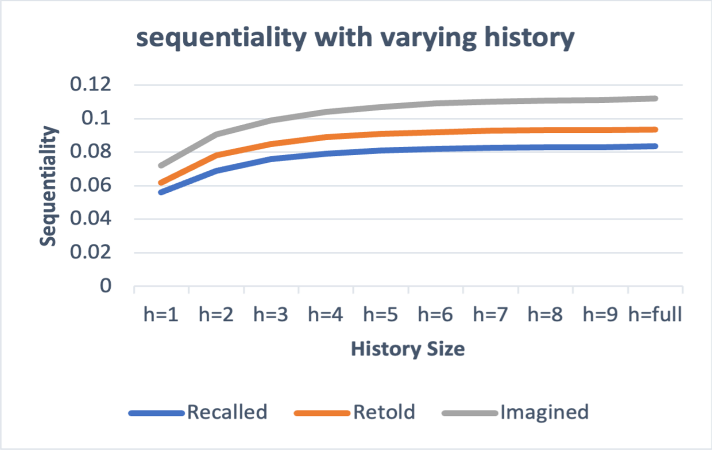
Background and Methods

The paper ”Quantifying the narrative flow of imagined versus autobiographical stories” explores how we can distinguish between different types of story events, namely ”Autobiographical (Recalled), Retold (recollecting the autobiographical stories by the same person using the story summary after a period of 6 months), and Imagined story (A random person describing an imagined story based on the Recalled story summary). It introduces the metric sequentiality, a measure of the narrative flow of events computed using the probability of story sentences given by Large Language models (LLMs). The sequentiality for a sentence within a story is computed by taking the difference of Negative log-likelihood (NLL) of the given sentence conditioned on just the topic (Topic-driven) and topic along with the history or previous sentences (Contextual). The topic and previous sentences would be fed in as the prompt for the LLM, and we then compute the NLL of the target sentence by taking the summation of token-level likelihood. Finally, the mean sequentiality for all the stories for a given type and history size.

The authors have used GPT-3 as the LLMs for computing sequentiality. Currently, GPT-3 is a paid service, and given that a huge number of inferences must be performed, GPT-3 did not seem as a viable option for replicating the results. The authors have mentioned that the sequentiality metric is LLM agnostic, given that the LLM is trained on a considerable amount of data. Hence, for the experiments, I went with GPT-2, which is open-source and available as part of the Transformers library. The experiments were done using the HIPPOCORPUS dataset, which has 6854 stories. In order to speed up the computation, the inferences for different history sizes were parallelized. The source code for the experiment can be found here.

Results

The main objective is to replicate the sequentiality metric for the different story types as in the paper. Similar to the experiments performed within the paper, for each of the story types, namely Imagined, Recalled, and Retold, a full history run was performed, and then the history size was varied from *h=1..9*. The Figure shows the plot of sequentiality for various story types (Recalled, Retold, and Imagined) for different history sizes. The results indicate that for any given history size, the Imagined stories have the highest sequentiality, followed by Retold and Recalled stories. This shows that the Imagined stories have a high narrative flow with respect to the previous context when compared with the retold and recalled. The pattern generated from the experiment using GPT-2 is consistent with the results mentioned in the paper, which was performed using GPT-3. This demonstrates that the sequentiality is not a property of a specific LLM but rather the language structure. The difference in the scale of the results of sequentiality can be attributed to the sentence probability variation between the LLMs used (GPT-2 instead of GPT-3). Along with the sequentiality metric, $\mathrm{NLL_c}$ and $\mathrm{NLL_t}$ for the different stories were also computed. The $\mathrm{NLL_t}$ values were higher for imagined stories than recalled stories, which followed the same trend as the paper.



General Discussion and Future Plans

Based on the experiments performed using GPT-2 on the HIPPOCORPUS dataset, I could replicate the sequentiality trend for the Imagined, Recalled, and Retold story types observed in the paper. Thus, sequentiality can be considered as an effective metric for quantifying how much a story follows the expected or common narrative flow for a specific story topic. As for the next phase, the main objective is to find linguistic features and use them to build a classifier that can distinguish between recalled and imagined events. The paper itself has explored a lot of linguistic features to differentiate the events. Following are some of the features and ideas that I would be exploring to build the classifier.

- Use Lexicon-centric measures such as the proportion of ”realis-events”, story length, and Linguistic Inquiry Word Count (LIWC) as features. The proportion of ”realis-events” is higher for recalled when compared with imagined stories.

- Use event annotations as features. The paper further explores the major, minor, surprising, and expected events within each story and compares how it vary within the story types. The paper shows that the Recalled stories have more minor events than imagined stories.

- Use the sequentiality as a feature, as we have observed that the sequentiality metric is relatively high for imagined stories on compared with Recalled stories.

- Using the story itself as a feature by embedding it using a language model such as BERT. This would be a naive way of building a classifier; it would act as the baseline.

Background and Methods

The objective is to identify novel linguistic features and build a classifier that can effectively distinguish between the **Recalled** and **Imagined** stories in the HIPPOCORPUS dataset with high accuracy. The **Sequentiality** metric, introduced in the paper ”Quantifying the narrative flow of imagined versus autobiographical stories,” is an LLM-based linguistic feature that effectively distinguishes the story types and would act as the base benchmark. The dataset comprises **5535 samples** with **2756 Imagined** and **2779 Recalled stories**. For a fair comparison, **20%** of the whole dataset will be set aside as the **’Test Dataset’** throughout all the classifier experiments. At a high level, the following are the main steps.

1. Feature Extraction and Aggregation
2. Feature selection
3. Top Feature interpretation
4. Traditional/Classical Classifiers Exploration
5. DNN-Based Classifier Exploration (BERT)

The source code for the experiment can be found here.

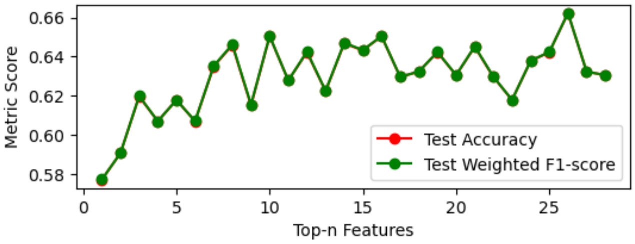
Feature Extraction and Aggregation

The Sequentiality paper provides many linguistic features that could potentially improve the accuracy of our classification task. Following are the features that are being computed for each story.

1. **Sequentiality** - Computed using GPT-2.
2. **Word Count** - Token count computed using nltk.
3. **Sentence Length** - The length of the story.
4. **Realis Score** - This is a proportion of ”realistic” (non-hypothetical) event words within each of the stories. The original paper used BERT-based NER-tagger, trained on a labeled dataset. Due to tagged data unavailability, this was calculated by using the ”spacy” library to identify the tense of the words, effectively computing an approximation of the Realis Score.
5. **Concreteness Score** - This is a proportion of the concrete words within each story and is computed based on the scores calculated based on the paper ”Concreteness ratings for 40 thousand generally known English word lemmas”.
6. **LIWC (Linguistic Inquiry and Word Count) Features** - LIWC is a linguistic analysis software that provides extensive attributes for textual data. The Sequentially paper provides the subset of LIWC attributes that are significant in distinguishing between ”Recalled” and ”Imagined Stories”, and only those attributes will be explored. Each of the following metrics (except for aggregation metrics) will use an internal dictionary that contains all the words associated with the corresponding metric.
   (a) **Summary metrics** - WordsPersentence, BigWords (Words longer than 6 letters), Analytic, Clout, Authentic, Tone
   (b) **Linguistic metrics** - Function, article, number, preposition, conjunction, negations
   (c) **Cognitive metrics** - CognitiveProcess, Insight, Discrepancy, Tentative, certitude, Differentiation
   (d) **Perception metrics** - Motion, Space
   (e) **Time Orientation metrics** - Time, Past focus, Present focus

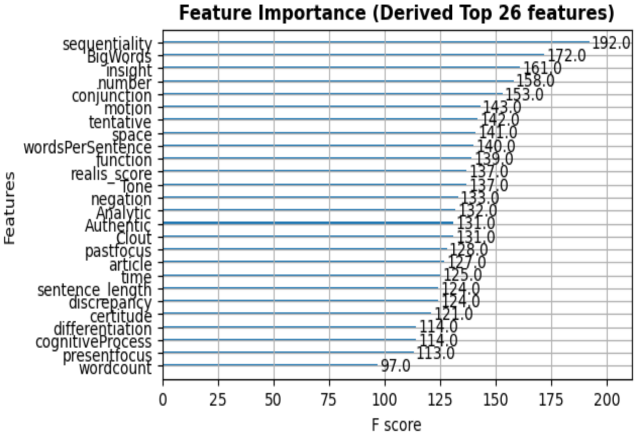
Thus we have 28 features to explore

Feature Selection



Modern classifiers can perform feature selection as part of training, but the training time and complexity would be significantly higher. In scenarios with large data points, pruning out the less important features will be computationally helpful. By leveraging the **Recursive feature selection (RFS)** technique as part of scikit-learn with the underlying classifier as XGBoostClassifier, we compare how performance metrics on the Test dataset vary with the top-n features selected. The figure shows how the **Test Accuracy** and **Test Weighted F1-score** change against the top features selected from n = 1 to 28. We can see that the Test Accuracy changes from **57.7%** when n = 1 to the maximum of **66.2%** when n = 26 features. The feature chosen by RFS when n=1 is **Sequentiality**, thus reinforcing the idea of it being a significant feature. **Thus, all the further experiments will use the selected combination of 26 features.**

Top Feature Analysis



The above figure illustrates the top 26 features, selected based on maximum test accuracy and arranged in descending order of importance, with sequentiality being the most important.

Features	**Recalled Stories**		**Imagined Stories**	
	Mean	SD	Mean	SD
Sequentiality	0.087	0.079	0.117	0.092
Bigwords	14.282	3.563	13.432	3.593
Insight	2.316	1.362	2.669	1.495
Number	1.458	1.070	1.189	0.893
Conjunction	7.307	1.760	6.849	1.909

The figure displays the top 5 features, along with their corresponding mean and standard deviation, categorized by Imagined and Recalled stories. The subsequent discussion provides an interpretation of the the scores.

- **Sequentiality** - Imagined stories exhibit more narrative flow than the Recalled stories, hence having a higher sequentiality mean.
- **Bigwords (Words longer than six letters)** - Recalled stories have more Bigwords than Imagined as they have more intricate details and thus require a more complex vocabulary to express them. Conversely, when imagining a story, individuals may opt for simpler and shorter words to convey the fundamental ideas of the summarized text.
- **Insight (Words such as think, know, consider)** - Imagined has higher insight-related words than recalled, which implies that people use these words to express hypothetical scenarios, which goes along with the creative nature of imagined stories.
- **Number (Words like one, thirty, forty)** - Recalled stories have a higher prevalence of number-related words ,and this can be attributed to recalled stories having more intricate details of an event, prompting people to provide more quantitative specifics such as count, time, dates, and monetary aspects associated with the recounted event.
- **Conjunctions** - The increased use of conjunctions in recalled stories indicates a more detailed and expressive narrative style characteristic of recalled stories.

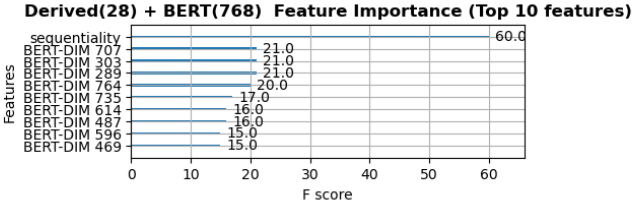
Traditional Classifier Methods

We employed various traditional classifiers, including XGBoost, Logistic Regression, and SVC, and utilized Grid Search for tuning hyperparameter such as learning rate, max depth, and regularization strength. The models were trained on 4428 samples, and their performance was evaluated on the Test Dataset of 1107 samples. Apart from the derived features, we converted the raw story text into embeddings using TF-IDF and Doc2Vec and used them as features. **In the Results figure, among the traditional classifiers, XGBoost, leveraging the derived 26 features, outperformed others with a Test accuracy of 66.2%**. This marks a notable 9% improvement over the baseline (57.7%) which used Sequentiality feature. Despite incorporating TF-IDF and Doc2Vec embeddings from the raw story text, no performance improvement was observed. This suggests that the additional information captured by these embeddings did not significantly contribute to the model's predictive accuracy in this context.

Classifier	**Feature**	**Accuracy**	**Weighted F1-score**
XGBoost	Sequentiality	0.577	0.583
XGBoost	**26 Features**	**0.662**	**0.653**
SVC	26 Features	0.654	0.656
Logistic Regression	26 Features	0.627	0.631
XGBoost	26 Features + TFIDF	0.649	0.649
XGBoost	26 Features + 100 DIM Doc2Vec	0.644	0.639
Fine-Tuned BERT	**Fine-Tuned BERT Embeddings**	**0.778**	**0.777**

Deep Learning Approaches

Transformer models are currently the state-of-the-art models for tasks associated with textual data. We added a linear classifier layer to a pre-trained BERT model and trained the BERTClassifier using the story text as input while fine-tuning the BERT weights. **The BERTClassifier achieved a Test accuracy of 77.7%, which is a substantial 11% improvement over derived features trained on traditional classifiers.** This raises the question:**Could BERT be uncovering and leveraging more effective features than those we derived earlier through traditional methods?**



To interpret the BERT performance, we extracted the 768-dimension embeddings of the fine-tuned BERT for the data samples and performed a feature importance analysis, integrating the earlier derived features, thus using 796 features **(768 BERT features + 28 derived features)**. The above figure shows the feature importance scores of the top 10 features out of 796. **We could see that Sequentiality still remained the most important feature, followed by the BERT dimension features, which outperformed our derived features.** Thus, BERT was able to generate more impactful features providing a plausible explanation for the significant jump in accuracy achieved.

General Discussion

Thus, we were able to derive significant features, which are evident from improving the classifier accuracy from the Sequentiality baseline of 57.7% to 66.2%. Using BERTClassifier, with the embeddings as the features, the accuracy further bumped up to 77.7 %, but at the expense of interpretability. Then, we performed a small study to compare the derived features against the BERT features. Sequentiality remained the most significant feature, followed by BERT features outperforming our derived features. As part of future work, we could explore the embeddings and understand what linguistic features the BERT model extracted. One method could be to understand the **”Attribution Scores”** associated with each input token and finding a pattern among the top tokens responsible for prediction of Imagined/Recalled stories.

References

Booth, J. W. P. E. F. J. (1999). *Linguistic inquiry and word count (liwc)*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). *Language models are few-shot learners.*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*.

Kuperman, M. B. . A. B. W. . V. (2013). *Concreteness ratings for 40 thousand generally known english word lemmas*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*.

Sap, M., & Jafarpour, A. (2022). Quantifying the narrative flow of imagined versus autobiographical stories..