Improved Restaurant Review Analysis using VADER-based Sentiment Analysis and Automatic Rating Matching

Alphonsus Joseph Bihag
College of Science and Computer Studies
De La Salle University - Dasmariñas
bad1246@dlsud.edu.ph

Justin Brian Abus
College of Science and Computer Studies
De La Salle University - Dasmariñas
aja0666@dlsud.edu.ph

Richard Tyrese Michio Uy
College of Science and Computer Studies
De La Salle University - Dasmariñas
urm0144@dlsud.edu.ph

ABSTRACT

This study presents an automatic review rating model for restaurant reviews using a rule-based sentimental analysis tool, VADER. The study aims to predict the rating of restaurant reviews based on their underlying sentiment. Sentiment analysis, a subfield of natural language processing, was used to determine the overall sentiment of a review, whether it is positive, negative, or neutral. This study demonstrates the effectiveness of using VADER for sentiment analysis to predict the actual rating of restaurant reviews as the findings of the study indicate. Utilizing LIME the researchers also explain the words that were most considered for (1) Highest Rated Reviews (2) Middle-rated reviews (3) Lowest rated reviews. The study also explores a theory in rule-based sentiment analysis of using language translation in order to make possible changes in accuracy. This study can be useful for businesses that rely on customer reviews, such as restaurants and food delivery services to gain insights into customer satisfaction and make data-driven decisions.

CCS CONCEPTS

• Artificial intelligence ~ Natural language processing • Machine learning ~ Supervised learning ~ Supervised learning by classification

KEYWORDS

Sentimental Analysis, Valence Aware Dictionary for Sentiment Reasoning (VADER), Artificial Intelligence, LIME, Language Translation

1 INTRODUCTION

Online reviews provide a valuable evaluation of a product or service's quality, serving as a trustworthy source of insight for both consumers and businesses. These reviews demonstrate their immense helpfulness in various commercial and social areas, influencing customer attitudes and choices regarding a company's goods and services. The restaurant industry, in particular, relies heavily on word-of-mouth from consumers. Statistics show that 76% of consumers consider online reviews to be highly important in ‘food and drink’ restaurant businesses, highlighting the dynamic role these reviews play in their improvement [9]. However, a critical challenge lies in these reviews, inconsistency between the actual content of the review and the provided rating by their author [12]. This discrepancy undermines the trustworthiness of reviews and hinders accurate interpretation. Sentiment analysis, a technique utilizing Natural Language Processing (NLP) to analyze textual sentiment, emerges as a potential solution to bridge the inconsistency gap [7].

Advancements in Natural Language Processing (NLP) have enabled computers to analyze and understand human language with increasing accuracy. This study utilizes VADER (Valence Aware Dictionary for Sentiment Reasoning), as a tool to address inconsistencies between review content and ratings provided. It is a rule-based sentiment analysis tool that follows grammatical and syntactical conventions for translating sentiment intensity. Most sentiment analysis models that use supervised learning algorithms these days consume loads of labeled data in the training phase to give satisfactory results which is usually expensive and leads to high labor costs in real-world applications [3]. However, VADER comes with its sentiment analysis lexicon, disregarding most of these costs. It is also a gold standard list of lexical features suitable for finding semantics in micro-blog text [1].

This study aims to classify various restaurant reviews using VADER-based sentiment analysis to provide matching ratings with restaurant reviews found online and determine the performance of the model.

2 METHODOLOGIES

2.1 Area of Study

The internet revolutionized how people interact with information and services, including the way they discover and share restaurant experiences. This research focuses on online restaurant reviews, specifically those found on social media platforms like Facebook and dedicated review websites like Zomato. Facebook, with its vast user base and ingrained social features, provides a unique platform for food reviews. Users can share their dining experiences with friends and followers, offering valuable insights and influencing the restaurant choices of others. Additionally, Facebook's search functionality allows users to discover reviews from a wide range of individuals, creating a comprehensive information pool on various restaurants and cuisines. Meanwhile, platforms like Zomato offer a wealth of restaurant-specific information, including menus, user reviews, and star ratings. This data allows researchers to delve deeper into consumer trends and conduct market research within the food industry. By analyzing Zomato reviews, we can gain valuable insights into consumer preferences, identify top-performing restaurants, and understand how factors like location and pricing influence a restaurant's success.

2.2 Data Gathering Procedure

Reviews amounting to 1150 were manually obtained by the proponents from Zomato and Facebook, listing all reviews from various restaurants which were compiled into an Excel (xlsx) file. The researchers provided the following information for each review: textual feedback, true rating, source, year written, and the restaurant for which the reviews were written. The 'true ratings' are derived from the evaluation of outside evaluators that were independent of both the original author and the proponents to label each review based on text connotations for model evaluation. For the selection of the reviews, the proponents focused on reviews of restaurants with a physical presence or local branch in the Philippines. Only reviews within a five-year range of the study's date of conduct were counted among the data for sentiment analysis. Also English, Filipino, and Taglish reviews were collected for this study. Further descriptions of the variables considered by the study in gathering data and other variables during the sentiment analysis procedure are provided in Table 1 below:

Variables	Description
Review	These are the feedback provided by customers of restaurants for their products and service either for the purpose of praise, suggestion, or expression of negativity.
True Rating	A label provided by external evaluators that classifies the reviews as either Positive, Neutral, or Negative for model evaluation.
Source	The website from which the reviews were obtained.
Year Written	The year in which the reviews were posted by their author in their respective source.
Recipient Restaurant	The restaurant for which the review was written by their author.
VADER Compound	The compound score produced by VADER that attributes the degree or score in negative, neutral, or positive altogether.
Star - Rating	The output of the model constructed in this study represents the scale of a review in its degree of negativity or positivity.

Table 1: Description of the Variables Considered and Used in the Study

2.3 Data Processing

Since VADER sentiment analysis operates primarily in English, reviews written in Filipino or Taglish required translation. To ensure consistency and efficiency, a batch translation approach was employed utilizing Google Translate. This involved grouping the Filipino and Taglish reviews together and submitting them for translation at once. While Google Translate offers a valuable tool for basic comprehension, it is important to acknowledge that nuances and cultural references within the reviews might not be perfectly captured in the translation process. The newly translated reviews were then combined with the reviews written in English.

Following the translation process, VADER-based sentiment analysis was applied to obtain a vader compound for each review. The vader compound was translated into matching ratings following a balanced distribution of scores that VADER could output from a range of -1 to +1 as discussed in Table 2 below:

Ratings and Sentiment	Compound Score Range
5 – Stars (★★★★★) (Very Positive)	0.60 to 1.0
4 – Stars (★★★★☆) (Positive)	0.21 to 0.59
3 – Stars (★★★☆☆) (Neutral)	-0.20 to 0.20
2 – Stars (★★☆☆☆) (Negative)	-0.59 to -0.21
1 – Star (★☆☆☆☆) (Very Negative)	-1.0 to -0.6

Table 2: VADER Compound Score to Matching Rating System

This system of assigning ratings based on the compound score is derived from their equivalent sentiment middle-ground where in the context of restaurant reviews, more stars depicted greater positivity [8].

2.4 Language and Model

Valence Aware Dictionary for Sentiment Reasoning or VADER relies on a dictionary that maps lexical features to emotion intensities called sentiment scores [2]. These scores are appropriately categorized into three categories that include neg (negative), neu (neutral), and pos (positive) to produce a compound score that factors in the previous categories based on the analysis of a given text. It computes the compound score using the formula below:

$$x = \frac{x}{\sqrt{x^2 + \alpha}}$$

Figure 1: Compound Score Formula for VADER

Where x = sum of valence scores of constituent words, and α = Normalization constant in which the default value is fifteen (15).

The compound score is the sum of the valence scores, adjusted according to the rules of the Sentiment Reasoning dictionary that is VADER, normalized to be between -1 for 'most extreme negative' and +1 for 'most extreme positive' [13].

However, in the case of the data produced by VADER, as it focuses on calculating and producing scores, it lacks proper explainability in the analysis process for humans to be able to understand. For this, an

algorithm called Local Interpretable Model-agnostic Explanations (LIME) may be used to help explain the prediction process of VADER [11]. It works by constructing a local interpretable model by finding the most important features [4][11] based on a set of calculated probabilities that are separated into provided classes based on a given sample of text. As a post-hoc method, it performs its processes after a prediction is made, meanwhile, LIME is able to present a visual model or aid of the probability calculated with the method separated into classes, top features, and textual evidence by highlighting the features from the text sample.

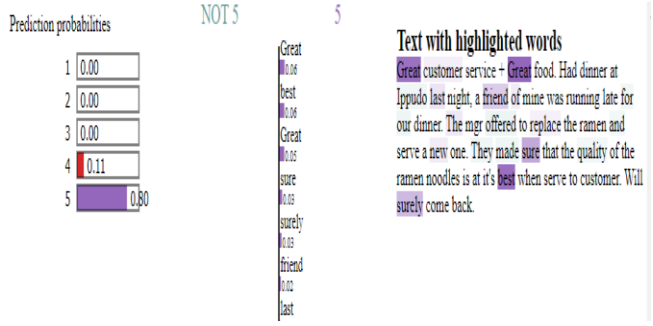

Figure 2: Visual Model produced by LIME

As exhibited by Figure 2, LIME utilizes the calculated probabilities and categorizes them into their appropriate class based on the method that has been factored in from the prediction of the sentiment classifier (VADER). It also shows a sorted graph of the features based on their relevance to the text sample fed to LIME while also providing a visual representation of the sample with highlights on the words shown in the sorted graph.

For the proponents of the study to interpret the data produced by their chosen method of sentiment analysis, LIME is used to explain certain samples of data. However, as rule-based methods such as VADER do not output class probabilities as in VADER's case that only outputs a single score (compound score), in order to utilize LIME to explain the results, it is needed to artificially generate the class probabilities.

```
def prediction(text):
    probs = []
    x = 0

    # First, offset the float score from the range [-1, 1] to a range [0, 1]
    offset = (vadar_sentiment(text) + 1) / 2.
    # Convert offset float score in [0, 1] to an integer value in the range [1, 5]
    binned = np.digitize(5 * offset, np.array([1, 2, 3, 4, 5])) + 1
    # Similate probabilities of each class based on a normal distribution
    simulated_probs = scipy.stats.norm.pdf(np.array([1, 2, 3, 4, 5]), binned, scale=0.5)

    while x < len(simulated_probs):
        probs.append(simulated_probs[x])
        x = x + 1
    this = np.array(probs)
    this

    return this
```

Figure 3: Artificial Class Probability Procedure

The procedure shown in Figure 3 utilizes a simple workaround to simulate the class probabilities using a continuous-valued sentiment score from the original range of '-1' to '1' by VADER to a normalized float score within the range of '0' to '1' that is scaled to five times in magnitude for each class in which for this case is based on the star-based rating system for reviews [10].

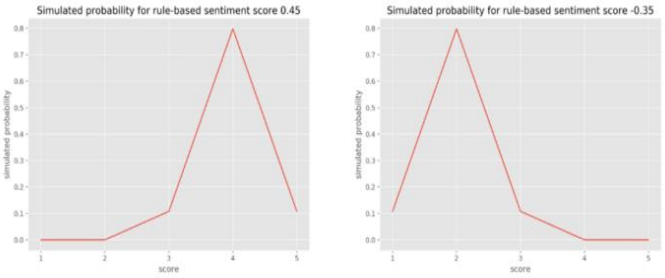

Figure 4: Examples of Simulated Probabilities Using the Work-around Procedure of Rao (2019): Artificial Class Probability Procedure

As shown in Figure 4, using the workaround procedure can adjust the compound score produced by a rule-based sentiment analyzer like VADER to the appropriate scale for the study, 0.45 was properly scaled into 4 in the graph from the left while -0.35 was appropriately scaled into 2 as found in the graph from the right.

2.4.1 Model Evaluation

Sentiment analysis relies on accurate results to ensure effectiveness. Metrics like precision, recall, and F1-score are calculated to assess this, considering how well the system classifies texts. These scores depend on correctly identifying positive, negative, and neutral sentiment, with this study expanding on existing metrics to include true Neutral (TN) and false Neutral (FN) as derived from Kanstren.

From this, it is possible to compute the following scores or metrics using the formulas that are summarized in the given Table 3:

Score/ Metric	Formula
Accuracy	$\frac{\text{True Positives + True Negatives}}{\text{True Positives + True Negatives + False Positives + False Negatives}} = \frac{\text{N. of Correct Predictions}}{\text{N. of all Predictions}} = \frac{\text{N. of Correct Predictions}}{\text{Size of Dataset}}$
Precision	$\frac{\text{True Positives}}{\text{True Positives + False Positives}} = \frac{\text{N. of Correctly Predicted Positive Instances}}{\text{N. of Total Positive Predictions you Made}}$:
Recall	$\frac{\text{True Positives}}{\text{True Positives + False Negatives + False Neutrals}}$
F1-Score	$2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$

Table 3: Summary of Metric and their Formulas

Sentiment analysis relies on several metrics to evaluate its effectiveness. Accuracy, the most fundamental metric, measures the overall proportion of correct predictions. Precision focuses on the exactness of positive predictions, while recall emphasizes the model's ability to identify all actual positive cases. Finally, the F1-score

combines both precision and recall for a more balanced assessment [5][6].

This study incorporates "Neutral" as a sentiment category. To account for this, an adjusted recall score will be calculated, penalizing the model for both false negatives (missing positive cases) and false neutrals (missing neutral cases). This adjustment ensures a more comprehensive evaluation of the model's performance.

3 RESULTS AND DISCUSSION

Actual vs Predicted Results			
TARGET / OUTPUT	Actual	Predicted	SUM
Actual	247 23.73%	188 18.06%	435 56.78% 43.22%
Predicted	16 1.54%	590 56.68%	606 97.36% 2.64%
SUM	263 93.92% 6.08%	778 75.84% 24.16%	837 / 1041 80.40% 19.60%

Figure 5: Checking Overall VADER Accuracy after translation

Figure 5 shows the accuracy of the sentiment analysis model in classifying the reviews. The model in classifying positive and negative reviews shows that it has an 80.40% accuracy.

```
print(f"True Positive count is {truePositiveCT}")
print(f"True Negative count is {trueNegativeCT}")
print(f"True Neutral count is {trueNeutralCT}")
print(f"False Positive count is {falsePositiveCT}")
print(f"False Negative count is {falseNegativeCT}")
print(f"False Neutral count is {falseNeutralCT}")
```

```
True Positive count is 590
True Negative count is 247
True Neutral count is 14
False Positive count is 188
False Negative count is 16
False Neutral count is 95
```

Figure 6: Identifying the True and False Prediction for the Translated and Combined Dataset

Figure 6 shows the total number of true or false positives, true or false negatives, and true or false neutrals after language translation. It was able to count a total of 590 for True Positive, 247 for True Negative, 14 for True Neutral, 188 for False Positive, 16 for False Negative, and 95 for False Neutral classifications. It can be observed that the majority of the predictions made fell under the True classification, showing that the model is significantly effective. After evaluation, the model performed with an F1-score of 0.85, a precision of 0.75, a recall of 0.97, and an overall accuracy of 74% when including neutrality.

Figure 7: LIME on Highest-Rated Reviews

By using the LIME Explainer model, ratings made by the model were easier to understand. Figure 7 shows the highest-rated review and how it was analyzed by VADER through LIME. It can be observed that is filled with positive words including 'pleasing' 'terrific' and 'better' leading to the review being the highest-rated. It shows the vast number of words that VADER considered to classify reviews.

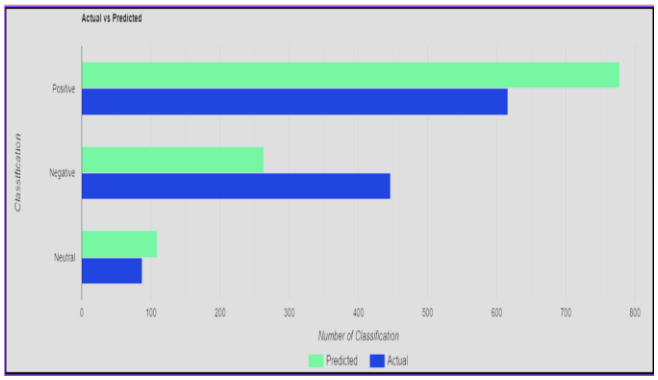

Figure 8: *Comparison of Actual and Predicted Classifications*

Figure 8 shows the comparison between the number of actual and predicted classifications. At the end of the sentiment analysis, the model classified 773 reviews as positive, 267 as negative, and 109 as neutral. In comparison to the actual classification of the data where 616 were positive, 446 negative, and 87 were neutral, the model appears to overestimate the positivity in the sentiment. There is a discrepancy of 157 classifications between positive and negative categories, with the model classifying 157 more reviews as positive than the actual data, classified by outside evaluator.

4 FUTURE WORK

This study acknowledges limitations due to VADER's English-centric nature. For multilingual data, translating reviews in English or using alternative NLP methods trained on the specific languages is recommended. Additionally, exploring state-of-the-art neural networks for sentiment analysis is suggested for potentially higher accuracy. Finally, the importance of a larger, balanced dataset with diverse sources is emphasized to enhance the overall analysis.

REFERENCES

[1] Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, *8*(S2), 1–6. https://doi.org/10.51983/ajcst-2019.8.s2.2037

[2] Calderon, P. (2018, March 31). *Vader sentiment analysis explained*. Medium. https://medium.com/@piocalderon/vader-sentiment-analysis-explained-f1c4f9101cd9

[3] Chiny, M., Chihab, M., Bencharef, O., & Chihab, Y. (2021). LSTM, Vader and TF-IDF based hybrid sentiment analysis model. *International Journal of Advanced Computer Science and Applications*, *12*(7). https://doi.org/10.14569/ijacsa.2021.0120730

[4] De Sousa Silveira, T., Uszkoreit, H., & Ai, R. (2019). Using aspect-based analysis for explainable sentiment predictions. *Natural Language Processing and Chinese Computing*, 617–627. https://doi.org/10.1007/978-3-030-32236-6_56

[5] Johnson, J. (2020, July 22). Precision, recall & confusion matrices in Machine Learning. BMC Blogs. https://www.bmc.com/blogs/confusion-precision-recall/

[6] Kanstren, T. (2020, September 12). A look at precision, recall, and F1-score. Medium. https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

[7] Kouadri, W. M., Ouziri, M., Benbernou, S., Echihabi, K., Palpanas, T., & Amor, I. B. (2020). Quality of sentiment analysis tools. *Proceedings of the VLDB Endowment*, *14*(4), 668–681. https://doi.org/10.14778/3436905.3436924

[8] Nielsen, N. (2024, February 29). *Restaurant rating system: Your guide to understanding reviews and stars*. Limepack Restaurant Rating System Your Guide to Understanding Reviews and Stars Comments. https://www.limepack.eu/blog/restaurant-rating-system-your-guide-to-understanding-reviews-and-stars

[9] Pitman, J. (2023, September 7). *Local consumer review survey 2022: Customer reviews and behavior*. BrightLocal. https://www.brightlocal.com/research/local-consumer-review-survey-2022/

[10] Rao, P. (2019, September 9). Fine-grained sentiment analysis in Python (part 2). Medium. https://towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-2-2a92fdc0160d

[11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939778

[12] Shan, G., Zhou, L., & Zhang, D. (2021). From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems*, *144*, 113513. https://doi.org/10.1016/j.dss.2021.113513

[13] Swarnkar, N. (2020, May 21). *Vader sentiment analysis: A complete guide, algo trading and more*. Quantitative Finance & Algo Trading Blog by QuantInsti. https://blog.quantinsti.com/vader-sentiment/#:~:text=that%20hot.%E2%80%9D.-,Compound%20VADER%20scores%20for%20analyzing%20sentiment,1%20(most%20extreme%20positive