Tone Analyzer for Online Customer Service: An Unsupervised Model with Interfered Training

Peifeng Yin
IBM Almaden Research
650 Harry Road
San Jose, CA 95120, US
peifengy@us.ibm.com

Zhe Liu
IBM Almaden Research
650 Harry Road
San Jose, CA 95120, US
liuzh@us.ibm.com

Anbang Xu
IBM Almaden Research
650 Harry Road
San Jose, CA 95120, US
anbangxu@us.ibm.com

Taiga Nakamura
IBM Almaden Research
650 Harry Road
San Jose, CA 95120, US
taiga@us.ibm.com

ABSTRACT

Emotion analysis of online customer service conservation is important for good user experience and customer satisfaction. However, conventional metrics do not fit this application scenario. In this work, by collecting and labeling online conversations of customer service on Twitter, we identify 8 new metrics, named as *tones*, to describe emotional information. To better interpret each tone, we extend the *Latent Dirichlet Allocation (LDA)* model to *Tone LDA (T-LDA)*. In T-LDA, each latent topic is explicitly associated with one of three semantic categories, i.e., tone-related, domain-specific and auxiliary. By integrating tone label into learning, T-LDA can interfere the original unsupervised training process and thus is able to identify representative tone-related words. In evaluation, T-LDA shows better performance than baselines in predicting tone intensity. Also, a case study is conducted to analyze each tone via T-LDA output.

1 INTRODUCTION

Due to the popularity of social web, e.g., Twitter, many companies set-up online agents to provide customer service. Unlike conventional telephone call, customers and agents interact asynchronously via the web, e.g., publish tweets and '@' the particular account.

While there is more convenience, customers tend to be quite emotional via online help desk [27]. Besides providing correct solution for the problem, it is equally important for online customer agents to properly pacify such emotion. In this scenario, an analysis would benefit in three folds: i) understand the user behavior in context of online customer service, ii) provide training materials for online customer service agents, and iii) shed insight in the development of human-like auto-response bot.

CIKM'17 , November 6–10, 2017, Singapore.

DOI: 10.1145/3132847.3132864

There are two major emotion metrics adopted by existing works. In [6, 9], five major categories, i.e., anger, sadness, joy, disgust, and fear, are used. On the other hand, a recent work [5] adopts a six-dimensional metric, including calm, alert, sure, vital, kind, and happy. However, in context of customer service, neither one fits. For example, showing empathy is not captured by either of them, but it is a common strategy to comfort the complaining customer.

In this work, we identify 8 new metrics for the scenario of online customer service conversation. To differentiate from existing ones, we name them as *tone*. Particularly, they are **anxious**, **excited**, **frustrated**, **impolite**, **polite**, **sad**, **satisfied**, and **sympathetic**. Then we crawl customer service conversation data from Twitter and ask people to label the intensity of each tone (at scale of 0 to 3) for 71,170 tweets. Details of tone definition and data labeling are shown in Section 3.1.

To understand each tone with labeled data, there are two main issues. The first one is how to handle disagreement labels. In the label process, to mitigate bias, each tweet is labeled by multiple people and the average is used as the true label. However, we also need to consider the variance. For instance, two tweets with the same average intensity label should be treated different if they have different variance. In Section 3.3, we design an adjustment strategy for the label considering factors such as average, variance and number of people who label it.

The second issue is how to model the label data to help understand the tone. Particularly, given one tone, we need to know what are its representative words. One natural way is to model it as a regression task with the labeled tone intensity. However, training on the raw bag-of-word feature may be quite noisy since not all words are tone-related. Another method is to use latent topic modeling, e.g., Latent Dirichlet Allocation (LDA) [4]. But it is an unsupervised method and the learned latent topic may not be associated with tone. There are a few existing works [18–20] integrating classification labels into LDA. However, they do not fit our scenario where the label is not binary but continuous values.

To address this issue, we develop an *Tone LDA (T-LDA)* model. It is essentially a generative model and follows a similar generative process with LDA. However, each latent topic has an explicit semantic meaning, falling into one of three categories: tone-related,

domain-specific and auxiliary. The first category aims to capture words that are highly associated with tone. The second one is to find words that are exclusive to a specific domain. Particularly in this work, each domain is corresponding to a company. Finally, the auxiliary topic contains meaningless words. Note that it is not equivalent to conventional "stop words". It also includes those that have a high frequency of appearance but is not related to either tone or domain. For example, the word "account" may appear in many areas such as bank, mobile carrier, online shopping, etc., and thus it becomes a common "meaningless" words that are representative of none. In Section 5.3, this design is demonstrated to be especially helpful in filtering out tone-irrelevant words.

In LDA, a prior Dirichlet distribution determines the latent topic vector. In T-LDA, however, this Dirichlet itself are model parameters that need to be learned. Specifically, the tone-related ones are fully known after human labeling the tone intensity. For domain-specific ones, we are only aware of which domain (company) the document (tweet) belongs to but not the intensity, thus they are half-known. Finally, the auxiliary is fully unknown. By fixing the tone-related Dirichlet parameters to be the known tone labels, we interfere the original unsupervised learning and make the model converge to our desired direction. Details of T-LDA and its training are introduced in Section 4.

In summary, our contribution contains three folds as below:

- We define new 8-dimension metric *tone* to describe emotional factors in scenario of online customer service.
- We develop new topic model Tone LDA (T-LDA) to learn the representative words for each tone.
- We conduct a case-study on the tone analysis over multiple companies' online customer service (via Twitter).

The rest of the paper is organized as follows. Section 2 gives a literature review of related fields. Section 3 describes preliminaries such as tone definition, data collection & preprocessing and label adjustment. Section 4 provides details of our proposed model and Section 5 shows evaluation as well as case study. Finally Section 6 concludes the whole work.

2 RELATED WORK

Traditional customer service often emphasizes the service quality [10], or exactly, whether the customers' informational needs are satisfied [14]. A recent study about online customer service has revealed that a significant amount of user requests on Twitter can be emotional and they are not necessarily intended to gain information [27]. For example, some tweets from our dataset are: "Comcast is the worst. Period", "BBQ afternoon ☼☼☼☼☼." and "Tesco they look amazing :) I ♡ the banoffee one!" Twitter gives users an opportunity to interact with a company through customer service agents, and these interactions extend beyond those expected from phone or email conversations. Previous works [15, 22] indicate that such emotional factors may directly affect their consumption satisfaction or experience, which motivates our work of tone modeling.

Different works have different measurements of emotion (also known as sentiment or mood). In [23, 24, 28], emotion is measured as a numeric value, where the sign ("+/-") indicates positive/negative and the absolute value suggests the intensity. In [2, 7], emotion is categorized into positive, neutral and negative. These measurements are easy to handle in computation but lacks the descriptive capability. In [5], 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy) of mood measurements are used. Furthermore, prior affective computing research in conversation analysis mainly focuses on five primary emotional categories [6, 9]: anger, sadness, joy, disgust, and fear. However, existing categories may not capture specific tones in customer care conversations. For example, empathy is an essential component of customer experience; yet, it is not considered in general emotional modeling research. We are the first to conduct a bottom-up analysis to identify tone dimensions that are perceived as important in the customer care domain.

If a tone is treated as a topic, it is natural to use latent topic model, e.g., LDA [4], to learn its semantic in terms of word. However, LDA is an unsupervised learning and the trained topic-word distribution may not match the tone. There have been a few extensive works on LDA variants to guide the model convergence and in general they fall into three categories: seed-word [3, 11], single-label [12, 13, 25, 29] and multi-label [18–20].

In [3, 11], seed words are provided for desired topics. During training, these words are restricted to be generated by particular topics. Therefore in the trained model, restricted latent topics is shaped to match the word distribution of desired ones. This variant works for the scenario where user has a start-up word set of desired topic and wants to explore other unknown words. It thus does not fit our problem.

For the second category, the Semi-LDA [25], sLDA [13], DiscLDA [12] and MedLDA [29] are representative examples. In these models, labels are mutual variables and each document has only one of them. Such groundtruth label is integrated into generative process. In our problem, each document can have multiple tones. Therefore these works can not be applied.

The multi-label variant is the closest to our work. In [18], Ramage et. al. developed a Labeled LDA, extending the LDA for multi-label classification. In this model, each label is corresponding to a latent topic and such label naturally participates in generating words. In [19], they make this model more flexible by allowing both latent and explicit (label) topic co-exist in the model. Rubin et. al. [20] further extends Labeled LDA by adding prior label distribution and label dependencies. These works treat label as binary variable generated by some prior multinomial distribution. In our problem, the tone has intensity and thus is a continuous number. Our method models it as the parameter of Dirchlet process that generates the resulted multinomial distribution.

3 PRELIMINARIES

3.1 Data Preparation

For a bottom-up analysis, there are in general three steps: data crawling, metric identification and tone labeling. The first step is to crawl raw tweet data, the second one is to identify proper metrics and the final one is to label data with the defined tone.

Data Crawing. To study tones involved in the customer service domain, we intentionally select 62 brands with dedicated customer service accounts on Twitter, which cover a large variety of industries, geographical locations and organization scales. We collect the conversational data by capturing both the brands' tweets and tweets

mentioning each of the brands using Twitter Streaming API. Conversations are reconstructed based on the "in_reply_to_status_id" and "in_reply_to_user_id" elements of the responses. Over 2.6M user requests were collected. The collected conversations happened between Jun. 1 and Aug. 1, 2016. We further conduct some preprocessing on the collected dataset by keeping only those conversations received at least one reply and involved only one customer and one agent. All non-English conversations or conversations containing requests or answers with images are also removed from the dataset. After data cleaning, we have 14,118 conversations / 71,171 conversational tweets with 31 brands.

Metric Identification. For tone identification, we pre-select a set of 53 emotional metrics (including aggressive, sincere, apologetic, relieved, sad, etc.) drawing from a number of literatures across different domains, such as marketing [1], linguistic [16], and psychology [8, 17]. We randomly sampled 500 conversations (about 2500 tweets) from the collected dataset and ask crowd workers on CrowdFlower[1] to rate the extent to which these 53 metrics can best describe the selected conversations.

To be more specific, workers are required to annotate on a conversation level, labeling all tweets involved in a conversation. To preserve privacy, we replace all the @mentions appeared in the each conversation as @customer, @[industry]_company (e.g. @ecommerce_company, @telecommunication_company), @competitor_company, and @other_users. For labeling, workers are asked to indicate on a 4-point Likert scale, ranging from "Very strongly" to "Not at all". Advantages of using Likert scale over binary yes/no include: higher tolerance for diverse perceptions and less sparsity for generated labels. Each conversation was labeled by 5 different crowd workers. We restrict the workers to be U.S. residence as well as maintaining some moderate level of accuracy in their previous annotation tasks. Validation questions are embedded in the real tasks to validate the quality of collected labels. The collected ratings were reliably aggregated via a reference-based method[26].

We next perform factor analysis on the labeled data using principal components analysis (PCA). The PCA of the 53 attributes (metrics) revealed a highly interpretable eight-dimensional solution, whose eigenvalues are greater than one. By analyzing the contribution of 53 attributes in each component, we summarize the eight emotional factors as: **anxious**, **excited**, **frustrated**, **impolite**, **polite**, **sad**, **satisfied**, and **sympathetic**.

Tone Labeling. With eight identified tones, again we ask crowd workers to annotate the extent to which each of the tone is demonstrated in a tweet on the 71,171 collected tweets. The process is the same with the one annotating original 53 metrics, as described earlier. In the end, there are 23,785 with non-zero anxious, 30,715 excited, 20,717 frustrated, 9,251 impolite, 63,434 polite, 30,439 sad, 22,758 satisfied, 31,289 sympathetic.

3.2 Word Segmentation

To facilitate modeling, usually the raw textual file is converted to bag-of-word feature vector. In this work, we use Standford NLP Parser [21] to do the word segmentation and part-of-speech tags. Additionally, we apply a few heuristic rules to reserve particular word types in Tweet such as name, hash tag, number sequence, emotion icon and word of negation.

Name & Hash tag. In Twitter, user name has prefix of "@" and hash tag of "#". In preprocessing, we use the name to identify different customer support accounts, which then become the identifier of Domain-specific topic for our Auxiliary Tone Model in Section 4. The hash tag is removed because it is a typical characteristic of Twitter and may bias our model.

Number sequence. In customer support's response, telephone number is sometimes provided for the user. This information is a strong Domain-specific signal. Such number sequence is sometimes segmented by the parser. To avoid this case, we apply heuristic pattern match to extract possible phone numbers in the whole sentence before segmentation. The list of patterns is shown below.

$$[1.]xxx.xxx.xxxx \qquad [1-](xxx)[-]xxx-xxxx$$
$$[1-]xxx-xxx-xxxx \qquad [1-]xxx\ xxx\ xxxx$$

where the x represents any number digits from 0 to 9, and content in square parenthesis stands for "optional" when matching patterns.

Word of negation. Sometimes negative words are used to represent opposite meaning. For example, "...is not helpful", "...hasn't yet arrived". Using bag-of-word feature after segmentation, we would lose such information. Therefore, we use the parser's POS tag to detect negation, then convert the next noun/adjective/verb to negative format, e.g., "neg_helpful", "neg_arrived". Such negative format is treated as a unique word in the feature space.

Emotion icons. In Twitter, some users would like to use emotion icons when publishing tweet. They are short and concise but contains many emotional information. For example, ":)" represents smile, ":(" means sad, "T_T" represents crying. We think each icon may be correlated with tone and thus treat it as a unique word. As the current parser does not support identifying emotion icons, we preprocess each tweet via string match based on a lexicon in Wikipedia [2] to label these emotion icons.

Repeated punctuation. In some cases, multiple punctuation, especially question mark (?) and exclamation mark (!) are used together. For example, "it's awesome!!!", "Things that used to be $129 are now $128!!!! ??????", etc. We think such repeated punctuation has a different meaning from the single one. Thus we transform them to "!_rep" and "?_rep" and treat them as two unique words.

3.3 Label Adjustment

One issue when processing human-labeled data is how much we should trust these labels. This issue is worth more attention especially when one item is labeled by multiple people and the results are quite different. For discrete choice of strength, e.g., ratings, one common method is to calculate the mean value and use it as a ground-truth label. One weakness of this method is that it fails to consider divergence. Consider such two scenarios: i) two people label the item as degree of 2, ii) one person label it as degree of 1

[1] https://www.crowdflower.com

[2] https://en.wikipedia.org/wiki/List_of_emoticons

while the other degree of 3. The mean value in both cases are the same. However, the divergence is different. Apparently the first one receives more agreement and should be more reliable.

In this work, we adopt a statistic method to adjust the labels. Intuitively, the true label of an item is the average value among all people. Obviously it is impossible to obtain such value since it is unrealistic to ask all people to label an item. Instead, a few people are randomly sampled to do the task, and the goal is to infer the real one from sampled values. Also, we make such an assumption that by default the tone label of an item is neural (degree of 0) unless there is a strong evidence. That means we need a mechanism to decay the sample average. This assumption complies with our goal of finding representative words for each tone. If there is big divergence, it suggests ambiguity and thus is not representative.

Formally, let s denote the number of people labeling the item and $\hat{\mu}, \hat{\sigma}^2$ the average and variance of the labels respectively. We decay the sample average $\hat{\mu}$ by the probability that it is no smaller than 0. With central limit theory, we know that the sample average satisfies a Gaussian distribution, i.e., $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/s)$, where μ, σ^2 are the average and variance over the whole population. The adjustment can be written by the following equation.

$$\begin{aligned}\tilde{\mu} = \hat{\mu} \cdot P(\hat{\mu} \le 0 | \mu, \sigma^2, s) &= \hat{\mu} \int_0^{+\infty} \frac{\sqrt{s}}{\sqrt{2\pi}\sigma} e^{-\frac{s(x-\mu)^2}{2\sigma^2}} dx \\ &\approx \hat{\mu} \int_0^{+\infty} \frac{\sqrt{s}}{\sqrt{2\pi}\hat{\sigma}} e^{-\frac{s(x-\hat{\mu})^2}{2\hat{\sigma}^2}} dx = \hat{\mu}\left(1 - \Phi(-\frac{\hat{\mu}}{\hat{\sigma}}\sqrt{s})\right)\end{aligned} \tag{1}$$

where the $\Phi(\cdot)$ represents the cumulative density function for standard Gaussian distribution (mean 0, variance 1).

As can be seen from Equation (1), the discount $1 - \Phi(-\frac{\hat{\mu}}{\hat{\sigma}}\sqrt{s})$ ranges from 0 to 1 and involves three factors: sample mean $\hat{\mu}$, variance $\hat{\sigma}^2$ and size s. Generally, small sample mean has small discount. This design is consistent with our assumption that we tend to believe neural tone (label 0) unless there is strong evidence. Similarly, small sample size leads to small ratio, indicating that we associate reliability proportional to the number of people labeling the item. Finally, the variance is negatively correlated with discount, where small value results in decay ratio close to 1 and vice versa. Particularly, if the variance is 0 (all people choose the same label), there is no decay for the sample mean.

4 TONE LDA MODEL

In this section we describe details of our Tone LDA (*T-LDA*) model. It origins from latent topic modeling, and its generative process follows that of Latent Dirichlet Allocation (LDA). However, by integrating the label information, the model is guided to converge toward our desired direction. For clarity, we list symbols and their meanings in Table 1.

4.1 Generative Process

Generally, the latent topic modeling has a discrete space of latent topic and each latent topic is represented by a distribution over word space. In T-LDA, there are in total three types of topics by design, and not all of them are latent. The first one is tone-related topic, which is known because of human labeling. The second one is domain-specific topic, aiming to capture the frequent words for different domains. Finally there is one auxiliary topic, related to words that support a grammar-correct sentence.

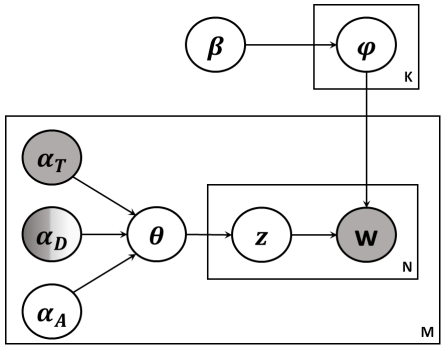

Figure 1: Graphic Representation of T-LDA. Observations are colored white and latent variables are gray. Partially observed variables are half white and half gray.

Figure 1 shows the graphic representation of T-LDA. As can be seen, the tone-related topic density α_T and the word W are the only known variables. Particularly, the domain-specific topic density α_D is half-white, suggesting that it is partially known. The reason is that we know for each document which domain it belongs to, but the density is unknown and needs to be learned from data. In other words, we know whether the α_D is zero or not but have no idea of the specific value if it is non-zero.

Algorithm 1 describes the generative process. Firstly, a prior Dirichlet parameter β generates a word-distribution vectors φ for each topic. Also, each document has its topic density, which generates a topic-distribution vector θ with a Dirichlet distribution. Then for each word a latent variable z is generated according to θ. It indicates which topic is used to generate the next word. Finally, the chosen topic generates a word w based on the corresponding word-distribution vector.

Algorithm 1 T-LDA Generative Process

for Each topic **do**
  choose $\varphi_k \sim \texttt{Dir}([\underbrace{\beta \cdots \beta}_{W}])$ ▷ Dirichlet distribution
end for
for Each document **do**
  choose $\theta_i \sim \texttt{Dir}(\langle \alpha_T^i, \alpha_D^i, \alpha_A^i \rangle)$
  for Each word in the document **do**
    choose $z_{ij} \sim \texttt{Multinomial}(\theta_i)$
    choose $w \sim \texttt{Dir}(\varphi_{z_{ij}})$
  end for
end for

4.2 Model Inference

The model is learned by maximizing the posterior log-likelihood. Formally, let $\mathcal{D}$ and α_T denote the corpus and corresponding tone labels, and $\Theta = \langle \varphi, \alpha_D, \alpha_A \rangle$ the collection of model parameters, the

Table 1: List of Symbols

Symbol	Dimension	Description
M	1	Number of documents
N	1	Number of words in a document
W	1	Total number of words in the whole data set
T, D, A	Set	Set of Tone-related, Domain-specific and Auxiliary topics
$K \equiv \|T\| + \|D\| + \|A\|$	1	Total number of topics
i, j, k	1	Index for document, word and topic
$\alpha_T^i, \alpha_D^i, \alpha_A^i$	vector of $\|T\|, \|D\|, \|A\|$	Topic density of Tone-related, Domain-specific and Auxiliary topics for i^{th} document
β	1	Prior Dirichlet distribution parameter for topic-word vector
φ_k	Vector of W	Word distribution for k^{th} topic
θ_i	Vector of K	Topic distribution vector for i^{th} document
z_{ij}	1	Topic index for j^{th} word in i^{th} document

posterior log-likelihood can be written as below.

$$\begin{aligned}
&\log P(\Theta|\mathcal{D}, \alpha_T, \beta) = \log P(\mathcal{D}|\Theta, \alpha_T) + \log P(\Theta|\alpha_T, \beta) - \log P(\mathcal{D}|\alpha_T) \\
&\propto \log P(\mathcal{D}|\Theta, \alpha_T) + \log P(\Theta|\alpha_T, \beta) = \sum_{i,j} \log P(w_{ij}, z_{ij}|\Theta, \alpha_T) + \log P(\varphi|\beta) \\
&= \sum_{i,j} \log \int P(w_{ij}, z_{ij}, \theta_i|\Theta, \alpha_T) d\theta_i + \sum_k \log P(\varphi_k|\beta) \\
&= \sum_{i,j} \log \Big(\int P(w_{ij}, z_{ij}|\theta_i, \varphi) \cdot P(\theta_i|\alpha_T^i, \alpha_D^i, \alpha_A^i) \Big) d\theta_i + \sum_k \log P(\varphi_k|\beta)
\end{aligned} \tag{2}$$

where we skip the data probability $P(\mathcal{D})$ as it is a constant value free of unknown parameters.

The topic-word distribution φ_k is determined by a prior Dirichlet distribution, thus the probability $P(\varphi_k|\beta)$ is shown below.

$$\log P(\varphi_k|\beta) \propto \log \Big(\prod_w (\varphi_k(w))^{\beta-1} \Big) = (\beta - 1) \sum_w \log \varphi_k(w) \tag{3}$$

Again we omit the parameter-free constant.

Note that by design each single word w_{ij} is generated by only one topic. We further introduce a binary variable $\eta_{ij}^k \in \{1, 0\}$ indicating whether the word w_{ij} is generated by k^{th} topic. In this case, the integral in Equation (2) can be rewritten as below.

$$\begin{aligned}
&\int P(w_{ij}, z_{ij}|\theta_i, \varphi) \cdot P(\theta_i|\alpha_T^i, \alpha_D^i, \alpha_A^i) \Big) d\theta_i \\
&= \prod_k \Big(\int P(w_{ij}, z_{ij} = k|\theta_i, \varphi) \cdot P(\theta_i|\alpha_T^i, \alpha_D^i, \alpha_A^i) d\theta_i \Big)^{\eta_{ij}^k} \\
&= \prod_k \Big(P(w_{ij}|\varphi_k) \int P(z_{ij} = k|\theta_i) \cdot P(\theta_i|\alpha_T^i, \alpha_D^i, \alpha_A^i) d\theta_i \Big)^{\eta_{ij}^k} \\
&= \prod_k \Big(\varphi_k(w_{ij}) \int \theta_i(k) P(\theta_i|\alpha_T^i, \alpha_D^i, \alpha_A^i) d\theta_i \Big)^{\eta_{ij}^k}
\end{aligned} \tag{4}$$

Now the integral is to calculate the expected value of k^{th} element in vector θ_i, which is a random variable satisfying Dirichlet distribution. Based on the probability density function, we can easily compute the expected value as below.

$$\begin{aligned}
\mathbf{E}\Big(\theta_i(k)\Big)_{\theta_i \sim \mathrm{Dir}(\langle \alpha_T^i, \alpha_D^i, \alpha_A^i \rangle)} &= \int \theta_i(k) P(\theta_i|\alpha_T^i, \alpha_D^i, \alpha_A^i) d\theta_i \\
&= \frac{\mathbf{1}_{k \in \alpha_T^i} \alpha_T^i(k) + \mathbf{1}_{k \in \alpha_D^i} \alpha_D^i(k) + \mathbf{1}_{k \in \alpha_A^i} \alpha_A^i(k)}{\sum \alpha_T^i + \sum \alpha_D^i + \sum \alpha_A^i}
\end{aligned} \tag{5}$$

where $\mathbf{1}_{k \in \alpha_*^i}$ is an indicator suggesting whether the k^{th} topic belongs to the particular topic category , i.e., Tone-related (T), Domain-specific (D) or Auxiliary (A).

By connecting all of these pieces, we finally define the objective function $\mathcal{L}(\Theta)$ as below:

$$\begin{aligned}
\mathcal{L}(\Theta) &= \sum_{i,j} \log \Big(\int P(w_{ij}, z_{ij}|\theta_i, \varphi) \cdot P(\theta_i|\alpha_T^i, \alpha_D^i, \alpha_A^i) \Big) d\theta_i \\
&+ (\beta - 1) \sum_w \log \varphi_k(w) \\
&= \sum_{i,j,k} \eta_{ij}^k \Big(\log \varphi_k(w_{ij}) + \log \sum_{\tau \in \{D,A\}} \mathbf{1}_{k \in \alpha_\tau^i} \alpha_\tau^i(k) - \log \sum_{\tau \in \{T,D,A\}} \alpha_\tau^i \Big) \\
&+ (\beta - 1) \sum_{k,w} \log \varphi_k(w)
\end{aligned} \tag{6}$$

The model parameters are learned via maximize the log-likelihood, i.e., $\Theta^* = \arg\max_\Theta \mathcal{L}(\Theta)$ with constraint $\forall k, \sum_w \varphi_k(w) = 1$. However, it is impossible to get the explicit solution due to the existence of unknown variable η_{ij}^k. Thus we use *Expectation-Maximization (EM)* to solve it.[3]

E-step. We first compute the expected value of η_{ij}^k with current parameter. Formally, let $\Theta^{(t)}$ denote the parameter value at t^{th}

[3] Alternatively, without introducing η_{ij}^k, one may use Gibbs-sampling.

iteration. The expected value $\mathbf{E}(\eta_{ij}^k)_{\Theta^{(t)}}$ can be computed as below.

$$\begin{aligned}
&\mathbf{E}(\eta_{ij}^k)_{\Theta^{(t)}} = 1 \cdot P(\eta_{ij}^k = 1|\Theta^{(t)}, \mathcal{D}) + 0 \cdot P(\eta_{ij}^k = 0|\Theta^{(t)}, \mathcal{D}) \\
&= P(z_{ij} = k|\Theta^{(t)}, w_{ij}) = \frac{P(w_{ij}|z_{ij} = k, \varphi_k^{(t)})P(z_{ij} = k|\alpha_*^{i(t)})}{\sum_{k'} P(w_{ij}|z_{ij} = k', \varphi_{k'}^{(t)})P(z_{ij} = k'|\alpha_*^{i(t)})} \\
&= \frac{\varphi_k^{(t)}(w_{ij}) \sum_{\tau \in \{T,D,A\}} \mathbf{1}_{k \in \tau} \alpha_\tau^{i(t)}}{\sum_{k'} \varphi_{k'}^{(t)}(w_{ij}) \sum_{\tau \in \{T,D,A\}} \mathbf{1}_{k' \in \tau} \alpha_\tau^{i(t)}}
\end{aligned} \tag{7}$$

where $\alpha_*^{i(t)}$ represents all topic density (i.e., T, D, A) for i^{th} document at t^{th} iteration.

M-step. With calculated $\mathbf{E}(\eta_{ij}^k)_{\Theta^{(t)}}$, the parameter at next iteration is updated by maximizing the objective function, i.e., $\Theta^{(t+1)} = \arg\max_\Theta \mathcal{L}(\Theta|\mathbf{E}(\eta_{ij}^k)_{\Theta^{(t)}})$ with constraint $\forall k, \sum_w \varphi_k^{(t+1)}(w) = 1$. Specifically, the updated rule for each parameter is given below.

$$\forall k, w \qquad \varphi_k^{(t+1)}(w) = \frac{\beta - 1 + \sum_{i,j} \mathbf{1}_{w_{ij}=w} \mathbf{E}(\eta_{ij}^k)_{\Theta^{(t)}}}{\sum_{w'} \left(\beta - 1 + \sum_{i,j} \mathbf{1}_{w_{ij}=w'} \mathbf{E}(\eta_{ij}^k)_{\Theta^{(t)}}\right)} \tag{8}$$

$\forall i, \tau \in \{D, A\}, k \in \tau$

$$\alpha_\tau^{i(t+1)}(k) = \begin{cases} (\sum \alpha_T^i) \cdot \frac{\mathbf{E}(\eta_{ij}^k)_{\Theta^{(t)}}}{\sum_{k' \in \alpha_T^i} \mathbf{E}(\eta_{ij}^{k'})_{\Theta^{(t)}}} & \text{if } \sum \alpha_T^i \neq 0 \\ \max\{\alpha_T^i | \forall i\} \cdot \frac{\mathbf{E}(\eta_{ij}^k)_{\Theta^{(t)}}}{\max_{k' \notin \alpha_T^i} \mathbf{E}(\eta_{ij}^{k'})_{\Theta^{(t)}}} & \text{otherwise} \end{cases} \tag{9}$$

As can be seen in Equation (9), there are special update rules when tone labels are all zero. These rules are necessary to guarantee unique solution. Recall in Algorithm 1, the generative process from α to θ satisfies Dirichlet distribution, where the probability density function is only affected by the relative scale of parameters. In other words, the topic density parameters are unbounded. In training phase, the α_D, α_A are bounded by the known tone labels α_T. But if the tone labels are all zero, there will be infinite solutions for domain-specific and auxiliary topic density. To avoid this situation, any arbitrary positive value can be used as "*anchor*". In this work, we choose the maximum value of tone topic density so that density of other topics fall within the same scale.

4.3 Model Application

With learned T-LDA, there are a number of applications in analyzing tones. We summarize all of them in this section.

Representative words for each tone. One goal of T-LDA is to help people understand different tones via representative words. With T-LDA, we obtain these words in two steps, i) calculate posterior probability and ii) adjust by Auxiliary topic. With no loss of generality, given a particular word w, we can compute the posterior probability for a topic $z \in \{T, D, A\}$ in Equation (10).

$$P(z|w) = \frac{P(z, w)}{P(w)} = \frac{P(w|z)P(t)}{\sum_{z'} P(w|z')P(z')} = \frac{\varphi_z(w)}{\sum_{z'} \varphi_{z'}(w)} \tag{10}$$

Note that in the equation, we assume a uniform probability for selecting a topic, i.e., $P(z_i) \equiv P(z_j)$. After step i), we obtain the normalized posterior probability for each topic. To obtain the representative words, we subtract it by the Auxiliary one, i.e., $P(z|w) - \sum_{z' \in A} P(z'|w)$. This adjustment is to remove words that are representative in both target topic and Auxiliary one. Recall that by model design and learning, Auxiliary topic is for meaningless words. To find representative words for meaningful topic (either tone or domain), we need to remove the noisy words in Auxiliary.

Example sentences for each tone. Each sentence can be treated as a bag-of-word set. With calculated posterior probability for each word, the ranking of the sentence is the sum of these probabilities, i.e., $\sum_w P(z|w)$. Here we do not use Auxiliary probability for adjustment. Because in practice we found that the impact of Auxiliary topic is reduced due to the consideration of multiple words.

Document topic distribution. Like all latent-topic models, one of the application is to infer a document's topic distribution vector, i.e., θ, given its bag-of-word feature. Naturally, given a document D, the corresponding topic distribution θ^D is obtained by maximizing the posterior likelihood, as shown in the equation below.

$$\theta^D = \arg\max_\theta P(\theta|D) = \arg\max_\theta \frac{\prod_{w \in D} \sum_z P(w|z)P(z|\theta)}{\prod_{w \in D} P(w)} \tag{11}$$

Obviously it is time-consuming. Thus a trade-off is to calculate each latent topic and word independently and then normalize them. Formally, let $\theta^D(k)$ denote the k^{th} element of the topic distribution vector, its value is calculated as below.

$$\theta^D(k) = \frac{\sum_{w \in D} P(z = k|w)}{\sum_{w \in D} \sum_{z'} P(z'|w)} \tag{12}$$

5 EVALUATION AND CASE STUDY

In this section we evaluate the proposed Tone LDA (T-LDA) model in two aspects, regression task and representation task. The former one checks its performance on predicting tone density. The latter one is a case study and analyzes its output of representative words and sentences for each tone.

5.1 Tone Intensity Regression

For regression, T-LDA is used to predict the topic distribution vector θ, then a separate regression model is used to predict the real tone density label. That means, the T-LDA is treated as a feature extraction in this process. Similarly, we can use other topic models, e.g., LDA as baseline for comparison. Also, instead of using topic models to extract feature, the raw bag-of-word feature can be used to train a regression model directly (denoted as *Raw*). With no loss of generality, in regression we normalize the tone density to the range of 0 - 1 and logistic regression is trained.

The *mean squared error (MSE)* is used as the evaluation metric. Formally, let y denote the labeled tone density and $\hat{y}$ is the predicted value. For each tone, the MSE on a test data $\mathcal{Y}$ is defined as follows.

$$MSE_{\mathcal{Y}} = \frac{1}{|Y|} \sum_{y \in \mathcal{Y}} (y - \hat{y})^2 \tag{13}$$

Table 2 shows the average results with 5-fold cross-validation. As can be seen, the T-LDA achieves lowest error for all tones except Excited. The regression model trained on raw bag-of-word has highest error in all. This suggests training on raw word is not as good as extracted features. Also, in topic model training, the LDA does not use tone density information while the T-LDA does. Although in the final regression training, such information is used

Table 2: MSE of Tone Regression

-	Raw	LDA	T-LDA
Anxious	0.00658	0.00653	**0.00652**
Excited	0.00698	**0.00647**	0.00669
Frustrated	0.01201	0.01194	**0.01175**
Impolite	0.00532	0.00563	**0.00555**
Polite	0.06140	0.06209	**0.05258**
Sad	0.02492	0.02425	**0.02307**
Satisfied	0.02205	0.02108	**0.01978**
Sympathetic	0.01286	0.01195	**0.01185**

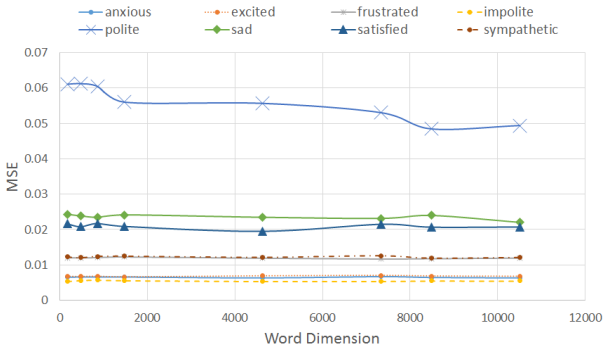

Figure 2: Impact of Word Dimension

in both methods, T-LDA shows better performance. Finally, since integrating the tone information during topic modeling, the original latent topic has explicit information and can be easily interpreted.

We then test the impact of word dimension. In the experiment, we define each word's *tweet coverage* as the number of tweets in which this word appears. With different tweet coverage thresholds, the number of words included will be different. Obviously, the smaller the threshold is, the larger the word dimension is. In previous experiment, we set the the threshold to be 4, where the word dimension is 8,488. Here we change it from 500 to 3, resulting in a word dimension from 178 to 10,504. The result is shown in Figure 2. As can be seen, in general the word dimension does not have too much impact on the regression error. For tone "polite", the performance is stable after 8,000+ words, i.e., 4 for tweet coverage threshold.

5.2 Representative of Tone

In this section we show a case study of how the T-LDA is used to find representative words and sentences for each tone. After trained on the whole data set with tweet coverage threshold as 4, we use equations in Section 4.3 to find representative words and sentences for each tone. The result is shown in Table 3.

From the table we can see that some tones, e.g., frustrated, impolite, sad, are easy to understand as there are strongly associated words. Some representative words in the table are not quite straightforward, because they have a high co-occurrence frequency with other sensible words. For instance, the words "us" and "know" have high score for tone polite. The reason is that they usually come with word "please", such as "please let us know", "please follow us", and so on. Furthermore, there are words that does not seem to correlate with the particular tone when viewed separately. However, once several words are combined, it highly indicates the tone. For example, in tone anxious, single representative words contain "anyone","need", "help". There is no obvious correlation between them and anxiety. However, it makes more sense if one sees sentences like "can anyone help me", "I need help".

As a side product of T-LDA, we can also analyze the top words of domain-specific topic, which in this work, represents a company. There are in total 31 companies in the data set. Due to the limited space, we only show some of them. Table 4 lists top words of some companies. As can be seen, in general these words can be categorized into three following aspects.

- Contact information. When customer service replies, common contact information, such as telephone, URL, email,etc, is attached. This seems to be response template.
- Company-related words. These words are highly associated with the company. Some are related to the company's service, e.g., package, tracking for @upshelp. Others are the company's product, e.g., kindle, prime for @amazonhelp, xbox, controller, console for @xboxsupport.
- Common issues. This category reflects common problems complained by customers. For instance, the "neg_arrive", "neg_arrived" for @amazonhelp represent the ordered package is not received. Also, "neg_service" and "neg_internet" for @attcares means the AT&T telephone has issues with internet connection.

5.3 Impact of Domain and Auxiliary topic

Recall that in our model, there are three types of topics, i.e., tone-related, domain-specific and auxiliary. The latter two types are used to capture words that are not tone-correlated but appear in document labeled with tones. Specifically the domain-specific contain words that have unique signature of that particular domain (company). And the auxiliary topic is to model frequent words that are neither tone-related nor domain-specific. In this experiment we show how these two types of topic help improve the tone analysis.

To study the impact of these two topic types, we switch off the domain and auxiliary topic during training. Then we can compare the resulted top words with that of full model. The comparison is shown in Table 5. For the model with tone topic, we denote as *tone-only*. For the one without auxiliary, we denote as *tone+domain*. Note that we skip tones where there is no significant difference among models. As can be seen, in general the model with all three topic types has a more sensible set of top words for tones. For other two, however, there are words that are obviously domain-specific or auxiliary, which we underline in the table. For tone *anxious*, word "xbox" is ranked high in tone-only model but it is actually a product (thus domain-specific). The model without auxiliary ranks "password", "email" and "account" as top words, which are more suitable in auxiliary category. For tone *excited*, auxiliary words "still", "not" and domain-specific words "tesco" appear in the list of tone-only model. And the auxiliary word "tomorrow" is ranked high by tone+domain model. Finally for tone sympathetic, the tone-only model includes a domain-specific URL link.

Table 3: Representative Words and Sentences of Tone

Tone	Representative words	Example sentences (tone density label)
anxious	need worried anyone confused help wondering address suddenly going ideas	• *@HPSupport My DVD drive stopped reading and burning, I just need someone to help me find a fix, I don't need Hector interrupting me.* (2.4) • *@BlackBerryHelp need help on the substandard phone iv got I need help. No one is ready to help!!* (1.2) • *@ArgosHelpers Are you going to sort the website out today, need to do serious shopping and need to log in?????* (1.2)
excited	! :D love sure excited awesome always best hearing fun	• *@HPSupport TY! I got Tech Support on ??; All IS Good! U Guys R The BEST In the Market Place! TY Always!* (2.6) • *@AmazonHelp its fine now just got a notfication that its been dispatched thanks! Alot any trouble i know i can relie on you guys! Im happy!* (2.4) • *Thanks for the update! If you run into anything else, let us know!* (3.0)
frustrated	ridiculous annoying useless worst annoyed terrible frustrating rude customer frustrated !_rep	• *@ATTCares just dm'd you. This is ridiculous. I've been on hold for over 1 hour and 40 minutes. I'm beyond frustrated.* (3.0) • *@Walmart is the worst!! They're customer service sucks so bad* (2.8) • *@SouthwestAir flight 1776 terrible customer service and attitude from flight attendant Terese. Rude with all passengers. Terrible.* (2.6)
impolite	sh*t s*cks f*ck f*cking wtf s*ck damn stupid hell ass	• *F*ck you Wells Fargo f*ck you so hard f*ck you f*ck you f*ck you* (2.8) • *@ChaseSupport yet again I wait in line at drive thru ATM for 30 minutes; the stupid fucking machine cannot take check deposits. F*CK YALL* (2.8) • *@RogersHelps f*cking u all msg me like 13 hours later wut kind of sh*tty customer service is this sh*t* (3.0)
polite	please us know reaching team assistance letting assist glad reach welcome	• *Having billing issues? We can help here. Please Follow/DM us your acct and contact info please.* (3.0) • *Please Alex, allow us to troubleshoot your DVR for you. Please follow/Dm your account info to begin* (3.0) • *Ok, cool! Please follow the steps to the link provided so you can secure your account. Please let us know! Thanks* (3.0)
sad	:(disappointed bad nothing neg_work neg_happy even problem really apparently	• *@VirginTrains really disappointed no sarni's on the 1240 Euston to Mancs :(* (2.8) • *@AmazonHelp really bad customer service on the phone, even more disappointed. What's an email address to complain to?* (2.8) • *My @amazon order was supposed to be delivered yesterday, now it ""should"" be here by Thursday. I'm not happy.* (2.6)
satisfied	thank thanks ok much great cheers sorted good helpful resolved :)	• *@SamsungSupport hey thanks! I'm able to log in. Thanks you rock. :) thanks for the fast service* (3.0) • *@AmazonHelp thank you problemo resolved. Thank you so much for your prompt reply* (3.0) • *@SamsungSupport Ok, thank you!! I will try to contact them! Thanks!* (3.0)
sympathetic	sorry hear inconvenience look apologies us frustration apologize experience trouble	• *Thanks for that. We'll take a look. Sorry to hear the item arrived broken.* (3.0) • *For assistance, pls call our Serve team @ 8009540559; they'll investigate. Sorry for any inconvenience. Enjoy your day.* (2.2) • *We apologize for any inconvenience caused. Please follow/DM us if we can be of any assistance. We're always glad to help.* (0.6)

For impolite, we use "*" to cover some letters of offensive words.

6 CONCLUSION AND FUTURE WORK

In this work we conduct a bottom-up emotion analysis on online conservations of customer service. Particularly we define 8 *tones* as new metric for emotional measurement. We also extend LDA to *Tone LDA (T-LDA)* to model tone intensity and find corresponding descriptive words as well as sentences. By evaluation on Twitter data, we demonstrate its better topic modeling than baselines.

Our current T-LDA works on single tweet. In reality, a complete conservation consists of multiple tweets between a customer and an agent. Thus tone of consequent tweets may have impact on each

Table 4: Top words of Domain-specific Topic

Company	Top words
@amazonhelp	prime amazon https://t.co/haplpmlfhn order kindle orders ordered neg_arrive
@ask_wellsfargo	banker compliment numbers neg_account location 1-800-869-3557 atm neg_acct
@deltaassist	delta https://t.co/6idgbjac2m standby cincinnati departing neg_confirm skymiles diamond captain neg_checked
@samsungsupport	model samsung s7 gear galaxy television edge verizon marshmallow carrier note
@xboxsupport	xbox console controller preview guide https://t.co/luv7xycgtq game 360 disc games

Table 5: Impact of Domain and Auxiliary Topic

	Tone only	Tone+Domain	Tone+Domain+Auxiliary (T-LDA)
anxious	worried wondering confused xbox ideas	password email account wondering need	need worried anyone confused help
excited sympathetic	excited still tesco fun not sorry apologies apologize inconvenience https://t.co/y5jpi9grhe	:D excited tomorrow fun neg_wait sorry hear inconvenience look apologize	! :D love sure excited sorry hear inconvenience look apologies

other. For future work, we plan to extend T-LDA to capture the whole conversation session and include such potential factors.

REFERENCES

[1] Jennifer L Aaker. 1997. Dimensions of brand personality. *Journal of marketing research* (1997), 347–356.

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, 30–38.

[3] David Andrzejewski and Xiaojin Zhu. 2009. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*. Association for Computational Linguistics, 43–48.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.

[6] Laurence Devillers, Lori Lamel, and Ioana Vasilescu. 2003. Emotion detection in task-oriented spoken dialogues. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, Vol. 3. IEEE, III–549.

[7] Brian Dickinson and Wei Hu. 2015. Sentiment analysis of investor opinions on twitter. *Social Networking* 4, 03 (2015), 62.

[8] Hans Jurgen Eysenck. 1953. The structure of human personality. (1953).

[9] Jonathan Gratch and Stacy Marsella. 2001. Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In *Proceedings of the fifth international conference on Autonomous agents*. ACM, 278–285.

[10] L Jean Harrison-Walker. 2001. The measurement of word-of-mouth communication and an investigation of service quality and customer commitment as potential antecedents. *Journal of service research* 4, 1 (2001), 60–75.

[11] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 204–213.

[12] Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. 2009. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*. 897–904.

[13] Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*. 121–128.

[14] Keith B Murray. 1991. A test of services marketing theory: consumer information acquisition activities. *The journal of marketing* (1991), 10–25.

[15] Richard L Oliver. 1997. Emotional expression in the satisfaction response. *Satisfaction: A behavioral perspective on the consumer* (1997), 291–325.

[16] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[17] Rosalind Wright Picard. 1995. Affective computing. (1995).

[18] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 248–256.

[19] Daniel Ramage, Christopher D Manning, and Susan Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 457–465.

[20] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine learning* 88, 1 (2012), 157–208.

[21] Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

[22] Amy K Smith and Ruth N Bolton. 2002. The effect of customersâĂŹ emotional responses to service failures on their recovery effort evaluations and satisfaction judgments. *Journal of the academy of marketing science* 30, 1 (2002), 5–23.

[23] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307.

[24] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2544–2558.

[25] Yang Wang, Payam Sabzmeydani, and Greg Mori. 2007. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. *Human Motion–Understanding, Modeling, Capture and Animation* (2007), 240–254.

[26] Anbang Xu and Brian Bailey. 2012. A reference-based scoring model for increasing the findability of promising ideas in innovation pipelines. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1183–1186.

[27] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. 3506–3510.

[28] Peifeng Yin, Nilam Ram, Wang-Chien Lee, Conrad Tucker, Shashank Khandelwal, and Marcel Salathé. 2014. Two sides of a coin: Separating personal communication and public dissemination accounts in Twitter. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 163–175.

[29] Jun Zhu, Amr Ahmed, and Eric P Xing. 2009. MedLDA: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th annual international conference on machine learning*. ACM, 1257–1264.