From Grammar to Geography: A Computational Analysis of Language Tool Ads Across Countries

Chenyun Wang, Yijia Bai

Summary

Introduction

With the growing demand for language tools across the globe, companies like *Grammarly* and *DeepL* have significantly expanded their digital advertising footprints. However, their ads often vary across countries—in content, format, and visibility. These variations are not only shaped by marketing strategies but may also reflect algorithmic preferences and cultural adaptation.

This project leverages data from the **Google Ads Transparency Center** to systematically collect and analyze ad creatives for Grammarly and DeepL across 14 countries and 3 ad formats (text, image, video). By comparing regional differences and using OCR to extract image-based ad content, we aim to explore how these companies tailor their messaging to local markets and whether ad recommendation algorithms exhibit geographical biases. Ultimately, this project contributes to a better understanding of the opaque mechanisms behind targeted advertising.

Motivation

As generative AI reshapes how we interact with language, it's easy to overlook the tools that laid the foundation for today's linguistic technologies. Prior to the rise of large language models, millions of users relied on traditional language tools such as Grammarly and DeepL for writing assistance and translation. These platforms not only played a critical role in improving communication but also pioneered strategies in reaching global audiences. Understanding their evolution offers a window into both the pre-AI landscape and the current digital marketing ecosystem.

We are particularly interested in how these tools, which were already influential before the generative AI boom, adapt their advertising to different regions and formats today. To explore this, we focus on the Google Ads ecosystem, where both Grammarly and DeepL maintain a strong presence- what are Grammarly's and DeepL's advertising strategies today? By collecting ad creatives across 14 countries and categorizing them by format (text, image, video), we aim to uncover patterns in content localization, cultural messaging, and algorithmic exposure.

This investigation helps us examine how global language apps market themselves in an increasingly competitive and AI-driven space. More broadly, it sheds light on how digital ad platforms may influence visibility across borders—raising questions about personalization, algorithmic bias, and transparency in global advertising.

Data Collection

To analyze the global advertising strategies of Grammarly and DeepL, we leveraged the Google Ads Transparency Center API through SerpAPI to systematically gather ad creatives across 14 countries, including the United States, Japan, Germany, and India. Our data spanned three formats—text, image, and video—and focused specifically on ads linked to each company's domain (e.g., "grammarly.com").

We developed an automated script to query all combinations of countries and formats, with each query capped at 100 results. Combining 14 countries, 3 content formats for the two companies, in total, we conducted 84 searches and collected 5,388 unique ad records. The data was organized into structured DataFrames, where we cleaned timestamps, removed duplicates, and prepared the dataset for further analysis. For image-based ads, we applied OCR (Optical Character Recognition) to extract embedded textual content, enabling deeper insights into messaging strategies. This comprehensive data acquisition pipeline sets the foundation for analyzing cross-regional advertising patterns and exploring the potential algorithmic dynamics influencing ad exposure.

The final dataset consisted of 16 columns, capturing both metadata and content-related details of each advertisement, exemplified in the graph below. Key fields include advertiser_id and advertiser, which identify the ad's sponsor; width and height, which indicate the visual dimensions of the ad; and total_days_shown, first_day_shown, and last_day_shown, which describe its lifespan. The country column specifies the geographic region where the ad was displayed.

For content analysis, each ad is associated with an ad_creative_id and includes either image or text URLs for text and image formats, or video URLs for video ads. We extracted text from images saved in the image_text column for further analysis . Each ad also contains a direct repository link from the Google Ads Transparency Center, allowing for external reference and verification.

	advertiser_id	advertiser	ad_creative_id	format	target_domain	image	width	height	total_days_shown
0	AR05288677867637440513	Grammarly, Inc.	CR04533100280257970177	text	grammarly.com	https://tpc.googlesyndication.com/archive/simg...	380.0	173.0	748
1	AR05288677867637440513	Grammarly, Inc.	CR03331623689333506049	text	grammarly.com	https://tpc.googlesyndication.com/archive/simg...	380.0	259.0	415
2	AR05288677867637440513	Grammarly, Inc.	CR10097413772969771009	text	grammarly.com	https://tpc.googlesyndication.com/archive/simg...	380.0	199.0	41
3	AR05288677867637440513	Grammarly, Inc.	CR03487015468663832577	text	grammarly.com	https://tpc.googlesyndication.com/archive/simg...	380.0	418.0	417
4	AR05288677867637440513	Grammarly, Inc.	CR09356532721863622657	text	grammarly.com	https://tpc.googlesyndication.com/archive/simg...	380.0	167.0	692

first_shown	last_shown	details_link	country	creative_format	link	image_text
2023-03-28	2025-04-13	https://adstransparency.google.com/advertiser/...	China	text	NaN	Sponsored\n\n© Grammarly\nwww.grammarly.com/\n...
2024-02-21	2025-04-13	https://adstransparency.google.com/advertiser/...	China	text	NaN	Sponsored\n\n© Grammarly\nwww.grammarly.com/\n...
2025-03-04	2025-04-13	https://adstransparency.google.com/advertiser/...	China	text	NaN	Sponsored\n\n© Grammarly\nwww.grammarly.com/\n...
2024-02-21	2025-04-13	https://adstransparency.google.com/advertiser/...	China	text	NaN	Sponsored\n\nGrammarly\nwww.grammarly.com/\n\n...
2023-05-23	2025-04-13	https://adstransparency.google.com/advertiser/...	China	text	NaN	Sponsored\n\n© wwarammarly.com/\n\nFree Plagi...

Methodology

1. Data preprocessing

As part of the data preprocessing step, we applied Optical Character Recognition (OCR) to extract textual content from image-based advertisements. Using the image URLs provided in the dataset, we downloaded each image and processed it with an OCR engine to identify and extract any embedded text. This allowed us to convert visual ad content into machine-readable text, enabling consistent content analysis across both text and image ad formats. The extracted text was then cleaned and lowercased to facilitate downstream comparisons of messaging strategies across regions and formats. URLs appearing in the image were also extracted into a new column, to make the cleaned text tidier for further analysis. However, OCR may extract URLs that are incomplete, outdated, or incorrect due to small font sizes, distortions, or truncations, affecting the cleanliness and utility of the new column created for URLs.

To ensure consistent analysis across languages, we translated all non-English text entries in the image_text_clean column into English. We used the langdetect library to automatically identify the language of each text sample. For those not originally in English, we applied the googletrans library (a Python interface to the Google Translate API) to translate the content into English. This preprocessing step was essential for enabling uniform sentiment analysis and keyword comparison across regions, especially given the multilingual nature of our dataset. While automatic translation provides a scalable solution, it may introduce minor inaccuracies, particularly in short phrases or domain-specific language. Despite these limitations, the translated text enabled more meaningful and coherent downstream analyses of global advertising content.

2. Sentiment analysis

To analyze the emotional tone of ad messaging, we used the RoBERTa transformer model for sentiment analysis on the cleaned text as an exploratory data analysis. Specifically, we adopted the cardiffnlp/twitter-roberta-base-sentiment model, which is fine-tuned on Twitter data. This choice was motivated by the similarity between our dataset and Twitter content: both consist of short, informal, and often context-dependent text. RoBERTa transformer has high classification accuracy, effective handling of informal language, and adaptability to real-world communication styles. However, this is based on the assumption that ad messages are similar to Twitter. While the Twitter-tuned model aligns well with our data, it may still underperform if ad language deviates significantly from typical social media usage. Additionally, RoBERTa is computationally demanding and slower than traditional models, especially when run without batching or on limited hardware.

To evaluate the validity of the sentiment classifications, we conducted a manual review of the text samples from each sentiment category. For each label—positive, neutral, and negative—we randomly selected rows and inspected the content of the corresponding image_text_clean entries to assess whether the assigned sentiment aligned with human judgment.

3. Semantic Alignment Analysis

To evaluate the alignment between Google Ads and official brand messaging, we conducted a semantic similarity analysis using GPT-generated brand descriptions as reference points. This approach aimed to quantify how closely ad copy from image ads reflected the core positioning of Grammarly and DeepL.

First, we generated concise, high-level brand descriptions using GPT-4, designed to capture each brand's core identity without directly copying text from their websites. These summaries served as the semantic baseline for comparison:

- Grammarly: *"Grammarly is a digital writing assistant that aims to revolutionize written communication by offering advanced grammar, punctuation, tone, and clarity corrections. Its primary target is anyone who writes and communicates in English, ranging from students and professionals to non-native English speakers and content creators."*

- DeepL: *"DeepL aims to serve individuals and businesses seeking high-quality, fast, and secure translation services, leveraging its advanced AI technology to ensure precise and nuanced language translation across various languages. Its key features include document translation, integrable solutions for businesses, and confidentiality of user data, underscoring its commitment to practical usability and user trust."*

Next, we extracted text from image ads using OCR and applied a multi-step cleaning process to remove noise. This included stripping "Sponsored" tags, URLs, special characters, and excessive whitespace. This preprocessing ensured that the resulting text accurately captured the core messaging without interference from platform-specific artifacts.

We then used the all-MiniLM-L6-v2 model from the SentenceTransformers library to compute cosine similarities between the cleaned ad texts and their respective brand summaries. This model was selected for its balance between computational efficiency and semantic accuracy, making it suitable for large-scale, multilingual text analysis. The final similarity scores were stored as a new column in our data, providing a quantitative measure of alignment for each ad.

4. Visual Style Clustering of Ad Images

To analyze the visual characteristics of Grammarly and DeepL ads, we employed a multi-step process involving image collection, feature extraction using a CLIP model, and dimensionality reduction with t-SNE. This approach enabled us to identify distinct visual styles and creative strategies used by each brand.

We sampled 1,000 ad images from the Google Ads Transparency dataset, covering both Grammarly and DeepL. These images were downloaded using public URLs and processed to ensure consistent format and resolution. This dataset provided a diverse set of ad creatives, reflecting each brand's visual identity across multiple regions and formats.

We then used the openai/clip-vit-base-patch32 model from Hugging Face to extract 512-dimensional image embeddings, capturing both style and content features relevant to semantic understanding. The embeddings were computed on CPU due to hardware constraints. This step allowed us to convert the high-dimensional pixel data into a compact, semantically meaningful representation for each ad image, preserving both visual style and content context.

To visualize the distribution of ad image styles, we reduced the high-dimensional embeddings to 2D using t-SNE. This technique projects similar ads closer together while maintaining meaningful distances between visually distinct clusters.

5. Cross-Country Variation in Semantic Alignment

To assess cross-country semantic divergence in ad texts, we analyzed how closely advertisements from different regions align with the official brand positioning for Grammarly and DeepL. This analysis aimed to capture regional variations in brand messaging and reveal strategic differences in global advertising. We first aggregated the semantic similarity scores for each country, calculating the mean and standard deviation to capture overall alignment trends. To ensure statistical reliability, we included only countries with at least 10 ads in this analysis. This filtering step reduced noise and emphasized meaningful regional differences, providing a clearer picture of how ad messaging varies across global markets.

To further explore brand-specific strategies, we grouped the data by both brand (Grammarly and DeepL) and country. This approach allowed us to isolate the alignment patterns for each brand, revealing how their advertising strategies differ across regions. By separating the data at this level, we aimed to capture both the general cross-country trends and the unique positioning approaches adopted by each brand.

For data visualization, we used two complementary methods. The first plot aggregates all brands to highlight the overall cross-country alignment patterns, while the second plot breaks down the data by brand, allowing for direct comparison of Grammarly and DeepL's regional strategies. Both visualizations use bar charts to present mean semantic similarity, facilitating straightforward cross-country and cross-brand comparisons.

6. Topic Modeling: What Do Ads Actually Say?

To uncover the underlying themes in ad texts and identify brand-specific content strategies, we employed BERTopic for topic modeling. This approach allowed us to group semantically similar texts into discrete topics, providing insights into the dominant messaging patterns for Grammarly and DeepL.

We began by extracting cleaned text from image ads, removing entries with missing or null values to ensure accurate topic clustering. Using the BERTopic library, we generated topics based on these texts, capturing both global and brand-specific themes. This method leverages transformer-based embeddings for semantic representation, combined with clustering algorithms to identify coherent topic groups. The resulting topics were then associated with each ad in the dataset, allowing for direct comparison between brands.

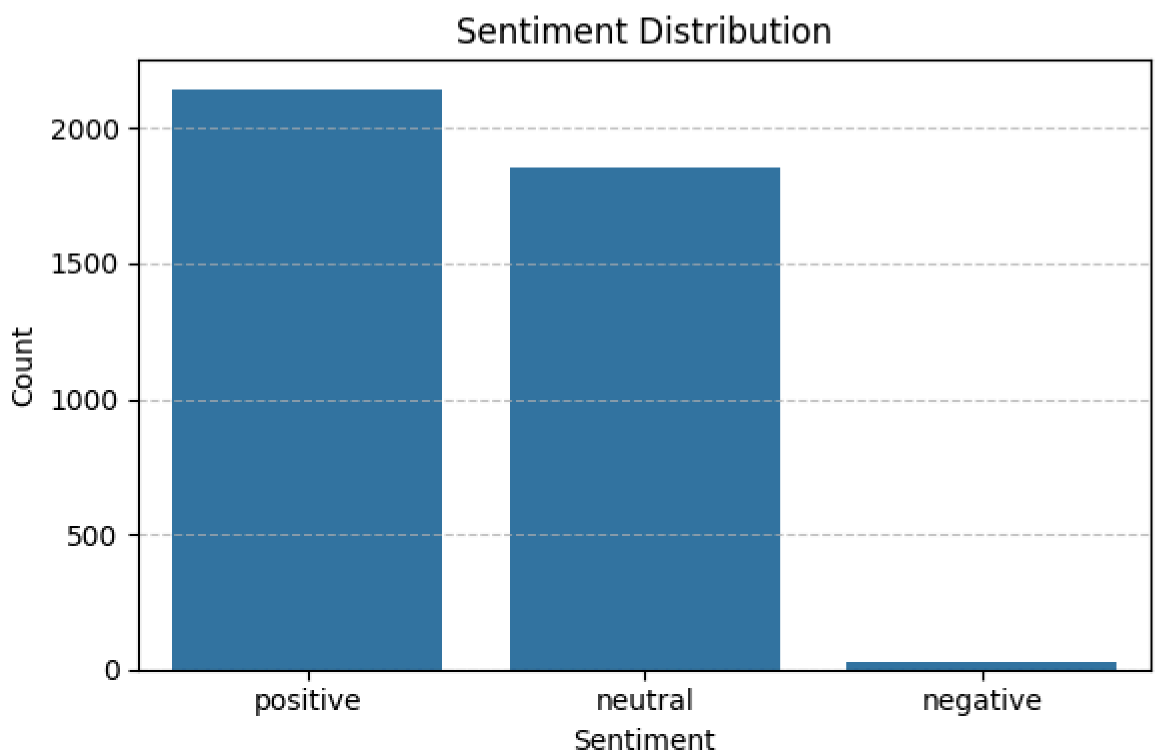
To visualize the overall topic structure, we generated an intertopic distance map, highlighting the spatial relationships between topics based on their semantic similarity (Figure 5.1). This visualization provided an overview of the topic distribution, revealing clusters of related content and outlier topics that may reflect niche or highly localized messaging.

We also extracted the top keywords for each topic to better understand the thematic focus of the ads. Additionally, we separated the data by brand, producing brand-specific topic frequency distributions to highlight differences in messaging strategy. These plots illustrate the relative emphasis each brand places on various content themes, offering a direct comparison of their communication priorities.

Preliminary Results

1. Sentiment analysis

The sentiment analysis of the cleaned advertisement text reveals a strong skew toward neutral or positive emotional tone. Out of the total classified ads, 2,163 (approximately 53%) were labeled as neutral, while 1,844 (about 45%) were identified as positive. In contrast, only 21 ads (less than 1%) were classified as negative. The distribution is displayed in the bar chart below. This distribution suggests that most ads in our dataset adopt either an informational or encouraging tone, aligning with the persuasive and brand-positive nature typical of digital marketing content. The near absence of negative sentiment is expected, as advertisers generally avoid language that could evoke adverse reactions from potential users.

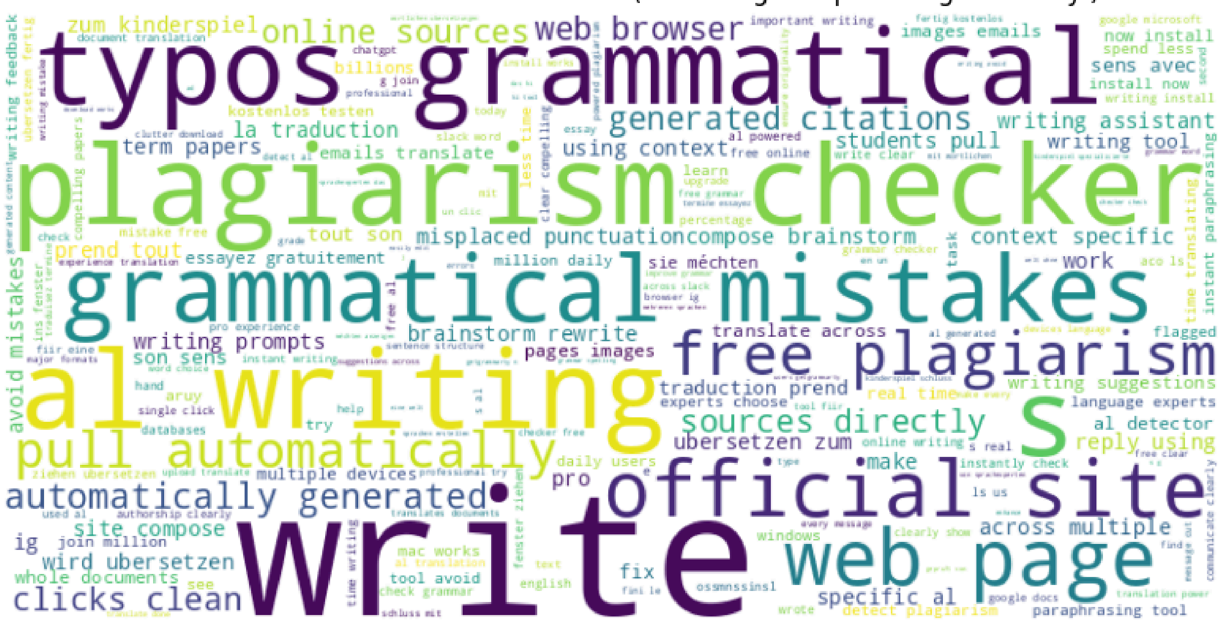


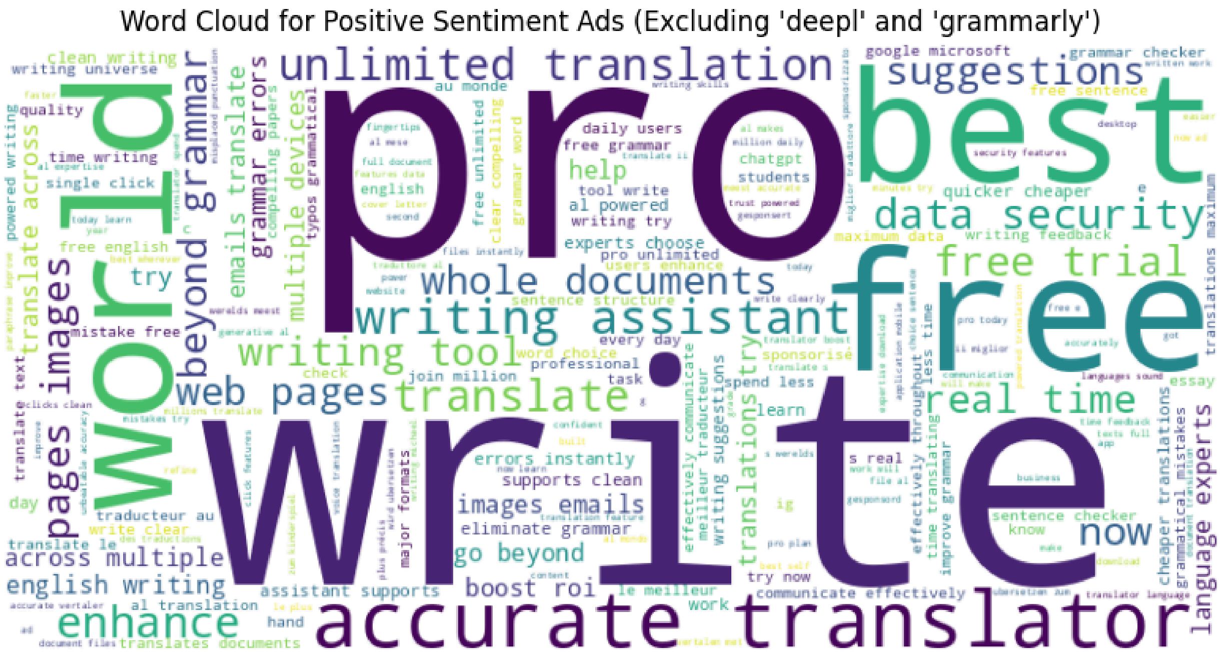
Graph1: Distribution of advertisement text sentiment

To further understand the sentiment distribution, we manually examined representative samples from each sentiment category. Negative ads were rare, and they were all from Grammarly. Upon closer inspection, we found that many of them focused on correcting grammar errors or avoiding plagiarism. These messages were functionally similar to those labeled as neutral or positive, often using promotional phrases such as “fix misplaced commas” or “check your work for plagiarism.” Negative sentiment is inherently uncommon in advertisements. Rather than expressing negative sentiment, these are classified as “negative” because of their focus on correcting mistakes and avoiding plagiarism in academic or professional writing.

The word clouds for positive and neutral sentiment ads show significant overlap in vocabulary, with frequent terms like “write,” “pro,” and “free” appearing in both. Neutral ads tend to highlight functional and compliance-oriented language, such as “plagiarism,” “checker,” and “grammatical mistakes,” reflecting an informational tone while demonstrating the main functions. Positive ads, by contrast, emphasize benefit-driven terms like “best,” “enhance,” and “suggestions,” aiming to promote action and improvement. Overall, the similarity between the two suggests that ad language is generally persuasive but emotionally moderate, making it challenging for sentiment models to clearly distinguish between neutral and positive tones.

Word Cloud for Neutral Sentiment Ads (Excluding 'deepl' and 'grammarly')





2. Semantic Alignment Analysis

The overall distribution of semantic similarity scores is presented in Figure 2.1. Most ads fall within the 0.3 to 0.5 range, indicating partial alignment with the official brand messaging. This distribution suggests that while many ads reflect certain brand values, they often include promotional or user-centric elements (e.g., discounts, speed, convenience) that diverge from the core brand narrative. Notably, a small number of ads with scores below 0.2 likely represent highly localized or off-brand messaging, while the rare cases above 0.6 reflect strong adherence to core brand values.

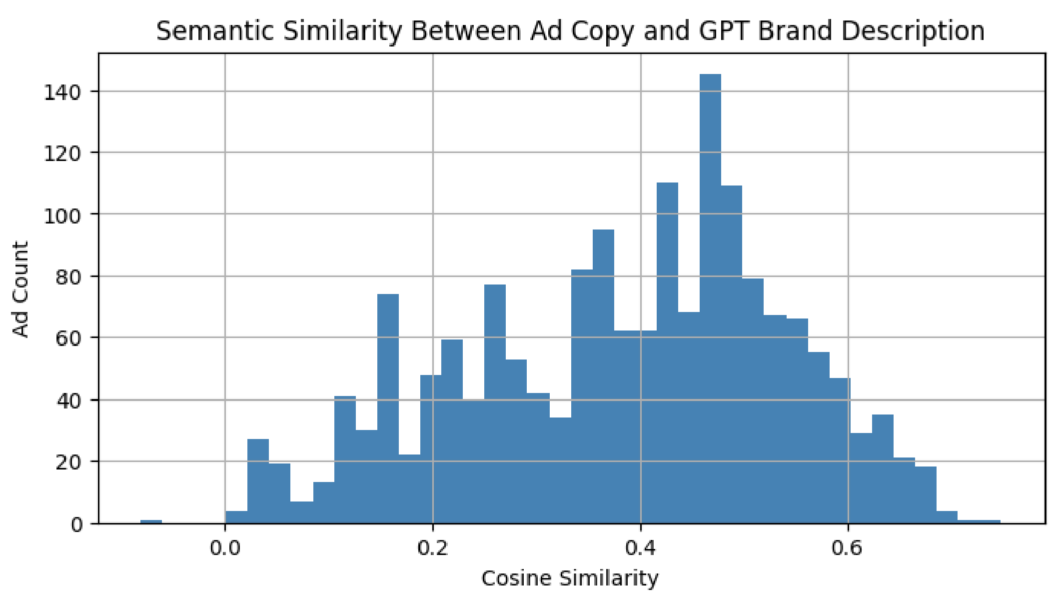


Figure 2.1

A brand-level comparison is provided in Figure 2.2, revealing distinct advertising strategies. Grammarly ads exhibit a broader distribution, with a noticeable peak around 0.35-0.45, reflecting more diverse messaging focused on user engagement and functionality (e.g., "free grammar check" and "writing improvement"). In contrast, DeepL ads show a narrower, higher-similarity distribution, reflecting a more consistent focus on translation quality, precision, and security. This difference likely reflects the brands' distinct marketing strategies, with Grammarly adopting a high-frequency, broad-target approach and DeepL emphasizing brand consistency and professional appeal.

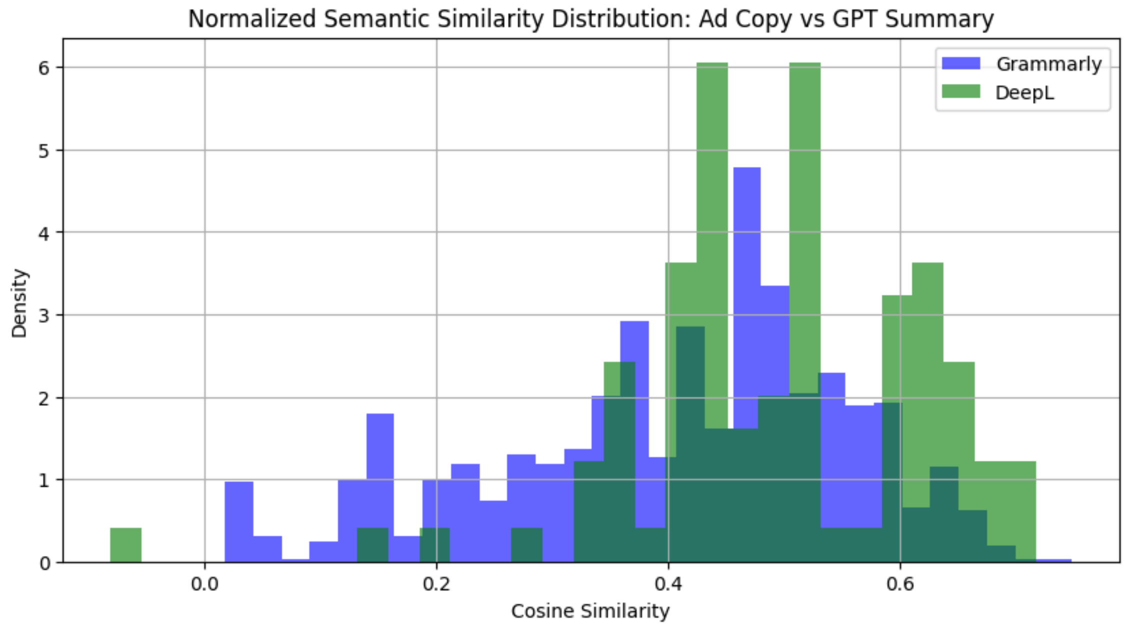


Figure 2.2

Finally, the ad volume comparison in Figure 2.3 highlights Grammarly's significantly larger ad output, aligning with its mass-market focus, while DeepL's smaller, more consistent set of ads reflects a more targeted, brand-focused strategy. Together, these findings illustrate two contrasting approaches to digital advertising: high-volume, conversion-oriented campaigns versus more focused, identity-preserving efforts.

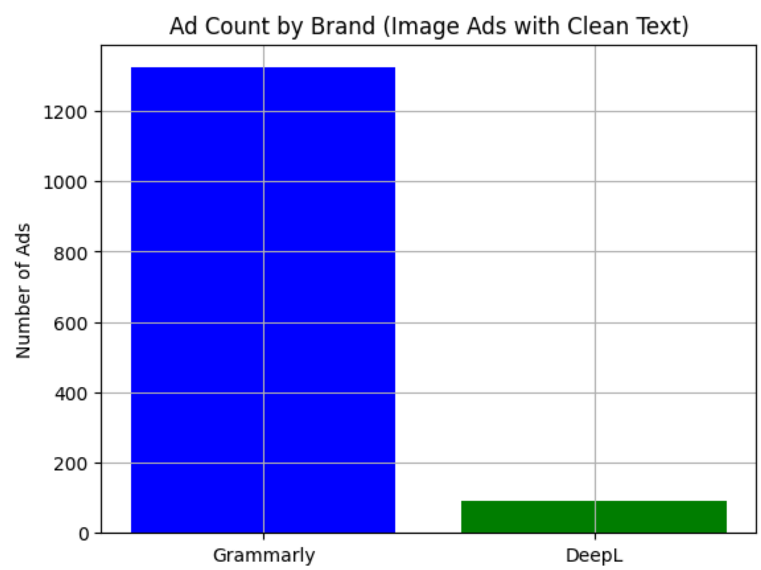


Figure 2.3

3. Visual Style Clustering of Ad Images

The t-SNE visualization (Figure 3.1) reveals a fundamental difference in the visual strategies of Grammarly and DeepL. Grammarly ads are widely dispersed, reflecting a more experimental and varied approach to visual communication. These ads often feature human-centric designs, diverse aspect ratios, and a mix of real-world and abstract backgrounds. This diversity likely reflects Grammarly's broad target audience, which includes both professional and everyday users seeking improved written communication.

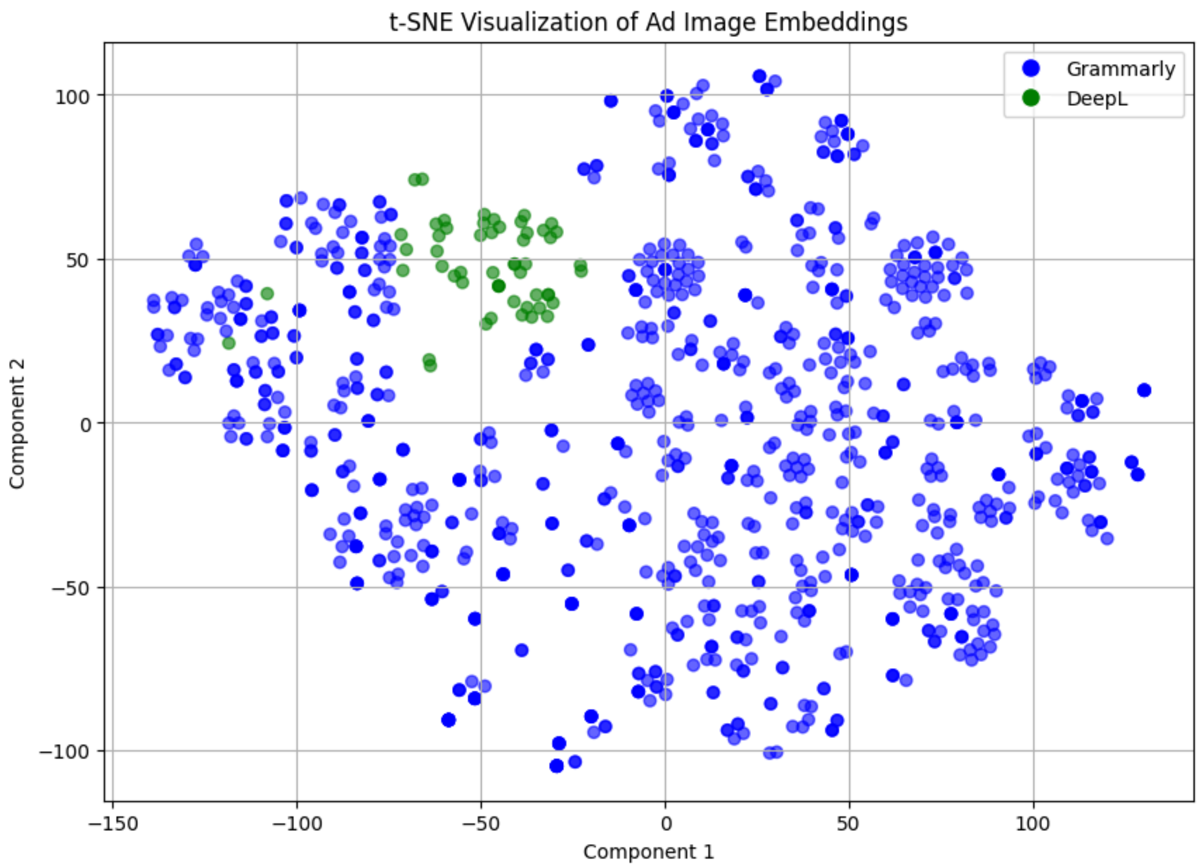


Figure 3.1

In contrast, DeepL ads form a more compact, localized cluster, indicating a more uniform design language. The DeepL cluster (Figure 3.2) showcases a consistent pastel color palette, frequent use of interface elements, and standardized call-to-action buttons like "Kostenlos testen" (Free to try). This cohesive style aligns with DeepL's professional, accuracy-focused branding, emphasizing functionality and user trust.

DeepL Cluster: Uniform Interface-Driven Visual Style

Figure 3.2

Meanwhile, the Grammarly cluster (Figure 3.3) demonstrates a much broader range of creative approaches, including customer testimonials, human portraits, and dynamic typography. This variety likely reflects Grammarly's strategy of engaging users through relatable and personalized messaging, contrasting sharply with DeepL's more product-centric approach.

Grammarly Cluster: Diverse Creative Visuals

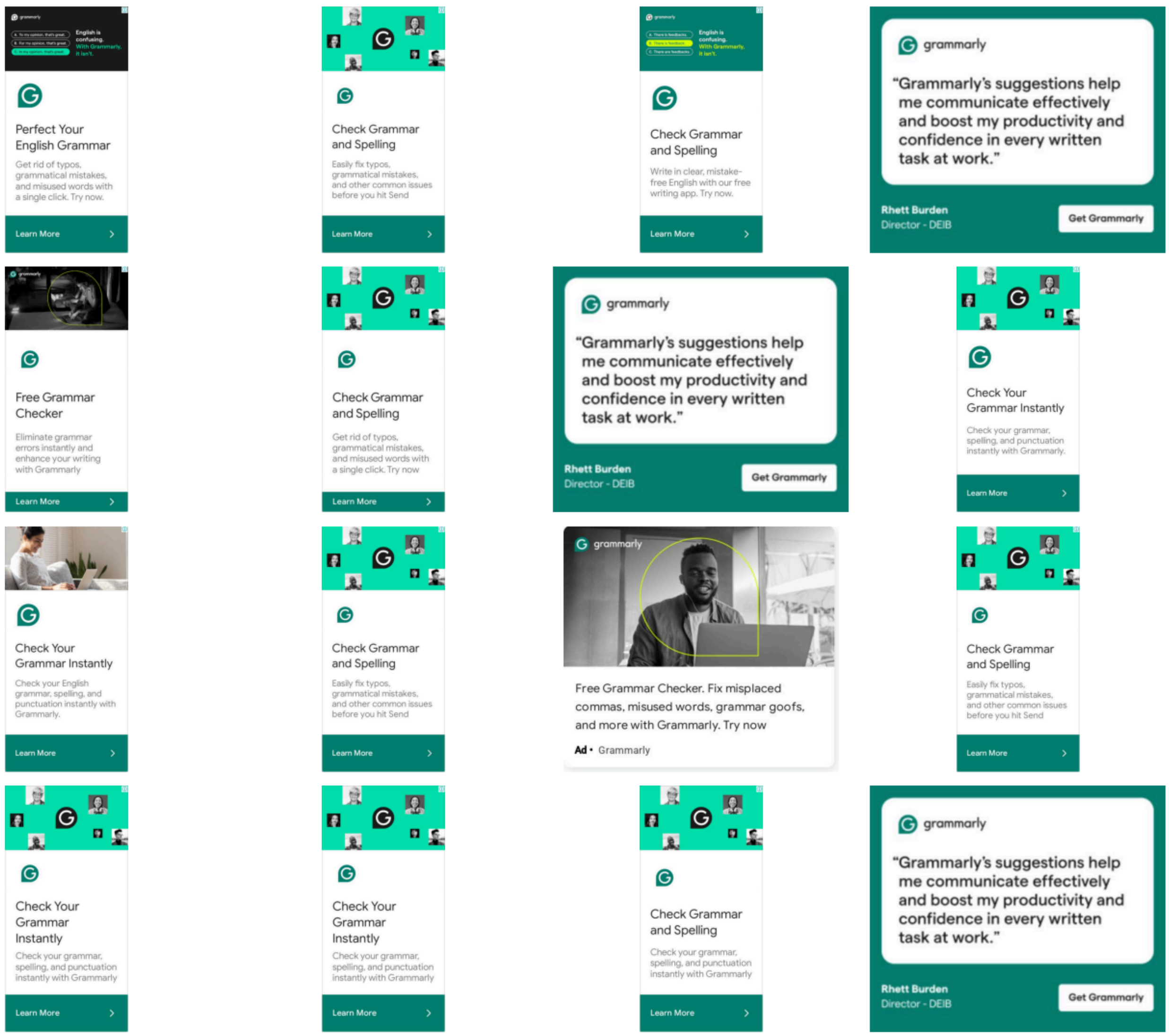


Figure 3.3

Together, these findings indicate a clear distinction in visual branding strategies: Grammarly opts for diverse, user-focused creatives that emphasize personalization and emotional connection, while DeepL maintains a more consistent, functional design aimed at reinforcing brand trust and technical precision.

4. Cross-Country Variation in Semantic Alignment

The cross-country analysis reveals several key insights. First, as shown in Figure 4.1, countries like India, Australia, Mexico, and Canada have the highest average semantic similarity, suggesting that ads in these regions are more closely aligned with the official brand voice. This may indicate that these markets are more receptive to standardized brand messaging or that the brands themselves have adopted a more globally consistent approach in these regions.

In contrast, the lower alignment observed in Germany, France, and the United States suggests that advertisers in these regions prioritize more localized, consumer-focused content that diverges from the brand's central narrative. This could reflect a strategic decision to emphasize practical benefits or regional preferences over core brand values.

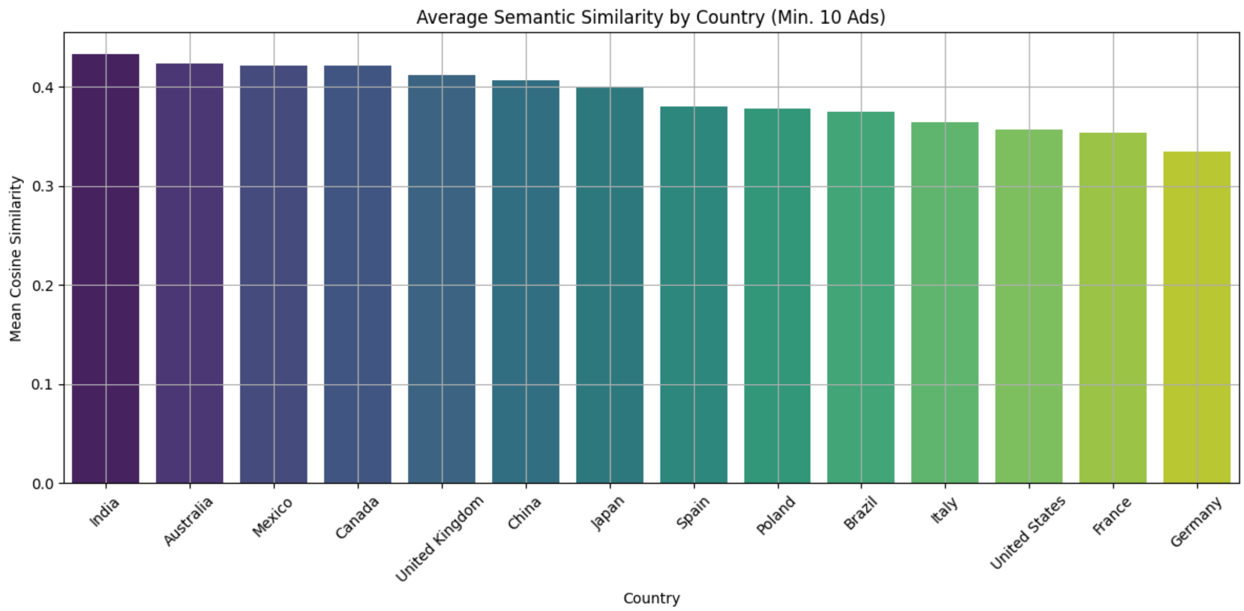


Figure 4.1

The brand-specific breakdown in Figure 4.2 further emphasizes these differences. DeepL consistently achieves higher alignment across almost all countries, reinforcing its reputation as a precise, professional translation tool. This consistency suggests a tightly controlled, globally unified branding strategy, potentially aimed at reinforcing trust and technical excellence.

Meanwhile, Grammarly's broader regional presence, including markets where DeepL has little or no measurable ad activity (e.g., India, Mexico, China), reflects a more diversified approach. Grammarly's ads appear to be more tailored to local market contexts, focusing on user benefits, calls to action, or practical use cases, rather than strictly reinforcing its core brand identity. This strategy likely supports its broader user base and mass-market appeal, in contrast to DeepL's more niche, professional positioning.

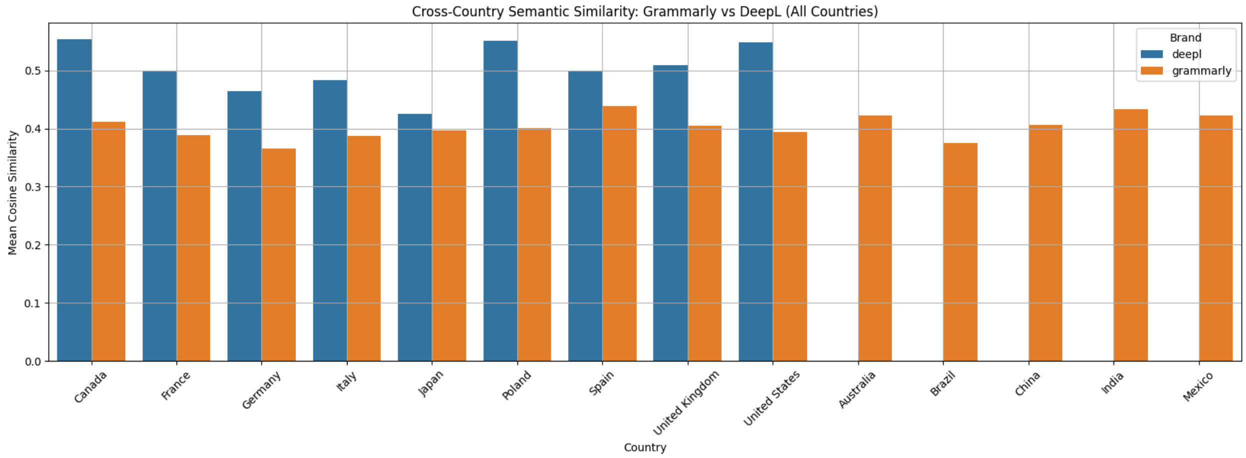


Figure 4.2

Overall, these findings highlight the contrasting approaches to global advertising: while DeepL prioritizes brand consistency and message control, Grammarly adopts a more flexible, audience-specific strategy, leveraging platform algorithms to reach a diverse set of users with varied needs.

5. Topic Modeling: What Do Ads Actually Say?

The intertopic distance map in Figure 5.1 reveals a wide spread of topics, indicating a diverse set of ad themes across the combined dataset. This visualization shows several dense clusters, suggesting areas where ad messaging is tightly focused, as well as more isolated topics that may reflect experimental or less common content.

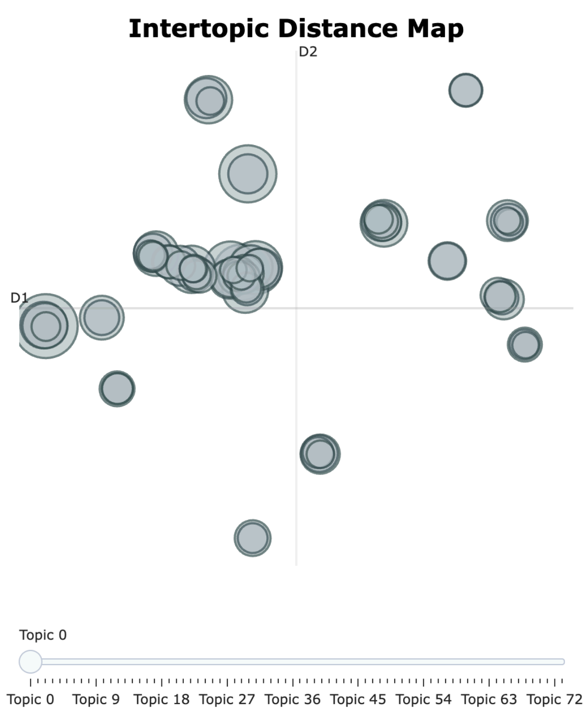


Figure 5.1

Brand-specific analysis reveals a stark contrast in content diversity. Grammarly, as shown in Figure 5.2, covers a broad range of topics, reflecting its multi-dimensional marketing strategy. Its most frequent topics include "help, apps, un," likely representing onboarding and product guidance, and "effectively, cs, throughout," which appears to target communication efficiency in professional or academic contexts. Other prominent topics include technical support ("where, install, mac") and core grammar-checking functions ("checker, grammar, avoid"), reinforcing Grammarly's position as a comprehensive writing tool.

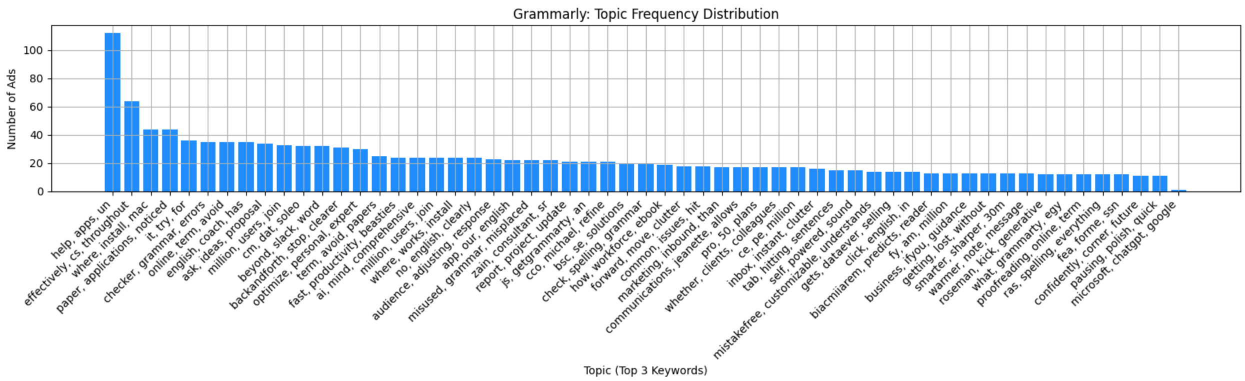


Figure 5.2

In contrast, DeepL's topic distribution, shown in Figure 5.3, is far more concentrated. The top three topics account for the vast majority of its ad volume, with keywords like "microsoft, chatgpt, google" highlighting its focus on AI integration and platform compatibility. This narrow thematic range aligns with DeepL's more specialized product positioning, emphasizing technical performance and reliability over broader, lifestyle-oriented messaging.

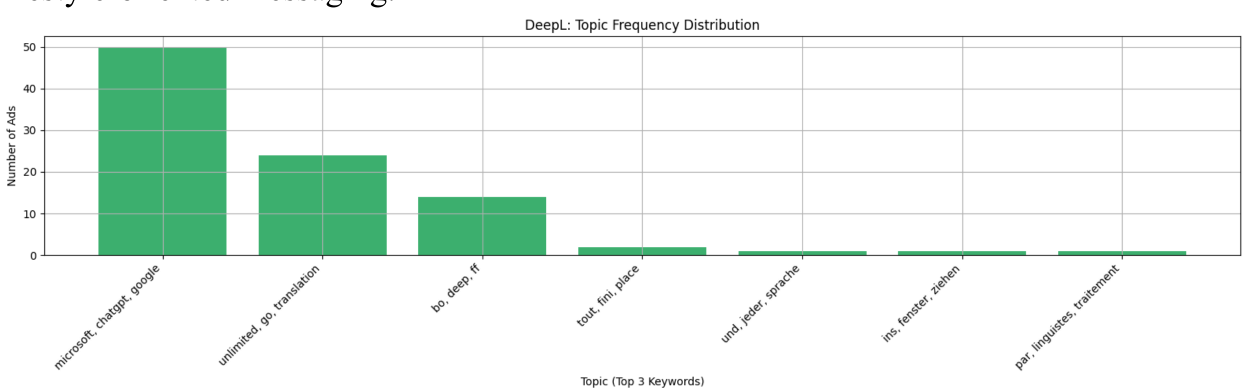


Figure 5.3

To further illustrate these thematic distinctions, we extracted the top 5 topics from each brand and reviewed their most representative ad texts. This detailed analysis provides additional context for the patterns observed in the intertopic distance map.

Grammarly: Diverse and User-Centric Messaging

1. **Onboarding / Productivity Flow (Topic -1)**
    - **Top Keywords**: got, traduire, document, help, download
    - **Sample Ads**:
        - "Stay in flow. Access top apps directly as you write, thanks to Grammarly's newest feature."
        - "Grammarly go from first draft to finished. Fast.
        - "Stay in flow. Access top apps directly as you write, thanks to Grammarly's newest feature. Grammarly."
    - **Interpretation**: This topic emphasizes workflow enhancement and productivity integration, particularly through in-app features. This is a critical aspect of Grammarly's product positioning as a versatile writing assistant.
2. **Academic or Professional Communication (Topic 0)**
    - **Top Keywords**: effectively, cs, throughout, universe, communicate
    - **Sample Ads**:
        - "Write more effectively. Communicate throughout your writing universe."
        - "Express yourself clearly and confidently in every message."
        - "Take your writing to the next level with Grammarly's advanced tools."
    - **Interpretation**: This topic highlights Grammarly's focus on clear and effective communication, appealing to students and professionals who prioritize precision in their writing.
3. **Business Writing & Fluency (Topic 2)**
    - **Top Keywords**: english, business, fluently, success, coach
    - **Sample Ads**:

        - "Set your team up for business success. Equip every member with helpful writing guidance."
        - "Grammarly provides quality suggestions to ensure your writing is clear and mistake-free."
        - "Write fluently in English. Ensure every sentence is professional."
    - **Interpretation**: This topic targets business users, emphasizing professional writing, fluency, and team communication, aligning with Grammarly's enterprise offerings.

4. **Installation and Platform Compatibility (Topic 3)**
    - **Top Keywords**: where, install, mac, windows, works
    - **Sample Ads**:
        - "Grammarly for Windows and Mac works where you do your most important writing. Install now."
        - "Seamless integration with your favorite tools."
        - "Write confidently on all your devices with Grammarly."
    - **Interpretation**: This topic focuses on technical compatibility and platform support, highlighting Grammarly's versatility across different devices and operating systems.

DeepL: Technically Focused and Consistent Messaging

1. **AI Comparison and Translation Superiority (Topic 1)**
    - **Top Keywords**: microsoft, chatgpt, google, accurately, language
    - **Sample Ads**:
        - "Translate more accurately than Google, ChatGPT, and Microsoft. Try Pro free for 30 days."
        - "Language experts choose DeepL translations over Google, Microsoft, and ChatGPT."

        - "Trust DeepL for precise translations, every time."
    - **Interpretation**: This topic reflects DeepL's strategic positioning as a technically superior alternative to mainstream AI platforms, emphasizing accuracy and reliability.

These findings reveal a stark contrast in content diversity. Grammarly's ads cover a broader range of themes, reflecting a more varied, context-driven marketing approach that targets multiple user segments. In contrast, DeepL's ads are tightly focused on a few core technical themes, reinforcing its position as a high-precision, professional translation tool. This difference in thematic breadth likely reflects each brand's distinct market positioning, with Grammarly aiming for mass-market appeal and DeepL targeting a more specialized, technically oriented audience.

Limitations

Our project faces several important limitations, the foremost being the nature of our data source. We relied on data from the Google Ads Transparency Center, which functions as a black box governed by proprietary and opaque algorithms. These algorithms likely influence which ads are surfaced and how frequently, potentially prioritizing certain advertisers, regions, or formats based on undisclosed criteria. As a result, the ads we collected may not provide a representative or exhaustive sample of each platform's global advertising activity. As a result, our dataset is inherently skewed, representing only a narrow glimpse of the broader advertising landscape. Moreover, our focus is specifically on the Google Ads ecosystem, where both Grammarly and DeepL have a notable presence, and our findings may not generalize to other platforms or contexts. This limited scope, combined with the opaque nature of Google Ads' selection mechanisms, means that our study's insights are contingent on the unique dynamics of this ecosystem, and caution should be taken when extrapolating our results to the overall market or other digital advertising environments.

Second, the methodologies we used may lead to information loss and subjectivity. We used GPT-4 to generate brand summaries, and these descriptions may not fully capture the evolving nuances and the comprehensive brand positioning. Additionally, while the all-MiniLM-L6-v2 model offers an efficient way to compute semantic similarity, it has limited capacity to detect all information and context. Therefore, when interpreting the results, they might be highly subjective. We also used CLIP and t-SNE as computational methods to realize visual clustering, and BERTopic for topic analysis, where information loss and interpretation subjectivity are also the challenges we faced.

Third, there is a notable imbalance in the volume of data collected across advertisers, which is particularly important when comparing Grammarly and DeepL. For example, we obtained over 1,300 ad entries for Grammarly, compared to only around 100 for DeepL. This discrepancy limits the fairness and comparability of our analysis, as conclusions drawn from the smaller DeepL sample may lack statistical robustness. Moreover, the significantly larger dataset for Grammarly contributes to greater variance in measures like topic diversity and semantic alignment, simply due to the increased opportunity for variation. As a result, some observed differences between brands may reflect differences in data volume rather than actual strategic or stylistic divergence.

Last but not least, our analysis primarily focuses on text and image formats. Although we collected video ad URLs, our ability to extract meaningful textual information from video content remains limited. Without reliable transcription or video content parsing, much of the potential data from video advertisements remains unutilized, constraining our sentiment and content analysis to only a subset of available formats.

Significance/Conclusion

Appendix