Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

9-2015

Did you expect your users to say this?: Distilling unexpected micro-reviews for venue owners

Wen-Haw CHONG
Singapore Management University, whchong.2013@smu.edu.sg

Bingtian DAI
Singapore Management University, btdai@smu.edu.sg

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

DOI: https://doi.org/10.1145/2700171.2791024

Citation

CHONG, Wen-Haw; DAI, Bingtian; and LIM, Ee-Peng. Did you expect your users to say this?: Distilling unexpected micro-reviews for venue owners. (2015). *HT '15: Proceedings of the 26th ACM Conference on Hypertext and Social Media: Guzelyurt, Northern Cyprus, September 1-4*. 13-22. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/3105

Did You Expect Your Users to Say This? Distilling Unexpected Micro-reviews for Venue Owners

Wen-Haw Chong, Bing Tian Dai, and Ee-Peng Lim
Singapore Management University
80 Stamford Road, Singapore 178902
whchong.2013@phdis.smu.edu.sg, btdai@smu.edu.sg, eplim@smu.edu.sg

ABSTRACT

With social media platforms such as Foursquare, users can now generate concise reviews, i.e. *micro-reviews*, about entities such as venues (or products). From the venue owner's perspective, analysing these micro-reviews will offer interesting insights, useful for event detection and customer relationship management. However not all micro-reviews are equally important, especially since a venue owner should already be familiar with his venue's primary aspects. Instead we envisage that a venue owner will be interested in micro-reviews that are *unexpected* to him. These can arise in many ways, such as users focusing on easily overlooked aspects (by the venue owner), making comparisons with competitors, using unusual language or mentioning rare venue-related events, e.g. a dish being contaminated with bugs. Hence in this study, we propose to discover unexpected information in micro-reviews, primarily to serve the needs of venue owners.

Our proposed solution is to score and rank micro-reviews, for which we design a novel topic model, Sparse Additive Micro-Review (SAMR). Our model surfaces micro-review topics related to the venues. By properly offsetting these topics, we then derive unexpected micro-reviews. Qualitatively, we observed reasonable results for many venues. We then evaluate ranking accuracy using both human annotation and an automated approach with synthesized data. Both sets of evaluation indicate that our novel topic model, Sparse Additive Micro-Review (SAMR) has the best ranking accuracy, outperforming baselines using chi-square statistics and the vector space model.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Applications**]: Data Mining

Keywords

Micro-review, tip, ranking, Foursquare, unexpected

HT '15 September 01 - 04, 2015, Guzelyurt, TRNC, Cyprus
©2015 ACM. ISBN 978-1-4503-3395-5/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2700171.2791024.

1. INTRODUCTION

In recent years, the prevalence and increased popularity of micro-blogging services have resulted in huge volumes of related data being collected. Various platforms such as Foursquare and Twitter provide rich context and fine-grained data. These have led to new possibilities for data mining applications and knowledge discovery tasks.

In particular, Foursquare[1] allows users to provide short reviews about specific venues. Unlike other platforms where review mining has been extensively explored [15, 18, 25, 17], Foursquare reviews are shorter and more succinct[2]. We call such short reviews as *micro-reviews* [27, 20] to differentiate them from traditional reviews. Foursquare micro-reviews are also known as tips and we use both terms interchangeably. While Foursquare is popular, this is not the only source of micro-reviews. Micro-reviews can also be found in micro-blogging platforms such as Twitter due to users writing about their experiences with products or businesses. Micro-reviews thus represent a wealth of information that can be exploited for applications in recommender systems and customer relationship management systems.

In this work, we approach micro-review mining with the aim of providing value to *venue owners*. We differentiate venue owners from venue visitors or customers which we shall simply refer to as users. With the growing popularity of social media, there is value for owners of physical venues, e.g. restaurants, to exploit social media for outreach, publicity and business improvement. In particular, micro-reviews serve as a form of feedback for venue owners, one that is gathered at little effort and cost. However, deriving useful information from the mass of data requires some effort. In fact, a popular venue can easily garner hundreds/thousands of tips over time. This justifies the need for content distillation techniques.

To the best of our knowledge, the needs of venue owners have been neglected in prior work [15, 18, 12, 19, 21] which largely focused on serving users. Intuitively, we expect venue owners and users to have very different needs in terms of useful information. While users may be deciding on venue visitations or trying to understand venue characteristics, venue owners should be more or less familiar with their own venue characteristics and what they are offering. However, given that users can write about almost anything, it is likely that *some micro-reviews will be relatively less expected to the venue owner*. This forms the primary motivation for our work. We envisage that from the venue owner's per-

[1] www.foursquare.com

[2] There is a character limit of 200 for Foursquare

spective, unexpected micro-reviews should be more useful in providing new nuggets of information.

Unexpected micro-reviews can be due to various reasons. Through analyzing the Foursquare data used in our experiments, we envisage some possibilities as follows:

1. **Problem-related**: Describing aspects that may have been overlooked or given little emphasis by venue owners, e.g. parking problems mentioned in relation to food venues
2. **Competition**: Making comparisons with competitors, e.g. mentioning that the same item at some competing restaurant tastes better.
3. **Event**: Describing new unexpected information or an event , e.g. finding bugs in a dish
4. **Linguistic**: Containing language or words that are less frequently used. Such language features are often found in micro-reviews with highly negative or highly positive sentiment.

Note that the above are not necessarily mutually exclusive. For example, a micro-review can be describing an event in a highly negative manner. As an illustration, Table 1 displays four selected tips for a cafe in Singapore. From the perspective of an average venue owner (one who wishes to attract customers), the first tip is least expected. The second tip describes construction noise that affects business, but is external to the venue and which may be easily overlooked, especially if the venue owner is not always physically present. Hence this tip should be relatively unexpected as well, compared to the bottom two tips. For more examples, refer to case studies in Section 4.

Really lazy staff here. Lotsa tables not cleared, litters all over the floor, and even cockroaches running around! Staff just hide in their cozy aircon counter and slack about. Is this Starbucks?!
It's a good place to watch the world go by but the double construction at the Royal Plaza Hotel and Shaw Centre makes it too noisy to enjoy the scene. Wait a few months for construction to stop.
Best part is that it's open 24 hrs.
great taste coffee...

Table 1: Sample tips for a cafe in Singapore.

Henceforth, we propose to compute scores for each micro-review for ranking. Formally our research task is: *For a venue of interest, assign scores to its micro-reviews such that micro-reviews that are unexpected have higher scores.* After this, we rank and extract the top ranking micro-reviews. Venue owners may then decide on further actions after examining high ranking micro-reviews. Also note that in the current work, we do not differentiate unexpected micro-reviews by the underlying reasons, deferring this to future research. In our setting, we also assume most micro-reviews to be relevant to their associated venues, i.e. containing related information. Should there be irrelevant micro-reviews, these can first be filtered out using existing techniques [8, 1].

We propose a topic modeling approach for computing micro-review scores. Our exploration is driven by how we expect a micro-review to be generated. Intuitively users talk about various topics in micro-reviews, related to the venues they visit. One can imagine that if a micro-review can be explained well in terms of a topic or attributed to some notion of a background, then it is not that unexpected after all. Hence with a model for the expected, we can offset the expected to derive the unexpected. With this intuition, we design a novel topic model based on a sparse generative framework [10]. The proposed topic model, **Sparse Additive Micro-Reviews (SAMR)** regards the generation of micro-reviews as a process jointly driven by the background, user-generated topics and the venue. We then score and rank each micro-review by combining appropriate model parameters.

Finally we recognise that ideally, venue owners should be included in the model. However this is currently not possible since micro-reviews are primarily written by users, not venue owners. Instead we have used topics and implicitly venue types to circumvent this. For example, the average owner of a dining venue should expect food and service-related topics to explain well the many micro-reviews he received. If not, then there may be cause for attention.

Contributions. In summary, our contributions are:

- We propose to mine unexpected micro-reviews to serve the needs of venue owners. This problem has been largely unexplored.
- We have designed a novel topic model, SAMR for the proposed problem. Coupled with an appropriate scoring scheme, SAMR outperforms compared baselines in both manual and automated evaluation.
- To assess ranking accuracy, we have conducted experiments with human annotators. Given a pair of micro-reviews ranked differently by competing approaches, annotators are tasked to vote for the micro-review that they deemed to be more unexpected. The SAMR model is superior, attracting 70% of the votes. (Refer to Section 5).
- For automated large scale evaluation, we construct *pseudo-venues* by mixing micro-reviews from different venues. In this manner, for each pseudo-venue, we obtain the ground truth set of unexpected micro-reviews for ranking. On measuring the ranking accuracy, the SAMR model is superior in precision. (Refer to Section 6).

2. RELATED WORK

Review Summarization. Instead of micro-reviews, much prior work [15, 18, 25, 17] had focused on mining longer reviews from e-commerce platforms, e.g. Amazon.com.

In review summarization, a well studied review mining problem, multiple reviews of an item are processed to select a set of sentences/phrases covering different representative aspects of an item, e.g. service quality for a restaurant. For example, Hu and Liu [15] selected sentences that express opinions on item aspects and compiled them into a summary, along with opinion statistics. Lappas and Gunopulos [18] exploited opinions differently and defined a review to be more confident if its expressed opinion agrees with the majority. Their objective is then to find a compact set of high-confidence reviews covering aspects selected by the user. Tsaparas et al. [25] reformulated summarization as a maximum coverage problem such that one selects a limited number of reviews that cover attributes from both positive/negative viewpoints. Lappas et al. [17] proposed a more constrained coverage problem in that reviews are selected to also preserve the distribution of different opinions.

The above summarization techniques have been successfully applied on non-micro reviews and should be easily extendible to micro-reviews. On the other hand, summarization results may be of less value to the average venue owner, since they should already be familiar with one's own venue. Instead, we provide a different value proposition by extracting unexpected micro-reviews. Our work is hence distinct in this aspect.

Supervised Review Mining. Previous works have also attempted to [12, 19, 16, 32] assign review scores based on a general notion of usefulness. These mainly rely on supervised techniques that fit a model using labeled data. The work by Ghose and Ipeirotis [12] used annotated data to fit regression models with review usefulness and sales impact as the dependent variables. Liu et al. [19] utilized rated reviews, i.e. reviews accompanied by online usefulness votes, to build a regression model. They determined that the reviewer expertise, writing style and timeliness are crucial factors affecting review usefulness. Kim et al. [16] applied SVM regression on rated reviews to model review usefulness, with the findings that review length, unigrams etc are relevant model input features. Similarly, Zhang and Varadarajan [32] applied regression on rated reviews for review usefulness scoring.

As can be seen, the targeted audience is again the users or customers, rather than the venue owners. The two roles will have different notions of usefulness. For example, while users may find a micro-review useful if it describes the signature dish of a restaurant, the restaurant owner already knows his signature dish and may derive less value. To date, there is also no collection of labels in terms of whether a micro-review is expected or not. Hence our problem is rather less straightforward and the absence of labeled data makes it difficult to apply supervised techniques.

Micro-blog/micro-review mining. Micro-reviews have some resemblance to micro-blogs, e.g. tweets, as both are short length in nature. There are also parallels if we regard the events or topics in micro-blogs to be analogous to items in micro-reviews. For micro-blog mining, a popular track [26, 14, 2, 30] is to model the interestingness/popularity of micro-blogs, where it is assumed that more popular/interesting micro-blogs will lead to more retweets. For example, Uysal and Croft [26] used a decision tree and a learned ranking function, to rank micro-blogs for each user based on his propensity to retweet each micro-blog. Both Hong et al. [14] and Alhadi et al. [2] employed classifiers to model and predict micro-blog interestingness/popularity. Lastly, Yang et al. [30] predicted retweeting behaviors as well, but in a semi-supervised framework using the factor graph.

Popularity has also been studied in the context of micro-reviews. Vasconcelos et al. [27] have utilized micro-review features and online user voting data to fit classification models for popularity. Other than popularity, other researchers [20, 7] focused on identifying sentiment polarity in micro-reviews. It is quite obvious that popular nor sentiment bearing micro-reviews cannot be equated to unexpected micro-reviews which are the subject of this work.

3. APPROACHES

We denote the number of users, venues and topics as U, V and K respectively. Also let the word vocabulary size be W. A topic is indicated by z, a venue by v and a word by w. A micro-review is represented by $\mathfrak{m}$ and has $|\mathfrak{m}|$ words. We also denote the micro-review count for user u as $n(u)$. We introduce other notations in an inline manner for easier reading.

We first discuss our proposed topic model, followed briefly by the baseline approaches.

3.1 Sparse Additive Micro-Reviews

To design a model for unexpectedness, it may be easier to start with *what is expected*. Hence our topic model's generative process is based on how we expect a micro-review to be formed. We then devise the micro-review unexpectedness scores for ranking. Readers seeking a quick understanding of the model can review the generative process below and the scoring schemes in Section 3.1.4.

Our proposed topic model: **Sparse Additive Micro-Reviews (SAMR)** utilizes the Sparse Additive Generative framework (SAGE) [10] introduced by Eisenstein et al. for topic modeling. Unlike traditional Latent Dirichlet Allocation (LDA)-based models [6], the SAGE framework is based on inferring *facet* deviations in log-space from a background distribution. In a Bayesian network, facets are parent nodes representing some factors which directly influence a probability distribution on a child node. For example, the word distribution in a document can be affected by facets such as topics and the authorship.

Eisenstein et al. [10] asserts several advantages of the SAGE framework which motivate our choice. In this framework, a child node, e.g. a word node, can be generated by multiple facets simultaneously without a latent switch to indicate which facets are active at any one time. It is also easy to enforce sparsity to avoid overfitting and generative facets can be combined additively in log-space. The latter facilitates the design of scoring schemes by facet addition or subtraction.

3.1.1 Generative process

We begin with a high level description of our model, which encapsulates how we expect a micro-review to be formed. Figure 1 presents the plate diagram while Table 2 summarizes model parameters which consist of all the facets.

We first assume a global background distribution over topics. Based on personal interest and the types of venues one visit, a user's topic set will differ to some extent from the global distribution, i.e. he 'deviates' from the background. For example, a user who frequents night spots may focus more on clubbing related topics. Thus in the model, the user generates topic deviations $\boldsymbol{\eta}^{USER}$ which are combined with the background topic facet $\boldsymbol{\eta}^0$ to obtain the conditional topic distribution.

At the next level, the topics generate deviations affecting the background distribution over venues and words. For words, we specify that the distribution is concurrently conditional on three facets: the background $\boldsymbol{\phi}^0$, topic $\boldsymbol{\phi}^{TOPIC}$ and venue $\boldsymbol{\phi}^{VENUE}$. Note that unlike LDA-based models, no latent switch is required per word to indicate the active facet. With the SAGE framework, we assume that all facets are jointly responsible for every word in an additive manner.

Due to the short lengths of micro-reviews, we also assume that each micro-review covers only one topic.

Intuitively the background word facet $\boldsymbol{\phi}^0$ implies that every word has some baseline occurrence frequency (in log space) at any venue while venue facets $\boldsymbol{\phi}^{VENUE}$ mean that venues elevate or depress word frequencies through positive/negative deviations. Similar interpretation can be applied for user topics $\boldsymbol{\phi}^{TOPIC}$. For venue frequencies, the topic facet $\boldsymbol{\theta}^{TOPIC}$ results in deviation from the background $\boldsymbol{\theta}^0$, e.g. a dining venue is more likely for food-related topics.

Note that deviations are computed in log space and additional facets can be included in an additive manner. The different facets are also readily combined to compute conditional distributions. For example, denote the background word facet as $\boldsymbol{\eta}^0$ with the k-th element η_k^0 corresponding to the k-th topic. Similarly let $\eta_{u,k}^{USER}$ be the (u,k)-th element of the user facet for topics $\boldsymbol{\eta}^{USER}$. Then the probability of topic k conditional on the background and user u is:

$$p(z = k|\eta_k^0, \eta_{u,k}^{USER}) = \frac{exp(\eta_k^0 + \eta_{u,k}^{USER})}{\sum_{i=1}^{K} exp(\eta_i^0 + \eta_{u,i}^{USER})} \qquad (1)$$

which is an element from a multinomial vector. Other component distributions can be similarly written out.

Model Parameters	Dimension	Symbol
Background topic facet	$1{\times}K$	$\boldsymbol{\eta}^0$
User-dependent topic facet	$U{\times}K$	$\boldsymbol{\eta}^{USER}$
Background venue facet	$1{\times}V$	$\boldsymbol{\theta}^0$
Topic-dependent venue facet	K$\times V$	$\boldsymbol{\theta}^{TOPIC}$
Background word facet	$1{\times}W$	$\boldsymbol{\phi}^0$
Topic-dependent word facet	K$\times W$	$\boldsymbol{\phi}^{TOPIC}$
Venue-dependent word facet	V$\times W$	$\boldsymbol{\phi}^{VENUE}$

Table 2: Model Parameters

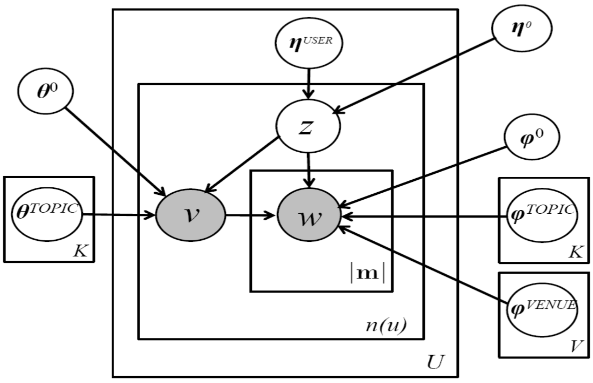
Formally for each micro-review $\mathfrak{m}$, the model's generative process is as follow:

1. The user first samples a topic: $p(k|\eta_k^0, \eta_{u,k}^{USER})$
2. The background venue facet and sampled topic jointly generate the venue v: $p(v|\theta_v^0, \theta_{k,v}^{TOPIC})$
3. The background word facet, sampled topic and observed venue jointly generate the bag of words in the micro-review: $\prod_{w\in\mathfrak{m}} p(w|\phi_w^0, \phi_{k,w}^{TOPIC}, \phi_{v,w}^{VENUE})$

We next discuss model regularization and inference. Readers less keen on technical details can skip directly to Section 3.1.4 describing the scoring scheme.

3.1.2 Regularization

Thus far, we have described an additive model for generating micro-reviews. However the model is not yet a sparse one. Sparsity is important to speed up parameter learning as well as assist model interpretation by retaining only those deviations that are more significant. To achieve sparsity, we utilize the 0-mean Laplace distribution as the prior, implying that the model likelihood function is L_1 regularized. Thus we penalize large parameter values and drive most of them towards 0. For example, the topic facet results in non-zero deviations from the background for only a small set of words instead of for all words.



$\boldsymbol{\eta}^{USER}$
$\boldsymbol{\eta}^0$
$\boldsymbol{\theta}^0$
z
$\boldsymbol{\varphi}^0$
$\boldsymbol{\theta}^{TOPIC}$
K
v
w
$|\mathbf{m}|$
$\boldsymbol{\varphi}^{TOPIC}$
K
$\boldsymbol{\varphi}^{VENUE}$
V
$n(u)$
U

Figure 1: The SAMR topic model in plate notation

3.1.3 Inference

Since the model likelihood and parameter gradients are readily computable, we apply EM learning for parameter inference. The EM-steps are iterated until the log likelihood converges. In the E-step, we hold the model parameters constant and sample topics for every micro-reviews conditional on the old topics. To sample a topic $z_{\mathfrak{m}}$ for micro-review $\mathfrak{m}$ with user $u_{\mathfrak{m}}$, venue $v_{\mathfrak{m}}$ and word vector $\boldsymbol{w}_{\mathfrak{m}}$, we use the following sampling equation:

$$z_{\mathfrak{m}} \sim p(z_{\mathfrak{m}} = k|\boldsymbol{\eta}_{u_{\mathfrak{m}}}^{USER}, \boldsymbol{\eta}^0) p(v_{\mathfrak{m}}|\boldsymbol{\theta}_{z_{\mathfrak{m}}=k}^{TOPIC}, \boldsymbol{\theta}^0) \\ p(\boldsymbol{w}_{\mathfrak{m}}|\boldsymbol{\phi}_{z_{\mathfrak{m}}=k}^{TOPIC}, \boldsymbol{\phi}^0, \boldsymbol{\phi}_{v_{\mathfrak{m}}}^{VENUE}) \qquad (2)$$

which is readily expanded in the same manner as eq(1). In the M-step, we fix the sampled topic assignments and maximize the model parameters via a proximal gradient based optimization technique.

First we present gradients which are required for optimization. All gradients take on an intuitive form of actual/observed frequencies minus the expected frequencies imposed by the model. The gradients for the topic facets are as follows:

$$\partial\eta_k^0 = \sum_{u=1}^{U} n(u,k) - \sum_{u=1}^{U} n(u) p(z = k|\eta_k^0, \eta_{u,k}^{USER}) \qquad (3)$$

$$\partial\eta_{u,k}^{USER} = n(u,k) - n(u) p(z = k|\eta_k^0, \eta_{u,k}^{USER}) \qquad (4)$$

where $n(u,k)$ is the number of micro-reviews from user u assigned to topic k and $n(u)$ is user u's total micro-review count. The gradients for the venue facets can be derived as:

$$\partial\theta_v^0 = \sum_{k=1}^{K} m(k,v) - \sum_{k=1}^{K} m(k) p(v|\theta_v^0, \theta_{k,v}^{TOPIC}) \qquad (5)$$

$$\partial\theta_{k,v}^{TOPIC} = m(k,v) - m(k) p(v|\theta_v^0, \theta_{k,v}^{TOPIC}) \qquad (6)$$

where $m(k,v)$ is the number of times venue v is assigned topic k and $m(k)$ is the total occurrences of topic k over all venues. Finally, there are 3 sets of gradients for the word facets. For brevity, denote $p(w|\phi_w^0, \phi_{k,w}^{TOPIC}, \phi_{v,w}^{VENUE})$ as $\alpha_{w|(k,v)}$. The gradients can be written as:

$$\partial\phi_w^0 = \sum_{v=1}^{V}\sum_{k=1}^{K} d(k,v,w) - \sum_{v=1}^{V}\sum_{k=1}^{K} d(k,v,w)\alpha_{w|(k,v)} \qquad (7)$$

$$\partial\phi_{k,w}^{TOPIC} = d(k,.,w) - \sum_{v=1}^{V} d(k,v,w)\alpha_{w|(k,v)} \tag{8}$$

$$\partial\phi_{v,w}^{VENUE} = d(.,v,w) - \sum_{k=1}^{K} d(k,v,w)\alpha_{w|(k,v)} \tag{9}$$

where $d(k,v,w)$ is the number of times word w at venue v is assigned to topic k.

With the computed gradients, we then derive an updating rule using the Iterative Shrinkage Thresholding Algorithm (ISTA)[5]. It has been shown [5] that the L_1 regularized problem can be solved iteratively by the following:

$$\mathbf{x}_t = \arg\max_{\mathbf{x}} \left\{ \frac{1}{2\tau}\|\mathbf{x} - (\mathbf{x}_{t-1} + \tau\triangledown f(\mathbf{x}_{t-1}))\|^2 - \lambda\|\mathbf{x}\|_1 \right\} \tag{10}$$

where $f()$ is a function to be maximized (i.e. likelihood in our case), $\mathbf{x}$ is the parameter vector, τ is the learning rate and λ is the regularization term. For our model inference, the parameters are the various facets described earlier.

By differentiating and solving eq(10), we obtain the following update rule for each element:

$$x_t = \begin{cases} \rho + \lambda\tau, & \rho > -\lambda\tau. \\ \rho - \lambda\tau, & \rho < \lambda\tau. \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

where $\rho = x_{t-1} + \tau\partial f/\partial x_{t-1}$.

3.1.4 Scoring schemes

With the inferred model, we now devise scoring schemes for ranking micro-reviews. Recap that we seek to extract micro-reviews that are unexpected. Now we consider how to offset the expected in order to derive unexpectedness score for each micro-review. Intuitively, if a micro-review can be explained well by a topic sampled from the venue, then we regard it as not that unexpected after all. For example, an Asian food venue will have many micro-reviews that are described well by, say an Asian food topic. Thus such micro-reviews will not be that unexpected. This implies that our scoring scheme should offset the effects of topics.

Recall from the generative process, that both the venue and micro-review topic lead to deviations of each micro-review word from the background. Deviations are computed in log-space, which facilitates addition and subtraction. Hence to offset the effect of the topic for each word, we can simply subtract the topic-dependent deviations from the venue-dependent deviation. Given a micro-review, we repeat this computation over all its words. By then averaging over all the words, we derive the micro-review score. Also note that the background is already implicitly accounted for in the deviations, hence we do not subtract the background again. We term this scoring scheme as SAMR(vt). Formally a micro-review score is computed as follows:

$$\text{SAMR(vt)} : score(\mathfrak{m}) = \frac{1}{|\mathfrak{m}|}\sum_{w\in\mathfrak{m}} (\phi_{v_{\mathfrak{m}},w}^{VENUE} - \phi_{z_{\mathfrak{m}},w}^{TOPIC}) \tag{12}$$

To ascertain the effect of topic subtraction, we compare the above scoring scheme with an alternate scheme, SAMR(v), where we use solely the venue-dependent word deviations:

$$\text{SAMR(v)} : score(\mathfrak{m}) = \frac{1}{|\mathfrak{m}|}\sum\nolimits_{w\in\mathfrak{m}} \phi_{v_{\mathfrak{m}},w}^{VENUE} \tag{13}$$

	Venue A (Corpus A)	Not venue A (Corpus B)	Total
Freq. of word w	a	b	$a+b$
Freq. of other words	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Table 3: Contingency table for venue analysis. The bracketed column headers illustrate traditional usage in corpora analysis.

The micro-review score can also be generated using other functions. One possibility is to use the maximum word score of constituent words such that micro-reviews that mix expected and unexpected words are not penalized by the former. For brevity in the current paper, we only present results with the averaging function.

3.2 Baselines

Our hypothesis is that by modeling and offsetting topics, we can better extract unexpected micro-reviews. Hence for comparison, we consider baseline approaches that do not consider topics at all.

3.2.1 Chi-squared Statistics

We start with the chi-squared statistic, which has a simpler expected model for word usage in micro-reviews. This has been used in corpora analysis to analyze linguistic differences between different text corpora, e.g. between British and American English [13]. For this task, works such as [28, 22] applied the chi-squared test to identify words with significant differences in usage frequency across different corpora. There is an expected usage frequency for each word based on proportions and without any notion of topics. Specifically the following test statistic is used:

$$\chi^2 = \sum\nolimits_i \frac{(O_i - E_i)^2}{E_i}, \quad E_i = \frac{M_i \sum_j O_j}{\sum_j M_j} \tag{14}$$

where O_i and E_i are respectively the observed and expected frequencies of the targeted word in corpus i, which in turn has M_i number of words. χ^2 follows the chi-squared distribution under the null hypothesis, i.e. targeted word has no difference in usage frequency across different corpora.

Hypothesis testing leads to rejecting/not rejecting the null hypothesis. This is not required in our problem and omitted. Instead we aim to compute a venue-specific score for each word and subsequently for each micro-review. We treat venues as analogous to corpora and utilize the test statistic χ^2 directly as word scores. The basic idea is that words deviating more from their expected frequencies for a given venue are more unexpected, thus giving larger χ^2 values. The score for each micro-review is then obtained by averaging its word scores. We refer to this scheme as χ^2 scheme.

For comparing a word usage across two corpora, it is convenient to represent information in a 2×2 contingency table. In our case, we treat the venue of interest as one corpus and all other venues as the second corpus. This results in a slightly modified contingency table as shown in Table 3. From the table, the test statistic is computed as:

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{15}$$

3.2.2 *Corrected Chi-squared Statistics*

In computing the chi-square statistic, one in fact approximates the *discrete* binomial distribution of observed frequencies with the *continuous* chi-squared distribution. It has been argued [31],[11, p.14] that the approximation error is overly large for small sample sizes, i.e. when at least one cell in the contingency table has expected frequency < 5. To mitigate this, the Yates' continuity correction [31, 11, 4] was proposed to reduce the error. The corrected statistic for the 2×2 contingency table is computed as:

$$Y^2 = \sum\nolimits_i \frac{(|O_i - E_i| - 0.5)^2}{E_i} = \frac{\zeta(|ad - bc| - 0.5\zeta)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{16}$$

where $\zeta = a + b + c + d$.

For large sample sizes, the correction effect is negligible. We refer to this scheme as the Y^2 scheme.

3.2.3 *Frequency based Weighting*

Lastly, we consider the case where each word is assigned its TF-IDF score [23, 29]. TF-IDF scores are computed by regarding each venue as a document made up of all its micro-reviews. To avoid introducing new terminology, we term this scoring and ranking approach as the *TF-IDF* scheme.

4. QUALITATIVE ANALYSIS

In this section, we cover qualitative analysis, giving examples of unexpected micro-reviews that are uncovered for sample venues. We apply the SAMR(vt) scoring scheme on Foursquare venues in Singapore. We include venues with at least 15 micro-reviews, thereby obtaining 3,150 venues with 56,997 micro-reviews generated by 25,189 users. We also remove stop words and rare words ($<$ 5 occurrences). Unexpected micro-reviews due only to rare words is a trivial problem, which we are not focusing on.

We set the number of topics at 10 and initialize them randomly. Following inference, our eventual model has a sparsity of 81.9% for the topic-word facet and 99.4% for venue-word facet, i.e. 99.4% of the word deviations due to venues are 0.

We then manually examined the results for dozens of venues. Since it is required to put oneself in the shoes of the venue owner, some subjectiveness may be involved in assessing whether a micro-review is unexpected or not. We also do not assert the fact that all venues have unexpected micro-reviews. This depends very much on how users write their micro-reviews. Through subsequent experiments with annotators and pseudo-venues, we seek a fair evaluation given the presence of any subjectiveness.

Table 4 illustrates the top 3 micro-reviews for 3 sample Foursquare venues, as ranked by the SAMR(vt) scheme. For the venue 'Fish & Co.', the first and third micro-reviews are rather unexpected for a food venue. The second tip on the swordfish collar is not mentioned by most customers and may be an unexpected positive comment. For 'Meng Kitchen Traditional Taste', users had focused on restaurant name and certain easily overlooked aspects of food and service. Lastly, one can see much comparison with competitors for 'Texas Chicken'. As described in the introduction, this is one reason where micro-reviews can be unexpected. Importantly, such micro-reviews can help the venue owner in improving his products/services to beat or at least match the competition.

Fish & Co. (Seafood/American Restaurant)
There's a cute girl here ;) hahaha!
Still the best in cooking the Swordfish Collar!
eat pray love
Meng Kitchen Traditional Taste (Asian/Chinese Restaurant)
True fact. Did u know that the correct name of meng's kitchen is actually Ming Fa.
Omg the uncle got some problem listening I said I don wan chili twice and end up with chili again
All the noodle is tasty! But the portion and ingredient always get lesser and lesser):
Texas Chicken (Fried Chicken Joint, Fast Food Restaurant)
Will never eat here again.. Popeye n KFC is better..
Get out of here. Their chicken's terrible. Cross over to Novena Square for KFC instead. If you have money, Kenny Rogers. If you have time, head to Popeye's at Toa Payoh
Chicken like rubber

Table 4: 3 sample Foursquare venues (names in bold) and their top 3 unexpected micro-reviews listed below respective venues. Venue categories are in brackets.

Topics. Next, we show that the SAMR model is able to surface topics discussed by users. The topics indicate user interest and are represented by positive deviations in certain words. Table 5 presents the top 10 positive deviating words for each topic. The topics are easily interpreted, hence indicating that the SAMR model is appropriate.

We also recap the notion of deviations. For example if the micro-review topic is on 'Service' in Table 5, the word 'rude' will deviate positively from the background such that it has higher occurrence probability.

Venue Types. Lastly our model provides a topic for each micro-review. Since each venue is described by multiple micro-reviews, it is easy for one to obtain a representation per venue in terms of distribution over topics. Such a representation is indicative of the types of venues. For example, one will expect a dining venue to have a mixture of food and service-related topics. In addition, the representation granularity is adjustable. If a more fine-grained representation is desired, one should define a larger number of topics, i.e. parameter K.

For example, Figure 2 illustrates the topic distribution for the venue 'Meng Kitchen Traditional Taste', where the topic labels are from Table 5. This venue has a Foursquare category of Asian/Chinese restaurant. As can be seen, topics on local cuisine are dominant, followed by service and dining style, thus providing a good indication of the venue type.

To conclude this section, qualitative analysis assures us that the topic model provides reasonable results that can be interpreted properly. We next cover quantitative evaluation.

5. EVALUATION WITH ANNOTATORS

5.1 Experiment Setup

In this section, we present quantitative evaluation results using human annotators. Our dataset is the set of micro-reviews from Singapore Foursquare venues, which we have

Service: rude slow waited attitude customer orders mins poor waitress customers
Relaxation: chill hang relax band live ambience quiet music atmosphere beers
Western Cuisine: garlic bacon mushroom cheese onion potato olio aglio ink fries
Desserts: cream chocolate ice vanilla strawberry caramel blended mocha choc milo
Local Cuisine 1: tom yong foo tau chop rice cutlet curry yam chicken
Local Cuisine 2: kway mee teow goreng hokkien nasi lemak hoon bee chor
Dining Style: xiao reasonable bao prices quality price western affordable sashimi buffet
Bubble Tea: bubble jelly pearl milk juice sugar gong tea koi pearls
Promotions: card discount call appreciated treasure license daily forget check save
Transport/Amenities: bus station mrt park marina exit interchange car wifi mall

Table 5: Top 10 positive deviating words for each topic. The topic labels are manually assigned.

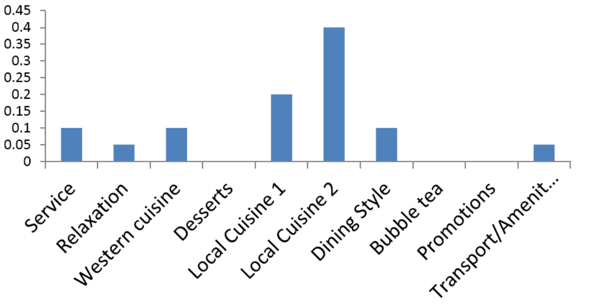


Figure 2: Topic distribution of 'Meng Kitchen Traditional Taste', as aggregated from micro-reviews. The venue is an Asian/Chinese restaurant.

covered in the previous section. We invite 6 participants not involved in this work to compare ranked pairs of micro-reviews. All participants have resided in Singapore for many years and are generally familiar with the characteristics of local venues.

For ease of voting, we limit voting options to two: χ^2 and SAMR(vt) scoring schemes, which we determine to be best performing from a separate experiment (see Section 6). First we rank micro-reviews using χ^2 and SAMR(vt) schemes, such that we obtain a ranked list for each venue per scheme. Then we select 30 venues whose results are least correlated in terms of the Spearman rank coefficient, i.e. venues where the schemes disagree the most. For each of this venue, we select 3 pairs of micro-reviews where the competing schemes disagree the most in terms of ranks. For each pair, the annotator is asked to put himself in the venue owner's perspective and decide which micro-review is relatively more unexpected. To assist the annotator, we also provide the venue category as indicated in Foursquare. The annotator can skip pairs for which he is undecided.

For example, Table 6 illustrates a sample pair for voting. The first micro-review is one that is ranked high by one scheme, say A and low by the other scheme, say B. The opposite is true for the second micro-review. Hence if the annotator regards the first micro-review as more unexpected, then it is equivalent to a vote for scheme A. Also note that annotators are not aware of the underlying scoring schemes.

cheese baked potato is a must-have!
It's the only second Halal Outlet!!

Table 6: Sample micro-review pair from a venue with category 'Diner and American Restaurant' for voting.

We also quantify agreement between annotator pairs with Bangdiwala's B-statistic [24]. Let x_{ij} be the number of pairs where the first annotator had voted for scheme i and the second annotator had voted for scheme j. The B-statistic ranges from 0 to 1 and is defined as:

$$B = \frac{\sum_{j=1}^{2} x_{jj}^2}{x_{1\cdot}x_{\cdot 1} + x_{2\cdot}x_{\cdot 2}} \qquad (17)$$

where $x_{1\cdot} = \sum_{j=1}^{2} x_{1j}$ and all summations are over decided pairs. We use the B-statistic instead of Cohen's Kappa [9] due to the following paradox [3]: when marginals are imbalanced, Kappa is low even when observed agreement is high. In our case, SAMR(vt) attracting the majority of the votes will push down the Kappa value.

5.2 Results

For the least correlated 30 venues between χ^2 and SAMR(vt) schemes, the average Spearman coefficient is -0.45. From these venues, we then derive 90 micro-review pairs for voting. Table 7 displays the voting results. Separately Table 8 displays the agreement between annotators.

Annotator	SAMR(vt)	χ^2	Undecided
1	66 (75.0)	22 (25.0)	2
2	68 (75.56)	22 (24.44)	0
3	61 (67.78)	29 (32.22)	0
4	33 (78.57)	9 (21.43)	48
5	58 (64.44)	32 (35.56)	0
6	54 (60.0)	36 (40.0)	0

Table 7: Voting results by annotators. Numbers are vote counts and bracketed numbers are percentage in favor of each scheme, considering each annotator's decided pairs.

Annotator	1	2	3	4	5	6
1	-	0.630	0.622	0.870	0.557	0.549
2	-	-	0.498	0.775	0.527	0.373
3	-	-	-	0.684	0.433	0.501
4	-	-	-	-	0.527	0.559
5	-	-	-	-	-	0.423

Table 8: B-statistic agreement between annotators, considering only decided pairs. We show only the upper half of the symmetric table for less cluttering.

Table 7 shows that all annotators favor SAMR(vt) in surfacing unexpected micro-reviews. For example, annotator 1 casted 66 votes for SAMR(vt) and 22 votes for χ^2, leaving 2 pairs where he's undecided. The vote proportion ranges from 60% to 70+% across all annotators. On average, SAMR(vt) is favoured with 70.2%, compared with 29.8% for χ^2, suggesting that SAMR(vt) is clearly superior. In tandem, Table 8 shows that the annotators' agreement with one another is substantially higher than what is expected by chance, i.e. value 0. On average, the B-statistic value is 0.568.

We also note that annotator 4 is highly conservative. Compared to the others, he had many more undecided pairs. Nonetheless considering only his decided pairs, the agreement rate with others is still high. This implies that he is voting only for pairs where his confidence is high and fast decisions can be made. For other annotators, votes were casted for most pairs, leading to more disagreements. However overall agreement is still high and in favor of SAMR(vt).

The annotators did comment that it is easier to decide for some micro-review pairs while certain pairs require more subjectivity to judge. This is also evident from examining the voting results since 'easier' pairs should experience more agreement. For example, for Table 6, 4 out of 6 annotators select the second micro-review as being relatively more unexpected (also ranked high by SAMR(vt)), supporting the notion that micro-reviews about an American restaurant being halal may be less expected than one describing a staple dish of cheese baked potato. In other comments, an annotator mentioned that if he is the venue owner, micro-reviews with negative sentiments will be more unexpected, since his goal would have been to provide good products and services.

Next, we move on to a second set of quantitative evaluation, one that is totally devoid of any annotator subjectivity.

6. EVALUATION WITH PSEUDO-VENUES

6.1 Pseudo-venues

Although we have conducted experiments with human annotators, we desire evaluation on an even larger scale, e.g. involving hundreds of venues and thousands of micro-reviews. Such an experiment is extremely expensive using annotators. To mitigate any human subjectivity, it is also desired to have as many annotators as possible. In view of these difficulties, we propose a more scalable experimental approach that can be easily applied on large corpora to assess ranking accuracies.

The idea is to mix micro-reviews from a pair of venues to construct *pseudo-venues*. Basically each pseudo-venue contains dominantly the micro-reviews of one venue and a small fraction of micro-reviews that are injected from another venue of a different function. The injected micro-reviews can then be treated as unexpected micro-reviews.

For example, consider a Chinese restaurant pseudo-venue. First sample a Chinese restaurant, and some of its micro-reviews. We then inject a small fraction of micro-reviews from another venue that is *not* a Chinese restaurant, e.g, an Ethiopian restaurant. The injected micro-reviews should then be relatively more unexpected (and thus ranked higher) since the focus should be on Ethiopian food or other aspects more specific to Ethiopian restaurants. We stressed that this applies in the probabilistic and not absolute sense, as injected micro-reviews can mention generic, widely applicable characteristics, e.g. slow service or opening hours. Some non-injected micro-reviews can be unexpected as well, e.g. bugs in the Chinese restaurant.

Pseudo-venue construction is easy since Foursquare venues are already categorized by functions. There are 9 coarse categories[3], e.g. shopping, food etc., which are broken down into more fine categories, e.g. shoe store, Turkish restaurant. At each level of the hierarchy tree, there are large and small categories with different number of member venues. In our experiments, we have constructed each pseudo-venue by mixing micro-reviews from a large and a small category.

In summary, we construct a pseudo-venue as follows:

1. Designate a set of small and a set of large categories. From each of the 2 sets, sample a category.
2. Sample a venue from the sampled small category→ $venue_s$ and a venue from the sampled large category→ $venue_l$.
3. Let $n_s \leq n_l$. Sample n_s micro-reviews from $venue_s$'s micro-reviews → $\{\mathfrak{m}_s\}$ and n_l micro-reviews from $venue_l$'s micro-reviews→ $\{\mathfrak{m}_l\}$
4. Construct $\{\mathfrak{m}_s\} \cup \{\mathfrak{m}_l\}$ as the pseudo-venue's micro-reviews. Regard $\{\mathfrak{m}_s\}$ as ground truth micro-reviews that are unexpected for the pseudo-venue.

Table 9 illustrates a pseudo-venue used in our experiments. The pseudo-venue was constructed using 5 micro-reviews from a theme park as $venue_s$ and 15 micro-reviews from a food venue as $venue_l$.

Universal Studios Singapore ($venue_s$), n_s=5
Ride the Mummy's revenge.
definitely have to try the mummy's revenge roller coaster ride!!!
Revenge of the Mummy, highly recommended! Shrek 4-D Adventure is pretty good as well!
Shrek 4D Adventure is a must-go for all families. A 3D adventure with 4D effects! Absolute fun but do return your Ogre-Vision glasses after the show!
The crews are all friendly!!! Nice management :) too bad cyclone-human hasn't opened yet :(
Beach Road Prawn Mee Eating House ($venue_l$), n_l=15
Great prawn & pork rib noodle (soup)
Nice but expensive. Can't eat it every week...
Have the Ngo Hiang as side dish with the prawn noodles... Great combination!
They should really start to serve soft drinks as well!
Prawn Noodles here is a Die Die Must Try!! Must have it with the tasty soup

Table 9: An example pseudo-venue constructed from 2 actual venues (in bold). For brevity, only a small sample of micro-reviews from the food venue ($venue_l$) are shown. Food items mentioned in its micro-reviews are local cuisine, e.g. Ngo Hiang is a form of minced pork roll.

6.2 Experiment Setup

We use Singapore and Malaysian Foursquare venues separately to generate datasets of 2 different settings: S5+15 and M5+15. For each setting, we generate 10 datasets which are different due to sampling conducted for their pseudo-venues. Each dataset consists of 100 pseudo-venues and remaining venues that have not been sampled for constructing pseudo-venues. Note that ranking accuracies can only be computed from pseudo-venues. We also filter off venues with less than 15 micro-reviews. In actual applications, ranking is unnecessary if a venue has too few micro-reviews.

[3] https://developer.foursquare.com/categorytree

S5+15: This uses Singapore venues. Each pseudo-venue consists of 20 micro-reviews with 5 as the ground truth unexpected micro-reviews, i.e. $n_s = 5, n_l = 15$. We designate the 3 largest and 3 smallest categories as large and small categories respectively. The large categories are 'Food', 'Shop & Service' and 'Residence'. The small categories used are 'Arts & Entertainment', 'College & University' and 'Nightlife Spot'. The unprocessed data consists of 2,827 and 20,769 venues in both category sets. Filtering out venues with too few micro-reviews and including the pseudo-venues, each dataset consists of 3,100+ venues with 56,000+ micro-reviews.

M5+15: We use Malaysian restaurant venues with $n_s = 5, n_l = 15$. The categories are now more fine-grained and in terms of restaurant types. We use the largest 3 restaurant categories ('Asian', 'Malay', 'Chinese') as large categories. This encompasses 3,178 venues. We regard other restaurant categories as small categories (1,974 venues), excluding 483 venues with unclear restaurant categories. On average, each processed dataset consists of around 1,900 restaurant venues with 35,500 micro-reviews.

For both settings, we apply all previously described approaches: TF-IDF, χ^2, Y and SAMR topic model with SAMR(v) and SAMR(vt) scoring schemes. For the SAMR settings, we use 20 topics. We omitted tuning the number of topics although this may potentially achieve even better ranking accuracies. For model inference, we use 25 EM iterations, with 50 gradient descent iterations in the M-step. The learning rate τ was set at 2.0E-4.

6.3 Accuracy Metrics

Mean Precision. We use the mean precision at position k, **MP(k)** to measure ranking accuracy. First we compute the precision at k for each pseudo-venue. Given a pseudo-venue's k highest ranked micro-reviews $\{\mathfrak{m}\}_k$, this is the proportion of micro-reviews that are unexpected, i.e. $Prec(k) = \{\mathfrak{m}\}_k \cap \{\mathfrak{m}_s\}/k$.

As described in the previous section, we generate multiple datasets for testing, each with 100 pseudo-venues. Hence, we average $Prec(k)$ over multiple pseudo-venues and datasets to obtain the mean precision.

Mean Average Precision. We denote the mean average precision over multiple pseudo-venues and datasets as **MAP**. This is based on Average Precision (AP), which has been widely used in document retrieval tasks. For a pseudo-venue, AP attains a perfect accuracy of 1 if all micro-reviews from $\{\mathfrak{m}_s\}$ are ranked higher than all micro-reviews from $\{\mathfrak{m}_l\}$. AP can be computed as:

$$AP = \sum_i Prec(i)\Delta r(i) \qquad (18)$$

where $\Delta r(i)$ is the change in recall from position $i-1$ to i. For each pseudo-venue, we also evaluate the summation over all its ranked micro-reviews (instead of just the top k).

6.4 Results

Tables 10 and 11 present the mean precision (MP) and mean average precision (MAP) figures for the settings: S5+15 and M5+15 respectively. Standard deviations are bracketed. For each setting, we applied the 5 discussed scoring schemes. Also recall that SAMR(v) and SAMR(vt) are different scoring schemes based on the same topic model SAMR.

Firstly we note the low accuracy figures obtained across the board. This can be explained by our observations that many injected micro-reviews cover common food items or generic issues, e.g. service, while some non-injected micro-reviews can be unexpected as well. These have the effect of lowering ranking precision in pseudo-venues. However we are primarily interested in *how the competing approaches perform relative to each other*. The experiment setup and results are already adequate for such an analysis.

	MP(1)	MP(5)	MAP
SAMR(vt)	**0.289 (0.037)**	**0.330 (0.024)**	**0.420 (0.016)**
SAMR(v)	0.095 (0.018)	0.108 (0.012)	0.272 (0.006)
TF-IDF	0.167 (0.027)	0.169 (0.018)	0.308 (0.013)
χ^2	0.277 (0.041)	0.253 (0.019)	0.369 (0.012)
Y^2	0.226 (0.032)	0.211 (0.016)	0.338 (0.012)

Table 10: Results on Singapore venues (S5+15). Standard deviations are bracketed. Best results are bolded.

	MP(1)	MP(5)	MAP
SAMR(vt)	**0.252 (0.056)**	**0.269 (0.029)**	**0.370 (0.025)**
SAMR(v)	0.098 (0.023)	0.122 (0.018)	0.276 (0.012)
TF-IDF	0.143 (0.031)	0.142 (0.014)	0.287 (0.013)
χ^2	0.226 (0.032)	0.193 (0.015)	0.323 (0.013)
Y^2	0.185 (0.033)	0.165 (0.016)	0.303 (0.015)

Table 11: Results on Malaysian restaurant venues (M5+15).

For all experiment settings, SAMR(vt) consistently outperforms other approaches across all metrics. The second best performer is the χ^2 scheme. Comparing these 2 approaches with the Wilcoxon signed rank test, the difference is *statistically significant* (beyond p-value of 0.05) for most settings and metrics: MP(5), MAP for S5+15 and M5+15.

χ^2 is based on differences against a simple model for expected word usage while SAMR(vt) offsets information from topics and the fact that topics and expected word usage are dependent on venues. Hence the latter is a richer model that is able to explain more of the expected, in deriving the unexpected. Both approaches also outperform the simple TF-IDF baseline.

We also note that Y^2, which is the continuity corrected version of χ^2, underperforms the latter in all settings. Thus approximation error in computing the χ^2 statistic for small data (i.e. rare words with limited usage) is not crucial for ranking accuracy here.

Interestingly SAMR(v) is consistently the worst performing approach in all settings. The single difference between SAMR(v) and SAMR(vt) is that in the former, we do not consider micro-review topics and use only the venue-dependent word deviations for scoring words. In contrast, SAMR(vt) offsets the effect of micro-review topics and achieves better accuracies. Simply put, if a micro-review contains words that are easily explained by just knowing its topic indicator, then the words are less likely to be unexpected. It is thus necessary to subtract off the topic-dependent word deviation as what SAMR(vt) has done.

In summary, supported by consistent results from 2 different settings, we now conclude that SAMR(vt) outperforms other approaches.

7. CONCLUSION

We have proposed the problem of identifying unexpected micro-reviews to serve the needs of venue owners. We envisage that the results can be used for various purposes, such as event detection, service improvement or identifying competitors. We then explore various approaches to solve the problem. Our best performing approach is scores derived from a novel topic model, SAMR, with which we use to account for the 'expected' and derive the unexpected.

Our work is extendible in several aspects. For example, it will be useful to explain why a micro-review is unexpected. This requires much deeper modeling and analysis of the context. Micro-reviews can then be clustered by their underlying reasons.

Related to the above, one can apply sentiment analysis on top of the current work. Obviously, a micro-review can be unexpected in either a positive, negative or neutral manner. A highly negative unexpected micro-review may be cause for concern. On the other hand, a highly positive unexpected micro-review may be useful in publicity or branding strategies.

Lastly, our proposed topic model can be adapted to other domains such as traditional reviews from ecommerce platforms. A key point is that traditional reviews are usually much longer than micro-reviews and hence the assumption of one topic per review will no longer hold. With some modifications, we can model multiple topics in a single review. The scoring schemes will need to be adjusted as well.

8. ACKNOWLEDGEMENTS

This research is partially supported by DSO National Laboratories, Singapore; and the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

9. REFERENCES

[1] A. Aggarwal, J. M. Almeida, and P. Kumaraguru. Detection of spam tipping behaviour on Foursquare. *WWW*, 2013.

[2] A. C. Alhadi, T. Gottron, J. Kunegis, and N. Naveed. Livetweet: Monitoring and predicting interesting microblog posts. *ECIR*, 2012.

[3] F. AR and C. DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*, 1990.

[4] M. Baroni and S. Evert. Statistical methods for corpus exploitation. *Corpus linguistics: An international handbook*, 2:777–802, 2009.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, March 2009.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3(993):1022, 2003.

[7] D. Carlone and D. Ortiz-Arroyo. Semantically oriented sentiment mining in location-based social network spaces. *FQAS*, 2011.

[8] N. Chen, J. Lin, S. C. H. Hoi, X. Xiao, and B. Zhang. AR-miner: mining informative reviews for developers from mobile app marketplace. *ICSE*, 2014.

[9] J. Cohen. A coefficient of agreement for nominal scales. *Educ Psychol Meas*, 1960.

[10] J. Eisenstein, A. Ahmed, and E. Xing. Sparse additive generative models of text. *ICML*, pages 1041–1048, 2011.

[11] B. S. Everitt. *The analysis of contingency tables.* Chapman and Hall, 1992.

[12] A. Ghose and P. G. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. *ICEC*, 2007.

[13] K. Hofland and S. Johansson. *Word frequencies in British and American English.* The Norwegian Computing Centre for the Humanities, 1982.

[14] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. *WWW*, 2011.

[15] M. Hu and B. Liu. Mining and summarizing customer reviews. *KDD*, 2004.

[16] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. *EMNLP*, 2006.

[17] T. Lappas, M. Crovella, and E. Terzi. Selecting a characteristic set of reviews. *KDD*, 2012.

[18] T. Lappas and D. Gunopulos. Efficient confident search in large review corpora. *ECML/PKDD*, 2010.

[19] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. *ICDM*, 2008.

[20] F. Moraes, M. Vasconcelos, P. Prado, D. Dalip, J. Almeida, and M. Gonçalves. Polarity detection of Foursquare tips. *SocInfo*, 2013.

[21] T.-S. Nguyen, H. W. Lauw, and P. Tsaparas. Using micro-reviews to select an efficient set of reviews. *CIKM*, 2013.

[22] D. Roland, D. Jurafsky, L. Menn, S. Gahl, E. Elder, and C. Riddoch. Verb subcategorization frequency differences between business-news and balanced corpora: the role of verb sense. *WCC*, 9, 2000.

[23] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 1988.

[24] V. Shankar and S. I. Bangdiwala. Observer agreement paradoxes in 2x2 tables: comparison of agreement measures. *BMC Medical Research Methodology*, 2014.

[25] P. Tsaparas, A. Ntoulas, and E. Terz. Selecting a comprehensive set of reviews. *KDD*, 2011.

[26] I. Uysal and W. B. Croft. User oriented tweet ranking: A filtering approach to microblogs. *CIKM*, 2011.

[27] M. Vasconcelos, J. Almeida, and M. Gonçalves. What Makes your Opinion Popular? Predicting the Popularity of Micro-Reviews in Foursquare. *SAC*, 2014.

[28] T. Virtanen. The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English. *ICAME*, 17, 1996.

[29] H. Wu, R. Luk, K. Wong, and K. Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26(3), 2008.

[30] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. *CIKM*, 2010.

[31] F. Yates. Contingency tables involving small numbers and the chi-squared test. *Journal of the Royal Statistical Society Supplement*, 1, 1934.

[32] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. *CIKM*, 2006.