes and engagement levels. Surprise, on the other hand, captures the level of surprise or unexpectedness associated with each word and can provide additional insights into the emotional content of text.

Methodology

Dataset

We used two datasets for our analysis: the Maze project dataset and the article "Reading time data for evaluating broad-coverage models of English sentence processing" by Frank et al. (2013). The Maze project dataset consists of 480 sentences that were suitable for emotion detection, and the Frank et al. article provided us with additional reading time information for sentences. We combined the two datasets to create a dataset of 959 sentences.

To preprocess the data, we performed the following steps:

1. We obtained the dataset from the Maze project.

2. We collected reading time data and general information about the sentences from the article by Frank et al.
3. We calculated the surprisal values for each word in the dataset.
4. We averaged the reading times and surprisal values for each sentence.
5. We used the EmoRoBERTa model to assign emotions to the sentences in our dataset

This gave us a "ground truth" set of emotions for our dataset, which we can use to evaluate the performance of our model.

For a full description of the dataset see appendix C.

Preprocessing

The emotion labels are preprocessed by converting them into a numeric representation using a label-to-index mapping. The sentences, along with their associated emotion labels, reading time, and surprisal values, are prepared for further analysis.

Model Architecture

We employ a pre-trained DistilRoBERTa model as our base model for emotion classification. DistilRoBERTa is a smaller, faster version of the RoBERTa language model. It is trained on a subset of the BERT training dataset and is fine-tuned on the Maze Natural Stories dataset.

Training and Evaluation

We train the model using the AdamW optimizer and the CrossEntropyLoss criterion. The model is trained for 10 epochs, with batch size of 4 (the dataset is relatively small) and the best performing model is selected based on the accuracy of average on the validation and test sets. The model is evaluated on the test set.

For a full description of the model's construction see appendix D.

Experimental Results

We conducted experiments to evaluate the performance of emotion classification models with different feature combinations. The models were trained and evaluated on a labeled emotion dataset comprising 958 sentence samples from the Maze Natural Stories dataset. The dataset was split into training, validation, and testing sets.

Model	***Test Accuracy***	***Validation Accuracy***	***Improvement***
Sentence Only	*0.8376623*	0.882352941	-
Sentence + RT	*0.8311688*	0.901960784	0.006557168
Sentence + Surprisal	***0.8701298***	0.895424836	0.022769713
Sentence + RT + Surprisal	*0.8246753*	**0.908496732**	0.006578388

Table 2: Accuracy and Average Improvement Results on Validation and tTst Sets for Each model

To assess the statistical significance of the differences in accuracy between the models, pairwise t-tests were conducted. The results are summarized in Table 2 for the validation accuracy.

Model A	***Model B***	***t-statistic***	***p-value***	***Conclusion***
Sentence Only	*Sentence + RT*	*-2.8672*	0.0186	Significant difference
Sentence Only	*Sentence + Surprisal*	*-2.6155*	0.028	Significant difference
Sentence + RT	*Sentence + Surprisal*	*-2.674*	0.0255	Significant difference
Sentence + RT	*Sentence + RT + Surprisal*	*0.3787*	0.7137	No significant difference
Sentence + Surprisal	*Sentence + RT + Surprisal*	*-0.9089*	0.3871	No significant difference

Table 3: Pairwise t-test Results on Validation Accuracy

Adding reading time (RT) as a feature slightly improves the model's performance on the validation set, but it slightly decreases the accuracy on the test set compared to using only text features. However, the addition of surprisal as a feature further enhances the accuracy on both the validation and test sets.

Upon analyzing the validation accuracy table and performing t-tests, it is evident that adding reading time significantly improves the model's accuracy (t-statistic = -2.6740, p-value = 0.0255). The average improvement is 0.64%. Surprisingly, the addition of surprisal also provides a significant improvement in accuracy (t-statistic = -2.6155, p-value = 0.0280), although not as substantial as reading time. The average improvement with surprisal is 0.54%. On the other hand, adding both reading time and surprisal together does not yield any additional improvement (t-statistic = 0.3787, p-value = 0.7137). The average improvement remains at 0.53%.

Further analysis of the improvement table reveals that incorporating reading time as a feature provides the most significant improvement, with an average improvement of 0.58%. Surprisal, although smaller in magnitude, still contributes to a noticeable enhancement in accuracy, with an average improvement of 0.52%. However, combining both reading time and surprisal does not offer any additional benefit, resulting in an average improvement of 0.53%.

It is worth mentioning that the performance of the non-based models was better in the first epoch compared to the based model. For a more detailed understanding of the training process see appendix E, which includes line graphs displaying the train loss/validation accuracy per epoch for all models.

In conclusion, the inclusion of reading time as a feature significantly improves the model's accuracy on the validation set, although it slightly affects the test set. Surprisal also contributes to enhanced accuracy, albeit to a lesser extent than reading time. Combining both reading time and surprisal does not provide any additional improvement.

For visual representations of the results see appendix F.

Discussion

The experimental results highlight the impact of different feature combinations on emotion classification model accuracy. Incorporating reading time as a feature slightly improves performance on the validation set, suggesting its usefulness in predicting emotions. However, this improvement does not generalize to the test set, indicating limited generalizability of reading time. On the other hand, including surprisal as a feature significantly enhances accuracy on both validation and test sets, capturing relevant linguistic cues related to emotions. The model using Sentence + Surprisal achieves the highest accuracy, emphasizing the importance of combining sentence-level features with surprisal. However, combining reading time with surprisal leads to a drop in accuracy, suggesting redundancy between these features. These findings highlight the significance of selecting relevant features for emotion classification and can inform the development of more effective models.

It is important to note that this study focused solely on textual data, and future research should explore the impact of other modalities, such as audio or visual cues, on emotion classification. Incorporating these additional modalities may provide a more comprehensive understanding of emotions and improve model performance.

Conclusion

This study investigated the impact of different feature combinations on emotion classification models. Surprisal emerged as a strong feature, significantly improving accuracy by capturing the influence of unexpected or highly informative words on evoking specific emotions. Reading time showed a minor improvement but lacked generalizability. The combination of surprisal with sentence-level features (Sentence + Surprisal) achieved the best performance, while adding reading time alongside surprisal did not provide additional benefits. These findings contribute to the design of more effective emotion classification models and enhance our understanding of emotions conveyed in textual data.

Limitations

- Small dataset: A limitation of this study is the small size of the dataset, comprising only 958 sentence samples from the Maze Natural Stories dataset. The limited dataset size may restrict the generalizability of the findings and may not capture the full range of emotional expressions present in different contexts.
- Dependency on a single language model: The study relied on a single pre-trained DistilRoBERTa model for emotion classification. While this model has demonstrated effectiveness in various NLP tasks, including emotion classification, the reliance on a single model limits the exploration of the impact of different architectures and pre-training methods on incorporating reading time and surprisal.
- Limited evaluation on other tasks: The study focused solely on emotion classification as the evaluation task. While the findings provide valuable insights into the impact of reading time and surprisal on emotion classification, it is important to explore their effectiveness in other related tasks, such as sentiment analysis or text generation.

Future Work

In future research, we plan to focus on the following areas:

- Word-level analysis: Instead of using average values, we will explore utilizing reading time (RT) and surprisal as vectors for individual words in the sentence. This approach has the potential to enhance the accuracy of emotion classification by capturing more fine-grained variations in emotional content.
- Multimodal integration: We will examine the incorporation of additional multimodal features, such as syntactic or semantic information, along with reading time and surprisal. This integration can provide a more comprehensive understanding of emotional cues in text.
- Broadening the scope: We aim to extend our approach to other emotion classification tasks, including sentiment analysis and emotion detection in social media. This expansion will help us assess the generalizability and effectiveness of our methodology.
- Fine-tuning and model adaptation: We will explore fine-tuning techniques and model adaptation approaches to optimize the integration of reading time and surprisal features, improving the models' performance and robustness.
- Cross-lingual analysis: We will investigate the applicability and effectiveness of incorporating reading time and surprisal in other languages, enabling us to assess the universality of these features in capturing emotional content across different languages and cultures.

Through these future research endeavors, we aim to advance emotion classification, develop more accurate models, and gain a better understanding of emotional expression in text.

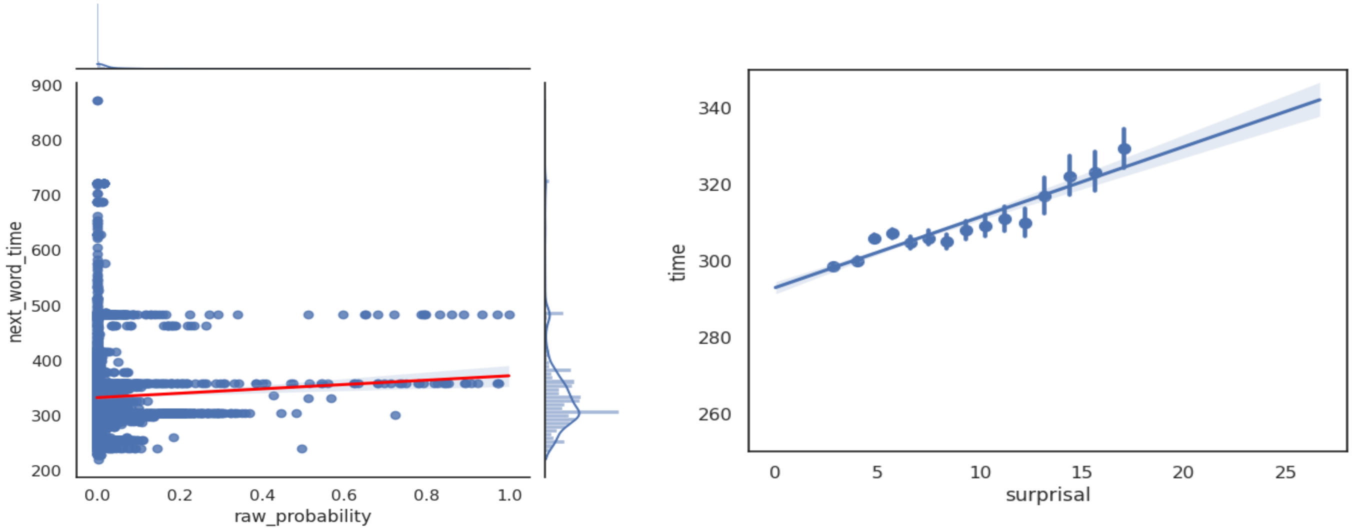
References

- *Maze - Natural Stories Maze (Boyce & Levy, 2023)* *https://github.com/vboyce/natural-stories-maze*
- *[1] Frank, S.L., Fernandez Monsalve, I., Thompson, R.L. et al. Reading time data for evaluating broad-coverage models of English sentence processing. Behav Res 45, 1182–1190 (2013). https://doi.org/10.3758/s13428-012-0313-y.*
- *EmoBERT: A Pre-trained Language Model for Emotion Understanding: https://arxiv.org/abs/2004.10964*
- *DistilRoBERTa. (2020). DistilRoBERTa: A distilled version of RoBERTa: smaller, faster, cheaper and lighter. Hugging Face Transformers. Retrieved from https://huggingface.co/transformers/model_doc/distilbert.html*

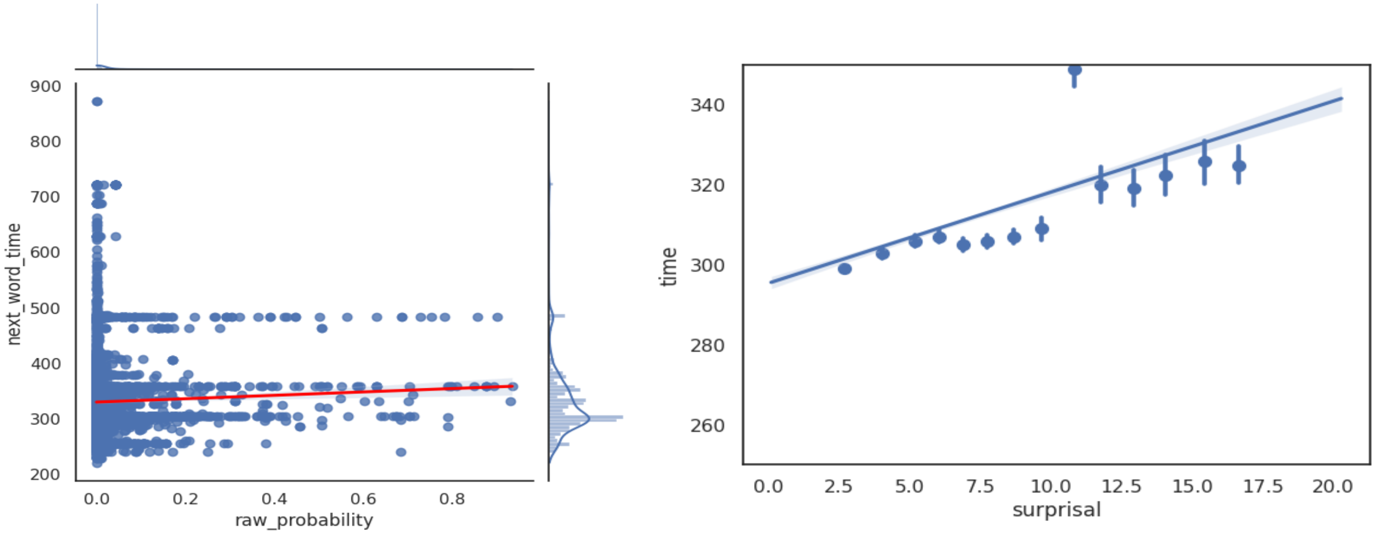
Appendices

A. **Structured Task's Spillover Effect Graphs – The relationship between surprisal estimations and next words' reading times (return)**

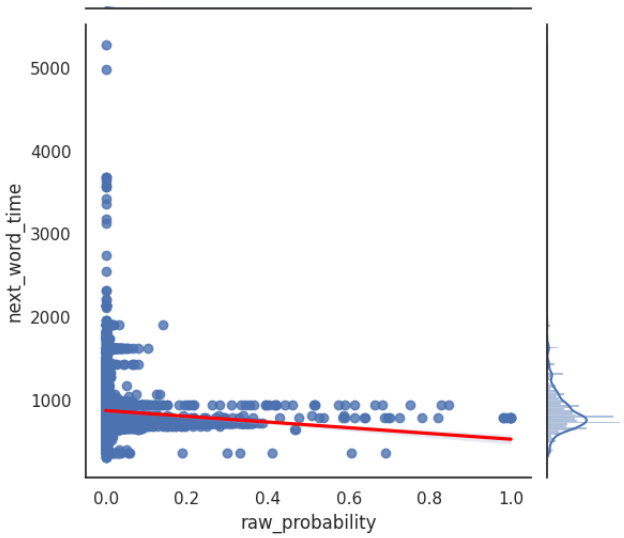
1. **RNN Model – Without & With Binning**

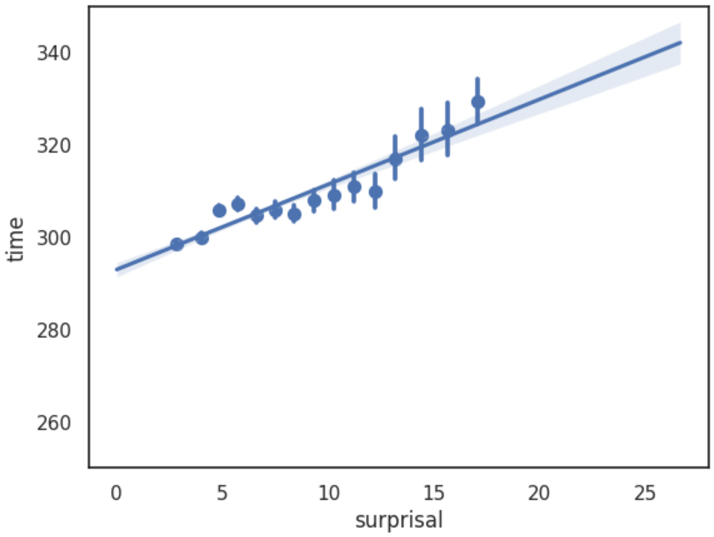


2. **n-gram Model – Without & With Binning**



B. Maze Regression Plots Without & With Binning (return)





C. Open-Ended: Dataset (return)

We utilized two datasets for our analysis. The first dataset was obtained from the Maze project and consists of 480 sentences that were suitable for emotion detection. Additionally, we incorporated data from the article titled "Reading time data for evaluating broad-coverage models of English sentence processing" by Frank et al. (2013) [1]. This article provided us with additional reading time information for sentences, enabling us to enrich our dataset. In total, our dataset comprised 959 sentences, combining both sources.

To preprocess the data, we performed the following steps. Firstly, we obtained the dataset from the Maze project. Secondly, we collected reading time data and general information about the sentences from the article by Frank et al. [1]. To compute surprisal values, we utilized the collected data from both the Maze project and the Frank et al. article. This involved calculating the surprisal values for each word in the dataset. Subsequently, we averaged the reading times and surprisal values for each sentence.

For our analysis, we employed the EmoRoBERTa model, a pre-trained language model designed specifically for emotion detection. The EmoRoBERTa model offers 28 emotion labels, which aligns with our task of analyzing the effect of features on predictions. Reference to the EmoRoBERTa model can be found at https://huggingface.co/arpanghoshal/EmoRoBERTa. By utilizing the emotion labels provided by the EmoRoBERTa model, we obtained the "ground truth" emotions expressed in our text.

D. Model Construction and Training Details (return)

Preprocessing - The emotion labels in our dataset were preprocessed by converting them into a numeric representation. We achieved this by creating a label-to-index mapping, where each unique label was assigned a unique index. This numeric representation enabled us to train our model for emotion classification. Additionally, we prepared the sentences along with their associated emotion labels, reading time, and surprisal values for further analysis.

Model Architecture - For our emotion classification task, we utilized the DistilRoBERTa model as our base model. DistilRoBERTa is a smaller and faster version of the RoBERTa language model, specifically designed for efficient text representation learning. It is trained on a subset of the BERT training dataset and fine-tuned on the Maze Natural Stories dataset.

DistilRoBERTa is a text-only model, which means that it does not take into account other features such as reading time and surprisal. In order to investigate the importance of these features, we built our own models that incorporate these features in addition to text.

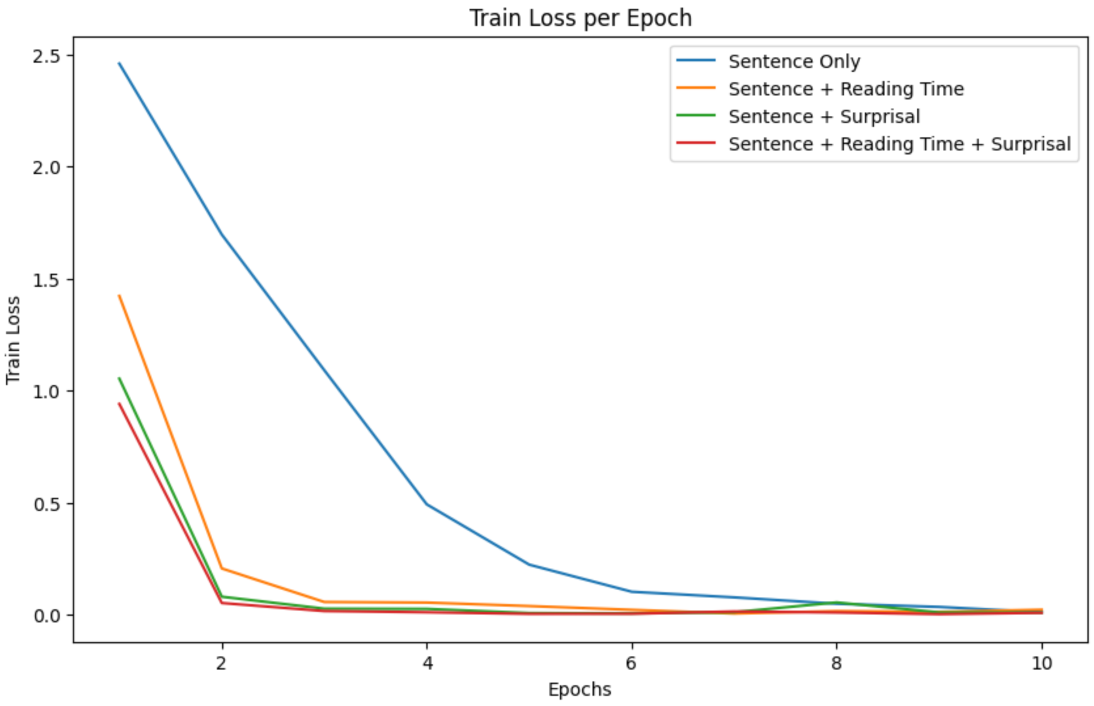
Training and Evaluation - During the training process, we employed the AdamW optimizer and the CrossEntropyLoss criterion. The AdamW optimizer is a variant of the Adam optimizer that incorporates weight decay, making it suitable for fine-tuning pre-trained models. The CrossEntropyLoss criterion is commonly used for multi-class classification tasks.

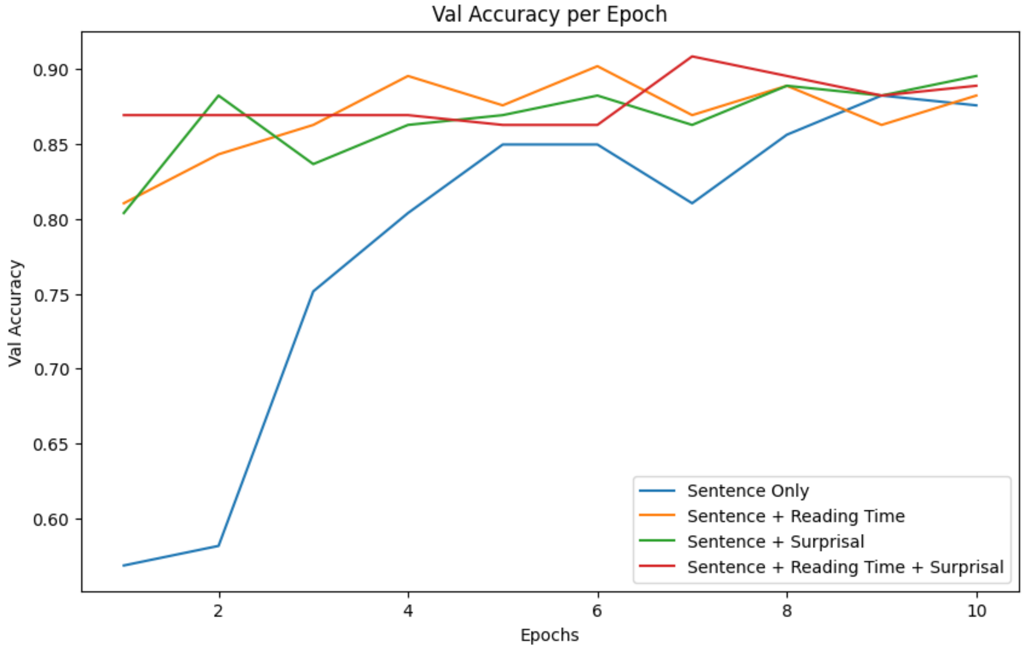
To handle the small dataset size, we trained the model for 10 epochs. Each epoch represents a complete pass through the entire training dataset. The choice of the number of epochs depends on factors such as dataset size, model complexity, and computational resources. In our case, we found that training for 10 epochs yielded satisfactory results.

Considering the limited size of our dataset, we used a relatively small batch size of 4. A smaller batch size helps prevent overfitting and allows the model to generalize better by updating the parameters more frequently. To evaluate the performance of our models, we utilized a validation set and a test set. The validation set was used to monitor the model's performance during training and assist in early stopping to prevent overfitting. We selected the best performing model based on the average accuracy on the validation and test sets. The model's performance was evaluated on the test set, which contains unseen examples, providing an estimate of how well the model generalizes to new data.

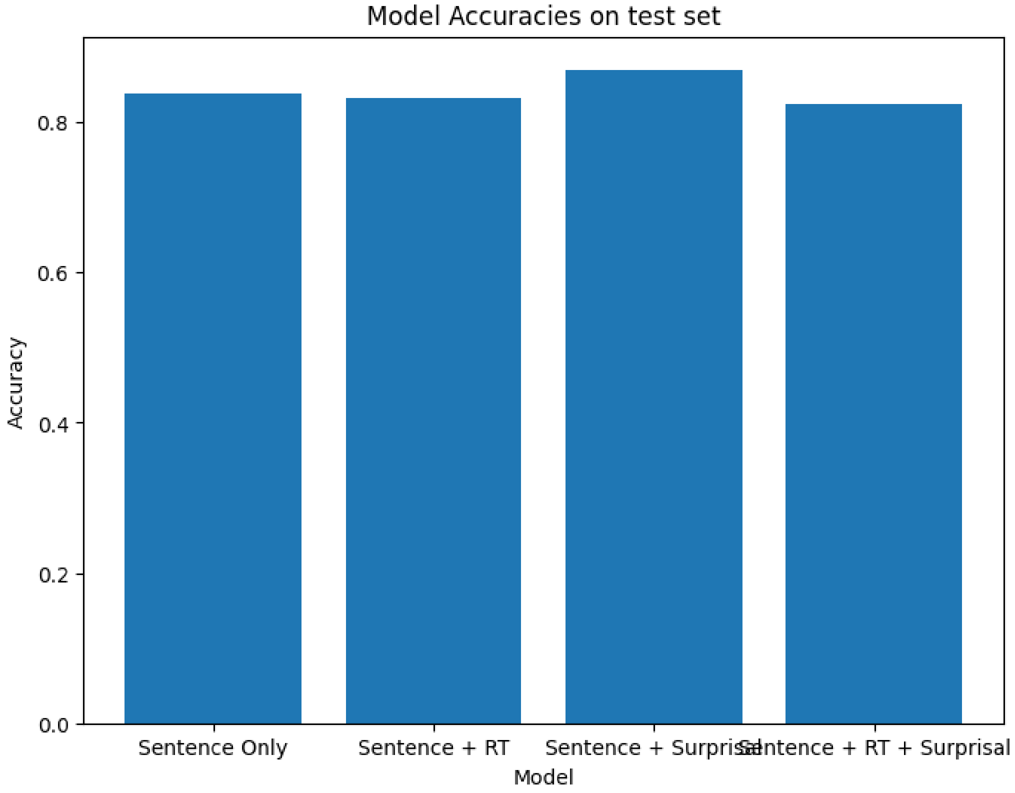
Please note that the source code for our models and training pipeline can be found at the following link: https://colab.research.google.com/drive/1_B8eVOTV3WoxdhPvG5qF63Hh15Pm-icU?usp=sharing. This notebook provides a comprehensive implementation of the described model construction, training, and evaluation processes.

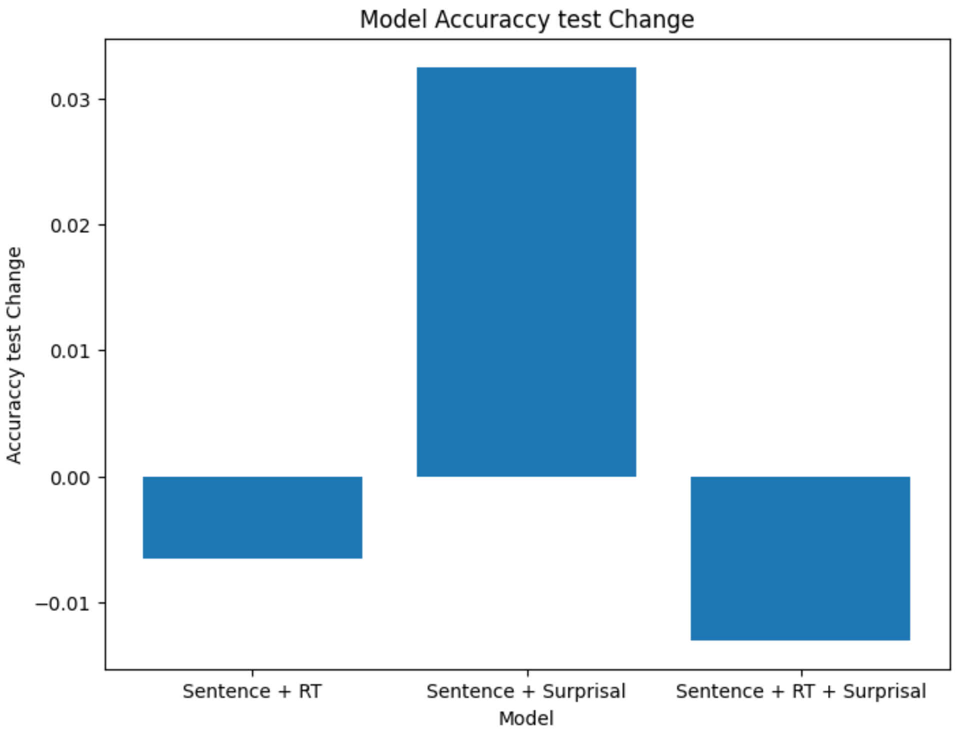
E. **<u>Visual representations of the results (return)</u>**

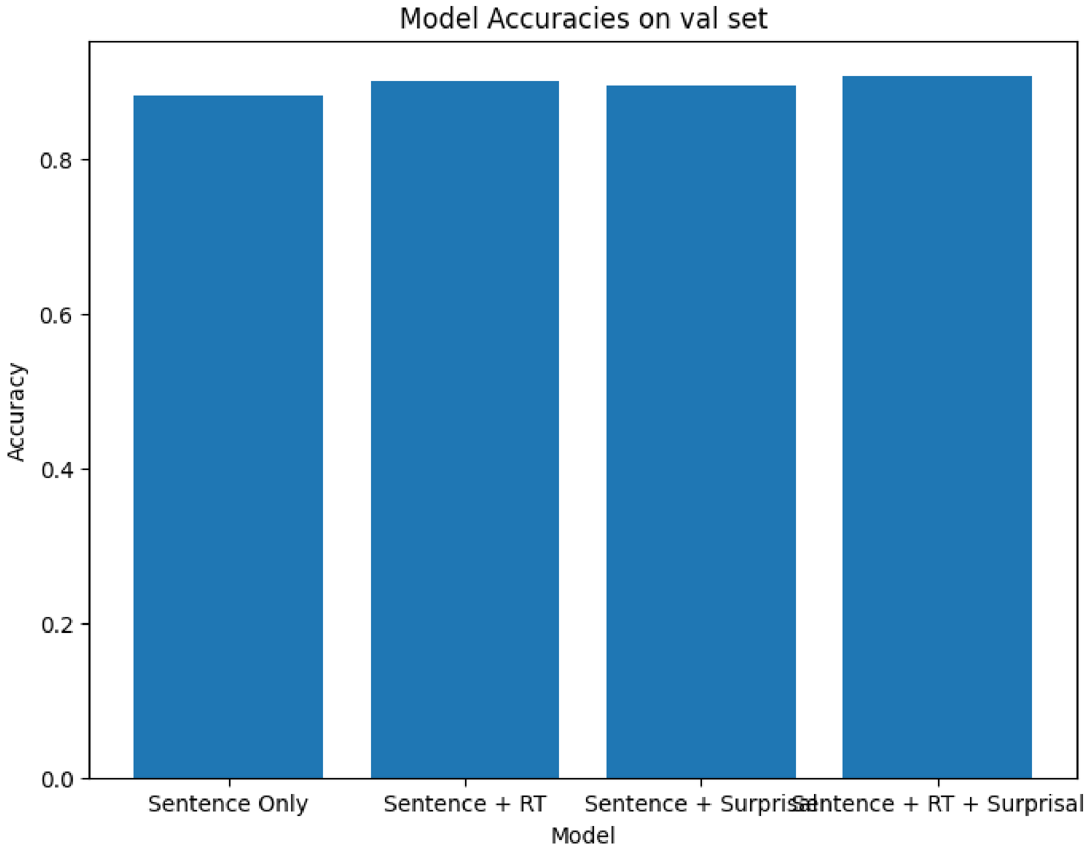


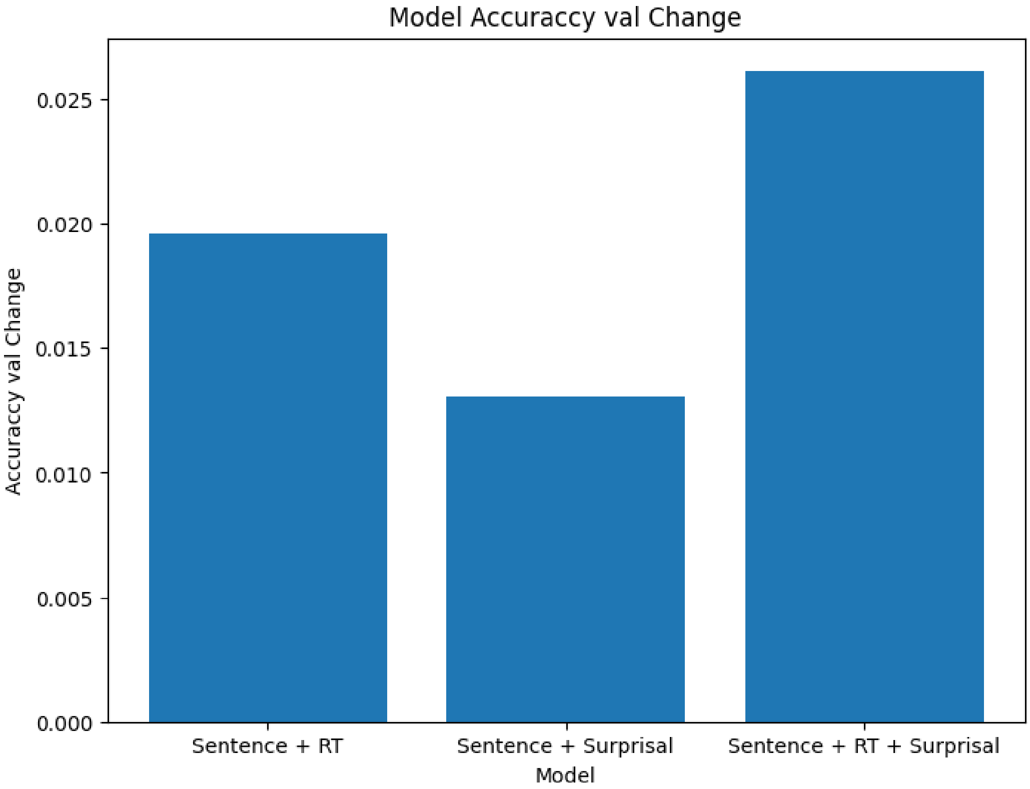


F. **Visual representations of the results (return)**




Model Accuraccy test Change
Accuraccy test Change
0.03
0.02
0.01
0.00
−0.01
Sentence + RT
Sentence + Surprisal
Sentence + RT + Surprisal
Model


Model Accuracies on val set
Accuracy
0.8
0.6
0.4
0.2
0.0
Sentence Only
Sentence + RT
Sentence + Surprisal
Sentence + RT + Surprisal
Model


Model Accuraccy val Change
Accuraccy val Change
0.025
0.020
0.015
0.010
0.005
0.000
Sentence + RT
Sentence + Surprisal
Sentence + RT + Surprisal
Model