MORNING EMAIL NEWSLETTER TEXT MINING ANALYSIS

INFO 4900: INDEPENDENT RESEARCH WITH PROFESSOR MATTHEW WILKENS

FALL 2023

Wally Chang
Cornell University
Ithaca, New York
wsc46@cornell.edu

ABSTRACT

Morning email newsletters have emerged as ubiquitous mediums delivering daily events, breaking news, and entertainment to a global audience. This study explores the distinctive features of these newsletters, characterized by attention-grabbing headlines, engaging stories, clean design, and an authoritative yet quirky tone. Leveraging text mining techniques, the research aims to unveil deeper insights within these succinct newsletters, with a focus on the "Morning Newsletter" from *The New York Times* and "Morning Brew" from *The Morning Brew*. The project encompasses two primary objectives: a classification task categorizing newsletters into "NYT" or "MB" categories and a regression task predicting newsletter publication dates. The exploration details the construction of a corpus, featuring newsletters from October 11, 2022, to October 11, 2023, and the subsequent basic processing and comparisons between the two newsletters. Visualizations using Truncated SVD highlight distinctions in features, leading to an unsupervised cluster analysis. Classification tasks using logistic regression with feature selection, including default paragraphs and 200-token chunks, reveal success in computational differentiation between the newsletters, with accuracy scores achieving a notable accuracy score of 0.89. However, regression tasks aiming to predict publication dates encounter substantial difficulties. Linear regression and neural-based approaches fail to establish meaningful correlations between input features and target variables, emphasizing the complexities inherent in modeling temporal aspects of newsletter text data.

***K*eywords** morning email newsletters · text mining · classification · regression · BERT

1 Introduction

Morning email newsletters are becoming ubiquitous mediums through which daily events, breaking news, and entertainment are distributed to millions across the world. Although email may initially appear outdated for disseminating news, the remarkable success of email newsletters in the past decade, amid the challenges faced by traditional channels, suggests otherwise. This achievement becomes more comprehensible when considering an audience with increasingly short attention spans and a reliance on checking email, whether for work or other purposes.

Morning email newsletters are characterized by their attention-grabbing headlines, brief, often abbreviated, but engaging stories, clean design, vivid imagery, authoritative tone, and often, bits of quirky personality injected by their authors [1], [2]. These characteristics, distinct from traditional news media offerings, are predicated on an audience that is familiar and comfortable with technology, quick to bore, and often, viewing on a mobile device.

Given the unique characteristics of morning email newsletters, an enticing prospect arises to utilize text mining techniques for revealing deeper insights within these concise pieces. This could involve gaining a nuanced understanding of the temporal significance in current event descriptions or discerning variations between different email newsletter outlets. In this exploration, *The New York Times's* "Morning Newsletter" and *The Morning Brew's* "Morning Brew" were compared.

This exploration primarily centered on achieving two overarching objectives: a classification task categorizing input newsletters into "NYT" or "MB" categories, and a regression task predicting the publication dates of the input newsletters.

This paper will offer a comprehensive examination of the proposed project—a text mining analysis of email newsletters. It will delve into the methodologies employed, examine the findings, and critically assess both successful aspects and areas that presented challenges.

1.1 Purpose of the Project

As an avid consumer of morning email newsletters, subscribed to more than seven distinct outlets, I contemplated the factors that drew audiences to each publication and contributed to their individual successes. With my background in Data Science and simultaneous enrollment in INFO 6350 - Text Mining History and Literature during the Fall 2023 semester, I realized a prime opportunity to answer these questions by utilizing text mining techniques to unveil distinguishing factors or to quantify the semantic differences between the newsletters that so strongly captivated my interest.

1.2 Project Pre-Registration and Beginnings

This project began with a pre-research project registration proposal in early September 2023. As described below, there was an initial focus on capturing sentiment differences between newsletters, though this was ultimately omitted from the investigation due to time constraints. The scope of newspapers included in the analysis would also be narrowed from the original proposition for the same reason:

> *This study will attempt to draw conclusions about general sentiment differences in response to landmark societal events between 2021 and 2023 between various news outlets' morning email newsletters. Morning newsletters are the subject of this exploration, as they are a unique repository of current events, intentionally dense in information and thus select in their word choice due to their necessitated brevity. In addition, morning newsletters often tout certain characteristics, including claims to be "unbiased", or having a specified "cheer me up" section, which will be interesting to analyze in the scope of this investigation. Data will be gathered from the New York Times's "The Morning Newsletter", the Morning Brew's morning newsletter, the The Know's "Daily ", and 1440's morning newsletter. Then, natural language processing techniques will be applied to the compiled dataset, possibly including but not limited to vectorization, VADER rule-based sentiment analyzer, and similarity scoring, in order to derive higher level insights about the sentiment characteristics, similarities, and differences between morning newsletters.*

Once the project proposal was approved, a meeting with Professor Wilkens was arranged to clarify objectives and solidify a starting direction. Work could then begin to identify the specific newsletters that would be included in the corpus used for analysis.

2 Related Work

The theoretical foundations of this work were grounded in basic text mining and text analysis methods. Within the scope of this project, the primary emphasis revolved around two key methods: classification and regression. Supplementary techniques, including topic modeling, clustering, and feature selection, were also employed to enhance the interpretability of results and to guide decision-making regarding subsequent courses of action as the study progressed.

This paper drew inspiration from Grimmer and Stewart (2013) to establish its overarching framework [3]. Following their lead, the structure initially employs unsupervised methods to discern underlying text features, generating clusters of input objects. Subsequently, the paper seeks to validate these clusters before transitioning to the application of supervised learning models. Finally, the exploration extends to neural methods, a facet not covered in Grimmer and Stewart's work.

Topic modeling played a pivotal role in this project, offering a valuable means to distill broad themes within the semi-large collection of newsletters constituting the corpus. The work of Boyd-Graber et al. (2017) [4] and Baumer et al. (2017) [5] provided contextual insights into addressing this "needle in a haystack" challenge, showcasing how Latent Dirichlet Allocation (LDA) could be effectively applied, particularly within the realm of social science exploration.

Drawing inspiration from these sources and guided by Professor Wilkens, the decision was later made to chunk input features, aiming to enhance the clarity of output topics derived from the LDA topic model.

Applications of classification and regression in this project were largely based on in-class lecture material from INFO 6350 - Text Mining History and Literature, though example experiments by Allison (2011) [6] and explanations to regression from the Princeton University Library [7] were certainly helpful as contextualization and support factors as well.

3 Corpus Creation

The construction of this corpus involved the collection and compilation of morning email newsletters. There were many newsletters that could have merited interesting comparison and research, including "The Skimm", "1440", *The Wall Street Journal's* "The 10-Point", *The Morning Brew's* "Morning Brew", and *The New York Times's* "The Morning Newsletter". Each was distinct with their own unique styles of news presentation - the "1440" carved out a niche as being "fact-focused", and claimed to "... share fact-focused information with more than 3 million people who trust us to be as unbiased as humanly possible". "The Skimm" offered its twist in its dedication to serving a female audience, as it stated, "theSkimm is a digital media company, dedicated to succinctly giving women the information they need to make confident decisions".

Amidst a myriad of choices, the selection of newsletters boiled down to a question of feasibility and efficiency, considering the demands of a progressing semester, and the challenge of actually gathering the newsletters, given the wide range of online newsletter archival methods. After researching the mentioned sources, two newsletters—*The New York Times's* "Morning Newsletter" and *The Morning Brew's* "Morning Brew"—distinguished themselves with easily accessible repositories of archived newsletters, suggesting a viable process for newsletter scraping and subsequent data loading. Building upon these two newsletters, the subsequent step focused on defining the scope of examination, particularly determining the timeframe covered by the corpus of newsletters. A 1-year timeframe was chosen, spanning from October 11, 2022 to October 11, 2023 for both newsletters. The choice of a 1-year timeframe was driven by the goal of offering a comprehensive overview of current events, coupled with a consideration for the possibility and practicality of manual data collection.

3.1 Morning Brew Newsletter Collection

The first task was to collect newsletters from "Morning Brew," a process that proved to be straightforward and efficient with the successful implementation of web scraping. URLs from each newsletter were aggregated into a list, then processed using the BeautifulSoup Python package to parse the HTML. This automatically extracted the title, text, and authors of each newsletter. The parsed data was then organized into a Pandas dataframe and passed through a data cleaning process. Notably, the dataset for the "Morning Brew" comprised 60 fewer newsletters than the dataset for *The New York Times*, because all Sunday newsletters were excluded. The decision to omit Sundays stemmed from the challenges posed by a distinct HTML format used for the "Morning Brew" Sunday newsletters, which would have significantly complicated the web scraping process.

3.2 New York Times Newsletter Collection

Initial attempts to collect newsletters from the "Morning Newsletter" archive were replications of the procedure used to collect "Morning Brew" newsletters. These attempts failed due to *The New York Times's* web scraping prevention measures. The workaround was to manually download each newsletter as an HTML file, saving them to a local folder. Then, BeautifulSoup could be applied by iterating through files in the folder to extract the title, text, and authors of each newsletter as before. Again, the parsed information was organized into a Pandas dataframe and cleaned.

3.3 Putting Everything Together

Upon loading all the text, additional cleaning and processing were executed to ensure uniformity between the two newsletters. This included converting date strings to Python datetime objects, converting non-year numbers into "num" tags, and fixing any other inconsistencies from the initial data pull. Finally, the two segregate dataframes were combined and shuffled to mitigate any effects that concatenation might have on the dataset (See Section 10). (See Figure 1)

	newsletter	title	date	authors	paragraphs
0	NYT	Herschel Walker's Polling Dip	2022-10-14	Nate Cohn	How the Senate race in Georgia is shaking out....
1	NYT	It's Coronation Day	2023-05-06	Melissa Kirsch	Come for the crown jewels and gold stagecoach;...
2	MB	Abercrombie is back	2023-05-25	[Molly Liebergall, Matty Merritt, Cassandra Ca...	\nGood morning. If you thought you were gettin...
3	MB	They got 50	2022-11-14	[Neal Freyman]	\nGood morning. If you're looking for a distra...
4	NYT	Brazil and Jan. 6	2023-01-10	German Lopez	How Brazil's riots compare to the Jan. num att...
5	NYT	What Happened to Monkeypox?	2022-10-13	German Lopez	Why cases suddenly began to decline.\nBy Germa...
6	NYT	The End of a Presidential Launchpad	2022-12-06	Peter Baker	Democrats stripped Iowa of its first-in-the-na...
7	NYT	History in the Rubble	2023-03-26	Ashley Wu	Documenting the damage of last month's earthqu...
8	MB	Breaking the seal	2023-04-05	[Sam Klebanov, Matty Merritt, Neal Freyman]	\nGood morning and Happy Passover to all those...
9	NYT	Compounding Disasters	2023-07-13	German Lopez	America's compounding natural disasters show t...
10	NYT	Welcome to Barbenheimer Weekend	2023-07-22	Melissa Kirsch	You have little choice but to surrender to the...
11	NYT	Debt Ceiling Fight Is Putting U.S. Economy at ...	2023-01-20	German Lopez	A political fight is again putting the economy...

Figure 1: Combined Dataset

4 Basic Processing and Comparisons

After loading and cleaning the data, the following step involved basic processing and comparison between the two newsletters. Various processes were undertaken, including the comparison of the 25th percentile, median, and 75th percentile number of tokens in each newsletter to evaluate differences in newsletter length (See Table 1). Additionally, an examination of the number of newsletters for each type and the identification of dates for oldest versus most recent publications was performed (See Table 2). Of note, the "Morning Brew" had about 30% more tokens, when comparing medians, than the NYT's newsletter. The "Morning Brew" also had 60 fewer newsletters than the NYT newsletter, as noted in Section 3.1.

	Morning Brew	New York Times
25th Percentile	2471	1847
Median	2549	1956
75th Percentile	2653	2054

Table 1: 25th, Median, and 75th Percentile # of Tokens in MB vs NYT

	Morning Brew	New York Times
# Newsletters in Corpus	305	365
Oldest Newsletter	10/11/2022	10/11/2022
Most Recent Newsletter	10/11/2023	10/11/2023

Table 2: # Newsletters & Oldest/Newest Dated Newsletters

Following this, discrete topic models were generated separately using exclusively "Morning Brew" and exclusively "NYT" text, respectively. The objective behind implementing discrete models was to visualize differences in topics between the two news outlets at an early stage. Latent Dirichlet Allocation (LDA) was utilized to produce probabilistic topic models here, harnessing its inherent strengths in capturing the diverse and sparse nature of newsletter content. The flexibility of the Dirichlet distribution in parameterizing concentration allowed for an accurate representation of topical distributions within each document, and a capturing of thematic variations between the two newsletters. The interpretability gained through LDA's sparse topic modeling enhanced analytical depth, allowing for a comparative examination of the prevalent themes encapsulated in each newsletter.

To prepare feature input, the newsletter documents were pre-processed, such that paragraphs were delineated based on HTML line breaks (\n), then flattened into a list, and finally vectorized using a Count Vectorizer. The Count Vectorizer was configured with a minimum document frequency of 0.001 and a maximum document frequency of 0.25. These settings, well applied to a document broken into paragraphs, served to filter out terms that were either too infrequent or

too common within the paragraphs. The vectorizer also accommodated both unigrams and bigrams while excluding stop word removal, tailoring the feature representation to capture meaningful patterns in the paragraphed newsletter documents. Ultimately, 50 topics were generated for each newsletter, both of which captured semi-realistic topics (See Appendix A.1 and A.2), but which would be improved on in later iterations of topic modeling on the full consolidated dataset in Section 6.1.

Finally, a visualization was created for the 700 most frequently occurring token unigrams between the "Morning Brew" and "NYT," employing Truncated SVD for dimensionality reduction. The choice of Truncated SVD was motivated by its effectiveness in handling high-dimensional and sparse datasets. Truncated SVD allowed for a more interpretable representation of the underlying patterns in the data while mitigating the computational challenges associated with large and sparse matrices, making it a suitable technique for this visual exploration (See Figure 2). Here, a TF-IDF vectorizer was employed with specific parameters: max_features = 700 and stop_words = 'english', as opposed to a Count Vectorizer. This choice was made to enhance the extraction of distinct characteristics from individual documents. The full combined text was then vectorized.

As evident from the visualization, a notable distinction existed between the features of "Morning Brew" and "NYT" when plotted in a dimension-reduced space. This observation prompted the initiation of an unsupervised cluster analysis.

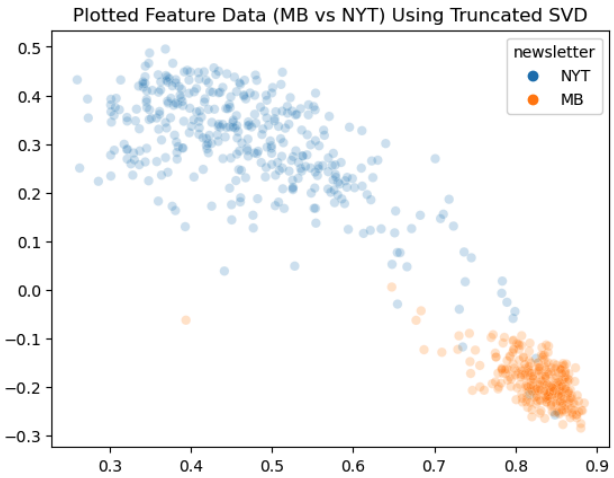

Figure 2: Dimension-Reduced 700 Most Frequently Occurring Token Unigrams Between MB and NYT

5 Unsupervised Cluster Analysis

Next, an unsupervised cluster analysis was performed in an attempt to further identify organizations of text within the data before undertaking validation or supervised learning measures. The K-Means algorithm - perhaps the most widely used method of Fully Automated Clustering (FAC) was used towards this objective [3]. In essence, the K-Means algorithm aimed to identify a partition of the documents that minimized the squared Euclidean distance within clusters.

First, the elbow method was used to identify an optimal value for "k", the number of clusters to use during the K-Means clustering [8]. Though there was no clear elbow point, k=3 was chosen as the optimal number of clusters given the drop off in the Within-Cluster Sum of Squares (WCSS), the measure of variability of observations within each cluster, after k=2.5 (See Figure 3). Then, K-Means was used to predict on the same vectorized feature set as was previously plotted, and its output clusters printed out (See Figure 4).

The clusters revealed by K-Means closely mirror the dimension-reduced feature data, as illustrated in Figure 2. Consequently, it was hypothesized that newsletter classification might be a straightforward task. High classification accuracy scores were anticipated, even with baseline classification methods, owing to the pronounced feature distinctions observed between the newsletters.

6 Supervised Classification Task for Newsletter Type

Following the completion of the unsupervised cluster analysis, a newsletter classification was conducted using logistic regression. Logistic regression, a statistical method employed for binary and multiclass classification tasks, is particularly

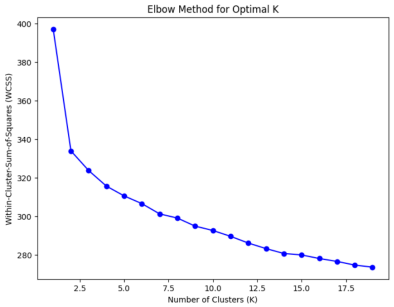

Figure 3: Elbow Method for Findng Optimal K

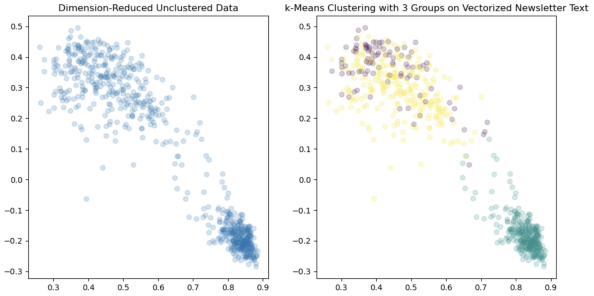

Figure 4: K-Means Clustering With 3 Groups

well-suited for this context. It models the probability of each newsletter belonging to a specific type based on the identified features, providing a probabilistic framework for classification. For this classification task, two main approaches for feature input were employed: **topic-based features** and **token-based features**. Within these methods, feature selection was further subdivided. In the first, **default paragraphs**, as described in Section 4 were used as token inputs and to inform topics. In the second, **200-token chunks** served as token inputs and were utilized to inform topics.

6.1 With Topic Based Features

The selection of topic-based features in the context of a supervised classification task for newsletter type was motivated by a desire to enhance the interpretability and semantic understanding of textual content. By utilizing topic modeling techniques, specifically Latent Dirichlet Allocation (LDA), the goal was to extract latent topics that encapsulated the underlying themes within newsletters. This approach offered a more abstract and contextual representation of the content, allowing for the identification of recurring topics and capturing the inherent structure of the documents. The uncovering of meaningful patterns in the data could facilitate more informed and nuanced predictions of newsletter types in a supervised learning framework. The exploration of topic-based features aimed to strike a balance between the interpretability of extracted topics and the predictive power required for effective classification tasks.

6.1.1 Using Default Paragraph Input

First, a topic model was generated using a similar process as the earlier discrete topic models seen in Section 4, utilizing LDA and default paragraph text input. In this iteration though, the complete amalgamation of texts was used as input, and a shift was made to use a TF-IDF vectorizer instead of the prior Count Vectorizer. The TF-IDF vectorizer was chosen for its capacity to discern terms that held significance not only in frequency but also in distinctiveness to specific documents, assigning elevated weights to rare terms that carried more meaningful information. This adjustment was anticipated to enhance the model's classification performance by emphasizing terms that contributed uniquely to each document's content. Default paragraphs were fed into this vectorizer. Then, the feature matrix was fed into a function that output 50 topics.

Despite occasional successes as seen in Table 3, the topic model predominantly generated uninterpretable results. Most topics presented confusing word combinations or simply consisted of stop words, highlighting limitations in the model's ability to discern meaningful themes. Regardless, proceeding to the next step, the vectorized input text was LDA transformed, returning a document topic matrix with shape (59839×50), then standard scaled and fed into a logistic regression predictor. After cross-validation, the classification of newsletters using topic features returned a mean accuracy of 0.663. Given the clear distinctions between newsletters seen in the earlier clustering analysis, it was evident that improvements in classification performance could be achieved through refinement of input features or other model modifications.

6.1.2 Using 200-Token Chunks as Input

Significant variation in the token lengths of default paragraphs led to the hypothesis that this variability might have adversely affected the performance of the topic modeling algorithm. Therefore, 200-token chunks were generated from the text to ensure a more uniform representation of the documents. After chunking, the same procedure was followed as with the default paragraph inputs.

The output of the topics appeared to be more meaningful overall (See Table 4, with fewer topics that were outright uninterpretable. Upon closer examination of the model, the document-topic matrix exhibited dimensions of (25363×50),

Quality	Topic #	Words	Hand-Labeled Interpretation
Good	6	july, prigozhin, wagner, russia, belarus, canada, pakistan, wounded, organized, yevgeny	Wagner Rebellion
Good	29	cup, women, world, the, france, soccer, spain, tournament, australia, argentina	Women's World Cup
Bad	9	bitcoin, monogram, alaska, independent, egypt, pool, no, knee, legally, change	?
Bad	26	the, to, you, and, of, it, your, for, in, is	? - Stop Words

Table 3: Selection of Good and Bad Topics Using Default Paragraph Inputs
See Appendix A.3 for Full Topic Model

Quality	Topic #	Words	Hand-Labeled Interpretation
Good	2	party, the, political, republican, voters, republicans, election, democrats, in, democratic	Politics
Good	13	ai, chatgpt, the, to, chatbot, artificial, san, openai,and, francisco	Artificial Intelligence
Medium	31	real, estate, the, art, cover, and, of, housing, saving, in	Real Estate?
Bad	44	we, it, that, you, my, to, they, what, do, so	? - Stop Words

Table 4: Selection of Good, Medium and Bad Topics Using 200-Token Chunk Inputs
See Appendix A.4 for Full Topic Model

indicating a considerable proportion of shorter paragraphs in the earlier methodology and a greater proportion in larger chunks in the current approach. When input into a logistic regression classifier and subjected to cross-validation, the mean accuracy score also demonstrated improvement, reaching 0.760. As such, the initial hypothesis suggesting a more uniform representation of documents as input seemed to be supported.

6.2 With Token Based Features

As topics were found to be suboptimal features for predicting newsletter types, the subsequent approach involved employing the more conventional method of utilizing tokens as features. Token-based features offer a granular representation of textual content, capturing detailed linguistic information and potentially improving the model's sensitivity to more nuanced patterns. The shift toward token-based features allowed for a comprehensive examination of individual words and their impact on classification outcomes.

6.2.1 Using Default Paragraph Input

Having previously generated a vectorized set of default paragraphs during the topic-based exploration, it was straightforward to repurpose and input it into a logistic regression classifier, this time without the LDA transformation. The shape of the feature matrix now stood at 59839×38726, signifying a more comprehensive representation of textual content input. Upon cross-validation and mean accuracy assessment, a score of 0.842 was obtained, marking a substantial improvement over the accuracy achieved through the topic-based methodology.

6.2.2 Using 200-Token Chunks as Input

Building on the success of enhancing classification performance in the topic-based methodology by using 200-token chunks as text input, a similar approach was employed here with the expectation that it would lead to further improvement in classification. The feature matrix now took the shape 25363×38726. Following cross-validation and mean accuracy assessment, an impressive score of 0.89 was attained, once more surpassing the accuracy achieved with the default paragraph text input.

The effectiveness of utilizing tokens as features for newsletter classification appeared to surpass that of using topics, aligning with the notion that more granular representations of the text would capture more nuanced information. This observation acknowledges the inherent trade-off between granularity and abstraction in feature representation. It is crucial to note that, in this case, a train-test split was not implemented, which could have posed the risk of biased results and potentially inflated accuracy scores. While the findings suggested the superiority of token-based features, a comprehensive evaluation with a more robust experimental setup, including a train-test split, is necessary to ensure the reliability of the classification model.

7 Supervised Regression Task for Publication Date Prediction

With newsletter classification achieving high accuracy scores, exploration moved on to attempting a regression task to predict the publication date for a given feature input. Linear regression, a statistical technique commonly used for predicting numerical values, was used to forecast the publication date of a newsletter based on input text. This method modeled the linear relationship between the identified features in the text and the target variable (publication date). Similarly to the classification task, topic and token based features were used, with further subdivision into default paragraph and 200-token-chunk raw text input.

This task involved an additional step of data processing. Since date data was saved in the dataframe in the format (yyyy-mm-dd), conversion to an integer representation was necessary for usage of linear regression, which predicts numerical values. Towards this purpose, dates were simply converted to integers counting up from the earliest date, 2022/10/11, expressed as 0.

To assess model performance, R-squared, also known as the coefficient of determination, was used, as it provided a measure of how well the predicted values matched the actual values. It quantified the proportion of the variance in the dependent variable that was predictable from the independent variable. A higher R-squared value (closer to 1) indicates a better fit, suggesting that the linear regression model captured a larger portion of the variability in the data, whereas a lower R-squared value indicates the converse. Therefore, using mean cross-validated R-squared helped to gauge the overall effectiveness of the linear regression predictor in explaining variability in the target variable based on the given feature matrix.

7.1 With Topic Based Features

7.1.1 Using Default Paragraph Input

With data pre-processing concluded from the previous task, inputting the feature matrix derived from default paragraphs into a linear regression predictor was straightforward. The predictor was trained on the default-paragraph-informed document-topic matrix and the integer representation of the dates. The document-topic matrix was then input back into the linear regression predictor, and evaluated on mean cross-validated R-squared. This attempt returned an R-squared score of -0.015. The negative value indicated that the model failed to capture any meaningful relationship between the variables, which is visualized in Figure 5.

7.1.2 Using 200-Token Chunks as Input

The same procedure was then performed again, but now with the 200-token-chunk-informed document-topic matrix. This returned a mean cross-validated R-squared score of -0.021, a non-improvement over the previous score. To further analyze and contextualize this score, a time-series was plotted to visualize individual topic probability over time, rather than to repeat the uninformative plot in Section 7.1.1 (See Figure 6).

The time series visualization revealed a few realistic and interpretable correlations, for example the spike in probability of Topic 9 around day 280 (July 18th, 2023), roughly corresponding with the FIFA Women's World Cup, which took place from 20 July to 20 August 2023. A spike in probability of Topic 33 around day 60 (December 10, 2022), roughly correlated with King Charles III's widely publicized Christmas message where he paid tribute to the late Queen

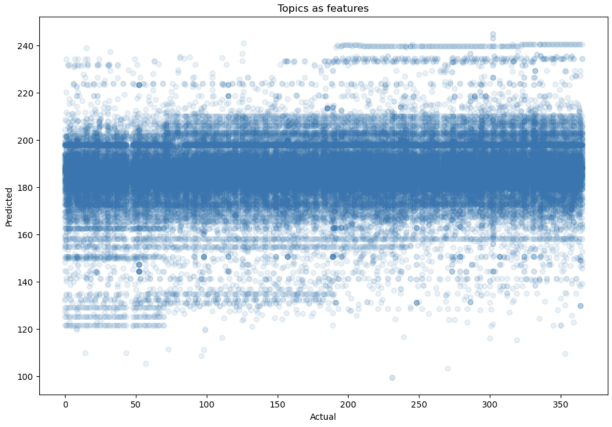

Figure 5: Scatterplot Showing Default-Paragraph Informed, Topic-Based Predictions Against Actual Values

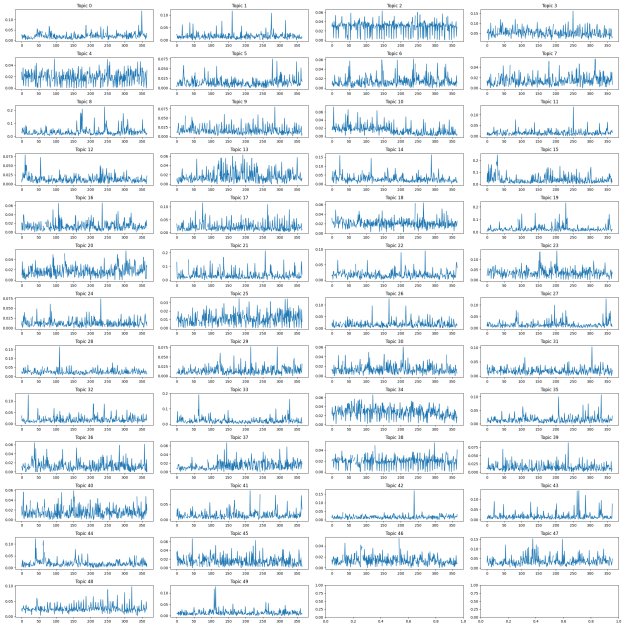

Figure 6: Time Series of 200-Token-Informed Topics

Elizabeth II, but did not mention either Prince Harry, Duke of Sussex, or Meghan, Duchess of Sussex. Though, these correlations were weak and far from obvious. Most topics appeared to be randomly apparent throughout the year.

7.2 With Token Based Features

The regression task was then executed using token-based features. However, considering the task's objective of predicting a single numerical value, it was hypothesized that performance might be compromised due to the diminished context of granular data and the limited training data for each corresponding label (only 2 newsletters per day).

7.2.1 Using Default Paragraph Input

Using a default-paragraph-informed feature matrix worsened performance, returning a mean cross-validated R-squared score of -6.52, supporting the initial hypothesis. The results were plotted, which returned an abnormal visualization as seen in Figure 7. It appeared as though the model was overfitting to some aspect of the data, leading to an artificially strong linear correlation. This explanation was backed by comparison to the non-cross-validated R-squared score of 0.634.

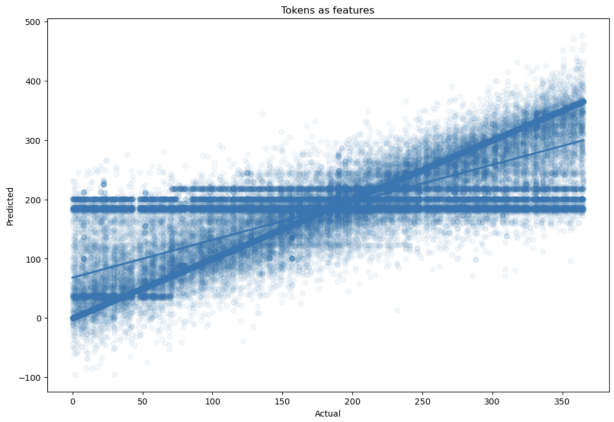

Figure 7: Scatterplot Showing Default-Paragraph Informed, Token-Based Predictions Against Actual Values

7.2.2 Using 200-Token Chunks as Input

Chunking tokens served only to worsen performance, returning a mean cross-validated R-squared score of -35.748. When plotted, the effects of overfitting and apparent linear correlation also worsened (See Appendix A.6).

8 Linear Regression Prediction of Month With 200-Token Chunk Features

Since the baseline linear regression prediction of integer date representations yielded null results, it was hypothesized that predicting integers might be too challenging, particularly considering the limitations to training data per date, and each date's singular occurrence over the course of a year. Therefore, it was attempted to predict by month rather than by specific date.

Implementation was straightforward, necessitating only the additional step of extracting the month value from each date and assigning those values as labels. 200-token chunks were chosen as inputs for the feature matrix. The same procedure was applied as previously, and the mean cross-validated R-squared value of -3.171 was returned. Though this result was an improvement from the prior -35.748, the negative value still indicated that the model failed to capture any meaningful relationship between the variables. (See plotted predictions in Appendix A.5).

Lastly, in an attempt to analyze which features were most influential to the outputs of the model, the predictor's coefficients were sorted by magnitude, and the top 20 most informative features were extracted, along with their predicted values, as seen in Figure 8.

Combined DataFrame:

	Feature	Coefficient	Predicted_Value
18	pegging	-1.386863e+13	6.298603
19	relist	1.386863e+13	6.542898
5	sportico	-3.993068e-01	6.515594
14	nonrambunctious	-3.375513e-01	8.115440
0	lvi	-3.061372e-01	6.955588
12	bot	2.674184e-01	6.542898
3	broadcasts	1.817614e-01	6.515594
16	draft	-1.526122e-01	6.515594
2	pills	-1.470380e-01	6.762895
17	moses	-1.360667e-01	6.884518
13	pill	-1.211410e-01	6.570202
8	nashville	-8.191928e-02	6.515594
9	newsroom	-6.622121e-02	6.515594
1	london	6.537748e-02	6.542898
6	devoting	-5.920290e-02	6.735591
4	writer	-4.699378e-02	6.515594
7	abortion	4.655960e-02	6.515594
11	nielsen	-4.048907e-02	6.515594
15	winter	3.395794e-02	5.574814
10	this	1.349765e-02	6.515594

Figure 8: Most Informative Features Based on Coefficients and Their Predicted Value

None of the terms stood out as being particularly meaningful or date-correlated, stymieing any further interpretation.

9 Neural-Network Based Regression Task for Publication Date Prediction

Following the unsuccessful attempts with baseline linear regression to establish a meaningful correlation between token or topic inputs and publication dates, a neural approach was pursued in an effort to enhance performance. The exploration involved the implementation of a BERT (Bidirectional Encoder Representations from Transformers) model for the task. BERT, known for its contextualized word embeddings and capability to capture intricate linguistic dependencies, was selected with the anticipation that its advanced architecture might better capture the nuanced relationships between input features and publication dates. This shift toward a neural method, leveraging pre-trained language models, aimed to uncover more intricate patterns and improve the overall predictive capabilities for the temporal aspect of the data.

The fine-tuning and BERT implementation procedures were derived directly from Galtier (2021) [9]. Modifications necessary from Galtier's work included the usage of BERT rather than CamemBERT, given the usage of English text rather than French, in the case of Galtier. Usage of pre-existing libraries made processes like corpus encoding, data splitting into train/test/validation sets, and data conversion for PyTorch, simple. Since BERT is typically used for classification tasks, Galtier modified the final hidden state in the model architecture such that the outputs might be utilized for regression instead, as stated below:

> *In BERT's paper, it is indicated that only the final hidden state of the first token in the output sequence should be used for classification tasks. In other words, only the vector representing the "[CLS]" token mentioned earlier, should be used. For our regression task, we will do the same thing but rather than add a dense pooling layer for classification behind it, we will add a dense linear layer with dropout that will serve as our final regression layer [9].*

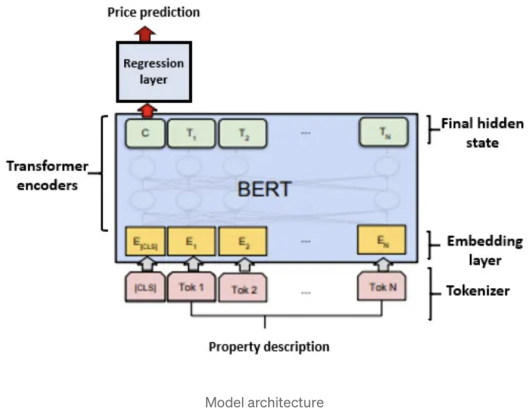

Figure 9: BERT Regression Model Architecture as Depicted in Galtier (2021)

To enable model training in a feasible amount of time, 100 premium credits were purchased through Google Colaboratory, which enabled access to and use of an Nvidia V100 GPU for training. The model was trained for 5 epochs, which took about 40 minutes.

When the results of both test loss and test R-squared were plotted, see Figures 10, 11, it was evident that the application of a neural-based approach also failed to capture any discernible relationship between the input features and the target variable, further emphasizing the complexity and challenges apparent in modeling the temporal aspects of the chunked newsletter text data. It was possible that the model performed poorly due to insufficient data, non-optimal fine-tuning, or a lack of enough training time, but it was also possible that 200-token chunks as features plainly failed to adequately capture patterns that correlated with temporality.

10 Procedural Review and Future Modifications

The investigative process and methodologies employed in the course of this project evolved organically as my comprehension and mastery of these methods transformed over time. This evolution was shaped by the progression through the

Figure 10: Test Loss Over Training Batches

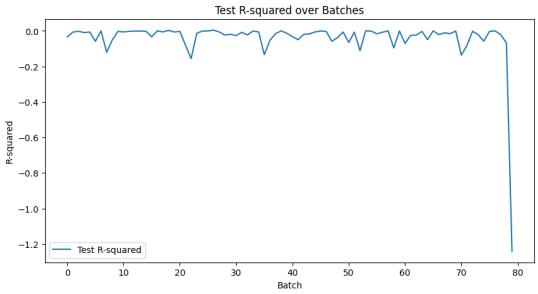

Figure 11: Test R-Squared Over Training Batches

class INFO 6350 - Text Mining History and Literature. Hence, despite ongoing incorporation of new methods into this project, a considerable amount of time was dedicated to revisiting earlier modules to rectify and edit inconsistencies or errors in the implementation that became apparent only later in the process. As such, the order of the investigative procedure outlined in this paper is significantly less evident in the Jupyter Notebook file. The organization of thoughts and detailing of procedural decisions were predominantly carried out after the initial implementation.

In retrospect, dedicating additional time and effort to thorough literature review in text mining, humanities, and the specific domain of newsletters before initiating the project would have provided a more comprehensive foundation from which to start. Such a review could have facilitated a deeper understanding of the investigation's subject matter and helped contextualize the computational results derived through various methodologies. Moreover, in the initial stages of processing and comparing the two newsletters, there could have been additional measures taken, such as the application of the "Fightin' Words" algorithm. This algorithm could have enhanced the understanding of the most distinguishing factors within each newsletter. Additionally, incorporating more visual aids, such as graphs, would have been beneficial for clearer visualization of disparities between the two newsletters. In the classification task, there could have been a heightened focus on enhancing explainability and thoroughly scrutinizing the model, its inputs and its outcomes. For example, midway through the exploration, a significant step function in label predictions became evident. However, strangely, this anomaly disappeared when the dataset was randomly shuffled. Alternative baseline classification methods could have been explored beyond logistic regression. Conforming to established practices in the field like implementing a train/test split would also have been advisable for both classification and regression tasks. Additionally, dedicating more time to fine-tune the BERT model and enhancing the visualization of its outputs beyond the stock implementation detailed by Galtier could have yielded valuable insights. Subsequent research could also involve additional efforts to refine the regression task focusing on months, like exploring the utilization of topic features, which demonstrated superior performance in regression compared to token features. The absences in the implementation of these changes were due to time limitations toward the end of the semester.

11 Conclusion

This exploration primarily sought to explore the characteristics and differences between morning email newsletters by employing text analysis and text mining techniques over the course of the Fall 2023 semester. After a corpus comprising one year of email newsletters spanning from October 11, 2022, to October 11, 2023, sourced from both the "Morning Brew" by *Morning Brew* and the "Morning Newsletter" by *New York Times* was compiled, preliminary processing and comparative analyses were carried out on the two datasets. Subsequently, the investigation proceeded to a classification task and a regression task.

Following the corpus creation process, it was quickly realized that there were significant differences between the two newsletters, as visualized in the K-Means clustering analysis. As outlined in the classification task, the process of categorizing newsletters proved to be both straightforward and successful. The model achieved an accuracy rate of nearly 90%, with the most effective performance observed when utilizing token-informed features rather than topic-informed features.

The next task, regression, proved to be immensely more challenging than initially expected. After an initial attempt using topics as features returned a negative R-squared score, subsequent token based attempts proved to be even less effective. Even efforts like chunking date by month, and applying neural-based methods failed to realize any significant improvement.

Though, the intricacies encountered during the regression task were perhaps the most thought-provoking and intriguing of all the stages of this exploration, consistently prompting the exploration of new ideas and avenues to address the challenges and gain insights into the factors hindering the model's performance. Through the iterative attempts to address these problems, my comprehension of the employed text mining methods substantially deepened, and I simultaneously broadened my awareness of both the capabilities and constraints inherent in computational text analysis methods. The regression task, unfortunately, remained unsolved within the time constraints. However, numerous possibilities for future research lie ahead, urging the exploration of a variety of alternative methods and strategies.

Acknowledgements

This paper would not be possible without the guidance of Professor Matthew Wilkens. Weekly meetings with Professor Wilkens served to provide clarity when there was confusion, and his constant suggestions and recommendations expanded my own thought horizon and made the research process exciting and enjoyable. Professor Wilkens was indispensable throughout the course of the semester as both a teacher and a mentor, and provided meaning, guidance, support, and context to the work that was being done.

References

[1] R. Bertin, *How On Earth Did Email Newsletters Become Popular Again?* 2017. [Online]. Available: https://medium.com/the-mission/how-on-earth-did-email-newsletters-become-popular-again-3fcee1addc7e.

[2] C. Monitor, *What Makes NYT's "The Morning" Newsletter One of the Most Popular in the World*, 2022. [Online]. Available: https://www.campaignmonitor.com/blog/email-marketing/what-makes-nyts-the-morning-newsletter-so-successful/.

[3] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis*, vol. 21, pp. 267–297, 2013, ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mps028.

[4] J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of topic models," 2017.

[5] E. P. S. Baumer, D. Mimno, S. Guha, E. Quan, and G. K. Gay, "Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?" en, *Journal of the Association for Information Science and Technology*, vol. 68, no. 6, pp. 1397–1410, 2017, ISSN: 2330-1643. DOI: 10.1002/asi.23786.

[6] S. Allison, "Quantitative formalism: An experiment," 2011.

[7] Princeton-University-Library and D. R. Adams, *Introduction to regression*, 1989. [Online]. Available: https://dss.princeton.edu/online_help/analysis/regression_intro.htm.

[8] B. Saji, *Elbow method for finding the optimal number of clusters in k-means*, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/.

[9] A. Galtier, *Fine-tuning bert for a regression task: Is a description enough to predict a property's list price?* 2022. [Online]. Available: https://medium.com/ilb-labs-publications/fine-tuning-bert-for-a-regression-task-is-a-description-enough-to-predict-a-propertys-list-price-cf97cd7cb98a.

Appendix

```
Topic  0: it company billion num billion its the company that for has is
Topic  1: here it you today play mini it here brew is play it
Topic  2: is tax an not is not financial offer facet investment this
Topic  3: for your with off use num off is for num code at
Topic  4: up on up to top to num tech future cash latest with
Topic  5: just just num minutes get leadership in just num minutes smarter get smarter smarter in
Topic  6: what these numbers mean here here what bank these numbers what these numbers mean
Topic  7: good morning what is good morning going more it this about
Topic  8: you your get can with you can it to get you re re
Topic  9: its search podcasts interested interested in in podcasts of its plans first plans to
Topic 10: has on been has been on the weekend biggest getting news the biggest
Topic 11: ai it on apple social an media that is users
Topic 12: world by the world written written by first team for on cup
Topic 13: as data as of data as close of num market stock et cryptocurrency
Topic 14: crypto house ftx debt deal the house exchange on sbf else
Topic 15: num num bps num bps is num num in num and minimum by num net returns
Topic 16: be it not but to be that musk will may will be
Topic 17: our your brew with learn for how it here on
Topic 18: from from the money note how founder note from personal how to makes
Topic 19: read for help with is this card offering no credit
Topic 20: without business education without the bs education without business education the bs one headlines
Topic 21: we you link give we ll ll we re re give you and more
Topic 22: his he million num million was at for that on with
Topic 23: they we you re if ve have it but do
Topic 24: of the most day one is the most one of the day for word
Topic 25: this week that it last for this week is month could
Topic 26: us as the us is that will on of the in the are
Topic 27: is this content this is advertising sponsored is sponsored sponsored advertising advertising content the
only
Topic 28: year than more more than according according to last num year than num per
Topic 29: your referral your referral daily others count or kid copy link
Topic 30: much analytics how analytics and behind how much for university water does
Topic 31: all rights reserved all rights rights reserved comes it that it that quiz
Topic 32: business programs essentials business essentials in business programs in wednesday age on wednesday febru
ary
Topic 33: sign up you sign up this email was to you was this you sign
Topic 34: brands dow game music video thanks games thanks to that the game
Topic 35: are that their out have people with companies for they
Topic 36: into back into the how on bring come office brand that
Topic 37: work take brew the brew to work take the brew to to take chance chance to
Topic 38: after on with two disney that for news at last
Topic 39: morning brew morning every brew valley about the week the week quick silicon valley
Topic 40: make to make same now the same your course you one invest
Topic 41: on based as number for half an is members number of
Topic 42: have been bitcoin have been best are russia ukraine list the best
Topic 43: neal neal freyman freyman matty matty merritt merritt 2023 check rubenstein morning
Topic 44: her was years she nasdaq num years who for police prison
Topic 45: new city the new new york york state for school washington year
Topic 46: you see can find real quote are word that there
Topic 47: become workers strike movie union is street about tv with
Topic 48: president that biden court for trump on against is case
Topic 49: for its markets first that yesterday since after investors the first
```

Figure A.1: Morning Brew Only Topic Model

```
Topic  0: sunday did paul question columns cup music biggest living local
Topic  1: season game num team age getting athletic football players night
Topic  2: workers care union strike killing jobs hollywood child writers didn
Topic  3: schools action modern based prime minister great british netanyahu london
Topic  4: argues children latest politics share hot soon storm families parents
Topic  5: trump jan donald charges prosecutors indictment hosts case election special
Topic  6: don know think going watch feel need like residents los
Topic  7: book man israel novel critics author tell decade conservative books
Topic  8: lauren jackson ian tom words hard prasad reach philbrick claire
Topic  9: news quiz week issue headlines fox followed sea ohio read
Topic 10: star good makes set want plan job senator needs democrat
Topic 11: old story shows year num fashion second accused came stars
Topic 12: life way art tiktok men line important spring bring focus
Topic 13: killed people texas death num shot medical hours said seen
Topic 14: page court 2022 today supreme judge case decision ruling law
Topic 15: days emily seven outside looking stay end close longer making
Topic 16: use better food work young fall ll today explain near
Topic 17: try billion thousands covers events major lopez german writer affect
Topic 18: trying right video india far buy cases winter games intelligence
Topic 19: ukraine russia war military ukrainian putin western troops vladimir east
Topic 20: wordle thanks spending tomorrow morning times today sudoku mini crossword
Topic 21: num young help general gave games game future friends friday
Topic 22: voters party election elections campaign 2024 republican trump popular presidential
Topic 23: new york city times help books using rules doing person
Topic 24: social media inflation prices prison ways george turn security avoid
Topic 25: company travel world john tour early building border southern created
Topic 26: read covid officials july pandemic chinese deaths michael continue issues
Topic 27: home million num look face violence nearly agreed justice department
Topic 28: biden daily president black water related critic street writes wall
Topic 29: police source health energy turkey brazil role film played public
Topic 30: change russian climate away things looks force used isn course
Topic 31: best wirecutter advice air heat family mass town florida clean
Topic 32: debt bank crisis drug limit money financial woman government growing
Topic 33: david leonhardt times editor podcast staff host magazine prize bureau
Topic 34: republican support debate tech democratic lost sex wants primary difficult
Topic 35: abortion administration rights biden james access legal policy different approach
Topic 36: clue letters crossword love mini today business movies short economy
Topic 37: died lives lived num known open french reporter tournament won
Topic 38: white states plans united fighting museum potential millions efforts research
Topic 39: washington desantis post history reports risk ron general governor gun
Topic 40: num years play percent 2021 ago wanted half rates rate
Topic 41: melissa south america kirsch culture holiday investigation global guide chicken
Topic 42: 2020 list twitter students data musk college documents classified red
Topic 43: says german lopez hit industry turned executive small chief place
Topic 44: california day start return university north readers reported britain office
Topic 45: bee yesterday spelling today pangram puzzle pangrams cities tv let
Topic 46: fans technology taking weather running held problems gave does changes
Topic 47: summer friday coming movie online released center announced saying places
Topic 48: china mexico live leader cover stories comes protests sign sports
Topic 49: house 2023 republicans women mccarthy west fight kevin speaker democrats
```

Figure A.2: NYT Only Topic Model

```
Topic  0: mexico china note nichols liquidpiston from memphis officers realpha mobile
Topic  1: turntable in george play santos hunt num choose game home
Topic  2: page front 2022 today here everything else offering read advertisement
Topic  3: the num in of to with and on for musk
Topic  4: podcasts interested note in journal founder from facet ill fidelity
Topic  5: workout va republican dec wells adventure savings li yield fargo
Topic  6: july prigozhin wagner russia belarus canada pakistan wounded organized yevgeny
Topic  7: wirecutter advice from best nvidia affirmative admissions your the weight
Topic  8: daily kid copy paste _code others referral link or your
Topic  9: bitcoin monogram alaska independent egypt pool no knee legally change
Topic 10: ours check out here 2010 member bank hawaii calif energy
Topic 11: _count count bee referral spelling your pangram yesterday today puzzle
Topic 12: nasdaq data as cryptocurrency et close stock 00am market of
Topic 13: leadership thanks spending part see times tomorrow your you with
Topic 14: lopez german leonhardt david by the times writer flagship more
Topic 15: quiz ace satisfying cruz gilbert fox the by news carlson
Topic 16: num search year bps million tesla deaths gulf per okla
Topic 17: written by 2020 bot our better use wordle get books
Topic 18: bret stephens thomas editorial discuss menendez haley innovation supported by
Topic 19: mass fi feb states netflix clinton wi united springs houston
Topic 20: chicken recipes guess the black listing cooking welcome open facts
Topic 21: money katie with show 1960 the finance your scoop brazil
Topic 22: dow international democrat port holmes winners elizabeth photography nigeria theranos
Topic 23: saudi arabia ice golf amazon cream pga tour liv sleep
Topic 24: day updates valley st silicon delivers insightful quick wall every
Topic 25: the num all reserved rights to of in and its
Topic 26: the to you and of it your for in is
Topic 27: neal freyman matty merritt rubenstein klebanov sam cassandra cassidy youtube
Topic 28: email forwarded sign up was this you to tips living
Topic 29: cup women world the france soccer spain tournament australia argentina
Topic 30: analytics and melissa clark by kirsch mitch mcconnell julian hispanic
Topic 31: apple 1940 moon japan republic boston nasa rail bay portugal
Topic 32: the in to of trump his and he on was
Topic 33: the to of in and that is for it num
Topic 34: numbers mean these what here digit israel africa netanyahu minister
Topic 35: casual missing business india pamela paul latino festival democracy peru
Topic 36: 2023 inc morning brew 1980 poland walker pope francis conservative
Topic 37: essentials programs business in courses accelerate career our your with
Topic 38: stream hip maine hop nota cytonics liquidpiston engines engine movies
Topic 39: voters utah registered flooding medium bear palestinian all tropical initiatives
Topic 40: today here wordle mini crossword and link we you sudoku
Topic 41: lauren jackson ian prasad philbrick ashley claire wu moses contributed
Topic 42: read full issue 2000 2015 the ukraine ohio nate cohn
Topic 43: los angeles 2002 offer san facet jackpot securities may other
Topic 44: york new times by the 2019 vows of wash population
Topic 45: here brands billion pence mike discussed britain hosts shallow tastytrade
Topic 46: brew work take minutes smarter get just to the num
Topic 47: sponsored content advertising this is white 2016 markets alibaba by
Topic 48: bs education without business 2021 the metropolitan diary odd faker
Topic 49: word rewards asian submit buffalo sept miss day hamlin fla
```

Figure A.3: Topic Model With Default Paragraph Inputs

```
Topic  0: your you to can how get with it help our
Topic  1: neal sign freyman matty written email merritt rubenstein minutes sam
Topic  2: the party political republican voters republicans election democrats in democratic
Topic  3: num your cash card you and with credit for get
Topic  4: the james awards tonight of golden in and baseball playoffs
Topic  5: the secret wedding sex to air marriage quality of in
Topic  6: russia putin the gun ukraine violence in vladimir russian to
Topic  7: los angeles the num bills la plane in crash buffalo
Topic  8: police in man her was prison she arrested death old
Topic  9: women the cup team world soccer in men moon of
Topic 10: the war of and germany in on europe ii columns
Topic 11: business brew morning wall every 2023 rights valley brands silicon
Topic 12: num data million year close as market stock 2022 numbers
Topic 13: ai chatgpt the to chatbot artificial san openai and francisco
Topic 14: german lopez events newsletter covers the affect world major how
Topic 15: play content sponsored advertising desantis is you name the this
Topic 16: school students action university the percent schools college of black
Topic 17: trump his the court he case to former donald president
Topic 18: the big picture growing revenue to apple mm sales in
Topic 19: chicken and the hair emily with thanksgiving sauce recipes of
Topic 20: abortion states access ban the to in state fda approval
Topic 21: the sense to roe and of mortgage wade are gains
Topic 22: leader the xi press jinping hosts to conference of stephen
Topic 23: lauren hard front reach jackson page the tom ian prasad
Topic 24: water california the in storm december of and christmas winter
Topic 25: the house biden to government bill debt mccarthy congress israel
Topic 26: music the her of pop song museum songs and in
Topic 27: the its inflation prices markets rate to in fed since
Topic 28: the city of in shooting people to and mayor killed
Topic 29: the george fashion and of to santos airlines airport paris
Topic 30: ukraine russia the in russian war of ukrainian western military
Topic 31: real estate the art cover and of housing saving in
Topic 32: china the chinese countries us and to india that secretary
Topic 33: the king charles in luxury of on to succession and
Topic 34: car electric invest the cars to and engine num uber
Topic 35: link we referral give ll share others saying friends your
Topic 36: the include open netflix to square price welcome foot house
Topic 37: book energy prime books climate minister oil the gas clean
Topic 38: tax tour the swift taylor missing rich saudi to tickets
Topic 39: google microsoft the to of in generation and betting industry
Topic 40: today here see part bee spending tomorrow thanks crossword times
Topic 41: twitter musk media social tiktok users the elon company app
Topic 42: fox the super network ice bowl coach in and of
Topic 43: movie the disney film of movies tv and barbie hollywood
Topic 44: we it that you my to they what do so
Topic 45: financial bank investment offer the value banks facet of sell
Topic 46: health the covid of and are to in people have
Topic 47: the num died lives lived game he at his in
Topic 48: david times washington writes the leonhardt chief post co podcast
Topic 49: the workers to that strike union pay could labor will
```

Figure A.4: Topic Model with 200-Token Chunk Inputs

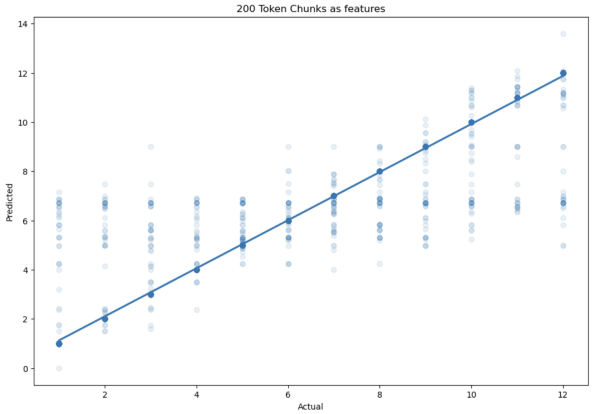

Figure A.5: Scatterplot Showing 200-Token Informed, Token-Based Month Predictions

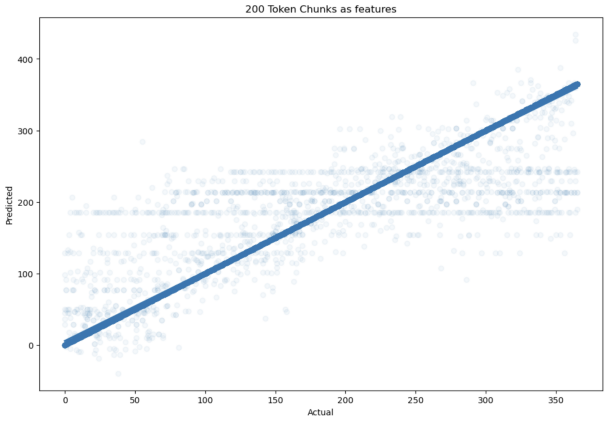

Figure A.6: Scatterplot Showing 200-Token Chunk Based Predictions Against Actual Values