Cognitively Inspired Feature Engineering for Predicting Online Product Review Polarity

Travis Vitello
College of Computing
Georgia Institute of Technology
Atlanta, USA
tvitello3@gatech.edu

***Abstract*—This work presents a cognitively inspired feature engineering framework for predicting online product review polarity, emphasizing interpretable and psychologically grounded representations of text. Using the UCSD Amazon "Electronics" dataset, a balanced subset of 40,000 reviews was constructed and formulated as a binary classification task (negative vs. positive).**

The proposed approach integrates multiple cognitively motivated lexical representations, including emotion mapping via the NRC Emotion Lexicon (EmoLex), sentiment polarity via Bing Liu's lexicon, and contextual sentiment progression using Valence Aware Dictionary and sEntiment Reasoner (VADER) compound scoring. To enable controlled comparisons of feature representations, twelve model configurations were constructed across Random Forest and XGBoost machine learning (ML) classifiers, including baseline and Optuna-optimized variants spanning emotion-only, emotion + sentiment, and emotion + sentiment + VADER feature sets.

Results show a consistent performance progression across feature sets, with emotion-only models demonstrating relatively worse results while the addition of contextual sentiment yielded the largest gains. The optimized XGBoost model achieved a multi-seed average 75.04% accuracy and 0.825 ROC-AUC, with comparable Random Forest performance, indicating that feature design may contribute more strongly to predictive outcome than differences in model architecture.

SHAP analysis reveals that global sentiment signals dominate predictions, while emotional features provide secondary refinement. Overall, the results demonstrate that structured, cognitively informed feature design can improve classification performance while maintaining interpretability.

***Keywords*—Sentiment Analysis, Feature Engineering, Explainable Machine Learning, Product Review Classification.**

I. Introduction

Online product reviews play a central role in modern consumer decision-making, with a majority of customers consulting reviews prior to purchase [1]. As a result, the ability to automatically interpret and classify review content has become an important problem in both natural language processing (NLP) and applied machine learning. However, review text is not simply a collection of words conveying polarity; it reflects underlying human cognition, including emotional expression, evaluative reasoning, and the progression of sentiment across a narrative [2]. Interpreting review text relative to an assigned rating can therefore be viewed as a complex cognitive task involving the integration of emotional, evaluative, and contextual signals.

Traditional sentiment analysis methods often rely on bag-of-words or lexicon-based representations that treat words as independent signals, ignoring structure, context, and temporal progression. While these approaches are interpretable and computationally efficient, they may fail to capture how sentiment evolves within a review. In contrast, modern deep learning methods can model contextual dependencies but often sacrifice interpretability, limiting their applicability in domains where transparency and explainability are required [3].

This work explores an intermediate approach: leveraging interpretable, lexicon-based features while structuring them to reflect cognitively meaningful signals. The approach is conceptually motivated by Thagard's Computational-Representational Understanding of Mind (CRUM) paradigm, which characterizes cognition as arising from interactions among representations, emotions, and decision processes [4]. In lieu of treating cognitive theory as a predictive model, this work uses it to guide feature construction, embedding interpretable proxies for emotional and evaluative reasoning into machine learning inputs.

To transact this idea, review text is represented through three complementary components: (1) emotion mapping via the NRC Emotion Lexicon (EmoLex) [5], aligned with structured models of emotional categories [6]; (2) sentiment polarity scoring using Bing Liu's lexicon [7]; and (3) contextual sentiment per Valence Aware Dictionary and sEntiment Reasoner (VADER) compound scoring, which incorporates linguistic modifiers, negation, and limited ordering effects [8]. These features are transformed into normalized lexical counts and aggregate sentiment measures, enabling consistent representation of emotional tone, polarity, and contextual sentiment across reviews. This approach evaluates alignment between linguistic content and user ratings, where accurate classification reflects effective sentiment alignment, and misclassification highlights ambiguity or complexity. Ekman's theory of basic emotions provides a foundational methodology for interpreting these outcomes [9].

Using the Amazon "Electronics" dataset from the University of California San Diego (UCSD), this study presents a binary classification task distinguishing negative (1–2 star) from positive (4–5 star) reviews [10]. A balanced subset of the data was constructed to support controlled evaluation, and the full experimental pipeline was repeated across multiple random

seeds to improve statistical reliability.

To evaluate the contribution of feature design and modeling choices, twelve model configurations were constructed across Random Forest and XGBoost classifiers, including baseline and Optuna-optimized variants spanning emotion-only, emotion + sentiment, and emotion + sentiment + VADER feature sets. This design enables controlled comparisons that isolate the effects of feature representation from those of model architecture and optimization. Tree-based models were selected to balance predictive performance with interpretability, allowing feature contributions to be analyzed directly [11]. Model performance is evaluated using standard classification metrics, while interpretability of the top-performing model is examined using SHapley Additive exPlanations (SHAP) [12].

The primary contribution of this work is the structured integration and evaluation of cognitively motivated, interpretable feature representations within a controlled experimental system. By combining emotion, sentiment, and contextual progression with systematic multi-model comparison and repeated sampling, this study demonstrates how feature design influences both predictive performance and model interpretability. More broadly, it examines whether structured feature engineering can approximate meaningful aspects of human sentiment judgment in an online review context.

II. Model and Tool Design

A. Data Preparation

This study utilizes the Amazon "Electronics" review dataset from UCSD, comprising approximately 1.7 million reviews. To construct a balanced and controlled experimental dataset, a stratified sampling approach was applied, randomly sub-setting 10,000 reviews each from 1-, 2-, 4-, and 5-star categories, resulting in a total population of 40,000 reviews. This subset balances computational efficiency with sufficient data diversity. Neutral (3-star) reviews were excluded to focus the task on binary classification of negative (1–2 stars) versus positive (4–5 stars), with labels encoded as 0 and 1, respectively. The experimental pipeline was repeated across five independent random seeds, with performance metrics averaged across trials.

Review text for EmoLex and Bing Liu scoring was prepared using standard natural language processing techniques. Text was tokenized and lemmatized using the NLTK *WordNetLemmatizer*, whereby part-of-speech tagging was applied to improve lemmatization accuracy (such as "received" to "receive") [13]. The *fix* method of the Python *contractions* library was applied to expand terms like "don't" into "do not". Then, the NLTK English stop word removal process was applied to eliminate high-frequency, low-information terms (such as "the" or "and") that contribute limited semantic value and may introduce noise [14]. This approach standardizes communicative forms while preserving semantic content needed for lexicon-based mapping. For VADER scoring, the original review text was used with minimal preprocessing (contraction expansion only), ensuring that word order, negations, and contextual modifiers were preserved.

B. Feature Engineering

Feature construction was designed to capture complementary aspects of sentiment and emotion while preserving interpretability and alignment with cognitively motivated representations of text. Guided by CRUM-inspired principles, features were organized to reflect interacting components of evaluation: emotional state (EmoLex categories), evaluative polarity (Bing Liu sentiment scores), and contextual sentiment progression (VADER compound scoring). This structure approximates how humans integrate contextual signals when forming judgments, consistent with CRUM's emphasis on combining representations in cognitive processing.

This study's feature groups were defined as:

1) Emotion Features (EmoLex): Each review was mapped to eight core emotional categories ("joy", "trust", "fear", "surprise", "sadness", "disgust", "anger", and "anticipation") using the NRC Emotion Lexicon. Token-level matches were counted and normalized by total token count to produce relative measures of emotional intensity, enabling comparison across reviews of varying length.

2) Sentiment Polarity Features (Bing Liu): Review tokens were matched against Bing Liu's positive and negative sentiment lexicons. Counts were normalized to produce relative proportions of positive and negative terms, capturing overall evaluative polarity based on the presence of sentiment-bearing words.

3) Contextual Sentiment Feature (VADER): Each review was scored using the VADER sentiment analyzer, which produces a normalized compound score in the range [-1, 1] reflecting overall sentiment. Unlike the count-based polarity features, VADER considers linguistic modifiers, negation, punctuation, and word order, providing a context-aware representation of how sentiment is expressed within a review.

While the EmoLex and Bing Liu features capture sentiment through aggregated lexical signals independent of order (akin to a bag-of-words approach), the inclusion of VADER introduces a complementary context-sensitive measure. Together, these features form a hybrid representation that balances interpretability with sensitivity to linguistic structure.

C. Analytical Model Design

Two supervised, tree-based ensemble methods were selected for classification: Random Forest (per *scikit-learn*) and XGBoost. These models were chosen for their strong performance on tabular data and their ability to support interpretable analysis of feature contributions. In contrast to transformer-based approaches (e.g., BERT or XLNet), which prioritize predictive performance at the cost of transparency, this study emphasizes controlled evaluation of cognitively inspired features within a simplified and interpretable modeling framework [11].

To explore the effects of feature design and model tuning, twelve model configurations were constructed across:

- **Feature sets:** Emotion only; Emotion + Sentiment; Emotion + Sentiment + VADER
- **Algorithms:** Random Forest and XGBoost
- **Model types:** Baseline and Optuna-optimized

This experimental design enables controlled comparisons that isolate the contribution of each feature group while assessing the impact of hyperparameter optimization. Optuna was used for tuning, employing adaptive sampling and pruning to identify improved configurations within each model [15]. Optimization was intentionally limited to model parameters rather than to the introduction of additional features, ensuring that performance gains reflect the effectiveness of the selected cognitively inspired feature representations *only*. For each experimental run, an 80–20 stratified train-test split was applied, resulting in 32,000 training samples and 8,000 test samples per trial. Given the balanced nature of the dataset, accuracy is considered to provide a meaningful primary evaluative metric. ROC-AUC offers a threshold-independent measure of discriminative performance of the top-performing model (#9).

D. Model Interpretability

To analyze feature contributions and support interpretability, SHAP was applied using the *TreeExplainer* method. SHAP values quantify the contribution of each feature to individual predictions, enabling both local and global analysis of model behavior. This analysis serves two purposes. First, it verifies that model predictions align with expected sentiment patterns, such as positive polarity contributing to positive classifications and negative polarity contributing to negative classifications. Second, it provides a structured template for evaluating the role of cognitively inspired features, illustrating how emotional and contextual signals influence classification outcomes. While SHAP does not model human cognition directly, it enables transparent examination of how cognitively motivated feature representations may be utilized within the model for review category prediction.

III. Results

Model performance was evaluated across five independent random seed trials, with all reported metrics representing the mean and standard deviation across runs. This repeated sampling approach reduces sensitivity to train-test splits and provides more stable estimates of model performance, consistent with established resampling and evaluation practices [16].

A. Model Configurations

Feature contributions and modeling choices were systematically evaluated across six baseline configurations spanning feature sets and algorithms, as defined in Table I. Corresponding Optuna-optimized variants were subsequently developed for each configuration; see Table II.

TABLE I: Baseline Model Configurations

Model	Algorithm	Features
1	XGBoost	Emotion Only
2	XGBoost	Emotion + Sentiment
3	XGBoost	Emotion + Sentiment + VADER
4	Random Forest	Emotion Only
5	Random Forest	Emotion + Sentiment
6	Random Forest	Emotion + Sentiment + VADER

TABLE II: Optimized Model Configurations

Model	Algorithm	Features
7	XGBoost	Emotion Only
8	XGBoost	Emotion + Sentiment
9	**XGBoost**	**Emotion + Sentiment + VADER**
10	Random Forest	Emotion Only
11	Random Forest	Emotion + Sentiment
12	Random Forest	Emotion + Sentiment + VADER

B. Performance Results

Tables III–VI represent average performance across feature sets, algorithms, and optimization strategies in percent (%) along with standard deviation across the trial population for all five random seeds.

TABLE III: XGBoost Baseline Performance (%)

Features	Acc.	Prec.	Recall	F1
Emotion	63.44 ± 0.34	61.87 ± 0.92	70.26 ± 3.86	65.74 ± 1.18
Emotion + Sent.	70.12 ± 0.72	69.04 ± 0.90	72.98 ± 0.31	70.95 ± 0.54
Emotion + Sent. + VADER	72.98 ± 0.99	72.65 ± 1.40	73.79 ± 1.52	73.20 ± 0.88

TABLE IV: Random Forest Baseline Performance (%)

Features	Acc.	Prec.	Recall	F1
Emotion	64.26 ± 0.40	62.76 ± 0.34	70.15 ± 1.10	66.25 ± 0.55
Emotion + Sent.	70.28 ± 0.71	69.35 ± 0.76	72.71 ± 1.02	70.99 ± 0.71
Emotion + Sent. + VADER	73.88 ± 0.46	74.97 ± 0.87	71.71 ± 1.09	73.30 ± 0.47

TABLE V: XGBoost Optimized Performance (%)

Features	Acc.	Prec.	Recall	F1
Emotion	64.98 ± 0.31	63.10 ± 0.41	72.14 ± 0.46	67.32 ± 0.17
Emotion + Sent.	71.14 ± 0.64	70.19 ± 0.87	73.53 ± 0.85	71.82 ± 0.52
Emotion + Sent. + VADER	**75.04 ± 0.52**	**75.26 ± 0.59**	**74.63 ± 0.58**	**74.94 ± 0.51**

TABLE VI: Random Forest Optimized Performance (%)

Features	Acc.	Prec.	Recall	F1
Emotion	64.93 ± 0.37	63.00 ± 0.49	72.38 ± 0.95	67.36 ± 0.34
Emotion + Sent.	70.95 ± 0.77	70.16 ± 0.96	72.93 ± 1.07	71.51 ± 0.71
Emotion + Sent. + VADER	74.82 ± 0.52	75.38 ± 0.60	73.73 ± 0.51	74.54 ± 0.51

Across all models, a consistent progression in performance is observed. In an ablation-style evaluation of feature sets, emotion-only models produce the weakest results, while the addition of sentiment polarity yields substantial improvements. The inclusion of contextual sentiment by means of VADER produces the largest performance gains across both algorithms, reflecting the added value of incorporating contextual and compositional information. The optimized XGBoost model using Emotion + Sentiment + VADER features achieves the best overall performance, reaching 75.04% accuracy, 75.26% precision, 74.63% recall, and 74.94% F1 score. Comparable Random Forest models produce slightly lower but competitive results, indicating that feature design may contribute more strongly to performance than algorithm selection alone. While overall performance plateaus near 75%, this likely reflects the inherent complexity of human language and the limitations of primarily lexicon-based representations.

C. ROC Analysis

The receiver operating characteristic (ROC) curve for model #9 is shown in Fig. 1. The average AUC of 0.825 ± 0.005 indicates strong discriminative capability well above random baseline, recognizing that an AUC for a binary response model between 0.8 and 0.9 is considered to embody "excellent discrimination" according to Hosmer et al. [17].

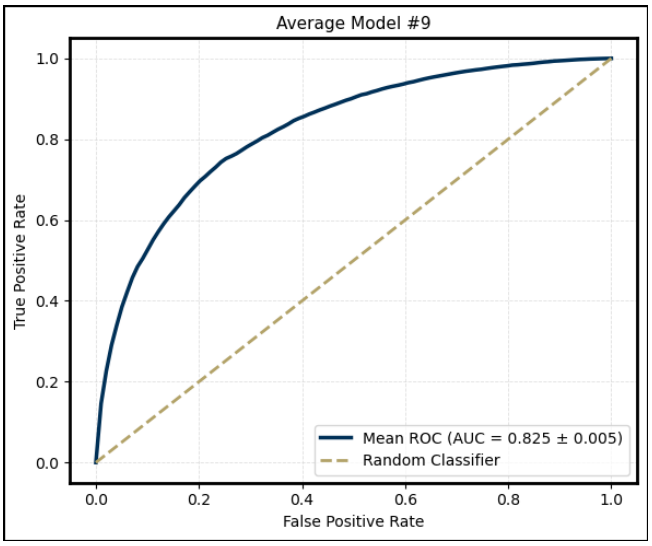

Fig. 1: Average ROC curve for the optimized XGBoost model using all features (model #9).

D. Confusion Matrix Analysis

The averaged confusion matrix for the optimized XGBoost model is shown in Fig. 2. The model demonstrates balanced classification performance, with comparable true positive and true negative rates.

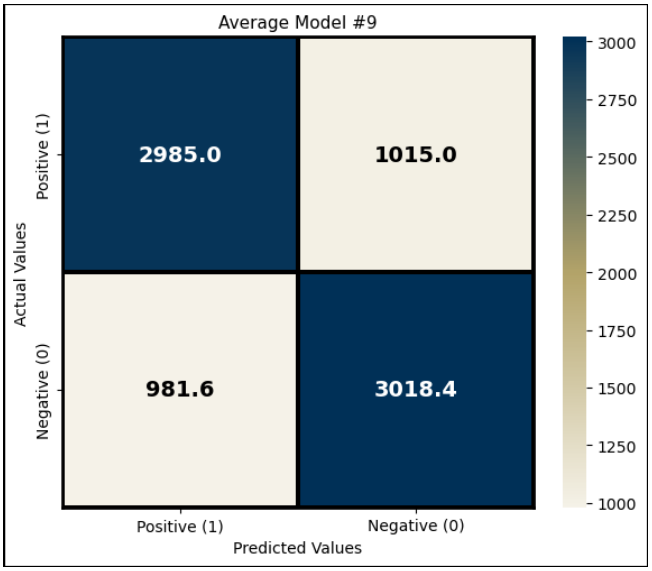

Fig. 2: Average confusion matrix for the optimized XGBoost model using all features (model #9).

E. Model Interpretability

SHAP analysis for the optimized model is shown in Fig. 3. Results indicate that global sentiment features, particularly the VADER compound score, serve as the primary drivers of model predictions, while emotional dimensions such as "sadness", "fear", and "joy" provide consistent secondary contributions. "Disgust" was found to be the most significant emotion in the average dataset considered.

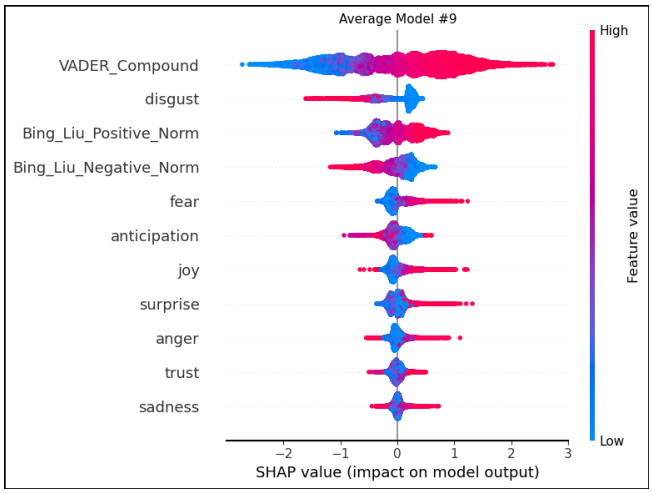

Fig. 3: SHAP summary plot showing feature contributions for the optimized XGBoost model using all features (model #9).

F. Feature Importance

Feature importance rankings for the optimized model are shown in Fig. 4. The VADER compound score is found to be the most influential model feature, again followed by "disgust" and normalized sentiment polarities derived from Bing Liu.

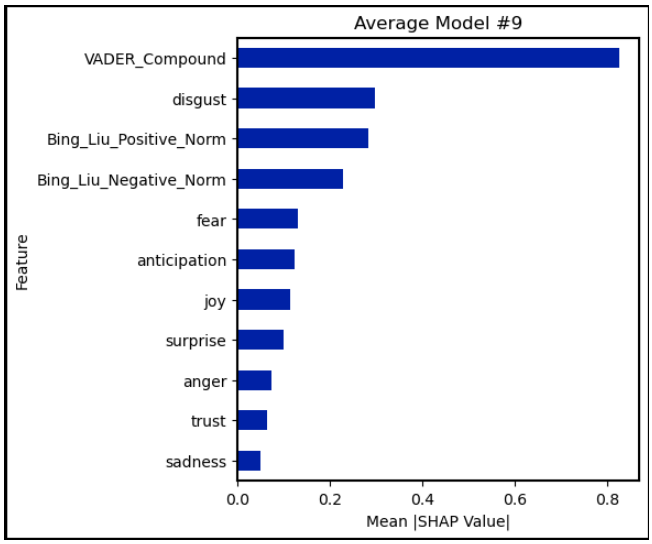

Fig. 4: Mean absolute SHAP values indicating feature importance for the optimized XGBoost model using all features (model #9).

G. Key Findings

Five primary findings emerge:

- Including sentiment (per Bing Liu mapping and contextually via VADER) features to emotion-only representations improves predictive performance by $>8\%$.
- Within the combined feature set, contextual sentiment emerges as the dominant predictor, exhibiting the highest contribution in SHAP analysis.
- Feature engineering plays a more critical role in predictive accuracy than model tuning or algorithm selection.
- The combined feature set not only enhances predictive results but also produces stable and interpretable importance patterns across varied random seed-based runs.
- Predictive outcomes are influenced by the interaction between local lexical cues and global contextual sentiment, with errors arising from conflicts between these signals.

IV. Discussion

The results of this study emphasize that feature representation, rather than model complexity, is the primary driver of performance in sentiment-based classification tasks and that online product review polarity can be predicted with reasonable accuracy based on a cognitively-inspired feature set. In particular, the inclusion of contextual sentiment by means of VADER compound scoring produced the largest gains in the study's primary evaluative metric, accuracy, across both Random Forest and XGBoost models.

This pattern suggests that the structure and progression of sentiment within an online product review provide more informative signals than static counts of emotional or polarity-based terms. While lexicon-based approaches such as EmoLex and Bing Liu offer interpretable and structured representations, they treat words as independent signals [2]. In contrast, VADER captures contextual effects such as negation, emphasis, and limited word ordering effects, allowing the model to account for how sentiment evolves across a review [8].

Examples drawn from the UCSD Amazon dataset illustrate this behavior. A review such as "This is a great little speaker and does what it is supposed to...however...this is not the case. The description is flawed" may be misclassified as positive due to strong positive lexical cues early in the text, despite an overall negative evaluation. This reflects a conflict between local lexical signals and global contextual sentiment; in contrast, clearly positive reviews such as "This is a nice mount for a smaller TV...the price is great" are consistently classified correctly, reflecting agreement across emotion, polarity, and contextual sentiment features [10].

It is also observed how "disgust" was an influential emotion per SHAP; this may align with frustration or disappointment with a product (negative review polarity) or a transition to relief or satisfaction (positive review polarity). "Disgust" may also function as a stronger signal of expectation violation in consumer contexts, where perceived product failure or deception evokes more critical language than more general dissatisfaction, thereby increasing its influence on polarity predictions.

From a cognitive perspective, these findings are consistent with the motivating principles of Thagard's CRUM concept [4]. Although the models do not explicitly simulate cognitive processes, the integration of emotional categories, polarity signals, and contextual sentiment constitutes a structured approximation of multi-signal cognitive evaluation, in which contextual sentiment establishes overall direction and affective cues refine judgment. The VADER compound score emerges as the most influential factor in SHAP analysis, while emotion features contribute meaningful affective signals that refine predictions.

At the same time, emotion features such as "sadness," "surprise," and "fear" remain important secondary predictors, providing discriminative value in cases where overall sentiment is less definitive. This hierarchical contribution pattern is consistent across both tree-based algorithms considered in this study, further reinforcing the stability of the feature representations across model architectures.

The experimental design strengthens these conclusions. By evaluating twelve configurations across two algorithms and averaging results over five random seeds, this study isolates the effects of feature augmentation and model optimization while reducing sensitivity to sampling variability. The relatively low variance observed in optimized models further suggests that the reported improvements are stable and reproducible compared to artifacts of a specific train-test split.

The comparison between XGBoost and Random Forest models further supports this conclusion. Although XGBoost achieved the highest performance, Random Forest produced closely comparable results, indicating that well-structured feature representations contribute more to predictive performance than differences between tree-based algorithms.

These results also have practical implications beyond predictive accuracy. Discrepancies between predicted sentiment and assigned ratings may indicate anomalous or low-quality reviews, particularly in cases involving sarcasm, colloquialisms, or ambiguous language [18]. More broadly, models built on transparent, cognitively informed feature representations may support applications such as content moderation, spam detection, and review validation in modern online systems.

V. Conclusion

This study presented a cognitively inspired feature-engineering architecture for predicting online product review polarity using Amazon "Electronics" review data. By integrating emotion-based representations (EmoLex), sentiment polarity features (Bing Liu), and contextual sentiment (VADER), interpretable machine learning models were developed and evaluated.

Across twelve model configurations and five independent random seed trials, results demonstrate that feature design, as opposed to model architecture, is the primary driver of predictive performance. Ablation-style evaluation shows that progressively richer feature representations yield consistent improvements, with the optimized XGBoost model achieving 75.04% accuracy and 0.825 AUC.

SHAP analysis reveals a stable and interpretable structure of feature contributions, with contextual sentiment serving as the dominant signal and emotion-based features providing complementary refinement. This pattern is consistent with cognitively motivated representations in which multiple signals are integrated to support evaluation and judgment [4]. Overall, the findings demonstrate that structured, interpretable feature engineering can achieve strong performance while preserving transparency, providing a viable alternative to possibly opaque or more complex deep learning approaches.

VI. Limitations and Future Work

While this study demonstrates the effectiveness of combining emotion, sentiment, and contextual features, several limitations highlight opportunities for further development. Although the inclusion of VADER introduces a degree of contextual awareness, reliance on simple lexicon-based feature mappings may not be sufficient to account for linguistic phenomena such as sarcasm, domain-specific vernacular, and nuanced phrasing which may not be fully captured. Future work could explore hybrid approaches that integrate transformer-based representations such as BERT [19] with the interpretable features developed in this study.

This study intentionally prioritizes interpretability and controlled feature design over state-of-the-art performance. The use of tree-based models and structured lexical features enables transparent analysis of feature contributions, versus direct competition with deep learning approaches. Future work may examine how these cognitively grounded features perform within more complex architectures and whether similar gains persist in higher-capacity models. Thresholding by prediction classification confidence and the impact on accuracy is another potential enhancement to this study.

This analysis is also limited to a single product domain ("Electronics") within the UCSD Amazon dataset, reflecting historical consumer language patterns [10]. Given that the dataset spans reviews from May 1996 through July 2014, it may not fully represent modern online discourse. This limitation may also introduce sampling bias, as the dataset may not capture broader consumer language patterns, product categories, or evolving review behaviors. Evaluating the proposed approach on more recent datasets, additional product domains, or cross-domain corpora would provide stronger evidence of generalizability.

Finally, this approach provides a foundation for applications beyond product review polarity classification. Future work could investigate the detection of misalignment between review text and assigned ratings to identify anomalous, mislabeled, or potentially misleading reviews, including through human-in-the-loop validation frameworks. Extending the approach to non-English languages is also recommended. Overall, these directions suggest that the proposed approach can be extended while preserving its core objective: demonstrating that structured, cognitively informed feature design can improve model performance while maintaining explainability.

Acknowledgments

The author would like to thank Professor Rafael Alvarado of the University of Virginia for inspiring the text analytics methodology used in this study. The author is also appreciative of Dr. Ashok Goel and Dr. Keith McGreggor of the Georgia Institute of Technology for motivating the cognitive science perspective that informed this work's feature engineering.

References

[1] T. Chen, P. Samaranayake, X. Cen, M. Qi, and Y. C. Lan, "The impact of online reviews on consumers' purchasing decisions: Evidence from an eye-tracking study," *Frontiers in Psychology*, vol. 13, Art. no. 865702, Jun. 2022, doi: 10.3389/fpsyg.2022.865702.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008. doi: 10.1561/1500000011.

[3] Z. C. Lipton, "The mythos of model interpretability," in *Proc. ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2016.

[4] P. Thagard, *Mind: Introduction to Cognitive Science*, 2nd ed. Cambridge, MA, USA: The MIT Press, 2005.

[5] S. Mohammad, "NRC Emotion Lexicon," National Research Council Canada. [Online]. Available: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. [Accessed: Mar. 25, 2026].

[6] R. Plutchik, *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*. Washington, DC, USA: American Psychological Association, 2003.

[7] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.

[8] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. Conf. Weblogs and Social Media (ICWSM)*, Ann Arbor, MI, USA, 2014. doi: 10.1609/icwsm.v8i1.14550

[9] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992. doi: 10.1080/02699939208411068

[10] J. McAuley, "Amazon datasets," Univ. of California San Diego, Comput. Sci. Eng., San Diego, CA, USA, 2024. [Online]. Available: https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html. [Accessed: Mar. 20, 2026].

[11] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *arXiv preprint arXiv:2106.03253*, 2021. [Online]. Available: https://arxiv.org/abs/2106.03253.

[12] A. M. Salih *et al.*, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *arXiv preprint arXiv:2305.02012*, May 2023. [Online]. Available: https://arxiv.org/abs/2305.02012.

[13] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, New York, NY, USA: ACM, 2004, pp. 168–177, doi: 10.1145/1014052.1014073.

[14] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLOS ONE*, vol. 16, no. 8, Aug. 2021, doi: 10.1371/journal.pone.0254937.

[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *arXiv preprint arXiv:1907.10902*, 2019. [Online]. Available: https://arxiv.org/abs/1907.10902.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[17] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[18] D. Maynard and M. Greenwood, "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis," in *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014, pp. 4238–4243.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.