Quantifying Temporal Change in Sentiment Analysis: A Study on the Amazon Reviews'23 Dataset

Anonymous ACL submission

Abstract

This paper investigates the impact of the temporality of data on the performance and generalization of sentiment analysis models trained on review data from different time periods. Using a large dataset of movie and TV show reviews labeled as positive or negative, we partition the data into 3 distinct scales: individual months, individual years and multi-year groups. We fine-tune separate DistilBERT models for each time period using a binary sentiment classification task and evaluate their performance through cross-period testing, examining how temporal gaps between training and testing data affect model accuracy. Performance is measured using accuracy. Our experiment aims to pinpoint the specific duration of the gap that leads to noticeable performance degradation. Our results show how model performance degrades when applied to out-of-period data, highlighting the evolving nature of sentiment language and the importance of temporal alignment in training and testing data for robust sentiment analysis models. This study offers a more nuanced understanding of how linguistic and contextual changes impact model generalizability over time.

1 Introduction

The internet is the main source of data for many NLP tasks, reflecting how humans communicate. This makes it important to study how language changes over time and how those changes impact NLP systems. One study shows that language on social media, evolves very quickly. (Eisenstein, 2013) They reasoned that out-of-vocabulary (OOV) phrases increase over time, with noticeable changes even within a single day. Monthly and yearly changes are even more significant. This can be due to changes in the style of communication, the use of new words and abbreviations, and a continuous rise and fall in trends.

Sentiment analysis is a fundamental task in NLP that determines the emotional tone or polarity of textual data. Applications involving social media monitoring, business analytics, and customer feedback systems all require sentiment analysis, making it one of the most critical technologies to understand user opinions and behavior. However, languages are dynamic-words, phrases, and the sentiment they convey evolve over time. As societal norms, cultural references, and trends shift, models trained on older data may struggle to capture sentiment accurately in contemporary text. This raises an important question: How large of a temporal gap between training and test data leads to noticeable degradation in performance, and does fine-tuning a sentiment analysis model on more recent data improve its ability to handle contemporary text compared to using older data?

We hypothesize that there is a critical temporal window between the training and test data, beyond which performance significantly degrades. By quantifying this gap, we aim to determine how the temporal misalignment between training and test data impacts sentiment classification performance. In this study, we investigate this hypothesis by examining the effect of temporal alignment on sentiment analysis performance. Using a large dataset of movie and TV show reviews spanning from 1997 to 2023,3 distinct scales: individual months, individual years and multi-year groups. For our experiments, we use DistilBERT, a lightweight and efficient transformer-based model derived from BERT. DistilBERT is pre-trained on extensive corpora and serves as the baseline for sentiment classification tasks. We fine-tune DistilBERT on data from each time period and evaluate its performance both within the corresponding period (in-period) and across other periods (out-of-period). By analyzing accuracy scores, we aim to quantify how temporal misalignment impacts performance and to explore whether fine-tuning on more temporally relevant data improves robustness and accuracy.

This work contributes to the understanding of

temporal adaptation in pre-trained language models, providing insights into the critical time period gap where linguistic changes—such as shifting connotations and evolving trends—begin to affect NLP model performance. The findings are particularly relevant for real-world applications where maintaining alignment with evolving language is essential for ensuring the accuracy and reliability of sentiment analysis models.

2 Related Work

The widespread adoption of pre-trained language models like BERT adds further nuance to this question. BERT and its variants are trained in two phases: an upstream pre-training phase using large-scale unlabelled corpora to learn general-purpose language representations, followed by a downstream fine-tuning phase on task-specific labelled data. While the choice of fine-tuning data has a significant impact on performance, prior work has demonstrated that the domain of pre-training data is equally critical. For instance, additional pre-training on domain-specific data has been shown to improve performance on downstream tasks, a process known as domain adaptation (Devlin et al., 2019).

Recent work has demonstrated that document classification performance often degrades when applied to data from different temporal contexts, on a variety of tasks. One study looked at the performance variation between six different corpora, including music reviews, hotel reviews, political platforms, and news articles, segmented into different yearly and seasonal intervals (Huang and Paul, 2018). By training classifiers on these various temporal segments and testing their performance, they discovered significant degradation when models were applied to out-of-period and out-of-season data.

(Lukes and Søgaard, 2018) investigated the impact of temporal shifts on sentiment analysis, showing that changes in lexical polarity (meaning sentiment or tone) of words over time lead to significant model degradation. Their study found that models trained on older product review datasets performed worse on more recent test sets, primarily due to shifts in word polarity rather than just reduced vocabulary overlap. Words with initially positive connotations could shift to neutral or negative meanings, a phenomenon known as semantic amelioration or pejoration. To address this, they introduced predictive feature selection techniques to counteract temporal polarity shifts, improving sentiment classifier accuracy by selecting features with consistent or predictable polarity changes. Their work underscores the importance of temporal adaptability in NLP systems to maintain reliability over time.

In contrast to previous studies that primarily focused on the effect of general temporal shifts in language, our experiment aims to pinpoint the specific time period gap between training and test data that leads to noticeable performance degradation. While past research has demonstrated that language evolves quickly, especially in terms of vocabulary and sentiment, our approach seeks to quantify the exact duration at which the performance of NLP models begins to degrade. By analyzing performance across various temporal gaps, we aim to provide a more nuanced understanding of how far back in time models can be trained before significant drops in accuracy are observed. This will allow us to pinpoint the critical window where temporal changes in language, such as shifting connotations and evolving trends, impact NLP tasks in a measurable way.

3 Method

3.1 Dataset

The dataset used in this study consists of 17,328,314 reviews spanning 26 years from 1997 to 2023. Each entry is structured with three key parts: the *text* of the review, a binary *label* indicating sentiment (1.0 for positive sentiment and 0.0 for negative sentiment), and the date the review was made. The distribution of reviews over time is uneven, with relatively few entries available in the early years (1997–2003), and starting in 2004, the volume of reviews increases significantly. To address this imbalance, we trimmed the data within a particular period to match the lowest number of entries among the periods being tested in each experiment. Additionally, we balanced the dataset to ensure an equal proportion (50%) of positive and negative reviews.

To label the sentiment of reviews, we categorized reviews rated 1–2 stars as negative (0.0) and reviews rated 4–5 stars as positive (1.0). Reviews rated 3 stars were excluded from the dataset and treated as neutral. We chose to exclude 3-star reviews because a 3-star rating often represents mixed or neutral sentiment, making it challeng-

ing to confidently categorize as either positive or negative without additional context. Furthermore, while many prior studies treat 3-star reviews as negative, this approach can introduce bias since such reviews frequently convey both praise and criticism. By excluding 3-star reviews, we ensure a clearer boundary between positive and negative sentiment classes, thereby improving the reliability of our dataset for downstream sentiment analysis tasks.

Figure 1: Example Sentiment Dataset Entries

Review Text	Label	Date
Amazon, please buy the show! I'm hooked!	1.0	2015-08-24
My Kiddos LOVE this show!!	1.0	2016-04-19
I think I could have written a better script.	0.0	2021-05-19

The comprehensive size and temporal span of the dataset make it well-suited for analyzing the impact of training models on time-aligned data.

3.2 Data Processing and Usage

To analyze the temporal impact of data on model performance, the dataset was divided into 3 different ways: individual months, individual years and 4-year groups.

Table 1: 4-Year Time Periods

Time Period	Entries
2004–2007	404,442
2008–2011	623,560
2012–2015	5,934,316
2016–2019	6,617,353
2020–2023	2,273,827

Table 2: Years

Years	Entries
2021	757884
2022	586695
2023	194246

For each time period, the data was split into training and testing sets. To address class imbalance, label balancing was applied by undersampling the majority class to ensure an equal number of positive (1.0) and negative (0.0) labels in both splits.

Table 3: Monthly Period for 2022

Month	Entries	Month	Entries
Jan	58079	July	53353
Feb	47806	Aug	54462
Mar	47424	Sep	57164
Apr	43466	Oct	47245
May	43533	Nov	40946
June	42774	Dec	50443

Table 4: 4-Year Dataset Shapes After Balancing

Time Period	Train Size	Test Size
2004–2007	99,106	24,464
2008–2011	152,662	37,882
2012–2015	993,950	249,328
2016–2019	1,377,136	345,168
2020–2023	808,838	202,000

Table 4 summarizes the size of each period after balancing. As an example, for the 2004–2007 period, the train set consisted of 49,553 instances of each label, while the test set included 12,232 instances of each label.

To ensure consistency and comparability across all time periods, uniform sampling was applied to standardize the size of training and testing sets. Each training set was sampled to contain 99,106 instances, while each testing set was sampled to 24,464 instances. This uniformity ensures that variations in model performance are not influenced by differing dataset sizes across time periods.

3.3 Model Implementation

For sentiment classification, DistilBERT, a transformer-based model, was selected for its lightweight architecture compared to the original BERT model. DistilBERT is pretrained on large text corpora and fine-tuned for binary sentiment classification tasks in this study. The transformer model was used to execute our experiment as follows:

- **Tokenizer:** The tokenizer from the `transformers` library was used to preprocess text inputs, ensuring consistent formatting.
- **Fine-tuning:** The model was fine-tuned separately on the training data for each time period, using a binary classification head to predict sentiment (positive or negative).
- **Model Evaluation:** After fine-tuning, the

model's performance was evaluated within the same period (intra-period testing) and across different periods (cross-period testing) using accuracy scores. This approach allowed for an assessment of how well the model generalizes to out-of-period data.

4 Results

4.1 Multi-Year Results

Train Period	2004-2007	2008-2011	2012-2015	2016-2019	2020-2023
2004-2007	90.25%	91.55%	93.67%	94.04%	92.92%
2008-2011	89.83%	91.68%	93.54%	93.94%	93.29%
2012-2015	87.73%	89.89%	94.71%	95.06%	94.16%
2016-2019	87.56%	89.65%	94.58%	95.00%	93.62%
2020-2023	88.15%	90.12%	94.16%	94.68%	94.00%

Figure 2: Accuracy matrix for models trained and tested on multi-year groupings.

When examining the 4-year grouping intervals (2004–2007, 2008–2011, 2012–2015, 2016–2019, 2020–2023), a clear trend emerges. Models generally achieve their highest accuracy when tested on data originating from the same time period as their training set. For instance, training on the 2012–2015 period yields a 94.71% accuracy when tested on 2012–2015 data, and training on 2016–2019 achieves 95.00% accuracy on its in-period test. Similarly, the 2020–2023 training period attains 94.00% when evaluated on 2020–2023 test data.

However, as the testing data moves further away from the model's training window, performance tends to decrease. For example, a model trained on 2004–2007 achieves only 90.25% accuracy when tested on in-period data but sees a gradual improvement when tested on more recent periods, up to 94.04% on 2016–2019, suggesting that while some generalizable sentiment features persist, the model does not consistently outperform models trained closer to the test period. Notably, training on older periods (e.g., 2004–2007 or 2008–2011) and then evaluating on newer time frames (2020–2023) results in lower accuracy (92.92–93.29%) compared to training and testing within the same recent period. In other words, models trained on older data still achieve reasonably high accuracy on newer data, but they do not match the performance of models trained on data from a closer temporal window.

Train Period	1999	2021	2022	2023
1999	88.01%	89.10%	90.33%	90.74%
2021	86.10%	89.78%	90.19%	89.65%
2022	86.10%	90.74%	90.87%	89.65%
2023	86.38%	89.10%	89.24%	89.37%

Figure 3: Accuracy matrix for models trained and tested on individual years.

4.2 Yearly Results

The single-year experiments (1999, 2021, 2022, and 2023) further support this pattern of temporal sensitivity. Models trained and tested on the same year exhibit the highest accuracy. For example, training on 1999 and testing on 1999 yields 88.01%, while training on 2022 and testing on 2022 reaches 90.87%. However, temporal distance reduces performance. A model trained on 1999 data scores lower on 2023 data (90.74%) than a model trained on 2023 itself (89.37% on 2023), and models trained on recent data (e.g., 2022 or 2023) do not fully recover performance when tested on distant past data (e.g., 1999), hovering around 86–89%. Although differences in accuracy across these yearly splits are less pronounced than in the 4-year groupings, the same principle applies: training closer to the test period typically yields stronger performance.

4.3 Monthly Results

At an even finer granularity, the monthly partitions of the 2022 dataset reveal that temporal shifts occur even on shorter timescales. Models trained on earlier months (e.g., January 2022) maintain relatively strong performance when tested on subsequent months of the same year (e.g., up to 92.31% on September data when trained in January), but subtle declines emerge as the temporal gap grows. Consistently, models achieve their peak performance on the month they are trained on or nearby months. For example:

- Training on January 2022 and testing on January 2022 yields 91.10%, while testing on later months like November or December still achieves strong results (91.92% and 92.28%, respectively), though rarely surpassing the performance of a model trained directly on those later months.

- Training on mid-year months (e.g., June 2022) shows similar patterns, with strong in-month

Train Period	Jan-22	Feb-22	Mar-22	Apr-22	May-22	Jun-22	Jul-22	Aug-22	Sep-22	Oct-22	Nov-22	Dec-22
Jan-22	91.10%	90.92%	91.80%	89.68%	91.80%	91.45%	90.60%	91.24%	92.31%	91.63%	91.92%	92.28%
Feb-22	90.98%	91.36%	91.72%	90.45%	91.72%	91.39%	91.33%	91.80%	92.16%	92.10%	92.18%	92.39%
Mar-22	90.74%	91.36%	91.36%	90.48%	91.92%	91.13%	90.83%	91.27%	91.92%	91.83%	91.07%	92.12%
Apr-22	91.12%	91.24%	90.89%	90.24%	92.39%	90.94%	90.94%	91.21%	92.16%	92.07%	92.01%	92.13%
May-22	90.30%	91.18%	91.04%	90.33%	92.10%	91.01%	91.30%	91.16%	91.33%	91.86%	92.28%	92.01%
Jun-22	90.68%	91.16%	91.48%	90.68%	92.10%	91.65%	90.77%	91.18%	91.74%	91.36%	91.92%	91.09%
Jul-22	90.68%	90.89%	91.33%	90.18%	91.95%	91.39%	91.04%	91.27%	92.27%	91.71%	91.54%	91.48%
Aug-22	91.00%	91.27%	90.98%	90.44%	91.39%	91.18%	91.04%	91.65%	92.07%	91.80%	91.54%	91.33%
Sep-22	91.30%	90.47%	91.03%	90.83%	92.27%	91.42%	91.24%	91.74%	92.25%	91.89%	91.57%	91.74%

Figure 4: Accuracy matrix for models trained and tested on monthly partitions of the 2022 dataset.

accuracy (91.65%) and high but slightly reduced accuracy for months at the year's end (around 92.09% in December).

- Across all monthly runs, the model maintains high accuracy within the same calendar year, suggesting that short-term linguistic shifts are present but more subtle than those observed across multiple years.

4.4 General Trends

In all temporal segmentations—multi-year, single-year, and monthly—several consistent trends emerge. Models achieve their best accuracy when trained and tested on the same or immediately adjacent time periods, underscoring the importance of temporal alignment. Although the model maintains relatively high accuracy when tested on data from different periods, there is a decline as the temporal distance grows. Older training data becomes less effective for classifying sentiment in more recent texts, and vice versa. The degradation is gradual rather than abrupt, unfolding over months or years. Even within a single year, performance remains high but peaks when tested close to the training window. Over longer spans (multiple years or decades), the impact of temporal misalignment is more pronounced.

These results strongly suggest that language use and sentiment expression evolve over time. While models retain a baseline level of performance across distant time frames, fine-tuning on data temporally closer to the test period yields better alignment with current linguistic norms, slang, cultural references, and product preferences. This highlights the potential need for continual or periodic re-training of sentiment analysis models to maintain peak performance in rapidly evolving data environments.

5 Discussion & Conclusion

Through this study, we have validated findings from prior research, confirming that data temporality significantly impacts sentiment analysis performance. Our results show that while temporal effects over shorter durations, particularly months within a single year, show minimal and variance in accuracy, these effects become more significant across larger timescales. Across years, we see noticeable variation in performance, and over 4-year periods, the degradation becomes even more evident, with declines in accuracy on the order of a few percentage points.

These findings highlight the evolving nature of language, emphasizing the importance of temporal alignment between training and testing data for robust sentiment analysis models. The results show the necessity for periodic updates to model training data to ensure optimal performance, especially in applications where maintaining high accuracy is critical.

Future work could explore this phenomenon in other domains and investigate approaches to mitigate the degradation caused by temporal misalignment.

6 Limitations

One of the key limitations of this study is that the dataset only contains movie and TV show reviews, which restricts the diversity of contexts in which sentiment analysis is applied. As a result, the findings may not fully generalize to other domains, such as product reviews, social media posts, or news articles, where language and sentiment can differ. The dataset is also very imbalanced, despite efforts to balance labels, it still may still affect model performance. Furthermore, the experiment only considers a binary sentiment classification, which might not capture the full range of sentiment expressed in reviews. Incorporating more nuanced sentiment labels (e.g., neutral, mixed sentiment)

could offer a deeper analysis of sentiment shifts over time. Additionally, investigating the temporal impact on other NLP tasks such as name entity recognition, word sense disambiguation, machine translation and so on, would provide a more comprehensive understanding of how language changes over time.

7 Statement of Contributions

Athmane implemented the data extraction from the Amazon reviews dataset and used it to fine-tune the BERT model adequately. Shoaib and Mohammad then extended the initial code to consider individual years and months. We then used the fine-tuned model to compute the different accuracies across the different samples. All three were actively involved in debugging and optimizing the code so that we could conduct the experiments. Also, all three members actively wrote and reviewed each section of the report.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. 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.

Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–71, Brussels, Belgium. Association for Computational Linguistics.