A Multi-View Framework for Cross-Domain Nutrition Misinformation Detection in Social Media

Vishwaa Shah[1], Indika Kahanda[1], Andrea Arikawa[2], Asal Abbaszadeh[2], Richard Loftis[2]

[1]School of Computing, University of North Florida, Jacksonville, FL, USA
[2] Department of Nutrition & Dietetics, University of North Florida, Jacksonville, FL, USA

{N01458714, indika.kahanda}@unf.edu
{a.arikawa, n01481527, n01501201}@unf.edu

Abstract

Nutrition misinformation on social media often arises from selective interpretation of scientific evidence rather than outright falsehoods, making it difficult to detect. We introduce a curated, expert-annotated Instagram dataset focused on seed oils and omega-6, two domains characterized by contested dietary claims. We evaluate feature-based, embedding-based, and transformer-based models under in-domain and cross-domain settings. Results show strong in-domain performance across all models, with Sentence-BERT achieving the highest AUPRC (up to 0.96). However, performance drops substantially under cross-domain transfer, indicating limited robustness to topic shift. Analysis suggests that while contextual embeddings capture strong in-domain semantic signals, linguistically and psychologically grounded features are more stable under distribution shift. These findings highlight the value of combining semantic and interpretable linguistic signals for robust misinformation detection.

1 Introduction

Nutrition misinformation has emerged as a significant public health concern, driven in part by the rapid dissemination of unverified or misleading dietary claims on social media (Diekman et al., 2023). Prior research shows that online health information frequently contains inaccurate or contextually distorted statements that can shape public understanding and influence health-related decision-making (Tan et al., 2025). Although recent advances in large language models (LLMs) have improved automated detection of misleading content, reliable identification remains challenging due to the domain-specific and context-dependent nature of discourse (Belkhouribchia and Pen, 2025).

Social media platforms play a central role in shaping public nutrition narratives, but their scale and heterogeneity make systematic monitoring difficult (Garg and Sharma, 2022). Among these platforms, Instagram (https://www.instagram.com) is particularly influential due to its large user base and image-centric design driven by engagement (Dean, 2025). However, most existing misinformation datasets/studies focus on platforms such as Twitter and Facebook, leaving Instagram comparatively underexplored, particularly in nutrition-related contexts (Nguyen et al., 2022). This limits our understanding of how nutrition-related claims are communicated in environments where information is often condensed into short captions, hashtags, and visually supported narratives.

This gap is particularly evident in emerging debates in nutrition, such as on seed oils and omega-6. While dietary guidelines recognize vegetable oils as part of healthy dietary patterns when consumed appropriately (Dietary Guidelines for Americans, 2020), social media narratives frequently reframe them as inflammatory or harmful (Ahmad et al., 2025). Such content often does not rely on entirely fabricated claims, but instead on selective interpretation or reframing of established scientific evidence (Lee and Kurniawan, 2025), making detection more subtle and context-dependent.

From a computational perspective, previous work has shown that linguistic and psycholinguistic features can capture stylistic and rhetorical signals associated with misinformation (Garg and Sharma, 2022). In particular, psycholinguistic cues have been found useful for identifying emotional intensification and rhetorical framing in health-related misinformation (Mahbub et al., 2022). However, generalization across domains remains an open challenge, with performance often degrading when models are applied to unseen topics or distributions (Xarhoulacos et al., 2021).

In this work, we explore a hybrid framework for nutrition misinformation detection in Instagram posts. As outlined in Figure 1, we construct a curated expert-annotated dataset of Instagram captions focusing on seed oils and omega-6. We pro-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 326–341

pose a multi-view representation that integrates linguistic complexity, rhetorical and stylistic features, psycholinguistic signals, and pretrained semantic embeddings. We evaluate both feature-based and transformer-based models under in-domain and cross-domain transfer settings to assess performance under domain shift.

Our experiments show that transformer-based models achieve strong in-domain performance, while feature-based models remain competitive and provide improved interpretability. However, all models exhibit performance degradation under cross-domain evaluation, particularly in ranking-based metrics such as AUPRC. These findings suggest that while both linguistic and neural representations capture useful signals, their ability to generalize across closely related nutrition topics needs further improvement. This highlights the value of combining interpretable linguistic signals with contextual embeddings for misinformation analysis.

2 Related Work

Health misinformation detection has been widely studied using natural language processing techniques, with transformer-based architectures and LLMs achieving strong performance in assessing the credibility of online content (Faruk, 2024; Tan et al., 2025). However, much of this work relies on general-purpose datasets or structured textual sources, which may not fully capture the informal, context-dependent nature of social media discourse, where health misinformation is often embedded in short, unstructured narratives (Belkhouribchia and Pen, 2025; Kauttonen et al., 2020).

A large portion of existing research has focused on platforms such as Twitter, Facebook, and YouTube (Yeung et al., 2022), while Instagram remains relatively underexplored despite its large user base and distinctive visual and caption-driven communication style (Nguyen et al., 2022). This is particularly important in health-related domains, where misinformation is often expressed through short captions, hashtags, and implicit claims embedded in multimodal content. Recent evidence suggests that nutrition misinformation is increasingly prevalent on Instagram, though systematic analysis remains limited due to the lack of curated datasets (Segado-Fernández et al., 2025).

Recurring nutrition misinformation narratives often center on dietary fats, particularly seed oils and omega-6 fatty acids (omega-6). These narratives frequently characterize such fats as harmful or pro-inflammatory, despite established dietary evidence supporting their role in healthy dietary patterns when consumed appropriately (Lee and Kurniawan, 2025; Ahmad et al., 2025). This gap between scientific consensus and public framing highlights the need for computational approaches that capture not only the content of claims but also their rhetorical and stylistic presentation.

Prior work in misinformation detection has demonstrated the usefulness of linguistic and stylistic features, including lexical diversity, readability, and syntactic complexity (Garg and Sharma, 2022). Misinformation is often associated with distinctive rhetorical patterns such as increased certainty, reduced hedging, and emotionally charged language (Mahbub et al., 2022). Hybrid approaches combining such features with dense embeddings, such as WELFake (Verma et al., 2021), have shown improvements in classification performance, though their effectiveness varies across domains and datasets (Khan et al., 2021).

Emotion and affective signals have also been widely studied as predictive indicators of misinformation. Prior studies suggest that misinformation content often exhibits stronger negative emotions such as fear, anger, and disgust compared to factual content (Farhoundinia et al., 2024). High-arousal emotional language has also been linked to reduced critical evaluation and increased susceptibility to misleading claims (Martel et al., 2020; Wurst et al., 2024), motivating their inclusion in computational models, although these signals may vary substantially across domains and contexts.

More recently, cross-domain misinformation detection has gained attention, particularly in health-related settings such as transferring models across disease topics (Goldani et al., 2025). These show that generalization across domains remains challenging, especially when moving between different thematic distributions or communication styles. However, most prior work focuses on disease-related misinformation, while broader lifestyle and nutrition domains remain relatively underexplored.

Despite these advances, existing approaches often treat linguistic features and neural representations separately, with limited systematic evaluation of their combination in domain-shift settings. This is particularly relevant in nutrition misinformation, where both interpretability and robustness are important. Motivated by this gap, we investigate a unified framework that integrates linguistic, rhetori-

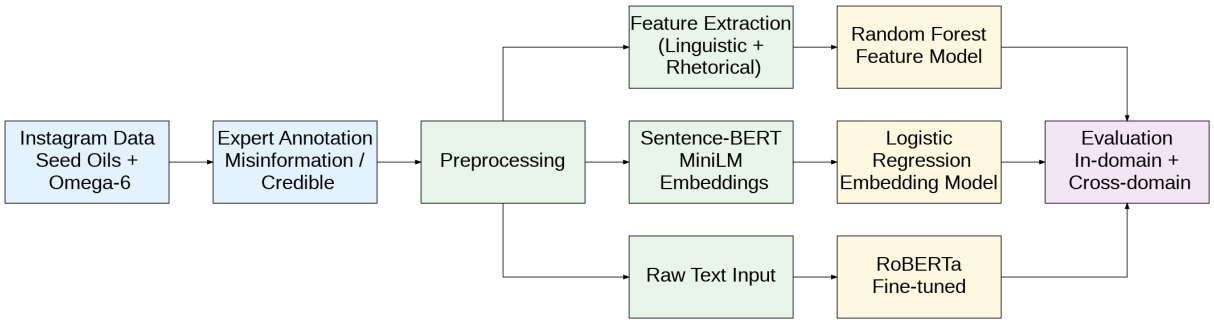

Figure 1: Overview of the proposed framework for nutrition misinformation detection. The pipeline includes data collection from Instagram, expert annotation, preprocessing, feature-based modeling (top branch), embedding-based modeling (middle branch), and transformer-based modeling (bottom branch). We evaluate robustness under domain shift, not just classification accuracy.

cal, and psycholinguistic features with transformer-based embeddings, and evaluate their effectiveness in both in-domain and cross-domain transfer.

3 Methodology

Figure 1 provides an outline of the proposed framework, summarizing the data collection, preprocessing, feature extraction, modeling approaches, and evaluation protocols used in this study.

3.1 Data Collection

A domain expert in nutrition identified relevant Instagram posts using the platform's native search interface. The sampling strategy was centered on two predefined fat-related topics: seed oils and omega-6. These topics were selected for focused analysis due to their frequent mischaracterization on social media as harmful or pro-inflammatory, despite established evidence supporting their role as beneficial unsaturated fats when consumed in appropriate dietary patterns, as reflected in the Dietary Guidelines for Americans 2020–2025 (Dietary Guidelines for Americans, 2020).

The keyword set was developed through expert consultation, iterative examination of nutrition-related content on social media, and preliminary testing of search outputs. Early exploration indicated that broad nutrition terms (e.g., "healthy eating") produced substantial amounts of irrelevant material, whereas targeted fat-specific queries yielded posts with clearer and more substantive health-related claims. The final keywords were chosen because they consistently surfaced high-engagement content and reflected areas where public misconceptions are especially prevalent.

Posts were considered eligible for inclusion if their captions explicitly referenced seed oils (e.g., soybean, sunflower, safflower, canola, or corn oil) or omega-6 in the context of nutrition, health, or disease-related claims. Only English-language posts were retained.

To mitigate the effects of algorithmic personalization, all searches were conducted using newly created Instagram accounts with no prior activity. Each account was used exclusively for one topic (seed oils or omega-6), ensuring that no cross-topic interaction influenced content retrieval. No engagement actions (e.g., likes, follows, or saves) were performed prior to or during data collection.

In addition, all search sessions were performed using a virtual private network (VPN) to minimize geographic and personalization biases. This approach was intended to approximate the content exposure of a new, unconditioned user encountering discussions related to these topics on Instagram.

Data collection took place between May 28, 2025 to June 4, 2025. Eligible posts were restricted to those published between January 1, 2020, and the end of the collection period. To focus on widely disseminated content, only posts from accounts with more than 5,000 followers were included.

All posts were publicly accessible at the time of collection, and no private accounts were accessed. Usernames and other identifying information were removed prior to analysis to preserve anonymity. Captions and associated metadata were extracted using the *Apify Instagram Post Scraper Actors*[1].

This study focuses exclusively on caption text rather than image or video content. This decision was made to enhance reproducibility and reduce ethical concerns, as textual data can be more readily anonymized, whereas visual content may contain identifiable individuals or sensitive information. By

[1] https://apify.com/actors, last accessed 06/04/2025.

isolating caption text, we attempt provide a lower-bound estimate of misinformation detectability, enabling clearer attribution of linguistic and rhetorical effects without visual confounds.

All preprocessing was implemented in Python within a Google Colab environment with fixed random seeds to ensure reproducibility. Preprocessing included removal of non-alphanumeric characters while preserving punctuation, normalization of whitespace, and retention of emojis as tokens. The annotated dataset and code are publicly available on Zenodo under a CC BY 4.0 license (https://doi.org/10.5281/zenodo.20247254).

3.2 Human Annotation

We define health misinformation in line with prior literature as "any health-related factual claim that contradicts established scientific consensus" (Sylvia Chou et al., 2020). In this work, scientific consensus is grounded in the 2020–2025 Dietary Guidelines for Americans (Dietary Guidelines for Americans, 2020). Content that contradicted or distorted these guidelines was annotated as *Misinformation*, whereas content that aligned with or did not conflict with the guidelines was labeled *Credible*. Consistent with prior public health misinformation research, our annotation framework evaluates claims relative to established evidence-based recommendations rather than isolated or emerging studies, and should therefore be interpreted as reflecting alignment with contemporaneous scientific consensus and public health guidance during the data collection period.

Three annotators with formal nutrition training independently labeled each post in two domains: seed oils and omega-6. The annotation team consisted of two graduate students in nutrition supervised by a PhD-trained nutrition faculty at University of North Florida. The annotation was guided by a detailed codebook derived from the Dietary Guidelines for Americans (see the Appendix A). To reduce potential bias, annotators worked independently, were blinded to each other's decisions, and reviewed posts in randomized order. The Inter-rater reliability (IRR) was calculated using Fleiss's κ ($= 0.814$), which accounts for agreement occurring by chance and indicates strong agreement (values above 0.80). Given this high agreement, we used majority voting without additional adjudication to preserve annotation independence. These consensus labels serve as the basis for the subsequent evaluation of model predictions.

Figure 2 presents the distribution of caption lengths by label across both domains. Overall, captions are relatively long, with mean lengths of 156.1 words (SD = 108.5) in the seed oil domain and 236.5 words (SD = 101.2) in the omega-6 domain, indicating that posts often require interpretation of extended context rather than isolated claims. In the seed oil domain, credible posts are longer on average (M = 179.5, SD = 113.9) than misinformation posts (M = 146.8, SD = 105.4), suggesting that misinformation content tends to be more concise in this domain. In contrast, in the omega-6 domain, caption lengths are highly similar across classes, with credible posts (M = 238.9, SD = 110.4) and misinformation posts (M = 235.8, SD = 98.9) showing nearly identical averages. Across both domains, caption lengths vary widely, with an overall mean of 193.6 words (SD = 112.4), reflecting substantial heterogeneity in how claims are presented.

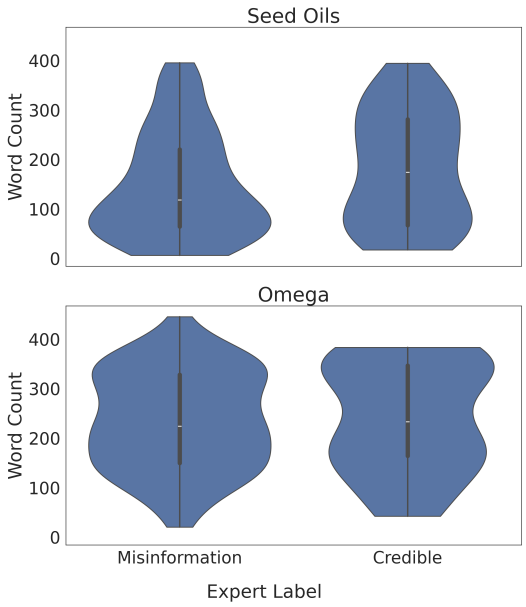

Figure 2: Caption length (word count) by expert label across seed oil and omega-6 domains. Captions in both domains are often long-form, requiring contextual interpretation beyond isolated claims.

The class distribution differs across domains but is consistently skewed toward misinformation. In the omega-6 dataset, 77.0% (114/148) of posts were labeled as *Misinformation*, while 23.0% (34/148) were labeled as *Credible*. In the seed oil dataset, 71.6% (121/169) of posts were labeled as *Misinformation*, compared to 28.4% (48/169) labeled as *Credible*. All captions were manually reviewed in full. Although some posts were brief or minimally informative, they were retained to

preserve the variability characteristic of real-world social media content.

3.3 Feature-Based Model

We develop a topic-agnostic feature-based framework to investigate how misinformation is linguistically and affectively distinguished from credible nutrition discourse. Instead of relying on surface lexical cues or topic-specific words, we focus on capturing how claims are framed across linguistic, rhetorical, and psychological dimensions.

3.3.1 Feature Space Construction

We construct a unified representation consisting of four complementary feature groups. Together, these four types of features form a comprehensive, topic-agnostic representation of linguistic style, persuasion strategy, and affective structure. A complete definition of all features is provided in Appendix B.

(1) Linguistic Complexity Features. We quantify structural and readability properties using `textstat` and `lexicalrichness`, including Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, lexical diversity, word count, and average word length. These features capture cognitive accessibility and structural complexity independent of topic.

(2) Rhetorical and Persuasive Features. We extract discourse-level signals using rule-based patterns to identify persuasive and stylistic markers. These include certainty and hedging expressions, rhetorical questions, exclamation usage, capitalization intensity, and citation or evidence indicators. These features model how claims are communicated rather than their content. Detailed rule-based patterns used for rhetorical feature extraction are provided in Appendix C.

(3) Affective Features. We use pretrained transformer models to extract sentiment and emotion signals from the transformers library (*cardiffnlp/twitter-roberta-base-sentiment-latest* and *j-hartmann/emotion-english-distilroberta-base*) to gauge the emotional tone of texts. These produce sentiment labels and emotion probability distributions. The full emotion label mapping and output structure of the classifier are provided in (Appendix D).

(4) Psychological Valence and Arousal Features. We operationalize affect using a lexicon-based approach grounded in the Research Domain Criteria (RDoC) framework (Insel et al., 2010), capturing Positive Valence (reward and approach motivation), Negative Valence (threat, loss, frustration), and Arousal/Regulatory systems (activation and intensity). These are combined with VADER (Hutto and Gilbert, 2014) sentiment scores to derive interpretable psychological dimensions reflecting emotional directionality, intensity, and motivational framing. Unlike categorical emotion labels, RDoC dimensions are designed to capture domain-independent affective processes such as threat sensitivity and approach motivation, making them well-suited for cross-topic misinformation analysis. The full lexicons used for RDoC feature extraction are provided in (Appendix E).

3.3.2 Feature-Based Classification Model

Random Forest achieved the best and most stable performance among several traditional machine learning classifiers, and was therefore selected as the final classifier. The model is implemented using `scikit-learn` and trained on the concatenated feature representation consisting of linguistic, rhetorical, and affective features.

Features are standardized prior to training, and model interpretability is examined using SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017). SHAP is used to quantify feature contributions at both global and instance levels, enabling analysis of which linguistic, rhetorical, and affective signals most strongly influence misinformation predictions.

SHAP values are computed using model-specific explainers: *TreeExplainer* for tree-based models, *LinearExplainer* for logistic regression, and *KernelExplainer* when required for non-linear models. We compute SHAP values on the full dataset using the final trained model and report mean absolute SHAP values for the positive class (misinformation) as global feature importance. We further compare SHAP-based rankings with built-in feature importance from tree-based models to assess consistency across interpretation methods.

3.4 Embedding-based model

We compute sentence-level embeddings using the Sentence-BERT model `all-MiniLM-L6-v2` (`https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`), which produces dense contextual representations of captions. These embeddings are used as features in a logistic regression classifier, allowing us to evaluate whether pretrained semantic representa-

tions improve generalization beyond handcrafted linguistic and psychological features. We select `all-MiniLM-L6-v2` due to its balance between computational efficiency and semantic performance, enabling fast inference while maintaining strong performance in sentence-level similarity tasks. This setup isolates the contribution of pretrained embeddings by using a simple linear classifier without additional modeling complexity.

3.5 Transformer-Based Model

We fine-tune `roberta-base` (Liu et al., 2019) for binary misinformation classification. RoBERTa is selected because it provides strong contextual language understanding and serves as a high-capacity baseline to evaluate whether deep semantic representations alone can capture misinformation patterns without explicit feature engineering. Captions are tokenized with a maximum length of 256 tokens. Approximately 39% of captions exceed the 256-token limit and are truncated during tokenization. Truncation primarily affects longer narrative-style posts, whereas shorter captions remain fully preserved. The model is trained using a learning rate of 2×10^{-5}, batch size of 16, weight decay of 0.01, and early stopping based on validation set performance.

3.6 Evaluation: In-Domain & Cross-Domain

We evaluate models under two settings: **In-Domain Evaluation:** For each dataset (omega-6 and seed oils), we perform stratified 5-fold cross-validation, ensuring balanced class distributions across folds. This procedure is applied consistently across all three modeling approaches. **Cross-Domain Evaluation:** We conduct bidirectional transfer experiments to evaluate cross-domain generalization. Specifically, models are trained on the full source-domain dataset and evaluated on the target-domain dataset without any adaptation. We consider omega-6$\rightarrow$ seed (training on omega-6, testing on seed oils) and seed $\rightarrow$ omega-6 (training on seed oils, testing on omega-6). This setup evaluates robustness under realistic domain shift, where topic-specific cues may not transfer across domains.

All experiments are conducted with a fixed random seed of 42 to ensure reproducibility. In the cross-domain setting, no information from the target domain is used during training or model selection, thereby preventing any form of data leakage.

3.7 Evaluation Metrics

We use evaluation metrics suited for imbalanced classification and misinformation detection, including precision, recall, and F1-score (for the misinformation class), as well as the area under the precision–recall curve (AUPRC). F1-score is the harmonic mean of precision and recall, providing a balanced measure under class imbalance. AUPRC is a threshold-independent metric that summarizes ranking performance and is particularly informative when the positive class is of primary interest. Although the class distribution is moderately imbalanced, we use AUPRC because it evaluates ranking performance independent of decision thresholds and provides a more informative measure of classifier behavior under non-uniform class distributions.

4 Results

4.1 In-Domain Performance

All models perform well on the misinformation class, with consistent trends across both domains (Table 1). To account for class imbalance, we compare against baselines based on majority-class and prevalence, which produce identical performance with AUPRC values of 0.77 (seed oils) and 0.70 (omega-6), reflecting the distributions of the underlying labels. All proposed models consistently outperform these baselines under in-domain settings and achieve higher ranking performance than baseline under most cross-domain transfers, indicating that improvements are not driven solely by class imbalance.

Sentence-BERT (MiniLM) achieves the highest overall performance, reaching 0.96 AUPRC on seed oils and 0.92 on omega-6, outperforming both RoBERTa and the feature-based Random Forest model in metrics. RoBERTa achieves perfect recall (1.00) across both datasets, but with lower precision (0.77 on omega-6, 0.72 on seed oils), resulting in reduced F1. This indicates a consistent tendency to over-predict the positive class.

The Random Forest model remains competitive, reaching F1 scores of 0.87 (omega-6) and 0.86 (seed oils). Despite using only engineered linguistic and rhetorical features, it performs close to neural approaches, indicating that structured features retain strong predictive value.

4.2 Cross-Domain Generalization

Across all models, AUPRC drops under cross-domain transfer (Table 2), indicating reduced rank-

Model	Domain	Prec	Rec	F1	AUPRC
MiniLM	omega-6	0.90	0.89	0.90	0.92
MiniLM	seed oils	0.87	0.89	0.88	0.96
RoBERTa	omega-6	0.77	1.00	0.87	0.86
RoBERTa	seed oils	0.72	1.00	0.83	0.84
Random Forest	omega-6	0.80	0.96	0.87	0.90
Random Forest	seed oils	0.79	0.95	0.86	0.84

Table 1: In-domain performance.

Train → Test	Model	Prec	Rec	F1	AUPRC
omega-6→ seed	MiniLM	0.72	1.00	0.83	0.74
omega-6→ seed	RoBERTa	0.72	1.00	0.83	0.69
omega-6→ seed	RF	0.72	1.00	0.83	0.74
seed → omega-6	MiniLM	0.77	0.95	0.85	0.86
seed → omega-6	RoBERTa	0.77	1.00	0.87	0.82
seed → omega-6	RF	0.77	0.90	0.83	0.80

Table 2: Cross-domain performance.

ing performance when models are applied to unseen topics. For example, MiniLM decreases from 0.92 in-domain on omega-6 to 0.74 when transferred from omega-6 to seed oils, while RoBERTa decreases from 0.86 to 0.69 in the same setting. The degradation is more pronounced for the omega-6→seed transfer direction (Figure 3).

Sentence-BERT (all-MiniLM-L6-v2) achieves the strongest overall in-domain performance and competitive cross-domain AUPRC across both transfer settings, suggesting relatively strong generalization ability compared to the other approaches.

The Random Forest model also demonstrates competitive cross-domain robustness, particularly in the seed→omega-6 setting, where it shows only a modest decrease in AUPRC (0.84 to 0.80). This suggests that topic-agnostic linguistic and rhetorical features may capture more transferable misinformation signals. In particular, rhetorical features (e.g., certainty markers) may generalize more consistently across topics than dense semantic representations.

4.3 Linguistic and Rhetorical Differences

Readability analysis indicates that captions across both classes are moderately complex, with no strong separation in Flesch Reading Ease scores. In the omega-6 domain, credible posts exhibit slightly higher readability (M = 52.95, SD = 11.76) than misinformation (M = 49.26, SD = 15.19). A similar pattern is observed in the seed oil domain (credible: M = 56.57, SD = 15.98; misinformation: M = 51.40, SD = 23.15). However, the substantial overlap and variance suggest that readability is not a primary distinguishing factor.

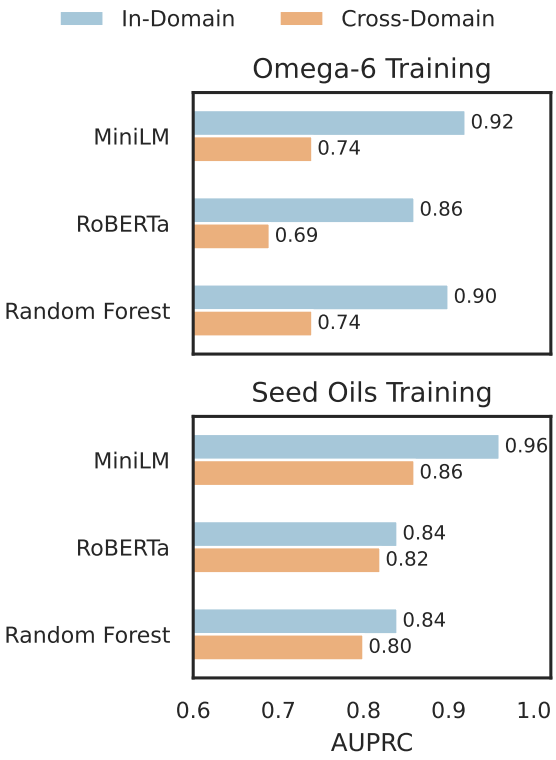

Figure 3: AUPRC comparison across in-domain and cross-domain transfer for omega-6 and seed oil datasets.

Clearer distinctions emerge in rhetorical structure (Figure 4). Across both domains, misinformation consistently exhibits lower levels of hedging, indicating a more assertive and less qualified communication style, while credible posts show more frequent use of hedging and exploratory language.

Misinformation is also characterized by stronger reliance on persuasive and heuristic framing, including more frequent appeal-to-nature arguments and the presence of conspiracy-related cues, which are largely absent from credible posts. In contrast, credible posts show slightly higher levels of exploratory questioning, suggesting a more open-ended or deliberative tone. Misinformation additionally exhibits more emphatic signaling (e.g., exclamation usage and capitalization), particularly in the seed oil domain. Overall, the differences are primarily rhetorical rather than lexical or readability-based, with misinformation distinguished by higher certainty, stronger framing devices, and reduced epistemic caution.

4.4 Affective/Psychological Feature Analysis

Affective analysis reveals a nuanced pattern suggesting affective simplification in misinformation. In both domains, misinformation shows a more con-

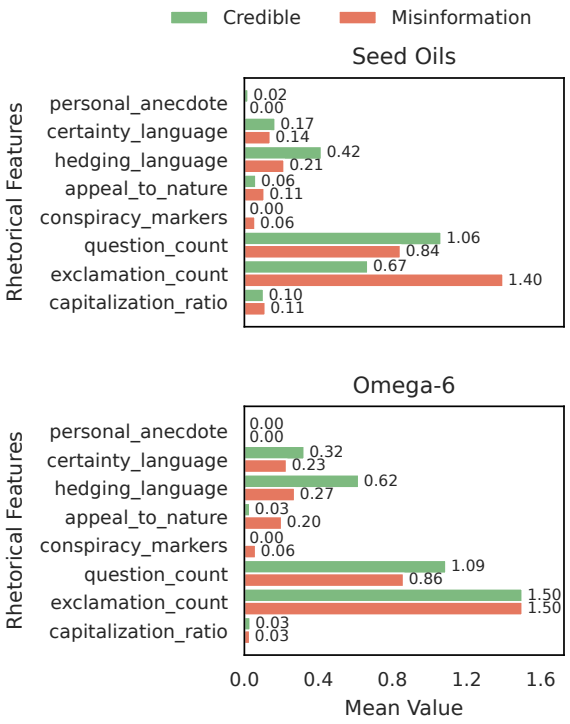

Figure 4: Rhetorical and linguistic feature distributions across credible and misinformation classes.

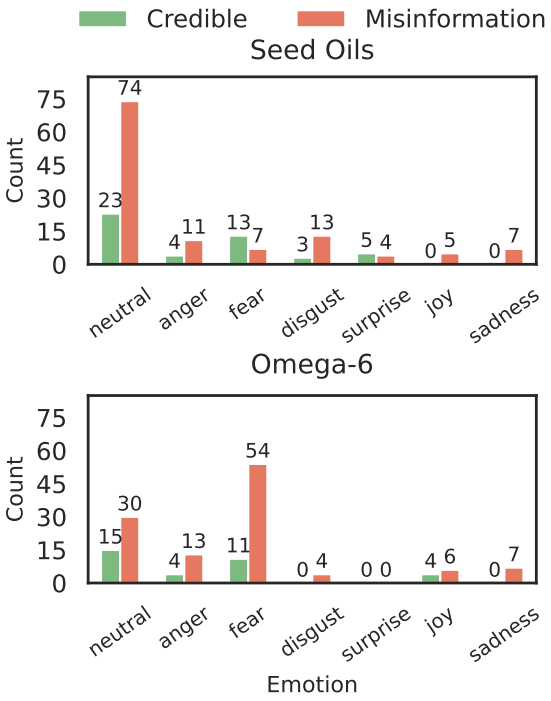

Figure 5: Emotion distribution across credible and misinformation posts in the seed oil and omega-6 domains.

centrated emotional profile, while credible content exhibits a less skewed distribution of affect (Figure 5). In the seed oil domain, misinformation is dominated by neutral affect with secondary anger and fear, whereas credible posts span a broader set of emotions, including fear and disgust. In the omega-6 domain, misinformation is strongly concentrated in fear-related language, while credible content distributes more evenly across negative and mixed emotions. Overall, misinformation is characterized by reduced emotional variability, while credible content shows a more heterogeneous affective distribution.

Across both domains, misinformation captions show lower scores across psychological dimensions, including valence, emotional intensity, and reward/threat systems, indicating reduced emotional richness. These findings suggest that misinformation is characterized not by emotional amplification, but by simplified and less varied affective expression, which may increase perceived clarity and directness.

4.5 Feature Importance and Interpretability

A total of 55 features were extracted per domain. SHAP analysis was conducted over the entire feature space without feature selection to provide an unbiased estimate of the importance of the feature. Figure 6 presents the top 10 features ranked by mean absolute SHAP values.

Across both domains, the most influential features are primarily psychological and affective. Top features include RDoC-based measures of threat and valence, emotion-neutral signals, and sentiment-related features. Rhetorical features, including appeal to nature and question frequency, also contribute to model predictions, highlighting the importance of communicative framing.

The consistency of these features across domains indicates that misinformation is characterized by stable, topic-agnostic signals, including reduced hedging, lower emotional complexity, and increased reliance on persuasive heuristics. These patterns help explain model behavior under domain shift: models likely maintain high recall by capturing generalizable stylistic signals, but may experience reduced AUPRC due to the loss of topic-specific semantic information.

5 Conclusion

In this work, we presented a multi-view framework for detecting nutrition misinformation in Instagram posts, focusing on seed oils and omega-6. We introduced a curated, expert-annotated dataset and eval-

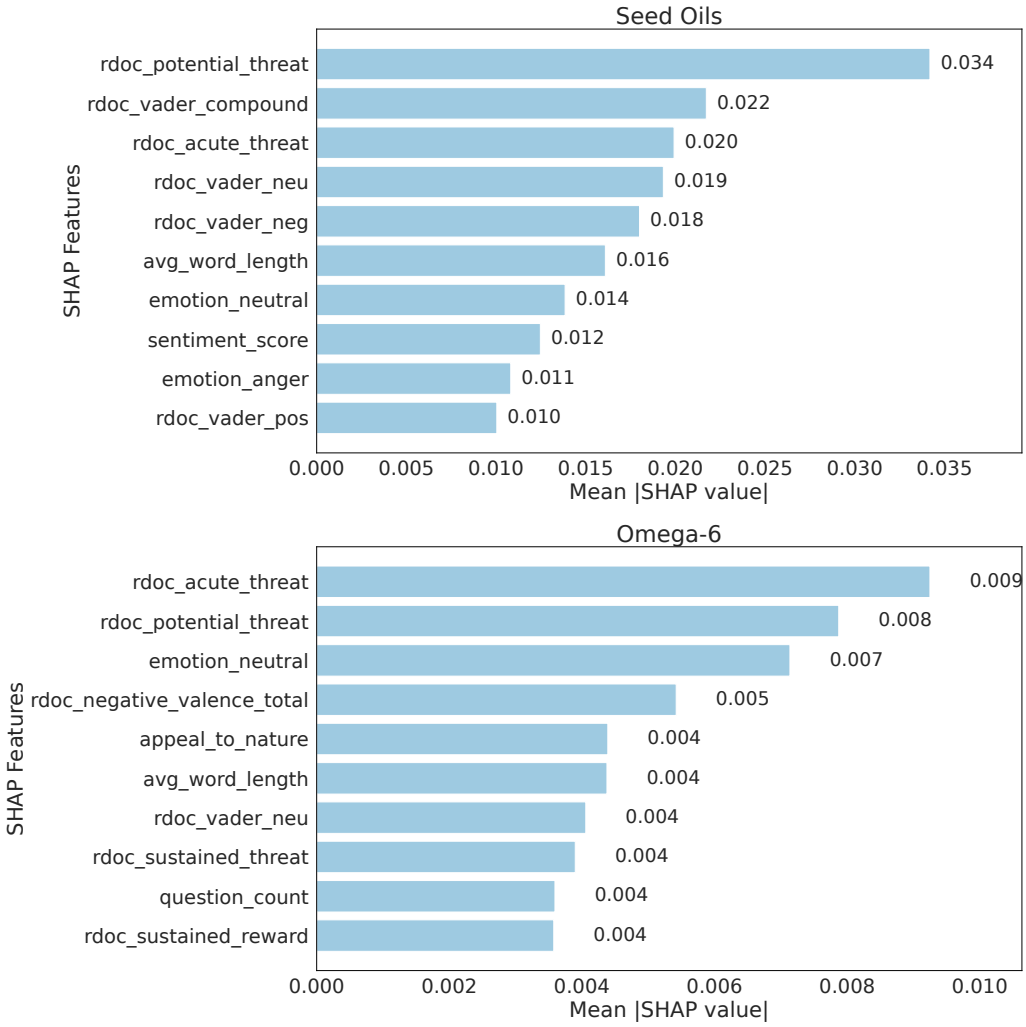

Figure 6: Top 10 features ranked by mean absolute SHAP values for the Random Forest model.

uated both feature-based and transformer-based models under in-domain and cross-domain transfer. Our results show that Sentence-BERT achieves the strongest in-domain performance, while RoBERTa prioritizes recall at the expense of precision. Notably, a feature-based Random Forest model leveraging linguistic, rhetorical, and psycholinguistic features performs competitively, highlighting the value of interpretable representations alongside deep semantic models.

We observe substantial performance degradation under cross-domain evaluation, even between closely related nutrition topics, underscoring limited generalization under domain shift. Beyond predictive performance, our analysis reveals consistent stylistic and affective patterns in misinformation, characterized by rhetorical certainty, reduced hedging, and simplified emotional expression. Feature importance analysis further confirms that psychological and affective signals are among the most informative predictors. Overall, these findings highlight the complementary strengths of interpretable linguistic features and contextual embeddings for robust and explainable nutrition misinformation detection, and motivate future work on multimodal and cross-platform generalization.

Limitations

The dataset used in this study is relatively small and restricted to two closely related nutrition topics, namely seed oils and omega-6. While this focused design allows for controlled analysis of domain shift, it may limit the broader applicability of the findings to other areas of nutrition or health misinformation.

The analysis is also limited to English-language Instagram captions, which excludes multilingual content and may not capture cultural or linguistic variations in how nutrition misinformation is expressed globally. In addition, only textual caption

data is considered, while Instagram is inherently multimodal and includes images, videos, and user engagement signals that may provide important contextual cues for misinformation detection.

Although the annotation process was guided by detailed expert-defined guidelines and inter-annotator agreement was measured, labeling misinformation inevitably involves a degree of subjectivity. Despite efforts to reduce bias through training and independent annotation, some ambiguity may persist, particularly in borderline cases where claims are partially supported or selectively framed. We also acknowledge that nutrition science is an evolving field and that dietary guidance may change over time as new evidence emerges. As with all evidence-based guidelines, the DGA reflects the prevailing scientific consensus at a particular point in time and is therefore subject to revision as the evidence base develops. Accordingly, the annotation framework used in this study should be interpreted as reflecting alignment with contemporaneous public health guidance during the data collection period rather than immutable scientific ground truth.

Finally, while the proposed feature set captures linguistic, rhetorical, and psycholinguistic signals, it is still based on predefined assumptions about informative cues. Transformer-based models are also constrained by the relatively small dataset size and may benefit from larger-scale or domain-adaptive training. Cross-domain evaluation is limited to topic transfer within Instagram, and broader cross-platform studies would be necessary to fully assess real-world generalization.

Ethical Considerations

This study uses only publicly available Instagram content and analyzes caption text without collecting private or sensitive user data. All identifiers were removed to ensure anonymity, and the dataset is fully de-identified. The models developed in this work are intended for research purposes and should be interpreted as probabilistic indicators rather than definitive judgments of truth or falsehood. Our annotation framework operationalizes misinformation relative to contemporaneous evidence-based public health guidance, specifically the 2020–2025 Dietary Guidelines for Americans and the accompanying Scientific Report of the Dietary Guidelines Advisory Committee. These guidelines are developed through systematic evidence review and expert synthesis of the nutrition literature and are intended to reflect prevailing scientific consensus at the time of publication. Consistent with prior public health misinformation research, we evaluate claims relative to established evidence-based recommendations rather than isolated or emerging studies. We acknowledge that nutrition science continues to evolve and that there may be some scientific disagreement; however, individual conflicting findings do not necessarily outweigh the broader consensus reflected in guideline-based recommendations. Care is required when applying misinformation detection systems in real-world settings, as misclassification risks may disproportionately affect nuanced or evolving scientific discussions. This work aims to support, rather than replace, expert evaluation of nutrition-related content.

Acknowledgments

We thank the nutrition experts, Dr. Charlotte Martin, DCN, RDN, and Dr. Alan Flanagan, PhD, for their valuable contributions to screening/extracting/annotating the original Instagram posts. The authors also gratefully acknowledge the School of Computing, College of Computing, Engineering and Construction, and Graduate School at the University of North Florida for their support and funding, which made this research possible.

References

Sharique Ahmad, Subuhi Anwar, and Pushpendra D. Pratap. 2025. The cod liver oil insights: From folklore to facts. *International Journal of Research - GRANTHAALAYAH*, 13(8):95–109.

Jamal Belkhouribchia and Joeri Jan Pen. 2025. Large language models in clinical nutrition: an overview of its applications, capabilities, limitations, and potential future prospects. *Frontiers in Nutrition*, 12:1635682.

Brian Dean. 2025. Instagram demographic statistics: How many people use instagram in 2024? Accessed: 2025-04-25.

Connie Diekman, Camille D Ryan, and Tracy L Oliver. 2023. Misinformation and disinformation in food science and nutrition: impact on practice. *The Journal of Nutrition*, 153(1):3–9.

Dietary Guidelines for Americans. 2020. Dietary guidelines for americans, 2020–2025 and online materials. https://www.dietaryguidelines.gov/resources/2020-2025-dietary-guidelines-online-materials.

Reyhaneh Farhoundinia, Daniel Munro, and Mark de Rond. 2024. Emotions unveiled: detecting COVID-19 fake news on social media. *Humanities and Social Sciences Communications*, 11(1):1–14.

Tanjim Bin Faruk. 2024. Evaluating the performance of large language models in scientific claim detection and classification. https://arxiv.org/abs/2412.16486.

Shubham Garg and Sonal Sharma. 2022. Linguistic features based framework for automatic fake news detection. *Computers & Industrial Engineering*, 172:108432.

Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. 2025. Fighting misinformation in health news: Dcnn-capsnet for cross-domain health misinformation detection. *ACM Transactions on Intelligent Systems and Technology*.

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Thomas R. Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S. Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (rdoc): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7):748–756.

Janne Kauttonen, Jenni Hannukainen, Pia Tikka, and Jyrki Suomala. 2020. Predictive modeling for trustworthiness and other subjective text properties in online nutrition and health communication. *PloS one*, 15(8):e0237144.

Jahanzaib Zain Khan, Basheer Alam, and Youngmoon Lee. 2021. An approach utilizing linguistic features for fake news detection. In *Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1–6. IEEE.

Kevin Lee and Keszya Kurniawan. 2025. Are seed oils the culprit in cardiometabolic and chronic diseases? a narrative review. *Nutrition Reviews*, 83(7):e2106–e2112.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Sinthia Mahbub, Eric Pardede, and A. S. M. Kayes. 2022. COVID-19 rumor detection using psycholinguistic features. *IEEE Access*, 10:117530–117543.

Cameron Martel, Gordon Pennycook, and David G. Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1):1–20.

Van Nguyen, Luke Testa, Andrea L Smith, Louise A Ellis, Adam G Dunn, Jeffrey Braithwaite, and Mitchell Sarkies. 2022. Unravelling the truth: Examining the evidence for health-related claims made by naturopathic influencers on social media–a retrospective analysis. *Health Promotion Perspectives*, 12(4):372.

Sergio Segado-Fernández, Beatriz Jiménez-Gómez, Pedro Jesús Jiménez-Hidalg, María del Carmen Lozano-Estevan, and Iván Herrera-Peco. 2025. Disinformation about diet and nutrition on social networks: a review of the literature. *Nutrición Hospitalaria*, 42(2).

Wen-Ying Sylvia Chou, Anna Gaysynsky, and Joseph N Cappella. 2020. Where we go from here: health misinformation on social media.

Dongmei Tan, Yi Huang, Ming Liu, Ziyu Li, Xiaoqian Wu, and Cheng Huang. 2025. Identification of online health information using large pretrained language models: Mixed methods study. *Journal of Medical Internet Research*, 27:e70733.

Pawan Kumar Verma, Priyanka Agrawal, Isabelle Amorim, and Radu Prodan. 2021. WELFake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Christiane Wurst, Paul Sauseng, Shruthi Prasad, Anchali Trakulp, Cameron Martel, and David Rand. 2024. Emotions in misinformation studies: distinguishing affective state from emotional response and misinformation recognition from acceptance. *Cognitive Research: Principles and Implications*, 9(1):1–18.

Constantinos-Giovanni Xarhoulacos, Argiro Anagnostopoulou, George Stergiopoulos, and Dimitris Gritzalis. 2021. Misinformation vs. situational awareness: The art of deception and the need for cross-domain detection. *Sensors*, 21(16):5496.

Andy Wai Kan Yeung, Anela Tosevska, Elisabeth Klager, Fabian Eibensteiner, Christos Tsagkaris, Emil D Parvanov, Faisal A Nawaz, Sabine Völkl-Kernstock, Eva Schaden, Maria Kletecka-Pulker, and 1 others. 2022. Medical and health-related misinformation on social media: bibliometric study of the scientific literature. *Journal of Medical Internet Research*, 24(1):e28152.

Appendix

A Annotation Guidelines

A.1 Purpose of These Guidelines

These guidelines ensure that all human coders classify social media nutrition posts consistently using:

- A two-level accuracy scale (0 or 1)

- Rules aligned with the *Dietary Guidelines for Americans, 2020–2025 (DGA)* and the *2020 DGAC Scientific Report*

Coders should reference only the content presented in the post and apply the rules exactly as described.

A.2 What Coders Evaluate

For each post, coders assign:

1. Accuracy classification (1 or 0)
2. An optional brief justification (2–3 sentences)

Coders do **not** evaluate:

- Creator's biography
- Hashtags
- Engagement metrics
- Intent or motivation
- User comments
- Information not shown in the post itself

A.3 Evidence Standard

Coders must evaluate claims solely against:

- *Dietary Guidelines for Americans 2020–2025*
- *DGAC Scientific Report (2020)*

Coders should **not** use:

- Personal knowledge
- Other nutrition guidelines
- Mechanistic physiology not discussed in the DGA/DGAC
- External sources, papers, or reputable organizations
- Internet searches

A.4 The Two-Level Accuracy Classification

1 - Accurate or Mostly Accurate A post is **Accurate** when:

- All nutrition claims fully align with DGA/DGAC
- Any minor inaccuracies do not change the overall message

Examples of Accurate Content

- Saying omega-6 fats are not inflammatory
- Correcting misinformation about seed oils

A post is **Mostly Accurate** when:

- It is evidence-aligned overall
- It contains some non-trivial inaccuracies, but the main message remains correct

Typical patterns

- Slight exaggeration without reversing the evidence

0 - Mostly Inaccurate or Inaccurate A post is **Mostly Inaccurate** when:

- It contains some correct information
- Misleading or incorrect claims dominate

Typical patterns

- Mixing evidence-based advice with unsupported mechanistic claims
- Overstating inflammation, hormone effects, oxidation, or fear-based messages about oils
- Encouraging elimination of seed oils while acknowledging they contain essential fats

A post is **Inaccurate** when:

- The core message contradicts DGA/DGAC
- Accurate statements, if present, are trivial

Examples

- "Butter is heart healthy; seed oils cause disease."
- "Seed oils are toxic poisons that cause inflammation."

A.5 Reasoning Requirements (2–3 Sentences)

Coders may provide a short justification for doubtful annotations. The justification must:

- Evaluate the overall message rather than line-by-line details
- Reference only areas covered by the DGA/DGAC
- Not mention excluded content (e.g., "the DGA doesn't cover . . . ")
- Avoid SEO phrases or quoting the creator directly
- Clearly explain why the classification was chosen

A.6 Special Situations

Mixed-accuracy posts Code based on the **overall message**, not on isolated statements.

Mechanistic claims (inflammation, oxidation, hormones) If the mechanism is not discussed in the DGA:

- Ignore the mechanism itself
- Evaluate whether the dietary conclusion contradicts DGA guidance

Reactive content If a post stitches another creator's video:

- Ignore the introductory clip
- Code only the creator's commentary

B Complete Feature Definitions

This section provides formal definitions of all features used in the feature-based model. Features are grouped into linguistic, rhetorical, affective (emotion and sentiment), psychological (RDoC-based), and interaction features. All continuous features are standardized prior to model training.

B.1 Linguistic Complexity Features

word_count Total number of words in the caption.

avg_word_length Average number of characters per word.

lexical_diversity Ratio of unique words to total words (type-token ratio).

flesch_reading_ease Readability score estimating ease of comprehension (higher values indicate easier text).

flesch_kincaid_grade Estimated U.S. grade level required to understand the text.

gunning_fog Readability index estimating years of formal education required to understand the text.

B.2 Rhetorical and Persuasive Features

certainty_language Count or binary indicator of certainty expressions (e.g., "always", "never", "prove", "guarantee").

hedging_language Count or binary indicator of uncertainty expressions (e.g., "may", "might", "possibly", "perhaps").

question_count Number of question marks in the text.

exclamation_count Number of exclamation marks in the text.

capitalization_ratio Proportion of uppercase characters relative to total characters.

appeal_to_nature Indicator for naturalistic fallacy framing (e.g., "natural is better", "chemical-free implies safe").

conspiracy_markers Indicator/count of conspiracy-related phrases (e.g., "big pharma", "cover-up", "they don't want you to know").

has_personal_anecdote Binary feature indicating presence of first-person experiential claims (e.g., "I tried", "my experience", "I noticed").

B.3 Interaction Features

hedge_x_exclaim Interaction between hedging intensity and exclamation usage.

certainty_x_caps Interaction between certainty language and capitalization ratio.

anecdote_x_exclaim Interaction between personal anecdote presence and exclamation usage.

word_count_x_lexical_div Interaction between text length and lexical diversity.

B.4 Emotion Features (Transformer-Based Model)

Emotion features are probability outputs from a pretrained DistilRoBERTa emotion classifier.

emotion_anger Probability of anger-related emotion.

emotion_fear Probability of fear or threat-related emotion.

emotion_disgust Probability of disgust or aversion.

emotion_joy Probability of positive affect or happiness.

emotion_sadness Probability of sadness or loss-related affect.

emotion_surprise Probability of surprise or unexpectedness.

emotion_neutral Probability of neutral or informational tone.

B.5 Sentiment Features (VADER-Based)

sentiment_score Overall sentiment polarity score ranging from negative to positive.

sentiment_positive Proportion of tokens contributing to positive sentiment.

sentiment_negative Proportion of tokens contributing to negative sentiment.

sentiment_neutral Proportion of tokens contributing to neutral sentiment.

rdoc_vader_pos Positive sentiment score mapped into RDoC positive valence space.

rdoc_vader_neg Negative sentiment score mapped into RDoC negative valence space.

rdoc_vader_neu Neutral sentiment score from VADER.

rdoc_vader_compound Normalized aggregate sentiment score combining all polarity signals.

B.6 RDoC-Based Psychological Features

B.6.1 Valence Features

rdoc_positive_valence_total Aggregate positive appraisal, reward, and pleasure-related language.

rdoc_negative_valence_total Aggregate threat, loss, and negative appraisal language.

rdoc_valence_dimension Net valence score computed as positive minus negative valence.

B.6.2 Arousal Features

rdoc_high_arousal High-energy emotional activation (urgency, excitement, intensity).

rdoc_low_arousal Low-energy emotional states (calmness, fatigue, neutrality).

rdoc_arousal_dimension Overall emotional activation level combining arousal signals.

rdoc_arousal_total Total magnitude of arousal-related activation.

rdoc_arousal_net Net difference between high and low arousal activation.

B.6.3 Threat and Reward System Features

rdoc_acute_threat Immediate or short-term threat-related language (fear, danger, urgency).

rdoc_potential_threat Uncertain or probabilistic threat framing.

rdoc_sustained_threat Chronic or long-term threat framing.

rdoc_loss Language indicating deprivation, harm, or negative outcomes.

rdoc_frustration Expressions of anger, irritation, or blocked goals.

rdoc_reward_expectancy Anticipation of positive outcomes or benefits.

rdoc_sustained_reward Stable or ongoing reward-related language.

rdoc_approach_motivation Goal-directed or desire-driven action language.

rdoc_reward_responsiveness Sensitivity or responsiveness to reward signals.

B.6.4 Composite Psychological Features

rdoc_emotional_intensity Overall magnitude of emotional expression derived from valence and arousal strength.

rdoc_emotional_complexity Entropy-based measure of diversity in emotional signals.

rdoc_threat_reward_ratio Ratio of threat-related to reward-related language intensity.

rdoc_valence_arousal_interaction Interaction between valence and arousal dimensions.

C Rhetorical and Persuasive Pattern Rules

We use rule-based pattern matching to extract rhetorical and persuasive features from text.

Certainty expressions:

- always, never, guarantee, prove, destroy

Hedging expressions:

- may, might, possibly, perhaps

Additional rhetorical markers:

- Question count: number of "?" in text
- Exclamation count: number of "!" in text
- Capitalization ratio: proportion of uppercase characters
- Citation indicators: presence of study, research, http, www, or year (e.g., 2020)

Conspiracy-related markers (used in exploratory analysis):

- terms such as "big pharma", "they don't want", "hidden", "cover-up"

D Emotion Model Output Mapping

Emotion features are extracted using a pretrained DistilRoBERTa emotion classifier. The model outputs probability distributions over the following emotion categories:

- anger
- disgust
- fear
- joy
- sadness
- surprise
- neutral

For each text, we store the probability score for each emotion class. These are used as continuous affective features in downstream classification.

E RDoC Lexicons for Valence and Arousal Features

We use manually constructed lexicons grounded in the Research Domain Criteria (RDoC) framework to operationalize psychological valence and arousal dimensions. These lexicons are used for feature extraction in the main analysis.

Positive Valence (Reward / Approach Motivation)

- reward_words: amazing, awesome, excellent, fantastic, great, wonderful, love, enjoy, happy, joy, excited, thrilled, delighted, perfect, best, incredible, outstanding, brilliant, superb, beautiful, gorgeous, lovely, pleasant, fabulous
- reward_expectancy: hope, expect, anticipate, look forward, optimistic, confident, believe, promise, guarantee, ensure, assure, certain
- approach_motivation: want, need, desire, wish, crave, seek, pursue, strive, goal, achieve, succeed, win, gain, obtain, acquire, try, attempt, effort, motivate, drive, determined

- reward_sustained: satisfaction, fulfillment, gratification, contentment, pleasure, reward, benefit, advantage, profit, success, achievement

Negative Valence (Threat / Loss / Frustration)

- acute_threat: fear, afraid, scared, terrified, frightened, panic, terror, alarmed, worried, anxious, nervous, dread, horror, dangerous, threat, risk, harm, danger, peril, hazard
- potential_threat: uncertain, unsure, doubt, suspicious, wary, cautious, concerned, apprehensive, uneasy, uncomfortable, tense, may, might, possibly, perhaps, suggest, potentially, risk
- sustained_threat: stress, chronic, ongoing, persistent, constant, continual, always, never-ending, relentless, unending, perpetual
- loss: lose, lost, loss, missing, gone, disappear, vanish, deprive, lack, without, absent, void, empty, sad, grief, mourn, regret, sorrow, heartbreak
- frustration: frustrate, annoyed, irritated, angry, mad, furious, rage, upset, disturbed, agitated, provoked, outraged, failed, failure, unsuccessful, denied, rejected, blocked

Arousal / Regulatory Systems

- high_arousal: intense, extreme, powerful, strong, overwhelming, explosive, urgent, critical, emergency, immediate, sudden, shock, surprise, astonish, startle, energy, active, lively, dynamic, vigorous, vibrant
- low_arousal: calm, peaceful, relaxed, tranquil, serene, quiet, slow, gentle, soft, mild, subtle, tired, fatigue, weary, exhausted, drained, lethargic, boring, dull, plain, mundane
- intensity_modifiers: very, extremely, incredibly, absolutely, totally, completely, utterly, highly, deeply, truly, so, quite, really

F Full SHAP Feature Importance Rankings

Seed Oils Domain
Feature
rdoc_acute_threat
rdoc_negative_valence_total
rdoc_emotional_intensity
rdoc_valence_arousal_interaction
emotion_sadness
emotion_disgust
rdoc_valence_dimension
rdoc_positive_valence_total
rdoc_threat_reward_ratio
emotion_fear
capitalization_ratio
rdoc_arousal_dimension
gunning_fog
rdoc_arousal_total
exclamation_count
rdoc_approach_motivation
emotion_surprise
rdoc_arousal_net
word_count_x_lexical_div
flesch_kincaid_grade
sentiment_score
hedging_language
rdoc_emotional_complexity
word_count
flesch_reading_ease
rdoc_potential_threat
emotion_neutral
rdoc_vader_compound
emotion_joy
rdoc_intensity_modifiers
lexical_diversity
avg_word_length
rdoc_vader_pos
rdoc_high_arousal
question_count
rdoc_vader_neg
emotion_anger
rdoc_vader_neu
rdoc_sustained_threat
rdoc_reward_expectancy
rdoc_loss
rdoc_sustained_reward
rdoc_reward_responsiveness
sentiment_neutral
hedge_x_exclaim
certainty_x_caps
conspiracy_markers
appeal_to_nature
has_personal_anecdote
sentiment_negative
certainty_language
rdoc_low_arousal
sentiment_positive
anecdote_x_exclaim
rdoc_frustration

Omega-6 Domain
Feature
rdoc_acute_threat
rdoc_potential_threat
emotion_neutral
rdoc_negative_valence_total
avg_word_length
appeal_to_nature
emotion_anger
rdoc_vader_neu
rdoc_sustained_threat
question_count
gunning_fog
rdoc_vader_neg
rdoc_sustained_reward
certainty_language
flesch_kincaid_grade
rdoc_high_arousal
hedging_language
rdoc_approach_motivation
rdoc_frustration
capitalization_ratio
rdoc_reward_responsiveness
rdoc_vader_pos
rdoc_valence_arousal_interaction
sentiment_positive
lexical_diversity
emotion_sadness
certainty_x_caps
word_count_x_lexical_div
rdoc_arousal_total
rdoc_arousal_dimension
hedge_x_exclaim
rdoc_reward_expectancy
rdoc_positive_valence_total
emotion_disgust
rdoc_valence_dimension
rdoc_threat_reward_ratio
rdoc_low_arousal
rdoc_arousal_net
sentiment_score
exclamation_count
word_count
rdoc_vader_compound
rdoc_emotional_intensity
sentiment_negative
emotion_joy
emotion_fear
rdoc_emotional_complexity
flesch_reading_ease
sentiment_neutral
emotion_surprise
rdoc_intensity_modifiers
rdoc_loss
conspiracy_markers
anecdote_x_exclaim
has_personal_anecdote

Table 3: Full SHAP feature importance rankings for seed oils and omega-6 domains (55 features per domain).