International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2024, 8 (2), 163-170
DOI: 10.36287/ijmsit.8.2.18
Received: November 8, 2024; Accepted: December 21, 2024

Word Frequency: New York Times Throughout the Times

Mehmet Aşıroğlu[1*] and Emre Atlıer Olca[2]

[1*]*Üsküdar American Academy, Istanbul, Turkey, (mehmet.asiroglu07@gmail.com) (ORCID: 0009-0006-1883-2245)*
[2]*Software Engineering Department, Maltepe University, Istanbul, Turkey (emreatlier@gmail.com) (ORCID: 0000-0001-6812-5166)*

Abstract – This paper investigates the evolution of the English language over the past century using a machine learning model trained on leading articles from The New York Times spanning from 1920 to 2020. The primary aim is to predict the year in which a given sentence could have been written based on linguistic patterns, including word usage and sentence structure. By analyzing these patterns, the model provides insights into the changing styles and trends in written English over time. The model's predictions are grounded in extensive data analysis and machine learning techniques, ensuring a high degree of accuracy. This study not only highlights the dynamic nature of language but also demonstrates the application of computational methods in linguistic research. The findings of this research are significant for historical linguistics and literature studies, as they provide a quantifiable method to track linguistic changes. Additionally, this work can aid in the development of tools for temporal text classification, benefiting fields such as digital humanities and archival studies. Understanding how language evolves is crucial for preserving cultural heritage and improving communication strategies in different communication platforms.

Keywords *– language evolution, machine learning, historical linguistics, text analysis, computational linguistics*

Citation: Aşıroğlu, M., Olca, E. (2024). Word Frequency: New York Times Throughout the Times, International Journal of Multidisciplinary Studies and Innovative Technologies, 8(2): 163-170.

I. INTRODUCTION

Language is constantly changing, reflecting the cultural, social, and technological shifts that occur over time. As societies grow and evolve, so does the way they communicate. This evolution is particularly noticeable in written language, where changes in word choice, sentence structure, and writing style can be observed across different periods. Understanding these changes can offer valuable insights into historical events and broader societal trends, helping us better understand how communication develops alongside human progress.

This paper explores the evolution of the English language over the past century by analyzing leading articles from The New York Times, published between 1920 and 2020[7]. The New York Times, as a prominent publication, has documented key events and cultural shifts throughout the last hundred years, making it an ideal source for studying changes in language. By examining these articles, we aim to identify and measure shifts in writing style, word usage, and sentence structure, providing a detailed view of how journalistic language has changed over time.

The core of this research is a machine learning model designed to predict the publication year of a given sentence. This model was trained using a large dataset of New York Times articles, learning from patterns in word choice, sentence length, and sentence structure. By doing so, the model captures subtle changes in language that might otherwise go unnoticed, giving us a clearer picture of how language usage in journalism has evolved over the years.

The model works by first analyzing the input sentence, identifying important linguistic features, and then predicting the most likely year of publication based on the patterns it has learned. This predictive capability not only demonstrates the effectiveness of the model but also introduces a new way to study linguistic change. By using advanced computational methods, our research offers a fresh approach to examining how language evolves over time.

The findings from this study have implications beyond linguistics. They can be applied in literature studies, where understanding the historical context of language can enhance literary analysis. Additionally, in digital humanities, this research contributes to the development of tools that can classify texts by their time period, helping with the preservation and analysis of historical documents. Overall, this study provides a new perspective on the evolution of language, highlighting the close connection between language, society, and time.

The paper is structured into several sections, each detailing crucial aspects of the study on the evolution of language. Section 1 provides an introduction to the research, outlining the significance of studying language change and the methods used. Section 2 reviews related works, comparing and contrasting similar studies on language evolution and predictive modeling. Section 3 delves into the development of the linear SVC-trained language model, explaining the methodology behind predicting the origin date of a sentence based on linguistic features. Section 4 presents the results, supported by graphs, and discusses the findings in the context of language change over the past century. Finally, Section 5 concludes the paper by summarizing the key insights and discussing potential future research directions.

II. MATERIALS AND METHOD

Several studies have explored the evolution of written language and its stylistic changes over time. This section discusses six significant works that relate to our research, highlighting their methodologies and findings, and comparing them to our approach.

In the study "Modeling the Development of Written Language"(Wagner at all., 2011)[1], the authors used confirmatory factor analysis to test different models of written composition and handwriting fluency among first- and fourth-grade students. The study identified five key factors affecting written composition: macro-organization, productivity, complexity, spelling and punctuation, and handwriting fluency. The correlation between handwriting fluency and written composition factors was examined, revealing significant developmental differences between the two grade levels. While this study focuses on early developmental stages of writing skills, our research differs by analyzing a broader timespan of 100 years and emphasizing the evolution of language in published media. Moreover, our model uses machine learning techniques to predict the temporal origin of sentences based on linguistic patterns, rather than developmental differences in young writers.

The article "Change and Constancy in Linguistic Change: How Grammatical Usage in Written English Evolved in the Period 1931-1991"(Geoffrey and Nicholas, 2009)[2] examines the evolution of grammatical usage in British English through the Lanc-31 corpus, a trio of corpora spanning 1931, 1961, and 1991. By analyzing frequency counts of various grammatical features, the study identifies trends of increasing or decreasing usage, providing insights into grammaticalization, colloquialization, Americanization, and densification. This research closely aligns with our project, as both studies aim to trace linguistic changes over an extended period. However, our approach focuses on American English and employs machine learning to predict the publication year of sentences from The New York Times articles between 1920 and 2020. Additionally, while the Lanc-31 corpus provides a static analysis of grammatical features, our model dynamically predicts temporal origins based on a combination of word usage, sentence length, and syntactic patterns, offering a more comprehensive understanding of language evolution in journalistic writing.

Another relevant work is the research titled "Learning to Predict U.S. Policy Change Using New York Times Corpus with Pre-Trained Language Model."(Zhang at all., 2020)[3]. This study focuses on predicting policy changes in the United States by analyzing large-scale news data from The New York Times. The researchers built a comprehensive news corpus covering the period from 2006 to 2018 and fine-tuned the pre-trained BERT language model [9] to detect shifts in newspaper priorities, which they argue correspond to changes in U.S. policy. The study introduces a BERT-based Policy Change Index (BPCI)[15] to measure these changes, offering a novel approach to understanding and predicting policy shifts based on media analysis. This research closely relates to my project in that it also leverages machine learning and large-scale textual data from The New York Times. Both studies aim to uncover patterns and trends over time by analyzing language usage. However, while their focus is on predicting policy changes, my research is centered on examining the broader evolution of language in journalistic writing. Where their model seeks to identify specific policy shifts based on news priorities, the model that is presented in this study is designed to predict the publication year of sentences by analyzing linguistic features. This difference highlights how similar methodologies can be adapted to address distinct research questions, demonstrating the versatility and potential of machine learning in the study of language and its applications.

In the study titled "A Framework for Analyzing Semantic Change of Words Across Time"(Adam and Kevin, 2014)[4] the authors present a comprehensive approach to understanding how the meanings of words evolve over extended periods. The framework they propose utilizes word representations from distributional semantics to explore lexical changes at various levels, including individual word meaning, contrastive word pairs, and sentiment orientation. Their method allows for a detailed analysis of semantic transitions by leveraging large-scale diachronic corpora, which enables the visualization of a word's evolution over time. This research is particularly relevant for fields such as computational linguistics, historical linguistics, and natural language processing (NLP)[12], where understanding semantic change is crucial. The work relates to our project in its focus on language change over time, specifically through the lens of semantic evolution. Both studies utilize large datasets and computational methods to analyze linguistic trends. However, while their research is centered on the semantic shifts of individual words, our study takes a broader approach by examining changes in writing style, word usage, and sentence structure within journalistic writing over a century. The primary difference lies in the level of analysis: theirs is focused on word-level semantics, while ours considers sentence-level patterns and temporal trends. Additionally, their framework is designed to provide visual insights into word evolution, whereas our model aims to predict the publication year of sentences based on linguistic features. Despite these differences, both projects share a common goal of advancing our understanding of language change through computational means.

The research titled "Measuring News Sentiment"(Shapiro at all., 2017)[5] presents a novel approach to assessing economic sentiment by extracting it directly from newspaper articles rather than relying on traditional survey-based measures. The study introduces a news sentiment index developed using computational text analysis on a large corpus of economic and financial news articles. The researchers employ sentiment-scoring models, primarily utilizing lexical techniques, to analyze sentiment within these articles. By combining existing lexicons and creating a new lexicon tailored specifically for economic news, the study enhances the accuracy of sentiment prediction, achieving a rank correlation of approximately 0.5 with human ratings. The result is a national time-series measure of news sentiment, labeled the "News PMI model"[13] which correlates strongly with survey-based consumer sentiment indexes and is used to predict macroeconomic[16] outcomes. This work is relevant to our research as both studies leverage large datasets of news articles to analyze linguistic patterns over time. While their focus is on measuring sentiment in economic news and its impact on macroeconomic variables, our study examines broader linguistic changes, such as word usage and sentence structure, to predict the publication year of journalistic content. The primary similarity lies in the use of text analysis and machine learning techniques to extract meaningful information from large corpora. However, the key difference is that their project

is centered on sentiment analysis and its implications for economic forecasting, whereas our research is focused on tracking linguistic evolution and temporal shifts in language use within the context of news media. Despite these differences, both studies contribute to the growing body of work that uses computational methods to derive insights from textual data.

The research project titled "Understanding the Influence of News on Society Decision Making: Application to Economic Policy Uncertainty"(Trust at all., 2023)[6] focuses on the use of digital text data to analyze the impact of news on economic decision-making. With the rise of digital documentation, ranging from social media posts to news articles, the study explores how computational methods can be employed to understand the correlation between language usage and economic policy uncertainty (EPU). The project builds upon the Economic Policy Uncertainty (EPU)[14] index developed by Baker et al., which uses keyword-based methodologies to extract EPU-related news articles. However, this traditional approach is prone to false positives and negatives, prompting the need for more advanced techniques. To address these challenges, the authors propose a novel approach using weak supervision combined with neural language models, specifically BERT, for the automatic classification of news articles related to EPU. This method reduces the false positive rate significantly compared to traditional keyword-based approaches and is more efficient and cost-effective than fully supervised methods, which require extensive manual annotation. The study also introduces an Irish weak supervision-based EPU index and demonstrates its predictive power through econometric analysis with Irish macroeconomic indicators. This project shares similarities with our research in its use of computational techniques to analyze large-scale textual data for understanding broader social and economic phenomena. Both studies leverage machine learning models to process and categorize textual information, albeit with different goals. While their project focuses on extracting economic signals from news articles and predicting macroeconomic indicators, our study centers on tracking linguistic changes over time to predict the origin date of journalistic content. Both approaches highlight the importance of text analysis in deriving insights from digital documents, and their use of machine learning models to enhance the accuracy and efficiency of these analyses is directly relevant to our work.

An example of the table is given below.

Table 1. Content of different studies vs this study

Study/Project	Usage of Model	Usage of Dataset	Includes NLP	Tracks Change Over Time	Focuses on Word Usage	Sentiment Analysis	Predictive Analysis
1. Modeling the Development of Written Language	X	✓	X	✓	✓	X	X
2. Change and Constancy in Linguistic Change: How Grammatical Usage Evolved	✓	✓	✓	✓	✓	X	X
3. Learning to Predict U.S. Policy Change Using NYT Corpus	✓	✓	X	✓	✓	✓	✓
4. A Framework for Analyzing Semantic Change of Words Across Time	✓	✓	✓	✓	✓	X	X
5. Measuring News Sentiment	✓	✓	✓	X	X	✓	✓
6. Understanding the Influence of News on Society Decision-Making (EPU)	✓	✓	✓	X	X	✓	✓
Word Frequency: New York Times Throughout The Times	✓	✓	✓	✓	✓	X	✓

III. RESULTS

A. The Purpose of the project

The primary objective of this work is to build a language model capable of accurately classifying which decade a given sentence was likely written in. By doing so, we aim to establish a foundation for leveraging machine learning models in the analysis of language evolution over time. This project is driven by the need to understand how word usage, sentence structure, and overall linguistic trends change across different periods, offering insights into cultural, societal, and historical transformations reflected in written texts.

The contribution of this work lies not only in developing a functional model for decade classification but also in creating a scalable framework for future research. The model can be expanded to encompass broader linguistic features and larger datasets, eventually leading to more precise predictions and a deeper understanding of the mechanisms behind language change. Furthermore, this project could serve as a springboard for applications in digital humanities, historical linguistics, and automated text analysis, where tracking language shifts over time can provide valuable context for interpreting historical documents and understanding linguistic innovation.

We present the development of a Linear Support Vector Classifier (SVC)[10] model to predict the origin date of a sentence based on New York Times article titles[7] from 1920 to 2020. We leverage text data processing techniques, including TF-IDF (Term Frequency-Inverse Document Frequency[8]), to transform text into numerical vectors, which serve as input to our machine learning model.

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic used to reflect the importance of a word in a document relative to a collection of documents or corpus. It is calculated by multiplying two components: Term Frequency (TF), which measures how often a word appears in a specific document, and Inverse Document Frequency (IDF), which gauges how common or rare the word is across all documents. The TF-IDF score increases with the number of times a word appears in a document while decreasing proportionally to the frequency of the word across the corpus. This approach highlights terms that are unique to each document, thus improving the model's ability to distinguish between different decades based on linguistic trends. TF-IDF is chosen in this study over other similar methods such as count vectorizer[17] or word embeddings[18] because it efficiently balances the significance of frequently occurring terms while downweighting common words that might not contribute to meaningful differentiation between decades. This results in a more refined feature set that enhances the performance of the Linear SVC model in predicting the temporal context of a sentence.

Our dataset is stored in Parquet format which is available on kaggle[8], which offers efficient storage and retrieval for large datasets, especially in structured or tabular data formats.

The process began with a clearly defined plan:

- load the dataset,
- preprocess the data to eliminate irrelevant entries,
- group the data by decades to reduce prediction complexity.

We intended to use a Linear SVC, a robust classification algorithm[19] well-suited for high-dimensional datasets like text data, and TF-IDF for feature extraction . By splitting the data into training and testing sets, we ensure model performance can be measured accurately. The code also allows users to input a new sentence and predict its likely decade.

B. Project Development Process

The development of this language model involved several iterations aimed at refining the model's performance and improving its ability to predict the decade in which a sentence was likely written. Initially, we began by training the model on a smaller dataset of New York Times article titles, which resulted in lower accuracy. The early trials used a random selection of fewer than 1,000 article titles per year. This yielded an accuracy of only 9%, largely due to the small training size and the lack of sufficient features for the model to generalize effectively.

We increased the sample size over time, eventually settling on 5,000 titles per year, as mentioned in the earlier code discussion. This improved the model's accuracy to 52%, indicating a more reliable relationship between the textual features of the article titles and their publication decade. The accuracy improvements were achieved by refining several steps, such as tweaking the TF-IDF vectorizer, using different random states, and optimizing the Linear SVC hyperparameters.

The code development in detail, from initial data processing to training the model and making predictions is explained as below.

Table 2. Getting data from New York Times

```
df = pd.read_parquet('/content/drive/MyDrive/
columbia-ml-data/nyt_data.parquet')
df = df.drop('excerpt', axis=1)
```

The code that is given in Table 2 starts by loading a dataset in Parquet format containing New York Times article titles from 1920 to 2020. The ‘drop('excerpt', axis=1)’ line removes the ‘excerpt’ column because it contains many empty values, which could confuse the model and negatively impact performance. By focusing on non-empty, relevant data (i.e., the article titles), the model is given cleaner, more consistent input.

Table 3. Grouping samples

```
sample=df.groupby('year').sample(n=5000,random_state
=84)
sample2 = sample['year'] // 10 * 10
sample2 = pd.concat([sample2, sample['title']], axis=1)
```

In table 3, the dataset is grouped by the year of publication, and a random sample of 5,000 titles per year is selected. The ‘random_state’ parameter ensures that the sampling process is reproducible. By dividing the years into decades (‘year // 10 * 10’), you simplify the prediction task: instead of predicting an exact year, the model predicts the decade in which a sentence might have been written. This reduces the complexity of the task and potentially increases the model’s accuracy.

Table 4. Labeling the axes

```
X = sample2['title']
y = sample2['year']
```

Table 4 shows the titles of the articles (‘X’) are used as the features, and the corresponding decades (‘y’) are used as the labels. These are the inputs and outputs that the model will learn to map during training.

Table 5. Splitting the data for train and test

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

In the code given in table 5, the dataset is split into training and testing subsets, with 80% of the data used for training the model and 20% reserved for testing its performance. The ‘random_state=42’ parameter ensures that the split is consistent each time the code is run, which is important for reproducibility.

Table 6. Using TF-IDF

```
tfidf = TfidfVectorizer()
X_train_vec = tfidf.fit_transform(X_train)
X_test_vec = tfidf.transform(X_test)
```

The TfidfVectorizer converts the text data into numerical vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) method in table 6. This process transforms the words in the article titles into a form that the model can process. The ‘fit_transform’ method is applied to the training data, while ‘transform’ is used on the test data, ensuring that the same transformation is applied consistently across both

datasets.

Table 7. Choosing the model

```
model = LinearSVC()
model.fit(X_train_vec, y_train)
```

In table 7, a Linear Support Vector Classifier (LinearSVC) is instantiated and trained on the vectorized training data ('X_train_vec') and the corresponding labels ('y_train'). The SVC is a popular choice for text classification tasks due to its effectiveness in high-dimensional spaces, such as text data.

Table 8. Using predict

```
y_pred = model.predict(X_test_vec)
```

The trained model is used to predict the decades for the titles in the test dataset in table 8. These predictions ('y_pred') will be compared against the actual decades ('y_test') to evaluate the model's performance.

Table 9. Accuracy score

```
print(accuracy_score(y_test, y_pred))
```

Table 9 shows the 'accuracy_score' function calculates how often the model's predictions match the actual labels. This gives you a sense of how well the model is performing in terms of predicting the correct decade for unseen data.

Table 10. Testing the model

```
new_sentence = ["insert input here"]
new_sentence_vec = tfidf.transform(new_sentence)
predicted_year = model.predict(new_sentence_vec)
print(predicted_year)
```

This code in table 10 allows you to input a new sentence and predict the decade in which it might have been written. The input sentence is vectorized using the same TF-IDF model ('tfidf.transform(new_sentence)') and then passed through the trained SVC model to obtain the predicted decade ('predicted_year').

This code implements a linear SVC model trained on New York Times article titles to predict the decade in which a given sentence could have been written. The process includes loading and cleaning the data, sampling and grouping by decade, vectorizing the text data, training the model, and evaluating its accuracy. Additionally, the model can be used to predict the decade of new, unseen sentences.

C. *Accuracy Score and Its Limitations*

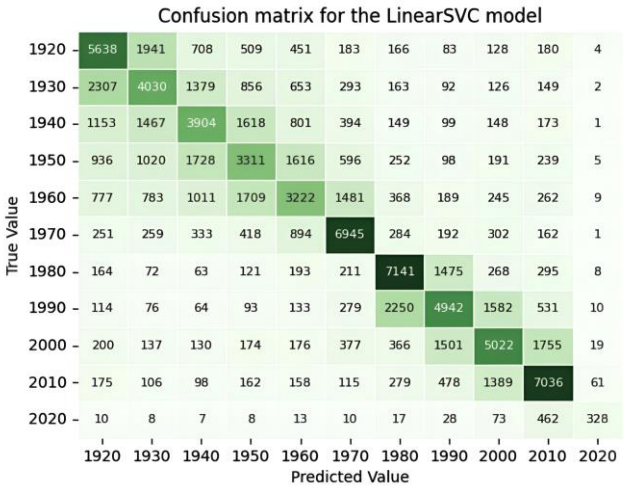
While an accuracy[20] of 52% is a significant improvement over the initial trials, it remains relatively low for real-world predictive applications. There are several reasons for this limitation. One major challenge lies in the overlapping nature of language use across decades. For example, while certain terms or phrases might be indicative of a specific era, many words and sentence structures remain constant or evolve slowly over time. This causes confusion in the model, especially for articles written in decades that are close to each other in time.

Another reason is the variability in the dataset. The New York Times articles cover a wide range of topics, from politics and economics to culture and science. Each topic may have its own unique linguistic features, and mixing them in a general model can dilute the predictive power. This variability contributes to the model's lower accuracy, as it must generalize over a vast array of subjects and styles.

IV. DISCUSSION

A. *Confusion Matrix Explanation*

Fig. 1. Confusion matrix for the model



The confusion matrix[11] shown in Figure 1 above provides a more detailed look at how well the model is performing across different decades. The matrix helps identify where the model tends to make incorrect predictions and gives insights into which decades are most easily confused with each other.

Diagonal values: These represent the correct predictions made by the model. For example, in the row corresponding to the 1920s, we see that 5,638 titles were correctly predicted to be from the 1920s. However, as we move down the diagonal, we notice that the number of correct predictions decreases for later decades, which is expected due to the increasing overlap in language use as we approach the present.

Off-diagonal values: These represent incorrect predictions, where the true value lies in one decade, but the model predicts another. For instance, the model frequently confuses the 1930s and 1940s, with 1,153 titles from the 1940s being incorrectly classified as from the 1930s. This indicates that the language in these two decades shares many similarities, making it

difficult for the model to distinguish between them.

Interestingly, the model performs much better for more recent decades, such as the 1980s and 2010s, where the diagonal values are much higher, indicating more accurate predictions. This suggests that the changes in language during these decades are more distinguishable from earlier periods, possibly due to the influence of modern technology and communication patterns.

B. Future Direction and Potential Applications

While the current model achieves moderate success, it sets the foundation for future work aimed at improving predictive accuracy and extending the analysis to other linguistic features, such as grammar patterns or word frequencies.

The long-term goal is to create a model that not only predicts the decade of a sentence but also identifies more granular linguistic shifts, such as the adoption of specific phrases or syntactic structures. Such a model could be useful in historical linguistic research, helping scholars track the evolution of language in various fields.

This study's model offers several promising applications in digital humanities and related fields. One notable contribution is its potential to aid archival research by serving as a temporal text classification tool for historians. By accurately predicting the decade a piece of text originates from, the model could help researchers organize large, unstructured corpora of historical documents, making it easier to identify patterns, trends, and shifts in language use over time. This would be particularly useful for analyzing underexplored periods or detecting chronological inconsistencies in archives.

In education, the model could be used to teach students about linguistic evolution and historical context. By analyzing texts from different decades, learners could explore how societal changes influenced language use, gaining insights into both history and linguistics. Additionally, this approach could enhance the development of historical language models, which often struggle to account for the nuances of older texts. Incorporating temporal context from models like this could improve their ability to process and understand historical documents, expanding their utility in fields like computational linguistics and cultural heritage preservation.

C. Example Predictions

To further illustrate how the model operates, let's consider a few sample predictions:

Input sentence: "The government enacted new legislation to regulate financial markets."

Predicted decade: 1980

Reasoning: The language model likely associates the terms "financial markets" and "regulate" with the economic policies and financial reforms of the 1980s.

Input sentence: "New technologies are shaping the future of communication."

Predicted decade: 2000

Reasoning: The prominence of "new technologies" and "communication" in this sentence aligns with the digital revolution and the rise of the internet, which began in earnest in the early 2000s.

These examples demonstrate the model's ability to link certain phrases and keywords with specific time periods. However, they also highlight some limitations. For example, if a sentence were written in a more ambiguous style, the model might struggle to provide an accurate prediction.

D. The Reasoning Behind TF-IDF

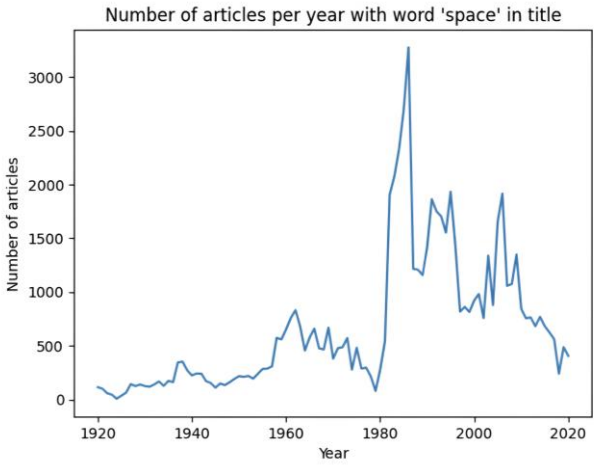
TF-IDF was selected as the feature extraction method for its simplicity, interpretability, and alignment with the study's objectives. Unlike advanced word embeddings such as Word2Vec or GloVe, which focus on capturing contextual semantics, TF-IDF emphasizes the relative importance of terms within the corpus. This makes it particularly suited for analyzing lexical and stylistic shifts over time, as it highlights changes in word usage patterns without introducing semantic complexities.

TF-IDF is computationally efficient, allowing the analysis of a large dataset like The New York Times corpus without excessive resource demands. While word embeddings could capture deeper relationships between words, they often introduce complexity and context that might obscure the stylistic and lexical trends this study seeks to investigate. Future research could include a comparative analysis of TF-IDF and embedding-based approaches to evaluate their respective impacts on capturing historical language trends and predicting temporal origins.

E. Cultural and Societal Factors

The model's accuracy variations across decades can be attributed to linguistic trends shaped by cultural and societal factors, which are often reflected in The New York Times' coverage of trending topics.

Fig. 2. Year by Number of Articles Graph



For example, as seen in figure 2, the sharp rise in the use of the word "space" during the 1960s and 1970s corresponds with the Space Race and the Apollo program, signaling a societal focus on space exploration. During these periods, the vocabulary, syntactic patterns, and frequency of related terms became more distinct, providing the model with stronger temporal markers and improving its predictive accuracy.

Conversely, in decades where trends are less sharply defined or topics overlap significantly with prior periods, the model may struggle to distinguish linguistic shifts, leading to reduced accuracy. For instance, the decline in articles about "space" after the 1980s reflects a cultural shift in priorities, resulting in less distinct linguistic signals for the model to learn from.

Since the corpus focuses on trending topics, it mirrors societal attention spans rather than steady linguistic evolution, which may introduce biases in the model. Fluctuations in accuracy highlight how the temporal coverage of events influences the distinctiveness of language features, reinforcing the importance of contextualizing model performance within cultural and historical dynamics.

F. Sampling Bias and Dataset Limitations

The dataset for this study was curated by selecting 5,000 article titles per year from the New York Times archive, creating a balanced representation of textual data across a century. While this sampling method ensured consistency in dataset size for each year, it may have introduced certain biases. The New York Times, as a major publication, tends to prioritize topics and writing styles that reflect its readership and editorial focus, potentially overrepresenting subjects of significant cultural, political, or economic importance at the expense of less mainstream topics.

This selection method also raises concerns about the diversity of linguistic styles captured in the dataset. Since the New York Times typically employs formal journalistic language, the model may have limited exposure to informal, regional, or genre-specific linguistic patterns that are also part of language evolution. Consequently, the model's predictions and insights may be skewed toward reflecting trends and styles unique to the New York Times rather than broader linguistic shifts.

Recognizing this limitation, future research could incorporate additional sources, such as regional newspapers or less formal publications, to create a more diverse and representative dataset. This would not only mitigate potential bias but also improve the generalizability of the model's findings to a wider range of linguistic contexts.

V. CONCLUSION

In conclusion, this study presents the development of a machine learning-based language model designed to predict the decade in which a given sentence from New York Times articles was written. We employed a Linear Support Vector Classifier (SVC) and used TF-IDF (Term Frequency-Inverse Document Frequency) to convert text into numerical vectors, focusing on analyzing language evolution over time. The core goal was to track shifts in word usage, sentence structures, and linguistic trends between 1920 and 2020, thereby offering insights into how language reflects cultural, societal, and historical changes.

Our methodology, which included data cleaning, decade-based grouping, and feature extraction using TF-IDF, allowed us to significantly improve the model's performance over time. By expanding the dataset to 5,000 titles per year and optimizing the model's parameters, we achieved an accuracy of 52%. Although this accuracy is moderate, it marks a substantial improvement from initial trials and demonstrates the potential of machine learning in understanding language change. TF-IDF was particularly effective in highlighting distinctive features that aided in decade classification, a decision that proved valuable in making the model more efficient.

Despite these advancements, several limitations persist. The overlap in language use across decades, especially between adjacent periods like the 1930s and 1940s, presented challenges for the model. Additionally, the wide range of topics covered by New York Times articles made it difficult for the model to generalize across different subjects. However, the model showed higher accuracy in more recent decades, suggesting that language changes were more discernible during periods of technological and communicative shifts.

Looking forward, future research should focus on incorporating additional linguistic features such as grammar patterns and word frequency analysis to improve predictive power. Expanding the dataset to include other text sources, like literature or social media, could offer a broader perspective on language evolution. Additionally, refining evaluation metrics beyond accuracy could help identify areas where the model struggles and provide more detailed insights into misclassifications.

While the current model is a foundational step, it opens the door to more advanced applications in digital humanities and historical linguistics. By tracking language shifts over time, future studies could offer deeper insights into how language adapts to societal changes, providing valuable tools for historians, linguists, and researchers in the field.

REFERENCES

[1] Wagner, Richard K., et al. "Modeling the development of written language." Reading and writing 24 (2011): 203-220.

[2] Leech, Geoffrey, and Nicholas Smith. "Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991." Corpus Linguistics. Brill, 2009.

[3] Zhang, Guoshuai, et al. "Learning to predict US policy change using New York Times corpus with pre-trained language model." Multimedia Tools and Applications 79 (2020): 34227-34240.

[4] Jatowt, Adam, and Kevin Duh. "A framework for analyzing semantic change of words across time." IEEE/ACM joint conference on digital libraries. IEEE, 2014.

[5] Shapiro, Adam Hale, Moritz Sudhof, and Daniel J. Wilson. "Measuring news sentiment." Journal of econometrics 228.2 (2022): 221-243.

[6] Trust, Paul, Ahmed Zahran, and Rosane Minghim. "Understanding the influence of news on society decision making: application to economic

policy uncertainty." Neural Computing and Applications 35.20 (2023): 14929-14945.
[7] https://www.kaggle.com/datasets/tumanovalexander/nyt-articles-data?resource=download Accessed 20 July 2024.
[8] Yun-tao, Zhang, Gong Ling, and Wang Yong-cheng. "An improved TF-IDF approach for text classification." Journal of Zhejiang University-Science A 6.1 (2005): 49-55.
[9] Sabharwal, Navin, et al. "Bert algorithms explained." Hands-on Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing (2021): 65-95.
[10] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html Accessed 6 Aug. 2024.
[11] Ohsaki, Miho, et al. "Confusion-matrix-based kernel logistic regression for imbalanced data classification." IEEE Transactions on Knowledge and Data Engineering 29.9 (2017): 1806-1819.
[12] Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." Journal of the American Medical Informatics Association 18.5 (2011): 544-551.
[13] Rao, Prahalad K., et al. "Process-machine interaction (PMI) modeling and monitoring of chemical mechanical planarization (CMP) process using wireless vibration sensors." IEEE Transactions on Semiconductor Manufacturing 27.1 (2013): 1-15.
[14] Yu, Jian, et al. "Economic policy uncertainty (EPU) and firm carbon emissions: evidence using a China provincial EPU index." Energy economics 94 (2021): 105071.
[15] Taylor, Joshua A., and Johanna L. Mathieu. "Index policies for demand response." IEEE Transactions on Power Systems 29.3 (2013): 1287-1295.
[16] Ramasamy, Ravindran, and Soroush Karimi Abar. "Influence of macroeconomic variables on exchange rates." Journal of economics, Business and Management 3.2 (2015): 276-281.
[17] Turki, Turki, and Sanjiban Sekhar Roy. "Novel hate speech detection using word cloud visualization and ensemble learning coupled with count vectorizer." Applied Sciences 12.13 (2022): 6611.
[18] Wadud, Md Anwar Hussen, M. F. Mridha, and Mohammad Motiur Rahman. "Word embedding methods for word representation in deep learning for natural language processing." Iraqi Journal of Science (2022): 1349-1361.
[19] Canty, Morton John. Image analysis, classification and change detection in remote sensing: with algorithms for Python. Crc Press, 2019.
[20] Li, Hongjian, et al. "The impact of protein structure and sequence similarity on the accuracy of machine-learning scoring functions for binding affinity prediction." Biomolecules 8.1 (2018): 12.
[21] Chen, Yuanyuan, Xuan Wang, and Xiaohui Du. "Diagnostic evaluation model of English learning based on machine learning." Journal of Intelligent & Fuzzy Systems 40.2 (2021): 2169-2179.
[22] Qi, Shi, et al. "An English teaching quality evaluation model based on Gaussian process machine learning." Expert Systems 39.6 (2022): e12861.
[23] Chang, Hui-Tzu, and Chia-Yu Lin. "Improving student learning performance in machine learning curricula: A comparative study of online problem-solving competitions in Chinese and English-medium instruction settings." Journal of Computer Assisted Learning (2024).
[24] Georgiou, Georgios P. "Comparison of the prediction accuracy of machine learning algorithms in crosslinguistic vowel classification." Scientific Reports 13.1 (2023): 15594.