Amazon's Product Review Analysis

Mrunal Pagnis
School of Informatics and Computing
Indiana University, Bloomington
mmpagnis@indiana.edu

Suraj Chandrakant Songire
School of Informatics and Computing
Indiana University, Bloomington
ssongire@umail.iu.edu

Pei-Ying Chen
School of Informatics and Computing
Indiana University, Bloomington
peiychen@iu.edu

Abstract

Product reviews are important sources for customers to evaluate the quality of a product they are thinking about purchasing on-line. In this study, we propose a method to classify Amazon's product reviews, and compare the results with customer's ratings as a validation test.

1 Introduction

According to Pew Internet & American Life Project's 2008 report on on-line shopping, 66% of American Internet users have the experience of purchasing a product on-line, showing that shopping on-line has become prevalent among on-line Americans mainly for its convenience and time-saving concerns. While 81% of Internet users have used the Internet to search for product-related information, 58% have encountered difficulties either as frustrated by the lack of information, confused by the information, or overwhelmed by the amount of information in the course of on-line shopping. At the same time, 30% and 32% of on-line users have respectively posted a comment/review about and rated the products they bought or services they received on-line (Horrigan, 2008), indicating that on-line product reviews may meet the information need of potential on-line customers by providing more personalized accounts of the products they are considering buying. To facilitate consumers' decision-making in on-line shopping, this study employs Support Vector Machines (SVM) and Naive Bayes to classify product reviews on Amazon, and compares the results with customer's ratings as a validation test.

2 Related Work

Pang and Lee (2008) provide the first comprehensive survey of the field of opinion mining and sentiment analysis with in-depth discussions of key concepts related to features as well as different approaches to classification. In an earlier paper, the same authors, with their collaborators, apply machine learning techniques on movie reviews to determine whether a review is positive or negative. They find that while the machine learning methods clearly outperform the human-produced baselines, with Naive Bayes tends to do the worst and SVM the best, these methods does not yield accuracies on sentiment classification comparable to traditional topic-based categorization (Pang et al., 2002).

In contrast, Turney (2002) proposes a simple unsupervised learning algorithm to classify reviews as *recommended* or *not recommended* based on the average semantic orientation of the phrases extracted from the review. In other words, while Turney's method relies on statistics gathered by a search engine, Pang *et al.* utilize completely prior-knowledge-free supervised machine learning methods. Other works focused on classify and summarize reviews include Hu and Liu's (2004) study of generating feature-based summaries of customer reviews of products online.

With the accessibility of larger datasets on the Internet, the review data are exploited to build ranking systems and recommender systems. For instance, Huang and Lee (2013) design a ranking system to summarize the pros and cons of a product by calculating product scores in terms of product reviews, product popularity, and product release month. More recently, Julian McAuley at UCSD has released the Amazon Product Data, and used the data to build a recommender system titled *Sceptre* (Substitute and Complementary Edges between Products from Topics in Reviews) by identifying topics in the product review texts through the combination of topic modeling and supervised link prediction (McAuley et al., 2015a). A visual and relational recommender system is also proposed modeling human preferences over styles and substitutes (McAuley et al., 2015b).

3 Data (Collection and Labeling)

Amazon Product Data, as mentioned in the previous section, is released by Julian McAuley at UCSD and made available online with product reviews and metadata from Amazon. It originally contains 143.7 million reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs) spanning from May 1996 to July 2014.[1] For the current study, we choose reviews from the 'Cell Phones and Accessories' category, which originally contains about 2.7 million reviews.

As can be shown in the sample data, the review consists of "reviewerID", "asin", "reviewerName" which are identifiers for the reviewer and the product. The most important fields in the review data are the "helpful", "reviewText", "overall" and "summary". We have used "reviewText" and "summary" to get the text features, "helpful" as a meta feature, and "overall" (Overall Rating gained in the review) as the label. The distribution of the overall ratings can be found in the figure 1. According to the given distribution, we take the label as positive if the overall rating is [0-4] and negative if the overall rating is 5. This is because the overall ratings are not uniformly distributed, as the figure 1 shows that the overall rating of 5 has almost half of the number of total reviews.

```
{
"reviewerID": "A1EWN6KZ4HMLT7",
"asin": "011040047X",
"reviewerName": "S. Prescott",
"helpful": [2, 2],
"reviewText": "The case pictured is a
soft violet color, but the case cover
I received was a dark purple. While
I'm sure the quality of the product
is fine, the color is very different.",
"overall": 1.0,
"summary": "Wrong color",
"unixReviewTime": 1344902400,
"reviewTime": "08 14, 2012"
}
```

Listing 1: Sample Data: Helpful Review

```
{
"reviewerID": "A1YX2RBMS1L9L",
"asin": "0110400550",
"reviewerName": "Andrea Busch",
"helpful": [0, 0],
"reviewText": "Saw this same case at a
theme park store for 25 dollars.
This is very good quality for a
great price.",
"overall": 5.0,
"summary": "Great product",
"unixReviewTime": 1353542400,
"reviewTime": "11 22, 2012"
}
```

Listing 2: Sample Data: Unhelpful Review

4 Method and Data Splitting

Data Cleaning: To receive better training results, we remove reviews which is deemed unhelpful by other customers, i.e., reviews voted as unhelpful by, or reviews received no voting at all from other users. In the former situation, the value of "helpful" would be [0, x], meaning that none of x users think the review is helpful, while in the latter, it would be rendered as [0, 0]. For instance, in the second sample data, even though the review is rated as 5, the vote for its helpfulness is zero. As a result, we do not include this review in our training data. After data cleaning, the number of reviews decreases from 2.7 million to 700,000. Out of these we use 6,000 reviews for fast processing.

Data Splitting: For the current project, we use 6,000 reviews, and split them into 80% training data and 20% test data.

Features: We use bag of words (BOW), part of speech (POS) tagging from NLTK,[2] and other features such as the field of ”helpful.” We do not remove stopwords, for it causes loss of information.

Training the Classifier: We use Support Vector Machines (SVM) and Naive Bayes to classify our data.

5 Evaluation

5.1 Support Vector Machines

Support vector machines are supervised learning models that analyze the data and recognize correct patterns. It is extensively used for classification and regression techniques. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap.

[1] http://jmcauley.ucsd.edu/data/amazon/

[2] http://www.nltk.org/api/nltk.tag.html

In our project we have thus used the SVM for the classification of reviews into positive and negative. We have used the SVM provided by the Natural Language Processing Toolkit (Bird et al., 2009).[3] This tool has a particular type of input and gives the correct output. Initially we parse our training set to generate a correct input file from that set. This file has a format with features and labels for each review separated with spaces. The same procedure is carried out on the testing data for the comparison.

As we can see from the result in the figure 1, the accuracy of the SVM is much better than the Naive Bayes because it uses a smart approach to classify the inputs while the Naive Bayes uses very basic approach for the classification. Here the SVM draws a very definite margin between the labels and classifies them.

5.2 Naive Bayes

Naive Bayes is another classifier based on the Bayes probabilistic theory. It uses Bayes theorem to approach the classification and applies the dependencies and independencies to the features.

Naive classifier is competitive to SVM, but in many cases it underperforms because of its simplistic approach. As we can see in the Table 2, the accuracy of this classifier is 79% for summary as a feature. Since most reviewers express their emotions in summary in a precise manner, summery is considered as the most strong feature. Compared to SVM which gives 83% accuracy to single feature summary, Naive is weaker. Similarly, it differs from SVM significantly in accuracy when all the features are considered. Table 1 and Table 2 show that the accuracy rates for Naive Bayes and SVM are 0.83 and 0.86 respectively.

5.3 Most Informative Features

For the creation of word cloud, we initially create a matrix called matrix termMatrix. Matrix termMatrix represents term frequency matrix whose rows represent words and columns represent documents.

The next step is scaling. We scale the termMatrix vector of length 2 indicating the range of the size of the words. Here the max word represents maximum number of words to be plotted. We dropped the least frequently used words. We then plot the words into random order. If false, they will be plotted in decreasing frequency.

6 Discussion and Conclusion

The paper employs a decent method to classify product reviews as negative or positive. We have successfully implemented two classifiers, SVM and Naive Bayes. The results show that SVM outperforms Naive Bayes in accuracy as well as speed.

In the future we would like to add more features by conducting the feature selection. Due to the time constraint, this project is conducted using the dataset containing 6,000 reviews despite the availability of a large dataset. We believe that if we use the entire dataset, the accuracy can significantly increase. We would like to take the advantage of the open huge dataset to build a lexicon and a database like SentiWordNet[5].

Baseline - Majority Class	78.13%
only summary as features	0.79
all features	0.83

Table 2: Naive Bayes Accuracy

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python.* " O'Reilly Media, Inc.".

Ian Fellows, 2014. *wordcloud: Word Clouds*. R package version 2.5.

John B. Horrigan. 2008. Online shopping. Technical report, Pew Internet American Life Project.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Yin-Fu Huang and Heng Lin. 2013. Web product ranking using opinion mining. In *Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on*, pages 184–190, April.

Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015a. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 785–794, New York, NY, USA. ACM.

[3] http://www.nltk.org

[5] http://sentiwordnet.isti.cnr.it

			Baseline	Majority Class	78.13%	
		precision	recall	f1-score	support	accuracy_score
only summary as feature	negative	0.79	0.54	0.64	28	
	positive	0.84	0.94	0.89	71	0.83
	avg / total	**0.82**	**0.83**	**0.82**	**99**	
All features	negative	0.71	0.50	0.59	20	
	positive	0.88	0.95	0.91	79	0.86
	avg / total	**0.85**	**0.86**	**0.85**	**99**	

Table 1: SVM Accuracy

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015b. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA. ACM.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karline Soetaert, 2014. *plot3D: Plotting multi-dimensional data.* R package version 1.0-2.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

7 Appendix

7.1 Response to the Instructor's Feedback

We create the majority class and add that as our baseline in response to Professor Abdul-Mageed's suggestion on modifying and putting the majority class as our baseline. In accordance to Professor's recommendation of focusing on better feature selection, we try to add better features beside simply adding more features. For example, we use metadata of "Summary" as our feature which provides precise emotion of the reviewer. Even though the feature we select is good enough, we think there still room to improve our accuracy by selecting more and better features.

7.2 Author Contributions

Conceived and designed the experiments: MP, SS. Performed the experiments: MP, SS. Wrote the paper: PC, MP, SS.

7.3 Author Information

Mrunal Pagnis: Master's Student in Computer Science.

Suraj Chandrakant Songire: Master's Student in Computer Science.

Pei-Ying Chen: Doctoral Student in Information Science. Class missed: 0.

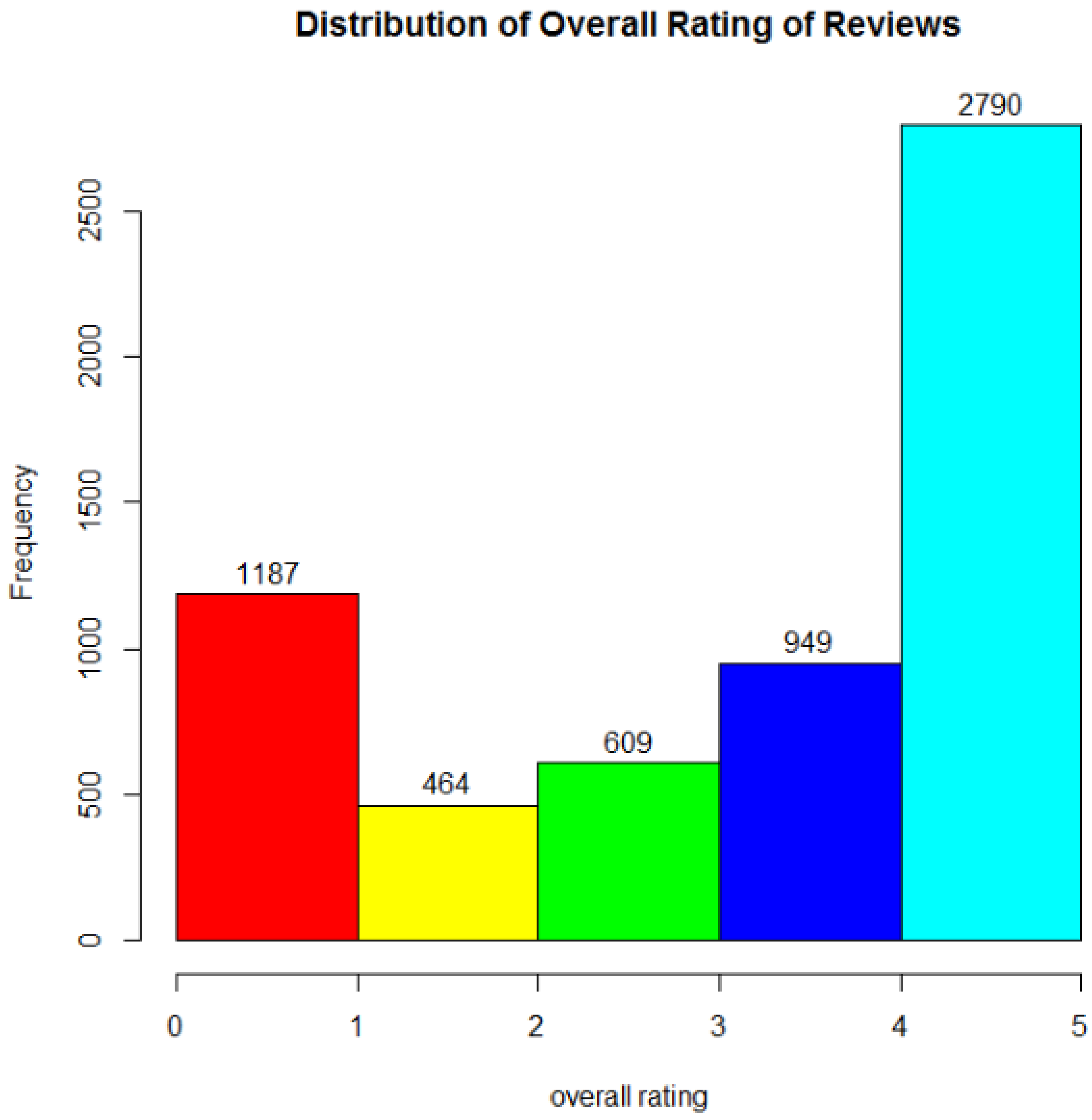

Figure 1: Distribution of Overall Ratings
(Soetaert, 2014)

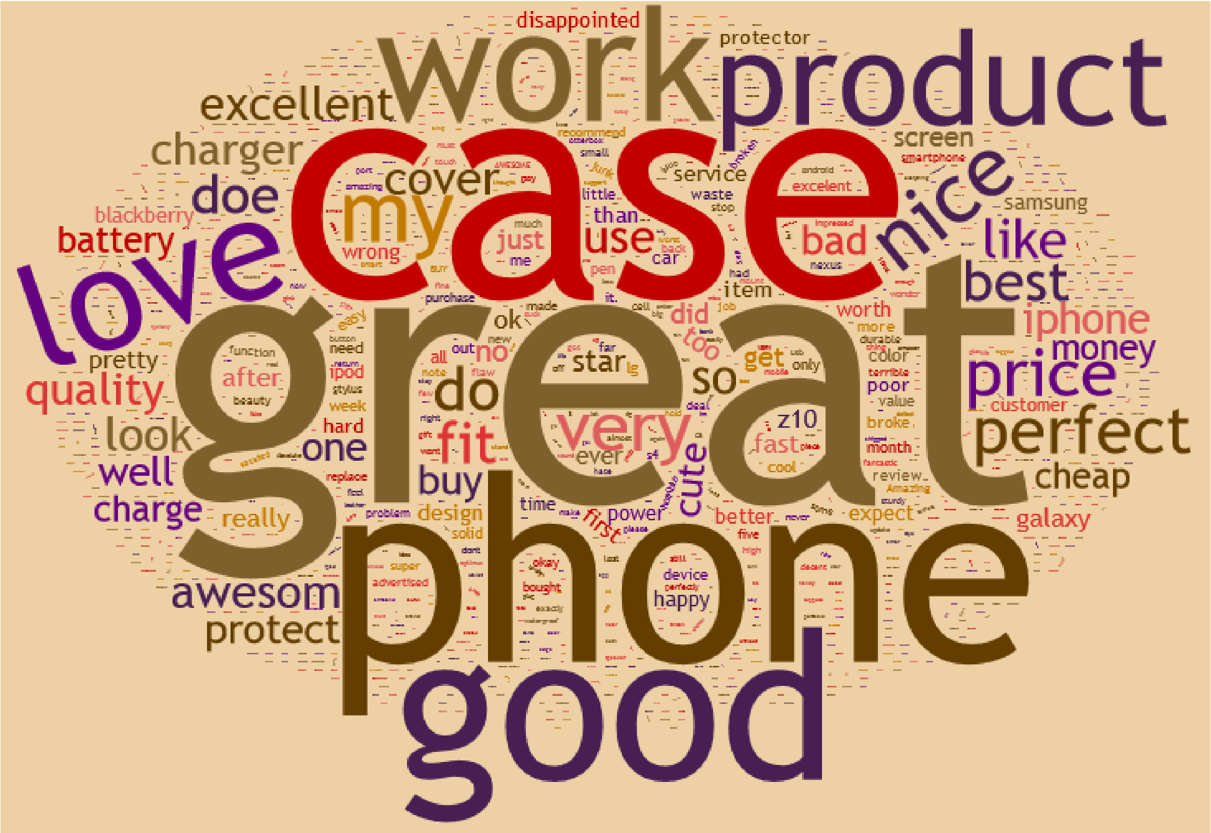

Figure 2: Word Cloud of Most Informative Words
[4] (Fellows, 2014)