NISTADS Tracks in Policy Research: NISTADS/DATA Analytics/2019/1

CSIR–National Institute of Science,
Technology and Development Studies

INDUSTRIAL DATA ANALYSIS

March 2019

Vinayak

Analysis of Product Reviews and Linguistic Statistics

Vinayak and P. Goswamy

CSIR-NISTADS

INTRODUCTION

Analysis of product reviews has been of general socioeconomic interests. Review data are usually available in sufficiently big volumes as required to withdraw substantial statistical conclusion over various issues. Reviews are supplied in natural language which may be processed and analysed using tools and theories developed in the context of Natural Language Processing (NLP) and Linguistic Statistics (LS). However, in the studies regarding LS, results have been reported mostly in the context of classic-literature. Reviews are in contrast to the literature-text in a sense that these are focused over the quality, usability and other related issues of a product, formally constrained with limited choice of words or sentences. Besides, review collecting agencies desire reviewers to highlight *pros and cons* of a product in few words a requirement which may not necessarily meet tidily.

Reviewers may use different words from vocabulary and make different types of sentences to express their opinions or suggestions about a product. They also decorate their reviews with star ratings facilitating relatively an easier comparison for the analyser. Normally a description is supplemented along with a remark justifying the star rating. The former in the dataset appears as body and the latter as the headline. In both, meaningful words express emotions and appear in appropriate order and related with the star ratings. In contrast, there appear less-relevant, or irrelevant, or jargon words representing bias or mental condition resulted from disappointment, annoyance, tantrum, or ecstasy, elation, gratification, due to expectations and use, or because of unfocussed review-writing. Study of such words might be a good measure in behavioural science, nevertheless has to be discarded. Empirically the former type of the word-category appears more frequently than the latter. It seems to be a good assumption to start with that the frequency associated with both types of categories follow different statistics. Based upon statistical findings, one may then attempt to classify the substantial part of the corpus and follow conventional methods for making sensible predictions. It is worth to mention that various methods and technics have been used so far to reduce the corpus size for such analysis—a worthy mention would be the data-size reduction method like principal component analysis on the count-vectorised matrix, what we represent in this report as X, encountered in the so-called “Bag-of-Words” approach.

Human tendency of using a word for a certain purpose is empirically repetitive, suggesting a pattern or universal law exhibiting such properties. In LS universal statistical behaviour has been much discussed of. Ample literature may be found discussing Herdan’s law and Zipf’s plot etc.

The Herdan's law underlines dependency of the number of distinct words on the number of texts, which shall be the number of reviews in our case. Zipf's law exhibit a power law in the frequency based ranking of words.

We show in this report two power laws; the first one in the number of appearances of a frequency vs frequency plot, and the other one in their contribution vs frequency plot, for small frequencies. For high frequencies we introduce three types of measures viz., the expectancy, the distinction and the weight, of a word based upon the use of a word per word, per distinct word, and per review. The expectancy is a positive rational number quantifying chance of a word to appear in a review. Similarly, the distinction measures how distinctively a word is used in the corpus. Finally, we use weight to understand contribution of a word in the corpus. We study these as a paradigm to identify words describing preference of decorating a product with stars. For instance, one may well use a word "good" to rate a product from 2 to 5 star values. Presumably a comparative analysis of these measures shall immediately reveal preference of a word to a star. These, however, allude to the properties imbedded in the matrix structure of X. For instance, expectancy hints about the column-wise sparsity associated with the matrix X. We must mention here the matrix X is populated with positive integers only and the dimension of the matrix depends on the size of the corpus. For instance, number of columns of X is determined by the number of distinctive words in the corpus, which increases with the corpus size in accordance with the Herdan's law.

Objective

In this report we use analyse Amazon review dataset in order to develop understanding of the technicalities and statistical properties of such complex datasets. Though various methodologies have been devised, and might have been in use by private industries with different objectives, we aim towards developing a general approach to understand reviews, available in bulk, within a framework of limited words with quantification as will be required for comparison and further studies.

Report Structure

The report is arranged as follows. We first describe the metadata. There are some information available along with the datasets, however, we have worked out further details as will be required in the analysis. For example, we detail number of reviews and number of reviews corresponding the verified purchase category in table 1. Next, we discuss different columns of the

datasets as available online resources. We further present a brief analysis describing tendency of reviewers to purchase an item based upon star ratings. Here we assume that every product with reported review have been sold at least for once. This assumption is justified with the reviews with verified purchase category as per the available explanation of the dataset.

We study statistical properties of the linguistic of the reviews of products. In this section we study universal laws, as discussed in the introduction, for the low frequency part of the bag-of-words. In this part we perform analysis of the four datasets of a product: Four datasets of a product come from the verified and unverified purchase categories of the review-body and review-headline as will be explained in the metadata section. We primarily focus upon products viz., toys and shoes for the total number of available reviews are close; See table 1. For the high frequency part we study the expectancy, distinction and weight measures of a word and compare results of the 8-corpus sets, i.e., four for each product. Finally we give details and link for the Amazon datasets used in this report in the last section.

THE METADATA

On the AWS webpage, the data is explained by 15 variables defining 15 data-columns, as listed below. Based upon online information the column for verified purchase is explained here in detail. In the analysis we treat both the verified and unverified categories separately. Reviews are explained in review_headline and review_body texts. This lives us with four types of corpus, viz. verified review headline (HV), verified review body (BV), unverified review headline (HUNV), and unverified review body (BUNV). Data for each product is available separately in the 15 columns, viz.

marketplace - 2-letters country code of the marketplace where the review was written. In our case this value is US in all rows.

customer_id - Random identifier that can be used to aggregate reviews written by a single author.

review_id - The unique ID of the review.

product_id - The unique Product ID the review pertains to. In the multilingual dataset the reviews for the same product in different countries can be grouped by the same product_id.

product_parent- Random identifier that can be used to aggregate reviews for the same product.

product_title - Title of the product.

product_category- Broad product category that can be used to group reviews (also used to group the dataset into coherent parts).

star_rating- 1-5 star rating of the review.

helpful_votes- Number of helpful votes.

total_votes - Number of total votes the review received.

vine - Review was written as part of the Vine program.

verified_purchase- The review is on a verified purchase. An "Amazon Verified Purchase" review means Amazon verified that the person writing the review purchased the product at Amazon and didn't receive the product at a deep discount. Reviews that are not marked "Amazon Verified Purchase" are valuable as well, but we either can't confirm that the product was purchased at Amazon or the customer did not pay a price available to most Amazon shoppers.

review_headline- The title of the review.

review_body- The review text.

review_date- The date the review was written.

A list of 46 products is given below in the first column of the table 1. In this table, we list number of readable reviews (readable in PYTHON3.5X) in the second column two and number of reviews corresponding the verified purchase category in the fourth column. Besides, we mention starting date of the reviews in the third column. All the products have the same last-review-date, i.e., 31 October 2015.

S. N.	Product Name	Number of Reviews[1]	First Review date	Number of Verified Purchase
1	WIRELESS	8991589	04-12-1998	7920181
3	WATCHES	960204	05-04-2001	830806
4	VIDEO GAMES	1780268	06-11-1997	1164785
5	VIDEO DVD	5049291	08-07-1996	3298590
6	VIDEO	380551	11-11-1995	95696
7	TOYS	4859607	05-01-1997	4017951
8	TOOLS	1740041	09-11-1999	1522771
9	SPORTS	4833094	09-10-1997	4268088
10	SOFTWARE	341249	21-09-1998	195088
11	SHOES	4358820	08-11-1999	3937353
12	PET PRODUCTS	2639853	23-08-1998	2320182
13	PERSONAL CARE APPLIANCES	85924	29-10-2000	63260
14	PC	6906869	01-07-1999	6047348
15	OUTDOORS	2299811	24-03-1999	2023594
16	OFFICE PRODUCTS	2640254	15-07-1998	2255942
17	MUSICAL INSTRUMENTS	904004	13-12-1999	780914
18	MUSIC	4740849	11-11-1999	1953378
19	MOBILE ELECTRONICS	104852	22-12-2001	88394

[1] Only readable reviews have been taken into account

20	MOBILE APPS	5008069	04-11-2010	4764408
21	MAJOR APPLIANCES	96834	26-08-2000	68800
23	LAWN AND GARDEN	2555288	27-11-1999	2250074
24	KITCHEN	4874890	20-01-2000	4094402
25	JEWELLERY	1766992	10-11-2001	1564514
26	HOME IMPROVEMENT	2629867	08-11-1999	2356374
27	HOME ENTERTAINMENT	705487	15-10-1998	523578
28	HOME	6216756	29-05-1998	5540659
29	HEALTH PERSONAL CARE	5312733	06-02-1999	4432821
30	GROCERY	2393379	09-05-1999	1969970
31	GIFT CARD	148309	14-10-2004	135289
32	FURNITURE	791673	17-03-2000	717808
33	ELECTRONICS	3091024	09-06-1999	2597626
34	DIGITAL VIDEO GAMES	144724	08-08-2006	123656
35	DIGITAL VIDEO DOWNLOAD	3998345	04-10-2000	2658368
36	DIGITAL SOFTWARE	101836	26-01-2008	70682
37	DIGITAL MUSIC PURCHASE	1681484	28-06-2000	1250178
38	DIGITAL EBOOK PURCHASE(0)	12444454	09-09-2013	9717488
39	DIGITAL EBOOK PURCHASE(1)	5100425	28-08-1999	3941371
40	CAMERA	1800845	20-11-1998	1493426
39	BOOKS _0	10236850	03-05-2012	7388390
41	BOOKS _1	6106190	14-10-2005	2240142
42	BOOKS _2	3105370	24-06-1995	229335
43	BEAUTY	5094307	31-10-2000	4212222
44	BABY (PRODUCTS)	1749148	13-07-1999	1388798
45	AUTOMOTIVE	3510895	24-10-1999	3224707
46	APPAREL	5881874	16-12-2014	5290304

Table 1: Details of the Amazon product review datasets where all 46 products are listed in the first column, total number of readable reviews are listed in the second column, starting date of reviews

corresponding each product are in the third column, and total number of reviews corresponding verified purchase are listed in the last column.

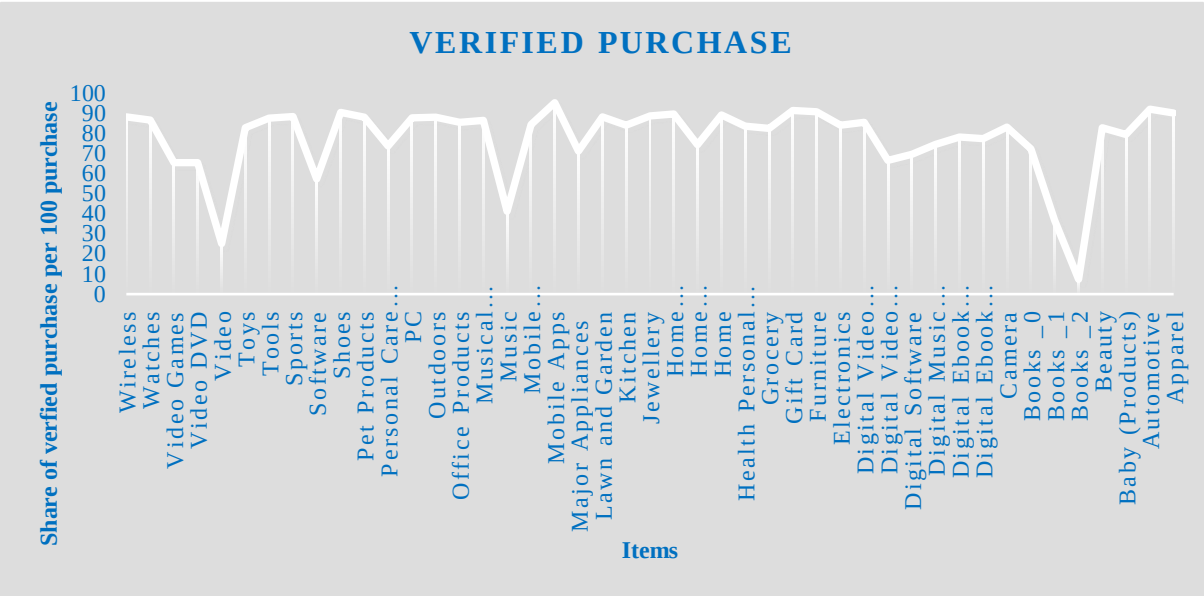

Figure 1: Illustration of the share of verified purchases per 100 purchases as listed in table 1.

In figure 1, we illustrate number of reviews in the verified purchase category in percentage of the total number of reviews of a product. Notice the dips in the curve of figure 1, representing the number of verified purchase per 100 purchase. These dips are pronounced for the products '*Video*' and '*Book_2*'.

Columns of product_title, product_id, and product_parent

For each product, based upon "*product_title*", "*product_id*" and "*product_parent*", there are different categories of sub-products. For instance, for "Softwares" there are 27741 types of sub-products with different *product_titles*. On the other hand there are 28717 types with different *product_ids*. This number is 28184 for the category *product_ parent*. The same hold true for other products.

Verified and Unverified Purchase

An analysis of "*star_rating*" categorised with verified and unverified purchases also presents an interesting picture. Mostly, there are $5-15\,\%$ reviews categorized with "*No*" for the column "*verified_purchase*". For instance, for *SHOES,* 3937353 number of reviews categorised as *verified_purchase* against 421467 *unverified_purchase*.We find another interesting fact while

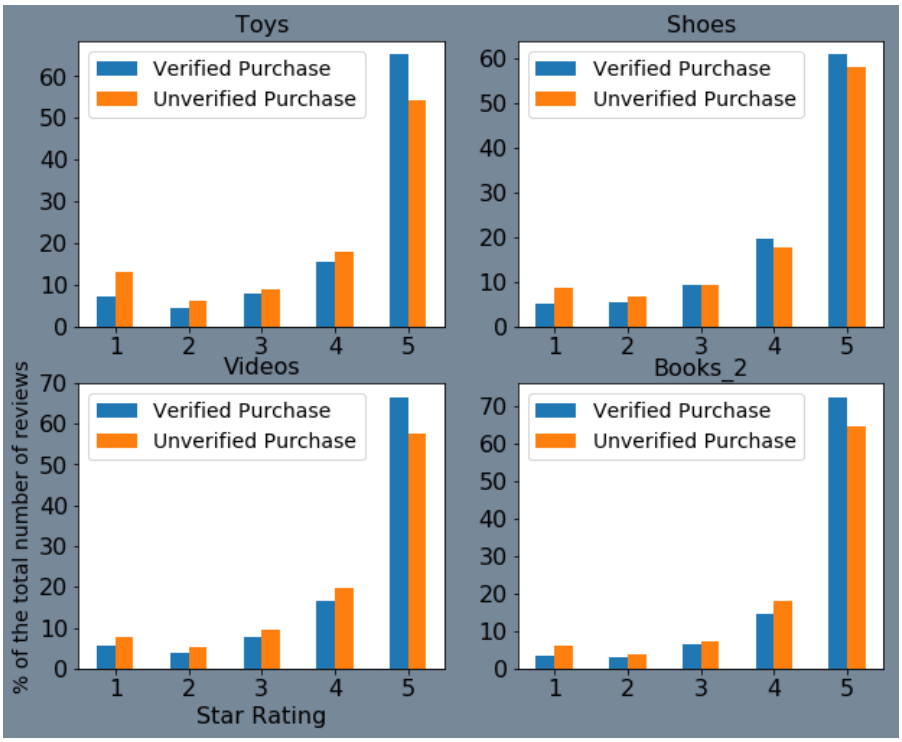

Figure 2: Category (a) as explained in the main text. Reviews of four products, viz. Toys, Shoes, Videos and Books_2 have been grouped with respect to their star ratings. Bars in blue and orange colors represents verified and the unverified purchase categories, respectively.

analysing data for the verified and unverified purchase categories separately: when reviews of *TOYS* are sectioned on the basis of *product_title*, there are 570539 sub-products out of 4017951 products corresponding the verified purchase category against 236172 sub-products out of 841657 products corresponding the unverified purchase category. Next, the order of the number of appearance of product titles change for the above two. For instance, for the verified purchase, *Card Against Humanity*, *Melissa & Doug*, *Card Against Humanity First* are the product titles which top the chart by appearing for 23124 times, 9947 times, and 5851 times, respectively, against *Melissa and DOUG*, *Fisher Price Ocean-Wonder* and *Card Against the Humanity*, appearing for 1704,1432 and 1141 times in the chart, corresponding the unverified purchase. Similarly for other products we find different ranking of the product for the two categories.

Tendency of Purchasing High Star-Rating Products:

It is natural for a buyer to choose online products with higher star ratings and the same should be reflected in the data. Commencing with the launch a product forms its reputation which increases with sale. Assume that each review represent a sale. Such quantification help to see the

relation of star rating with sale and time. There are two types of figures that we show in this section, viz. (a) Star rating vs percentage of reviews, and (b) percentage of reviews vs average star rating.

For the category (a), we use Toys review data. As shown here the peaks in both types of purchases, most reviews are either for 5-star rating or for 1-star rating. However, we clearly see a small bias towards the higher ratings; slightly higher in the verified-purchase category that unverified ones. We have obtained similar figures from the analysis of *Shoes, Camera*, *Softwares*,

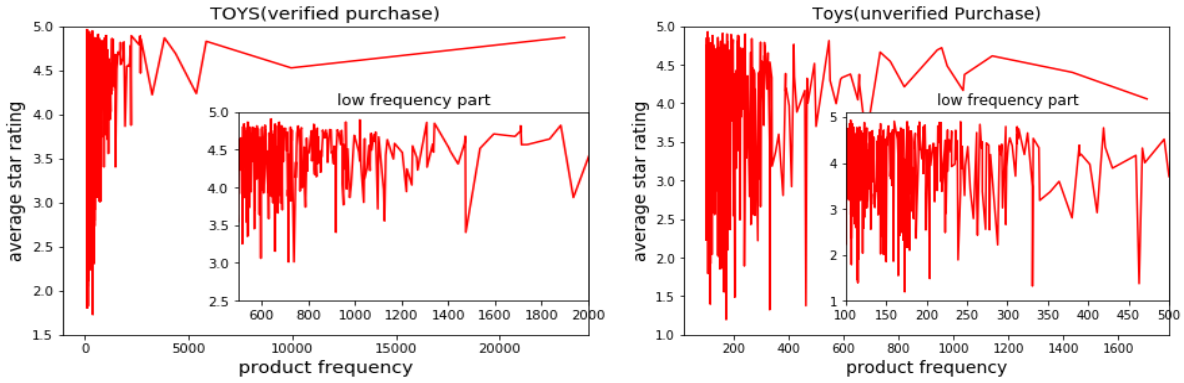

Figure 3: Average star rating vs product frequency, i.e., how many time a product categorized with respect to the product_title, has been reviewed. We show this analysis for Toys-reviews for the verified (left figure) and unverified (right figure) purchase. In the inset of both the figures we show the low frequency part.

Videos etc. However, for Book_2 we get the same bias but slightly more for 5-star and 1-star ratings for the unverified purchase category.

In category (b), we plot averaged review-star ratings against the number of times reviews of a product where the latter has been grouped with respect to the product title. For instance, for *TOYs* a product titled as "*Card Against Humanity*" has a sale of 23124 in the verified purchase category and on average it has a rating of 4.87 with standard deviation 0.49. This sale-number we refer to in this section as the frequency. On the other hand, for the unverified category "*Melissa & Doug*" topped the list with a frequency of 1704 and average star rating 4.056 with a standard deviation 1.34. In the figures below we show plots for the verified purchase and unverified purchase category, respectively. In the inset we show products with lower frequencies. Here frequency means that for how many times the product corresponding the product title has been reviewed. As can be seen in the plot, a decrease in the randomness for higher frequency products. These plots allude to a tendency towards high ratings for higher frequency product. Moreover, one

may also conclude the same for number of sales if frequency genuinely represents the former. This seems valid at least for the valid purchase category.

STATISTICAL PROPERTIES OF THE LOW FREQUENCIES

We expect some well-known statistical laws, such as Herdan's law and power-law, to hold for all four corpus of each product. The power law in the linguistic statistics has been a subject of interest for almost 80 years [Ziph 1936]. Altman and Gerlach quote Gustav Herdan's [1964] as"…'language in use' can not be studied without statistics". The LS has been demonstrated studied on some classic literatures such as *War and Peace, Pride and Prejudice, The Voyage of the Beagle*, etc. One the other hand, reviews are focused elaborating properties or aspects of a product to justify an opinion rated as star, while literatures usually assert upon details, description or discussion and stories. We expect differences in LS of the reviews, viz. in *review_headline and review_body corpus.* In this sense we may well expect a difference in parameter-values characterizing such laws. Besides, we need to derive a range using which one can demarcate generic properties against the specific ones for the further analysis.

THE HERDAN'S LAW

Herdan's law or Heap's law explains number of distinct words in a document as a function of the length of document increases. It is given by

$$d = \delta\, j^{\gamma},$$

for d number of distinct words in j number of reviews. Prefix δ and exponent γvalues are to be determined. We are interested to learn how the number of distinct words increases with the number of reviews d. In particular we are interested only in the stem-word count, completely ignoring the stop-words, as a function of j. As expected, if we remove the constraints of distinction among words, the rise is linear.

For Amazon reviews, we proceed as follows. We chose first j of reviews of a product and then section them in the verified (V) and unverified (UNV) categories. Next, we drop all *stop-words*, such as 'is', 'of', 'it' etc., and create a 'bag of words'. This is a standard practice. Further, we choose all in one bin, distinct ones in another bin, and count how the size d (or the word-count) of these bins increase with respect to j. We use Porter Stemming algorithm and then construct the matrix X. By construction the number of columns of the matrix is the number of distinct words. Sum of all matrix elements provides total number of words.

As shown below, for Shoes review corpus for BV and HV, the word-count rise seems consistent with the form given above up to $j \sim 17000$. As shown in the top left (BV) side figure

the rise is closely described by $\gamma \sim 0.4$ while it almost linear when all the words are considered. In the bottom figure $\gamma \sim 0.5$ represents the same analysis for HV. Here we have chosen $n = 100000$ reviews at random however the result is consistent with the ordered reviews.

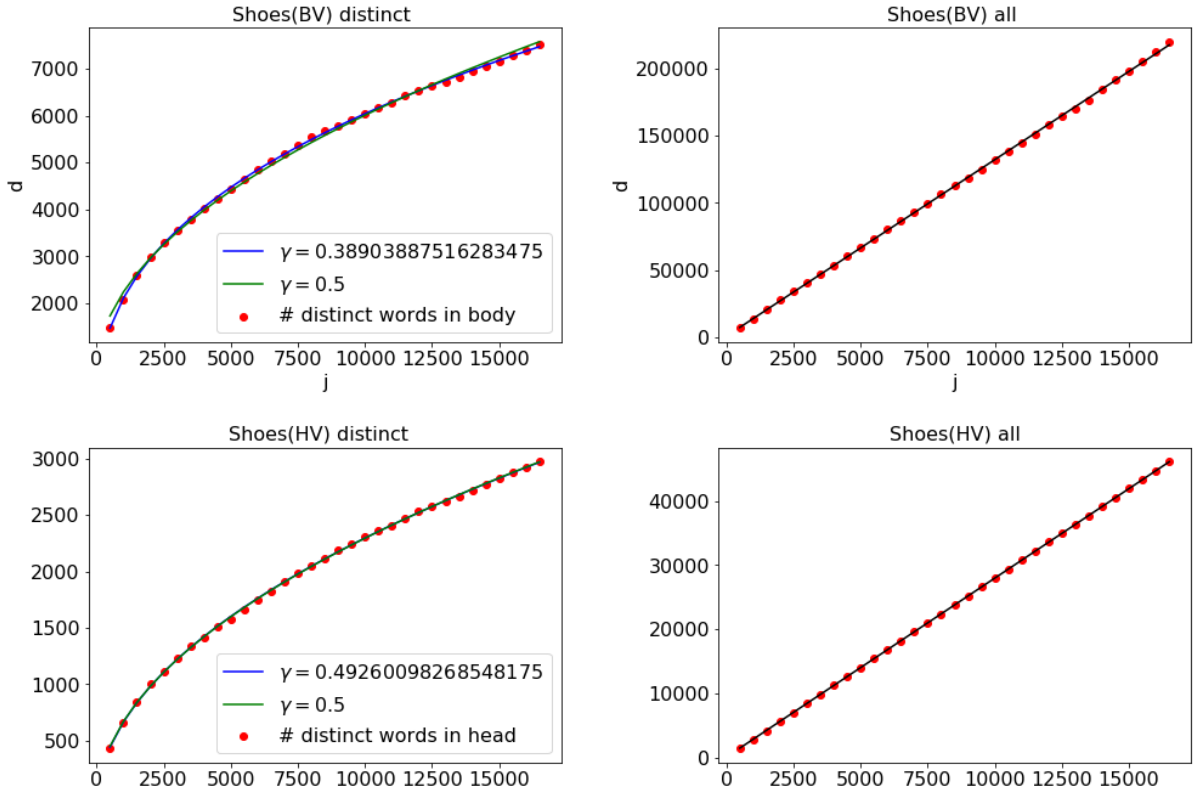

Figure 4: Increase of the number of words,d d, vs number of reviews, j, for the verified_purchase of Shoes. In the left panel d represents number of distinct words while in the right panel it represents total number of words. Top panel shows this analysis for the review_body and bottom panel is for the review_head. Notice that the exponent, γ, value in the left panel is close to 1/2 for the review_head.

The same analysis has been repeated for the reviews corresponding unverified purchase. This is shown below in figure 4. While, qualitative results are consistent with the sequentially chosen reviews for the verified purchase, inconsistency in the BUNV and HUNV cases are quite noticeable. For the random case we see a hump which is present also in the first 17000 UNV reviews near $j \sim 12500$. This hump is not visible in the next 17000 UNV reviews chosen in increasing order.

We have also done this analysis for the verified and for the unverified purchases of toys. For the verified purchase category we find no difference between the sequentially chosen and randomly chosen reviews. However for the unverified category we see the inconsistency here as

well. However, we find this feature less prominent for the toys. We find the same γ-values matched with the *Shoes* case for the BV as well as for the HV in ~3% of relative error.

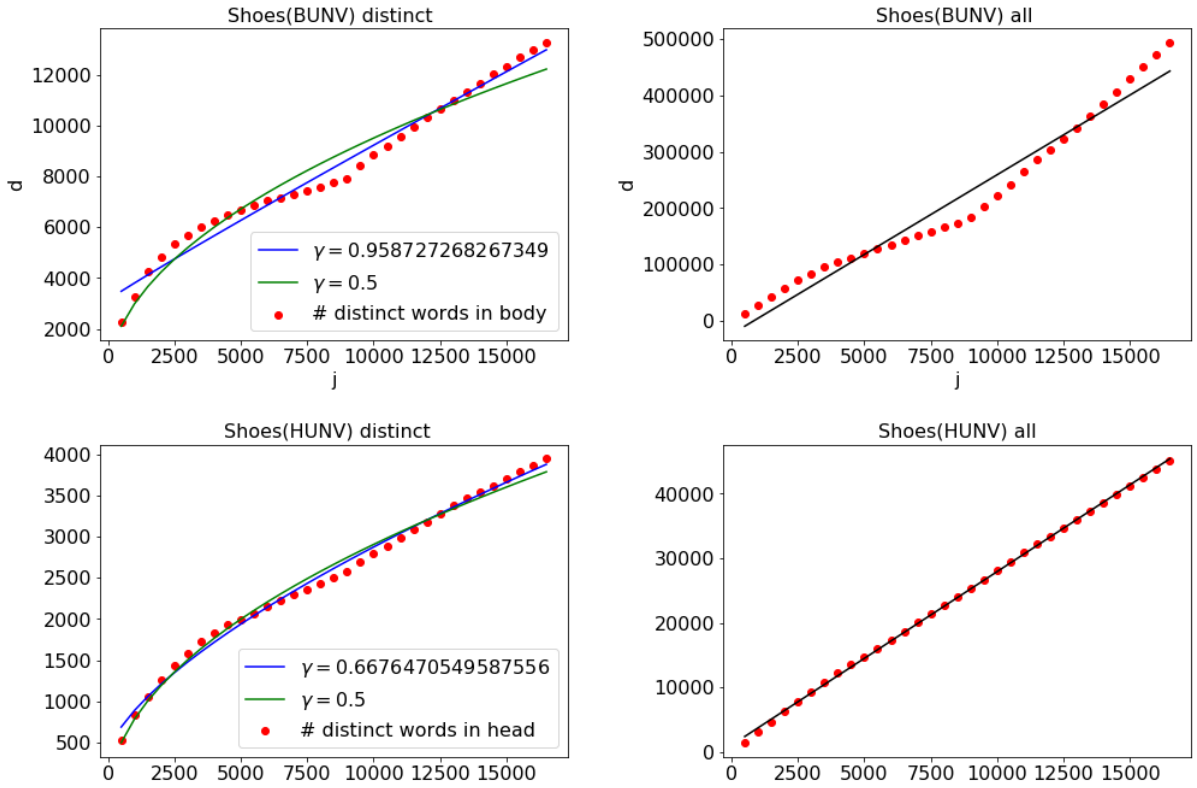

Figure 5: Same as the previous figure but for the unverified category.

POWER LAW

Power law has a long history with linguistic statistics. One end goes with the Zipf rank plot which states that the frequency f of the r'th most frequent word is inversely proportional to its rank: $f \sim r^{-\alpha}$. Since frequency itself will depend upon the size of the corpus, one need to define a scale starting from 1. For example, if it starts with the top most frequent word with 1 then for the second most frequent word f is proportional to $2^{-\alpha}$, and $3^{-\alpha}$ for the third most frequent, and so on. It was later advocated by Mandelbrot argued that an improved distribution must be of the form $\sim(r+\beta)^{-\alpha}$. Plenty of literature have been devoted on this topic. Power law fit usually demonstrated on a log-log plot where it is seen as a straight line. The success of the fit is usually justified by visualization. This technique has been popular amongst scientist until its validity and mathematical rigor have been questioned critically. It has a valid ground as the power law which in most cases appears a heavy tail is badly affected by the statistics. Even worse is the fitting on

log-log axis where the variance is shown maximum in the farther tail. Through examples of some well-studied distribution, e.g. power-law distribution in continuous case or in discrete case, it has been shown that instead of the least-square fitting the maximum likelihood approach yields a more reliable result. Besides, there remains issues with the uncertainty associated with the histogram which may be corrected to some extent.

We are rather interested in the frequency distribution. First, we choose n number of reviews randomly from all m reviews of a product, where $n < m$. Then the review are categorised with respect to purchase. We create an ensemble of such samples. For each sample we order words with respect to their frequency in the sample. We do not expect that the sequence be the same, especially for smaller frequencies. Next, instead of the most frequent words we consider the least frequent ones and plot how many times these frequencies have appeared in the sample. Explicitly we evaluate the ensemble averaged number of the appearance of the frequency, p, as function of the frequency, q, for $1 \leq p, q \leq k$ where $k \leq L$ and L is the corpus size (size of the smallest sample). We assume that the p's for any two frequencies are independent of each other. We expect that the fit function should be a straight line on the log-log axis, as

$$p = \varepsilon\, q^{-\theta},$$

where p, q are positive integers and the parameters, ε and θ define the power-law. We consider the least-square fitting method to fit this function with data. We consider all four VB, BUNV, HV, and HUNV corpus of the *Toys*, and *Shoes*. We choose $n = 50000$ randomly for 400 times and for 1000 times, respectively for *Toys* and *Shoes* reviews. Consider k number of words corresponding a frequency q on average with variance v. We find that for the top k values, the distribution is closely described by Gaussian with mean k and variance v. For instance, for BV corpus of *Shoes*, we find on average ~ 28860 words, with standard deviation 246 in the sample of size 500, having frequency 1. Distribution of this number and numbers corresponding frequencies is closely described by Gaussian. Below we show our result for *Toys* and for *Shoes* in side by side figures. In these figures we have $\sum_{q}^{maxim\_q} p(q) = 1$, $maxim\_q$ is the maximum value of q in the corpus. The exponent values are given in the same figure. It is evident that power describes well the low frequency behavior. Clearly for both products exponent values are almost same for the BV and BUNV corpus while HV and HUNV corpus are described by different exponents. We also have performed the Kolmogorov-Smirnov test for each fit with running index for the number of points

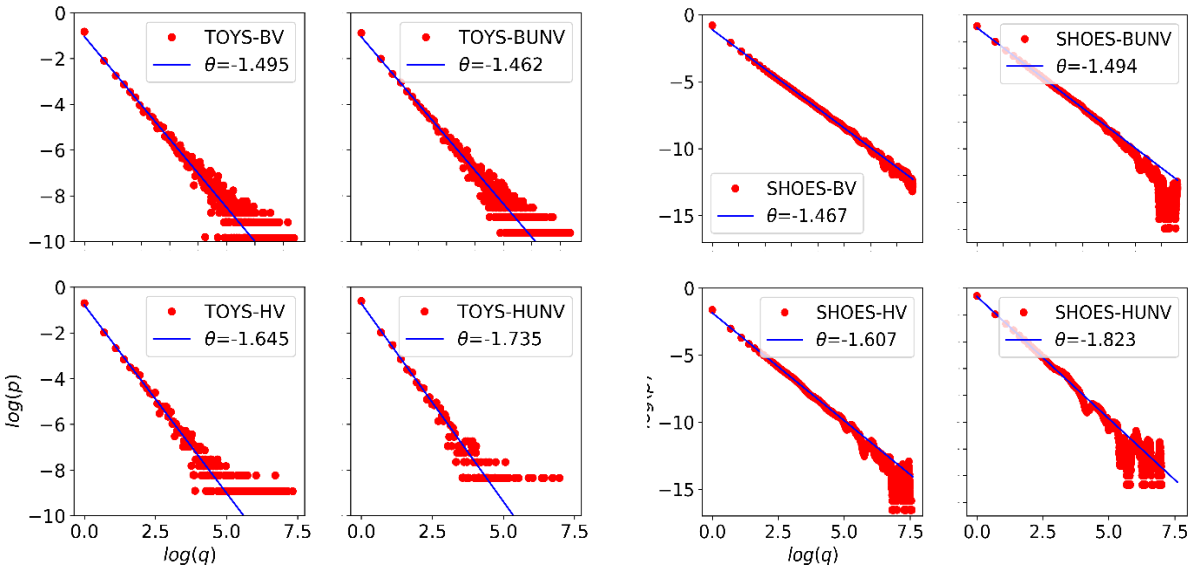

Figure 6: Log-Log plot illustrating power-law for the number of appearance of the frequency, p, vs frequency, q, of words. In the left and right panel we illustrate results for the toys and shoes respectively. All four corpus results are shown in each panel viz. verified and unverified purchase categories respectively in the left and right part and review body and headlines respectively in top and bottom panels.

starting from 1. For instance, in the table 2, for the p vs q plot of the BV corpus of Shoes, we enlist the maximum frequency $qmax$ up to which the curve has been fit with the power law, the KS-statistic vale, the p-value, and fitting parameters, θ and ε respectively from left to right columns.

qmax	KS-Statistic	p-value	θ	ε
64	0.03125	1	1.46667	1.11997
63	0.031746	1	1.46743	1.11819
62	0.032258	1	1.46798	-1.1169
31	0.032258	1	1.50706	1.03915
61	0.032787	1	1.46884	1.11488
60	0.033333	1	1.46992	1.11239
30	0.033333	1	1.50896	1.03583
118	0.033898	1	1.45606	1.14724
59	0.033898	1	1.47046	1.11114

117	0.034188	1	1.45649	1.14598

Table 2: Table listing fitting parameters for p vs q *plot. The leftmost column is for the qmax, i.e., followed in right by maximum frequency up to which the function is fit to data starting from* $q = 1$. *In the right-increasing order columns represent the corresponding KS-statistic,p-value and the parameters* θ *and* ε, *respectively.*

It is also interesting to note that contribution of the number of word with frequency q to the total size of the corpus also described by a power law for small frequencies. For instance, let r represents the contribution of the frequency q and we rescale $\sum_{q}^{maxim\_q} r(q) = 1$. Then we expect that for small frequencies,

$$r = \rho\, q^{-\sigma},$$

which is to be estimated by fit. It is not difficult to deduce that

$$r = \alpha\, p^{-\beta},$$

where $\beta = \sigma / \theta$ and $\alpha = \rho\varepsilon^{-\sigma/\theta}$. The last function informs about the contribution of number of appearance of the frequency to the corpus where the number of the appearances are ordered with respect to frequency. Below in table 3 we compare values of β for different corpus calculated from the above relation and β from the fitting of the corresponding data with the power law. A nice agreement between the two is evident.

Corpus	σ	θ	β-theory	β-result
BV	0.4688527	1.46667	0.3196713	0.337838654
BUNV	0.5171551	1.49445	0.3460494	0.341221211
HV	0.6042803	1.60718	0.3759874	0.381938775
HUNV	0.8246143	1.82274	0.4524046	0.452331923

EXPECTANCY, DISTINCTION AND WEIGHT OF HIGH FREQUENCY WORDS

By *Expectancy* we mean that how many times a word is expected to appear in one review in a corpus. Similarly *Distinction* is described as how many times a word appear per distinctive word, and *Weight* is defined as how many times a word appear per word. Using frequency of a word, i.e. the number of time a word appears in n number of reviews, $f(w,n)$, we define the above three quantities. We define i, as

$$i_k = f\ (w,n)/k,$$

where k is n, $N(n)$, or $M(n)$, defining respectively the expectancy, distinction and weight. We evaluate these by choosing n number of reviews at random for l times and for each of such realization we calculate $i_k, M(n)$ and $N(n)$. We finally perform the ensemble averaging and refer to the evaluated number as *Averaged Expectancy* (I_n), *Averaged Distinction*(I_N) and *Average Weight* (I_M). Notice that larger will be the expectancy for larger I_n-*values* but smaller will be distinction for larger I_M. Consider the matrix X where matrix columns are defined by distinct words and matrix rows represents the j'th chosen review. As mentioned above quantity I_M is the measure with respect to the total "*weight*" of the matrix and I_n describes sparsity along each column. I_n measures how repetitive a word is with respect to reviews while I_N measures that how distinctly a word is chosen in the corpus. I_M describes contribution of the word in the corpus and thus named as weight.

To understand sparsity of X further we need to evaluate sum along the rows. However, high sparsity is expected row-wise as few words are available for each reviews and it increases with the number of reviews. So we suspect that for sufficiently large number of reviews matrix X might be close to a highly-singular-structure.

In figures (8), in the top we illustrate I_n, I_N , and I_M for *Toys* review-body corpus and compared the verified and unverified categories. We have chosen top 48 words from a sequence ordered with respect to I_n. The number 48 is our choice. As seen from the curves, reviews for both categories have differences for almost all words. For I_M, however, the two almost overlap. For I_n, the BUNV graph lies above the BV graph. It explains that for the first category, for verified (V) category, words are less frequently chosen. On the other hand the role reversed in I_N where

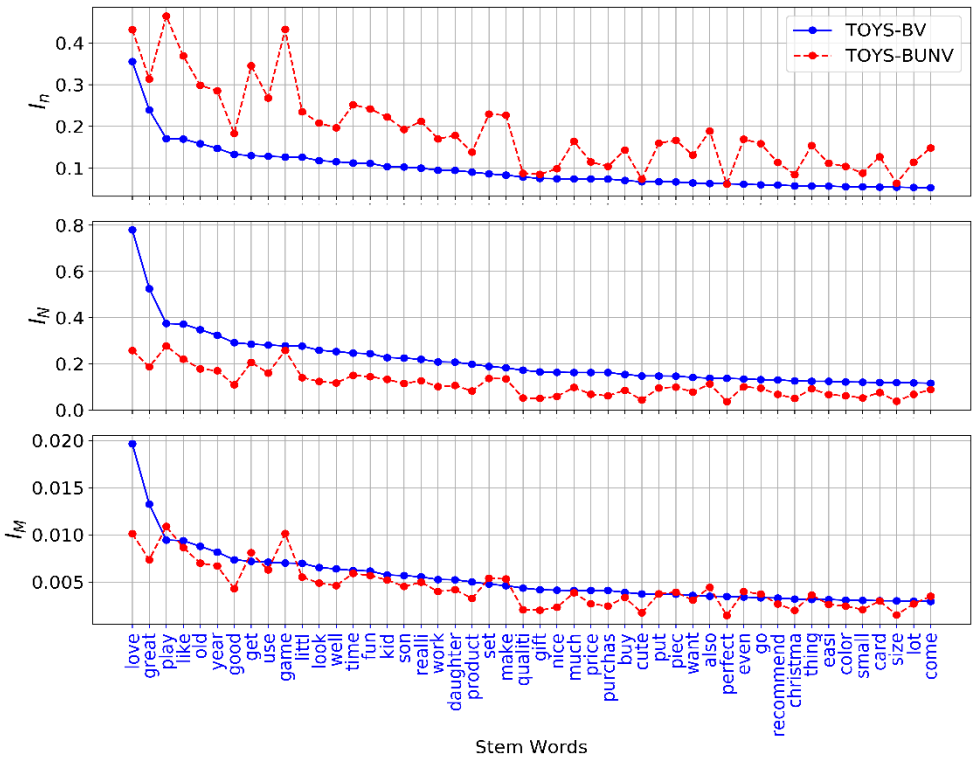

Figure 8: Expectancy, viz I_n (top), distinction I_N (middle), and weight I_M (bottom), of words,. Words are ordered with respect to I_n from the review body of Toys. In blue we show results for the verified purchase and in red we show results for the unverified purchase.

the second category, i.e. for the unverified (UNV), lies below the first category. Combining the two figures we may draw a conclusion that for the V, words are chosen less frequently and less distinctly than UNV. The top figure explains that on average *love* is the most used word in all reviews followed by *great*, *one, play, like.* Words like *play, year, fun* seem neutral and product specific.

In figure (9), we show the same analysis for the review-headline corpus. As can be seen here the two categories almost overlaps for I_n, and I_M but for I_N. Most favoured and the second most favoured words viz., *love* and *great* have their rank interchanged. Apart from these two, *play, like, old, year, good, get, use, game, little, look, well, time, fun, kid, son, really, work, daughter, product, set, quality, gift, nice, much, price, buy, cute, perfect* are the common words with different rankings between the two sets.

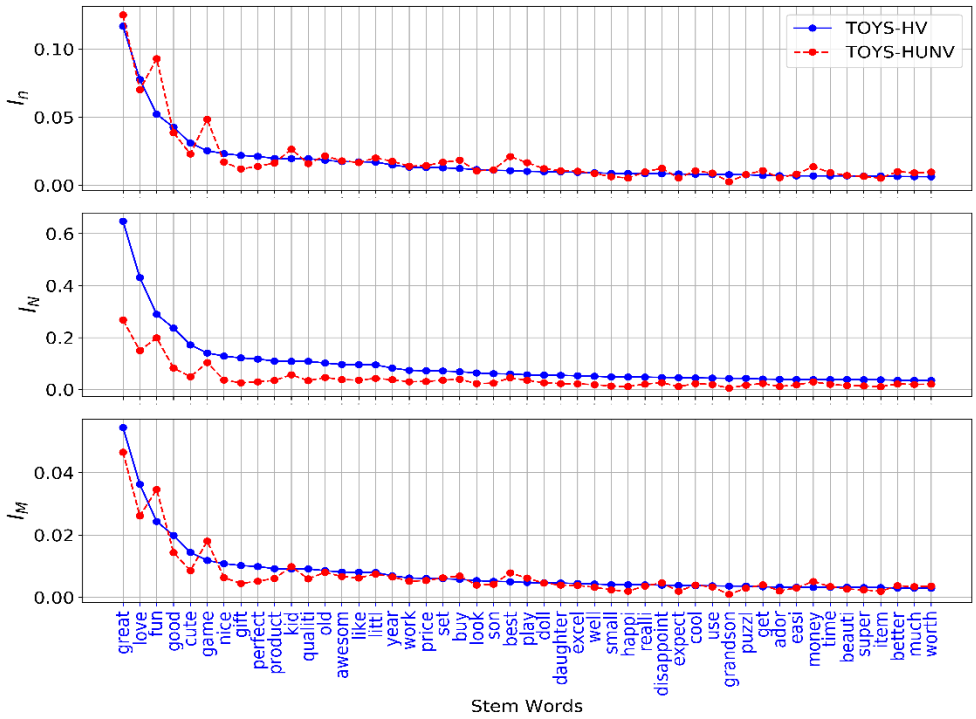

Figure 9: Same as the figure 8 but for the HV and HUNV corpus generated from the Toy-reviews dataset.

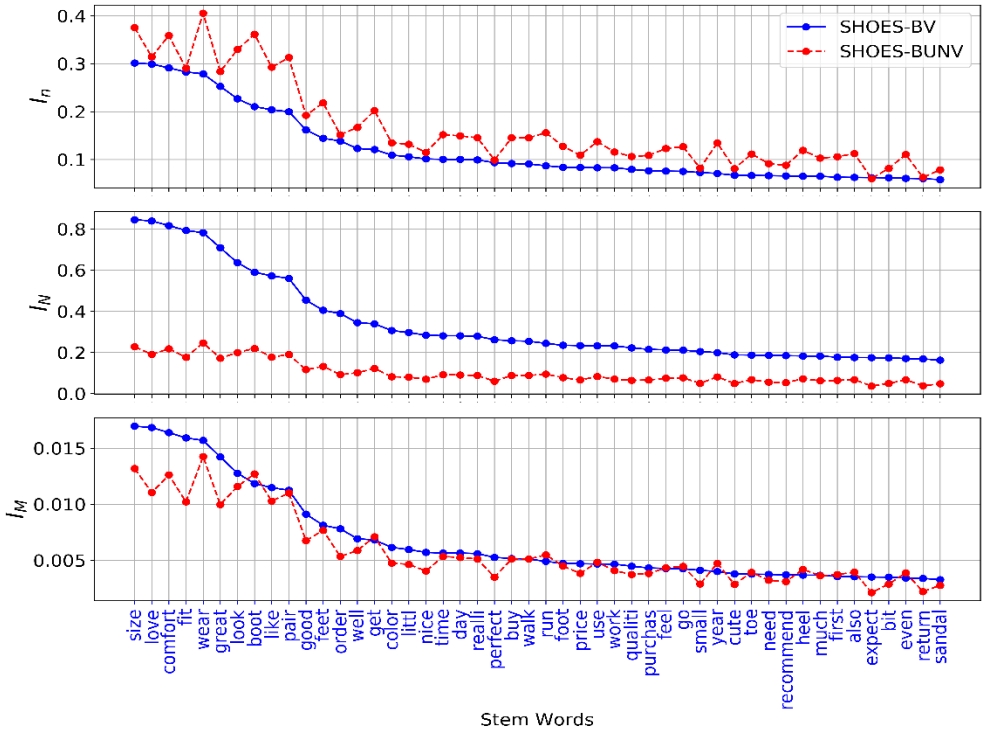

Figure 10: Same as figure 8 but for the BV and BUNV corpus generated from the Shoes-review dataset.

In figure (10) we show result for the review-body corpus of *Shoes*. For the expectancy we

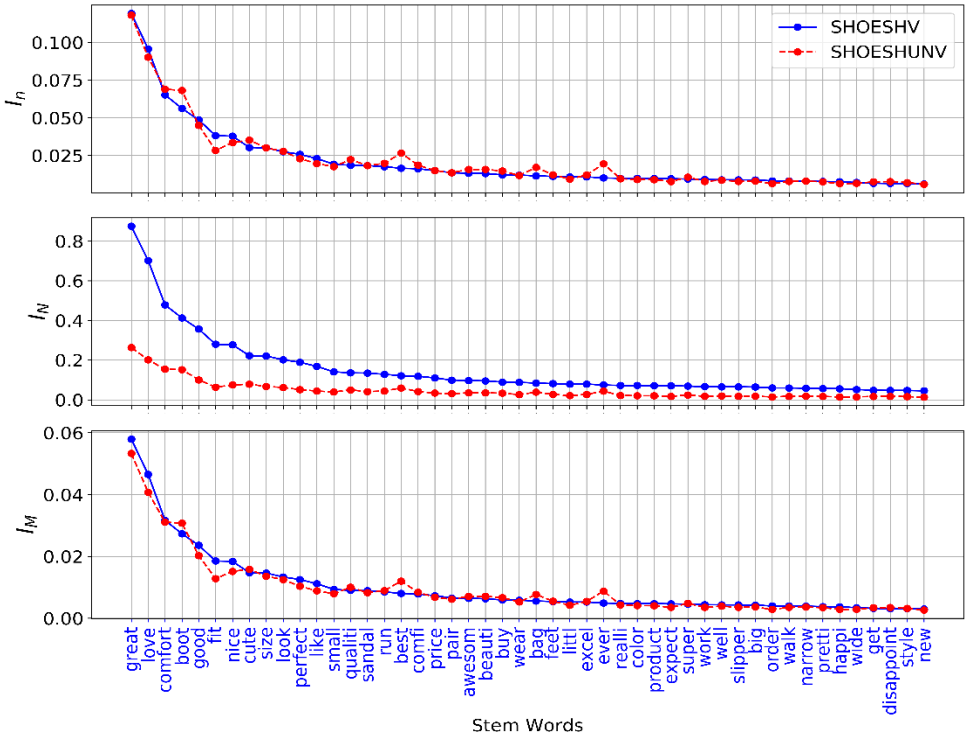

Figure 11: Analysis of HV and HUNV corpus. Results are shown in the same pattern of the previous three figures.

fine the same qualitative behavior as for the toy case however it is distinction where difference is more visible. It is clear that for shoes we do not see a rapid fall in expectancy for the verified purchase while for the unverified purchase the graph is similar to that for the toys. Weight is almost on the same pattern as for the toys. For the review headline we find a very close overlap of the two categories in expectancy and weight measures, however a clear mismatch in the distinction measure. Finally, in figure (11) we show results for the analysis of HV and HUNV corpus generated from the Shoes-review dataset. As for the toys, here as well find I_n and I_M overlapped for HV and HUNV but I_N show differences.

Amazon Data and License Policy for Academic Use

We have considered Amazon (US) data available online as amazon-reviews-pds S3 bucket in AWS US East Region. This data is subjected to non-commercial use. All reviews are in English, stored in Tab Separated Values (TSV) format, for 46 different products. All datasets are available with a license permitting their use only for academic purpose. Further details may be traced by hitting the link in the text highlighted in blue.

DISCUSSION

In this report we have studied a type of secondary datasets available in public domain. The basic idea of this work to analyse such datasets in order to develop understanding of the complexity associated with the data, technicalities and statistical properties. We aimed towards developing a general approach to understand reviews, available in bulk, within a framework of limited words with quantification as will be required for comparison and analysis. We have considered four types of corpus generated from a product review dataset and analysed two type of words viz., words that appear frequently and words which rarely appear. For the first type of words we have introduced three measures, viz., *Expectancy*, *Distinction* and *Weight*. We have explicitly compared these measures for the products corresponding verified and unverified categories and demonstrated the differences. These measures have been evaluated and compared for the four corpus separately. Highlights of our analysis are

- Star ratings for the verified purchase category is more bias towards high ratings that for the unverified purchase category.
- The power law exponent corresponding the review-headline are larger than that for the review-body, for small frequencies, implying thereby a wider-small-frequency spectrum for the former and sharper decline in their contribution in the corpus. The first remark comes from the comparison of p vs q plots of the respective corpus and the second remark comes from the comparison of r vs q plots.
- Although with the same contribution in the corpus, words in the BUNV are chosen more distinctly than for the BV. The same comparison for HV and HUNV assert the same.
- Although with the same contribution in the corpus, words in BUNV are more repetitive than in BV. For HUNV and HV words are almost equally repeated.
- The same word has different ranks in the review body and headlines, and thus underlines requirement of further analysis.

Disclaimer: This report is largely descriptive in nature. It is based on secondary data and information composed from the data available in public domain, and the location of items may change as menus and webpages are reorganized.