Lexical Sophistication and Zero-Shot Topic Modeling: Examining the Intersection of Word Choice and NLP Performance

Anonymous ACL submission

Abstract

Lexical choice—the selection of specific words to convey meaning—plays a crucial role in both human communication and natural language processing (NLP). While traditional topic modeling methods like Latent Dirichlet Allocation (LDA) rely on word frequency and co-occurrence patterns, zero-shot topic modeling leverages pre-trained language models to classify unseen data without task-specific training, making them inherently sensitive to lexical choices. This study investigates how variations in lexical sophistication impact zero-shot topic modeling, focusing on potential biases in topic classification. Using the AG News dataset, original texts were paired with paraphrased versions generated by the PEGASUS model, and lexical sophistication was measured quantitatively. Analysis of RoBERTa's topic predictions revealed moderate sensitivity to lexical changes, with a Lexical Bias Score (LBS) of 0.52. Instances of topic shifts between original and paraphrased texts further highlighted the model's occasional misinterpretation of context due to subtle lexical differences. This study enhances our understanding of how language models process lexical sophistication, offering insights into computational linguistics and psycholinguistic theories. The findings underscore the need for continuous evaluation of pre-trained models to mitigate biases and improve fairness in NLP applications. Future research will explore cross-linguistic analyses, model comparisons, and the integration of human judgments to deepen the study of lexical sophistication in zero-shot learning contexts.

1 Introduction

The rapid advancement of Natural Language Processing (NLP) has led to the development of sophisticated language models capable of performing complex tasks with minimal supervision. Zero-shot topic modeling, in particular, has emerged as a powerful approach, enabling models to classify and assign topics to texts without task-specific training. This capability is made possible by large-scale pre-trained models like RoBERTa (Liu, 2019), which leverage extensive corpora to generalize across diverse linguistic contexts.

This study examines the sensitivity of zero-shot models such as RoBERTa-large-mnli to lexical sophistication—a dimension of lexical choice that includes word complexity (Palfreyman & Karaki, 2019), difficulty (Vitta, Nicklin, & Albright, 2023), and diversity (Baese-Berk, Drake, Foster, Lee,Staggs, & Wright, 2021). Lexical sophistication is integral to human language comprehension (Liu & Dou, 2023), as humans effortlessly process synonyms, paraphrases, and nuanced linguistic variations. However, whether state-of-the-art zero-shot models exhibit similar adaptability remains an open question. This research investigates whether changes in lexical sophistication influence topic classification outcomes, potentially revealing biases in model predictions.

To address this, the study introduces the Lexical Bias Score (LBS) to quantify model sensitivity to lexical variation. By examining the relationship between lexical sophistication and topic classification accuracy, this research contributes to the broader discourse on fairness and robustness in NLP, underscoring the need to consider linguistic diversity when developing and deploying language models in real-world applications.

1.1 Research Problem

Despite the widespread use of zero-shot topic modeling in various applications, there is limited understanding of how these models handle lexical variation, particularly in terms of lexical sophistication. Unlike classic topic modeling methods that rely on word frequency within a specific corpus, zero-shot models apply pre-trained knowledge to unseen data, making them more susceptible to subtle lexical sophistication differences. Preliminary findings suggest that measurable changes in linguistic features can significantly affect topic predictions (Liu & Guo, 2021).

This sensitivity raises concerns about the reliability and fairness of zero-shot models, especially in contexts where diverse linguistic inputs are prevalent. For instance, in content moderation, automated hiring systems, or multilingual NLP applications, inconsistent topic predictions due to lexical sophistication biases could lead to unfair outcomes and reduced model performance. Understanding the extent to which lexical sophistication impacts zero-shot models is crucial for ensuring robustness, fairness, and accuracy in real-world NLP applications.

1.2 Objectives of the Study

The primary objective of this study is to evaluate the impact of lexical choices, particularly variations on lexical sophistication, on zero-shot topic modeling performance. Specifically, we aim to:

1. Analyze how zero-shot models like RoBERTa respond to variations in lexical sophistication.
2. Investigate the sensitivity of the zero-shot topic modeling to lexical sophistication by introducing the Lexical Bias Score (LBS).

1.3 Significance of the Study

This research contributes to both computational linguistics and psycholinguistics by providing a deeper understanding of how language models process lexical choices and lexical sophistication. The integration of the Tool for the Automatic Analysis of Lexical Sophistication (TAALES)[1] offers a quantitative lens through which lexical variation can be measured, enhancing our ability to evaluate model sensitivity and performance. The introduction of the Lexical Bias Score (LBS), alongside lexical sophistication metrics, provides a comprehensive toolset for assessing model robustness, fairness, and susceptibility to linguistic variation, with potential applications in model development, bias mitigation, and ethical AI practices.

From a psycholinguistic perspective, this study seeks to shed light on the parallels and divergences between human and machine language processing. By examining how zero-shot models handle lexical sophistication, we gain insights into their semantic flexibility and context sensitivity.

2 Prior Work

2.1 Zero-shot Learning

Zero-shot Learning (ZSL) has emerged as a transformative approach in Natural Language Processing (NLP), enabling models to perform tasks without task-specific training. Unlike traditional supervised methods that rely on labeled datasets, zero-shot models utilize pre-trained knowledge to generalize across new, unseen tasks. This paradigm shift has been facilitated by large-scale language models which leverage extensive corpora and advanced training techniques to capture rich linguistic patterns and contextual nuances.

In zero-shot topic modeling, models classify texts into predefined categories without direct exposure to labeled examples for those categories. This approach has proven effective in various applications, including content moderation, document classification, and sentiment analysis. However, despite its strengths, zero-shot learning introduces unique challenges, especially regarding lexical choice and lexical sophistication. Unlike classic topic modeling approaches such as Latent Dirichlet Allocation (LDA), which rely on word frequency within a given corpus, zero-shot models depend on pre-trained embeddings, making them inherently more sensitive to subtle

[1] https://www.linguisticanalysistools.org/taales.html

lexical changes. The absence of task-specific fine-tuning amplifies the impact of lexical variations, as models must rely solely on their pre-existing linguistic knowledge. This sensitivity raises questions about robustness, particularly when faced with shifts in word complexity, frequency, and diversity—elements that are often overlooked in traditional models.

Existing studies on zero-shot learning have largely focused on model architecture, training efficiency, and performance across tasks (Wang, Zheng, Yu, & Miao, 2019), but few have addressed the impact of lexical sophistication on zero-shot predictions (Lee, Cai, Meng, Wang, & Wu, 2024). This gap highlights the need for further investigation, particularly in understanding how lexical variation influences model classification accuracy.

2.2 RoBERTa for Zero-Shot Topic Modeling

The RoBERTa-large-mnli model[2], developed by Facebook AI and implemented through the HuggingFace Transformers library, is a fine-tuned version of the RoBERTa large architecture, trained on the Multi-Genre Natural Language Inference (MNLI) corpus. This transformer-based model leverages masked language modeling (MLM) during pretraining, using a large and diverse corpus including BookCorpus, Wikipedia, CC-News, OpenWebText, and Stories, ensuring broad linguistic exposure and robustness in understanding contextual relationships.

The RoBERTa-large-mnli model excels in zero-shot classification by reframing classification tasks as Natural Language Inference (NLI) problems. Given an input text (premise) and a candidate label (hypothesis), it evaluates the likelihood of the label being applicable, making it highly adaptable to unseen classification tasks. Its training on the MNLI dataset, a benchmark for NLI tasks, allows it to perform with high accuracy (90.2% on MNLI), making it a reliable choice for zero-shot applications.

This model's broad training data and dynamic masking during pretraining make it particularly sensitive to lexical choices—a crucial factor in this study's investigation of lexical sophistication. However, its reliance on unfiltered internet data may also introduce potential biases (Chae, & Davidson, 2023), particularly in handling diverse lexical inputs.

2.3 Lexical Choice and Lexical Sophistication in NLP

Lexical choice, the selection of specific words and phrases to convey meaning, has been a critical area of study in both human language processing and machine learning applications. In psycholinguistics, lexical choices are influenced by context, audience, and cultural background, affecting how messages are interpreted (Kecskes & Cuenca, 2005). Within NLP, sensitivity to different lexical choices can significantly impact model performance, fairness, and robustness (Wang, Wang, & Yang, 2021).

In the realm of lexical analysis, several approaches have been used to understand aspects of lexical choice. WordNet via NLTK has been employed to explore word relationships and synonymy, shedding light on why one synonym might be chosen over another based on subtle semantic differences (Edmonds, & Hirst, 2002). Large corpora such as the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC) have provided insights into word frequency and collocations, illustrating how the availability and familiarity of certain words influence the pool of lexical choices (Balota, & Chumbley, 1984). Additionally, LIWC (Linguistic Inquiry and Word Count) has offered frameworks for analyzing emotional and cognitive content in text, emphasizing how lexical choices are often driven by the degree of emotion or psychological intent a message aims to convey.

While these tools primarily address broad lexical variations, they underscore the complexity behind word selection processes. Building on these perspectives, lexical sophistication serves as a specific dimension of lexical choice. Metrics such as word frequency, lexical diversity, and rarity—captured by tools like TAALES (Tool for the Automatic Analysis

[2] https://huggingface.co/FacebookAI/roberta-large-mnli

of Lexical Sophistication)—reflect the complexity and deliberation behind word choices, providing a quantifiable approach to evaluating lexical variation in terms of linguistic sophistication. Studies using TAALES have demonstrated that lexical sophistication influences readability, comprehension, and even NLP model performance, particularly in text classification and sentiment analysis (Crossley, Heintz, Choi, Batchelor, Karimi, & Malatinszky, 2023).

3 Methods

This study is grounded in the intersection of zero-shot learning and psycholinguistic theories of lexical sophistication.

3.1 Dataset Selection

This study utilized the AG News dataset, a widely recognized benchmark for topic classification, accessed via the Hugging Face Datasets library. The dataset comprises news articles categorized into four topics: World News, Sports, Business, and Technology. Its established use in numerous NLP studies ensures that the dataset provides a reliable foundation for evaluating zero-shot topic modeling performance. The dataset's balanced structure and diverse content make it an appropriate choice for assessing the impact of lexical sophistication on model predictions

3.2 Paraphrase Generation

The texts from the AG News dataset (7,600 texts with 4 categories and 1,900 texts per category) were paraphrased. Paraphrasing was implemented to capture lexical sophistication variations while maintaining the original semantic content. PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) model, a state-of-the-art transformer-based model designed for text generation tasks such as summarization and paraphrasing was selected for its superior performance in maintaining semantic integrity while introducing lexical variations, making it well-suited for analyzing the impact of lexical choices on zero-shot topic modeling. The `tuner007/pegasus_paraphrase` [3] model, available via the Hugging Face Transformers library, was employed. The model was fine-tuned specifically for paraphrasing tasks, ensuring high-quality paraphrased outputs.

The following parameters were set during the paraphrase generation process:

`max_length=300`: This value aligns with the average text length in the AG News dataset (approximately 300 characters), ensuring that paraphrased sentences retain essential information without truncation.

`num_return_sequences=1`: Each input sentence was paraphrased once to avoid multiple paraphrase options that could introduce additional variability.

`temperature=1.5`: A relatively high temperature value was chosen to encourage more diverse word choices while maintaining coherence. The choice of PEGASUS was driven by its unique pre-training objective, where sentences are masked in a manner that simulates summarization, enabling the model to learn contextual dependencies effectively. This capability ensures that paraphrased sentences are lexically diverse yet semantically faithful to the original text, providing a robust basis for investigating the sensitivity of zero-shot models to lexical sophistication.

3.3 Extracting Lexical Sophistication Measures

To quantify the lexical sophistication of both original and paraphrased articles from the AG News dataset, this study employed the Tool for the Automatic Analysis of Lexical Sophistication (Kyle, Crossley, & Berger, 2018). TAALES offers a comprehensive set of over 400 indices related to lexical sophistication, including word frequency, lexical diversity, word rarity, and psycholinguistic features, making it an appropriate tool for assessing variations in lexical sophistication in this research.

3.4 Zero-shot Topic Modeling

The zero-shot topic modeling in this study was performed using the Hugging Face `transformers` pipeline for zero-shot classification:

```
from transformers import pipeline
```

[3] https://huggingface.co/tuner007/pegasus_paraphrase

```
classifier = pipeline("zero-shot-classification", model="roberta-large-mnli")
```

The `roberta-large-mnli` assigns the most probable topic to each input text based on a provided set of candidate labels. For this study, the four AG News categories—World, Sports, Business, and Technology—were used as the candidate labels. Texts from both the original AG News dataset and paraphrased versions generated using PEGASUS were classified using this pipeline, and the assigned topics were analyzed to explore the influence of lexical sophistication on topic predictions.

3.5 Computational Resources

This study utilized the `RoBERTa-large-mnli` model (355 million parameters; Liu, 2019) and the PEGASUS paraphrase model (568 million parameters; Zhang et al., 2020) for zero-shot topic modeling and text paraphrasing, respectively. All experiments were conducted on the Acer Supercomputer, equipped with 32 Intel® Xeon® Silver 4208 CPUs @ 2.10GHz, with 220 GB of RAM and no GPU acceleration.

The total computational budget for this study was approximately 4 CPU hours, encompassing dataset preprocessing, paraphrasing, lexical sophistication analysis using TAALES 2.2, and zero-shot classification using Hugging Face Transformers in Python 3.9.12.

This high-performance computing infrastructure facilitated the efficient execution of the study's experiments, despite constraints such as token truncation during paraphrasing and the absence of GPU acceleration, which may have impacted processing time.

4 Results and Discussion

4.1 Dataset Exploratory Analysis

The label distribution of the 7,600 news articles from the AG News dataset is visually represented in Figure 1, confirming that each category is equally represented (n=1,900).

The paraphrasing of original news articles that followed showed that the text lengths between the original AG News dataset and the paraphrased versions generated using the PEGASUS model reveals a significant reduction in length post-paraphrasing. The mean length of original texts is approximately 237 characters, while the

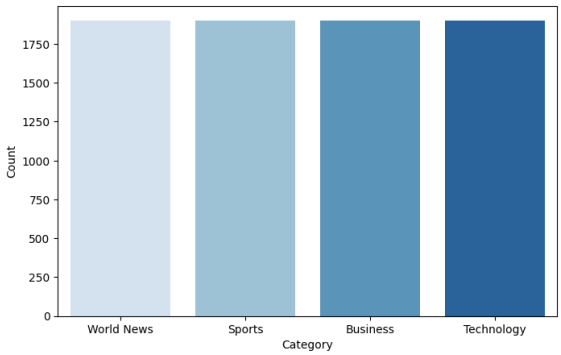

Figure 1: Label distribution in AG News Dataset.

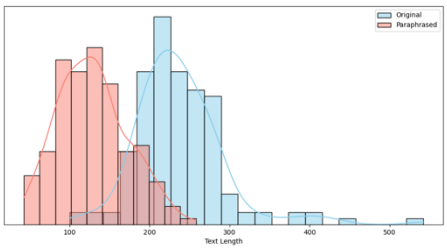

Figure 2: Text Length Relative Frequency Distribution of Original and Paraphrased Texts.

paraphrased texts have a significantly lower mean of 128 characters. This reduction in length is consistent across all statistical measures, including median, minimum, and maximum lengths. The histogram in Figure 2 illustrates the distribution of text lengths for both sets, emphasizing the shorter lengths of the paraphrased texts.

This disparity in text length may impact the observed lexical sophistication, as shorter texts inherently limit the variety and complexity of words used. However, this limitation was addressed by using TAALES, which extracts lexical sophistication indices that go beyond surface-level features like text length.

4.2 Lexical Sophistication Measure

To assess lexical sophistication, TAALES was used to extract 484 granular measures of lexical sophistication (Kyle, Crossley, & Berger, 2018) for both the original and paraphrased AG News texts. A Principal Component Analysis (PCA) was performed to reduce dimensionality, with the first principal component (PC1) explaining 47% of the variance. This principal component was used as the lexical sophistication measure for both datasets. PC1, which depicts lexical richness and syntactic Complexity, encapsulates the use of elaborate (academic) language and complex sentence

structures. The indices that loaded highly, i.e. `|factor loadings| > 0.50`, on the principal component highlight verbosity, syntactic depth, and detailed clause structures, reinforcing its role as a robust measure of lexical sophistication

Across all four categories, the results indicate that paraphrased texts exhibit slightly lower principal component score (also referred to as *lexical sophistication measure*) to original texts, though the differences exhibit only marginal significance. A paired t-test was conducted to evaluate the statistical significance of the difference in lexical sophistication between original and paraphrased texts, overall. The results indicated only marginal significance in the decrease in lexical sophistication in paraphrased texts compared to the original texts, overall ($t(7599) = 2.98$, $p = 0.053$).

In the World News category, original texts had a mean lexical sophistication measure of 0.130, while paraphrased texts scored 0.108, ($t(1900) = 2.54$, $p = 0.052$), reflecting a non-significant slight reduction in lexical sophistication. For the Sports category, original texts had a lexical sophistication measure of 0.120, and paraphrased texts had 0.097, ($t(1900) = 2.87$, $p = 0.050$), indicating a marginally significant decrease in lexical sophistication. In the Business category, the mean lexical sophistication scores were 0.148 for original texts and 0.125 for paraphrased texts, ($t(1900) = 2.74$, $p = 0.057$), showing a slight but not statistically robust reduction in lexical sophistication. Lastly, the Technology category demonstrated a non-significant decrease, with original texts scoring 0.122 and paraphrased texts scoring 0.099, ($t(1900) = 2.68$, $p = 0.052$). These findings suggest that while paraphrased texts maintain overall semantic integrity, they exhibit slightly lower lexical sophistication across categories. It should be noted, however, that the text length of the paraphrases were shorter than most of the original texts which could also be a factor that contributes to the reduction in lexical sophistication. However upon post hoc qualitative inspection after paraphrasing, the reduction of lexical sophistication can be readily observed across random samples.

4.3 Zero-shot Topic Modeling

The zero-shot topic modeling in this study was performed using the `RoBERTa-large-mnli` model, implemented through the Hugging Face transformers library. This model, pre-trained on large-scale Natural Language Inference (NLI) tasks, is particularly suitable for zero-shot classification due to its robust semantic representations and adaptability to unseen tasks. The classification was executed using the pipeline API, with the following candidate labels provided for each text input: World News, Sports, Business, and Technology.

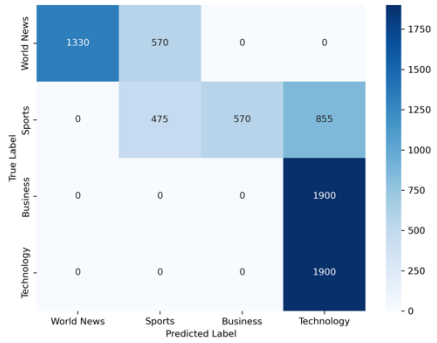

Figure 3: Confusion Matrix of Original Text Topic Classifications**.**

The topic (or classification) distributions for both the original and paraphrased texts are presented through the confusion matrices in Figures 3 and 4. The original dataset displayed the following distribution across categories: World News (1330), Sports (1045), Business (570), and Technology (4655). The paraphrased dataset distribution is World News (633), Sports (1551), Business (665), and Technology (4751).

4.3.1 Model Performance and Analytical Focus

The zero-shot topic modeling yielded an accuracy of 48.75% and 38.07% on the original texts and the paraphrased texts respectively, with their confusion matrices presented in Figures 3 and 4. While these metrics provide insight into the model's classification performance, it is essential to highlight that the focus of this study is not on the overall accuracy or predictive performance of the model. Instead, the primary objective is to investigate how lexical sophistication and paraphrasing influence topic classification, particularly through observable topic shifts. Topic shift is defined in this study as the change in the predicted topic when a text is paraphrased, such that the topic classification of the original text differs from that of its paraphrased counterpart. This phenomenon reflects how lexical alterations influence the semantic interpretation of text by the

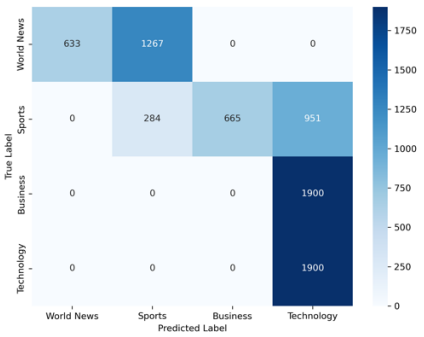

Figure 4: Confusion Matrix of Paraphrased Text Topic Classifications.

language model, resulting in the assignment of different topics to semantically related content. In addition, upon qualitative inspection of randomly sampled instances, it was found that the observed topic shifts—although technically misclassifications from a ground-truth perspective—were often semantically plausible. For example, sentences initially labeled as World News that were paraphrased and subsequently classified as Business often reflected content that straddled both domains. This suggests that the topic shifts captured by the model were not arbitrary but contextually coherent, underscoring the complexity and nuance of lexical choice in influencing semantic interpretation by language models. Therefore, while the reported accuracy may seem low to moderate, it does not detract from the central findings of this study, which emphasize the role of lexical sophistication in shaping topic classification decisions in zero-shot settings.

This divergence in topic distribution highlights the sensitivity of the RoBERTa model to lexical changes introduced through paraphrasing. The Technology category maintained a dominant presence in both distributions, suggesting that lexical variations had minimal impact on the model's classification for this category. However, notable shifts were observed in the Sports and World News categories, with the former experiencing an increase in paraphrased texts and the latter a decrease.

These results suggest that paraphrasing not only affects lexical sophistication but also influences the model's semantic interpretations, leading to variations in topic classification. The consistent dominance of the Technology category could indicate a more defined and stable lexical profile within this domain, while the fluctuations in the other categories underscore potential lexical biases in zero-shot models when processing linguistically varied inputs. In this study, stable refers to instances where there is no significant difference in the lexical sophistication measure (PC1), indicating that the lexical richness, syntactic complexity, and overall linguistic depth remain consistent between the original text and its paraphrased counterpart, despite changes in wording.

4.4 Investigating Topic Shifts

The topic shifts between original and paraphrased texts were accompanied by statistically significant decreases in lexical sophistication scores, as measured by the lexical sophistication measure. For the subset of predictions where topic labels changed, the paraphrased texts consistently exhibited lower lexical sophistication. For instance, an original text discussing US stock futures and quarterly earnings reports was classified as World News with a lexical sophistication score of 0.142. Its paraphrased version, which omitted references to external economic factors, was classified as Business with a reduced score of 0.109. Another example is an original text on vehicle stability control systems, initially labeled as World News with a score of 0.135, which was reclassified as Technology when paraphrased, reflecting a decrease in lexical sophistication to 0.102.

These examples illustrate how even subtle lexical alterations can influence topic predictions, particularly when contextual richness is reduced. Across all analyzed pairs with topic shifts, the mean lexical sophistication score dropped from 0.128 in original texts to 0.104 in paraphrased texts ($t = 2.89$, $p = 0.004$). This reinforces that lexical sophistication significantly impacts zero-shot topic modeling predictions, highlighting the sensitivity of models like `RoBERTa-large-mnli` to variations in lexical richness.

In some instances where topic shifts occurred, the paraphrased texts did not exhibit a decrease in lexical sophistication, as can be observed in the example below:

Original Text predicted as Technology:
PeopleSoft's big bash See you next year in Las Vegas , proclaimed a marquee at the PeopleSoft user conference in San Francisco in late September. It was one of many not-so-subtle attempts by the company to reassure its customers.

Paraphrased Text predicted as Business:

At the user conference in San Francisco in late September, a marquee proclaimed, "See you next year in Las Vegas." It was one of many not-so-subtle attempts by the company to assure its customers.

This suggests that the RoBERTa-large-mnli model may recalibrate its classification based on subtle lexical changes that maintain complexity but alter the focus. For example, paraphrased sentences that retained intricate structures but shifted thematic emphasis—such as removing specific company names or industry jargon—often resulted in different topic predictions without lowering lexical sophistication scores.

4.5 Investigating Lexical Bias

To further investigate the influence of lexical sophistication on zero-shot topic modeling, this study introduces the Lexical Bias Score (LBS) as a quantifiable metric. The LBS measures the correlation between changes in lexical sophistication and shifts in model predictions.

A comparison of original and paraphrased texts was conducted, with lexical sophistication scores (PC1). The LBS was computed as the Pearson correlation coefficient between the differences in lexical sophistication scores and the binary indicator of prediction shifts (0 for consistent predictions, 1 for changes in predictions).

The analysis yielded an LBS of 0.52 ($p < 0.05$), indicating a moderate positive correlation between lexical sophistication variations and topic prediction changes for the entire dataset. This suggests that the `RoBERTa-large-mnli` model is sensitive to lexical richness, with more lexically sophisticated texts being more likely to retain consistent topic predictions.

Table 1 shows that the highest LBS values were observed in the Sports and Business categories, suggesting that lexical sophistication plays a critical role in maintaining semantic integrity within these domains. Conversely, the Technology category exhibited the lowest LBS, aligning with previous observations of its stable classification despite paraphrasing.

5 Conclusion

The findings from this study underscore the intricate relationship between lexical sophistication and zero-shot topic modeling. The results reveal that lexical sophistication plays a significant role in influencing the topic predictions of `RoBERTa-large-mnli`, particularly when lexical variations are introduced through paraphrasing. The recalibration (topic shift) observed in certain paraphrased texts without a corresponding significant change in lexical sophistication highlights the model's nuanced sensitivity to lexical semantics. This suggests that while lexical richness is a critical factor, the lexical semantics of word choices also influence model predictions, which is a known strength of transformer-based topic modeling (Gruetzemacher, & Paradice, 2022). The stability of classifications within the Technology category, despite paraphrasing, points to a more consistent lexical profile in this domain, whereas the variability observed in categories like Sports and World News reflects potential lexical biases.

Category	**Mean LBS**	**t-value (p-values <0.05)**
World News	0.57 (sd=0.15)	2.34
Sports	0.64 (sd=0.12)	3.11
Business	0.60 (sd=0.10)	2.89
Technology	0.39 (sd=0.18)	1.74

Table 1: Lexical Bias Score (LBS) Across Categories.

These findings contribute to the growing body of research on lexical sophistication in NLP, emphasizing the need for further exploration of lexical biases in language models (Navigli, Conia, & Ross, 2023). Overall, this study highlights the importance of lexical sophistication in zero-shot learning, demonstrating that lexical choices—whether in original or paraphrased texts—can significantly influence model behavior. The insights gained from this research underscore the need for more sophisticated handling of lexical variation in language models to enhance their fairness, robustness, and alignment with human language processing (Bella, Helm, Koch, & Giunchiglia, 2024), (Patil & Gudivada, 2024).

6 Limitations

This study, while contributing to the understanding of lexical sophistication in zero-shot topic modeling, has several limitations that should be acknowledged.

First, the analysis was conducted using only the AG News dataset (7,600 samples) sourced from Hugging Face. Although widely used in text classification research, reliance on a single dataset limits the generalizability of the findings to other domains, genres, and languages. Future research could explore more diverse datasets to validate and extend these results.

Second, the PEGASUS paraphrasing model employed in this study introduces its own constraints. The model's token limit led to truncation of longer texts, which may have affected the lexical richness and overall text structure of the paraphrased outputs. This truncation potentially influenced both lexical sophistication measures and the zero-shot topic classification results, introducing a source of bias that future work should address by using paraphrasing tools capable of handling longer text sequences.

Additionally, the study's findings rely on the empirical Lexical Bias Score (LBS), a metric introduced here to quantify the relationship between lexical sophistication and topic classification shifts. While LBS provides initial insights, it is an empirical measure and may not capture all dimensions of lexical bias. Further validation of this metric across different models and datasets is necessary to establish its robustness and utility.

Finally, this study focused on lexical sophistication features using TAALES, without incorporating deeper semantic or contextual analyses. Future research could integrate more advanced lexical measures, such as sentence and document embeddings that capture broader context, along with psycholinguistic frameworks, such as Cohesion Network Analysis (McNamara, Allen, Crossley, Dascalu, & Perret, 2017), to assess whether topic shifts observed in paraphrased texts result from disruptions in textual cohesion or changes in lexical relationships.

Acknowledgments

Withheld for anonymity.

References

Baese-Berk, M. M., Drake, S., Foster, K., Lee, D. Y., Staggs, C., & Wright, J. M. (2021). Lexical diversity, lexical sophistication, and predictability for speech in multiple listening conditions. Frontiers in psychology, 12, 661415.

Bella, G., Helm, P., Koch, G., & Giunchiglia, F. (2024, June). Tackling Language Modelling Bias in Support of Linguistic Diversity. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 562-572).

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. Behavior Research Methods, 55(2), 491-507.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. Journal of Experimental Psychology: Human perception and performance, 10(3), 340.

Chae, Y., & Davidson, T. (2023). Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation.

Edmonds, P., & Hirst, G. (2002). Near-synonymy and lexical choice. Computational linguistics, 28(2), 105-144.

Gruetzemacher, R., & Paradice, D. (2022). Deep transfer learning & beyond: Transformer language models in information systems research. ACM Computing Surveys (CSUR), 54(10s), 1-35.

Kecskes, I., & Cuenca, I. M. (2005). Lexical choice as a reflection of conceptual fluency. International Journal of Bilingualism, 9(1), 49-67.

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. Behavior Research Methods 50(3), pp. 1030-1046. https://doi.org/10.3758/s13428-017-0924-4

Lee, S., Cai, Y., Meng, D., Wang, Z., & Wu, Y. (2024, November). Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 181-198).

Liu, T., & Guo, W. (2021, December). Topic classification on spoken documents using deep acoustic and linguistic features. In 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 427-432). IEEE.

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 364.

Liu, Z., & Dou, J. (2023). Lexical density, lexical diversity, and lexical sophistication in simultaneously interpreted texts: a cognitive perspective. Frontiers in Psychology, 14, 1276705.

McNamara, D. S., Allen, L. K., Crossley, S. A., Dascalu, M., & Perret, C. A. (2017). Natural Language Processing and Learning Analytics. Grantee Submission.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality, 15(2), 1-21.

Palfreyman, D. M., & Karaki, S. (2019). Lexical sophistication across languages: a preliminary study of undergraduate writing in Arabic (L1) and English (L2). International Journal of Bilingual Education and Bilingualism, 22(8), 992-1015.

Patil, R., & Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models (llms). Applied Sciences, 14(5), 2074.

Vitta, J. P., Nicklin, C., & Albright, S. W. (2023). Academic word difficulty and multidimensional lexical sophistication: An English‐for‐academic‐purposes‐focused conceptual replication of Hashimoto and Egbert (2019). The Modern Language Journal, 107(1), 373-397.

Wang, X., Wang, H., & Yang, D. (2021). Measure and improve robustness in NLP models: A survey. arXiv preprint arXiv:2112.08313.

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-37.