Exploring Language in Different Daily Time Segments Through Text Prediction and Language Modeling

Kennedy Roland
St. Francis Xavier University
Antigonish, NS, Canada
x2022gds@stfx.ca

Milton King
St. Francis Xavier University
Antigonish, NS, Canada
mking@stfx.ca

Abstract

Temporal-aware language models have proved to be effective over longer time periods as language and its use changes, but little research has looked at how language use can change at different times of the day. We hypothesize that a person's usage of language varies at different times of day. We explore this concept by evaluating if models for language modeling and next word prediction improve their performance when considering the time of day. Specifically, we explore personalized temporal-aware models for next-word prediction and language modeling and compare them against baseline models, including non-temporal-aware personalized models. Specifically, our proposed model considers which of the 8, 3-hr daily time segments that a text snippet was written during for a given author. We found that our temporal-aware models tend to outperform temporal-agnostic models with respect to accuracy and perplexity.

1 Introduction

Language models are often trained on large amounts of text from many different people but do not necessarily consider the time that the text was written. In this work, we tailor a general pre-trained language model (GPT2) (Radford et al., 2019) toward a single person and the time of day that their text was written. The intuition behind this is that we believe that the same individual's use of a language varies throughout different times of the day and we can use models to explore that. For example, a person's text snippets in morning could differ from their text snippets in the evening. As far as we have found, there has been no prior work considering the use of language at different times of day. Improving the performance of a language model through considering the time of day that the text was written, could potentially assist authorship attribution models (Fabien et al., 2020), although additional experiments are required to validate this.

Classifiers have been found to have improved performance when tested on the same time period they were trained on compared to other intervals, both annually and seasonally (Huang and Paul, 2018). Considering the month a post was made in has been shown to have an impact on document classification using a pre-trained BERT model when looking at Reddit posts to classify which political subreddit they belonged to (Röttger and Pierrehumbert, 2021). Models trained on an even shorter time frame, like day of the week, have also been seen to outperform temporal-agnostic models for tasks like word sense disambiguation (Wei and King, 2024). We now aim to look at the potential of time impacting next-word prediction and language modeling in an even shorter duration. Many applications that support sharing text often include timestamps with text and therefore it is reasonable to consider this type of data in real scenarios. Due to the size of our dataset, we had segmented time of day into 8, 3-hour segments, but acknowledge that the size of the time segments could be a hyperparameter that could be tuned to select the best performing time range size.

2 Related Work

Although there is a lot of previous work looking at language models with a temporal context, relatively low amounts of research have focused on authors speaking differently at various times in the context of improving next-word prediction and language modeling. Some research showed that there is a benefit in using temporal data for various other tasks including document classification (Huang and Paul, 2018; Röttger and Pierrehumbert, 2021). It's suggested that classifiers perform better when they are applied to the time period they were

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 120–125, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-011

trained on, both on a seasonal level and by year (Huang and Paul, 2018). They recommend training a classifier from the most current chronological samples instead of randomly in order to get the best performance (Huang and Paul, 2018). Word sense disambiguation models — models that are tasked with assigning a sense to a word in context — have also been shown to perform better when tailored toward temporal segments (Wei and King, 2024).

Rosin et al. (2021) proposed tempoBERT, a model that considers the context of time and used that alongside prepended text indicating the year that it was written to initially help with time based facts; they also explored semantic change and sentence time prediction. They found that by using smaller language models, they were able to produce state-of-the-art results, outperforming the larger models (Rosin et al., 2021). Rosin and Radinsky (2022) worked on temporal attention by creating a matrix of the input and embedding the time vector into that, allowing for the language model to consider time without changing the input at all, so there isn't a need to prepend text based on what year it is. King and Cook (2020) applied a similar technique, which they refer to as priming, where they input tokens from text from the author to an LSTM-based language models before being evaluated on some text for testing. They evaluated their models using adjusted perplexity (a variation of perplexity), accuracy@k, and accuracy@k given the first c number of characters from the target token.

Another contribution that involves temporal information is tempLAMA, a dataset which contains queries with time-sensitive answers, making it a good tool to test temporal language models (Dhingra et al., 2022). They used tempLAMA to train language models and found similar results, where language models that consider time performed better than time-agnostic models that were trained on more text and temporal-aware models have consistently performed better on these time-sensitive questions compared to time-agnostic models (Dhingra et al., 2022).

There has been work and discussion around keeping human nature in the field of natural language processing. Hovy and Yang (2021) suggest that there are many things that change the way the same person will write text such as who the receiver of the message is, what the event/occasion is or, based on the topic. They suggest that we need to get closer to social understanding of humans to better language prediction (Hovy and Yang, 2021). We agree with the idea that social modeling can benefit with the addition of more context and our work focuses on exploring if time of day has any impact on this.

Soni et al. (2022) expanded on the idea that we need human social understanding and had a similar idea to our hypothesis in the sense that humans will use language differently in various situations. They segmented things by the human state and were able to make the model aware of what state of being the individual was in, personifying a bit of the machine part of natural language processing (Soni et al., 2022). In our case, the different states can be comparable to what time of day the author is writing during.

3 Dataset

We randomly selected 50 subreddits[1] from a list of popular communities[2]. We then extracted the 25 most recent posts from each of those subreddits. We reselected a new subreddit if the originally selected subreddit primarily consisted of images. We collected all public comments and posts from the users that were authors of each selected post from the original list of subreddits, which includes comments and posts from other subreddits. The intent of gathering text from different subreddits by the same user was to increase topic diversity for a given user. We removed any authors that had less than 25 posts or comments total (across all subreddits, selected originally and otherwise) from the dataset. We converted the times of all posts and comments from UTC into our local time (ADT). This was done because there was no associated location and time zone info publicly available and it is important to mention that we do not have access to the actual time of day for the timezone of the author when the text was posted. Therefore, our analysis and experiments will be considering the time relative to each individual author and we cannot compare the same time segment across authors. We then separated all posts and comments from each author into time segments (3-hour periods beginning at 12AM ADT). This resulted in 8 different time segments within a day. Authors that had not posted across a minimum of two time segments were removed.

[1] Subreddits are topic-specific forums or sub-sites where users can post and comment on the social media platform, Reddit.

[2] https://www.reddit.com/best/communities/1/

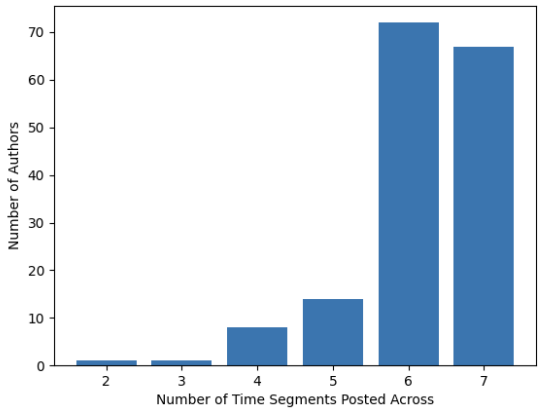

Figure 1: The distribution of the number of time segments the authors had posted across.

This resulted in 163 different authors and 1,008 author/time segment combinations across those authors — up to 8 time segments per author. The number of time segments per author can be seen in Figure 1). Most authors had posted across 6 or 7 time segments, with no authors posting across all eight segments. There are very few authors that only posted in 2 or 3 time segments. There are 2.34 posts and 101.01 comments per author on average, with an average length of 36.56 words. All text included in this dataset is in English. Text that was detected to be in multiple languages were removed from the dataset before ensuring they met the minimum post and time segment requirements.

4 Experimental Setup

The tasks that we focus on are language modeling and next-word prediction. The same models will be applied to both tasks. The models we used were not trained specifically on temporal-aware data. Each model varies by selecting different types of text to prepend the text to be used for testing (or not prepending any text in one case seen in Section 4.2). From hereon, we refer to every piece of text the author has written as a post, regardless of whether it was a post or comment.

4.1 Models

Due to the lack of location-based information on posts, and therefore timezone information, each model is tailored toward a time segment for a given author, instead of similar time segments across authors. This is to capture that the timestamp for the text is relative to the author's time of day. For example, morning in one location could be night in another location, but the existing information with the text does not provide enough location-based context to determine the region-specific time of day. Therefore, we can't have a model tailored towards text from a specific time segment from multiple authors. We used GPT2 (Radford et al., 2019) with a 1024 token limit from Huggingface [3] as our general pretrained language model and its corresponding tokenizer.

Each of the following models use GPT2 and we discuss the variations of the models in the following subsections. Each model will be given text to evaluate their language modeling capabilities and next-word prediction.

4.2 *Nothing*

In this model, we use GPT2 without any modifications. Specifically, we pass each post on its own to GPT2. We refer to this model as *Nothing* in the later sections.

4.3 *Random*

In this model, each given post is prepended with text selected randomly from any of the other authors in our dataset and any relative time segment. We refer to this model as *Random* in the later sections. For example, if the test text was from author A in time segment 7, then the *Random* text prepended to that could be from any time segment (1-8) and any author that isn't author A.

4.4 *Author*

In this model, each given post is prepended with text from the same author and different time segments to the post that was being evaluated. For example if the test text is from author A in time segment 7, then the *Author* text prepended to it would be from author A and any available time segment other than 7. We refer to this model as *Author* in the later sections.

4.5 *Temporal*

In this model, each post is prepended with text from the same author and the same time segment. The text can be from different days, but it is the same daily time segment relative to the given author. For example, if the test text is from author A and time segment 7, *Temporal* text prepended to it would include all other posts from time segment 7 for

[3] https://huggingface.co/docs/transformers/en/model_doc/gpt2#openai-gpt2

author A. We refer to this model as *Temporal* in the later sections.

We chose these experimental setups to explore the influence of time in the models. For example, if someone were to write in the same manner regardless of the time of day, we would expect the improvement in performance to be similar between *Author* and *Temporal*. However, if *Temporal* outperforms *Author*, then we would expect that it is due to same person changing their manner of writing throughout the day. *Random* was chosen so that we could test if simply adding more prepended text could help as much as the personalized models. Since this dataset was limited and token length can vary within time segments of an author, we recorded the length of text used for the *Temporal* model and limited the amount of text for *Author* and *Random* to have the same token length so that results would not be affected by one model simply having more text to benefit from.

5 Results

In this section, we discuss our results from the experiment.

5.1 Evaluation Metrics

The performance of Each of the models were measured in terms of perplexity per post and accuracy at k, with k ranging from 1 through 5. The Perplexity score per post was averaged to be one value for each author/time segment combination. The accuracy at k for every post was summed and divided by the sum of all tokens in a time segment to be the percentage of tokens accurately predicted in that time segment. These values were then compared across all four of the models for each time segment.

For every time segment for every author, each post had a perplexity value and accuracy@k — the number of tokens accurately predicted within the top k predictions – for each value 1-5. Accuracy@k lends itself as a more extrinsic evaluation and resembles the desired performance on systems with next-word prediction, such as messaging applications. For example, the application recommends k words as a prediction when writing text.

Both the values for perplexity and accuracy@k from our models were then compared against all our other models to see how frequently one model produced a better value (lower perplexity or greater accuracy@k) than the model it was being compared against.

5.2 Evaluation Results

Table 1 shows each model's weighted average perplexity and accuracy@k. The weighted average incorporates the number of tokens in each post. The weighting was done so that the impact of perplexity scores was relative to the length of a post, for example, a short post could dominate the overall score if it performed exceptionally poorly on it. *Temporal* consistently outperforms all other considered models on both perplexity and accuracy@k. Although showing relatively large improvements over *Author* for perplexity, the difference with respect to accuracy@k is marginal. *Author* consistently outperformed *Nothing* and *Random*, which shows that personalizing models is beneficial.

To directly compare between models, we calculate the percentage of instances that one model outperforms another model in Table 2. The values in the table represent the percentage of all instances that one model (row) beat the other model (column) in regards to average perplexity. This further shows that the temporal-aware model, *Temporal* more often outperforms the other temporal-agnostic models with respect to perplexity.

Similarly, Table 3 shows the percentage of time segments that one model outperformed another model on next-word prediction with respect to accuracy@1. Table 4 and Table 5 also show next-word prediction results for a model comparison at k being equal to 3 and 5, respectively. The values in these table show that *Temporal* is more likely to outperform the other temporal-agnostic models than be outperformed by them. Interestingly, *Author* outperforms *Nothing* more often than *Temporal* outperforms *Nothing*.

Random outperforming *Nothing* on most instances shows that more text typically does help a model improve accuracy in next-word prediction and lowers perplexity. However, seeing that both *Temporal* and *Author* beat *Random* a majority of the time supports the idea that performance can be improved with human context, simply giving the model more text doesn't have as significant of an impact as targeting that text to represent the human who it is testing on. Author-specific text is better than random posts, but if you can get time segment info, that does tend to perform better, supporting our hypothesis.

Our results show that the type of text provided to the model can influence the performance, which is demonstrated by *Author* and *Temporal* outperform-

	Nothing	Random	Author	Temporal
Perplexity	198.21	180.61	160.31	**129.14**
% Acc@1	0.25	0.27	0.29	**0.30**
% Acc@2	0.34	0.36	0.38	**0.39**
% Acc@3	0.39	0.41	0.44	**0.45**
% Acc@4	0.43	0.45	0.47	**0.48**
% Acc@5	0.45	0.48	0.50	**0.51**

Table 1: The overall weighted average perplexity (lower is better) and accuracy@k (higher is better).

	Nothing	Random	Author	Temporal
Nothing		17.76	8.23	7.54
Random	82.24		13.59	12.4
Author	91.77	82.14		37.8
Temporal	92.46	83.13	57.94	

Table 2: The percentage of time that one model (row) outperformed the other (column) with regards to perplexity.

ing *Random*. This finding supports existing work regarding personalized models. Lastly, the manner in which a person uses a language during different times of day is potentially captured by our model and presented as *Temporal* outperforming *Author*.

6 Conclusions

At the root of human nature is communication, and much of that exists in the context that language is used. In this work, we focused on the time segment that the text was written in for our context. We examined if the usage of language for an individual author changes at different times of day. To explore this phenomena, we compared the performance of different models with respect to language modeling (perplexity) and next-word prediction (accuracy@k). Our *Temporal* model outperformed the other temporal-agnostic models on both perplexity and accuracy@k. This demonstrates the difference in the usage of language for an individual author at different times of day (time segments). Furthermore, in a direct comparison, *Temporal* outperformed the other temporal-agnostic models more than half the time for all accuracy@k and perplexity on each segment. Our results also reinforce the benefit of personalized models, since both the personalized temporal-aware and temporal-agnostic models outperform both non-personalized models. An additional benefit to the use of prepending text for our models is that it is not relatively computationally expensive and it does not require high-end GPUs as our experiments were conducted primarily on a computer with modest hardware.

Unfortunately, the dataset could not access the user's location data, and therefore we could not compare time segments across authors as we are unable to know what time it is relevant to their timezone, which could allow us to explore if people use language differently at different times of day regardless of the specific author. For example, exploring if people generally use language differently in their morning compared to their evening. This would be a reasonable and interesting direction for future work.

	Nothing	Random	Author	Temporal
Nothing		25.50	5.26	5.85
Random	74.11		8.73	9.03
Author	94.44	82.34		38.19
Temporal	93.85	83.33	51.49	

Table 3: The percentage of time that one model (row) outperformed the other (column) with regards to accuracy@1.

	Nothing	Random	Author	Temporal
Nothing		14.88	4.27	4.76
Random	84.92		9.62	6.75
Author	95.44	83.13		34.42
Temporal	95.04	85.62	55.06	

Table 4: The percentage of time that one model (row) outperformed the other (column) with regards to accuracy@3.

	Nothing	Random	Author	Temporal
Nothing		12.3	3.97	4.56
Random	87.5		8.33	6.25
Author	95.83	83.83		33.43
Temporal	95.24	86.21	55.95	

Table 5: The percentage of time that one model (row) outperformed the other (column) with regards to accuracy@5.

7 Ethical Considerations

All text was acquired through Reddit's public API and anything posted on private subreddits was not included in this research.

8 Acknowledgements

This research was supported by the Alley Heaps Undergraduate Research Internship through St. Francis Xavier University.

References

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.

Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2461–2469, Marseille, France. European Language Resources Association.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2021. Time masking for temporal language models. *CoRR*, abs/2110.06366.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Yuchen Wei and Milton King. 2024. Sense of the day: Short timeframe temporal-aware word sense disambiguation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14676–14686, Torino, Italia. ELRA and ICCL.