

Author Declaration for Group 14

Assignment Number: Final Term Essay
Module Number: CS7IS4- Text Analytics
Title of Assignment: Sentiment Analysis: Predicting Overall Ratings of Products using Customer Reviews
Word Count: 2470





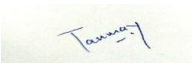
Student Number	Student Name	Nature of Contribution	Percentage contribution
19302722	Ankit Taparia	Verifier - Contributed to dataset collection, pre-processing and analysis. Contributed to feature extraction, polarity detection and experimental setup. Implemented Naive Bayes, Logistic Regression classifiers to predict star ratings. Contributed to report creation in Latex by adding experimental results and evaluation.	20
19300875	Ashish Kannur	Accountant - Contributed to writing introduction. Contributed to sentiment polarity detection. Contributed in finding correlation between reviews and ratings. Contributed in writing experiment results and conclusion.	20
19302709	Hemlata Sharma	Recorder - Read papers on related work and helped draft the Literature review. Performed analysis of peer reviews. Contributed to the report creation.	20
19300733	Himanshu Gupta	Chair - Contributed to literature review and abstract. Contributed to the polarity detection of the reviews. Contributed to Evaluation, experiment results and Conclusion writing.	20
19300702	Tanmay Bagla	Ambassador - Contributed to feature extraction from textual reviews. Contributed to the experimental setup. Performed analysis of peer reviews. Contributed to the report creation in Latex.	20

We have read and we understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: <http://www.tcd.ie/calendar>. We have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at <http://tcd-ie.libguides.com/plagiarism/ready-steady-write>.

We declare that this assignment, together with any supporting artefact is offered for assessment as our original and unaided work, except insofar as any advice and/or assistance from any other named person in preparing it and any reference material used are duly and appropriately acknowledged.

We declare that the percentage contribution by each member as stated above has been agreed by all members of the group and reflects the actual contribution of the group members.

Signed and dated: 13-04-2020

	Ankit Taparia
	Himanshu Gupta
	Ashish Kannur
	Hemlata Sharma
	Tanmay Bagla

Sentiment Analysis: Predicting Overall Ratings of Products using Customer Reviews

Ankit Taparia
Trinity College Dublin
tapariaa@tcd.ie

Ashish Kannur
Trinity College Dublin
kannura@tcd.ie

Hemlata Sharma
Trinity College Dublin
sharmah@tcd.ie

Himanshu Gupta
Trinity College Dublin
guptah@tcd.ie

Tanmay Bagla
Trinity College Dublin
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. Across all businesses, these data are being used to find insights that help reduce operational costs and predict various future outcomes and trends. 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, analysed, and how these textual reviews justify the star ratings is studied. Features derived from textual reviews are used to predict star ratings of the reviews. To accomplish it, the prediction problem is transformed to a classification task to classify reviews to one of the five classes corresponding to its star rating. The features which affect the ratings the most are also discussed.

Keywords: sentiment analysis, bag of words, word cloud, polarity estimation, tf-idf, multinomial naïve bayes, logistic regression, multi-class classifier roc curves

1 Introduction

The Internet is the best source nowadays for any company to know public opinions about their products. 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. Often, people use even positive words to express sarcasm. Now, this is easy for humans to interpret but is difficult for machines to do the same.

Sentiment analysis of product reviews helps to improve/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 overall ratings of the products do not capture the exact polarity of the sentiments. Just the ratings and the price of the product are simple heuristics used by the users to decide over the final purchase of the product. But it would be more effective that a user understands the sentiments for other reviews before deciding on the purchase.

Thus ratings and sentiments play independent roles. This leads us to some important research question: How do sentiments and ratings affect each other? From the data collected from Amazon.com, we

would extract the sentiments from the set of product reviews and build a model to evaluate the above research question. Firstly, the association between scores and sentiments will be determined dependent on increasing one's likelihood values. Then the impact of different elements of sentiments, like strong sentiments and weak sentiments on ratings would be evaluated.

When trying to make a final decision about the purchase of a product, user tries to simplify the decision-making by adopting less demanding processes like choosing simple attributes as ratings and prices initially and then move to the more demanding process of reading the reviews about the product, derive sentiments from it and finally make a decision. Thus even though the sentiment analysis requires more time and effort, it may have a greater impact on the final decision as compared to ratings. Thus in this research study, we extract various features from customer reviews and evaluate how they correlate and can be used to predict overall ratings of the products. The rest of the paper is organised as follows: [Section 2](#) reviews the related work in the domain. We defined our hypothesis in [Section 3](#). In [Section 4](#), the proposed research framework has been described. [Section 5](#) discusses the evaluation and results. Finally, [Section 6](#) gives the concluding remarks.

2 Literature Review

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

In [\(Kavousi and Saadatmand, 2019\)](#), SVM and Naive 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, [\(NithyaKalyani et al., 2018\)](#), presented a model which predicts the star rating of a review. Sentiment polarity for an individual review is can be obtained from the difference of total positive words and total negative words. A positive resultant value signifies a positive polarity, a negative value signifies a negative polarity and zero signifies a neutral polarity. A Sentiment Degree dictionary is also used to calculate polarities. The words in the dictionary are separated into different levels, L1 through L5. Words with the highest sentiment degree fall into the L1 category while words with the lowest sentiment degree fall into the L5 category: ‘absolute’ falls into L1 and ‘bit’ falls into L5. A Naive based classifier also predicts the polarity. The output of these three methods is combined mathematically into a formula to determine the star rating corresponding to a review. 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, [\(Qu et al., 2010\)](#) 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 copora 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. Test results showed that the Bag of Opinions model outperformed traditional techniques used for rating prediction.

In another study [\(Yuan et al., 2019\)](#), 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, [\(Geetha et al., 2017\)](#) 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 Naive 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 Hypothesis

Generally, it is being observed that highly rated reviews tend to have more positive sentiments than low rated ones. In other words, more negative is the sentiment of the review, lesser will be its rating. Thus, our hypothesis is: “Sentiment of the reviews decides the ratings.”

4 Proposed Framework

We used the below framework for our research:

4.1 Dataset Collection and its Features

We used a 5-core Amazon review dataset provided by (Ni, 2018). 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

4.2 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.

We used the Bag of Words approach to analyse the reviews. For this approach, data is pre-processed using the following techniques:

- | 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 | | DONT 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 |

4.4 Visualizing Sentiment Analysis using Word Cloud

screen_protector_not
fit_phone_perfectly
not_go_wrong
highly_recommend
definitely_recommend
would_definitely_recommend
would_recommend_anyone
typehidden_name_valuehttpsimagesna
not_beat_price
last_long_time
love_love_time
not_add_much
protects_phone_well
name_valuehttpsimagesna
would_recommend_case
north_every_gunny
fit_iphone_perfectly
theapple_datahubprodclickidmefassilinnomel
would_definitely_recommend
typehidden_name_valuehttpsimagesna
samsung_galaxy_note
typehidden_name_valuehttpsimagesna
not_big_deal
work_like_chang
great_product_great
really_like_case
fit_like_glove
highly_recommend_product
thing_not_like
one_screen_protector
case_fit_perfectly
case_fit_well
great_product_great
highly_recommend_case
get_job_done
would_recommend_product
apple_creditwell_apple_creditwell_apple_creditwell
case_fit_phone_fit_phone_well
samsung_galaxy_note
typehidden_name_valuehttpsimagesna
typehidden_name_valuehttpsimagesna

Figure 2: Word Cloud for positive customer reviews



4.5 Polarity Detection and Feature Extraction

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	DONT 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.	Cutel		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

To transform the obtained cleaned reviews text data into numerical data to make it machine readable, Term Frequency–Inverse Document (TF-IDF) (Hiemstra, 2000) 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 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.

Table 1: Correlation between Ratings and Selected Derived Features

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 tends to decrease.

5 Experimental Results and Evaluation

The experiment is conducted on 100,000 customer reviews randomly sampled from the dataset. The reviews are analysed and pre-processed using our proposed framework. 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 Naïve Bayes Classifier, Logistic Regression Classifier and Linear SVM 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 4.5](#). The results obtained and the metrics used for evaluation are described below.

5.1 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).

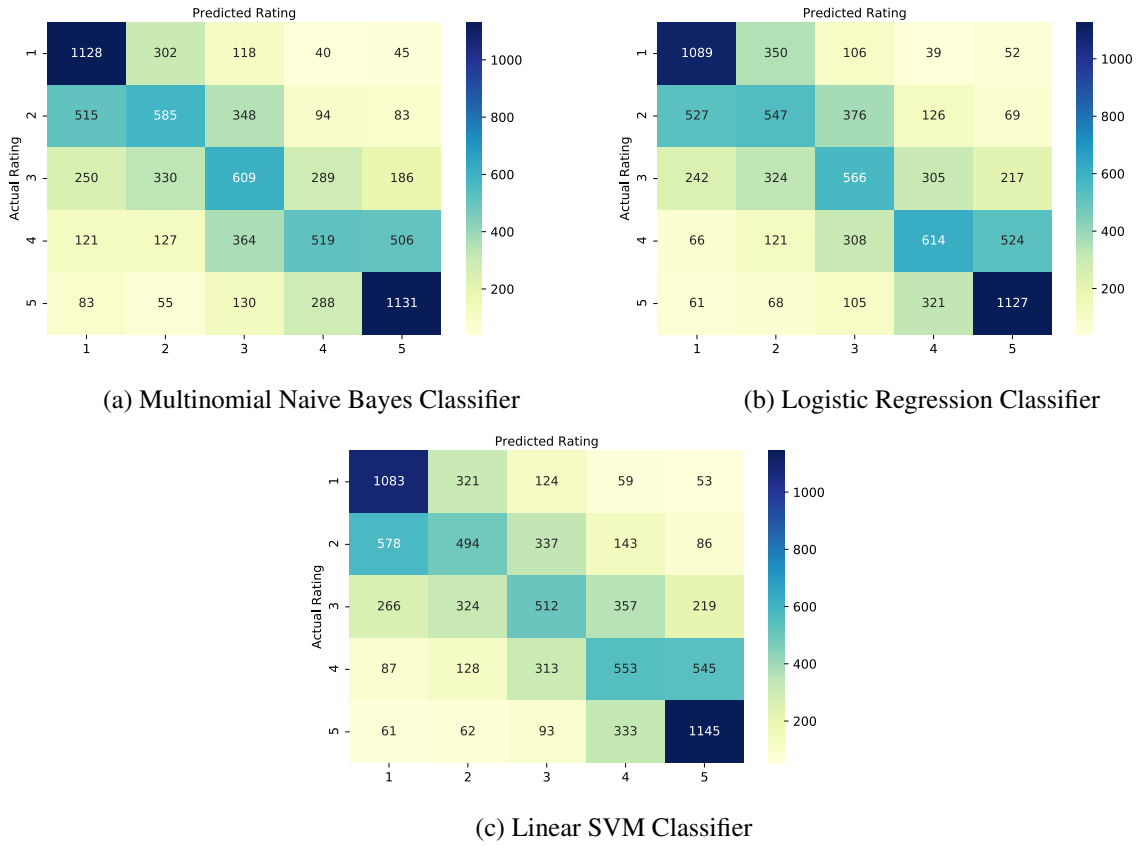


Figure 5: Confusion matrices of various classifiers

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 51.4%.

Table 2: Classifier and its accuracy

	Naive Bayes	Logistic Regression	Linear SVM
Accuracy	48.1 %	51.4 %	46.03 %

5.2 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 considered a better metric than accuracy for multi class classification models as ROC curve visualizes all possible thresholds whereas accuracy measures model's accuracy on a single threshold. 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 AUC is 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.

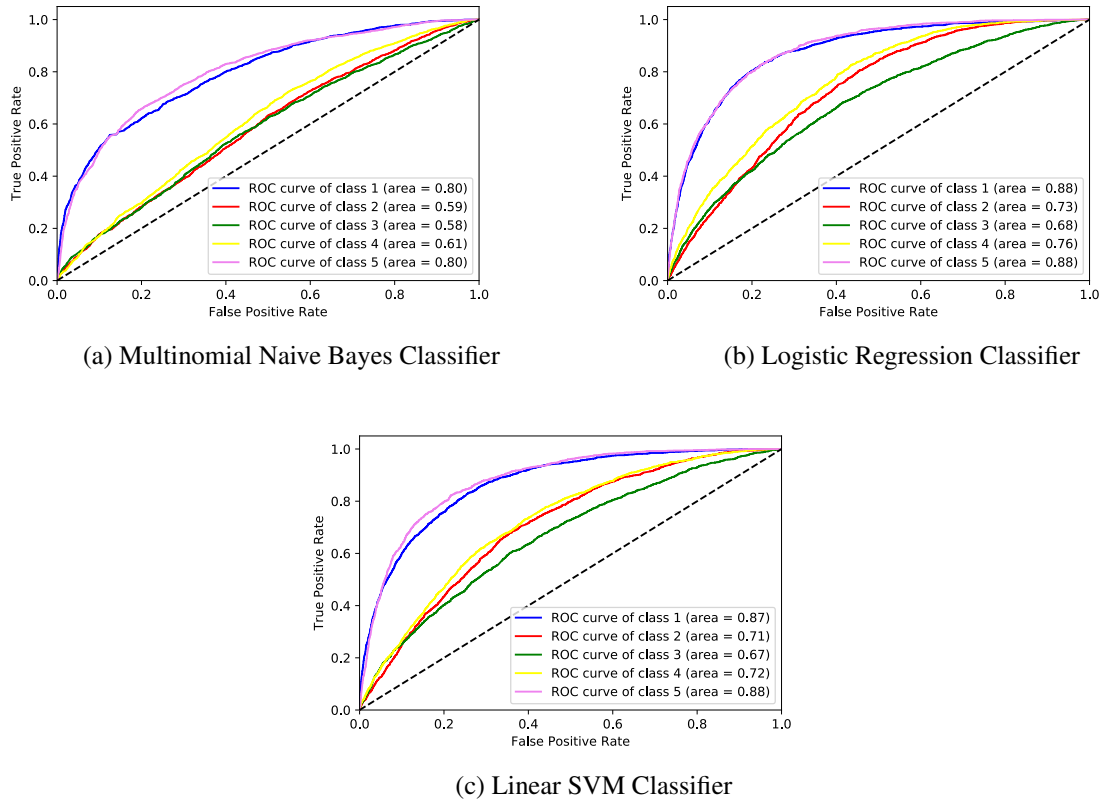


Figure 6: AUC - ROC curves of various classifiers

5.3 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**.

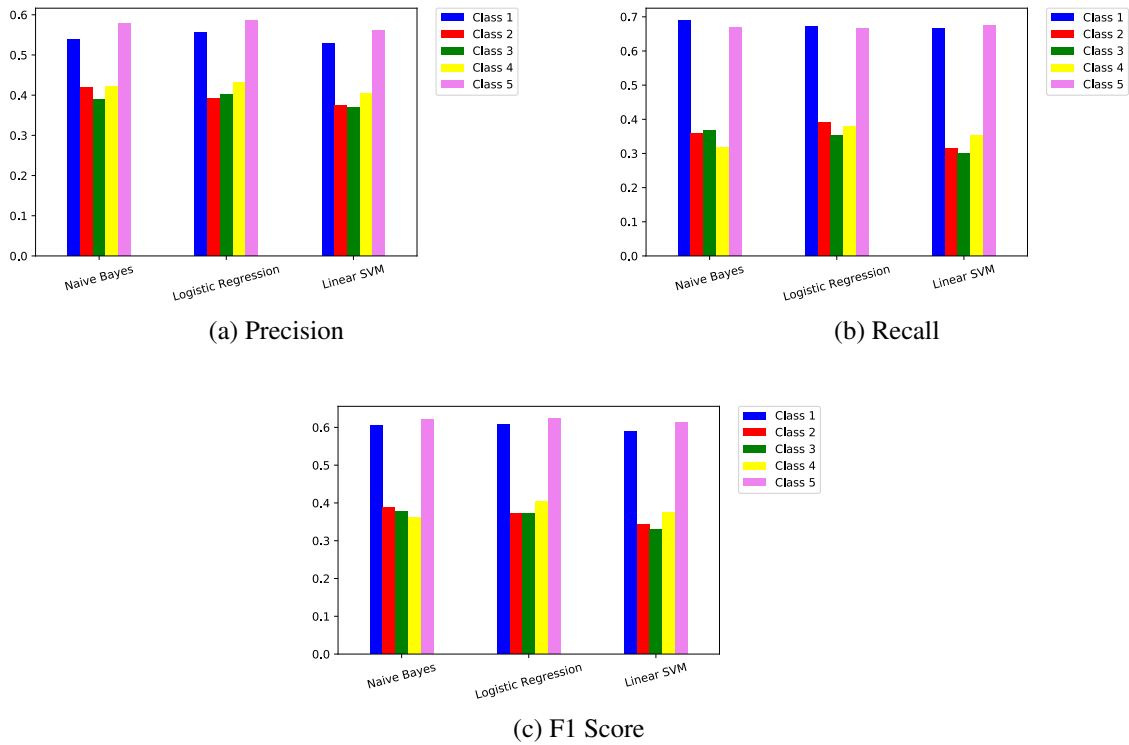


Figure 7: Precision, Recall and F1 Score of various classifiers

6 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 relation between the in-text property (customer reviews) and out-of text property (ratings) we considered for our analysis. Hence, it justifies our hypothesis: Sentiments of the text reviews decide 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 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.

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.

References

- Geetha, M., P. Singha, and S. Sinha (2017). Relationship between customer sentiment and online customer ratings for hotels-an empirical analysis. *Tourism Management* 61, 43–54.
- Hiemstra, D. (2000). A probabilistic justification for using $\text{tf} \times \text{idf}$ term weighting in information retrieval. *International Journal on Digital Libraries* 3(2), 131–139.
- Kavousi, M. and S. Saadatmand (2019). Estimating the rating of the reviews based on the text. In *Data Analytics and Learning*, pp. 257–267. Springer.
- Ni, J. (2018). Amazon review data (2018): <https://nijianmo.github.io/amazon/index.html>.
- NithyaKalyani, A., S. Ushasukhanya, T. Nagamalleswari, and S. Girija (2018). Rating prediction using textual reviews. In *Journal of Physics: Conference Series*, Volume 1000, pp. 012044. IOP Publishing.
- Qu, L., G. Ifrim, and G. 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*, pp. 913–921. Association for Computational Linguistics.
- Yuan, Z., F. Wu, J. Liu, C. Wu, Y. Huang, and X. Xie (2019). Neural review rating prediction with user and product memory. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2341–2344.