Sentiment Analysis: Predicting Product Reviews' Ratings using Online Customer Reviews

Ankit Taparia
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
tapariaa@tcd.ie

Tanmay Bagla
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
baglat@tcd.ie

ABSTRACT

In the present era, most of the data on the internet is in the form of raw text. These gold mines of data are invaluable since it contains lots of underlying information which can be extracted using natural language processing or text analytics techniques. The data from these text-based documents disclose users' sentiments and opinions about a particular subject. In this paper, customer reviews from Amazon.com are pre-processed, analyzed using our proposed framework, and how these textual reviews justify the star ratings is studied. Features derived from textual reviews are used to predict its corresponding star ratings. To accomplish it, the prediction problem is transformed into a multi-class classification task to classify reviews to one of the five classes corresponding to its star rating. The performances of various classifiers used are evaluated and compared. Evaluation results on ground-truth dataset show that Logistic Regression Classifier outperformed other models. Our study also reveals that among the various factors, polarity of the review and length of the review showed a significant effect on its rating.

KEYWORDS

sentiment analysis, bag of words, word cloud, polarity estimation, n-grams, tf-idf, multi-class classifiers, multi-class classifier roc curves

1 INTRODUCTION

Internet is the best source nowadays for any organisation to know public opinions about their products and services. Many consumers form an opinion about a product just by reading a few reviews. Everyday many reviews get generated and it is difficult to handle such a huge chunk of data and analyse it. However, it is critical that these textual reviews are analysed for understanding the sentiments from them. Many researchers have studied the impact of the online reviews on the sales and concluded that positive reviews help to uplift the profit of an organization. The existing customer reviews strongly affect the purchasing decision of new potential customers in absence of true feel of the product to be purchased.

Sentiment analysis of product reviews also helps to enhance product features, improve customer service by understanding their expectations, improve marketing strategies, etc. Sentiment analysis particularly is helpful when it comes to negative reviews. It helps discover the exact shortcomings of the products. Generally, the ratings and the price of the product are simple heuristics used by the customers to decide over the final purchase of the product. But often, the overall star ratings of the product reviews may not capture the exact polarity of the sentiments. This makes rating prediction a hard problem as customers may assign different ratings for a particular review.

Thus sentiments (in-text property) and ratings (out-of-text property) play somewhat independent roles. But this leads us to an important research question: Do sentiments and ratings affect each other in someway? From the data collected from Amazon.com, we extracted the sentiments from the set of product reviews and build predictive models to evaluate the above research question by predicting reviews' ratings from sentiments of the textual reviews.

In this research study, we extract various textual and numerical features from customer reviews, evaluate how they correlate, and try using them to predict the reviews' star ratings for the products. The rest of the paper is organised as follows: Section 2 reviews the related work in the domain. In Section 3, the proposed research framework has been described. Section 4 discusses the evaluation and results. Finally, Section 5 gives the concluding remarks.

2 LITERATURE REVIEW

In recent times, several opinion mining and sentiment analysis studies of reviews have been conducted.

In [3], SVM and Naïve Bayes classifiers are trained to classify the movie ratings as either "high" or "low" based on its reviews. Various linguistic features are extracted from the textual reviews and feature selection is performed using TF-IDF and information gain. The results found that the SVM classifier modeled using features selected based on information gain was most accurate. However, the model lacks granularity as it cannot distinguish between "bad" (with 2 star rating) and "worst" (with 1 star rating) reviews.

On the other hand, [7], presented a model which predicts the star rating of a review using sentiment dictionaries. Like most polarity-determining approaches, the paper uses the unigram model to represent text. The unigram model often fails to capture phrase patterns properly which leads to polarity incoherence. To overcome this drawback, the authors also employ a n-gram model. But this way of representing text vectors leads to the creation of large sparse matrices that are space inefficient known as n-gram sparsity bottlenecks.

To overcome the problem of polarity incoherence, [8] introduces the Bag of opinions model. This model overcomes the limitations of the unigram and n-gram models. It has 3 components: root word, modifiers and negation words. Each opinion from the corpora of reviews is assigned a score using ridge regression method. At the end, a final score is calculated by combining all the independent scores of all the opinions.

In another study [9], neural networks have been used for sentiment classification. It takes into account the product features, given users'

information in addition to the textual reviews. The authors claim that a review cannot be generalised for all users as people might have varying perceptions and opinions about it. The output states that the implemented model is better than the other classification methods such as SVM, HAN especially in the case where enough data is not available.

A number of studies have also been done to demonstrate how sentiment analysis can be leveraged in the tourism and hospitality domain. For example, [1] found that there exists a correlation between the reviews and ratings provided by customers for a hotel. The authors analysed the hotels reviews and used Naïve Bayes classifier against a lexicon of words to determine the polarity of the reviews. A linear regression model was developed to predict the ratings based on reviews and other hotel features such as price, location, etc.

3 PROPOSED FRAMEWORK

We used the below framework for our research:

3.1 Dataset Collection and its Features

We used a 5-core Amazon review dataset provided by [5, 6]. The chosen dataset contains product reviews of Cell phones and Accessories purchased from Amazon.com. It includes 1,128,437 rows and 11 features as explained below. Each row corresponds to a customer review, and includes the feature variables:

- reviewerID - ID of the reviewer
- asin - ID of the product
- reviewerName - name of the reviewer
- vote - helpful votes of the review
- style - a dictionary of the product metadata
- reviewText – customer review text
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)
- image - images that users post after they have received the product

3.2 Preliminary Feature Selection

Since our research is focused towards studying the sentiments from customer reviews and how it corroborates to the ratings; relevant features are selected for the analysis. Features – "reviewText", "overall" and "summary" are considered.

3.3 Data Pre-processing

Data is pre-processed (Figure 1) using the following techniques:

- Convert the text to lowercase
- Expand the contractions so we do not miss out any relevant sentiments such as haven't etc.
- Remove the punctuations, digits and special characters
- Tokenize the text, filter out the adjectives used in the review and create a new column in data frame
- Use Negation handling to append preceding and successive negation clauses, before removing the stop words. For example – convert "not worth" to "not_worth"
- Remove the stop words

	overall	reviewText	summary	cleanReviewLength	cleanReview	adjectives
0	5.0	Looks even better in person. Be careful to not drop your phone so often because the rhinestones will fall off (duh). More of a decorative case than it is protective, but I will say that it fits perfectly and securely on my phone. Overall, very pleased with this purchase.	Can't stop won't stop looking at it	23	looks even better person careful not drop phone often rhinestones fall duh decorative case protective say fits perfectly securely phone overall pleased purchase	careful more decorative protective overall pleased
1	5.0	When you don't want to spend a whole lot of cash but want a great deal.. this is the shop to buy from!	1	11	not want spend whole lot cash want great deal shop buy	whole great
2	3.0	so the case came on time, i love the design. I'm actually missing 2 studs but nothing too noticeable the studding is almost a bit sloppy around the bow, but once again not too noticeable. I haven't put in my phone yet so this is just what I've notice so far	Its okay	25	case came time love design actually missing 2 studs nothing noticeable studding almost bit sloppy around bow not noticeable not put phone yet notice far	noticeable sloppy noticeable
3	2.0	DON'T CARE FOR IT. GAVE IT AS A GIFT AND THEY WERE OKAY WITH IT. JUST NOT WHAT I EXPECTED.	CASE	7	not care gave gift okay not expected	
4	4.0	I liked it because it was cute, but the studs fall off easily and to protect a phone this would not be recommended. Buy if you just like it for looks.	Cute!	13	liked cute studs fall easily protect phone would not recommended buy like looks	cute studs

Figure 1: Data after pre-processing

3.4 Visualizing Sentiment Analysis using Word Cloud

In order to study the sentiments and nature of words used in the review texts, the reviews are divided into 3 subgroups – positive, neutral and negative based on their corresponding given ratings. Reviews with overall rating greater than 3, less than 3 and equal to 3 are labelled as positive, negative and neutral respectively. Word Cloud (Figure 2 and Figure 3) is used to visualise the most frequent words used in positive and negative reviews. It is observed that negative reviews contain more negative words (sentiments) than the positive and neutral reviews.

Figure 2: Word Cloud for positive customer reviews

Figure 3: Word Cloud for negative customer reviews

3.5 Polarity Detection and Feature Extraction

After completing data pre-processing, cleaned review texts, summary, and adjectives extracted from reviews are used to estimate the polarity of the reviews. A Python library - TextBlob [4] is used to

predict the polarities of the sentiments. It assigns each review a score between 0 and 1. Values closer to 1 represents positive sentiments whereas values close to 0 indicate presence of negative sentiments in the review. Figure 4 represents the textual reviews and their corresponding polarities determined using TextBlob.

	overall	reviewText	summary	cleanReviewLength	cleanReview	adjectives	polarity
1	5.0	When you don't want to spend a whole lot of cash but want a great deal...this is the shop to buy from!	1	11	not want spend whole lot cash want great deal shop buy	whole great	0.600000
2	3.0	so the case came on time. i love the design. I'm actually missing 2 studs but nothing too noticeable the studding is almost a bit sloppy around the bow, but once again not too noticeable. I haven't put in my phone yet so this is just what I've notice so far	Its okay	25	case came time love design actually missing 2 studs nothing noticeable studding almost bit sloppy around bow not noticeable not put phone yet notice far	noticeable sloppy noticeable	-0.004167
3	2.0	DON'T CARE FOR IT. GAVE IT AS A GIFT AND THEY WERE OKAY WITH IT. JUST NOT WHAT I EXPECTED.	CASE	7	not care gave gift okay not expected		0.200000
4	4.0	I liked it because it was cute, but the studs fall off easily and to protect a phone this would not be recommended. Buy if you just like it for looks.	Cute!	13	liked cute studs fall easily protect phone would not recommended buy like looks	cute studs	0.511111
5	2.0	The product looked exactly like the picture and it was very nice. However only days later it fell apart. I'm very disappointed with the quality of the product.	Not so happy	14	product looked exactly like picture nice however days later fell apart disappointed quality product	nice disappointed	0.011000

Figure 4: Polarity detection of sentiments in reviews

To transform the obtained cleaned reviews text data into numerical data to make it machine readable, Term Frequency–Inverse Document (TF-IDF) [2] vectorization is used. TF-IDF quantifies each word present in reviews and assigns weight to it which denotes the importance of the word in review and whole corpus. If a word appears in almost all the reviews, it is deemed as less significant and is given less weightage. On the other hand, if a word occurs only in many reviews, then it is assumed to be more significant and thus TF-IDF assigns greater weight to it. To avoid losing about important information by tokenizing each individual word in the reviews, n-grams (n=[1,3]) technique is used along with TF-IDF vectorization. It retains the contextual meaning and captures multi-word expressions occurring in the text that is ignored by Bag-of-Words approach.

After extracting features, correlation between ratings and some of the derived features such as polarity, clean review length are analysed before building the prediction models.

Table 1: Correlation between Ratings and Selected Derived Features

	polarity	cleanReviewLength
overall_rating	0.571315	-0.116626

It can be clearly observed from Table 1, reviews with high polarity values (having positive sentiments) tends to have higher ratings. And also as the length of review increases, ratings tend to decrease.

4 EXPERIMENTAL RESULTS AND EVALUATION

The experiment is conducted on 200,000 customer reviews randomly sampled from the dataset. To overcome the class imbalance problem, 40,000 reviews from each of the five ratings classes are considered. The reviews then are analysed and pre-processed using our proposed framework. 80% of the reviews are used for training and rest 20% are used for testing the models.

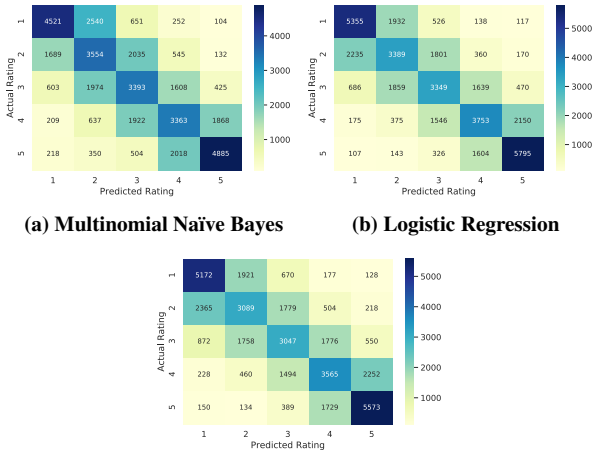
Thereafter, for ordinal values of overall ratings ranging from 1 to 5, five discrete categorical classes are created to treat the rating prediction as a multi-class classification problem. Three multi-class classifier models used are Multinomial Naïve Bayes Classifier, Logistic Regression Classifier and Linear SVC (Support Vector Classifier). These models are trained and utilized to classify reviews to one of the 5 classes using TF-IDF vector features created from review text; and other derived features such as polarity, length of the review as mentioned in Section 3.5. The results obtained and the metrics used for evaluation are described below.

Table 2: Classifier and its accuracy

	Multinomial Naïve Bayes	Logistic Regression	Linear SVC
Accuracy	49.2 %	54.1 %	51.1 %

4.1 Confusion Matrix and Accuracy

Confusion matrix is an important metric used to measure classifier efficiency. It represents the number of samples correctly classified and is used to determine precision, recall and F1 score. Confusion matrices are shown in Figure 5. It is clearly observable that all the classifiers predicts 1 star and 5 star reviews with much higher accuracy than the neutral reviews (3 star).



(a) Multinomial Naïve Bayes **(b) Logistic Regression**

(c) Linear SVC

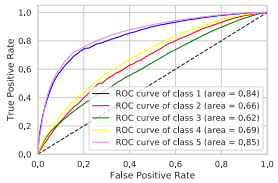
Figure 5: Confusion matrices of various classifiers

Accuracy of the models are recorded in the Table 2. As we have five rating classes, any random guessing would be correct only 20% of the time which is our baseline. It can be observed from Table 2 that all the classifiers outperforms the baseline comfortably and Logistic Regression classifier performs best with an overall accuracy of 54.1% with an improvement of around 170% over the random classifier.

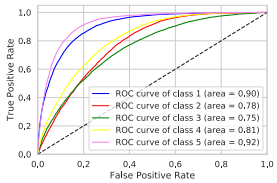
4.2 Area under Receiver Operating Characteristics Curve (ROC) Curve

Receiver Operating Characteristics Curve (ROC) is a probability curve and area under curve (AUC) indicates measure of separability. This notes how well the model can discriminate between classes. ROC curve is an important metric than accuracy for multi class

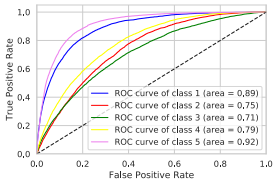
classification models as it visualizes model's accuracy across all possible thresholds. Area under ROC with value 0.5 is considered to be baseline which represents a random classification model.
It can be observed from Figure 6 that AUCs are well above the baseline and it is particularly higher for ROC curves representing 1 star and 5 star ratings. It suggests these two classes are much more separable and are easily classified whereas class 3 representing neutral reviews with 3 star ratings is least separable.



(a) Multinomial Naïve Bayes



(b) Logistic Regression



(c) Linear SVC

Figure 6: AUC - ROC curves of various classifiers

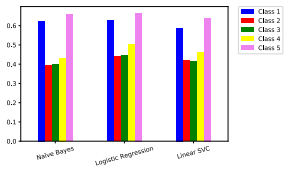
4.3 Precision, Recall and F1-Score

i Precision: Precision indicates what proportion of predicted positive reviews is truly positive.
ii Recall: Recall represents what proportion of actual positive reviews are correctly classified by the classifier.
iii F1-Score: Comparing two models with low accuracy and high recall is complicated, or vice versa. Therefore, the F1-Score is used to make them analogous. F1-score is known as a harmonic mean of Recall and Precision.

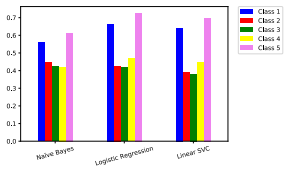
Evaluation metrics of precision, recall and f1-score of classifiers implemented are shown in Figure 7.

5 CONCLUSION

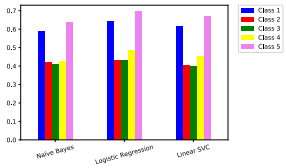
From the experimental results obtained, we find that classifiers models are able to predict ratings of the reviews using various features derived from its textual content with a decent accuracy. This suggests a strong correlation between the customer reviews (in-text property) and ratings (out-of text property) we considered for our analysis. Hence, it answers our research question: Sentiments of the text reviews do affect its corresponding ratings. It has been observed that among the features used for classification, polarity of the review and length of the review are more influential and highly correlated with its rating. Moreover, it can also be observed and concluded that it is relatively easy to predict 1 star and 5 star ratings owing to presence of strong sentiments in them than in neutral reviews with 3 star ratings.



(a) Precision



(b) Recall



(c) F1 Score

Figure 7: Precision, Recall and F1 Score of various classifiers

As part of future work, review date can also be incorporated to give higher weightage to recent reviews than older reviews to predict its rating. Moreover, the pre-processed dataset obtained from the proposed framework can be used along with the sales data of the products to study the impact of sales of the products due to its reviews, ratings provided by the customers and vice-versa. Finally, it would be also interesting to leverage our proposed framework over customer reviews from other e-commerce websites.

REFERENCES

[1] M Geetha, Pratap Singha, and Sumedha Sinha. 2017. Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. *Tourism Management* 61 (2017), 43–54.
[2] Djoerd Hiemstra. 2000. A probabilistic justification for using tf× idf term weighting in information retrieval. *International Journal on Digital Libraries* 3, 2 (2000), 131–139.
[3] Mohammadamir Kavousi and Sepehr Saadatmand. 2019. Estimating the Rating of the Reviews Based on the Text. In *Data Analytics and Learning*. Springer, 257–267.
[4] Steven Loria. 2018. textblob Documentation. *Release 0.15* 2 (2018).
[5] Jianmo Ni. 2018. Amazon Review Data (2018): https://nijianmo.github.io/amazon/index.html. https://nijianmo.github.io/amazon/index.html
[6] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 188–197.
[7] A NithyaKalyani, S Ushasukhanya, TYJ Nagamalleswari, and S Girija. 2018. Rating prediction using textual reviews. In *Journal of Physics: Conference Series*, Vol. 1000. IOP Publishing, 012044.
[8] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 913–921.
[9] Zhigang Yuan, Fangzhao Wu, Junxin Liu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2019. Neural Review Rating Prediction with User and Product Memory. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2341–2344.