ONLINE REVIEW HELPFULNESS PREDICTION

1.0 INTRODUCTION

Online marketplaces have exploded in popularity in recent years for their convenience. Online shoppers are four times more likely to order from an online marketplace than from a retailer in the UK. One benefit of online shopping is that potential customers can easily access a pool of reviews provided by other consumers (Akbarabadi & Hosseini, 2018). Helpful reviews allow users to grasp more clearly the features of a product they are about to buy and thus understand whether it fits their needs (Passon et al., 2018). Product reviews can be viewed as a passive recommendation process, and these reviews could reflect user satisfaction with the experience of a product or service purchased. More people resort to digital reviews before buying a good or deciding where to eat or stay (Passon et al., 2018).

Given the importance of product review helpfulness, a growing body of studies has been dedicated to exploring what factors may affect the helpfulness of online reviews (Zhou & Yang, 2019). However, the enormous volume and vastly varying quality of available reviews provide a significant obstacle to seeking out the most helpful reviews (Lee et al., 2020). In order to understand which factors are used to determine the review helpfulness, it is necessary to perform a semantic analysis of the review content. LIWC-22 (Linguistic Inquiry and Word Count dictionary) is a text analysis tool that compares each word in the review text against a user-defined dictionary, and the dictionary identifies which words are associated with which psychologically-relevant categories. After analyzing and accounting for all words in a given review content, it calculates the percentage of total words that match each dictionary category (LIWC (Linguistic Inquiry and Word Count), 2019). Therefore, our research decided to use LIWC-22 to generate factors associated with reviews, and these quantitative factors can then be ranked by importance using Random Forest, Mutual Information, and RFE methods.

In this project, XGBoost, Random Forest, Support Vector Machine, Naïve Bayes, and Linear Regression models were the selected machine learning algorithms to predict the review helpfulness with the feature selection on the Amazon review dataset. The rest of the article is organized as follows: Section 2 briefly summarizes related research. Section 3 presents the metadata and summarizes the research steps used in the study. Section 4 presents the model results and discussion. Finally, section 5 concludes the article and proposes some limitations and future research directions.

1.1 Problem Statement

Given the importance of review helpfulness in supporting customer purchasing decisions, more and more people are referring to digital reviews before buying a good or deciding where to eat or stay (Passon et al., 2018). Various research has been done to predict helpful reviews, such as Zhou & Yang (2019), Lee et al. (2020), and Akbarabadi & Hosseini (2018). However, they used only one feature selection method in the combination of their machine learning techniques, which may lead to biases in feature importance. Furthermore, some researchers, such as Zhou & Yang (2019), used an outdated text analysis tool, LIWC, released in 2015. It is outdated and not comprehensive enough for today's text analysis requirement, which may lead to incomprehensive results and affect the helpfulness. predictions.

1.2 Objectives

1. To identify suitable machine learning techniques for predicting helpful reviews.
2. To find significant features in identifying helpful reviews.
3. To conduct an assessment of machine learning models for prediction review helpfulness.

Hean Kuan Kang (S2005669)   Chen Yayi(S2110521)   Siyi Zhou(S2031274)
Chenchen Liang(S2113246)   Qiyi Liu(S2106632)

2.0 LITERATURE REVIEW

2.1 Previous Research

Haque et al. (2018) used four feature extraction techniques to extract reviews with star ratings as lexical, structural, semantic, and combined LS2F features, and the combined LS2F features outperformed other individual features in the following prediction module. Their proposed feature extraction technique (LS2F) can quantify the helpfulness of each product based on user-submitted reviews to address the problem of review accumulation in helpfulness prediction. Lee et al. (2020) suggested a helpful prediction model using NBN. They used the conditional probability of the binned determinants to suggest the importance of each determinant in the expanded list of predictors, including product, reviewer, and other textual factors through NBN. To validate their model, the authors compared the predictive performance of NBN, multivariate discriminant analysis (MDA), k-nearest neighbors (kNN), and neural network (NN), and the prediction accuracy of NBN performed much better.

Zhou & Yang (2019) studied the impact of the numerical and textual factors of reviews on review helpfulness prediction. Their findings indicated that text sentiment has a negative effect on review helpfulness and numerical factors are more important than textual factors. Finally, the authors used random forest to predict the review helpfulness based on its numerical and textual factors. Akbarabadi & Hosseini (2018) mainly studied the influence of review title features on predicting online review helpfulness. First, they proposed a new method for classifying action verbs into four main categories of text, reviewer, readability, and title features, and then they tested their proposed prediction model on two real Amazon datasets. The results showed satisfactory performance but also revealed that the title feature is not important enough to be a key factor in determining the usefulness of reviews.

Park (2018) examined the possibility of predicting Amazon review helpfulness using features from MARGOT, an automatic argumentation mining system. To evaluate their proposed approach, the authors conducted a prediction task using Support Vector Regression, Linear Regression, Random Forest, and M5P Tree on 117,000 Amazon reviews to confirm that it is indeed feasible. Please refer to the appendix for more detailed comparison information.

2.2 Feature Selection

Mutual information: It is one of the most popular feature selection methods in machine learning and is used to measure the degree of interdependence between two random variables. Its value is always greater than or equal to zero, and the larger the value, the greater the relationship between the two variables (Brownlee, 2020). In other words, the variables are independent if the mutual information is zero.

Random Forest: Each tree in a random forest can calculate the importance of features based on its ability to increase the purity of leaves. The higher the increment in leaf purity, the higher the importance of the feature (Malato, 2022).

Recursive feature elimination: RFE selects the best subset of features for the supplied estimator by removing 0 to N features (where N is the number of features). It eliminates n features from a model by fitting the model multiple times and at each step, removing the weakest features, determined by feature importance attribute of the fitted model. The RFE will repeat this process till we get the number of the most important features we would like to use in the prediction model.

3.0 RESEARCH METHODOLOGY

Figure 3.4 shows the proposed conceptual framework for helpfulness prediction whereby consists of 4 significant steps: data preprocessing, feature selection, prediction model, and evaluation. These steps will be discussed in the following sections.

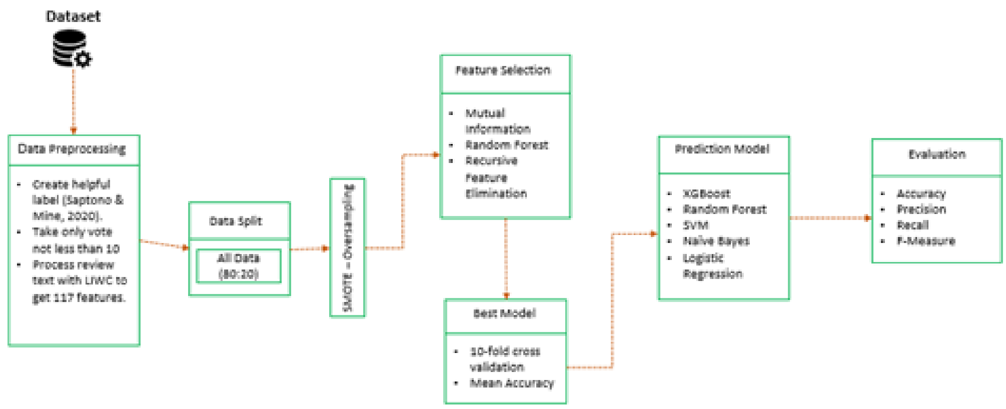

Figure 3.1: Proposed Conceptual Framework for Helpfulness Prediction

3.1 DATASET

The Amazon dataset used in this study is from https://nijianmo.github.io/amazon/index.html, spanning May 1996 to Oct 2018 (Ni et al., 2019). More specifically, this research selected the Video Games product category dataset, which contains 497,577 reviews and 12 features, as presented below:

1. reviewerID - ID of the reviewer, e.g., A2SUAM1J3GNN3B
2. asin - ID of the product, e.g., 0000013714
3. reviewerName – the name of the reviewer
4. verified - verified user "True" or "False"
5. vote - helpful votes of the review
6. style - a discretionary of the product metadata, e.g., "Format" is "Hardcover"
7. reviewText - text of the review
8. overall - rating of the product
9. summary - summary of the review
10. unixReviewTime - time of the review (unix time)
11. reviewTime - time of the review (raw)
12. image - images that users post after they have received the product

3.2 DATA PREPROCESSING

3.2.1 Helpful Definition and label

Due to the significant amount of reviews on Amazon, the website provided a mechanism for users to vote on reviews that users think are helpful. This information is captured in the Amazon review dataset, shown in figure 3.2, whereby vote: 10 means the review has received ten votes from users. This vote feature is used to define the helpful label for classification in this research with the following conditions (Saptono & Mine, 2020):

- helpful (1): Review with number of votes > the first quartile of the total vote in a product
- unhelpful (0): Review with number of votes <= the first quartile of the total vote in a product

reviewerID	A2XFRYFB5CVH2G
asin	788602144
reviewerName	Skip
verified	FALSE
reviewText	I initially bought Vol 3 on VHS, because I'd never seen the Supremes perform "Nothing But Heartaches," being that during Motown's greatest era, I was living in Germany, and didn't have much opportunity to catch any of their appearances shown stateside. Watching the entire clips brings back great memories. I miss the days of singers who occasionally danced...
overall	5
reviewTime	05 15, 2002
summary	Aaah...the 60's
unixReviewTime	1021420800
style	{'Format:': ' DVD'}
vote	10
image	

Figure 3.2: An example of the Amazon review

3.2.2 Data Cleaning

This research will execute the data cleaning on the dataset with the steps below to improve the data quality. First, all the reviews with less than ten votes are removed to ensure the robustness of the prediction (Akbarabadi & Hosseini, 2018). Then 23531 rows are left. Furthermore, the unwanted columns are dropped. Finally, the dataset contains 18185 rows with 1 (helpful) and 5346 rows with 0 (unhelpful) reviews. Then to conduct the model building stage, the dataset is split into 80 percent for model training and 20 percent for model testing.

3.2.3 SMOTE

As mentioned in section 3.2.2, the data set is imbalanced after the data cleaning step. Therefore, this research utilized SMOTE to create synthetic samples for the training dataset to resolve the data imbalance issue. This approach symmetrically oversamples minority-class data, producing the same occurrences for the positive and negative classes in the training data. After the SMOTE, we got the training dataset with 14568 rows of helpful reviews and 14568 rows of non-helpful reviews.

3.2.4 LIWC

LIWC (Linguistic Inquiry and Word Count) is a text analysis software developed by Pennebaker et al. to evaluate text samples' psychological and structural components (Park, 2018). Now fifth (LIWC-22) version updated the original application with increasingly expanded dictionaries and sophisticated software design (Boyd et al., 2022). This research used LIWC to analyze the review text and obtain different features for prediction. Figure 3.3 shows an example of the outcome of LIWC-22. Figure 3.4 shows all the variables after being processed by the LIWC. Columns 1 to Column 12 are the original variables from the Amazon data. Columns 13 until 130 are the features derived from LIWC.

filler	AllPunc	Period	Comma	QMark	Exclam	Apostro	OtherP
0.41	20.25	6.20	4.55	0.83	0.00	2.89	5.79

Figure 3.3: Sample outcome of LIWC-22

1	overall	35	prep	69	prosocial	103	fatigue
2	verified	36	auxverb	70	polite	104	reward
3	reviewTime	37	adverb	71	conflict	105	risk
4	reviewerID	38	conj	72	moral	106	curiosity
5	asin	39	negate	73	comm	107	allure
6	reviewerName	40	verb	74	socrefs	108	Perception
7	reviewText	41	adj	75	family	109	attention
8	summary	42	quantity	76	friend	110	motion
9	unixReviewTime	43	Drives	77	female	111	space
10	vote	44	affiliation	78	male	112	visual
11	style	45	achieve	79	Culture	113	auditory
12	image	46	power	80	politic	114	feeling
13	Segment	47	Cognition	81	ethnicity	115	time
14	WC	48	allnone	82	tech	116	focuspast
15	Analytic	49	cogproc	83	Lifestyle	117	focuspresent
16	Clout	50	insight	84	leisure	118	focusfuture
17	Authentic	51	cause	85	home	119	Conversation
18	Tone	52	discrep	86	work	120	netspeak
19	WPS	53	tentat	87	money	121	assent
20	BigWords	54	certitude	88	relig	122	nonflu
21	Dic	55	differ	89	Physical	123	filler
22	Linguistic	56	memory	90	health	124	AllPunc
23	function	57	Affect	91	illness	125	Period
24	pronoun	58	tone_pos	92	wellness	126	Comma
25	ppron	59	tone_neg	93	mental	127	QMark
26	i	60	emotion	94	substances	128	Exclam
27	we	61	emo_pos	95	sexual	129	Apostro
28	you	62	emo_neg	96	food	130	OtherP
29	shehe	63	emo_anx	97	death	131	quantile_1
30	they	64	emo_anger	98	need	132	helpful
31	ipron	65	emo_sad	99	want		
32	det	66	swear	100	acquire		
33	article	67	Social	101	lack		
34	number	68	socbehav	102	fulfill		

Figure 3.4: Attributes of the data set

3.3 FEATURE SELECTION

This research utilized three feature selection methods, mutual Information, Random Forest, and recursive feature elimination, to determine the best number of features to use in the prediction model. A top-N (from 10 features until 100 features, with a step of 10) experiment is conducted to collect the accuracy of each combination of the number of features, feature selection, and machine learning algorithms. The accuracy result of each experiment iteration is collected via 10-fold cross-validation using training data to have more generalized results and avoid overfitting for the final prediction model (will use only test data).

Table 3.1: Combination of Machine Learning Algorithm and Feature Selection Techniques for Top N Experiment

Machine Learning Algorithm	Feature Selection Techniques
XGBoost	Mutual Information
Random Forest	Mutual Information
SVM	Mutual Information
Naïve Bayes	Mutual Information
Logistic Regression	Mutual Information
SVM	Random Forest
Naïve Bayes	Random Forest
Logistic Regression	Random Forest
Random Forest	Recursive Feature Elimination
XGBoost	Recursive Feature Elimination

3.3.1 Best Model Based on Result of Feature Selection

Based on the top-N experiment result, the best model or combination of the number of features, feature selection techniques, and machine learning algorithm will be selected based on the following criteria.

1) The mean accuracy (cross-validation) of the model is larger than the 0.6 quantiles of the overall result of the model
2) The least number of features is used for prediction.

3.4 PREDICTION MODELING

This research will use XGBoost, Random Forest, SVM, Naive Bayes, Logistic Regression, and ten-fold cross-validation to make the final prediction. However, only the best model (by machine learning algorithm and feature selection methods) will undergo the final prediction. The best model definition is defined in section 3.3.1. To further enhance the accuracy, the top 3 models with the highest mean accuracy (from cross-validation) will be selected for further hyperparameter optimization, whereby a randomized search technique will be used with 10-fold cross-validation and 60 iterations to find the best performing (accuracy) parameter.

3.5 MODEL EVALUATION

This research aims to find the performance of the selected machine learning model. Therefore, four evaluation matrix has been selected for model performance comparison: accuracy, precision, recall, and F1-score.

4.0 RESULT AND DISCUSSION

4.1 TOP N FEATURE SELECTION

Figure 4.1 shows the accuracy of each model for the top 100 features. The x-axis is the number of features, and the y-axis is the model accuracy. The polylines with different colors represent different models. The three models with the highest accuracy are Random Forest RFE (around 85%), Random Forest MI (around 80%), and XGboost RFE (around 78%). Moreover, the model with the lowest accuracy is Naive Bayes MI, with an accuracy lower than 55%.

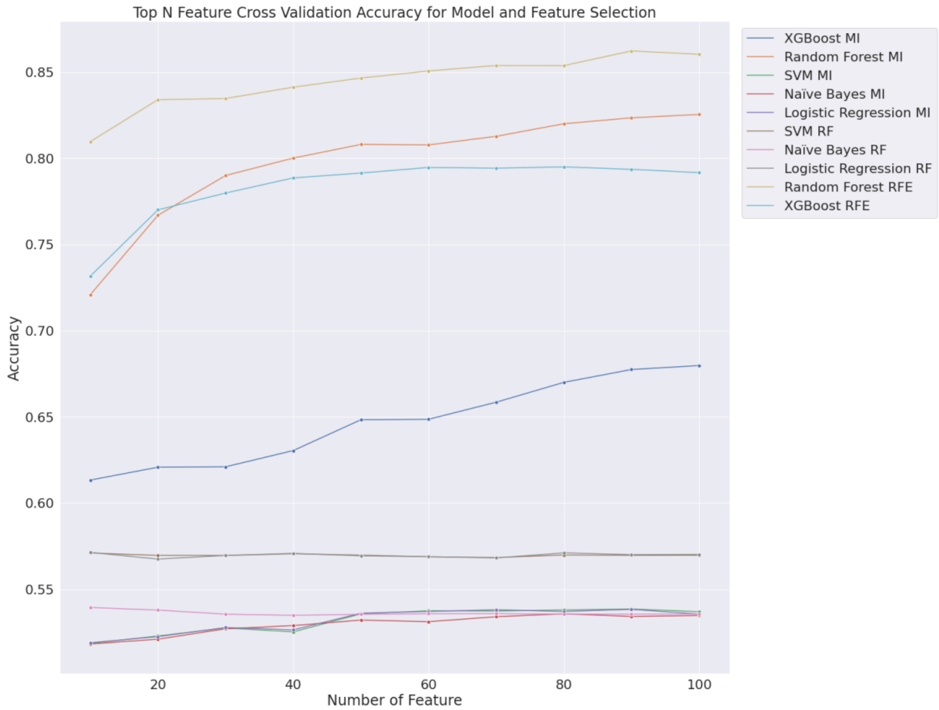

Figure 4.1: The N feature Cross-validation Accuracy for Model and Feature Selection

Table 4.1 illustrates the result of the top n features for each model. The best model is selected based on the definition in section 3.3.1. From the result, the maximum number of features needed by models to get an acceptable accuracy is 70. Some algorithms, like SVM, Naive Bayes, and Logistic regression, only need the top 10 features to get the average accuracy, but the accuracy is poor. It is clear that the Random Forest RFE model has the highest accuracy of 0.8538, followed by Random Forest MI with an accuracy of 0.8127, and XGBoost seems to be the third-best model with an accuracy of 0.7947.

Table 4.1: The best ten models selected from the Top N Experiment

Model	Num feature	Accuracy	Quantile
XGBoost MI	70	0.6585	0.6585
Random Forest MI	70	0.8127	0.8099
SVM MI	60	0.5376	0.5372
Naive Bayes MI	70	0.5341	0.5330
Logistic Regression MI	60	0.5372	0.5370
SVM RF	10	0.5711	0.5698
Naive Bayes RF	10	0.5395	0.5357
Logistic Regression RF	10	0.5714	0.5702
Random Forest RFE	70	0.8538	0.8302
XGBoost RFE	60	0.7947	0.7925

4.1.2 Best Models Prediction Result

The ten best models from the above experiment are then retrained with the training dataset (80% of total data) and tested on the test dataset (20% of total data) to check the model performance. Table

4.2 shows the number of features, model, and their respective result: accuracy, precision, recall, and f1. The Random Forest RFE, Random Forest MI, and XBboost RFE performed the best in general; the Random Forest RFE has the highest accuracy, 0.7226. Random Forest MI has the second-highest accuracy, 0.6956, while XGBoost RFE received the third-highest accuracy, 0.6690.

Table 4.2: The prediction results of 10 prediction models

Model	Num Feature	Accuracy	Precision	Recall	F1
XGBoost Mutual Information	70	0.5791	0.7462	0.6698	0.7059
Random Forest Mutual Information	70	0.6956	0.7696	0.851	0.8083
SVM Mutual Information	60	0.5449	0.7712	0.5638	0.6514
NB Mutual Information	70	0.524	0.7842	0.5088	0.6172
LG Mutual Information	60	0.5451	0.7722	0.5629	0.6511
SVM Random Forest	10	0.5481	0.7728	0.5676	0.6545
NB Ranfom Forest	10	0.4778	0.7864	0.4222	0.5494
LG Random Forest	10	0.5483	0.7737	0.5667	0.6543
Random Forest Rfe	70	**0.7226**	**0.7706**	**0.9**	**0.8303**
XGBoost Rfe	60	0.669	0.7609	0.8182	0.7885

4.1.3 COMPARISON BETWEEN THE HYPERPARAMETER TUNING

From the result of table 4.2, the top three models are chosen to conduct the hyperparameter tuning. Table 4.3 shows the difference before and after hyperparameter tuning. Random Forest RFE hyper still obtains the highest result in accuracy with 0.7341, followed by XBGBoost, Random Forest MI, with 0.7197, 0.7093. XGBoost had a more significant (0.05) accuracy increment compared to Random Forest.

Table 4.3: The result after the hyperparameter tuning

Model	Num Feature	Accuracy	Precision	Recall	F1
Random Forest RFE Hyper	70	**0.7341**	**0.7686**	**0.9262**	**0.8401**
Random Forest RFE	70	0.7226	0.7706	0.9	0.8303
XGBoost RFE Hyper	60	0.7197	0.7737	0.888	0.8269
XGBoost RFE	60	0.669	0.7609	0.8182	0.7885
Random Forest Mutual Information Hyper	70	0.7093	0.7676	0.8815	0.8206
Random Forest Mutual Information	70	0.6956	0.7696	0.851	0.8083

In summary, the best performing model is the combination of Random Forest and Recursive Feature Elimination with an accuracy of 0.7341. This finding also aligns with previous research such as Zhou & Yang (2019) and Akbarabadi & Hosseini (2018), whereby Random Forest has shown good performance. RFE is a greedy approach that removes the least important features in each selection iteration. However, this also leads to a more optimized feature selection than Mutual Information and random forest embedded feature selection. Based on the table 4.3, the random forest improved lesser than XGBoost. This is because of the great performance of the default random forest parameter (Fernandez-Delgado et al., 2014).

5.0 Conclusion

In this section, the three research objectives are relooked back. Furthermore, the limitation and future work are discussed as well.

5.1 Relook at Research Objective

5.1.1 To identify suitable machine learning techniques for predicting helpful reviews

This research used XGBoost, Random Forest, SVM, Naive Bayes, and Logistic Regression to predict the helpfulness of Amazon reviews. However, Random Forest is the most suitable algorithm because it has outperformed other techniques. Furthermore, RFE has outperformed other techniques; therefore, it is a suitable feature selection technique.

5.1.2 To find significant features in identifying helpful reviews

The Random Forest RFE with 70 features can lead to the determination of helpful reviews.Those 70 features are 'overall', 'WC', 'Analytic', 'Clout', 'Authentic', 'Tone', 'WPS', 'BigWords', 'Dic', 'Linguistic', 'pronoun', 'ppron', 'i', 'you', 'ipron', 'det', 'article', 'number', 'prep', 'auxverb', 'adverb', 'conj', 'negate', 'verb', 'adj', 'quantity', 'Drives', 'achieve', 'power', 'Cognition', 'allnone', 'insight', 'cause', 'discrep', 'tentat', 'certitude', 'differ', 'Affect', 'tone_pos', 'tone_neg', 'emotion', 'emo_pos', 'emo_neg', 'socbehav', 'Culture', 'tech', 'Lifestyle', 'leisure', 'money', 'Physical', 'need', 'want', 'acquire','curiosity', 'allure', 'Perception', 'space', 'visual', 'auditory','feeling', 'time', 'focuspast', 'focuspresent', 'Conversation', 'AllPunc', 'Period', 'Comma', 'Exclam', 'Apostro', 'OtherP'. These features have been chosen because they are the most significant and important in terms of prediction performance. Data collection can be focused on these critical traits, lowering the model's computational cost and improving its performance.

5.1.3. To conduct an assessment of machine learning models for prediction review helpfulness

As shown in table 4.4, among the three models, Random Forest-RFE with 70 features outperformed XGBoost, and Random Forest MI has the highest accuracy, precision, recall, and f1. Random XGBoost RFE is the second-best, while Random Forest MI is the third best.

Table 4.4: The best three prediction models results

Model	Num Feature	Accuracy	Precision	Recall	F1
Random Forest RFE Hyper	70	0.7341	0.7686	0.9262	0.8401
XGBoost RFE Hyper	60	0.7197	0.7737	0.888	0.8269
Random Forest Mutual Information Hyper	70	0.7093	0.7676	0.8815	0.8206

5.2 Limitations and Future Works

In this research, only the video game dataset is selected for prediction due to the limitation of hardware availability to do this project. Other product data have a significantly larger dataset and could not be processed. Furthermore, this research only investigates the feature provided by the LIWC analysis. The different product categories can be used for future research to have a more dynamic scenario, such as CDs, and vinyl, Books. Furthermore, the user behavior might differ due to the volume and the product category. Another aspect/feature can be considered as the psychological aspect of users for the helpfulness prediction.

6.0 REFERENCE

Akbarabadi, M., & Hosseini, M. (2018). Predicting the helpfulness of online customer reviews: The role of title features. *International Journal of Market Research*, *62*(3), 272–287. https://doi.org/10.1177/1470785318819979

Boyd, R. L., Ashwini, A., Sarah, S., & James W., P. (2022). The Development and Psychometric Properties of LIWC-22. Austin, TX: University of Texas at Austin. https://www.liwc.app

Brownlee, J. (2020, December 9). *Information Gain and Mutual Information for Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/information-gain-and-mutual-information/

Dey, D., & Kumar, P. (2019). A Novel Approach to Identify the Determinants of Online Review Helpfulness and Predict the Helpfulness Score Across Product Categories. *Big Data Analytics*, 365–388. https://doi.org/10.1007/978-3-030-37188-3_21

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Haque, M. E., Tozal, M. E., & Islam, A. (2018). Helpfulness Prediction of Online Product Reviews. *Proceedings of the ACM Symposium on Document Engineering 2018*. https://doi.org/10.1145/3209280.3229105

Lee, S., Lee, K. C., & Choeh, J. Y. (2020). Using Bayesian Network to Predict Online Review Helpfulness. *Sustainability*, *12*(17), 6997. https://doi.org/10.3390/su12176997

LIWC (Linguistic Inquiry and Word Count). (2019). 知乎专栏. https://zhuanlan.zhihu.com/p/72418905

Malato, G. (2022, January 5). *Feature selection with Random Forest - Towards Data Science*. Medium. https://towardsdatascience.com/feature-selection-with-random-forest-3e1979b3a2dc

Ni, J., Li, J., & McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. https://doi.org/10.18653/v1/d19-1018

Park, Y. J. (2018). Predicting the Helpfulness of Online Customer Reviews across Different Product Types. *Sustainability*, *10*(6), 1735. https://doi.org/10.3390/su10061735

Passon, M., Lippi, M., Serra, G., & Tasso, C. (2018). Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining. *Proceedings of the 5th Workshop on Argument Mining*. https://doi.org/10.18653/v1/w18-5205

Saptono, R., & Mine, T. (2020, December). Time-based Sampling Methods for Detecting Helpful Reviews. In 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (pp. 508-513). IEEE.

Weiran, S. (2021, December 11). *Hyper Parameter Tuning with Randomised Grid Search - Towards Data Science*. Medium. https://towardsdatascience.com/hyper-parameter-tuning-with-randomised-grid-search-54f865d27926

Zhou, Y., & Yang, S. (2019). Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews. *IEEE Access*, *7*, 27769–27780. https://doi.org/10.1109/access.2019.2901472