Dynamical Properties Of E-Commerce Website

Nikky Kumari
14CS60R11

Supervised by:
Prof. Niloy Ganguly
Prof. Animesh Mukherjee

May 5, 2016

Overview

- **The role of outsiders in consensus formation: A case study of Yelp.**
- **Amazon sales-rank prediction.**

Introduction

Problem Statement

To find the presence of consensus in adoption of products on e-commerce sites at different geographical levels.

Overview of Project

- For analyzing the consensus at different geographical level, we have used online reviews and ratings to see the adoption of a particular product or service by the people.

Overview of Project

- For analyzing the consensus at different geographical level, we have used online reviews and ratings to see the adoption of a particular product or service by the people.
- For getting online reviews, we have used Yelp, a popular reviewing site where user can give reviews and ratings for various services like restaurant, music clubs, gym etc.

Overview of Project

- For analyzing the consensus at different geographical level, we have used online reviews and ratings to see the adoption of a particular product or service by the people.
- For getting online reviews, we have used Yelp, a popular reviewing site where user can give reviews and ratings for various services like restaurant, music clubs, gym etc.
- By analyzing those reviews, we have studied consensus in various geographical level and have come to the following conclusions:
  - There is Consensus among non local(outside) reviewers.

Overview of Project

- For analyzing the consensus at different geographical level, we have used online reviews and ratings to see the adoption of a particular product or service by the people.
- For getting online reviews, we have used Yelp, a popular reviewing site where user can give reviews and ratings for various services like restaurant, music clubs, gym etc.
- By analyzing those reviews, we have studied consensus in various geographical level and have come to the following conclusions:
  - There is Consensus among non local(outside) reviewers.
  - Same does not happen for local(inside) reviewers.

Overview of Project

- For analyzing the consensus at different geographical level, we have used online reviews and ratings to see the adoption of a particular product or service by the people.
- For getting online reviews, we have used Yelp, a popular reviewing site where user can give reviews and ratings for various services like restaurant, music clubs, gym etc.
- By analyzing those reviews, we have studied consensus in various geographical level and have come to the following conclusions:
  - There is Consensus among non local(outside) reviewers.
  - Same does not happen for local(inside) reviewers.
  - Inherent differences lie between the inside and outside reviewers which can be the possible reasons for above conclusions.

Data collection and Description

- Yelp Data has been collected by two different ways i.e Yelp Dataset Challenge and crawling.
- The data consisted of information about 61,184 different Business objects, 366715 different user objects and 1569264 review objects.
- For the above three objects following information were present:

Business Objects:

Business id
Name of Business
Unit
City
Latitude-Longitude
Review Count

Review Objects:

Business id
User id
Stars
Review Text
Date Of Review

User Objects:

User id
Name of User
Review Count
Average Stars

- The data provided by the challenge consists of the data from year 2006 to year 2015.

User to Locality Mapping

- Mapped cities of only those users who have reviewed those business units which we have taken into consideration.

User to Locality Mapping

- Mapped cities of only those users who have reviewed those business units which we have taken into consideration.
- Crawled reviews for those business units in time period 2006-2015.

User to Locality Mapping

- Mapped cities of only those users who have reviewed those business units which we have taken into consideration.
- Crawled reviews for those business units in time period 2006-2015.
- Crawled review data, users name, city, date of review and stars rating.

User to Locality Mapping

- Mapped cities of only those users who have reviewed those business units which we have taken into consideration.
- Crawled reviews for those business units in time period 2006-2015.
- Crawled review data, users name, city, date of review and stars rating.

- Mapped the user-id given in the dataset to city based on users name, date and stars.

Categorization

- With respect to a particular business unit, reviewers are categorized based on location of business units and location of reviewers.
  - Inside Reviewers:Reviewers who belong from same city as that of business.
  - Outside Reviewers:Reviewers who belong from different city.

Categorization

- With respect to a particular business unit, reviewers are categorized based on location of business units and location of reviewers.
  - Inside Reviewers:Reviewers who belong from same city as that of business.
  - Outside Reviewers:Reviewers who belong from different city.
- We have broadly categorized businesses into two broad category on the basis of its reviewers:
  - Category 1:Approximately 90-97% of reviewers are outsiders .
  - Category 2:Approximately 50-60% of reviewers are outsiders.

Result of Categorization

- We have taken 9 classes of business into consideration and categorized them into category1 and category2.

Category 1			
Business Class	Total Reviewers	insiders %	outsiders %
Arts and Entertainment	21480	3.80	96.20
Event planning	22120	3.80	96.20
Hotels and Travel	36120	9.01	90.99
Restaurant	22720	4.15	95.85
Category 2			
Grocery	2800	47.20	52.40
Gym	1520	40.19	59.81
Auto Repairs	2160	48.75	51.25
Health and Medical	1040	55	45
Music Videos	1160	44	56

Table: The different business types.

Entropy method

- We have used entropy of ratings to see the consensus.

Entropy method

- We have used entropy of ratings to see the consensus.
- Calculated six month moving average of entropy of ratings for a particular business unit over the period of 9 years.

Entropy method

- We have used entropy of ratings to see the consensus.
- Calculated six month moving average of entropy of ratings for a particular business unit over the period of 9 years.
- Entropy is $-\sum_1^5 p_i \log p_i$ where p_i denotes the fraction of reviewers who have given rating i.

Entropy method

- We have used entropy of ratings to see the consensus.
- Calculated six month moving average of entropy of ratings for a particular business unit over the period of 9 years.
- Entropy is $-\sum_1^5 p_i \log p_i$ where p_i denotes the fraction of reviewers who have given rating i.
- It is then averaged over top 10 business units of each class of business unit.

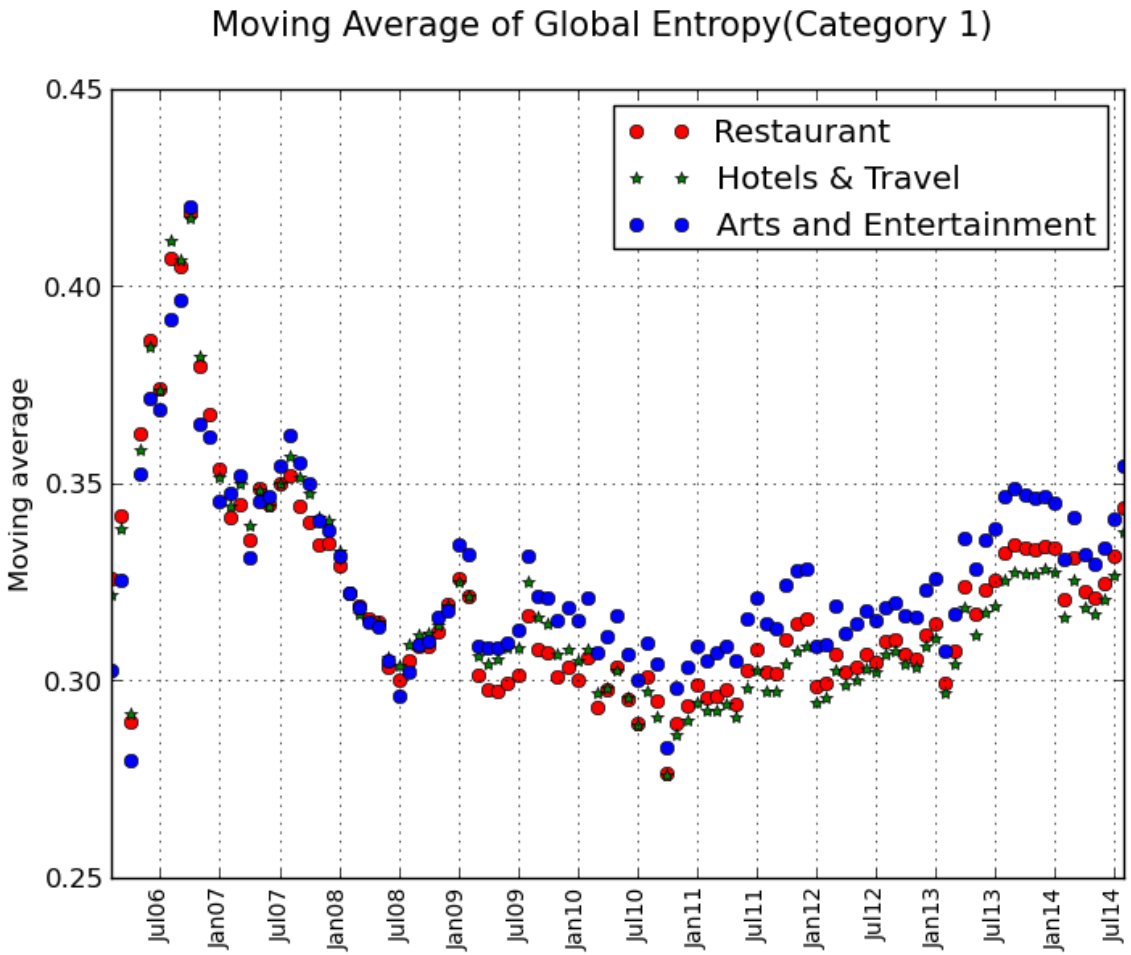
Entropy method

- We have used entropy of ratings to see the consensus.
- Calculated six month moving average of entropy of ratings for a particular business unit over the period of 9 years.
- Entropy is $-\sum_1^5 p_i \log p_i$ where p_i denotes the fraction of reviewers who have given rating i.
- It is then averaged over top 10 business units of each class of business unit.
- The decline in entropy depicts the agreement between reviewers.

Entropy method

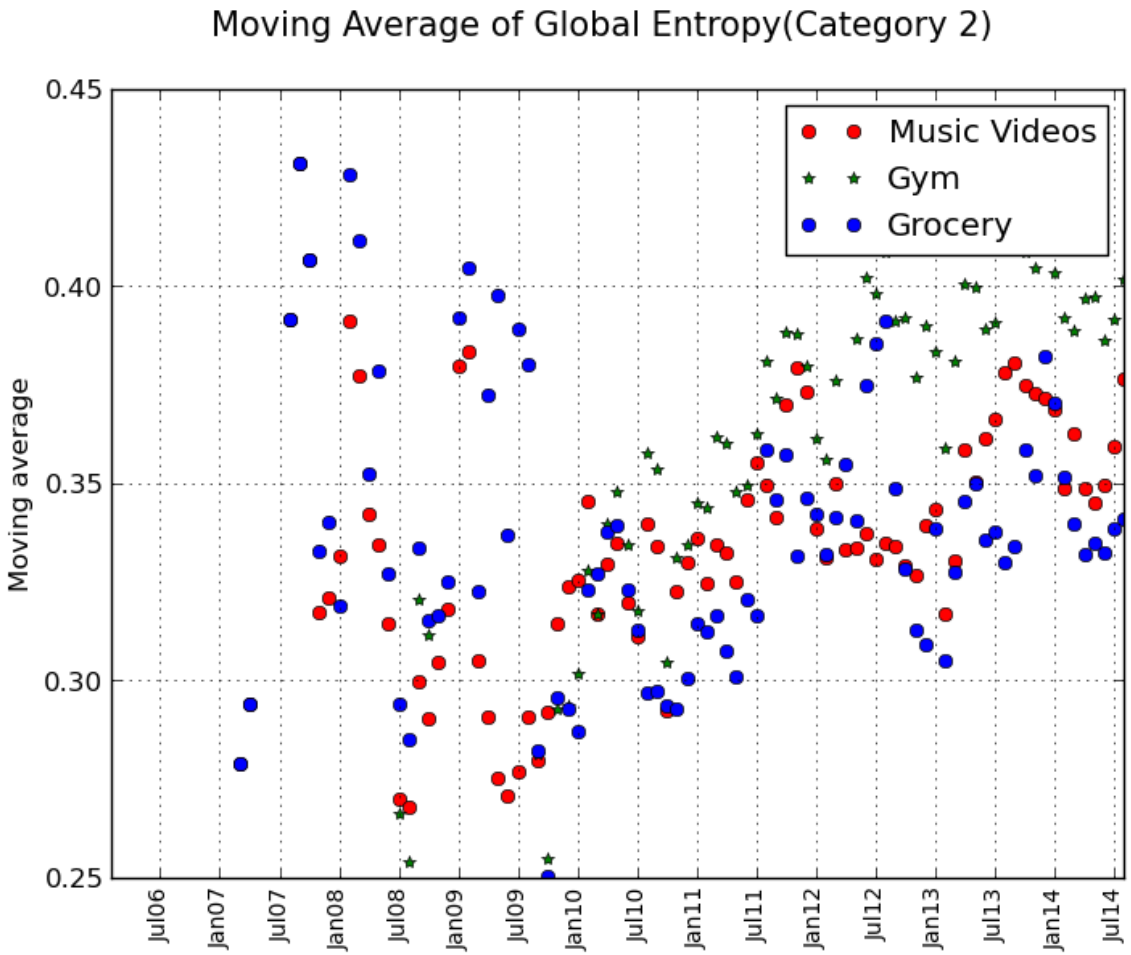
- We have used entropy of ratings to see the consensus.
- Calculated six month moving average of entropy of ratings for a particular business unit over the period of 9 years.
- Entropy is $-\sum_1^5 p_i \log p_i$ where p_i denotes the fraction of reviewers who have given rating i.
- It is then averaged over top 10 business units of each class of business unit.
- The decline in entropy depicts the agreement between reviewers.
- No such decline would imply that reviewers do not agree to a same choice.

Global entropy for category 1



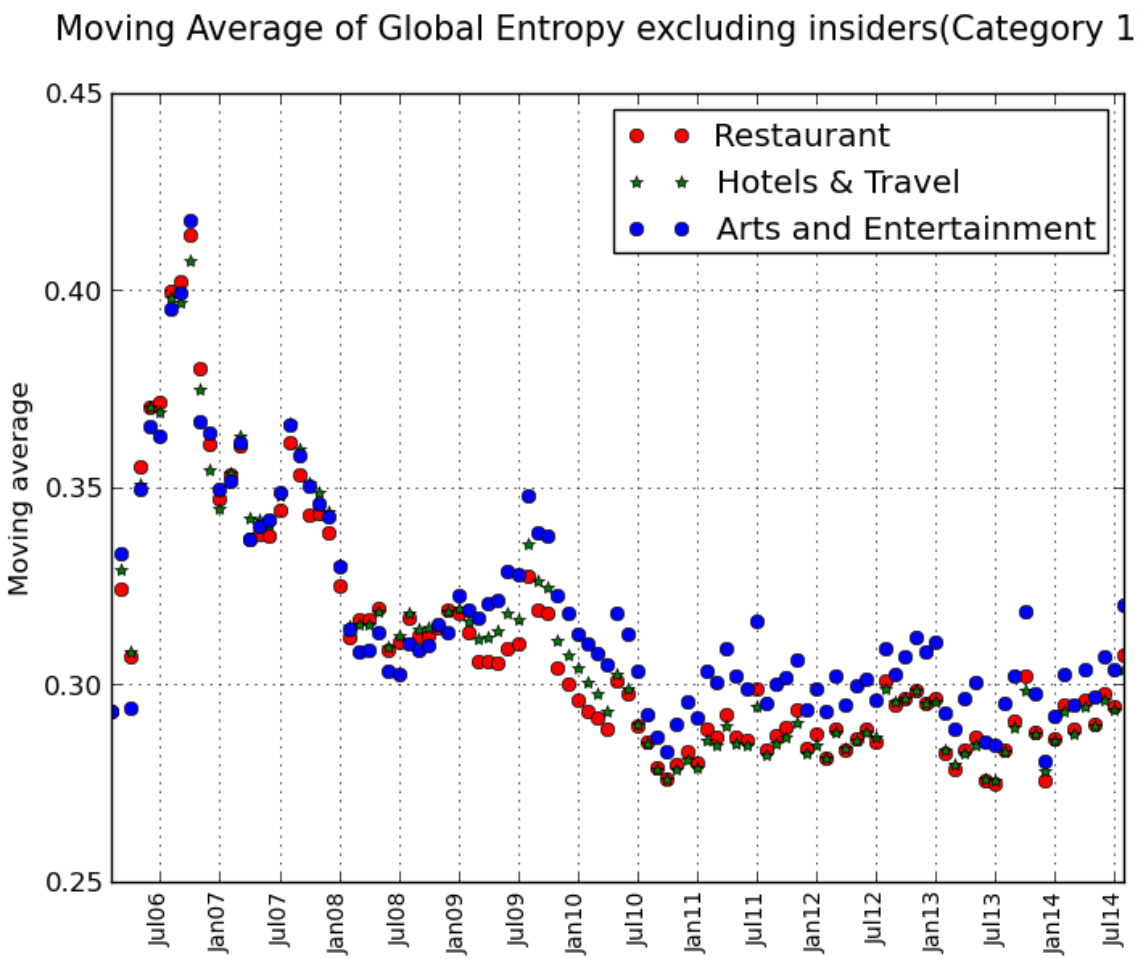
The decline of entropy for category 1 which is outsider driven depicts that ratings of reviewers are converging to a single value over time.

Global entropy for category 2



For category 2 no such decline can be seen and so we can say that reviewers have mixed opinion.

Entropy for category 1 considering only outsiders



Entropy for all business types is on decline for Category 1 when considered only outsiders implies that outsiders are responsible for consensus formation.

Result and analysis

- We have seen from above plots that there is decline in entropy where outsiders are more actively participated.
- So we can make a conclusion that outsiders are responsible for consensus formation.
- By further analysis we have found out that there are inherent differences between outsiders and insiders in following areas which can be the possible reason for the above conclusion:
  - Fake Reviews
  - Early Adopters
  - Elite Users

Fake Reviews

- Crawled fake reviews(as identified and blacklisted by Yelp) from Yelp.
- Calculated the percentage of reviews which are fake for both insiders and outsiders.
- On an average, percentage of insiders producing fake reviews is more than outsiders, with more pronounced difference for Category 1.
- Insiders could possibly have vested interests in promoting particular business unit in their local neighbourhood (by writing fake reviews) while demoting others.
- For an outsider, such interests are naturally absent.

Fake Reviews Results

Category 1			
Business class	overall fake %	insiders fake %	outsiders fake %
Arts and Entertainment	21.37	35.33	20.70
Event planning	21.06	35.52	20.35
Restaurant	10.27	45.20	7.73
Category 2			
Grocery	17.65	19.55	15.83
Gym	20.04	25.40	16
Health Medical	30.52	32.23	28.33

Table: Percentage of fake reviews in the two categories.

Early Adopters

- One of potential causes for opinion formation is attributed to presence of early adopters.
- We assume the first 25% reviewers who have rated business unit to be early adopters.
- From result, we observe outsiders are far more likely to be early adopters than insiders.
- This observation is more pronounced for category 1 as it has a huge number of outsiders.

Early Adopters Results

Category 1		
Business class	insiders early adopter%	outsiders early adopters %
Arts and Entertainment	15.03	20.19
Event planning	14.30	20.21
Restaurant	18.98	20.10
Category 2		
Grocery	16.87	20.65
Health and Medical	19.23	20.94
Gym	12.11	14.74

Table: Percentage of early adopters for the two categories.

Elite Users

- Yelp endows elite status to reviewers depending on quality and trustworthiness of review content generated by them.
- We crawled elite user information from Yelp to see the proportion of insiders and outsiders in elite users.
- We observed that outsiders in general are more elite reviewers than insiders.

Category 1			
Business class	overall elite%	insiders elite%	outsiders elite%
Restaurant	30.95	20.79	31.9
Category 2			
Grocery	27.03	12.22	40.4

Table: Percentage of elite users in the two categories.

Conclusion

- According to our study, we have come to the conclusion that outsiders are more active in writing reviews.

Conclusion

- According to our study, we have come to the conclusion that outsiders are more active in writing reviews.
- Outsiders are more genuine reviewers as compared to the insiders as they are are less involved in writing fake reviews.

Conclusion

- According to our study, we have come to the conclusion that outsiders are more active in writing reviews.
- Outsiders are more genuine reviewers as compared to the insiders as they are are less involved in writing fake reviews.
- Outsiders are mainly responsible for information cascade and consensus formation.

Amazon sales-rank prediction

Amazon Data Description

- The Amazon dataset which we have used contains product reviews and metadata about the products.
- It consisted of nearly 140 million reviews spanning may 1996 to july 2014.
- Amazon data is divided into multiple categories like books, electronics, clothing, cds etc.

Review objects:-

review id
product id
review text
review time
rating

Metadata objects:-

product id
name of product
salesrank
also bought products
also viewed products

Salesrank

Salesrank is a number with 1 to 8 digits and it is given for every product on Amazon to depict the popularity of that product.

Few general facts about salesrank:-

- Better products have lower sales rank and worse products have higher sales rank.
- Sales ranks are different in various categories.Products have overall ranks and also specific ranks within subcategories.

Problem Statement

Given the data of a product item on Amazon, make an early prediction of its sales-rank (low/high) in a given category.

Data Selection

- We have taken the category "CDs" of amazon for our study.
- There were total 492799 number of cds present in our data set.
- We have selected cds which is having average of 1 reviews per month from 2010 to 2013.That accounted for 2600 cds.
- We have sorted the cds according to sales rank and taken top 200 cds and bottom 200 cds as two different classes as popular or unpopular products.
- On the basis of different features, we have tried to predict whether a cd will belong to a popular class or unpopular class after certain period of time.

Feature Analysis

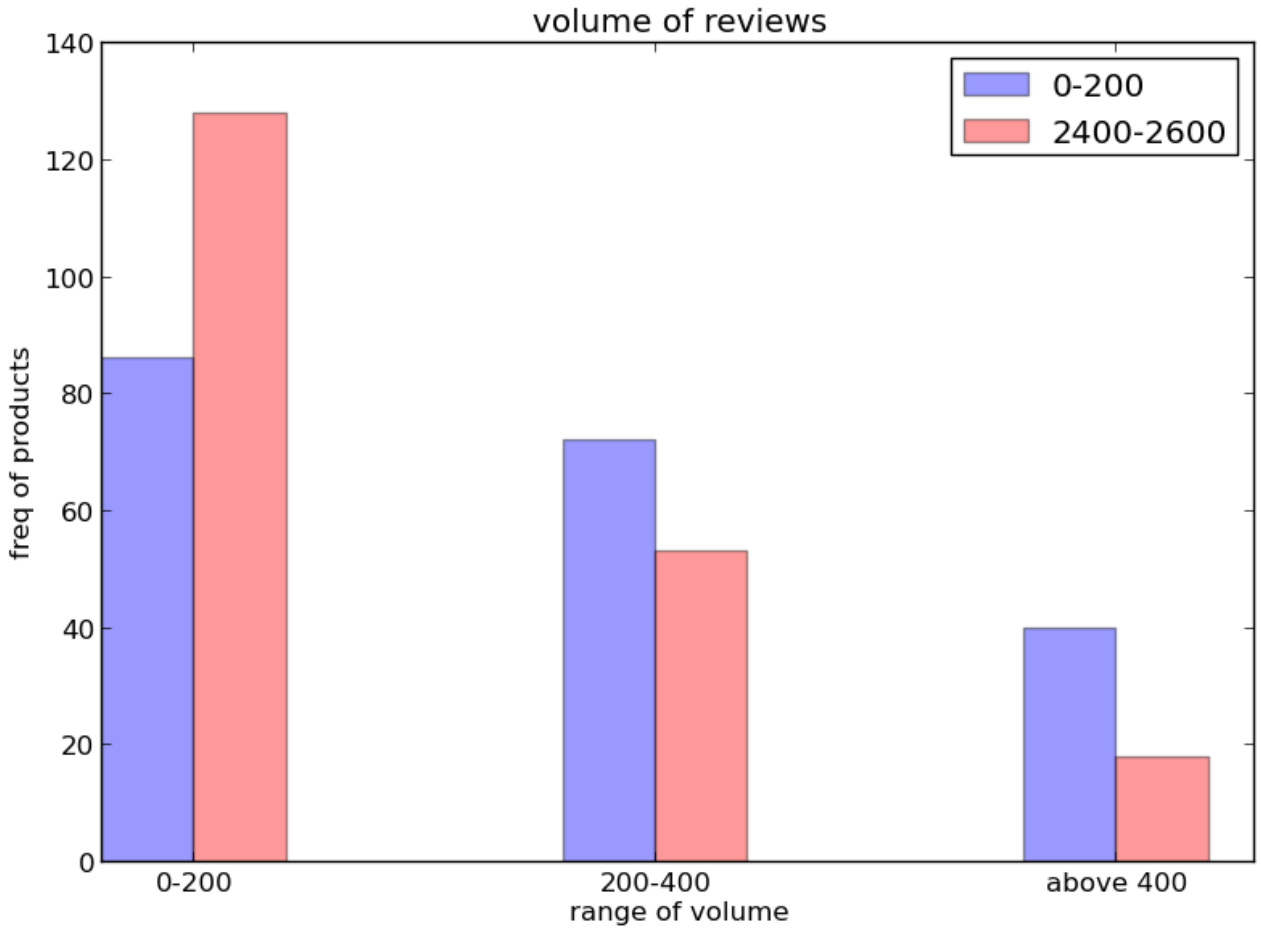
General Features:-

- volume of reviews
- percent of fake reviews
- dwell time
- word diversity
- number of words
- sales rank of also bought products

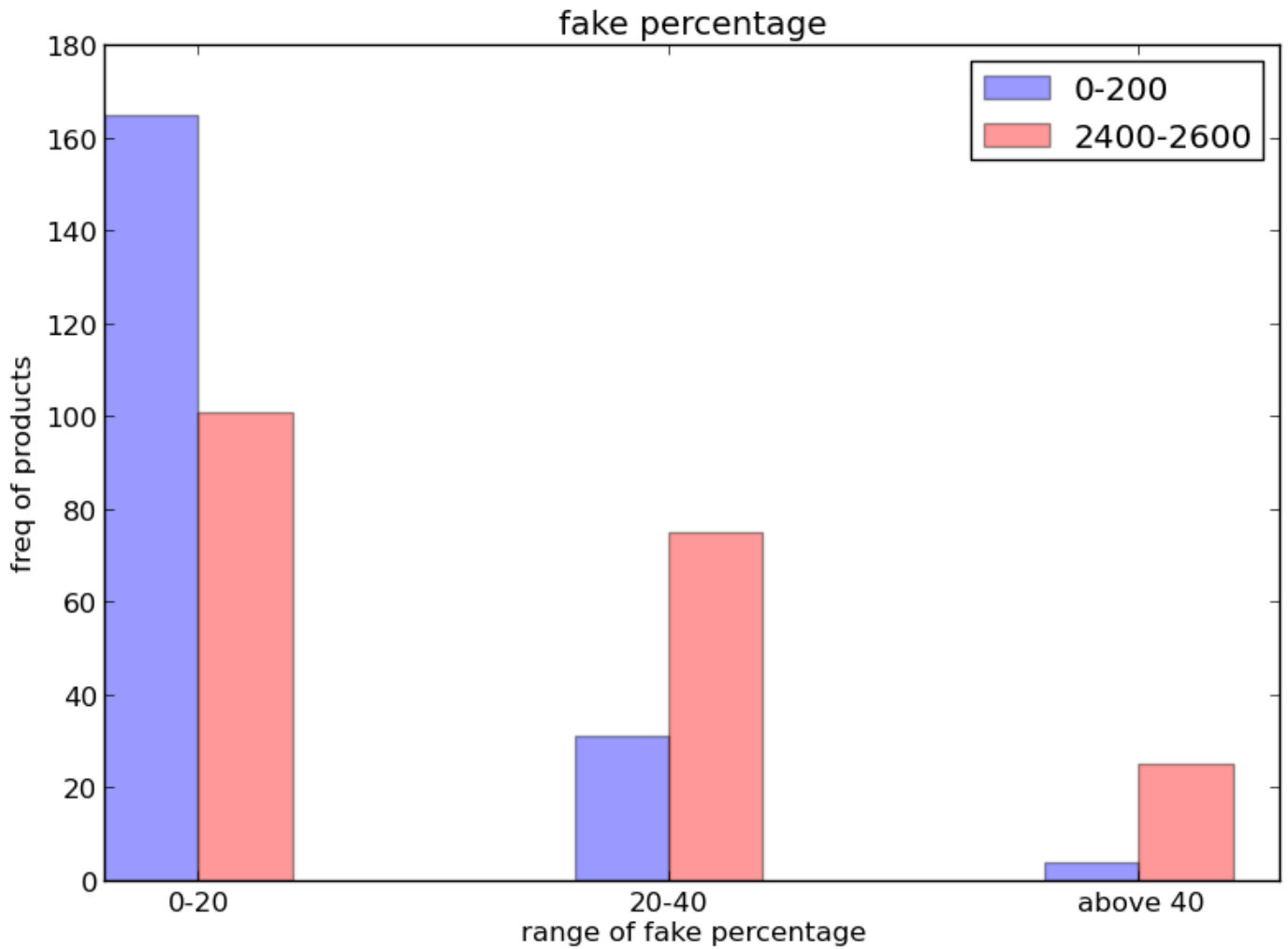
Linguistic Features:-

- sad
- anger
- negative emotion

Volume of Reviews

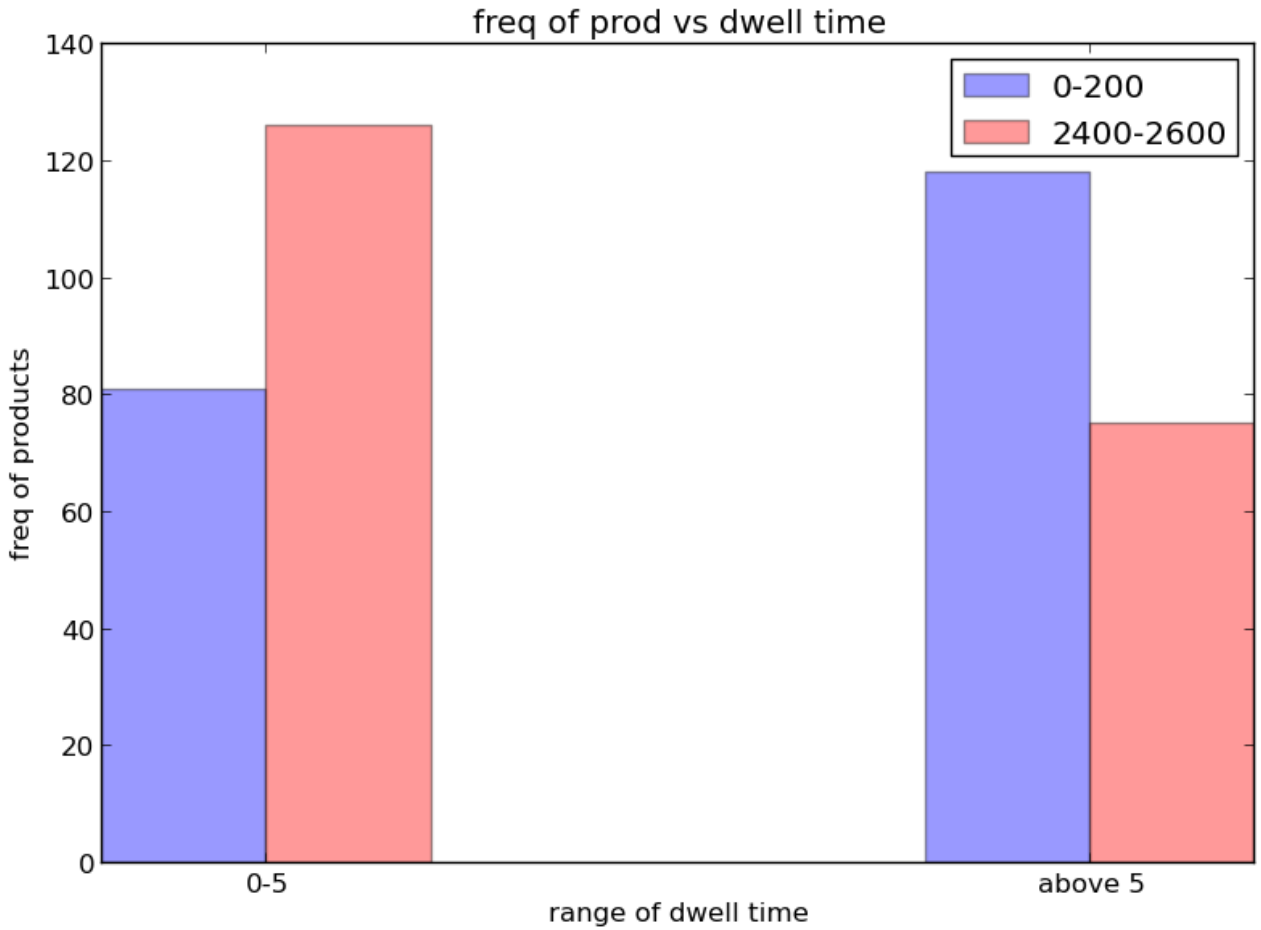


Percentage of Fake Reviews

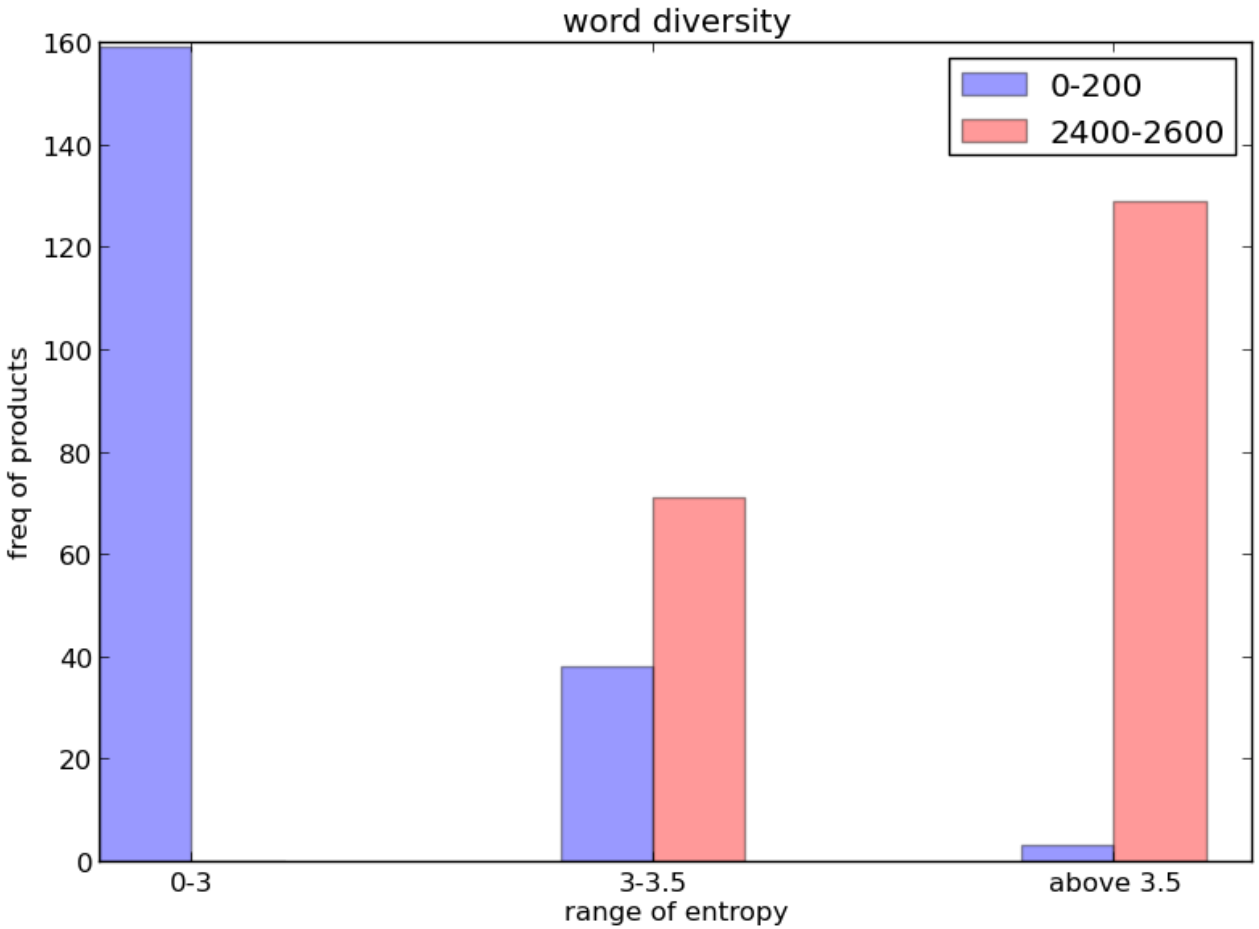


Using open-source implementation of Bernoulli supervised learning approach

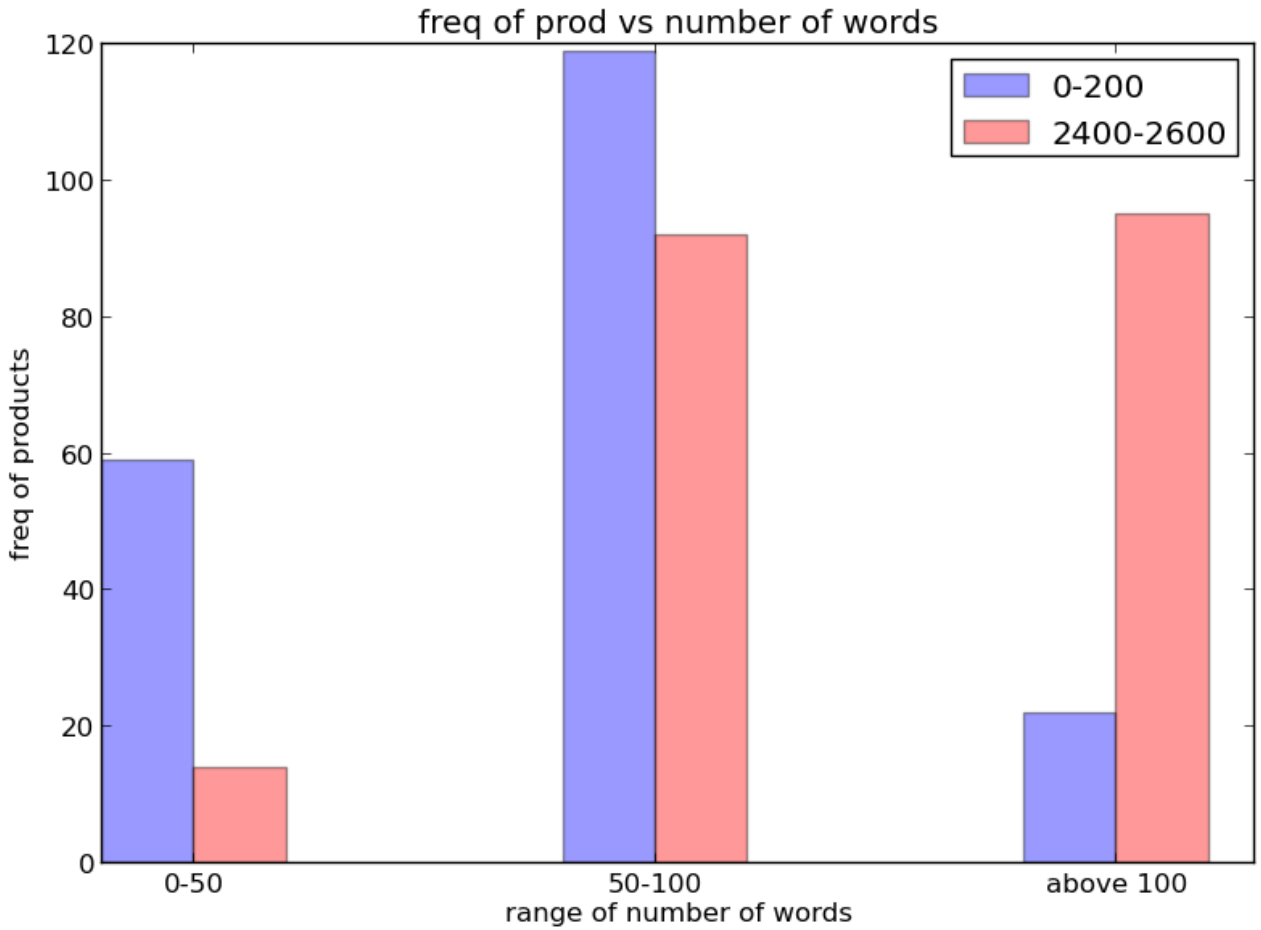
Average Dwell Time



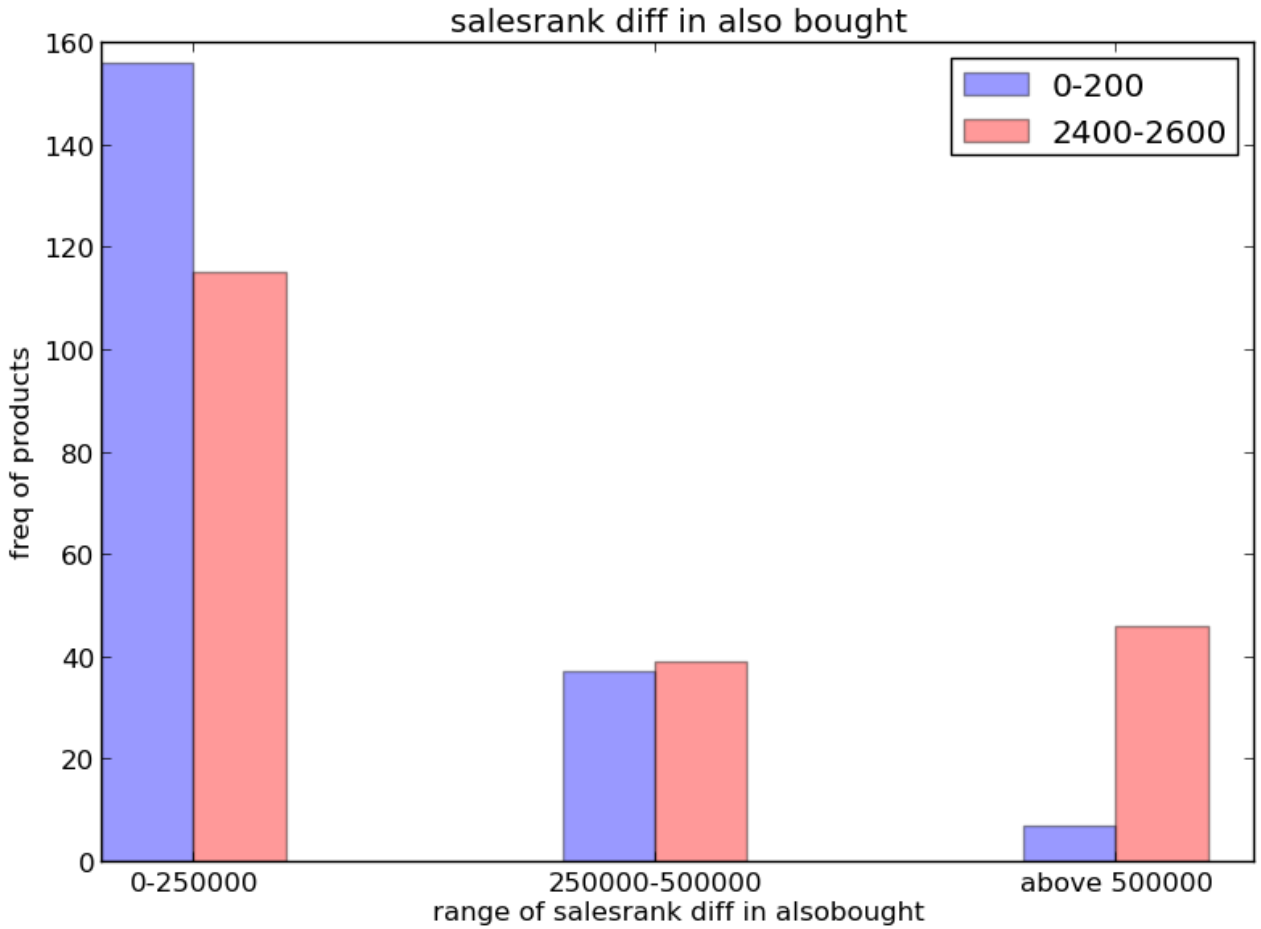
Word Diversity



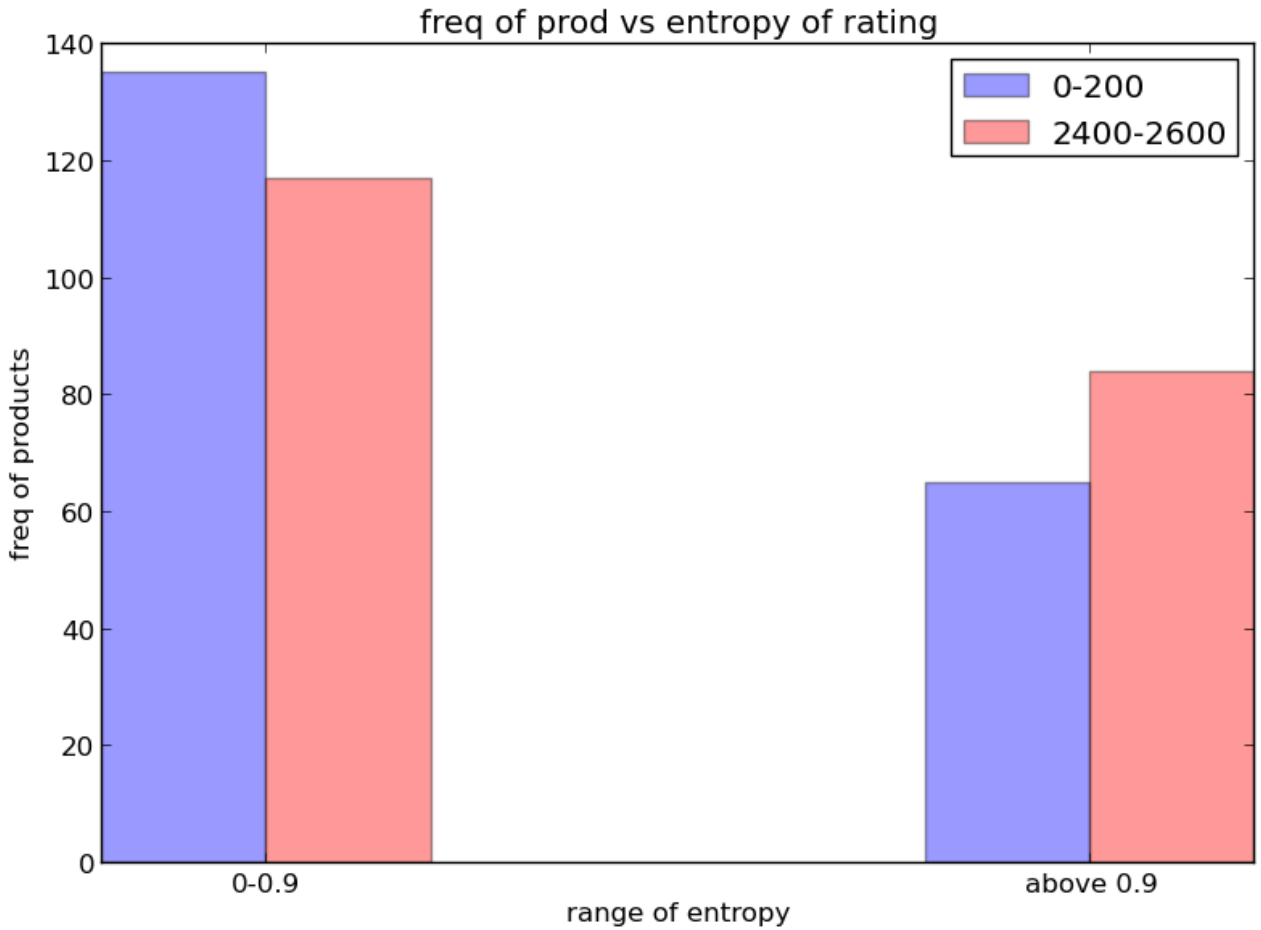
Number of Words



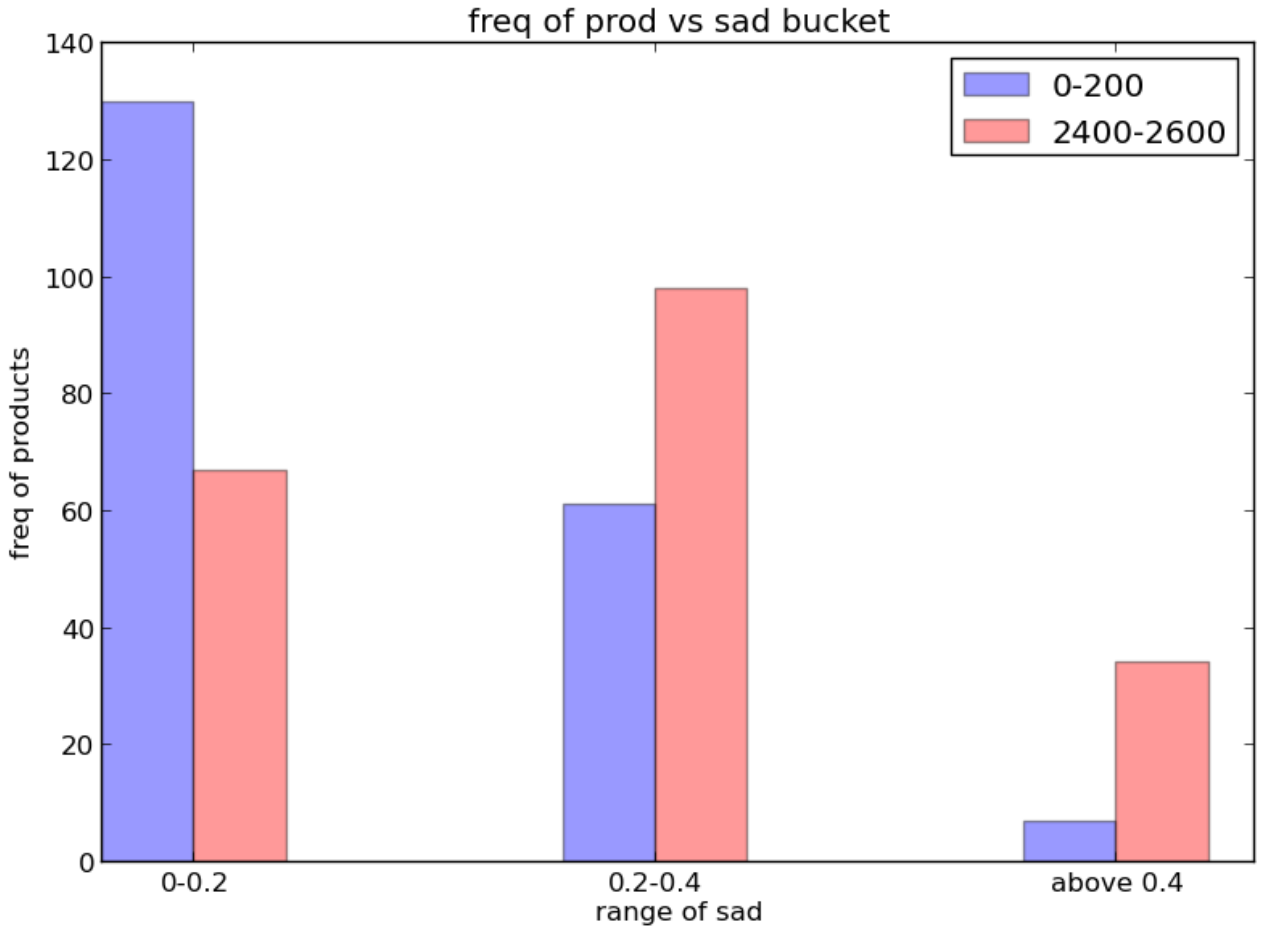
Sales Rank of Also Bought Products



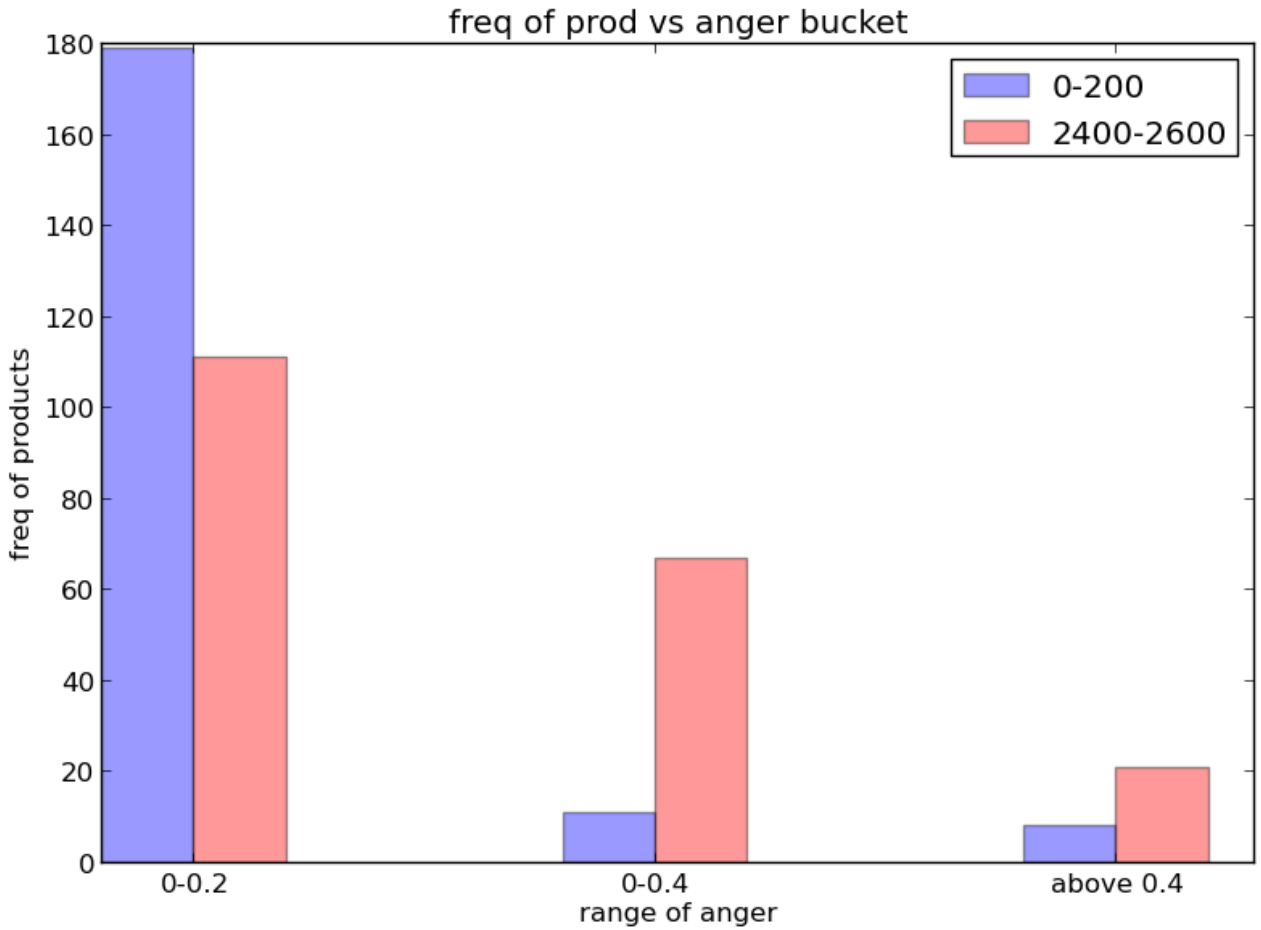
Entropy of ratings



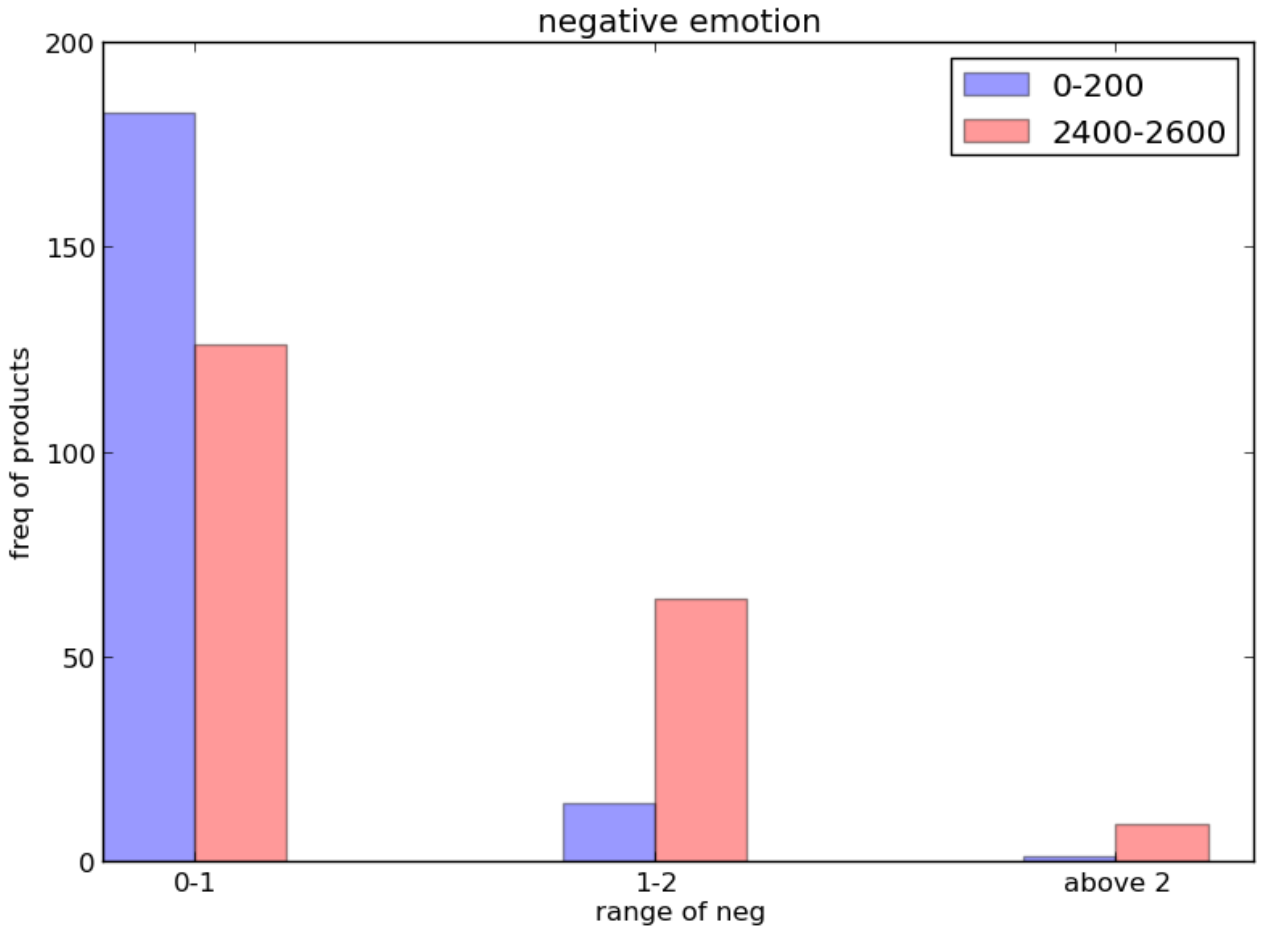
LIWC feature:Sad



LIWC feature:Anger



LIWC feature:Negative Emotion



Classification of products based on sales-rank

- We used linear SVM classifier for classification.
- We used test set of size 50(25 from either class) in each fold of 10-fold cross validation.
- Used values of above mentioned features until 2013 to predict the salesrank after July 2014.
- Achieved an accuracy of 78%.

Conclusion

- We have tried to make the early prediction of sales-rank and achieved an accuracy of about 78%.
- We can increase the accuracy by using more efficient features.
- We can also try with different classifiers to see if accuracy goes high.

Thank you