Going Beyond Corr-LDA for Detecting Specific Comments on News & Blogs

Mrinal Das† mrinal@csa.iisc.ernet.in
Trapit Bansal† trapit@csa.iisc.ernet.in
Chiranjib Bhattacharyya† chiru@csa.iisc.ernet.in

† Department of Computer Science and Automation, Indian Institute of Science, India

ABSTRACT

Understanding user generated comments in response to news and blog posts is an important area of research. After ignoring irrelevant comments, one finds that a large fraction, approximately 50%, of the comments are very specific and can be further related to certain parts of the article instead of the entire story. For example, in a recent product review of *Google Nexus 7* in ArsTechnica (a popular blog), the reviewer talks about the prospect of "*Retina equipped iPad mini*" in a few sentences. It is interesting that although the article is on *Nexus 7*, but a significant number of comments are focused on this specific point regarding "*iPad*". We pose the problem of detecting such comments as *specific comments* location (SCL) problem. SCL is an important open problem with no prior work.

SCL can be posed as a correspondence problem between comments and the parts of the relevant article, and one could potentially use Corr-LDA type models. Unfortunately, such models do not give satisfactory performance as they are restricted to using a single topic vector per article-comments pair. In this paper we propose to go beyond the single topic vector assumption and propose a novel correspondence topic model, namely SCTM, which admits multiple topic vectors (MTV) per article-comments pair. The resulting inference problem is quite complicated because of MTV and has no off-the-shelf solution. One of the major contributions of this paper is to show that using stick-breaking process as a prior over MTV, one can derive a *collapsed Gibbs sampling* procedure, which empirically works well for SCL.

SCTM is rigorously evaluated on three datasets, crawled from *Yahoo! News* (138,000 comments) and *two blogs*, ArsTechnica (AT) Science (90,000 comments) and AT-Gadget (160,000 comments). We observe that SCTM performs better than Corr-LDA, not only in terms of metrics like *perplexity* and *topic coherence* but also discovers more unique topics. We see that this immediately leads to an order of magnitude improvement in F1 score over Corr-LDA for SCL.

WSDM 2014, February 24 - 28 2014, New York, NY, USA
Copyright 2014 ACM 978-1-4503-2351-2/14/02 ...$15.00.
http://dx.doi.org/10.1145/2556195.2556231.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

Keywords

specific, correspondence, comments, news, blogs

1. INTRODUCTION

Comments on news and blogs naturally promote user participation and play a pivotal role in increasing the popularity of host websites. In this paper we investigate the correlation structure between the content, such as news or blog-postings, and user generated comments.

Consider the example in Figure 1 that shows an excerpt from the review on *Google Nexus 7*, dated July 31 2013[1]. It is one of the well commented recent reviews with more than 200 comments. Some comments are related to the entire article, like *... good to see tablets are becoming cheaper* (black colored in Figure 1), which are *general comments.* Some comments are *irrelevant*, for example *I work on a laptop* (blue in Figure 1).

In the same article, in about 8% of the sentences, the author talks about retina display and compares with "*iPad*" (colored green in Figure 1). It was found that, leaving irrelevant comments, around 22% comments (colored green in Figure 1) are made only on that specific point. The fact of 22% comments on a topic covered in 8% of sentences, and those sentences being far different from the central theme of the article, is quite astonishing. These comments are not useless, rather they provide valuable user feedback. We call such comments as *specific comments.* Discovering *specific comments* and the associated part of the content which is specific to the comment will be called the *specific comment location* (SCL) problem.

Solving SCL will open up many opportunities such as accumulating user feedback, analysing market trends, improving user experience etc. There is no prior work related to SCL, furthermore there are no off-the-shelf techniques which could be adapted to solve this problem. Discriminative techniques are often accurate but require lot of labeled training data which in this case is not available. Past approaches such as [11] on comment understanding demonstrates that accuracy of discriminative approaches depend

[1] arstechnica.com/gadgets/2013/07/the-2013-nexus-7-review-meet-the-new-standard-for-android-tablets/

on cleverly crafted features. Keeping this in mind we pursue topic models, an unsupervised approach, to the SCL problem. Correspondence LDA (Corr-LDA) [1] is an interesting topic model which could be a potential candidate for SCL. Existing topic models, including Corr-LDA, assign a single topic vector to each document. This makes Corr-LDA more suitable for discovering general comments and as empirical evidence suggests it is not suitable for SCL. In this paper we address the issue of SCL and make the following contributions.

Contributions: We introduce *specific correspondence topic models* (SCTM), based on the notion of *multiple topic vectors* (MTV) as opposed to single topic vector in the state of the art models. Moreover, in order to handle wide variety of comments we enhance the diversity among topics through sparsity. The inference becomes non-standard due to MTV and sparsity. We explore a *stick-breaking process* (SBP) as a prior over MTV. One major contribution of this paper is a collapsed Gibbs sampling inference procedure for SCTM. The resultant algorithm converges fast and also leads to simple update equations which are easy to implement. Using three real world datasets, we evaluate the proposed approach in two aspects. (i) Using perplexity and topic coherence we show that SCTM models data better than the state of art. Then, we demonstrate that SCTM discovers more diverse set of topics and converges faster in terms of likelihood. (ii) Using precision-recall we compare SCTM with the baseline on the task of discovering specific comments as well as aligning them to the respective sentences in the article. Finally, we show various use cases of SCTM on some interesting practical applications.

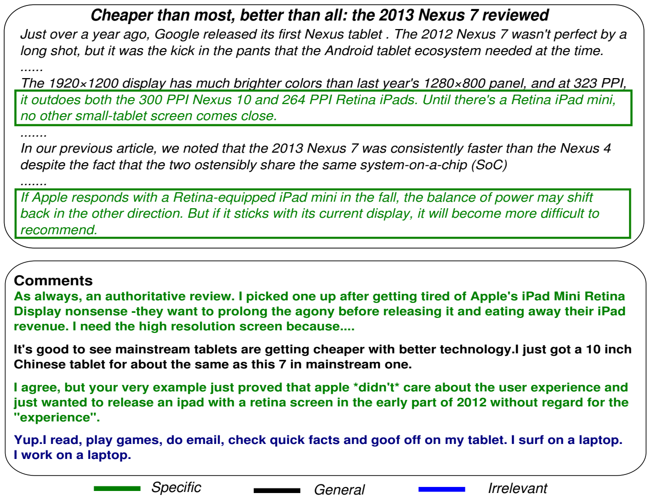
The paper is organized as follows. We formulate the problem of SCL and discuss the challenges and related works in section 2. Then in section 3 we present the proposed model and the Gibbs sampling inference algorithm. We describe the algorithm for SCL in section 4. In section 5, we provide empirical evaluation. Finally, section 6 presents some use cases of SCL.

2. THE PROBLEM OF SPECIFIC COMMENT LOCATION

In this section we introduce the problem of specific comment location (SCL) and discuss the difficulties in resolving SCL. We begin by introducing relevant notation.

Notation: The set of element wise positive d dimensional vectors will be denoted by $\mathbb{R}_+^d$. $\sim$ means "distributed as", 1_x is a x dimensional vector with all entries as 1. K is the number of topics and V is the number of words in the vocabulary. β_k is a V dimensional vector such that $\sum_{j=1}^{V}\beta_{kj} = 1$, popularly called as a "topic". $x.y$ is element wise product of two vectors x and y of same dimension. *Dir* denotes Dirichlet distribution, U denotes discrete uniform distribution and *mult* represents multinomial distribution. $[n] = \{1, 2, \ldots, n\}$ and $|R|$: cardinality of set R. $\tilde{x}$ means a set of variables of same type. $I[.]$ is the indicator function.

Definition: A news or a blog article A_d, indexed by d, is a collection of S_d number of sentences. More explicitly $A_d = \{s_{da} | a \in [S_d]\}$, where each sentence is denoted by $s_{da} = \{w_{dai} | i \in [n_{da}]\}$. The ith word in the ath sentence s_{da} is denoted by w_{dai} and the number of words in s_{da} is denoted by n_{da}. Corresponding to each article, A_d there is a set of comments denoted by $C_d = \{c_{dej} | j \in [n_{de}],\ e \in [E_d]\}$ where c_{dej} is the jth word in the eth comment related to article A_d. Number of comments on A_d is denoted by E_d and n_{de} is the number of words in the eth comment. Furthermore $w_{dai}, c_{dei} \in [V]$, where V is the number of unique words in the entire corpus. Let $R_{de} \subseteq A_d$ denote the set of sentences related to the eth comment of dth article. In Figure 1 the boxes show the sentences indexed by R_{de} for the green colored comment. The eth comment can be a general, irrelevant or specific comment depending on the size of R_{de}. It is a general comment when $|R_{de}|$ is *almost equal or equal* to S_d, and is an irrelevant comment when R_{de} is empty. Comment e is a specific comment if $0 < |R_{de}| \leq N_d$, where the parameter N_d is preset by the user and determines the granularity of specific comments. Notice that general comments and irrelevant comments are two extreme cases.



Cheaper than most, better than all: the 2013 Nexus 7 reviewed

Just over a year ago, Google released its first Nexus tablet . The 2012 Nexus 7 wasn't perfect by a long shot, but it was the kick in the pants that the Android tablet ecosystem needed at the time.

......

The 1920×1200 display has much brighter colors than last year's 1280×800 panel, and at 323 PPI,

it outdoes both the 300 PPI Nexus 10 and 264 PPI Retina iPads. Until there's a Retina iPad mini, no other small-tablet screen comes close.

....

In our previous article, we noted that the 2013 Nexus 7 was consistently faster than the Nexus 4 despite the fact that the two ostensibly share the same system-on-a-chip (SoC)

....

If Apple responds with a Retina-equipped iPad mini in the fall, the balance of power may shift back in the other direction. But if it sticks with its current display, it will become more difficult to recommend.

Comments

As always, an authoritative review. I picked one up after getting tired of Apple's iPad Mini Retina Display nonsense -they want to prolong the agony before releasing it and eating away their iPad revenue. I need the high resolution screen because....

It's good to see mainstream tablets are getting cheaper with better technology.I just got a 10 inch Chinese tablet for about the same as this 7 in mainstream one.

I agree, but your very example just proved that apple *didn't* care about the user experience and just wanted to release an ipad with a retina screen in the early part of 2012 without regard for the "experience".

Yup.I read, play games, do email, check quick facts and goof off on my tablet. I surf on a laptop. I work on a laptop.

Figure 1: Example from ArsTechnica. For specific comments (in green), the corresponding sentences in the article (in green, referred as R_{de}) are few and are not contiguous. The box shows *hot-spot* (most commented sentences). Irrelevant comment (in blue) does not have any corresponding sentence in the article, whereas general comments (in black) are related to the entire article.

SCL formulation: The problem of specific comment location can now be formulated as identifying R_{de} given A_d and the eth comment whenever $|R_{de}| \leq N_d$.

For a specific comment e, the set R_{de} can be understood as *specific correspondence* between the comment and the article A_d. In this paper we concentrate on articles and comments but it is conceivable that similar problems exist in many other domains such as *technical paper and bibliography*, *image and tags* etc.

The difficulty in resolving SCL: It is not easy to discover R_{de}. Specific comments and general comments look very similar, and there are no distinguishing features such as length (number of words), presence of proper-nouns, or selection of words in comments. Absence of distinguishing features makes it impractical to apply rule based approaches.

In addition, due to lack of labeled data, supervised models are inapplicable. For example, [11] made an attempt to align comments to paragraphs in an article, however their method being supervised is restrictive and does not apply generally. An immediate alternative could be the use of *keyword* based searches.

Very low textual overlap between specific comments and the article: To measure the overlap in words between a specific comment and the main article, we define, $\Delta_{dae} = \frac{n_{dae}}{\sqrt{n_{da}}\sqrt{n_{de}}}$, where n_{dae} is the number of words in common between a^{th} sentence and e^{th} comment for article indexed by d. Δ_{dae} can be interpreted as cosine distance or correlation coefficient which computes the normalized similarity between a sentence in the main article with a comment. n_{da} and n_{de} are as defined before. Let D be the total number of (A_d, C_d) pairs then the average textual overlap over all sentences and the articles can be measured by $\Delta = \frac{1}{D}\sum_d \frac{1}{E_d}\sum_e \frac{1}{|R_{de}|}\sum_{a\in\{a|s_{da}\in R_{de}\}}\Delta_{dae}$.

From our study based on ArsTechnica gold-standard (see section 5), we find the following. For specific comments, considering all sentences of the article (i.e. using A_d instead of R_{de} in Δ above), Δ is 0.07 and for non-specific comments it is 0.06. However, considering only the relevant sentences for the specific comments, the value of Δ is 0.08. Notice that, not only the value is very small but also they are very similar for both specific and non-specific comments. For example in Figure 1, "*iPad*" is the only word which probably can link but due to other words the overlap becomes low. This makes keyword based approaches difficult to apply.

Topic models: Over the last decade Topic modeling has become ubiquitous in text analysis. It is an unsupervised approach rooted in Bayesian modeling. In the following we review some of the related topic models and discuss their suitability for SCL. There has been an increasing interest in using topic models such as [9, 15, 6] for understanding user generated content. [6] considered the problem of finding episodic tweets about an event, tweets which are related to one of the segments of an event. Their model is similar in spirit to [15], where event and tweet are exchangeable. As noted in [1] these models [15, 6] are not suited for understanding the dependency between a news article and comments. MG-LDA [12] has been developed to model local topics. A set of sliding windows are used across the sentences in a document which uses local topics. MG-LDA is inappropriate because: (i) local topics scatter across the corpus which is not the case here, and (ii) every segment or a set of contiguous sentences may not correspond to a comment. Among the existing models the most suited seems to be Corr-LDA [1] and we will use it as a baseline.

Review of Corr-LDA: Corr-LDA [1] is a topic model for understanding correspondences. It was initially proposed for modeling annotations on images. Corr-LDA uses a bag-of-words and there is no notion of sentences, i.e. $S_d = 1$. The generative process of Corr-LDA is as follows.

For each (A_d, C_d)

- Sample a topic vector $\theta_d \sim Dir(\alpha 1_K)$
- For each word w_{di}, $i \in [n_d]$
  - sample topic $z_{di} \sim mult(\theta_d)$
  - sample word $w_{di} \sim mult(\beta_{z_{di}})$
- For each comment $e \in [E_d]$
  - For each word $i \in [n_{de}]$
    - sample topic $y_{dei} \sim U(\tilde{z}_d)$
    - sample word $c_{dei} \sim mult(\beta_{y_{dei}})$

Notice that comment topic indices are sampled uniformly at random from article topic indices. Hence, y_{de} corresponds to the complete set $\tilde{z}_d$ and comment e becomes related to the entire article d. This lacks the *specific correspondence* focus of the current paper. Using Algorithm 1 (to be described in section 4) Corr-LDA can be used for SCL.

Limitations of Corr-LDA for SCL: We summarize some major limitations of Corr-LDA for SCL as below.

A.) Very low correspondence: Corr-LDA models R_{de} as the entire article, however R_{de} is not known a priori and is very small for specific comments. Based on the gold-standard on ArsTechnica (see section 5), we observe that on average only 3% of the sentences in an article are related to a specific comment, i.e. $|R_{de}|$ is very small. In Figure 1, for the specific comment on "*iPad*", $\frac{|R_{de}|}{S_d}$ is 0.08.

B.) Topical difference in R_{de}: A specific comment e is related to a small part R_{de} and is less relevant to the rest of the article. Therefore, R_{de} has a different topic proportion than the rest of the article. For example, in Figure 1, R_{de} should have high probability for "*iPad*" topic, although that is a topic with low overall probability in the article.

Note that in Corr-LDA there is a single topic vector θ_d for document d. Hence topic proportions in R_{de} is also θ_d, i.e. probability of "*iPad*" topic, for example, in Figure 1 is also low in R_{de}. This is a major drawback which we address in our proposed model SCTM.

3. SCTM: A CORRESPONDENCE TOPIC MODEL

In this section we define *specific correspondence topic models* (SCTM) and the associated inference algorithm.

3.1 SCTM

The need for multiple topic vectors (MTV): As pointed out in the last section, existing correspondence topic models use a single topic vector θ_d for each pair of article and the associated comments. This makes the topic proportions constant throughout the article and comments. Though this maybe suitable for general comments, where one might have similar topic proportions, but specific comments have different proportion over topics than the main article. Precisely, a specific comment e and R_{de} should have similar topic proportions. It is thus clear that one needs to go beyond the single topic vector assumption to resolve SCL. To this end we introduce the novel concept of using multiple proportions over topics per article-comments pair to model *specific correspondence.*

Modeling topical difference in R_{de} using MTV: Note that the challenge is that R_{de} is not known a priori, so that we can extract R_{de} from the main article and model the correspondence between R_{de} and comment e following Corr-LDA. We resolve this by introducing multiple topic proportions to be called as MTV (multiple topic vectors). Recently, [5] has used the concept of MTV in a different context to find subtle topics.

```
For each vocabulary term $v \in [V]$
    • Sample $\kappa_v \sim Beta(\frac{\nu\lambda}{V}, \lambda)$
For each topic $k \in [K]$,
    • For each term of the vocabulary $v \in [V]$
        – Sample word selector $\phi_{kv} \sim Bernoulli(\kappa_v)$
    • Draw topic distribution $\beta_k \sim Dir(\eta 1_{V+1}.\phi_k)$
For each article-comments pair $(A_d, C_d),\ d \in [D]$,
    • For $j \in [J_d]$, draw $\theta_{dj} \sim Dir(\alpha 1_K)$
    • For each article sentence $a \in [S_d]$,
        – Draw $\rho_{da} \sim stick(\tau, \iota)$
        – For each word $i \in [n_{da}]$,
            ∗ Draw $b_{dai} \sim mult(\rho_{da})$
            ∗ Draw $z_{dai} \sim mult(\theta_{db_{dai}})$
            ∗ Draw word $w_{dai} \sim mult(\beta_{z_{dai}})$
    • For each comment, $e \in [E_d]$
        – Sample $\pi_{de} \sim Beta(\varrho_1, \varrho_2)$
        – Sample $\epsilon_{de} \sim Beta(\varsigma_1, \varsigma_2)$
        – For each sentence $a$ of the article
            ∗ Draw $\xi_{dea} \sim Bernoulli(\pi_{de})$
        – Set $\varphi_{de} = \{z_{dai},\ \forall\ (a, i) \mid \xi_{dea} = 1\}$
        – For each comment word $i \in [n_{de}]$,
            ∗ Draw $t_{dei} \sim Bernoulli(\epsilon_{de})$
            ∗ If $t_{dei} = 1$, $y_{dei} = K+1$ (irrelevant topic)
            ∗ Else $y_{dei} \sim U(\varphi_{de})$
            ∗ Draw word $c_{dei} \sim mult(\beta_{y_{dei}})$
```

Figure 2: Generative process of SCTM. θ_{dj}: topic-proportion vector, ρ_{da}: distribution over topic-vectors, b_{dai}: topic-vector index, z_{dai}: topic assignment, ξ_{dea}: sentence selector, t_{dei}: irrelevant topic selector. *stick* refers to construction as in (1).

Therefore, per article-comments pair there are $J_d \geq 1$ topic vectors denoted by $\{\theta_{dj}\}_{j=1}^{J_d}$. While generating a word in a sentence, one θ_{dj} is selected randomly. So, proportions over topics in each sentence can be different and a random selection of R_{de} from article d can have very different proportions over topics than that of article d. For example, even if "*iPad*" topic has low probability in the entire article in Figure 1, it may have high probability in some R_{de}.

Choice of a suitable prior for MTV: The choice of a suitable prior for MTV is an important question. The challenge is that, (i) in the SCL problem, it may occur that some topic has an extremely low probability in the entire document but might have a very high probability in a specific region and can be useful in detecting specific comments. (ii) Topic vectors should be shared across the sentences in an article because for a specific comment e on article d, R_{de} may span over multiple non-contiguous sentences. Important to note that, over-estimation of J_d will try to model many general comments as specific in nature, whereas under-estimation will try to model specific comments as general ones. Nonparametric priors such as dependent Dirichlet process can mitigate this issue. One could potentially use hierarchical Dirichlet process (HDP) which is suited to share topics as required in (ii). However the difficulty with HDP is its *rich getting richer property* which tends to make a widely appearing topic vector highly probable in all the sentences, contradicting the requirement (i). We will explore SBP as an alternative prior which could address these issues.

Stick-breaking process (SBP): SBP[7] is defined as follows. Let $\tau, \iota \in \mathbb{R}_+^J$ and Γ is a diffuse probability measure on a measurable space $(\Omega, \mathcal{B})$. A random probability measure G on $(\Omega, \mathcal{B})$ is a stick-breaking prior if $G = \sum_{j=1}^{J} \rho_j \delta_{\theta_j}$, $\theta_j \overset{iid}{\sim} \Gamma$ and (ρ_j) are constructed as

$$\rho_1 = v_1, \quad \rho_j = v_j \prod_{l=1}^{j-1}(1 - v_l),\ j > 1 \tag{1}$$

where $v_j \overset{iid}{\sim} Beta(\tau_j, \iota_j)$, $j < J$, and $v_J = 1$. To keep the exposition simple we assume J to be finite. δ_{θ_j} denotes an atomic distribution where the entire probability mass is concentrated at θ_j. Γ is a measure defined on vectors θ_j, commonly referred as atoms of G. Furthermore by construction $\sum_{j=1}^{J} \rho_j = 1$. For each sentence s_{da} in article A_d we use a distribution G_{da} with topic vectors (θ_{dj}) as atoms. In our case the topic vectors are sampled from a Dirichlet prior $Dir(\alpha 1_K)$. For a fixed d assume that (ρ_{da}) is defined as in equation (1) then

$$\theta_{dj} \overset{iid}{\sim} Dir(\alpha 1_K),\ G_{da} = \sum_{j=1}^{J_d} \rho_{da} \delta_{\theta_{dj}} \tag{2}$$

Notice that, the proportion over the topic vectors for a sentence does not depend on those of other sentences, nor does it depend on any document wide property. Unlike Dirichlet distribution, J_d acts as an upper limit on number of topic vectors, where higher indexed topic vectors get lower prior probabilities. Thus SBP allows to provide high weight on a topic vector for some sentences, whereas for most of the sentences that topic vector has low probability. Unfortunately though SBP is suitable, as it is not commonly known in machine learning, makes the inference non-standard.

Generation of R_{de} We randomly select a set of sentences in article A_d as R_{de}. Random selection of R_{de} and use of MTV vary the topic proportions in R_{de} making SCTM appropriate for SCL. Note that, alternatively, R_{de} could be sampled first and then corresponding to each R_{de} a topic vector could be used. But this has several problems: (i) even if R_{de} is same for two comments it will have two different θ, (ii) articles with large number of comments will create too many topic vectors, (iii) modeling article will become dependent on comments. So we have avoided this choice.

Enhancing diversity in topics: Note that, topics related to specific comments contain words which are very rare in the article. For example, "*iPad*" has rarely appeared in the article. Now considering a web scale data we need to consider a vast and diverse set of article-comments. Hence, it is required that we detect a diverse set of topics which have high probability for different sets of words.

We model this fact by modifying the definition of topic slightly. Instead of defining a topic over all the words in the vocabulary, similar to the selection of R_{de} we randomly select a subset of words using ϕ and define a topic over that subset of words. ϕ is a Bernoulli random variable and when $\phi_{kv} = 0$, $\beta_{kv} = 0$ with probability one, where β_{kv} is the probability of word indexed by v in the k^{th} topic. To avoid accidental situations, we use ϕ_k of length $V + 1$, and set $\phi_{k,V+1} = I[\sum_{i=1}^{V} \phi_{ki} = 0]$. Similar concept has been explored by [13] for decoupling sparsity and smoothness in topics. Unlike [13], we have two parameters λ and ν, where λ is a *repulsion parameter*. When λ increases, diversity across ϕ increases, in turn increasing diversity across β, i.e. topics.

Irrelevant topic: Irrelevant topic is used for modeling comments to take care of words which generally appear in comments but not in main articles. These words are uninformative words, motivating the name of the topic. For example in Figure 1, "Yup" in the irrelevant comment is one such word. Irrelevant topic models a topic which is irrelevant to any type of correspondence between article and comment. It has been observed that irrelevant comments mostly contain this topic. Similar technique has been explored in [9].

3.2 Collapsed Gibbs sampling inference

In this section we describe an efficient and easy to implement Gibbs sampling based inference procedure for SCTM. In order to achieve accelerated convergence, we need to marginalize out the real valued random variables β, θ, ϵ, ρ, π and κ, in the generative process (Fig. 2) and infer the latent variables z, y, b, ξ and ϕ. The major challenges are: (1) sampling b, due to SBP the inference becomes non-standard and (2) ξ, ϕ are binary random variables demanding novel inference mechanisms.

We address the first challenge by noting the relationship between *SBP* and *generalized Dirichlet distribution* $\mathcal{GD}$ [4], and conjugacy of $\mathcal{GD}$ with the multinomial distribution. That allows us to integrate out ρ, see appendix for details. We address the second challenge by integrating out the parameters using the Beta-Bernoulli conjugacy and directly sampling the binary random variables. Following counting notations will be used in the inference.

Count notation: (i) "dot" in the suffix represents marginalization at the corresponding index, (ii) $(-x)$ in the superscript means counting without x.

$\bar{m}_{kv}$ denotes number of times word v is assigned to topic k, and $\bar{m}_{k.}$ is number of times topic k has occurred. $\dot{m}_{deak} = \sum_{r=1}^{n_{da}} I[z_{dar} = k, \xi_{dea} = 1]$, i.e. number of times topic k has occurred in the article considering only those sentences selected by ξ_{dea}. $\hat{m}_{djk}$ is number of times topic k has been used from topic vector j of article d. $\check{m}_{daj}$ is number of times topic vector j has appeared in sentence a in document d. $\mathring{m}_{dk} = \sum_{e=1}^{E_d} \sum_{i=1}^{n_{de}} I[y_{dei} = k]$ is the number of times topic k has appeared across all the comments.

Sampling b: We compute the conditional probability $p(b_{dai} = j|b^{-dai}, \tau, \iota)$, for $j < J_d$, as

$$\frac{\tau_j + \check{m}_{daj}^{-dai}}{\tau_j + \iota_j + \sum_{r=j}^{J_d} \check{m}_{dar}^{-dai}} \prod_{l<j} \frac{\iota_l + \sum_{s=l+1}^{J_d} \check{m}_{das}^{-dai}}{\tau_l + \iota_l + \sum_{s=l}^{J_d} \check{m}_{das}^{-dai}} \qquad (3)$$

& $p(b_{dai} = J_d|b^{-dai}, \tau, \iota) = 1 - \sum_{l=1}^{J_d-1} p(b_{dai} = l|\tilde{b}^{-dai}, \tau, \iota)$. For any j the above equation can be expressed as $u_j \prod_{l=1}^{j-1}(1-u_l)$ for suitably defined u. The probability of $b_{dai} = j$ directly depends on probability of $b_{dai} \neq l, l < j$. This property is absent in standard priors e.g. finite dimensional Dirichlet distribution, nonparametric DP. Finally, $p(b_{dai} = j|\mathbf{b}^{-dai}, \mathbf{z})$ is computed as:

$$\begin{aligned} &\propto \quad p(z_{dai}|b_{dai} = j, \tilde{z}^{-dai})p(b_{dai} = j|\tilde{b}^{-dai}) \\ &= \quad \frac{\alpha + \hat{m}_{djz_{dai}}^{-dai}}{K\alpha + \hat{m}_{dj.}^{-dai}} \; p(b_{dai} = j|\tilde{b}^{-dai}) \end{aligned} \qquad (4)$$

Sampling ϕ: Note that, ϕ_{kv} is a binary selector, so if $\bar{m}_{kv} > 0$ then $\phi_{kv} = 1$ a.s., otherwise we compute $p(\phi_{kv} = 1|\tilde{w}, \tilde{z}, \tilde{\phi}^{-kv})$ as below.

$$\begin{aligned} &\propto \quad p(w|\phi_{kv} = 1, \tilde{z}, \tilde{\phi}^{-kv})p(\phi_{kv} = 1|\tilde{\phi}^{-kv}) \\ &= \quad \frac{\Gamma(\sum_{u\neq v} \phi_{ku}\eta + \eta)}{\Gamma(\sum_{u\neq v} \phi_{ku}\eta + \eta + \bar{m}_{k.})} p(\phi_{kv} = 1|\tilde{\phi}^{-kv}) \end{aligned} \qquad (5)$$

where $p(\phi_{kv} = 1|\tilde{\phi}^{-kv})$ can be computed as

$$\int d\kappa_v \; p(\phi_{kv} = 1|\kappa_v)p(\kappa_v|\tilde{\phi}^{-kv}) = \frac{\frac{\nu\lambda}{V} + \sum_{j\neq k} \phi_{jv}}{\frac{\nu\lambda}{V} + \lambda + K} \qquad (6)$$

Sampling ξ: The inference equation is derived from uniform (first part) and Beta Bernoulli conjugacy (second part). We compute $p(\xi_{dea} = 1|\tilde{y}_{de}, \tilde{\xi}^{-dea}, \tilde{z}_d, \varrho)$ as below.

$$\begin{aligned} &\propto \quad \prod_{i=1}^{n_{de}} p(y_{dei}|\tilde{z}_d, \xi_{dea} = 1, \tilde{\xi}^{-dea}) \;\; p(\xi_{dea} = 1|\tilde{\xi}^{-dea}, \varrho) \\ &= \quad \prod_{i=1}^{n_{de}} \frac{\dot{m}_{de.y_{dei}}^{-a} + \dot{m}_{deay_{dei}}}{\dot{m}_{de..}^{-a} + \dot{m}_{dea.}} \;\; \frac{\varrho_1}{\varrho_1 + \varrho_2} \end{aligned} \qquad (7)$$

Sampling z: Sampling topic indices z follows from Dirichlet multinomial conjugacy as follows. $p(z_{dai} = k|\tilde{w}, \tilde{z}^{-dai})$ is computed as below.

$$\begin{aligned} \propto \quad & p(w_{dai}|z_{dai} = k, \tilde{\phi}, \eta)p(z_{dai} = k|\tilde{z}^{-dai}, \tilde{b}, \alpha) \\ & \times \; p(\tilde{y}_d|\tilde{z}^{-dai}, z_{dai} = k) \\ = \quad & \frac{\phi_{kw_{dai}}\eta + \bar{m}_{kw_{dai}}^{-dai}}{\sum_v \phi_{kv}\eta + \bar{m}_{k.}^{-dai}} \frac{\alpha + \hat{m}_{db_{dai}k}^{-dai}}{\alpha K + \hat{m}_{db_{dai}.}^{-dai}} \prod_{l\neq k} \dot{m}_{de.l}^{\mathring{m}_{dl}} (\dot{m}_{de.k} + 1)^{\mathring{m}_{dk}} \end{aligned} \qquad (8)$$

Sampling y: Comment indices depend on topic-word Dirichlet-multinomial conjugacy (first part) and uniform distribution (second part) for the comments-article correspondence. We compute $p(y_{dei} = k|\tilde{c}, \tilde{y}^{-dei}, \tilde{z}_d, \tilde{\xi}_{de})$ as below.

$$\begin{aligned} &\propto \quad p(c_{dei}|y_{dei} = k, \tilde{y}^{-dei}, \tilde{\phi}, \eta) \;\; p(y_{dei} = k|\tilde{y}^{-dei}, \tilde{z}_d, \tilde{\xi}_{de}) \\ &= \quad \frac{\phi_{kc_{dei}}\eta + \bar{m}_{kc_{dei}}^{-dei}}{\sum_v \phi_{kv}\eta + \bar{m}_{k.}^{-dei}} \;\; \frac{\dot{m}_{de.k}}{\dot{m}_{de..}} \end{aligned} \qquad (9)$$

Inference algorithm: Equations (4), (5), (7), (8) and (9) together form the inference algorithm. Inference of all variables depend on others, so we need to solve iteratively. The procedure starts by initializing the variables randomly.

Relationship with other algorithms: [7] and [5] have given two different algorithms for solving SBP. The algorithm in [7] samples v explicitly, while the algorithm in [5] is similar to that used here. On the other hand, [13] uses a different mechanism for inducing sparsity over words, they do

Algorithm 1 Classification & alignment

```
Input: $\mathbf{y}$, $\mathbf{z}$; thresholds: $t_{cos}, t'_{cos}, t_\xi$
Compute $\tilde{\mathbf{y}}$, $\tilde{\mathbf{z}}$ from $\mathbf{y}$, $\mathbf{z}$.
for $d \leftarrow 1, \dots, D$ do
    Initialize $\Phi_d \leftarrow \{\}$, $\Psi_d \leftarrow \{\}$, $\Gamma_d \leftarrow \{\}$    ▷ empty
    for $e \leftarrow 1, \dots, E_d$ do
        $R_{de} \leftarrow \{\}$, $S_{de} \leftarrow \{\}$    ▷ empty
        for $a \leftarrow 1, \dots, S_d$ do
            $\Delta(a, e) \leftarrow \frac{\tilde{y}_{de}^T \tilde{z}_{da}}{\|\tilde{y}_{de}\| \|\tilde{z}_{da}\|}$,    ▷ cosine distance
            if $\Delta(a, e) \geq t_{cos}$ then
                $R_{de} \leftarrow R_{de} \cup \{a\}$    ▷ insert $a$
            end if
            if $\Delta(a, e) \geq t'_{cos}$ & $p(\xi_{dea} | y_{de}, z_d) \geq t_\xi$ then
                $S_{de} \leftarrow S_{de} \cup \{a\}$    ▷ insert $a$
            end if
        end for
        if $|R_{de}| == 0$ then    ▷ cardinality is 0
            $\Psi_d \leftarrow \Psi_d \cup \{e\}$    ▷ Irrelevant
        else if $|R_{de}| \leq N_d$ then
            $\Phi_d \leftarrow \Phi_d \cup \{e\}$    ▷ Specific
        else $\Gamma_d \leftarrow \Gamma_d \cup \{e\}$    ▷ General
        end if
    end for
end for
Output: $\Phi_d, \Psi_d, \Gamma_d, R_{de}, S_{de}$
```

not sample the binary random variables ϕ explicitly. The algorithm given here is novel and much simpler to implement, yet efficient. Use of ξ has not been studied before.

4. ALGORITHM FOR SCL

In this section we propose an algorithm based on SCTM for SCL. The algorithm considers the inferred latent variables z, y and ξ obtained from SCTM inference procedure and uses them for labeling each comment as general, specific or irrelevant and recover the set R_{de} for specific comments.

$\mathbf{z}$ and $\mathbf{y}$ are topic indices for the article and comment respectively [2] $\tilde{\mathbf{z}}$, $\tilde{\mathbf{y}}$ are topic frequency vectors[3] for article and comment respectively. Following the definition in section 2, we need to compute R_{de} for each comment e corresponding to article d. Computation of R_{de} follows from cosine similarity, between the topic assignment counts of the comment $\tilde{\mathbf{y}}$ and the sentence $\tilde{\mathbf{z}}$, above a fixed threshold. The decision will depend on threshold N_d, i.e. specific if $|R_{de}| \leq N_d$. The details are given in Algorithm 1. For SCTM we can also use the posterior probability of ξ, i.e. $p(\xi_{dea} = 1 | y_{de}, z_d)$, to improve the alignment to specific sentences. Thus, for SCTM, we have a separate set S_{de} in the algorithm which is obtained by using this information .

Selecting thresholds: We set the threshold as $N_d = min(N_g, \lfloor 0.6 * S_d \rfloor)$. N_g limits the threshold for very large documents. $N_g \in [6, 10]$ is found to work well.

5. EMPIRICAL EVALUATION

In this section we will evaluate the proposed model SCTM empirically on various aspects using real life datasets[4].

Table 1: Properties of the datasets.

	AT-Science	AT-Gadgets	Yahoo! News
#(Articles)	1,369	2,186	730
#(Comments)	90,654	160,761	138,538

5.1 Datasets

We show some basic properties of the datasets in Table 1 and provide a description below. Note that the data consists mostly of articles with a large number of comments which introduces further challenges in discovering specific comments as their number will be expected to be very small compared to the total comments.

ArsTechnica Science (AT-Science): The dataset consists of articles and comments crawled from Science section of the site ArsTechnica[5]. ArsTechnica is a science and technology blog whose writers consist mostly of academicians. Its articles and readership leads to opinionated discussions in the comments which makes it a perfect testbed for our problem. We crawled 1500 articles and their comments over approximately a two year timeline (June 2011 to March 2013) and removed articles with less than 5 comments.

Gold-standard: In order to quantitatively evaluate our model, we developed a gold standard by manually annotating articles over a one year timeline (March 2012 to March 2013) after filtering out articles which either had very few comments or had only general and irrelevant comments. For each of these article we manually labeled specific and non-specific comments and also created the alignment of specific comments to the relevant sentences. The gold standard consists of 501 articles with a total of 3176 comments, in which there are 1443 specific comments with an average of 2.9 specific comments per article.

ArsTechnica Gadgets (AT-Gadgets): The dataset consists of articles and comments crawled from Gadgets section of the site ArsTechnica[6]. This dataset consists mostly of product reviews on latest gadgets. We crawled about 2200 articles and their comments over a two year timeline (August 2011 to August 2013) and removed articles with less than 5 comments.

Yahoo! News: The dataset consists of articles and comments crawled from the most-commented[7] and archive section[8] of the site Yahoo! News, one of the most popular news site. We crawled about 1150 articles along with their comments, going chronologically backwards from 31 March 2013. We then removed all those articles in the dataset which had fewer than 5 lines or had fewer than 5 comments. We were left with 730 articles with more than 100,000 comments.

5.2 Experimental setup

We have used the same pre-processing of the dataset and exactly similar parameter settings for Corr-LDA as well as SCTM. We have removed stop words and transformed all

[2] $\mathbf{z} = \{\{z_{dai} | i \in [n_{da}], a \in [S_d]\}_{d=1}^D\}$

[3] $\tilde{\mathbf{z}} = \{\{\tilde{z}_{da} | a \in [n_{da}]\}_{d=1}^D\}$, $\tilde{z}_{da} = (\tilde{z}_{dak})_{k=1}^K \in \mathbb{R}^K, \tilde{z}_{dak} = \frac{1}{n_{da}} \sum_{i=1}^{n_{da}} I[z_{dai} = k]$, $\tilde{\mathbf{y}}$ is defined similarly

[4] Relevant resources at: mllab.csa.iisc.ernet.in/sctm

[5] arstechnica.com/science

[6] arstechnica.com/gadgets

[7] news.yahoo.com/all-sections-most-commented/most-popular

[8] news.yahoo.com/archive

Table 2: Perplexity on test data (lower is better).

Model	AT-Science	AT-Gadgets	Yahoo! News
SCTM	**12,327**	6,445	**6,880**
Corr-LDA	14,029	6,443	9,673

Table 3: Topic coherence for top 50 topics (greater is better).

Model	AT-Science	AT-Gadgets	Yahoo! News
SCTM	**-52.08**	**-50.47**	**-65.83**
Corr-LDA	-88.22	-87.27	-93.68

Table 4: Topic diversity (greater is better).

Model	AT-Science	AT-Gadgets	Yahoo! News
SCTM	**96.34**	**96.78**	**99.45**
Corr-LDA	9.0	23.34	21.5

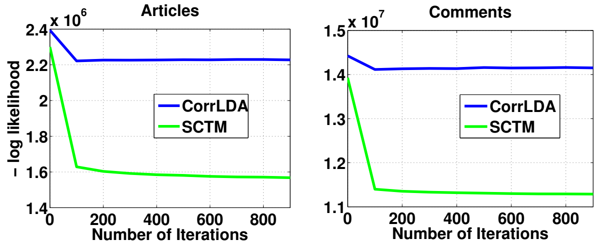


Figure 3: Convergence of inference: change in negative log-likelihood with number of iterations of the sampling algorithm on Yahoo! News (lower is better). SCTM converges faster and better.

characters into small case. Blank spaces are treated as the delimiter between words. However, we have not used stemmer or POS-taggers.

We used the following parameter setting: $\alpha = 1$, $\eta = 1$, $\tau_j = 0.01$, $\iota_j = 0.1$ $(\forall j)$, $\nu = 1$, $\lambda = 10$, $\varrho_1 = 1$, $\varrho_2 = 4$, and $J_d = S_d$. For α, η we use uninformative value. For other parameters specific to SCTM, low held-out data perplexity (Table 2) and high training likelihood (Figure 3) indicate that model is less sensitive to these hyperparameters. We used 300 topics and 20,000 vocabulary in our experiments.

Evaluation criteria: First we evaluate SCTM against Corr-LDA in terms of fitness to the dataset. We use perplexity, topic coherence, topic diversity on all three datasets and convergence of the algorithm on Yahoo! News. Then we focus on the main question of this paper, by evaluating the performance for SCL on AT-Science (subset with gold-standard) and a limited evaluation on Yahoo! News.

5.3 Comparison of SCTM and Corr-LDA

We quantitatively evaluate SCTM in the task of modeling article-comments dataset, comparing with the baseline Corr-LDA. For this task, we use perplexity, topic-coherence, diversity among discovered topics and convergence of the inference algorithm as evaluation metrics. The results are presented in Table 2, 3, 4 and Figure 3.

Table 5: Precision, recall and F1 score for discovering specific comments.

Model	Precision	Recall	F1
SCTM	**0.60**	**0.61**	**0.60**
Corr-LDA	0.23	0.21	0.22

Perplexity: Perplexity is a standard quantitative measure in topic modeling literature to compare the performance of various topic models [2]. A lower value of perplexity indicates better generalizability of the topic model. As shown in Table 2, SCTM performs far superior than Corr-LDA.

Topic coherence: By approximating the user experience of *topic quality* on top σ words of a topic, [10] proposed that topic coherence (TC) can be measured as: $TC(\sigma) = \sum_{i \leq \sigma} \sum_{j<i} \log \frac{D(w_i, w_j)+\epsilon}{D(w_j)}$. $D(w)$ is the document frequency of any word w, and $D(w_i, w_j)$ is the document frequency of w_i and w_j together. ϵ is a small constant to avoid log zero. Values closer to zero indicate better coherence. We have used $\sigma = 5$ to compute coherence of a topic. Table 3 contains the comparison between SCTM and Corr-LDA which shows supremacy of the proposed approach.

Topic diversity: We want to evaluate the ability of SCTM in discovering diverse set of topics. We consider top 5 words for topics which have maximum probability for a word greater than 0.01. Then compute textual difference between topic i and topic j as $Tdiv_{ij} = 1 - |tw_i \cap tw_j|/5$, where tw_i contains top 5 words for topic i. Averaging over all pairs we report in Table 4. Note that higher value of $Tdiv$ is better and empirical results show that SCTM is superior.

Convergence of inference: We plot negative log likelihood of SCTM and Corr-LDA against iterations in Figure 3. Note that SCTM mixes much better than Corr-LDA. Both the models, SCTM as well as Corr-LDA, converge almost in the same time. However, as fitness of Corr-LDA is worse, the likelihood of the dataset remains low throughout.

Discussion: Interestingly, SCTM outperforms Corr-LDA in modeling aspects commonly used in topic modeling literature. The main reason behind this is that SCTM finds more unique topics, whereas Corr-LDA finds relatively mixed topics by mixing up many unique topics. On the other hand, supremacy of SCTM over Corr-LDA in topic diversity is expected as that is explicitly ensured in the modeling. Low perplexity on held-out dataset and high likelihood on training data affirms that although SCTM is a far more complex model than Corr-LDA, it is able to learn from the dataset appropriately.

5.4 Accuracy of SCTM over Corr-LDA on SCL

We evaluate the main task of this paper in this section, i.e. the task of discovering specific comments. We use the manually annotated gold-standard from AT-Science dataset for evaluation. Using Algorithm 1 we classify comments into specific and non-specific (general and irrelevant) for both Corr-LDA and SCTM. Using the gold-standard we compute precision-recall and report them in Table 5. SCTM outperforms Corr-LDA significantly.

Table 6: Precision, recall and F1 score for aligning specific comments to the sentences.

Model	Precision	Recall	F1
SCTM	**0.365**	**0.442**	**0.400**
Corr-LDA	0.019	0.038	0.025

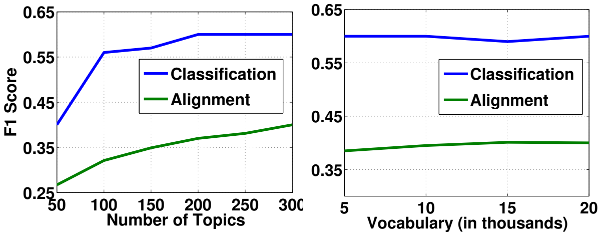


Figure 4: Sensitivity analysis: F1 score variation with respect to number of topics and vocabulary size for SCTM.

Similarly, from Algorithm 1 we get the article sentences corresponding to each specific comment detected by the models, for both Corr-LDA and SCTM. Then using gold-standard we compute precision, recall and F1. The results are reported in Table 6. SCTM is far superior than Corr-LDA.

We can see that Corr-LDA is unable to give any satisfactory result in discovering specific comments. This is due to the fact that there is one topic vector per article-comments pair which fails to model specific comments properly.

Evaluation on Yahoo! News: Due to the unavailability of ground-truths and the large number of comments in the dataset, we have done a limited evaluation on all articles which had upto 50 predicted specific comments (a total of 604 articles out of 730). Out of a total of 138,538 comments in the dataset, SCTM predicted 23,624 comments to be specific comments and the per document average accuracy is **63%**. In comparison, Corr-LDA discovers only a total of 4,704 comments in the dataset which is even less than 50% of the *correctly* discovered specific comments by SCTM.

Sensitivity analysis of SCTM: We analyse the sensitivity of SCTM towards number of topics and vocabulary size, two important hyperparameters, in Figure 4. F1 score is observed to be not sensitive to the vocabulary size, however it gets better with the number of topics and stabilizes gradually. This is justified as we expect to get large number of topics due to diversity across the articles and comments.

6. USE CASES OF SCL

Comments are a great source of public response. Whereas specific comments are more useful to mine information from comments. To demonstrate that SCTM is useful in a wide variety of applications, we do a detailed analysis of two specific articles. (i) Regarding *President Obama*'s visit to Middle-Eastern countries and (ii) a product review on *Google Nexus*. The first article was the *most commented* article in our Yahoo! News dataset and the second article was the most recent well commented article in ArsTechnica's Gadget section. The two articles demonstrate the efficacy of SCTM in the diverse scenarios of large and small number of comments, different categories of politics and product reviews and different hosting platforms of news and blogs. Note that, such analysis is subject to efficacy in SCL and hence is beyond the scope of the existing methods.

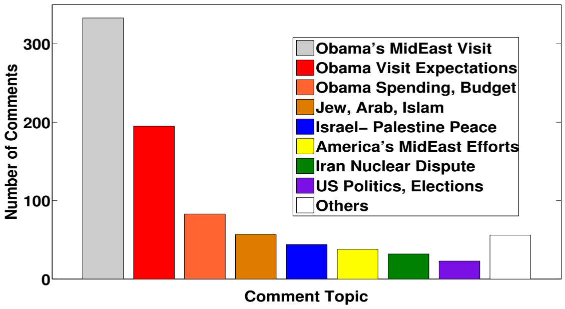


Figure 6: Popularity of topics based on number of comments. The first topic (gray-colored) corresponds to general comments. Except that, all other topics correspond to specific comments and are missed by the state of the art methods.

6.1 Analysing comments using topics discovered by SCTM

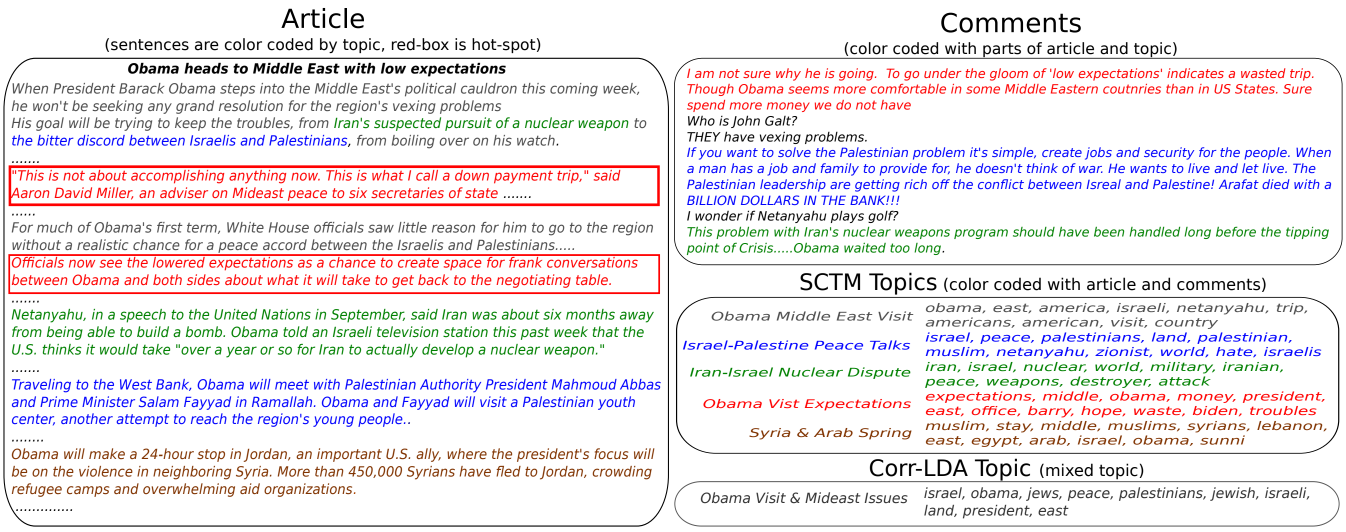
We analyse the comments on the news reporting *President Obama*'s visit to the Middle East during March 19-23, 2013[9]. The event was widely covered in the media and was highly commented upon. Figure 5 shows excerpts from one such article at Yahoo! News. At the time of crawling, the article had 1919 comments. Analysing the comments one can accumulate a survey of opinion.

Note that the article in example touches upon various issues like Iranian nuclear program, Syria etc in different segments of the article. The number of such issues covered in the article is large but the amount of text contributing to each of them is small. Interestingly, many comments focus on such issues minorly described in the article.

Figure 5 shows some of the topics discovered by SCTM on this event and the comments corresponding to those topics. Sentences, topics and the comments are color coded such that they signify a link among them. Note that, SCTM is able to discover fine topics like "*Obama's MidEast Visit*", "*Israel-Palestine Peace*" and "*Iran Nuclear Dispute*" where Corr-LDA mixes them together. SCTM is able to retrieve comments which are related to such topics described in specific parts of the article.

Using the ability to detect precise topics SCTM can analyze the public response quantitatively, on the basis of the topic being discussed in the comment. We categorize comments based on their major topic. Figure 6 shows the number of comments classified in this way into the different topics. As expected, we find that most of the comments are general and about the topic "*Obama's MidEast Visit*" (gray colored). The remaining comments are specific and related to other topics. Among the specific comments, the most discussed topic is "*Obama Visit Expectations*" (red colored) with 195 comments marking the "*hot-spot*" in the article (box in Figure 5).

[9] news.yahoo.com/obama-heads-middle-east-low-expectations.html


Article

(sentences are color coded by topic, red-box is hot-spot)

Obama heads to Middle East with low expectations

When President Barack Obama steps into the Middle East's political cauldron this coming week, he won't be seeking any grand resolution for the region's vexing problems

His goal will be trying to keep the troubles, from Iran's suspected pursuit of a nuclear weapon to the bitter discord between Israelis and Palestinians, from boiling over on his watch.

....

"This is not about accomplishing anything now. This is what I call a down payment trip," said Aaron David Miller, an adviser on Mideast peace to six secretaries of state ....

......

For much of Obama's first term, White House officials saw little reason for him to go to the region without a realistic chance for a peace accord between the Israelis and Palestinians.....

Officials now see the lowered expectations as a chance to create space for frank conversations between Obama and both sides about what it will take to get back to the negotiating table.

....

Netanyahu, in a speech to the United Nations in September, said Iran was about six months away from being able to build a bomb. Obama told an Israeli television station this past week that the U.S. thinks it would take "over a year or so for Iran to actually develop a nuclear weapon."

....

Traveling to the West Bank, Obama will meet with Palestinian Authority President Mahmoud Abbas and Prime Minister Salam Fayyad in Ramallah. Obama and Fayyad will visit a Palestinian youth center, another attempt to reach the region's young people..

....

Obama will make a 24-hour stop in Jordan, an important U.S. ally, where the president's focus will be on the violence in neighboring Syria. More than 450,000 Syrians have fled to Jordan, crowding refugee camps and overwhelming aid organizations.

....

Comments

(color coded with parts of article and topic)

I am not sure why he is going. To go under the gloom of 'low expectations' indicates a wasted trip. Though Obama seems more comfortable in some Middle Eastern coutnries than in US States. Sure spend more money we do not have

Who is John Galt?

THEY have vexing problems.

If you want to solve the Palestinian problem it's simple, create jobs and security for the people. When a man has a job and family to provide for, he doesn't think of war. He wants to live and let live. The Palestinian leadership are getting rich off the conflict between Isreal and Palestine! Arafat died with a BILLION DOLLARS IN THE BANK!!!

I wonder if Netanyahu plays golf?

This problem with Iran's nuclear weapons program should have been handled long before the tipping point of Crisis.....Obama waited too long.

SCTM Topics (color coded with article and comments)

Obama Middle East Visit	obama, east, america, israeli, netanyahu, trip, americans, american, visit, country
Israel-Palestine Peace Talks	israel, peace, palestinians, land, palestinian, muslim, netanyahu, zionist, world, hate, israelis
Iran-Israel Nuclear Dispute	iran, israel, nuclear, world, military, iranian, peace, weapons, destroyer, attack
Obama Vist Expectations	expectations, middle, obama, money, president, east, office, barry, hope, waste, biden, troubles
Syria & Arab Spring	muslim, stay, middle, muslims, syrians, lebanon, east, egypt, arab, israel, obama, sunni

Corr-LDA Topic (mixed topic)

Obama Visit & Mideast Issues	israel, obama, jews, peace, palestinians, jewish, israeli, land, president, east

Figure 5: An example from Yahoo! News. SCTM is able to find issues under-reported in the article and specific comments on them. SCTM also detects *hot-spot* (red box), although "*Obama-Visit-Expectations*" topic has appeared only in those two sentences, it has received the most number of comments leaving general ones. Corr-LDA finds only one mixed topic failing to analyse user response effectively.

It is important to note that sentences in the main article related to a specific comment can be very few and may not be contiguous (which can be seen in Figure 5 and Figure 1). This demonstrates the elegance of SCTM in performing such a hard task of SCL.

6.2 Capturing surprising market trends

We analyse the comments on the review of a recently launched tablet, *Google Nexus 7*, published in AT-Gadget. An excerpt is given in Figure 1. This article is one of the well commented reviews in the recent times with 180 comments at the time of crawling. In review articles, the story generally covers a wide range of aspects and many of them will be less described. However, people might be sensitive to a certain aspect which is difficult to guess a priori.

The author discussed only in approximately 8% of the sentences the "screen display quality", or more precisely, the comparison with "*iPad* Retina display" and the future prospect of "*iPad mini*" equipped with such retina display. SCTM correctly found that this point received a great deal of attention from users with 22% of comments leaving the irrelevant ones. It is surprising because an article about a new *Google* tablet has attracted a large number of comments related to another tablet, "*iPad mini*", which has not even been released. Examples of such specific comments can be seen in Figure 1, all of these were correctly discovered by SCTM. We hope, this provides an important feedback to both *Google* as well as *Apple*. Note that none of the existing methods are capable to capture such trends.

6.3 Enhancing user satisfaction

We demonstrate two applications of SCL which can improve user experience. In these cases accuracy on SCL is critical and hence beyond the scope of the existing methods.

Hot-spot detection: A *hot-spot* is the set of sentences in the article which have received most attention in the comments. For example, in Figure 5 the portions in *red box* are the hot-spots discovered by SCTM. Similarly, in Figure 1 the portions in *green box* are the hot-spots in that article. SCL enables one to find such hot-spots in articles automatically. This can help in summarization, advertisements, better user engagement etc.

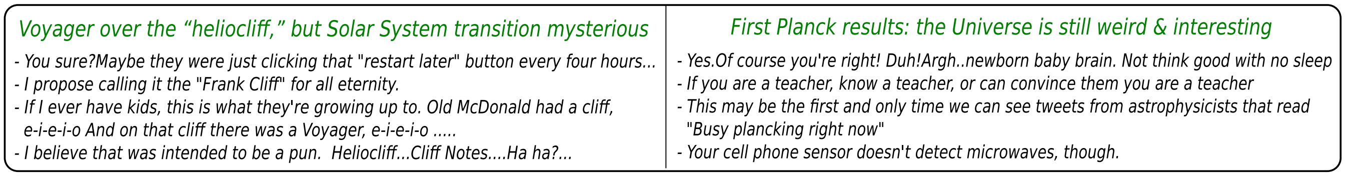
The ability of SCTM in categorizing comments into various issues covered in the article and connecting with the sentences in the article is key to discover hot-spots, where existing methods fail.

Comment cleaning: Comment cleaning is one of the important tasks today for online media sites. Being equipped with SCL, we propose a novel method of cleaning up irrelevant comments without using any external resource or supervision. Existing methods of comment cleaning focus on special features like presence of URL or certain keywords. However, even after filtering out such comments, there will be some comments which are absolutely irrelevant and in many cases indecent (see Figure 7). But it is difficult for existing methods to clean up such irrelevant comments as they look very much similar to the normal comments, many of them may contain words present in the article. Moreover, most of the existing methods are either supervised or semi supervised (refer to [8] for a complete review) or use text enrichment [14].

The algorithm for detecting irrelevant comments is in Algorithm 1. The comments shown in Figure 7 on two articles of AT-Science[10] are all examples of *irrelevant comments* correctly detected by SCTM, where existing methods failed.

We evaluated the performance of SCTM in finding irrelevant comments on 100 randomly selected articles from the Yahoo! News dataset. For each of these articles, the algorithm was applied to predict top 20 irrelevant comments

[10] arstechnica.com/science/2013/03/first-planck-results-the-universe-is-still-weird-and-interesting/, arstechnica.com/science/2013/03/voyager-probes-key-transition-remains-mysterious/



Voyager over the “heliocliff,” but Solar System transition mysterious

- *You sure?Maybe they were just clicking that "restart later" button every four hours...*
- *I propose calling it the "Frank Cliff" for all eternity.*
- *If I ever have kids, this is what they're growing up to. Old McDonald had a cliff, e-i-e-i-o And on that cliff there was a Voyager, e-i-e-i-o .....*
- *I believe that was intended to be a pun. Heliocliff...Cliff Notes....Ha ha?...*

First Planck results: the Universe is still weird & interesting

- *Yes.Of course you're right! Duh!Argh..newborn baby brain. Not think good with no sleep*
- *If you are a teacher, know a teacher, or can convince them you are a teacher*
- *This may be the first and only time we can see tweets from astrophysicists that read "Busy plancking right now"*
- *Your cell phone sensor doesn't detect microwaves, though.*

Figure 7: Examples of irrelevant comments on two articles from AT-Science (showing only headline and irrelevant comments). Notice that these comments look normal and do not have any specific feature except lack of correspondence to the main article making it challenging for existing methods. These comments are correctly marked by SCTM to be irrelavnt, whereas the existing methods failed.

which were manually annotated as correct or incorrect. The mean average precision at 20 (**MAP@20**) was found to be **0.87** which is quite high. Moreover, most of the mistakes that the model makes fall in a grey area, where humans will differ in their opinion of the comment being irrelevant.

7. CONCLUSION

In this paper we explored an interesting problem of specific comment location which is beyond the scope of the state of the art. A novel correspondence topic model, namely SCTM, has been proposed which admits an efficient collapsed Gibbs sampling algorithm. Following [3] the proposed inference can be easily modified to be scalable. On three real life datasets we evaluated SCTM against Corr-LDA to demonstrate efficacy of the proposed approach. Finally, we demonstrated four different use cases with practical importance. We believe similar study can be done using SCTM on other datasets like image-tag, paper-bibliography etc.

8. ACKNOWLEDGMENTS

We are thankful to all the reviewers for their valuable comments. The authors were partially supported by DST grant (DST/ECA/CB/1101).

9. REFERENCES

[1] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR*, pages 127–134. ACM, 2003.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Jan. 2003.

[3] K. R. Canini, L. Shi, and T. L. Griffiths. Online inference of topics with latent Dirichlet allocation. In *AISTATS*, volume 5, 2009.

[4] R. J. Connor and J. E. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[5] M. Das, S. Bhattacharya, C. Bhattacharyya, and G. Kanchi. Subtle topic models and discovering subtly manifested software concerns automatically. In *ICML*, pages 253–261, 2013.

[6] Y. Hu, A. John, D. D. Seligmann, and F. Wang. What were the tweets about? Topical associations between public events and twitter feeds. In *ICWSM*, 2012.

[7] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[8] R. Kant, S. H. Sengamedu, and K. S. Kumar. Comment spam detection by sequence mining. In *WSDM*, pages 183–192. ACM, 2012.

[9] Z. Ma, A. Sun, Q. Yuan, and G. Cong. Topic-driven reader comments summarization. In *CIKM '12*, pages 265–274. ACM, 2012.

[10] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *EMNLP*, pages 262–272. Association for Computational Linguistics, 2011.

[11] D. K. Sil, S. H. Sengamedu, and C. Bhattacharyya. Supervised matching of comments with news article segments. In *CIKM*, pages 2125–2128. ACM, 2011.

[12] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, pages 111–120. ACM, 2008.

[13] C. Wang and D. Blei. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *NIPS*, pages 1982–1989. 2009.

[14] J. Wang, C. T. Yu, P. S. Yu, B. Liu, and W. Meng. Diversionary comments under political blog posts. In *CIKM*, pages 1789–1793. ACM, 2012.

[15] T. Yano, W. W. Cohen, and N. A. Smith. Predicting response to political blog posts with topic models. In *NAACL*, pages 477–485. ACL, 2009.

APPENDIX

Collapsing SBP random vector ρ.

From the relation between SBP and $\mathcal{GD}$ [4] we get that, if ρ_{da}s are constructed as equation (1), then they are equivalently distributed as $\mathcal{GD}$ [7]. The density of ρ_{da} is:

$$f_{\rho_{da}} = \prod_{j=1}^{J_d-1} \frac{\rho_{daj}^{\tau_j-1}(1-\sum_{l=1}^{j}\rho_{dal})^{\kappa_j}}{B(\tau_j,\iota_j)} \tag{10}$$

where $B(\tau_j,\iota_j) = \frac{\Gamma(\tau_j)\Gamma(\iota_j)}{\Gamma(\tau_j+\iota_j)}$. $\kappa_j = \iota_j - \iota_{j+1} - \tau_{j+1}$ for $j = 1, 2, \ldots, J_d - 2$ and $\kappa_{J_d-1} = \iota_{J_d-1} - 1$. Note that, $\rho_{daJ_d} = 1 - \sum_{l=1}^{J_d-1}\rho_{dal}$. Note that, by setting $\iota_{j-1} = \tau_j + \iota_j$, $2 \leq j < J_d$, $\mathcal{GD}$ reduces to standard Dirichlet distribution.

Like Dirichlet distribution, $\mathcal{GD}$ is also conjugate to the multinomial distribution, and hence we can integrate out ρ's and υ's. If $\rho_{da} \sim \mathcal{GD}_{J_d-1}(\tau_1,\ldots,\tau_{J_d-1},\iota_1,\ldots,\iota_{J_d-1})$, and b_{daj}s are sampled from $mult(\rho_{da})$, then the posterior distribution of ρ_{da} given (b_{dal})s $(l \neq i)$ is again a $\mathcal{GD}$ with density $\mathcal{GD}_{J_d-1}(\bar{\tau}_1,\ldots,\bar{\tau}_{J_d-1},\bar{\iota}_1,\ldots,\bar{\iota}_{J_d-1})$, where $\bar{\tau}_j = \tau_j + \check{m}_{daj}^{-dai}, \bar{\iota}_j = \iota_j + \sum_{l=j+1}^{J_d}\check{m}_{dal}^{-dai}$.