Yonathan Perez yperez@stanford.edu Nora Tarano ntarano@stanford.edu

CS 229 Project Final Report:

Learning Convention Propagation in BeerAdvocate Reviews from a Network Perspective

Abstract

We look at the way conventions propagate between reviews on the BeerAdvocate dataset, and try to predict whether a specific convention will be adopted by a user in his coming review. Learning and prediction of convention adoption are done based on exposure of a review, or a reviewer at a specific time point, to the convention. In this project, we define the criteria for exposure of one review to another, which in turn define an implicit network structure over the reviews. We then use features extracted from this network to learn and predict convention adoption.

1. Introduction

BeerAdvocate is a website in which users write reviews on various brands of beer. The BeerAdvocate dataset contains over 1.5 million review records, made by more than 33K users. Some of the users adopt a unique "jargon" in their review text, and use certain conventions (specific words, phrases or abbreviations) which are shared by multiple reviewers and across multiple beers.

In the scope of this project, a convention is an element from a pre-defined set of pieces of text C. We do not address the semantic meaning of a convention, or what makes a piece of text to become a "convention". In this project, we look at the binary classification problem of learning when a convention $c \in C$ is used ("adopted") and try to predict whether a new review r would adopt that convention. In addition to its content, a review r is characterized by three components: The reviewer $user(r)$, the product $beer(r)$, and the time of the review $time(r)$. The hypothesis behind this project is that high "exposure" to a convention c by $user(r)$ while reviewing $beer(r)$ at $time(r)$ increases the likelihood of review r adopting c.

One key question is how to define "exposure" in a review website such as BeerAdvocate. In problems that deal with information propagation in networks (e.g. social networks), the network is given in advance and determines whether a node (which in most cases represents a user) is exposed to another node. In contrast, a review website such as BeerAdvocate does not explicitly define a network structure. Instead, it is up to us to define when are two reviews (or two reviewers) are exposed to one another, and at what times. The definition of exposure then defines an underlying "exposure network" that can be used to reason about information propagation between its nodes. The established exposure relations between reviews and the network structure they define are used to define features for learning and prediction of convention adoption by new reviews that are added to the network.

There are three categories of features (attributes) we use for convention adoption learning: The extent of the exposure the review has to the convention, user bias, and convention bias. User bias captures the tendency of the user to adopt conventions, and convention bias captures the tendency of the convention to be adopted. Features related to the embedding of the sub-graph induced by the convention adopting reviews within the general exposure graph are included under convention bias as an implicit measure of correlation between exposure and convention propagation for that convention.

2. Exposure Network Model

Definition: a review r is *exposed* to review r' if r' is either an earlier review by $\text{user}(r)$, or one of the k preceding reviews on $\text{beer}(r)$. Review r is exposed to a convention c if one of the reviews r is exposed to uses c. Note that there is no requirement for usage by r itself.

Formally, the set of reviews r is exposed to is $Exp[r] = Exp_U[r] \cup Exp_B[r]$ where:

$$Exp_U[r] = \{\text{review } x \mid user(r) = user(x) \wedge time(x) < time(r)\}$$
$$Exp_B[r] = \{\text{review } x \mid beer(r) = beer(x) \wedge rank(r) - k \le rank(x) < rank(r)\}$$

$rank(x)$ is the chronological rank of review x among the reviews of $beer(x)$. Note that $rank(x) < rank(r)$ also implies $time(x) < time(r)$.

The reasoning behind this definition is that "exposure" *between reviews* originates from previous usage of a convention ("user-based exposure") or from contagion from one of the immediate preceding reviews of the same product which are immediately visible to the reviewer ("product-based exposure"). We set *the product exposure parameter* k to be 25, which is the number of reviews in a page on the BeerAdvocate website.

While the above definition is binary (r is either exposed to r' or not), the *extent of exposure* of r to the reviews is not uniform - but a decreasing function of the time difference between the reviews in the case of same-user exposure (using the same conventions as a recent review is more probable than using a convention from an "ancient" one), or a decreasing function of the rank difference between the reviews in the case of same-product exposure (the reviews close in rank to the current review are more easily visible to the reviewer, and probably more relevant).

The above definition of exposure induces a directed network structure $G = (N, E)$ over the dataset, where the nodes represent reviews, and an edge $(r' \to r)$ exists in the network if and only if review r is exposed to review r'. The extent of exposure of r to r' then defines a weight for the edge $(r' \to r)$ which is a decreasing function of the time / rank difference between r and r'. We noticed that the likelihood of convention propagation decreases far more drastically for product-based exposures as rank difference increases than it does for user-based exposure as time difference increases. Thus the edge weights are modeled in the following manner:

$$w(r' \to r) \propto \begin{cases} \left(time(r) - rime(r')\right)^{-1} & ;\ \text{if } (r' \to r) \text{ is due to user-based exposure} \\ \exp\left(-\left(rank(r) - rank(r')\right)\right) & ;\ \text{if } (r' \to r) \text{ is due to product-based exposure} \end{cases}$$

The exposure and network model discussed here deal with reviews as basic entities (i.e. nodes in the network), and not reviewers. This is important in order to incorporate temporal considerations into the model. A review is an instantaneous event - and exposure for the purpose of convention propagation is only relevant at the moment of the review. Thus, we cannot discuss absolute exposure between reviewers (who write multiple reviews at different times), but only exposure at specific times, which is equivalent to discussing exposure between reviews.

3. Features

The basic features (attributes) we use for learning and predicting adoption of convention c are described in the table below. They divided into three categories: The extent of exposure to convention c of the review r (features 1,2), bias of $user(r)$ at $time(r)$ (features 3,4), and the bias of the convention c at $time(r)$ (features 5-9). The last category also includes features that come to capture the embedding of the sub-graph G_c induced by the reviews that adopted c within the general exposure network G (features 7-9). These features come to capture the continuity and linearity of the spread of the convention within the exposure network as an implicit measure of correlation between exposure and convention propagation for that convention.

For more compact formulation, the following sets are defined:

$$T[r] = \{x \mid time(x) < time(r)\}\ ;\ \forall c \in C : S[r,c] = \{x \mid x \in T[r] \wedge x \text{ uses } c\}\ ;\ E[r] = \{(x,y) \in T[r] \times T[r] \mid (x \to y) \in E\}$$

$$\text{In}[x] = \{y \mid (y \to x) \in E\}\ ;\ \text{out}[x] = \{y \mid (x \to y) \in E\}$$

These sets are in turn used for feature formulation:

	Description	Formulation
1	Extent of user-based exposure of review r to convention c	$score(r,c)=\frac{\sum_{x\in Exp_U[r]}\mathbf{1}\{x \text{ uses } c\}\cdot w(x\to r)}{\sum_{x\in Exp_U[r]} w(x\to r)}$
2	Extent of product-based exposure of review r to convention c	$score(r,c)=\frac{\sum_{x\in Exp_B[r]}\mathbf{1}\{x \text{ uses } c\}\cdot w(x\to r)}{\sum_{x\in Exp_B[r]} w(x\to r)}$
3	The fraction of conventions adopted by $user(r)$ up to $time(r)$	$score(r)=\frac{1}{\lvert C\rvert}\cdot\sum_{c\in C}\mathbf{1}\{\exists x\in Exp_U[r]: x \text{ uses } c\}$
4	The fraction of reviews by $user(r)$ to adopt a convention up to $time(r)$, maximized over all possible conventions	$score(r)=\max_{c\in C}\frac{1}{\lvert Exp_U[r]\rvert}\cdot\sum_{x\in Exp_U[r]}\mathbf{1}\{x \text{ uses } c\}$
5	Likelihood of the convention c to get adopted at $time(r)$ - the fraction of reviews that adopted c by $time(r)$	$score(r,c)=\frac{\lvert S[r,c]\rvert}{\lvert T[r]\rvert}$
6	The likelihood of propagation of c given that a review is exposed to c - the weighted fraction at $time(r)$ of exposures (edges in the network) that represent propagations of c	$score(r,c)=\frac{\sum_{(x,y)\in E[r]}\mathbf{1}\{x \text{ uses } c\wedge y \text{ uses } c\}\cdot w(x\to y)}{\sum_{(x,y)\in E[r]} w(x\to y)}$
7	The fraction of adoptions at $time(r)$ that serve as sources in G_c (i.e. start a propagation flow)	$score(r,c)=\frac{1}{\lvert S[r,c]\rvert}\cdot\sum_{x\in S[r,c]}\mathbf{1}\{\text{In}[x]\cap S[r,c]=\varnothing\}$
8	The fraction of adoptions at $time(r)$ that serve as sinks in G_c (i.e. end a propagation flow)	$score(r,c)=\frac{1}{\lvert S[r,c]\rvert}\cdot\sum_{x\in S[r,c]}\mathbf{1}\{\text{out}[x]\cap S[r,c]=\varnothing\}$
9	Average "propagation fan-out " at $time(r)$ - the average fraction of adopters among the out-neighbors of an adopter	$score(r,c)=\frac{1}{\lvert S[r,c]\rvert}\cdot\sum_{x\in S[r,c]}\frac{\lvert\text{out}[x]\cap S[r,c]\rvert}{\lvert\text{out}[x]\rvert}$

The exposure network G is a massive graph of over 1.5M nodes and over 600M edges, with attribute data on both nodes and edges. Even in a compact binary representation, the object representing G is over 30GB in size. Therefore it is crucial that all feature extraction will be performed extremely efficiently. Many of the features depend on a sum of the form $\sum_{x|time(x)<time(r)} F(x)$.

When computed naively, such a sum requires $O(N^2)$ steps where N is the number of reviews, which makes the computation infeasible for such a large graph. However, by visiting reviews in their chronological order and using dynamic programming to incrementally compute the features, we were able to extract all features in a linear time $O(N)$.

For programming convenience and computational efficiency, in our analysis we also only address reviews for which all feature values are available by traversing only the edges - i.e. reviews that both serve as a source node and a destination node of some edge (i.e. have in-degree and out-degree of at least one). This still results in a massive dataset of 829,066 reviews.

4. Learning and Evaluation

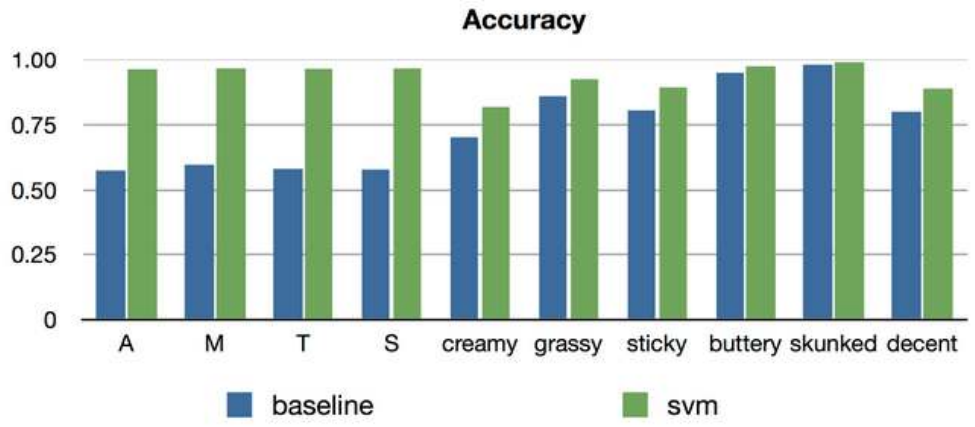
We processed the dataset, constructed the exposure network G , and extracted the feature values using SNAP - a high-performance library for analysis of massive networks (http://snap.stanford.edu/snap/). We learn adoption of the following set of conventions:

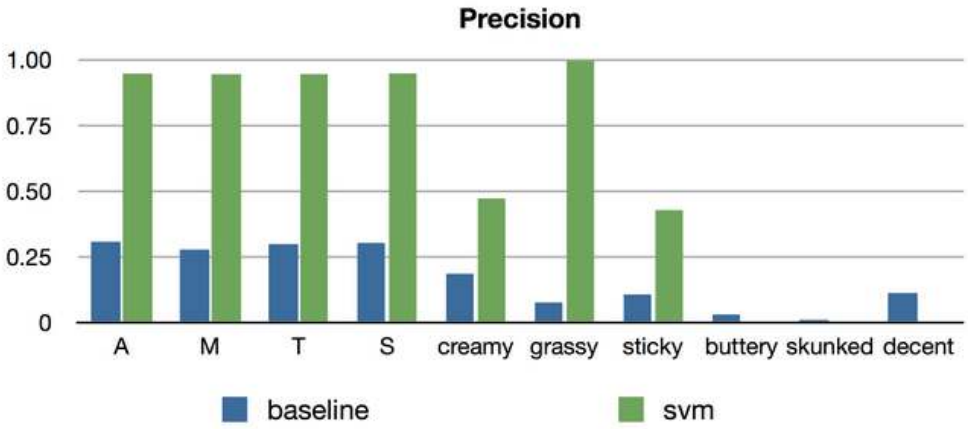
"A", "M", "S", "T", "decent", "stick", "cream", "grass", "butter", "skunk".

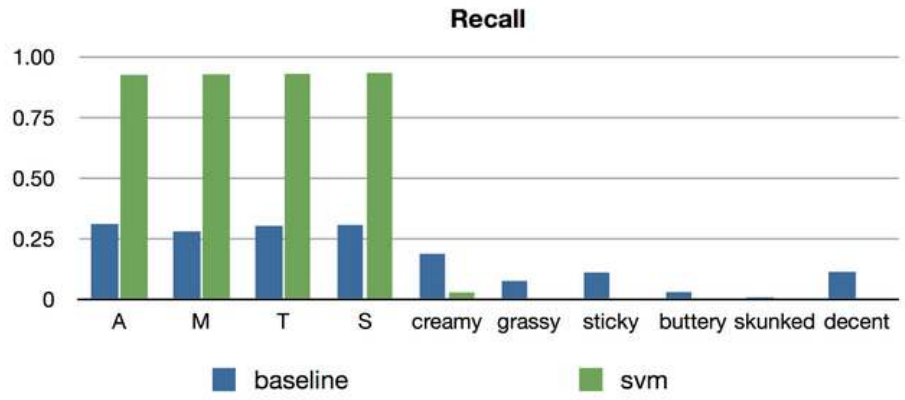
The first four are conventions for abbreviations in the BeerAdvocate community (stand for "Aroma", "Mouth-full", "Smell" and "Taste"). We only counted the appearances of these abbreviations in the text when they were used as conventions - when capitalized and followed by a colon or a hyphen. The rest are commonly repeating not obvious word roots used in beer descriptions. The frequencies of these conventions across the entire dataset are given in the following table:

"A"	"M"	"S"	"T"	"decent"	"stick"	"cream"	"grass"	"butter"	"skunk"
30.79%	28.03%	30.32%	29.98%	11.34%	11.04%	18.49%	7.66%	2.71%	1.03%

We partitioned the dataset (829,066 reviews) into a training set containing 70% of the data points (580,346 reviews) and a test set containing the remaining 30% (248,720 reviews). We then trained an SVM (using a liblinear SVM with a linear kernel with L_2 regularization) to learn and predict the adoption of the above conventions using the features extracted from the adoption network. Our baseline for evaluation was naive prediction based on convention frequency: a prediction scheme that considers each review r and convention c independently and predicts that r would use c with probability p_c which is the frequency of convention c across the entire dataset. We compare accuracy, precision and recall for both prediction methods for each of the conventions. The results are described in the following figures:

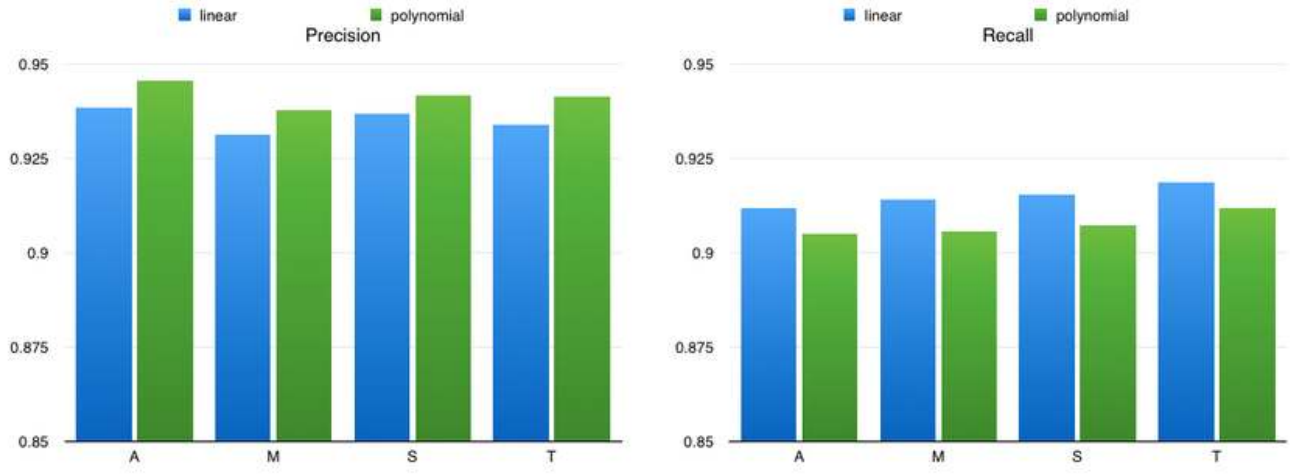






Accuracy is a more significant for the higher-frequency conventions than for lower-frequency conventions. But even then, the positive and negative classes are highly skewed (the vast majority of reviews do not use a given convention). Thus, precision and recall should be taken into account as more significant performance metrics of the prediction scheme. We can see that the SVM performs surprisingly well when a convention appears frequently enough in the dataset. It also seems that the SVM predicts more conservatively (less prone to assign a positive label) than the baseline when the convention has low frequency, which explains the low recall for infrequent conventions. This can be attributed to the fact that SVM captures dependency between the data points, whereas our baseline predicts for each point independently.

The number of training examples by far exceeds the number of feature used, so it makes sense to use more features and construct a richer model by mapping the existing nine features into a higher dimensional feature space using a kernel. However, trying to do so using the entire dataset (using libsvm SVMs with a higher-degree polynomial kernel) proved to be extremely computationally expensive. We compared the precision and recall scores for the most commonly used conventions using an SVM with a linear kernel vs. one with a 3rd -degree polynomial kernel (both using L_2 regularization) on a sample of 10K reviews. Despite our initial assumption, the results showed that the performance of both kernels is very similar (we didn't try higher degree kernels due to long runtimes):



5. Conclusion and Potential Future Work

We were surprised to see how well classification based on exposure network features performed in comparison to naive frequency-based prediction. Yet, we believe that exposure network based features can be further improved - for instance by using statistical inference to determine values for network edges weights, or by leveraging network statistics (for instance degree distribution or weakly-connected component decomposition) of the entire graph and convention-adopting sub-graphs. Additionally, such prediction could be further improved by incorporating non-network based features, such as user and product bias, and semantic/linguistic features of the conventions.