The Enrollment Effect: A Study of Amazon's Vine Program

Dinesh Puranam
Samuel Curtis Johnson
Graduate School of Management
Cornell University
dp457@cornell.edu

Claire Cardie
Department of Computer Science
Department of Information Science
Cornell University
cardie@cs.cornell.edu

Abstract

Do rewards from retailers such as free products and recognition in the form of status badges[1] influence the recipient's behavior? We present a novel application of natural language processing to detect differences in consumer behavior due to such rewards. Specifically, we investigate the "Enrollment" effect, i.e. whether receiving products for free affect how consumer reviews are written. Using data from Amazon's Vine program, we conduct a detailed analysis to detect stylistic differences in product reviews written by reviewers before and after enrollment in the Vine program. Our analysis suggests that the "Enrollment" effect exists. Further, we are able to characterize the effect on syntactic and semantic dimensions. This work has implications for researchers, firms and consumer advocates studying the influence of user-generated content as these changes in style could potentially influence consumer decisions.

1 Introduction

In 2007 Amazon introduced its Vine program[2]. According to Amazon, "Amazon invites customers to become Vine Voices based on their reviewer rank, which is a reflection of the quality and helpfulness of their reviews as judged by other Amazon customers. Amazon provides Vine members with free products that have been submitted to the program by participating vendors. Vine reviews are the "*independent opinions* of the Vine Voices."[3] There could be potential concerns as to whether this enrollment affects the way reviews are written, introducing, for example, a positive bias.[4]

In this work, we investigate whether enrollment in the Vine program results in changes in the linguistic style used in reviews. We investigate this by looking at reviews by individuals before and after enrollment in the program. Following Feng et al. (2012) and Bergsma et al. (2012), we conduct a stylometric analysis using a number of syntactic and semantic features to detect differences in style. We believe that detecting changes in consumer behavior due to intervention by a firm is a novel natural language processing task. Our approach offers a framework for analyzing text to detect these changes. This work is relevant for social scientists and consumer advocates as research suggests that product reviews are influential (Chevalier and Mayzlin, 2006) and changes in style could potentially influence consumer decisions.

2 Related Work

Our work lies at the intersection of research in four broad areas — Product Reviews, Product Sampling, Status and Stylometry.

Product Reviews Product reviews have received considerable attention in multiple disciplines including Marketing, Computer Science and Information Science. Research has addressed questions such as the influence of product reviews on product sales and on brands (Gopinath et al. (2014); Chevalier and Mayzlin (2006)), detection of deceptive reviews (Ott et al., 2011) and sentiment summarization (Titov and McDonald, 2008).

[1] A status badge is a special identification usually placed next to a username in online content.

[2] http://blog.librarything.com/main/2007/08/amazon-vine-and-early-reviewers/

[3] http://www.amazon.com/gp/vine/help, words italicized by authors.

[4] http://www.npr.org/blogs/money/2013/10/29/241372607/top-reviewers-on-amazon-get-tons-of-free-stuff.

Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pages 17–27,
Baltimore, Maryland USA, 27 June 2014. ©2014 Association for Computational Linguistics

This list is by no means comprehensive, but it is indicative of the extensive work in this domain.

Product Sampling Here, consumers receive products for free — as a marketing tactic. This is also a well-studied phenomenon. Research in this area has indicated that consumers value free products (Shampanier et al. (2007); Palmeira and Srivastava (2013)); that product sampling affects brand sales (Bawa and Shoemaker, 2004) and that sampling influences consumer behavior (Wadhwa et al., 2008).

Status Research shows that status can influence writing style. Danescu-Niculescu-Mizil et al. (2012) study discussions among Wikipedia editors and transcripts of oral arguments before the U.S. Supreme Court and show how variations in linguistic style can provide information about power differences within social groups.

Stylometry focuses on the recognition of style elements to identify authors (Rosen-Zvi et al., 2004), detect genders and even determine the venue where an academic paper was presented (Bergsma et al., 2012).

Our work draws from each of these research areas and in turn hopes to make a contribution to each in return. Our primary objective is to establish a framework to detect behavioral change due to a decision by a firm (in this case enrollment to the Vine program characterized by free products and Vine membership status) by analyzing product reviews. Further, we hope to understand the dimensions on which this behavior may have changed. Consequently, we pursue a novel stylometric task. This type of work is especially important when the traditional numerical measure (rating) suggests there is no difference in the review pre and post-enrollment (see Section 4).

3 Data & Pre-processing Steps

We gathered all reviews by the top 10,000 reviewers ranked by Amazon as of September, 2012. These rankings are partly driven by helpfulness and recency of reviews[5]. The data collected includes the review text, review title, rating assigned, date posted, product URL, product price, whether the reviewed product was received for free via the Vine program (also referred to as "Vine Review"), "helpfulness" votes and badges received by the reviewer .

We collected a total of 2,464,141 reviews of which 282,913 reviews were for products received for free via the Vine program. These reviews covered a total of 9,982 reviewers[6] of which 3,566 were members of the Vine program. Approximately half the reviews belonged to Vine members. We eliminated reviews that did not have a rating. We further excluded reviews where the review text was less than 20 words in length. We were left with 1,189,704 reviews by Vine members.

The date of enrollment to the Vine program for each reviewer is not explicitly available. We infer the date of enrollment in the following manner. We sort in ascending order all the "Vine Reviews" for each reviewer by posted date. We assume the earliest posted date for a "Vine review" is the enrollment date. This is an important assumption, as potentially reviewers could have moved in and out of the program at varying points of time. Reviewers can be moved out of the program for reasons such as not posting a "Vine Review" within 30 days of receipt of the product. In our data set we found 47,510 "Vine Reviews" by 163 reviewers who were not actively on the Vine program [7]. We can view these reviewers as having been dropped from the Vine program. Given the small volume of this type of reviews and reviewers, our assumption on date of enrollment appears reasonable.

Member Type	Free/ Paid	Enrollment Timing	Review Count
Non Vine	Paid	NA	1,169,561
Non Vine	Free	NA	47,510
Vine	Paid	Post	452,729
Vine	Paid	Pre	503,688
Vine	Free	Post	233,287

Table 1: Data Summary

4 Enrollment Effect

This research seeks to answer the question: does enrollment in the Vine program change the writing styles of reviewers. One naive theory is that

[5] http://www.amazon.com/review/guidelines/top-reviewers.html/

[6] During the crawling, ranks changed resulting in fewer than 10,000 reviewers in our data set.

[7] As these reviewers were not enrolled to Amazon's Vine Program as of September, 2012, they are excluded from our analysis.

perhaps receiving products for free and receiving status badges will result in Vine members posting more positive reviews. Interestingly, the average rating for reviews by Vine members posted before enrollment is 4.22 and after enrollment is 4.21 and this difference is not statistically significant. In contrast, the length of reviews significantly increased from 251 words prior to enrollment to 306 words post-enrollment. Natural language techniques are the only option to further investigate possible effects of enrollment. Consequently we focus on the review text posted by Vine members.

4.1 Approach

Following Ashok et al. (2013) and Bergsma et al. (2012) we construct features that represent writing style from each review (discussed in more detail in the next section). We incorporate these features in a classification algorithm that attempts to classify each review as having been written pre or post-enrollment to the Vine program. We report whether the difference in accuracy for this classifier vs. a majority vote classification is statistically significant or not. In order to detect differences in style pre and post-enrollment, we need to address certain confounding factors — Reviewer Specificity , Product Specificity and Time Specificity.

Reviewer Specificity It may be possible that certain users post more reviews post-enrollment than pre-enrollment. Consequently the classifier may simply end up learning the differences in style *between* reviewers. To avoid this, we construct a balanced sample where we randomly select 25 reviews for each reviewer prior to and post-enrollment (see Table 2). This also sets our baseline accuracy at 50%.

Product Specificity As the program started in 2007, the post-enrollment reviews are likely to predominantly contain products released in after 2007. This might result in the classifier simply learning the differences *between* products (say I Phone vs Palm). Given our focus on style, we do not use word tokens as such - thus avoiding the use of product specific features. However, for some products, the product specific details may result in the use of specific syntactic structures. We assume this is not a significant contributor to the prediction performance. A post-hoc analysis of the top features supports this assumption. A second source of change in writing style could be due to simply whether the product was bought or received for free. We exclude "Vine Reviews" [8] to eliminate this confounding factor.

Time Specificity A similar concern as *Product Specificity* exists for date references. By focusing on syntactic and semantic style, we avoid the use of time specific features.

Another concern is that perhaps post enrollment, reviewers receive writing guidelines from Amazon. This does not appear to be the case, as the writing guidelines [9] appear to be for all members.We now turn to the extraction of style features.

Data Type	Number of Reviews	Number of Reviewers
Training	113,250	2,265
Test	2,500	50

Table 2: Experiment Data

4.2 Feature Extraction

We consider three different features — "Bag of words/ unigrams", "Parse Tree Based Features" and an umbrella category consisting of genre and semantic features (see Section 4.2.3).

4.2.1 Bag of Words

Bag of Words/Unigrams (UNIGRAMS) Unigrams have often been found to be effective predictive features (Joachims, 2001). In our context, this serves as a competitive baseline for the classification task.

4.2.2 Parse Tree Based Features

Following Feng et al. (2012) and Ashok et al. (2013) we use Probabilistic Context Free Grammar (PCFG) to construct a parse tree for each sentence. We then generate features from this parse tree and aggregate features to a review level.

All Production Rules (Γ) This set of features include all production rule features for each review, including the leaves of the parse tree for

[8] Reviews where product was received for free via the Vine program.

[9] http://www.amazon.com/gp/community-help/customer-reviews-guidelines

each sentence in the review. This effectively represents a combination of production rules and unigrams as features and represents an additional competitive baseline.

Non Terminal Production Rules (Γ^N) This excludes the leaves and hence restricts the feature set to non-terminal production rules. This allows us to investigate purely syntactic features from the text.

Phrasal/ Clausal Nodes (PHR/CLSL) We also investigate features that incorporate phrasal or clausal nodes of the parse trees. Please see Table 5 and Table 6 for examples of these features.

Parse Tree Measures (PTM) We construct a set of measures for each sentence based on the parse tree. These measures are maximum height of parse tree, maximum width of the parse tree and the number of sentences in each review.

Latent Dirichlet Allocation (LDA) We also apply Latent Dirichlet Allocation (Blei et al., 2003) to the production rules extracted from the Probabilistic Context Free Grammar. We use the topics generated as features in our prediction task. Our objective was to determine whether certain co-occurring production rules offered better classification accuracy. Our implementation includes hyper-parameter optimization via maximum likelihood. The number of topics is selected by maximizing the pairwise cosine distance amongst topics. We used the Stanford Parser (Klein and Manning, 2003) to parse each of the reviews and the Natural Language Toolkit (NLTK) (Bird et al., 2009) to post process the results.

4.2.3 Genre and Semantic Features

Style Metrics (STYLE) This includes three distinct types of metrics. *Character Based* - This includes counts of uppercased letters, number of letters, number of spaces and number of vowels. *Word Based* - This includes measures such as number of short words (3 characters or less), long words (8 characters or less), average word length and number of different words. *Syntax Based* - This includes measures such as number of periods, commas, common conjunctions, interrogatives, prepositions, pronouns and verbs.

Parts of Speech (POS) features have often been surprisingly effective in tasks such as predicting deception (Ott et al., 2011). Consequently we test this feature set as well.

Domain-independent Dictionary We make use of the Linguistic Inquiry and Word Count (LIWC) categorization (Tausczik and Pennebaker, 2010). One key advantage of this categorization is that it is domain independent and emphasizes psycho-linguistic cues. We run two variants of this set of features. The first (LIWC ALL) includes all the categories — both sub-ordinate and super-ordinate categories. The second (LIWC SUB CATEG.) only includes the sub-ordinate categories, thus ensuring the features are mutually exclusive.

Subjectivity Measures (OPINION) We measure number of subjective, objective and other (neither subjective nor objective) sentences in each review. We use the "OpinionFinder System" (Wiebe et al., 2005) to classify each sentence with these measures. We aggregate the count of subjective, objective and other sentences at the review level and use these aggregates as features.[10] We also report results on experiments where multiple feature types are included simultaneously in the model.

5 Experimental Methodology

All experiments use the Fan et al. (2008) implementation of linear Support Vector Machines (Vapnik, 1998). The linear specification allows us to infer feature importance. We learn the penalty parameter via grid search using 5 fold cross-validation and report performance on a held-out balanced sample of reviews from 50 randomly selected users (all of whom were excluded from the training set) from the group of reviewers with at least 25 reviews in pre and post enrollment periods. While reporting the results, for some features we report the threshold (Thr) value set to exclude the least frequent features. These thresholds were also learned via the 5 fold cross validation process. Finally, text features can be binarized, mean centered and/or normalized. Each of these options were also selected via 5 fold cross validation.

6 Results & Analysis

All of the feature sets perform statistically better[11] than a majority vote (50%).

Baselines Unsurprisingly, the feature set containing all production rules (Γ) yields the best ac-

[10] One drawback is that the classifiers are trained on sentences from the MPQA corpus. Domain specificity is likely to yield poorer classification performance on our data.

[11] as indicated by a paired t-test at p=0.05 on the held out sample

Baselines		
Style Features	**Feature Count**	**Accuracy**
Unigrams	796,826	60.9 %
Γ (Thr =50)	29,362	62.0 %

By Feature Type		
Style Features	**Feature Count**	**Accuracy**
Γ^N (Thr=200)	2,730	59.2 %
Phr/Clsl	23	57.4 %
PTM	3	55.8 %
LDA	200	54.0 %
Style	26	57.6 %
POS	45	57.5 %
LIWC All	76	59.8 %
LIWC Sub Categ.	67	60.3 %
Opinion	3	56.3 %

Feature Combinations		
Style Features	**Feature Count**	**Accuracy**
Γ^N (Thr=200) + Style	2,756	57.9 %
Γ^N (Thr=200) + Opinion	2,733	56.2 %
Phr/Clsl + Opinion	26	58.0 %
Phr/Clsl + Style	49	57.5 %
LIWC + Style	93	60.2 %
LIWC + Phr/Clsl	90	60.2 %
LIWC + Γ^N (Thr=200)	2,797	59.1 %
LIWC + Opinion	70	60.3 %
PTM + Opinion	6	57.2 %
Style + Opinion	29	58.7 %
Style + PTM	29	57.4 %
LIWC +Style+Phr/Clsl	116	60.1 %

Table 3: Experiment Results

curacy (62.0 %). Unfortunately, as expected, the top features all included terminal production rules that signal time or product specificity. For example in the pre-enrollment reviews the top 10 features for Γ include NNP → 'Update', CD → '2006', NNP → 'XP' and NNP → 'Palm'. In the post-enrollment reviews the top 10 features include CD → '2012',CD → '2011', NN → 'iPad' and NN → 'iPhone'. We observe the same issue with the Unigrams feature set. This supports our contention that the analysis should restrict itself to style and domain-independent features. The best performing style feature set is LIWC Sub Categ. followed by Non Terminal Production Rules (Γ^N). Opinion is the most parsimonious feature set that performs significantly better than a majority vote.

Non Terminal Production Rules (Γ^N) Table 7 presents the top Non Terminal Production Rules. We observe the following: First, pre-enrollment reviews have noun phrases(NP) that contain fewer leaf nodes than in the post-enrollment reviews. This appears to be due to the inclusion of determiners (DT), adjectives (JJ), comparative adjectives (JJR), personal pronouns (PRP $) or simply more nouns (NN). This might indicate that topics are discussed with more *specifics* in post-enrollment reviews. Second, clauses(S) begin with action oriented verb phrases (VP) in the pre-enrollment reviews. In contrast in the post-enrollment reviews clauses connect two clauses using coordinating conjunctions(CC) or prepositions(IN). One possibility is that reviewers are offering more *detail/concepts per sentence* (where each clause is a detail/concept) in the post-enrollment reviews. Finally, we observe that pre-enrollment reviews include adjectival phrases (ADJP) connect to superlative adverbs (RBS)which convey *certainty*. We will revisit this finding when we review the results from the LIWC model below.

Phrasal/Clausal (Phr./Clsl.) Tables 5 and 6 suggest that post-enrollment reviews emphasize information using *descriptive phrases* — adjectival phrases (ADJP) and adverbial phrases (ADVP) — and *quantifier phrases* (QP). Pre-enrollment reviews appear to have more *complex* clause structures (SBAR, SINV, SQ, SBARQ - see table 5 for definitions).

Parse Tree Metrics (PTM) The three features used are number of sentences, maximum height of parse tree and the maximum width of the parse tree, listed here in descending order of importance for the post-enrollment reviews. As mentioned earlier in section 4 the *average review length* is higher in the post-enrollment reviews so the finding that the number of sentences predict post-enrollment reviews is consistent. Maximum tree width predicts the pre-enrollment reviews. This flat structure indicates a more *complex* communication structure.

Latent Dirichlet Allocation (LDA) This model did not perform very well, being statistically marginally better than majority vote. As mentioned before, we selected the number of topics by maximizing the average cosine distance amongst topics. Even with 200 topics, this measure was 0.39, suggesting that the topics were themselves not well separated. In the limit, each topic would be a non-terminal production rule. This is the same as Non Terminal Production Rules (Γ^N) feature set discussed earlier in this section.

Predicts PRE Enrollment
'number of different words', 'uppercase', 'alphas', 'vowels' , 'short words', 'words per sentence', 'to be words' , 'punctuation symbols', 'long words', 'common prepositions'
Predicts POST Enrollment
'average word length', 'spaces', 'verbs are', 'chars per sentence' , 'verbs be', 'common conjunctions', 'verbs were', 'personal pronouns' , 'verbs was', 'verbs am'

Table 4: Style Metrics: Top Features

Style (STYLE) Table 4 presents the top features for this feature set. The features suggest that reviewers used a more varied vocabulary (number of different words), more words per sentence (words per sentence) and more long words (long words) in pre-enrollment than in post-enrollment reviews. This might indicate that sentences in the pre-enrollment reviews were longer and more *complex*. Interestingly, the average word length did go up in the post-enrollment reviews as did the characters per sentence. In addition, more personal pronouns and conjunctions are used — a finding replicated in the model using LIWC features (see below).

Parts of Speech (POS) The top features for post-enrollment are commas, periods, comparative adjectives, verb phrases and coordinating conjunctions. The top features for pre-enrollment are nouns, noun phrases, determiners , prepositions and superlative adverbs. These results are more difficult to interpret though the use of comparative adjectives suggest more *comparisons* between different objects in the post enrollment reviews.

LIWC SUB CATEG. The top 10 LIWC features are shown in Table 8. LIWC features are categories that are contained in broader categories. For example POSEMO (see Table 8, first feature for "Predicts POST enrollment") refers to the class of positive emotion words. POSEMO itself is contained in a category called "Affective Features" which in turn is classified as a Psychological Process (abbreviated to Pscyh.). The analysis of the categories of features is in itself interesting. Psych./ Cognitive Features occur higher up in features predictive of pre-enrollment reviews than in the features predictive of post-enrollment reviews. "Psych./ Affective Features" occurs as a top feature for the post-enrollment reviews. The actual feature from the "Psych./ Affective Features" category is POSEMO suggesting that the *positive emotion* is more strongly conveyed in the post-enrollment reviews than in the pre-enrollment reviews. Interestingly the corresponding negative feature NEGEMO is in the top 10 features predicting the pre-enrollment reviews. This is especially intriguing since the average rating for reviews in the pre and post-enrollment reviews is the same (see 4). We were concerned that possibly our sampling had induced a bias in the ratings. But the average ratings in our sample are 4.18 and 4.19 pre and post-enrollment respectively (difference is not statistically significant).

FUNCTION WORDS occur extensively in the post-enrollment reviews. We also observe that inclusive (INCL) and exclusive (EXCL) terms are used more in the post-enrollment reviews. Its possible that reviewers are seeking to be more *balanced*. Products are described in personal (I), perceptual (FEEL) and relativistic (SPACE) terms. Pre-enrollment reviews discuss personal concerns (LEISURE, RELIG) , indicate a level of *certainty* (CERTAIN) and opinions are presented in terms of thought process (INSIGHT). Interestingly, the pre-enrollment reviews address the reader (YOU).

Opinions (OPINION) Features predicting post-enrollment are number of objective sentences, number of subjective sentences and finally number of other (neither subjective nor objective) sentences. This suggests that reviewers try to write somewhat more objectively in the post-enrollment reviews.

Feature Combinations With the exception of the combinations STYLE + OPINION , PHR/CLSL +OPINION and PTM + OPINION which improve on either feature set used alone, none of the other combinations improved performance over all component feature sets modeled individually. Overall, none of the combinations improved over LIWC SUB CATEG. Hence we do not delve further into features from these models.

Summary Overall pre-enrollment reviews are more complex (complex clauses, wide parse trees, varied vocabulary, more words per sentence), have fewer concepts per sentence, contain negative emotions, addresses the reader directly and are more certain. Post-enrollment reviews are longer, more descriptive, contain comparisons, contain quantifiers, have more positive emotion and describe the product experience in physical and personal terms.

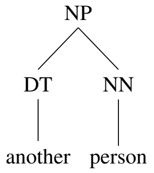
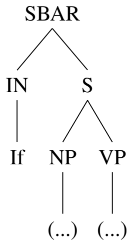
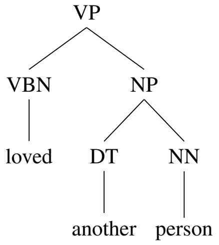
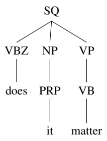
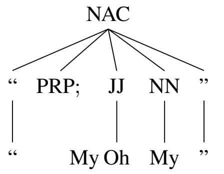
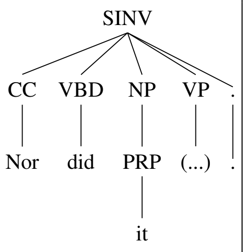
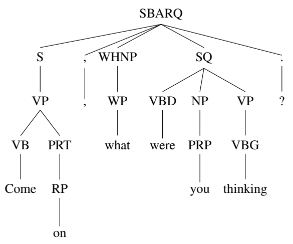
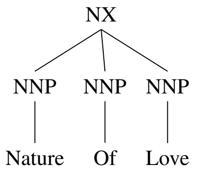
Predicts PRE Enrollment	
1 NP (Noun Phrase)	6 LST (List marker. Includes surrounding punctuation)
EXAMPLE 	EXAMPLE (3)
2 SBAR (Clause introduced by a (possibly empty) subordinating conjunction)	7 VP (Verb Phrase)
EXAMPLE 	EXAMPLE 
3 SQ (Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ)	8 PRN (Parenthetical)
EXAMPLE 	EXAMPLE (p. 73)
4 NAC (Not a Constituent; used to show the scope of certain prenominal modifiers within an NP)	9 SINV (Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal)
EXAMPLE 	EXAMPLE  SINV CC VBD NP VP . Nor did PRP (...) . it
5 SBARQ (Direct question introduced by a wh-word or a wh-phrase)	10 NX (Used within certain complex NPs to mark the head of the NP)
EXAMPLE  SBARQ S , WHNP SQ . VP , WP VBD NP VP ? VB PRT what were PRP VBG Come RP you thinking on	EXAMPLE 

Table 5: Phr/Clsl: Top Features PRE

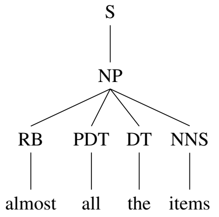
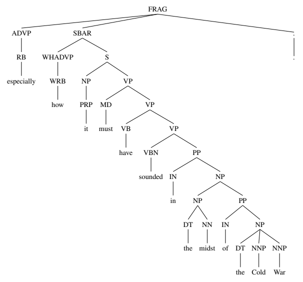
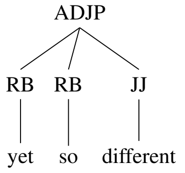
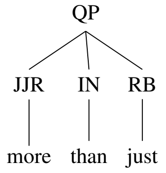
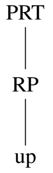
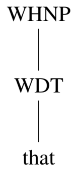
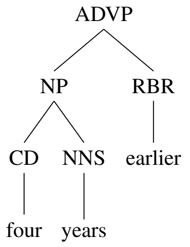
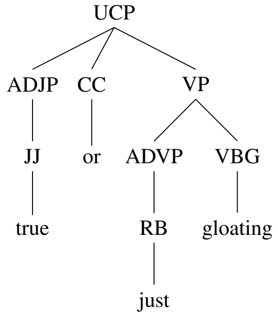
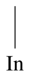
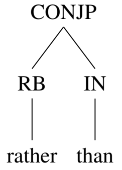
Predicts POST Enrollment	
1 S (Simple declarative clause)	6 FRAG (Fragment)
EXAMPLE  S NP RB PDT DT NNS almost all the items	EXAMPLE  FRAG ADVP SBAR RB WHADVP S especially WRB NP VP how PRP MD VP it must VB VP have VBN PP sounded IN NP in NP PP DT NN IN NP the midst of DT NNP NNP the Cold War .
2 ADJP (Adjective Phrase)	7 QP (Quantifier Phrase)
EXAMPLE 	EXAMPLE 
3 PRT (Particle. Category for words that should be tagged RP)	8 WHNP (Wh-noun Phrase)
EXAMPLE 	EXAMPLE 
4 ADVP (Adverb Phrase)	9 UCP (Unlike Coordinated Phrase)
EXAMPLE 	EXAMPLE  UCP ADJP CC VP JJ or ADVP VBG true RB gloating just
5 X (Unknown, uncertain, or unbracketable)	10 CONJP (Conjunction Phrase)
EXAMPLE 	EXAMPLE 

Table 6: Phr/Clsl: Top Features POST

Predicts PRE Enrollment

Feature	Examples
ROOT → S	(1) And nearly every single item seemed cute and usable to me. (2) Look closely, (...) overwhelming personal and cultural upheaval.
NP → NNP NNP	(1) Tim Bess (2) Jennifer Fitch
PP → IN NP	(1) for its psychological and emotional richness (2) of loyalty
NP → DT NN	(1) the price (2) a book
NP → NNP POS	(1) Frost 's (2) Clough 's
ADJP → RBS JJ	(1) most assuredly (2) most entertaining
WHNP → WP	(1) who (2) what
NP → NNP	(1) Blessed (2) India
PP → TO NP	(1) to the crime (2) to me
S → VP	(1) linking Pye to the crime scene (2) Gripping due to (...)

Predicts POST Enrollment

Feature	Examples
S → S , IN S .	(1) It is functionally the same as Apple's 10 watt charger which outputs 2.1 A , so it is also suitable for charging the iPad. (2) It has 3 levels of trays that spread as you open the box, so you can easily access contents in all trays.
S → IN NP VP .	(1) So I don't think the investment in graphics (...) enjoyability in the game. (2) So we decided to try it again this year.
ROOT → NP	(1) Some kind of (...) disorder ? (2) Proper Alignment and Posture; This segment (...) .
S → S CC S .	(1) Mage and Takumo (...) but lacking in depth.(2) The light feature is great and it powers off (...).
NP → PRP$ NNP NN	(1) your Alpine yodeling (2) my MacBook Pro
S → VP .	(1) Enough negativity. (2) Suffice it to say that (...) .
NP → DT JJR NN	(1) a better future (2) a slower flow
NP → DT JJ , JJ NN	(1) an immediate , visceral reaction (2)a roots-based, singer-songwriter effort
NP → DT NNP NNP NNP NNP	(1) the Post-Total Body Weight Training (2) The Gunfighter DVD Gregory Peck
WHADVP → WRB RB	(1) How far (2) how well

Table 7: Γ^N : Top Features (PCFG Non Terminal)

Predicts PRE Enrollment

Feature	Category	Examples
leisure	Personal Concerns	Cook, chat, movie
verb	Function words	Walk, want, see
certain	Psych./Cognitive Processes	always, never
insight	Psych./Cognitive Processes	think, know, consider
negemo	Psych./Affective Processes	Hurt, ugly, nasty
exclam	Exclamation	!
period	Period	.
you	Function words	2^{nd} person , you, your
preps	Function words	to, with, above
relig	Personal Concerns	2^{nd} synagogue, sacred

Predicts POST Enrollment

Feature	Category	Examples
posemo	Psych./Affective Processes	Love, nice, sweet
article	Function words	a, an, the
i	Function words	1^{st} person singular.
space	Psych./Relativity	Down, in, thin
ingest	Psych./Biological Processes	Dish, eat, pizza
ipron	Function words	Impersonal Pronouns, it its , those
incl	Psych./Cognitive Processes	Inclusive, and, with , include
conj	Function words	and, but, whereas
excl	Psych./Cognitive Processes	Exclusive but, without, exclude
feel	Psych./Perceptual Processes	feels , touch

Table 8: LIWC Sub Category : Top Features

These reviews are are specific, balanced and contain more objective sentences as well.

Discussion on Readability One possibility is that the "Enrollment" effect leads to reviewers writing more readable reviews. To test this hypothesis we performed a paired t-test between readability scores for pre and post-enrollment reviews. Table 9 suggests that indeed this is the case. Flesch Reading Ease is the only measure where a higher score indicates simpler text. For the rest of the measures a higher score implies more complex text. All of the measures are within the average readability range and the magnitude of the differences are small. Nevertheless, these differences are statistically significant [12] with one exception lending support to the idea that "Enrollment" effect might lead to reviewers writing more readable reviews.

[12] The cell size for each class is 57,875, making the modest difference in magnitude statistically significant.

Reading Measure /Cite	Pre Mean	Post Mean	t Value
ARI /(Senter and Smith, 1967)	9.16	9.15	(0.45)
Coleman Liau /(Coleman and Liau, 1975)	8.76	8.68	(6.39)*
Flesch Kincaid /(Kincaid et al., 1975)	8.75	8.71	(2.19)*
Flesch Reading Ease /(Kincaid et al., 1975)	65.63	66.18	6.61*
Gunning Fog /(Gunning, 1952)	11.75	11.70	(2.18)*
LIX /(Anderson, 1983)	38.24	38.07	(2.89)*
RIX /(Anderson, 1983)	3.74	3.71	(3.05)*
SMOG /(McLaughlin, 1969)	10.59	10.56	(2.56)*
* Significant at 5% level			

Table 9: Readability Measures

7 Discussion

So far we have ignored the possibility that writing styles of reviewers may simply continuously evolve with experience and we are simply detecting a difference due to this underlying trend. [13] To address this question we investigated the subperiods within the the pre and post enrollment periods.

We split the post enrollment period (i.e. from date of enrollment to the date the most recent review was posted) further into two equal time periods for each reviewer. As before, we learn a classifier to discriminate between the sub periods. Interestingly the classifier performed the same as chance at p=0.05 (Test Accuracy= 51.0%).[14] [15] However a similar analysis in the pre-enrollment period results in a test set accuracy of 63.3% (significant at p=0.05). So there is a change in writing style within the pre-enrollment period, but there is no continued change post-enrollment. This is not consistent with the continuous style evolution hypothesis. One account would be that Amazon enrolls reviewers whose styles have stabilized. This remains a possibility as Amazon actively selects the members (and we are not aware of the specific rules used by Amazon). The trends (see Figure 1)suggest that there are changes right upto the enrollment dateand some levelling out in the post enrollment period , providing some evidence against this hypothesis.

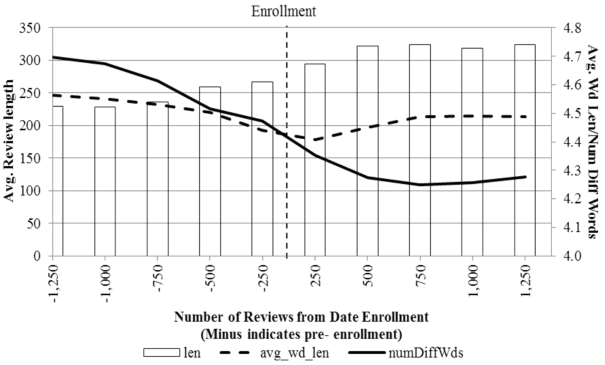


Figure 1: Feature Trends

	Train Size	Test Size	Accuracy
Within Pre-Enrollment	44,800	1000	63.3%
Within Post-Enrollment	59,250	1000	51.0%
Pre vs Post Enroll. Down Sampled	53,840	1000	57.5%

Table 10: Sub Period Results

8 Conclusion

We view this work as a first step toward investigating this phenomenon further. In particular, we plan to test the robustness of our results w.r.t. product specificity, to investigate stylistic differences (a) between reviews for purchased products versus for products received for free amongst Vine members and (b) between reviews by Vine reviewers and non-Vine reviewers. Another line of inquiry involves decomposing the “Enrollment” effect into a reputation/status effect (the influence of the status badge - Vine membership) and a product sampling effect (the influence of receiving goods for free). Finally, investigating the temporal dynamics of style for these reviewers might prove interesting as would determining whether these subtle differences in style affect the *readers* and influence purchase decisions.

9 Acknowledgements

We would like to thank our reviewers for insightful comments that we sought to address here. We would also like to thank Myle Ott for generously sharing the data.

[13] Ideally, if a) the enrollment date had been the same for all reviewers and b) the enrollment was random, we would have a clean experimental framework to detect whether a similar trend exists for non-vine reviewers. Unfortunately, this is not the case.

[14] We report the results only on POS for conciseness. The other feature sets performed similarly.

[15] As before the test sample includes 50 users. However we sampled only 10 reviews in each sub period. Corresponding down sampled performance for Pre vs Post enrollment accuracy is 57.5% (significant at p=0.05)using POS features.

References

Jonathan Anderson. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, pages 490–496, 1983.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. *Poetry*, 580(9): 70, 2013.

Kapil Bawa and Robert Shoemaker. The effects of free sample promotions on incremental brand sales. *Marketing Science*, 23(3):345–363, 2004.

Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337. Association for Computational Linguistics, 2012.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3): 345–354, 2006.

Meri Coleman and TL Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM, 2012.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9: 1871–1874, 2008.

Song Feng, Ritwik Banerjee, and Yejin Choi. Characterizing stylistic elements in syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1522–1533. Association for Computational Linguistics, 2012.

Shyam Gopinath, Jacquelyn S Thomas, and Lakshman Krishnamurthi. Investigating the relationship between the content of online word of mouth, advertising, and brand performance. *Marketing Science*, 2014.

Robert Gunning. Technique of clear writing. 1952.

Thorsten Joachims. A statistical learning learning model of text classification for support vector machines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 128–136. ACM, 2001.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

G Harry McLaughlin. Smog grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

Mauricio M Palmeira and Joydeep Srivastava. Free offer$\neq$ cheap product: A selective accessibility account on the valuation of free offers. *Journal of Consumer Research*, 40(4):644–656, 2013.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

RJ Senter and EA Smith. Automated readability index. Technical report, DTIC Document, 1967.

Kristina Shampanier, Nina Mazar, and Dan Ariely. Zero as a special price: The true value of free products. *Marketing Science*, 26(6):742–757, 2007.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Monica Wadhwa, Baba Shiv, and Stephen M Nowlis. A bite to whet the reward appetite: The influence of sampling on reward-seeking behaviors. *Journal of Marketing Research*, 45 (4):403–413, 2008.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.