Predicting Amazon Product Review Polarity Using a CRUM-Based Approach

Travis Vitello
Online Masters in Computer Science
Georgia Institute of Technology
Atlanta, GA, USA
tvitello3@gatech.edu
CS6795, Spring 2025: Computation Model/Tool Term Project

***Abstract*—In this investigation, Amazon "Electronics" product review data from May 1996-July 2014 (comprising 1,689,188 entries) was studied to anticipate if negative product scores (1-2 stars) and positive product scores (4-5 stars) could be accurately predicted using supervised XGBoost and Random Forest machine learning (ML) models. Predictions explored emotional mapping using the NRC EmoLex dictionary, Bing Liu's polarity scoring methods, and finally VADER Compound effects to mirror the spirit of Thagard's *CRUM* model of cognition, showing that these ancillary predictor fields generally improved model predictive performance. Models were optimized using the Optuna Python library, with fields' contributory predictive performances evaluated using SHapley Additive exPlanations (SHAP) to highlight influential model features. The optimized XGBoost model that included EmoLex emotion scores, Bing Liu sentiment polarities, and VADER Compound scoring effects was found to have an accuracy of 75.4%, exceeding the performance for all other models considered in this study.**

I. Introduction

As a behemoth in e-commerce, Amazon is recognized as "one of the largest retail platforms" of consumer goods today [1]. Research has found that relationships exist between products' online review content, including those of Amazon, and consumer opinion, whereby as many as "93% of customers read online reviews before buying a product" [2]. Positively scored reviews have been shown to influence consumer spending by at least 31% while a single negatively scored review may dissuade as many as 22% of potential customers [2]. For businesses or brands aiming to succeed economically, understanding the cognition and psychology of reviews and reviewers, respectively, may lead to improved accurate marketing strategies and more authentic published review content; however, consumers are known to sometimes have difficulty identifying "fake" reviews [3]. While it is considered beyond the scope of this study to develop methods for moderating content so as to distinguish between real and fake reviews such as previously explored by Doan, et al. [4], as well as by Jindal and Liu [5], this project will attempt to demonstrate that the cognitive response to a review's text classification may or may not be aligned with its review score. This study seeks to explore if the integration of emotions, sentiments, and scoring methods intended to mimic the characteristics of Thagard's "Computational-representational understanding of mind" (*CRUM*) [6] model improves a machine learning model's capability to anticipate human review scoring. Finally, a further motivation for this study is Julian McAuley's University of California San Diego (UCSD) repository of Amazon reviews dating from May 1996 through July 2014 across several product disciplines; herein, "Electronics" reviews, consisting of 1,689,188 entries, were used [7]. This decision was based on the large quantity of available data and the perceived categorical relevance to Georgia Tech's OMSCS program, as opposed to alternate product classes such as "Beauty" or "Musical Instruments". McAuley's reviews were subset to a simple, two-column dataframe consisting of "review text" and "review score" (in stars), with no supplementary UCSD-sourced data or metadata being applied.

II. Model/Tool Design

A. Data Preparation

This study only considered positive (4-5 stars) and negative (1-2 stars) reviews, thereby dropping all neutral (3-star) reviews. Prior to applying machine learning (ML) methods in this study, positive reviews were mapped to a response score of 1 while negative reviews were mapped to a response score of 0 to facilitate binary classification. Upon ingestion of the UCSD data, the number of reviews were reduced to 10% of the resultant entries, begetting the dataframe shown in Figure 1 where "reviewText" implicative of the original text of the review (inclusive of the unaltered linguistic nuances, such as *in situ* spelling, grammar, capitalization, punctuation, and spacing) and "overall" represents the review score in stars. Reviews were sub-sampled to an even distribution of 1, 2, 4, and 5-star reviews, or 8,207 reviews each (32,828 reviews overall). Next, review text was tokenized with lemmatization using the **nltk** *WordNetLemmatizer* method to adjust word forms without stemming, as seen in Figure 2. Note that the corresponding **nltk** *WordNet* method was applied for tagging words per their corresponding part of speech, such as "nouns", "adjectives", "verbs", and "adverbs," based on content

in order to most appropriately lemmatize tokens [8]. The *fix* method of the Python **contractions** library was applied to expand terms like "don't" and "y'all" into "do not" and "you all", respectively. Then, the **nltk** English stop word removal method was applied to eliminate uninteresting or insignificant words from the token list (such as "the" or "and") [9]. As observed, lemmatizing allowed for word form simplifications, such as "received" to "receive" and "giving" to "give". This process also stripped the tokens of casing and punctuation, elements typically absent from the NRC Emotion Lexicon (EmoLex) and Bing Liu's positive and negative sentiment word lists, which were applied in this study.

	reviewText	overall
0	what I received didn't match the picture shown...	1.0
1	It is a great tablet for the price. I use it f...	4.0
2	I was hesitant to buy a TV online in case I ne...	5.0
3	Works great, no problems. Very tiny and cool ...	5.0
4	I've owned a Canon Powershot SD880IS for sever...	2.0
...	...	...
32823	I'm giving this 1 star in terms of design for ...	2.0
32824	The price and the picture are the only good th...	2.0
32825	This item just arrived this morning, so I have...	5.0
32826	Broke in less than six months with only modera...	2.0
32827	Horrible sound quality. If you have to buy a F...	1.0

32828 rows × 2 columns

Fig. 1: Sample Dataframe After Reduction.

	reviewText	overall	tokens
0	what I received didn't match the picture shown...	1.0	[receive, match, picture, show, item, picture,...
1	It is a great tablet for the price. I use it f...	4.0	[great, tablet, price, use, test, run, questio...
2	I was hesitant to buy a TV online in case I ne...	5.0	[hesitant, buy, tv, online, case, need, return...
3	Works great, no problems. Very tiny and cool ...	5.0	[work, great, problem, tiny, cool, look, amaze...
4	I've owned a Canon Powershot SD880IS for sever...	2.0	[own, canon, powershot, sd880is, several, year...
...	...	...	...
32823	I'm giving this 1 star in terms of design for ...	2.0	[give, 1, star, term, design, underwater, purp...
32824	The price and the picture are the only good th...	2.0	[price, picture, good, thing, toshiba, tv, off...
32825	This item just arrived this morning, so I have...	5.0	[item, arrive, morning, much, time, evaluate, ...
32826	Broke in less than six months with only modera...	2.0	[broke, less, six, month, moderate, use, scree...
32827	Horrible sound quality. If you have to buy a F...	1.0	[horrible, sound, quality, buy, fm, transmitte...

32828 rows × 3 columns

Fig. 2: Typical Tokenized and Lemmatized Text.

EmoLex counts, inclusive of the emotions "trust," "sadness," "fear," "anger," "anticipation," "joy," "surprise," and "disgust", were cumulated as raw totals based on the tokens comprising the review content [10]. Said eight emotions are derived from the work of Plutchik, who created a so-called "wheel" as shown in Figure 3 [11]. Plutchik's concept provides a comprehensive reduction that attempts to illustrate the inter-relationships of emotions, which may align with Thagard's *CRUM* model as a tool for explaining the underlying motivation of human thoughts and actions [6]. Thagard, whose *CRUM* model is considered an "approach to understanding the mind", holds that emotions are foundational to cognitive processes, motivating reasoning and decision making in addition to forming "judgments about a person's general state" [6]. While Thagard does not provide an explicitly-defined catalog of core or fundamental emotions like Plutchik or Ekman [12], his *CRUM* model may nevertheless be considered useful for analyzing emotional expressions in online review content [6].

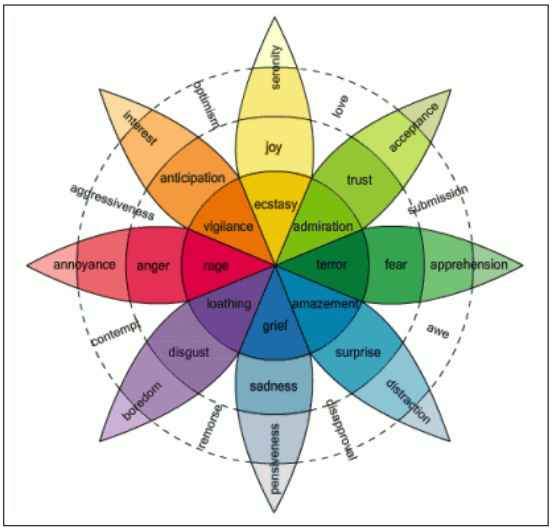
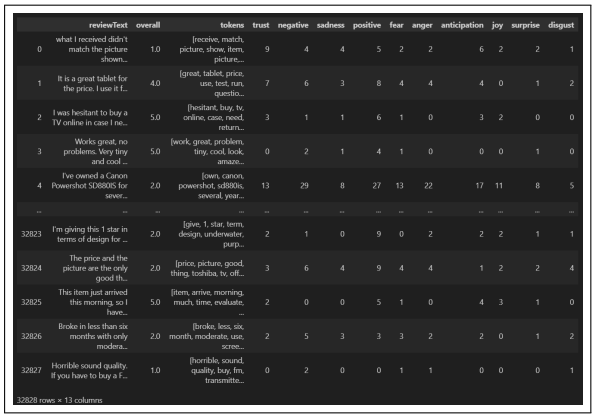


Fig. 3: Plutchik's Emotion Wheel [11].

Upon counting such content against the EmoLex word list, columns were appended to the dataframe previously shown in Figure 2; this is demonstrated in Figure 4. A limitation to this approach is that typos or misspellings in reviews will likewise result in unmatched emotions, thereby providing inaccurate dataframe values. Future study would benefit from editing reviews to correct such deficiencies, while also potentially including a meta engineered feature to the dataframe like "number_of_text_errors" to explore the relationship between such phenomena and review classification (for example, is an error-prone review more or less likely to be correlated to a negative classification or to a distinct emotion class?).

Further, the next aspect of data preparation was to compare the review content to Bing Liu's "positive" and "negative" sentiment word lists [13]. Like the application of EmoLex, the same tokenized and lemmatized set of words per review were compared to Liu's documents, whereby a cumulative sum of positive and negative scores were appended to our dataframe as the columns "Bing_Liu_Positive" and "Bing_Liu_Negative" respectively. This can be seen in Figure 5, which provides the top five rows of our modified dataframe. However, to ameliorate potential bias in the analysis due to reviews of especially long or short content word-wise, it was considered pragmatic to **normalize the EmoLex columns by the size of each review's list of tokens** (recognizing that the **nltk**

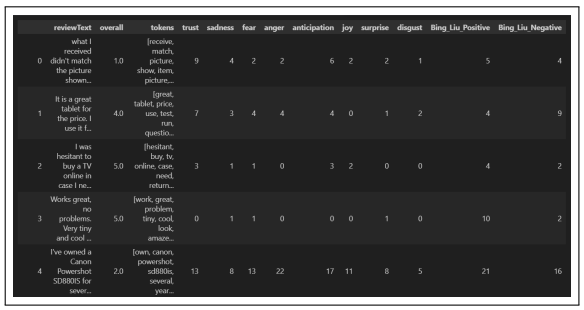


	reviewText	overall	tokens	trust	negative	sadness	positive	fear	anger	anticipation	joy	surprise	disgust
0	what I received didn't match the picture shown...	1.0	[receive, match, picture, show, item, picture,...	9	4	4	5	2	2	6	2	2	1
1	It is a great tablet for the price. I use it f...	4.0	[great, tablet, price, use, test, run, questio...	7	6	3	8	4	4	4	0	1	2
2	I was hesitant to buy a TV online in case I ne...	5.0	[hesitant, buy, tv, online, case, need, return,...	3	1	1	6	1	0	3	2	0	0
3	Works great, no problems. Very tiny and cool ...	5.0	[work, great, problem, tiny, cool, look, amaze...	0	2	1	4	1	0	0	0	1	0
4	I've owned a Canon Powershot SD880IS for sever...	2.0	[own, canon, powershot, sd880is, several, year...	13	29	8	27	13	22	17	11	8	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...
32823	I'm giving this 1 star in terms of design for ...	2.0	[give, 1, star, term, design, underwater, purp...	2	1	0	9	0	2	2	2	1	1
32824	The price and the picture are the only good th...	2.0	[price, picture, good, thing, toshiba, tv, off...	3	6	4	9	4	4	1	2	2	4
32825	This item just arrived this morning, so I have...	5.0	[item, arrive, morning, much, time, evaluate, ...	2	0	0	5	1	0	4	3	1	0
32826	Broke in less than six months with only modera...	2.0	[broke, less, six, month, moderate, use, scree...	2	5	3	3	3	2	2	0	1	2
32827	Horrible sound quality. If you have to buy a F...	1.0	[horrible, sound, quality, buy, fm, transmitte...	0	2	0	0	1	1	0	0	0	1

32828 rows × 13 columns

Fig. 4: Dataframe with EmoLex Emotion Columns.

word_tokenize method does not exclude redundant words, unlike a *set* in Python); see Figure 6. Likewise, the Bing Liu sentiment was normalized against the total positive proportion ($S_{positive\_norm}$) and negative proportion ($S_{negative\_norm}$) sentiment scores per review, per Equations 1 and 2; the resultant columns were identified as "Bing_Liu_Positive_Norm" and "Bing_Liu_Negative_Norm" respectively.



	reviewText	overall	tokens	trust	sadness	fear	anger	anticipation	joy	surprise	disgust	Bing_Liu_Positive	Bing_Liu_Negative
0	what I received didn't match the picture shown...	1.0	[receive, match, picture, show, item, picture,...	9	4	2	2	6	2	2	1	5	4
1	It is a great tablet for the price. I use it f...	4.0	[great, tablet, price, use, test, run, questio...	7	3	4	4	4	0	1	2	4	9
2	I was hesitant to buy a TV online in case I ne...	5.0	[hesitant, buy, tv, online, case, need, return...	3	1	1	0	3	2	0	0	4	2
3	Works great, no problems. Very tiny and cool ...	5.0	[work, great, problem, tiny, cool, look, amaze...	0	1	1	0	0	0	1	0	10	2
4	I've owned a Canon Powershot SD880IS for sever...	2.0	[own, canon, powershot, sd880is, several, year...	13	8	13	22	17	11	8	5	21	16

Fig. 5: Dataframe with Bing Liu Sentiments Columns Included.

$$S_{\text{positive\_norm}} = \frac{\sum n_{\text{positive}}}{\sum n_{\text{positive}} + \sum n_{\text{negative}}} \quad (1)$$

$$S_{\text{negative\_norm}} = \frac{\sum n_{\text{negative}}}{\sum n_{\text{positive}} + \sum n_{\text{negative}}} \quad (2)$$

Where:

- $\sum n_{positive}$ is the sum of Bing Liu positive sentiment tokens in a review.
- $\sum n_{negative}$ is the sum of Bing Liu negative sentiment tokens in a review.

Finally, reviews were considered in their original word order in anticipation of applying VADER (Valence Aware Dictionary and sEntiment Reasoner) to capture how sentiment may evolve progressively, resulting in a final normalized "compound" score (where a score > 0 suggests a positive holistic sentiment, $0 <$ suggests a negative holistic sentiment, and 0 suggests a neutral holistic sentiment) [14]. Whereas Bing Liu's conventional dictionary mapping approach considers an elementary one-to-one alignment of listed terms akin to an orderless bag-of-words paradigm, VADER is designed to understand the ordered context and sequence of words, including how modifiers (like "not") and term progressions can influence the sentiment of a text; similarly, this allows VADER to evaluate texts (such as reviews) that begin negatively but end positively, or vice versa [14]. An example of this sort of transition from positive to negative which VADER would score negatively may be a sentence like: "Initially, I really liked this Zune, however it broke after 5 uses and now I can't use it." The motivation for applying VADER to score review content is to simulate a symbolic representation of sentiment (akin to Bing Liu's historically more simplified, binary positive or negative sentiment approach [13]), which is intended to mirror or otherwise conceptually align with the *CRUM* philosophy of abstract reasoning, whereby concepts (in this case, review text) are structured into cognitive classes, such as sentiment categories or — like in this study — Amazon product preferences. While Thagard's *CRUM* model is not necessarily a traditional tool for natural language processing (NLP) applications, it is believed that approaches like VADER theoretically may be leveraged for capturing linguistic context in a manner that mimics the cognitive complexities of human emotion and communication; thus, a hypothesis of this study is that by integrating VADER Compound score to a dataframe, ML scoring (such as predictive accuracy) will improve beyond basic, reductive emotion scoring derived from EmoLex alone or even as supplemented with Bing Liu sentiment polarities. It is noted that by using these approaches or a combination thereof, the resulting models will be transparent, interpretable, and explainable as opposed to less clear blackbox ML models or single-source lexicons like AFINN, whose inability to effectively consider emotion and context have been demonstrated to warrant more mature approaches [15].

B. Analytical Method

Two ML model classes were considered for this study: a simple Random Forest Classifier using the **scikit-learn** Python library and a gradient boosted decision tree classifier, XGBoost, per the **xgboost** Python library. These tree ensemble models were chosen due to their relatively fast solution speed and interpretable results along with general robustness when operating on tabularly-structured data, such as what was considered in this study. Indeed, in a separate investigation, Shwartz-Ziv and Armon found that such tree ensemble models have advantages over neural network-based deep learning models for most structured datasets [16]. Note that for both classes, models were trained on unoptimized (**baseline**) hyperparameters and ones **optimized** using the **Optuna** library. The motivation to use **Optuna** in lieu of more traditional and cumbersome

Grid Search or Bayesian Search techniques was based on the advanced "pruning strategy," efficient sampling optimization, computational time savings, and overall efficacy of this approach, as discussed by Akiba, et al. [17]. **Baseline** cases utilized human-provided hyperparameter values considered as pragmatic "best guesses" with the expectation that optimization will improve predictive performances across evaluative metrics (including accuracy, precision, recall, and F1 score).

Finally, the contribution of each model's predictor fields on model output impact was evaluated using SHapley Additive exPlanations (SHAP), primarily its *TreeExplainer* method as appropriate for Random Forest and XGBoost applications. SHAP was elected over LIME for this study due to author preference, however each strategy has been demonstrated to effectively provide insight into explaining ML model behaviors [18]. By employing SHAP, we aim to illustrate how model features interact and contribute to predictions. This form of feature attribution can be viewed as loosely analogous to the process of analogical reasoning in *CRUM*, in the sense that both involve identifying patterns of influence across structured inputs. While not a direct comparison, this framing allows us to explore whether classifier behavior mirrors aspects of human emotional reasoning. Although this may not be an exact match to the inner-workings of human cognitive processes, using explainable AI methods like SHAP provides an opportunity to explore the decisions made by models in classifying and predicting their responses. Models were constructed per the prior section, with the addition of supplemental columns to complement each prior model's fields; refer to Tables I and II. In all cases, "Sentiment" is considered the normalized Bing Liu polarity (positive and negative), while "VADER" is considered the compound score derived from the **nltk** *SentimentIntensityAnalyzer* method, while "Emotion" is the scoring of a review's unordered words against the EmoLex dictionary against total word count in a review.

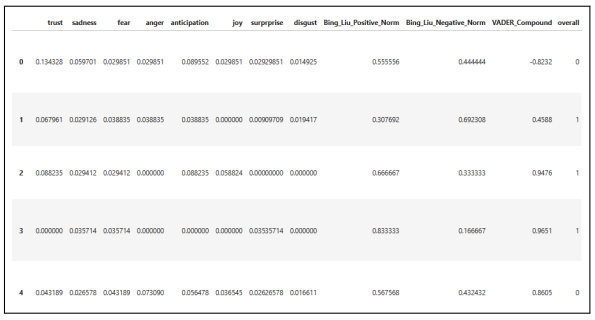
TABLE I: Baseline Models

Model #	ML Algorithm	Fields
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

A *stratified* 80-20 training-test data split was applied to the prepared dataframes per each model examined (as comprised of the fields given in Tables I and II) using the *same random seed*. Quantitatively, this approach resulted in 26,262 reviews considered in the training dataset (evenly divided as 13,131 positive and negative reviews, respectively); further, 6,566 reviews were considered in the test dataset (evenly divided as 3,283 positive and negative reviews, respectively). The typical structure of the final prepared train-test dataframe having all columns (emotion + sentiment + VADER, from left to right) is provided in Figure 6, recognizing that the rightmost column designated "overall" represents the set of binary responses which were isolated for predictive evaluation. Notably, text-related fields, including each set of tokens per line, were omitted from the ML models.

TABLE II: Optimized Models

Model #	ML Algorithm	Fields
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



	trust	sadness	fear	anger	anticipation	joy	surprprise	disgust	Bing_Liu_Positive_Norm	Bing_Liu_Negative_Norm	VADER_Compound	overall
0	0.134328	0.059701	0.029851	0.029851	0.089552	0.029851	0.02929851	0.014925	0.555556	0.444444	-0.8232	0
1	0.067961	0.029126	0.038835	0.038835	0.038835	0.000000	0.00909709	0.019417	0.307692	0.692308	0.4588	1
2	0.088235	0.029412	0.029412	0.000000	0.088235	0.058824	0.00000000	0.000000	0.666667	0.333333	0.9476	1
3	0.000000	0.035714	0.035714	0.000000	0.000000	0.000000	0.03535714	0.000000	0.833333	0.166667	0.9651	1
4	0.043189	0.026578	0.043189	0.073090	0.056478	0.036545	0.02626578	0.016611	0.567568	0.432432	0.8605	0

Fig. 6: Dataframe with EmoLex Emotion, Bing Liu Sentiment, and VADER Compound Score Columns.

The **Optuna** optimization for the champion XGBoost model #9 discussed in the following section had the following hyperparameters applied:

- **n_estimators:** 109
- **learning_rate:** 0.045906820392350396
- **max_depth:** 6
- **min_child_weight:** 7
- **subsample:** 0.7114069803698975
- **colsample_bytree:** 0.5685143036994255

The second-best performing model in this study as discussed in the following section was Random Forest model #12; its **Optuna**-optimized hyperparameters consisted of:

- **n_estimators:** 57
- **max_depth:** 20
- **min_samples_split:** 3
- **min_samples_leaf:** 7

III. Results

As the dataset was *stratified* and therefore balanced between binary positive (1) and negative (0) response cases, it was felt that **accuracy** would be an acceptable metric for determining a "champion" model of the

12 cases considered (refer to Tables I and II). These results are summarized in Tables III-VI, where each top scoring metric is shaded gray. Note that model cross-validation was not applied during this study as balanced binary class representation, as noted above, along with having over 6,000 reviews in the test set was considered as providing sufficient statistical power to reliably evaluate all 12 models; however, future study may seek to improve this approach by averaging results across multiple random seeds and sub-samples.

TABLE III: XGBoost Evaluation Metrics (Baseline)

Model	Acc.	Prec.	Recall	F1
Emotion Only	63.3%	63.7%	61.9%	62.8%
Emotion + Sentiment	69.9%	68.3%	74.4%	71.2%
Emotion + Sentiment + VADER	72.4%	71.2%	75.4%	73.2%

TABLE IV: Random Forest Evaluation Metrics (Baseline)

Model	Acc.	Prec.	Recall	F1
Emotion Only	64.3%	62.7%	70.4%	66.3%
Emotion + Sentiment	70.8%	70.2%	72.4%	71.3%
Emotion + Sentiment + VADER	74.6%	75.6%	72.9%	74.2%

TABLE V: XGBoost Evaluation Metrics (Optimized)

Model	Acc.	Prec.	Recall	F1
Emotion Only	64.5%	63.0%	70.1%	66.4%
Emotion + Sentiment	71.1%	70.4%	72.6%	71.5%
Emotion + Sentiment + VADER	75.4%	76.2%	74.0%	75.1%

TABLE VI: Random Forest Evaluation Metrics (Optimized)

Model	Acc.	Prec.	Recall	F1
Emotion Only	65.1%	63.3%	72.0%	67.4%
Emotion + Sentiment	71.3%	70.4%	73.4%	71.9%
Emotion + Sentiment + VADER	75.2%	75.7%	74.4%	75.0%

In all cases, including Bing Liu sentiment polarity tended to improve accuracy-based predictive performance compared to using just the EmoLex emotion values for the Amazon reviews in the test data. The further addition of VADER Compound score resulted in a continued increase in predictive performance, with the **Optimized XGBoost (Emotion + Sentiment + VADER)** model achieving the *highest overall accuracy at **75.4%*** (as seen in Table V). However, with an accuracy of 75.2% (as seen in Table VI), the Optimized Random Forest (Emotion + Sentiment + VADER) model also performed well in this study, outscoring the "champion" **Optimized XGBoost (Emotion + Sentiment + VADER)** model in recall (74.4% vs. 74.0%). These top performers represent models #s 9 and 12, respectively, per Table II. The confusion matrix of the champion model (#9) is provided in Figure 7. While the top-performing model was able to identify 2,517 True Negative and 2,434 True Positive reviews correctly, it also misclassified 855 reviews as False Negatives and 760 reviews as False Positives. This indicates that there is further improvement in classification accuracy with more advanced modeling methods and algorithms.

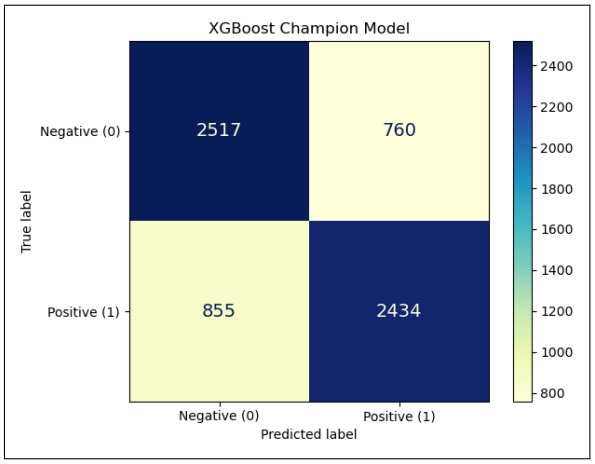


Fig. 7: Champion Model Confusion Matrix.

Examples of reviews [7] which the model correctly classified as True Negatives include:

- "came on too much. eats battery life. hard to move when stuck on wall so cant try new location. Disappointed."
  - VADER Compound score: -0.4767
- "It was good while it lasted, but it did fail and I lost data. I don't think I'll buy this brand again."
  - VADER Compound score: -0.8100
- "VERY WEAK UNIT! Worked about two months, Harddrive went out..."No HDD". These people have no service centers anywhere. BUY ANYTHING ELSE! Don't even think about buying this DEFENDER!"
  - VADER Compound score: -0.7887

Examples of reviews [7] which the model correctly classified as True Positives include:

- "Working nice from 2 years, although I have minimal use but good product for the amount of money. I recommend this for domestic usage."
  - VADER Compound score: 0.8402
- "This is a nice mount for a smaller TV. It has a nice range of motion and the price is great!"
  - VADER Compound score: 0.8748
- "I was worried about how well these would lock shut, but they have a surprisingly solid locking mechanism. I'd definitely buy these again."
  - VADER Compound score: 0.8020

Conversely, examples of reviews [7] which the model incorrectly classified as False Positives include:

- "This is a great little speaker and does what it is supposed to...however...I was under the premise

that it was powered...and had an amp in it...this is not the case. The description is flawed....this should be made clear."
  - VADER Compound score: 0.7717
- "Only works in G mode with Vista. Spent most of 2 days on phones with technical service, could not get to work with Vista-32 or Vista-64 in N mode. Might work with XP?"
  - VADER Compound score: 0.3818
- "it works, but its not easy to configure and the audio quality is ok, at best (mp3 quality but NOT cd quality; not even close)."
  - VADER Compound score: 0.5486

Finally, examples of reviews [7] which the model incorrectly classified as False Negatives include:

- "This is the best cable killer out there. Now featuring youtube. I use this twice as much as my cable box and plan on cutting the cord completely after my dish contract ends"
  - VADER Compound score: -0.5423
- "WORKS BUT DUE TO MY COMPUTER ILLITERACY OR YAHOO CHATS STUPID "CHANGES" IN MY VIDEO CALLING, I CANNOT GET THE CAMERA/VOICE FEATURE TO WORK!!!! NOT THE PRODUCTS FAULT."
  - VADER Compound score: -0.8285
- "Bought this for my GoPro. Can hold more 1080p footage than the fully charged battery can record before dying! Probably over kill in terms of speed for the GoPro, but I'm not complaining! Love it"
  - VADER Compound score: -0.8303

The prior examples highlight how ML models integrating cognitive qualities like emotion, sentiment, and VADER Compound scores can often mimic human interpretation, but are still limited by nuanced or mixed-language reviews, especially through the interpretation of capitalization, exclamation marks, and other aspects that may be more intuitive to a human reader than AI. While the negative and positive reviews appear more direct, notable language like "cutting" and "killer" (or "ILLITERACY", "STUPID", and "FAULT") appear to have been misinterpreted as being associated with negativity (in what was actually a positive review) or "great" and "best" appear to have been misinterpreted as being associated with positivity (in what were actually negative reviews). Such complexities and other context-dependent language pushed the classifications, explicitly the VADER Compound scores, in incorrect directions, thus demonstrating certain limitations in how human thoughts, in applications like Amazon review classification, are complex phenomena to model computationally. Yet the improvements observed from the Baseline to the Optimized models underscore the idea that cognitively-grounded approaches appear to improve predictive performance in spite of such human complexity. This is further considered through a comparison of the effects of optimization between the XGBoost (Emotion + Sentiment + VADER) models (Baseline vs. Optimized) as captured in Figure 8; this plot shows a comparison of each models' corresponding ROC-curve and subsequent AUC scores. In this case, the optimized champion (model #9) had an AUC of 0.8285 while the associated baseline case (model #3) had a less performant AUC of 0.8044, further demonstrating the improvement that the **Optuna** optimization method, in this case, had on fine tuning hyperparameters for increased functionality. According to Hosmer, et al., an AUC for a binary response model between 0.8 and 0.9 is considered to embody "excellent discrimination" [19]. This indicates that the top-performing model developed in this study demonstrated capability to identify both negative reviews from positive reviews in a manner considered much better than a proverbial coin toss.

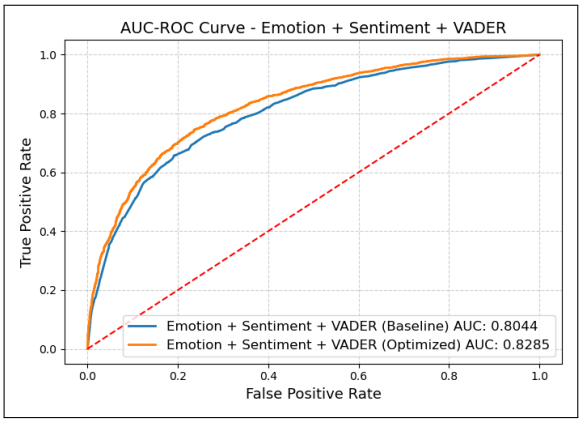


Fig. 8: ROC Curve Comparison of XGBoost Champion Model vs. Corresponding Baseline Case.

Further, an exploration of the fields contributing to the champion model's predictions was performed using SHAP; refer to Figure 9. In this figure, higher-scored values are shown in red, which nudge the model more strongly toward the positive class (representing the binary value of 1, or a positive Amazon review), while blue values denote a relatively lower feature value (i.e., towards the negative predicted class, representing the binary value of 0, or a negative Amazon review). As observed in Figure 9, the overwhelmingly most impactful feature of the champion model was the "VADER_Compound" field, followed next by Bing Liu's normalized sentiment scoring, with the "positive" sentimentality being more influential on accurate review prediction than "negative" sentiment. Among the EmoLex emotional features, "disgust" and "joy" were found to be the most influential, with "disgust" skewing more strongly towards negative predictions and "joy" skewing more strongly towards positive predictions. Possibly more mixed or ambiguous emotions like "trust", "sadness", and "fear" appeared to have a lower overall impact, yet still demonstrated characteristics of directional influence on model predictions.

Having intermediate influence on model prediction were the emotional features of “anticipation”, “surprise”, and “anger”, each of which may be associated with both positive and negative review types.

Consider the following two cases from the review data [7]:

- “Used it a couple of times, read the instructions over and over, didn’t feel like I got a good result, finally stopped using it.”
- “Arrived in perfect order and on time, does what it says on the box and was wrapped in an appropriate manner.”

Both of these reviews had similarly high EmoLex “anticipation” emotion scores in the model, with the first being a negative review and the second being a positive review. Interestingly, the first of these reviews received a -0.1053 VADER Compound score likely due to the change of emotion from having used the product to stopping use of the product; conversely, the second review achieved a VADER Compound score of 0.5719 as it appears to stay positive throughout the entirety of the given text. Thus, “anticipation” alone may not serve as a sufficient indicator for review classification as it is an emotion that is plausible to apply to both positive and negative human opinions, such as shown in the above examples. Yet in general, the VADER Compound score — which was intended to follow Thagard’s *CRUM* model — was the most influential feature of the champion model, which may suggest that the evolution of feeling in a sentence is more significant than simple disconnected dictionary-based word mappings for emotion or sentiment alone.

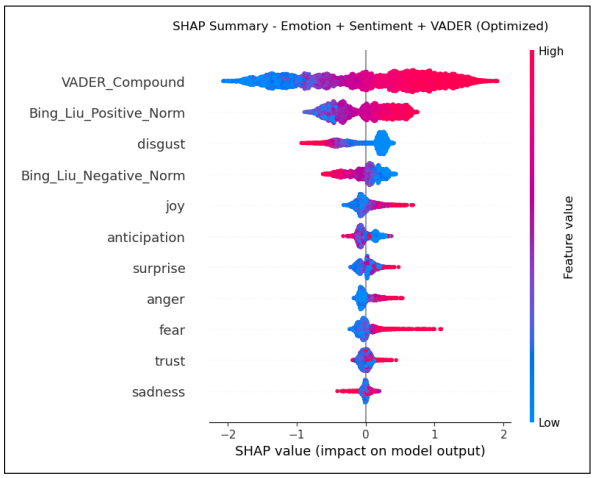


Fig. 9: Champion Model SHAP Plot.

In summary, of the twelve ML models tested, model #9 demonstrated that the highest performance was achieved through a combination of emotion, sentiment, and VADER Compound score fields. Indeed, the inclusion of VADER Compound scoring consistently improved model performance for both the baseline and optimized XGBoost and Random Forest cases, thus supporting the hypothesis that incorporating ordered sentiment context adds predictive value beyond static emotion or polarity lexicons alone. These results reinforce the notion that cognitive-inspired feature engineering in ML models — such as representing emotional tone and sentiment flow — enhance classification accuracy in review-based sentiment analysis. Yet, as observed by the False Positive and False Negative examples, having a strong positive or negative VADER Compound score may not be sufficient (despite its significance per Figure 9), along with the other fields used in this study, to correctly classify *all* human-generated Amazon “Electronics” reviews in this dataset. While a classification accuracy of **75.4%** and an AUC of 0.8285 exceeds guessing, further opportunities for improvement clearly exist.

IV. Discussion

The results of this study demonstrate that adding emotionally intelligent features (chiefly the VADER Compound score) made an appreciable difference in how well the tree-based ML models considered in this study were able to predict review scoring. The success of integrating VADER Compound scoring to enhance prediction accuracy is attributed to its ability to track how emotions unfold throughout a sentence, much like how we, as humans, process feelings and ideas in text. This approach is quite in tune with Thagard’s *CRUM* model of cognition, which suggests our thoughts and emotions are deeply interconnected. Instead of just counting words related to emotions or sentiments, VADER looks at the whole structure and flow of the text, recognizing that how we write and arrange our words is key to understanding the real emotions behind them. This makes VADER especially good at analyzing review scores, whether for Amazon’s ”Electronics” or anything else, because it captures the genuine human touch in the reviews.

Yet, while SHAP, in particular, may offer insights into a model’s predictive behaviors via feature attribution, future study may benefit from further attempting to align SHAP outputs to the emotional logic and response of pure human judgment. Nevertheless, a potential real-world implication of these findings is that through the integration of psychology-inspired concepts to data science models, ML or AI models may be enhanced to better predict and ultimately mimic human thought and behavior. The implications for product marketing, review content moderation, and human sociolinguistic analysis are considered significant enough to warrant further study on this topic in the opinion of the author. This study suggests that the emotions of a review’s author are neither irrelevant nor simply noise, but rather cogent information capable of enhancing overall predictive model performances.

V. Conclusion

From this study, it was learned that ML models developed for predicting human opinion-based review

scores can be substantially enhanced by the inclusion of engineered features that leverage a combination of emotion, sentiment, and VADER Compound score (the latter of which is intended to roughly mimic cognitive science-based principles of human psychology). The top performing model of this study blended such features, consistently outperforming models that were not as fully featured. A secondary takeaway from this study reinforced that optimized XGBoost-based models tended to have a stronger predictive capability than comparatively more basic Random Forest-models, with such optimization via Optuna hyperparameter tuning (focused on accuracy maximization) resulting in models that typically outperformed un-optimized baseline models in most metrics (see Tables III-VI).

VI. Limitation

As noted, lexicon matching (such as EmoLex and Bing Liu sentiment) divorces reviews from their context and order, thus potentially missing linguistic nuances. This can be further exacerbated if text has phenomena like typos, slang, or sarcasm, some of which may be deliberately adversarial (such as phenomena like "trolling" or "astroturfing" [20]). Such bag-of-words style approaches may miss deeper semantic meaning, which appeared somewhat alleviated through more thoughtful approaches like VADER. Nevertheless, future investigation would benefit from using more modern NLP methods that utilize **transformer architectures** like BERT [21] to better capture emotional tone. Incorporating BERT or XLNet (which builds upon BERT), would further enhance a model's ability to understand the context of words in sentences much like humans do, thus providing cognitively richer analysis better-suited for capturing subtleties and nuances in reviews that comparatively simpler models, such as those of this study, might overlook [22]. Further, iterating models and averaging results across multiple random seeds would be recommended. Models may also benefit from run-time optimization, which was not considered for this study but which may be critical in models used for possible near real-time inferencing. Alternate methods for hyperparameter optimization and an expanded dataset may also be focus areas for deeper research; limiting the explored dataset to just the UCSD Amazon set of "Electronic" product reviews from May 1996-July 2014 is likely not fully representative of consumer culture and psychology, thereby introducing a path for potential **sampling bias**. Future study may benefit from human-in-the-loop test validation to ensure fidelity of review labels with respect to review content (e.g., someone who may love a product yet mis-categorizes it as being 1-star, and so on). Nevertheless, this study sought to explore the hypothesis that the inclusion of supplemental fields representative of human cognition would enhance an ML model's predictive capability, which appeared to be supported in this study's results. It is felt by the author that blending psychology-inspired features with machine learning methods measurably improves predictive performance.

References

[1] G. Lai, H. Liu, W. Xiao, and X. Zhao, "Fulfilled by Amazon: A Strategic Perspective of Competition at the e-Commerce Platform," Manufacturing & Service Operations Management, vol. 24, no. 3, pp. 1406–1420, 2022, doi: 10.1287/msom.2022.1078.

[2] D. Kaemingk, "Online Customer Review Statistics You Should Know", Qualtrics, Jan. 20, 2025. [Online]. Available: https://www.qualtrics.com/blog/online-review-stats. [Accessed: Mar. 16, 2025].

[3] M. Kollewe, "Fake Reviews: Can We Trust What We Read Online as Use of AI Explodes?," The Guardian, Jul. 15, 2023. [Online]. Available: https://www.theguardian.com/money/2023/jul/15/fake-reviews-ai-artificial-intelligence-hotels-restaurants-products. [Accessed: Feb. 5, 2025].

[4] A. Doan, et al., "Online Review Content Moderation Using Natural Language Processing and Machine Learning Methods," in 2021 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2021, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9483739

[5] N. Jindal and B. Liu, "Opinion Spam and Analysis," in Proceedings of the International Conference on Web Search and Web Data Mining - WSDM '08, 2008, doi: 10.1145/1341531.1341560.

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

[7] J. McAuley, "Amazon Datasets," University of California San Diego Computer Science and Engineering (CSE), 2024. [Online]. Available: https://cseweb.ucsd.edu/ jmcauley/datasets/amazon/links.html. [Accessed: Mar. 16, 2025].

[8] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), New York, NY, USA: Association for Computing Machinery, 2004, pp. 168–177. doi: https://doi.org/10.1145/1014052.1014073.

[9] A. Vidhya, "Removing Stop Words with NLTK Library in Python," Medium, Jun. 22, 2020. [Online]. Available: https://medium.com/analytics-vidhya/removing-stop-words-with-nltk-library-in-python-f33f53556cc1. [Accessed: Mar. 16, 2025].

[10] S. Mohammad, "NRC Emotion Lexicon," National Research Council Canada. [Online]. Available: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. [Accessed: Mar. 25, 2025].

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

[12] P. Ekman, "An Argument for Basic Emotions," Cognition & Emotion, vol. 6, no. 3-4, pp. 169-200, 1992.

[13] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, 2012.

[14] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI, 2014.

[15] C. Compagner, C. Lester, and M. Dorsch, "Sentiment Analysis of Online Reviews for Selective Serotonin Reuptake Inhibitors and Serotonin-Norepinephrine Reuptake Inhibitors," Pharmacy, vol. 9, no. 1, p. 27, Jan. 2021, doi: 10.3390/pharmacy9010027.

[16] R. Shwartz-Ziv and A. Armon, "Tabular Data: Deep Learning is Not All You Need," arXiv preprint arXiv:2106.03253, Jun. 2021. [Online]. Available: https://arxiv.org/abs/2106.03253

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," arXiv preprint arXiv:1907.10902, 2019.

[18] 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.

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

[20] J. MacDonald, “Is That Online Review a Fake?” The Christian Science Monitor, 18 Oct. 2012. [Online]. Available: www.csmonitor.com/Business/2012/1018/Is-that-online-review-a-fake.

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv preprint arXiv:1810.04805, Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[22] Z. Yang, et al., “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” in Advances in Neural Information Processing Systems (NeurIPS 2019), 2019.

APPENDIX

The project plan, including tasks and hours, is provided in Table VII below.

Note that the "Optional Midterm Check-in" was not completed.

TABLE VII: Term Project Plan: Anticipated vs. Actual Hours

Week	**Task**	**Anticipated Hours**	**Actual Hours**	**Complete?**
5	Prepare and Execute: Term Project Pitch	5	5	Y
6	Literature review on sentiment analysis in the context of online customer and cognitive science reviews	15	16	Y
7	Dataset collection and pre-processing	10	8	Y
8	Exploratory data analysis and sentiment scoring	10	7	Y
9	Train Baseline ML models	15	8	Y
9	Tune and re-train ML models	15	8	Y
10	Model evaluation	5	3	Y
10	Prepare and Execute: Optional Midterm Check-in	5	0	N
11	Analysis of results	5	5	Y
12	Prepare: Term Project - Final Report	5	15	Y
13	Execute: Term Project - Final Report	10	21	Y
14	Prepare: Term Project - Final Presentation	10	12	Y
15	Execute: Term Project - Final Presentation	5	5	Y
Total		**115**	**113**	Y