Predicting Usefulness of Yelp Reviews: A Comparative Study of Algorithms

Lola Akinsehinwa
Department of Data Science
Carolina University
Winston-Salem, USA
akinsehinwak@carolinau.edu

Abstract—Online customer reviews have become increasingly popular over the years and are often the deciding factor in decision-making. Online platforms such as Yelp, which hosts millions of user reviews for products and services, incorporate user interactivity by allowing users to rate reviews as "useful," "funny," or "cool." Among these, the "useful" voting feature is the most impactful as it plays a key role in forming prospective customers' impressions of a company. For organizations, understanding what constitutes a 'useful' review provides the opportunity to develop strategies that can increase growth and revenue, improve customer experiences, and strengthen brand perception. This study evaluates the performance of traditional machine learning models, including a Support Vector Machine (SVM), Naive Bayes, and a transformer-based model (DistilBERT), in predicting review usefulness. The analysis of the Yelp Open Dataset found that the SVM model achieved the highest overall performance in classifying review usefulness. Future research could explore additional machine learning and more advanced transformer-based models to enhance prediction accuracy.

Keywords—Review Usefulness, DistilBERT, Support Vector Machine, Naive Bayes.

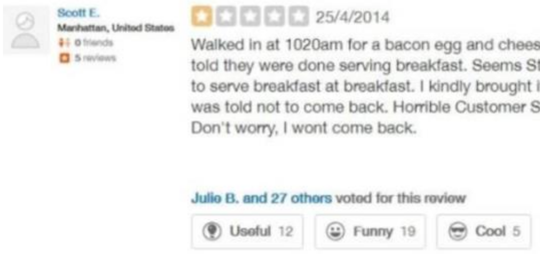
I. Introduction

Over the years, the digital landscape has transformed how customers engage with products and services. Consequently, online consumer reviews have created new paradigms for exchanging information, offering consumers indirect experiences with products and services, and swaying purchasing behavior and business operations [1]. With a strong web and mobile presence, platforms such as Yelp have become essential digital resources that shape consumer decisions. The Yelp platform includes interactive features that allow users to vote for reviews as "useful," "cool," or "funny."

Among the feedback options, the "useful" vote is particularly impactful, as it directly influences the perceived credibility of reviews and can affect how customers arrive at a decision [3]. Therefore, it is important to explore what makes reviews "useful." As described in [4], review usefulness "represents the subjective valuation of the review judged by others and is also the aggregate perceived utility of the information contained in the review". Essentially, reviews that are considered useful add perceived value to the products and services being offered [4].

According to [2], Yelp has approximately 76 million monthly site visits, and gaining insights into useful reviews can benefit businesses. For instance, such insight can be an avenue through which an organization can improve its products and service offerings, which can drive customer conversion. Moreover, patterns of constructive criticism can help businesses improve their daily operations and customer service, directly influencing customer retention and loyalty [4]. The concept of review usefulness is complex and multifaceted, as what qualifies a review as useful is subjective [4]. For example, one might conclude that a review is helpful based on its length, whereas another might not reach the same conclusion. Through exploratory data analysis, this study aims to uncover the trends and attributes that contribute to the perceived usefulness of a review.

Figure 1. – Review Rating Interface on Yelp.com



Scott E.
Manhattan, United States
0 friends
5 reviews

25/4/2014

Walked in at 1020am for a bacon egg and cheese. I was told they were done serving breakfast. Seems Strange not to serve breakfast at breakfast. I kindly brought it up and was told not to come back. Horrible Customer Service. Don't worry, I wont come back.

Julio B. and 27 others voted for this review

II. Related Work

Online reviews are considered a new age of word-of-mouth, known as electronic word-of-mouth, creating a digital repository of opinions that is accessible to both businesses and potential consumers [5]. Research has shown that between 73% and 87% of consumers actively search for restaurant reviews before making dining decisions [5], demonstrating the influence of user reviews on consumer behavior.

Past research on review usefulness have explored a wide range of factors that contribute to its perceived value. For example, Mudambi and Schuff (2013) revealed a positive correlation between word count and review usefulness but did not determine the ideal review length [4]. Essentially, the length of a review can influence users' perceptions of its usefulness, but it is not the primary factor. According to Cao et.al (2011), while the semantic characteristics of a review have a greater impact on usefulness, reviews with strong opinions are the key drivers of perceived usefulness [4]. The implications of extreme opinions raise concerns about the potential for biased or misleading reviews. According to [4], extreme reviews do not indicate review quality, as they may include irrelevant information that does not help potential readers seeking to make a decision. Furthermore, researchers have found that the characteristics of a review text, including its structure, language style, and contextual features, directly influence the perception of a review as useful [4].

Despite the research on review usefulness, there are noted gaps in the literature. While prior research has identified contributing factors, limited research has focused on developing predictive models that compare traditional machine-learning techniques with newer deep-learning approaches. This study aims to address this gap by evaluating the performance of three models, specifically, Support Vector Machine (SVM), Naive Bayes (NB), and DistilBERT, a transformer-based deep learning model, to predict review usefulness. Although the Yelp Open Dataset provides data on multiple industries and sectors, this study focuses exclusively on restaurant reviews.

III. Methodology

A. Data Collection

The publicly available Yelp Open Dataset was sourced from Yelp.com and contains information about businesses, reviews, and users across multiple metropolitan areas. The review file contained approximately 6.99M records with features including review_id (a unique identifier for each review), user_id (a unique identifier for the user who wrote the review), business_id (the identifier of the business being reviewed), stars (the rating a user gave to a business), useful (the number of users who found the review useful), funny (the number of users who found the review funny), cool (the number of users who found the review cool), text (the content of the review), date (the date the review was posted), and name (the name of the user who wrote the review). The business file contained 150,346 records with features, including business_id (a unique identifier for each business), name (name of the business), address (physical address of the business), city (city where the business is located), state (state where the business is located), latitude and longitude (geographical coordinates), stars (average rating of the business), review_count (total number of reviews for the business), is_open (binary indicator that represents whether a business is operational), categories (business type classifications), and hours (business hours of operations).

B. Data Preprocessing

Data preprocessing is necessary before implementing machine-learning models to ensure optimal model performance. Because of memory limitations, the Yelp review dataset was processed in chunks (e.g., 1M records per chunk). First, the business dataset was filtered to include only restaurants based on the categories field and operational businesses (is_open = 1). The resulting records from the business and review files were then merged using the business_id field to create a unified dataset. To ensure distinct entries, duplicate records based on review_id were excluded. Stratified sampling was applied based on review star ratings and the "useful" field for balanced representation. Thereafter, a sample size of 30,000 records were selected for the analysis. Furthermore, irrelevant columns were excluded from the dataset, including unique identifiers (e.g., review ID and user ID), funny, cool, geographical fields (e.g., address, longitude, and city), and detailed business information (e.g., hours and attributes). The raw review text (text field) was processed using several cleaning steps, including converting the text to lowercase and removing punctuation, numbers, and extra whitespace. In addition, vectorization was applied using Term-Frequency – Inverse Document Frequency (TF-IDF) to convert the data into numerical features. New features were created from the text field, including "review length," "word count," and "sentence count." The "useful" field was categorized into two classes including, not useful (zero votes) and useful (one or more votes). Although stratified sampling was applied, the dataset remained highly imbalanced as non-useful reviews dominated. To address this, random undersampling was applied to reduce the majority class (non-useful reviews) to 12,100 samples, resulting in a balanced dataset. The final dataset was split into 80% training and 20% testing sets for model evaluation.

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) includes techniques applied to a dataset to understand the patterns, trends, and relationships among features. EDA is an invaluable method for gaining insight into data distribution, making sense of complex data, and providing an effective method for detecting outliers. It can also help inform data preprocessing and model selection. Figure 1.2 illustrates the distribution of useful and non-useful reviews. The results show that approximately 60% of users categorized reviews as not useful, whereas 40% categorized reviews as useful. An imbalance in the dataset presents a challenge for classification tasks because most reviews are not useful.

Descriptive statistics revealed that the average review and business rating is 3.5 on a 5-star scale, which suggests that customers are generally satisfied with the restaurants they visit. Furthermore, the average review length was 522 words, with a maximum sentence count of 7.5 per review. As shown in Figure 1.2, the spread of useful votes ranges from 0 to 28, with most votes at zero, indicating non-usefulness. Additionally, review ratings show an upward trend over the years, supported by past research (Figure 1.3). Further analysis revealed that five-star reviews had the most extended review length, followed by four-star reviews (Figure 1.4). Moreover, the analysis revealed that one-, two-, and three-star reviews are generally shorter. In addition, the relationship between review length and usefulness was analyzed, and the results showed that the distribution of review lengths for both useful and non-useful reviews shared a similar pattern, with most reviews falling between zero and 250 (Figure 1.5).

Figure 1.1

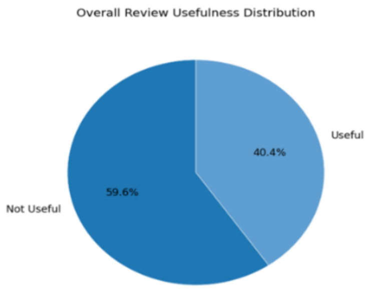


Figure 1.2

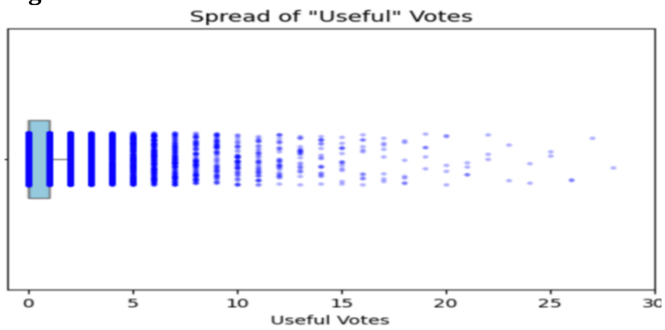


Figure 1.3

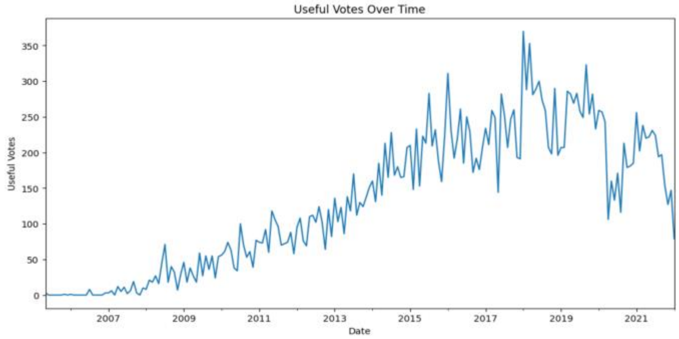


Figure 1.4

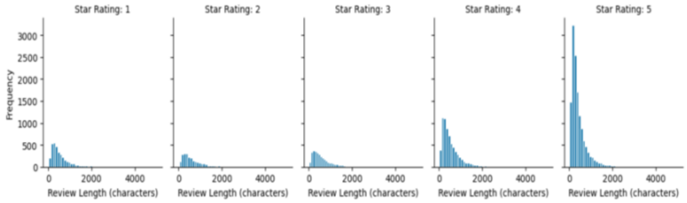
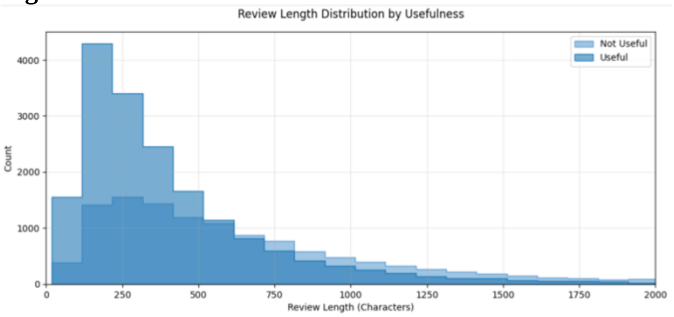


Figure 1.5



D. Machine Learning and Transformer-Based Models

a) Multinomial Naïve Bayes): Multinomial Naive Bayes (MNB) is a variant of Naive Bayes designed for classification tasks where features are discrete and countable, such as word counts in text classification [8]. The MNB model estimates the likelihood of each class (useful and non-useful) based on the frequency of words in the training data. It also assumes conditional independence between features, which can affect model performance [8]. The likelihood of observing each class is defined as:

$$P(C|X) = \frac{P(C)\prod_{i=1}^{n} P(x_i|C)}{P(X)}$$

where, P (C) is the prior probability of class C (e.g., the likelihood a review is "useful"), P(xi|C), the probability of observing feature xi (a word) given the class C, P(X) is a constant that makes the probabilities across classes sum to 1[7]. In this study, the MNB model was implemented using a pipeline method. Text data were first transformed into numerical features using Term Frequency–Inverse Document Frequency (TF-IDF) vectorization. Hyperparameter tuning was then applied using grid search with cross-validation to optimize model performance.

b) Support Vector Machine (SVM): The SVM model is a supervised machine learning algorithm that is effective for classification tasks. In text classification, SVM is effective for handling high-dimensional data. The SVM model works by finding the best decision boundary, the hyperplane, in a feature space that maximizes the margin between classes [6]. The key components of SVM include a hyperplane that separates classes, support vectors, the observations from each class closest to the hyperplane, margins that measure the distance between the support vectors and the hyperplane, and a kernel function (e.g., linear, radial basis function (RBF)) that transforms data into a higher-dimensional space to handle data that is not easily separable [6]. In this study, the RBF kernel was used to transform the data into a higher-dimensional space, which allowed the model to separate the two review classes. The model was trained using TF-IDF vectorized features from the review text, and the hyperparameters were fine-tuned to achieve optimal performance.

c) DistilBERT: DistilBERT is a smaller and faster version of BERT (Bidirectional Encoder Representations from Transformers) [9], a transformer-based model used for a wide range of natural language processing (NLP) tasks, including text classification. It is a pre-trained model that uses self-supervised learning on a large corpus of text data [9]. Unlike traditional machine learning models, transformer-based architectures such as DistilBERT are bidirectional, meaning they consider the context before and after a target word [10]. This is achieved through Masked Language Modeling (MLM), where the model randomly masks a percentage of input words and learns to predict the masked words [9]. Furthermore, DistilBERT uses pretrained contextual embeddings to represent words as vectors in n-dimensional space [10][11]. Positional encodings are added to the token embeddings [10][11] to keep the sequential order of words, which allows the model to understand the sequence of words effectively. The DistilBERT model's self-attention mechanism allows it to weigh the importance of different words relative to each other [10], another key difference from the traditional machine learning model. This attribute allows the model to capture meaning from text. This study used and fine-tuned the pre-trained distilbert-base-uncased model from the Hugging Face Transformers library to predict review usefulness. The review text data was tokenized using the DistilBERT tokenizer, and early stopping was implemented to prevent overfitting.

IV. Results

Developing a robust model for predicting review usefulness can help businesses develop strategies to improve their services and customer experience. In this study, the metrics used to assess the performance of the models include accuracy, precision, recall, and F1-score. Accuracy is a metric that measures the proportion of total correct classifications [12]. Precision measures how often a model's positive predictions are correct (i.e., useful reviews). Recall represents the proportion of actual positives identified as positives [12]. The F1-Score is the harmonic mean of precision and recall, indicating how well the model balances these two metrics. Table 1 summarizes the performance of each model for Class 1("useful" reviews).

Table 1

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SVM	0.86	0.87	0.86	0.86
Naïve Bayes	0.72	0.73	0.70	0.72
DistilBERT	0.82	0.83	0.83	0.82

V. Discussion

The results of this study indicate that both traditional and deep learning models can effectively predict review usefulness. Among the three models evaluated, the Support Vector Machine (SVM) demonstrated the best overall performance across all evaluation metrics; this can be attributed to the RBF function, which effectively maps data into high-dimensional space and allows for the effective separation of classes. DistilBERT also maintained strong performance, which could be the result of the pre-trained model on a large corpus of text data, the use of contextual word embeddings, positional encoding, and its bidirectional processing abilities. In contrast, Naïve Bayes did not perform as well as the SVM and DistilBERT. This can be attributed to the assumption of independence between features. As Cao et al. [4] noted, semantic characteristics and relationships between words play important roles in determining review usefulness, which the Naïve Bayes model may not fully capture. Overall, each model maintained consistent performance across all metrics, but SVM had the best overall performance, followed by DistilBERT and Naïve Bayes.

While these models performed well in predicting overall usefulness, there is room for improvement. Future research could explore additional resampling techniques, such as random oversampling or the Synthetic Minority Oversampling Technique (SMOTE), to assess whether these techniques could further improve the performance of each model. Despite its strong performance, this study has several limitations. First, the "useful" feature was limited to two categories: useful (more than one vote) and non-useful (zero votes). Future research could include multiclass labeling to capture the nuanced nature of usefulness. Another limitation is that the original dataset was highly imbalanced, with non-useful votes dominating. While the random undersampling technique achieved a balanced dataset, future research should consider alternative techniques to retain more of the original data distribution.

VI. Conclusion

Online reviews strongly influence consumer behavior, often swaying customer decisions to engage with businesses. In the digital age, reviewing businesses has steadily increased over the years and is often an important factor for prospective customers seeking more information before engaging with products and services. With millions of users, platforms such as Yelp provide features that allow users to upvote reviews as useful. The useful upvote feature is highly influential as it can shape the perception of how prospective customers view businesses. Several factors, including review length, text structure, and language quality, influence the perceived usefulness of reviews. As such, businesses should take reviews that are considered useful seriously and use them to improve customer experiences. Businesses can better understand the underlying factors that drive customers to leave useful reviews by developing predictive models to classify review usefulness.

REFERENCES

[1] S. Bae and T. Lee, "Product type and consumers' perception of online consumer reviews," *Electronic Markets*, vol. 21, no. 4, pp. 255–266, Nov. 2011, doi: https://doi.org/10.1007/s12525-011-0072-0.

[2] Yelp Inc, "Study shows high-intent consumers are contacting businesses quickly on Yelp," *Yelp for Business*, Jan 16, 2023. https://business.yelp.com/resources/articles/study-shows-high-intent-consumers-are-contacting-businesses-quickly-on-yelp/?domain=local-business (accessed Apr. 06, 2025).

[3] Y. Lim and B. Van Der Heide, "Evaluating the Wisdom of Strangers: The Perceived Credibility of Online Consumer Reviews on Yelp," *Journal of Computer-Mediated Communication*, vol. 20, no. 1, pp. 67–82, Aug. 2014, doi: https://doi.org/10.1111/jcc4.12093.

[4] A. H. Huang, K. Chen, D. C. Yen, and T. P. Tran, "A study of factors that contribute to online review helpfulness," *Computers in Human Behavior*, vol. 48, pp. 17–27, Jul. 2015, doi: https://doi.org/10.1016/j.chb.2015.01.010.

[5] E. S. Alamoudi and N. S. Alghamdi, "Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings," *Journal of Decision Systems*, pp. 1–23, Jan. 2021, doi: https://doi.org/10.1080/12460125.2020.1864106.

[6] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. "O'Reilly Media, Inc.," 2019.

[7] E. Frank and R. R. Bouckaert, "Naive Bayes for Text Classification with Unbalanced Classes," *Springer Link*, 2006. https://link.springer.com/chapter/10.1007%2F11871637_49

[8] Scikit-learn, "sklearn.naive_bayes.MultinomialNB — scikit-learn 0.22 documentation," *Scikit-learn.org*, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

[9] "distilbert/distilbert-base-uncased · Hugging Face," *huggingface.co*, Mar. 11, 2024. https://huggingface.co/distilbert/distilbert-base-uncased

[10] A. Vaswani *et al.*, "Attention Is All You Need," 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[11] Transformer Explainer: LLM Transformer Model Visually Explained," *Github.io*, 2017. https://poloclub.github.io/transformer-explainer/

[12] google, "Classification: Accuracy, recall, precision, and related metrics," *Google for Developers*, 2024. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall