Public Dialogue: Analysis of Tolerance in Online Discussions

Arjun Mukherjee[†] **Vivek Venkataraman**[†] **Bing Liu**[†] **Sharon Meraz**[‡]
[†]Department of Computer Science [‡]Department of Communication
University of Illinois at Chicago
arjun4787@gmail.com {vvenka6, liub, smeraz}@uic.edu

Abstract

Social media platforms have enabled people to freely express their views and discuss issues of interest with others. While it is important to discover the topics in discussions, it is equally useful to mine the nature of such discussions or debates and the behavior of the participants. There are many questions that can be asked. One key question is whether the participants give reasoned arguments with justifiable claims via constructive debates or exhibit dogmatism and egotistic clashes of ideologies. The central idea of this question is *tolerance*, which is a key concept in the field of communications. In this work, we perform a computational study of tolerance in the context of online discussions. We aim to identify tolerant vs. intolerant participants and investigate how disagreement affects tolerance in discussions in a quantitative framework. To the best of our knowledge, this is the first such study. Our experiments using real-life discussions demonstrate the effectiveness of the proposed technique and also provide some key insights into the psycholinguistic phenomenon of tolerance in online discussions.

1 Introduction

Social media platforms have enabled people from anywhere in the world to express their views and discuss any issue of interest in online discussions/debates. Existing works in this context include recognition of support and oppose camps (Agrawal et al., 2003), mining of authorities and subgroups (Mayfield and Rosè, 2011; Abu-Jbara et al. (2012), dialogue act segmentation and classification (Morbini and Sagae, 2011; Boyer et al., 2011), etc.

This paper probes further to study a different and important angle, i.e., the psycholinguistic phenomenon of *tolerance* in online discussions. Tolerance is an important concept in the field of communications. It is a subfacet of deliberation which refers to critical thinking and exchange of rational arguments on an issue among participants that seek to achieve consensus/solution (Habermas, 1984).

Perhaps the most widely accepted definition of tolerance is that of Gastil (2005; 2007), who defines tolerance as a means to engage (in written or spoken communication) in critical thinking, judicious argument, sound reasoning, and justifiable claims through constructive discussion as opposed to mere coercion/egotistic clashes of ideologies.

In this work, we adopt this definition, and also employ the following characteristics of tolerance (also known as "code of conduct") (Crocker, 2005; Gutmann & Thompson, 1996) to guide our work.

Reciprocity: Each member (or participant) offers proposals and justifications in terms that others could understand and accept.

Publicity: Each member engages in a process that is transparent to all and each member knows with whom he is agreeing or disagreeing.

Accountability: Each member gives acceptable and sound reasons to others on the various claims or proposals suggested by him.

Mutual respect and civic integrity: Each member's speech should be morally acceptable, i.e., using proper language irrespective of agreement or disagreement of views.

The issue of tolerance has been actively researched in the field of communications for the past two decades, and has been investigated in multiple dimensions. However, existing studies are typically qualitative and focus on theorizing the sociolinguistic aspects of tolerance (more details in §2).

With the rapid growth of social media, the large volumes of online discussions/debates offer a golden opportunity to investigate people's implicit psyche in discussions quantitatively based on the real-life data, i.e., their tolerance levels and their arguing nature, which are of fundamental interest to several fields, e.g., communications, marketing,

politics, and sociology (Dahlgren, 2005; Gastil, 2005; Moxey & Sanford, 2000). Communication and political scholars are hopeful that technologies capable of identifying tolerance levels of people on social issues (often discussed in online discussions) can render vital statistics which can be used in predicting political outcomes in elections and helpful in tailoring voting campaigns and agendas to maximize winning chances (Dahlgren, 2002).

Objective: The objective of this work is two-fold:

1. Identifying tolerant and intolerant participants in discussions.
2. Analyzing how disagreement affects tolerance and estimating the tipping point of such effects.

To the best of our knowledge, these tasks have not been attempted quantitatively before. The first task is a classification/prediction problem. Due to the complex and interactive nature of discussions, the traditional n-gram features are no longer sufficient for accurate classification. We thus propose a generative model, called DTM, to discover some key pieces of information which characterize the nature of discussions and their participants, e.g., the arguing nature (agreeing vs. disagreeing), topic and expression distributions. These allow us to generate a set of novel features from the estimated latent variables of DTM capable of capturing authors' tolerance psyche during discussions. The features are then used in learning to identify tolerant and intolerant authors. Our experimental results show that the proposed approach is effective and outperforms several strong baselines significantly.

The second task studies the interplay of tolerance and disagreement. It is well-known that tolerance facilitates constructive disagreements, but sustained disagreements often result in a transition to destructive disagreement leading to polarization and intolerance (Dahlgren, 2005). An interesting question is: What is the tipping point of disagreement to exhibit intolerance? We take a Bayesian approach to seek an answer and discover issue-specific tipping points. Our empirical results discover some interesting relationships which are supported by theoretical studies in psychology and linguistic communications.

Finally, this work also produces an annotated corpus of tolerant and intolerant users in online discussions across two domains: politics and religion. We believe this is the first such dataset and will be a valuable resource to the community.

2 Related Work

Although limited work has been done on analysis of tolerance in online discussions, there are several general research areas that are related to our work.

Communications: Tolerance has been an active research area in the field of communications for the past two decades. Ryfe (2005) provided a comprehensive survey of the literature. The topic has been studied in multiple dimensions, e.g., opinion and attitude (Luskin et al., 2004; Price et al., 2002), public engagement (Escobar, 2012), psychoanalysis (Slavin and Kriegman, 1992), argument repertoire (Cappella et al., 2002), etc.

Tolerance has also been investigated in the domain of political communications with an emphasis on political sophistication (Gastil and Dillard, 1999), civic culture (Dahlgren, 2002), and democracy (Fishkin, 1991). These existing works study tolerance from the qualitative perspective. Our focus is quantitative analysis.

Sentiment analysis: Sentiment analysis determines positive or negative opinions expressed on topics (Liu, 2012; Pang and Lee, 2008). Main tasks include aspect extraction (Hu and Liu, 2004; Popescu and Etzioni, 2005; Mukherjee and Liu, 2012c; Chen et al., 2013), opinion polarity identification (Hassan and Radev, 2010; Choi and Cardie, 2010) and subjectivity analysis (Wiebe, 2000). Although related, tolerance is different from sentiment. Sentiments are mainly indicated by sentiment terms (e.g., *great*, *good*, *bad*, and *poor*). Tolerance in discussions refers to the reception of certain views and often indicated by agreement and disagreement expressions and other features (§5).

Online discussions or debates: Several works put authors in debate into support and oppose camps. Agrawal et al. (2003) used a graph based method, and Murakami and Raymond (2010) used a rule-based method. In (Mukherjee and Liu, 2012a), contention points were identified, in (Mukherjee and Liu, 2012b), various expressions in review comment discussions were mined, and in (Galley et al., 2004; Hillard et al., 2003), speaker utterances were classified into agreement, disagreement, and backchannel classes. Also related are studies on linguistic style accommodation (Mukherjee and Liu, 2012d) and user pair interactions (Mukherjee and Liu, 2013) in online debates. However, these works do not consider tolerance analysis in debate discussions, which is the focus of this work.

In a similar vein, several classification methods have been proposed to recognize opinion stances and speaker sides in online debates (Somasundaran and Wiebe, 2009; Thomas et al., 2006; Bansal et al., 2008; Burfoot et al., 2011; Yessenalina et al., 2010). Lin and Hauptmann (2006) also proposed a method to identify opposing perspectives. Abu-Jbara et al. (2012) identified subgroups. Kim and Hovy (2007) studied election prediction by analyzing online discussions. Other related works studying dialogue and discourse in discussions include authority recognition (Mayfield and Rosè, 2011), dialogue act segmentation and classification (Morbini and Sagae, 2011; Boyer et al., 2011), discourse structure prediction (Wang et al., 2011).

All these prior works are valuable. But they are not designed to identify tolerance or to analyze tipping points of disagreements for intolerance in discussions which are the focus of this work.

3 Discussion/Debate Data

For this research, we used discussion posts from Volconvo.com. This forum is divided into various domains: Politics, Religion, Science, etc. Each domain consists of multiple discussion threads. Each thread consists of a list of posts. Our experimental data is from two domains, Politics and Religion. The data is summarized in Table 1(a). In this work, the terms users, authors and participants are used interchangeably. The full data is used for modeling, but 436 and 501 authors from Politics and Religion domains were manually labeled as being tolerant or intolerant (Table 1(c)) respectively for classification experiments.

Two judges (graduate students) were used to label the data. The judges are fluent in English and were briefed on the definition of tolerance (see §1). From each domain (Politics, Religion), we randomly sampled authors having not more than 60 posts in order to reduce the labeling burden as the judges need to read all posts and see all interactions of each author before providing a label. Given all posts by an author, a and his/her associated interactions (posts by other authors replying or quoting a), the judges were asked to provide a label for author a as being *tolerant* or *intolerant.* In our labeling, we found that users strongly exhibit one dominant trait: tolerant or intolerant, as our data consists of topics like elections, immigration, theism, terrorism, and vegetarianism across politics and religion domains, which are often heated and thus attract people with pre-determined, strong, and polarized stances[1].

Domain	Posts	Authors
Politics	48605	1027
Religion	66835	1370

(a) Full Data

Cohen's κ
0.74
0.77

(b) Agreement

Tol.	Intol.	Total
213	223	436
207	294	501

(c) Labeled data

Table 1: Data statistics (Tol: Tolerant users; Intol: Intolerant users. Total = Tol. + Intol).

The judges worked in isolation (to prevent bias) during annotation/labeling and were also asked to provide a short reason for their judgment. The agreement statistics using Cohen's kappa are given in Table 1(b), which shows substantial agreements according to the scale[2] in (Landis and Koch, 1977). This shows that tolerance as defined in §1 is quite decisive and one can decide whether a debater is exhibiting tolerant vs. intolerant quite well. To account for disagreements in labels, the judges discussed their reasons to reach a consensus. The final labeled data is reported in Table 1(c).

4 Model

We now present our generative model to capture the key aspects of discussions/debates and their intricate relationships, which enable us to (1) design sophisticated features for classification and (2) perform an in-depth analysis of the interplay of disagreement and tolerance. The model is called Debate Topic Model (DTM).

DTM is a semi-supervised generative model motivated by the joint occurrence of various topics; and agreement and disagreement expressions (abbreviated AD-expressions hereon) in debate posts. A typical debate post mentions a few topics (using similar topical terms) and expresses some viewpoints with one or more AD-expression types (Agreement and Disagreement) using semantically related expressions. This observation forms the basis of the generative process of our model where documents (posts) are represented as admixtures of latent topics and AD-expression types (Agreement and Disagreement). This key observation and the

[1] These hardened perspectives are theoretically supported by the polarization effect (Sunstein, 2002), and the hostile media effect, a scenario where partisans rigidly hold on to their stances (Hansen & Hyunjung, 2011).

[2] Agreement levels are as follows. $\kappa \in [0, 0.2]$: Poor, $\kappa \in (0.2, 0.4]$: Fair, $\kappa \in (0.4, 0.6]$: Moderate, $\kappa \in (0.6, 0.8]$: Substantial, and $\kappa \in (0.8, 1.0]$: Almost perfect agreement.

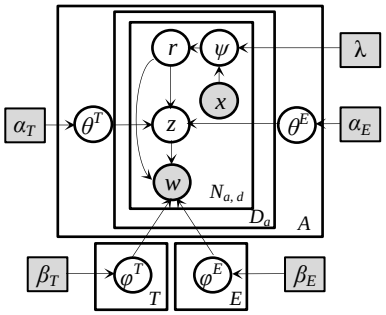

Figure 1: Plate notation of DTM

Variable/Function	Description
$a; A; d$	An author a; set of all authors; document, d
$(a, d); D_a$	Post d by author a; Set of all posts by a
$T; E; V$	# of topics; expression types; vocabulary
$w_{a,d,j}; N_{a,d}$	j^{th} term in (a, d); Total # of terms in (a, d)
$\psi_{a,d,j}$	Distribution over topics & AD-expressions
$\overrightarrow{x_{a,d,j}}$	Associated feature vector of observed $w_{a,d,j}$
λ	Learned Max-Ent parameters
$r_{a,d,j} \in \{\hat{t}, \hat{e}\}$	Binary indicator/switch variable (topic ($\hat{t}$) or AD-expression ($\hat{e}$)) for $w_{a,d,j}$
θ_a^T; $\theta_a^E(\theta_{a,Ag}^E$, $\theta_{a,DisAg}^E)$	a's distribution over topics ; expression types (Agreement: $\theta_{a,Ag}^E$, Disagreement: $\theta_{a,DisAg}^E$)
$\theta_{a,d}^T; \theta_{a,d,t}^T$	Topic distribution of post d by author a; Probability mass of topic t in $\theta_{a,d}^T$.
$\theta_{a,d}^E$; $\theta_{a,d,e\in\{Ag,DisAg\}}^E$	Expression type distribution of post d by author a; Corresponding probability masses of Agreement: $\theta_{a,d,e=Ag}^E$ and Disagreement in $\theta_{a,d,e=DisAg}^E$.
$z_{a,d,j}$	Topic/Expression type of $w_{a,d,j}$
φ_t^T; φ_e^E	Topic t's ; Expression type e's distribution over vocabulary terms
$\alpha^T; \alpha^E; \beta^T; \beta^E$	Dirichlet priors of θ_a^T; $\theta_a^E; \varphi_t^T$; φ_e^E
$n_{a,t}^{AT}; n_{a,e}^{AE}$	# of times topic t; expression type e assigned to a
$n_{t,v}^{TV}; n_{e,v}^{EV}$	# of times term v appears in topic t; expression type e

Table 2: List of notations

motivation of modeling debates are from our previous work in (Mukherjee and Liu, 2012a). In the new setting, we model topics and debate expression distributions specific to authors as this work is concerned with modeling authors' (in)tolerance nature. Making latent variable θ^E and θ^T author specific facilitates modeling user behaviors (§5.3).

Assume we have $t_{1...T}$ topics and $e_{1...E}$ expression types in our corpus. In our case of debate posts, based upon reading various posts, we hypothesize that $E = 2$ as in debates as we mostly find 2 dominant expression types: Agreement and Disagreement. Meanings of variables used in the following discussion are detailed in Table 2. In this work, a document/post is viewed as a bag of *n*-grams and we use terms to denote both words (unigrams) and phrases (*n*-grams)[3]. DTM is a switching graphical model performing a switch between topics and AD-expressions similar to that in (Zhao et al., 2010). The switch is done using a learned maximum entropy (Max-Ent) model. The rationale here is that topical and AD-expression terms usually play different syntactic roles in a sentence. Topical terms (e.g., "U.S. elections," "government," "income tax") tend to be noun and noun phrases while expression terms ("I refute," "how can you say," "I'd agree") usually contain pronouns, verbs, wh-determiners, and modals. In order to utilize the part-of-speech (POS) tag information, we place the topic/AD-expression distribution, $\psi_{a,d,j}$ (the prior over the indicator variable $r_{a,d,j}$) in the term plate (Figure 1) and set it using a Max-Ent model conditioned on the observed context $x_{a,d,j}$ associated with $w_{a,d,j}$ and the learned Max-Ent parameters λ (details in §4.1). In this work, we use both lexical and POS features of the previous, current and next POS tags/lexemes of the term $w_{a,d,j}$ as the contextual information, i.e., $x_{a,d,j} = [POS_{w_{a,d,j-1}}, POS_{w_{a,d,j}}, POS_{w_{a,d,j+1}}, w_{a,d,j-1}, w_{a,d,j}, w_{a,d,j+1}]$, which is used to produce feature functions for Max-Ent. For phrasal terms (*n*-grams), all POS tags and lexemes of $w_{d,j}$ are considered as contextual information for computing feature functions in Max-Ent. DTM has the following generative process:

A. For each AD-expression type e, draw $\varphi_e^E \sim Dir(\beta^E)$
B. For each topic t, draw $\varphi_t^T \sim Dir(\beta^T)$
C. For each author $a \in \{1 \dots A\}$:
   i. Draw $\theta_a^E \sim Dir(\alpha^E)$
   ii. Draw $\theta_a^T \sim Dir(\alpha^T)$
   iii. For each document/post $d \in \{1 \dots D_a\}$:
      I. For each term $w_{a,d,j}$, $j \in \{1 \dots N_{a,d}\}$:
         a. Set $\psi_{a,d,j} \leftarrow MaxEnt(x_{a,d,j}; \lambda)$
         b. Draw $r_{a,d,j} \sim Bernoulli(\psi_{a,d,j})$
         c. if ($r_{a,d,j} = \hat{e}$) // $w_{d,j}$ is an AD-expression term
            Draw $z_{a,d,j} \sim Mult(\theta_a^E)$
            else // $r_{a,d,j} = \hat{t}$, $w_{a,d,j}$ is a topical term
            Draw $z_{a,d,j} \sim Mult(\theta_a^T)$
         d. Emit $w_{a,d,j} \sim Mult(\varphi_{z_{a,d,j}}^{r_{a,d,j}})$

[3] Topics in most topic models (e.g., LDA (Blei et al., 2003)) are unigram distributions and a document is treated as an exchangeable bag-of-words. This offers a computational advantage over models considering word orders (Wallach, 2006). As our goal is to enhance the expressiveness of DTM (rather than "modeling" word order), we use 1-4 grams preserving the advantages of exchangeable modeling.

Disagreement expressions ($\varphi^E_{e=Disagreement}$)
I, disagree, I don't, I disagree, **argument**, reject, **claim**, I reject, I refute, **and**, **your**, I refuse, **won't**, **the claim**, nonsense, *I contest*, dispute, **I think**, completely disagree, don't accept, don't agree, incorrect, **doesn't**, *hogwash*, *I don't buy your*, *I really doubt*, your nonsense, **true**, *can you prove*, argument fails, *you fail to*, **your assertions**, *bullshit*, *sheer nonsense*, *doesn't make sense*, *you have no clue*, *how can you say*, *do you even*, *contradict yourself*, …
Agreement expressions ($\varphi^E_{e=Agreement}$)
agree, **I**, correct, yes, true, accept, I agree, **don't**, indeed correct, **your**, **point**, **that**, I concede, is valid, **your claim**, **not really**, *would agree*, **might**, *agree completely*, yes indeed, absolutely, you're correct, *valid point*, **argument**, **the argument**, proves, *do accept*, support, agree with you, *rightly said*, **personally**, well put, *I do support*, *personally agree*, **doesn't necessarily**, exactly, *very well put*, absolutely correct, *kudos*, *point taken*,...

Table 3: Top terms (comma delimited) of two expression types. **Red** (**bold**) terms denote possible errors. *Blue* (*italics*) terms are newly discovered; rest (black) terms have been used in Max-Ent training.

4.1 Inference

We employ posterior inference using Monte Carlo Gibbs sampling. Denoting the random variables $\{w, z, r\}$ by singular subscripts $\{w_k, z_k, r_k\}$, $k_{1\ldots K}$, where $K = \sum_a \sum_d N_{a,d}$, a single iteration consists of performing the following sampling:

$$p(z_k = t, r_k = \hat{t} | W_{\neg k}, Z_{\neg k}, R_{\neg k}, w_k = v) \propto \frac{\exp(\sum_{i=1}^{n} \lambda_i f_i(x_{a,d,j,\hat{t}}))}{\sum_{y \in \{\hat{t},\hat{e}\}} \exp(\sum_{i=1}^{n} \lambda_i f_i(x_{a,d,j,y}))} \times \frac{n^{AT}_{a,t\ \neg k} + \alpha^T}{n^{AT}_{a,(\cdot)\ \neg k} + T\alpha^T} \times \frac{n^{TV}_{t,v\ \neg k} + \beta^T}{n^{TV}_{t,(\cdot)\ \neg k} + V\beta^T} \quad (1)$$

$$p(z_k = e, r_k = \hat{e} | W_{\neg k}, Z_{\neg k}, R_{\neg k}, w_k = v) \propto \frac{\exp(\sum_{i=1}^{n} \lambda_i f_i(x_{a,d,j,\hat{e}}))}{\sum_{y \in \{\hat{t},\hat{e}\}} \exp(\sum_{i=1}^{n} \lambda_i f_i(x_{a,d,j,y}))} \times \frac{n^{AE}_{a,e\ \neg k} + \alpha^E}{n^{AE}_{a,(\cdot)\ \neg k} + E\alpha^E} \times \frac{n^{EV}_{e,v\ \neg k} + \beta^E}{n^{EV}_{e,(\cdot)\ \neg k} + V\beta^E} \quad (2)$$

where $k = (a, d, j)$ denotes the j^{th} term of document d by author a and the subscript $\neg k$ denotes assignments excluding the term at (a, d, j). Omission of the latter index denoted by $(\cdot)$ represents the marginalized sum over the latter index. Count variables are detailed in Table 1 (last two rows). $\lambda_{1\ldots n}$ are the parameters of the learned Max-Ent model corresponding to the n binary feature functions $f_{1\ldots n}$ for Max-Ent. The learned Max-Ent λ parameters in conjunction with the observed context, $x_{a,d,j}$ feed the supervision signal for updating the topic/expression switch parameter, r in equations (1) and (2).

The hyper-parameters for the model were set to the values $\beta^T = \beta^E = 0.1$ and $\alpha_T = 50/T$, $\alpha_E = 50/E$, suggested in (Griffiths and Steyvers, 2004). Model parameters were estimated after 5000 Gibbs iterations with a burn-in of 1000 iterations. The Max-Ent parameters λ were learned using 500 labeled terms in each domain (politics:- topical: 376 & AD-expression: 124; religion:- topical: 349 & AD-expression: 151) appearing at least 10 times in debate threads other than the data in Table 1 (we do so since the data in Table 1(c) is later used in the classification experiments in §6.1).

Table 3 lists some top AD-expressions discovered by DTM. We see that DTM can cluster many correct AD-expressions, e.g., "I disagree", "I refute", "don't accept", etc. in disagreement; and "I agree", "you're correct", "agree with you", etc. in agreement. Further, it also discovers highly specific and more distinctive expressions beyond those used in Max-Ent training (marked *blue* in italics), e.g., "I don't buy your", "can you prove," "you fail to", and "you have no clue" in disagreement; and phrases like "valid point", "rightly said", "I do support", and "very well put" in agreement. In §6.1, we will see that these AD-expressions serve as high quality features for predicting tolerance.

Lastly, we note that DTM also estimates several pieces of useful information (e.g., AD-expressions, posterior estimates of author's arguing nature, θ^E_a; latent topics and expressions, φ^T_t; φ^E_e, etc.). These will be used to produce a rich set of user behavioral features for characterizing tolerance in §5.3.

5 Feature Engineering

We now propose features which will be used for model building to classify tolerant and intolerant authors in Table 1(c). We use three sets of features.

5.1 Language based Features of Tolerance

Word and POS n-grams: As tolerance in communication is directly reflected in language usage, word *n*-grams are obvious features. We also use POS tags (obtained using Stanford Tagger[4]) as features. The rationale of using POS tag based features is that intolerant communications are often characterized by hate/egotistic speech which have pronounced use of specific part of speech (e.g., pronouns) (Zingo, 1998).

Heuristic Factor Analysis: In psycholinguistics, factor analysis refers to the process of finding groups of semantically similar linguistic constructs (words/phrases). It is also called meaning extraction in (Chung and Pennebaker, 2007). As toler-

[4] http://nlp.stanford.edu/software/tagger.shtml

Factor: Reasoning words/phrases
because, because of, since, reason, reason being, reason is, reason why, due to, owing to, as in, therefore, thus, henceforth, hence, implies, implies that, implying, hints, hinting, hints towards, it follows that, it turns out, conclude, consequence, consequently, the cause, rationale, the rationale, justification, the justification, provided, premise, assumption, on the proviso, in spite, …
Factor: Sourcing words/phrases
presence of hyperlinks/urls, source, reference, for example, for instance, namely, to explain, to detail, to clarify, to elucidate, to illustrate, to be precise, furthermore, moreover, apart from, besides, we find, …

Table 4: Heuristic Factor Analysis (HFA). Words/Phrases in each factor compiled from prior works in psycholinguistics.

ance in discussions is characterized by reasoned expressions which often accompany sourcing (e.g., providing a hyperlink, making an attempt to clarify with some evidence, etc.), we compiled a list of reasoned and sourced expressions (shown in Table 4) from prior works (Chung and Pennebaker, 2007; Flor and Hadar, 2005; Moxey and Sanford, 2000; Pennebaker, et al., 2007).

5.2 Debate Expression Features

AD-expressions: As we have seen in §4, DTM can discover specific agreement and disagreement expressions in debates. We use these expressions as another feature set. Estimated AD-expressions (Table 3) serve as a principled way of performing factor analysis in debates instead of heuristic factor analysis as in Table 4 used in prior works.

As the AD-expression types are modeled as Dirichlet distributions ($\varphi^E \sim Dir(\beta^E)$), due to the smoothing effect, each term in the vocabulary has some non-zero probability mass associated with the expression types. To ensure that the discovered expressions are representative AD-expressions, we only consider the terms in φ^E with $p(v|e) = \varphi^E_{e,v} > 0.001$ as probability masses lower than 0.001 are more due to the smoothing effect of Dirichlet distribution than true correlation.

5.3 User Behavioral Features

Here we propose several features of user interaction which reflect the socio-psychological state of tolerance while participating in discussions. We note that these features rely on the posterior estimates of latent variables θ^E, z, and r in DTM (§4) and are thus difficult to obtain without modeling.

Overall Arguing Nature: The posterior on θ^E_a (Table 2) for each author, a gives an estimate of a's overall arguing nature (agreeing or disagreeing). We use the probability mass assigned to each arguing nature type as a user behavioral feature. This gives us two features f_1, f_2 as follows:

$$f_1(a) = \theta^E_{a,Ag}\ ;\ \ f_2(a) = \theta^E_{a,DisAg} \quad (3)$$

Behavioral Response: As intolerant users are likely to attract more disagreement, it is naturally useful to estimate the response (agreeing vs. disagreeing) a user receives from other users. For computing behavioral response, we first use the posterior on z to compute the distribution of AD-expressions (i.e., the relative probability masses of agreeing and disagreeing expressions) in a document d by an author a as follows:

$$\theta^E_{a,d,Ag} = \frac{|\{j|z_{a,d,j}=Ag, 1\le j\le N_{a,d}\}|}{|\{j|r_{a,d,j}=\hat{e}, 1\le j\le N_{a,d}\}|};$$

$$\theta^E_{a,d,DisAg} = \frac{|\{j|z_{a,d,j}=DisAg, 1\le j\le N_{a,d}\}|}{|\{j|r_{a,d,j}=\hat{e}, 1\le j\le N_{a,d}\}|} \quad (4)$$

Now to get the overall behavioral response of an author, a we take the expected value of the agreeing and disagreeing responses that a received from other authors a' who replied to or quoted a's posts. The expectations below are taken over all posts d' by a' which reply/quote posts of a.

$$f_3(a) = E[\theta^E_{a'\,d',Ag}];\ f_4(a) = E\left[\theta^E_{a'\,d',DisAg}\right] \quad (5)$$

Equality of Speech: In communication literature (Dahlgren, 2005; Habermas, 1984), equality is theorized as an essential element of tolerance. Each participant must be able to participate on an equal footing with others without anybody dominating the discussion. In online debates, we can measure this phenomenon using the following feature:

$$f_5(a) = E\left[\left(\frac{\#\ of\ posts\ by\ a\ in\ thread\ l}{\#\ of\ posts\ in\ thread\ l}\right) E[\theta^E_{a,d,DisAg}]\right] \quad (6)$$

where the inner expectation is taken over all posts of a in thread l and the outer expectation is taken over all threads l in which a participated. The above definition computes the aggressive posting behavior of author a whereby he tires to dominate the thread by posting more than others. The aggressive posting behavior is weighted by author's disagreeing nature because a person usually exhibits a dominating nature when he pushes hard to establish his ideology (which is often in disagreement with others) (Moxey and Sanford, 2000).

Topic Shifts: An interesting phenomenon of human (social) psyche is that when people are unable

to logically argue their stances and feel they are losing the debate, they often try to belittle/deride others by pulling unrelated topics into discussion (Slavin and Kriegman, 1992). This is referred to as topic shifts. Topic shifts thus have a relation with tolerance in deliberation. Stromer-Galley (2005) reported that if the discussion is off topic, then tolerance or deliberation cannot meet its objective of deep consideration of an issue. Hence, the average topic shifts of an author, a across various posts in a thread can serve as a good feature for measuring tolerance. We use the posterior on per-document topic distribution, $\theta^T_{a,d,t} = \frac{|\{j|z_{a,d,j}=t, 1 \le j \le N_{a,d}\}|}{|\{j|r_{a,d,j}=\hat{t}, 1 \le j \le N_{a,d}\}|}$ to measure topic shifts using KL-Divergence as follows:

$$f_6 = E\left[\text{avg}_{d,d' \in thread\ l}\left(D_{KL}\left(\theta^T_{a,d}||\theta^T_{a,d'}\right)\right)\right] \quad (7)$$

We first compute author, a's average topic shifts in a thread, l which measures his topic shifts in l. But this only gives us his behavior in one thread. To capture his overall behavior, we take the expected value of this behavior over all threads in which a participated. We take average KL-Div (KL-divergence) over all pairs of posts by a in a given thread to account for the asymmetry of KL-Div.

Finally, we note that by no means do we claim that the mere presence and a large value of any of the above features imply that a user is intolerant or tolerant. They are indicators of the phenomenon of tolerance in discussions/debates. The actual prediction is done using the learned models in §6.1.

6 Experimental Evaluation

We now detail the experiments that investigate the strengths of features in §5. In particular, we first consider the task of classifying whether an author is tolerant or intolerant in discussions. Then, we analyze how disagreement affects tolerance.

6.1 Tolerant and Intolerant Classification

Here, we show that the features in §5 can help build accurate models for predicting tolerance. We employ a linear kernel[5] SVM (using the SVM^{Light} system (Joachims, 1999)) and report 5-fold cross validation (CV) results on the task of predicting the socio-psychological nature of users' communication: tolerant vs. intolerant in politics and religion domains (Table 1(c)). Note that for each fold of 5-fold CV, DTM was run on the full data of each domain (Table 1(a)) excluding the users (and their associated posts) in the test set of that fold for generating the features of the training instances (users). The learned DTM was then fitted (using the approach in (Hofmann, 1999)) to the test set users and their posts for generating the features of the test instances.

To investigate the effectiveness of the proposed framework, we incrementally add feature sets starting with the baseline features. Word unigrams and bigrams (inclusive of unigrams)[6] serve as our first baseline (B1a, B1b). Word + POS bigrams is our second baseline (B2). "Word" in B2 uses bigrams as B1b gives better results. B2 + Heuristic Factor Analysis (HFA) (Table 4) serve as our third baseline (B3). Table 5 shows the experiment results. We note the following:

1. Across both domains, adding POS bigrams slightly improves classification accuracy and F_1-score beyond standard word unigrams and bigrams. Feature selection using information gain (IG) does not help much.
2. Using heuristic factor analyses (HFA) of reasoned and sourced expressions (Table 4) brings about 1% and 2% improvement in accuracy in politics and religion domains respectively.
3. Debate expression features (DE) in §5.2 and user behavioral features (UB) in §5.3 produced from DTM progressively improve classification accuracies by 4% and 8% in politics domains and 5% and 6% in religion domains. The improvements are also statistically significant.

In summary, we can see that modeling made a major impact. It improved the accuracy by about 10% than traditional unigram and bigram baselines. This shows that the debate expressions and user behaviors computed using the DTM model can capture various dimensions of (in)tolerance not captured by n-grams.

6.2 How Disagreement affects Tolerance?

We now quantitatively study the effect of disagreement on tolerance. We recall from §1 that tolerance indicates constructive discussion and allows disagreement. Some level of disagreement is often times an integral component of deliberation and tolerance (Cappella et al., 2002).

Disagreements, however, can be either constructive or destructive. The distinction is that the for-

[5] Other kernels (rbf, poly, sigmoid, etc.) did not perform as well.

[6] Higher order *n*-grams did not result in better results.

Feature Setting	Politics				Religion			
	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1	Accuracy
B1a: Word unigrams	64.1	86.3	73.7	70.1	61.9	86.8	72.6	71.9
Word unigram + IG	64.5	86.2	73.9	70.2	62.7	86.9	72.9	71.9
B1b: Word bigrams	66.8	87.8	75.9	72.4	64.9	89.1	75.9	75.1
B2: W+POS bigrams	68.5	86.8	76.4	73.7	66.6	88.4	76.8	76.7
B3: B2 + HFA(Table 4)	69.2	90.5	78.1	75.2	66.4	90.6	76.8	77.5
B3 + DE (§5.2)	74.7	91.3	**82.4**†	**79.5**†	70.2	92.8	**80.8**†	**82.1**†
B3 + DE + UB (§5.3)	76.1	92.2	**83.1**‡	**83.2**‡	71.7	93.4	**82.1**‡	**83.3**‡

Table 5: Precision, Recall, F_1 score on the tolerant class, and Accuracy for different feature settings across 2 domains. DE: Debate expression features (AD-expressions, Table3, §5.2). UB: User behavioral features (§5.3). Improvements in F_1 and Accuracy using DTM features (beyond baselines, B1-B3) are statistically significant (†: p<0.02; ‡: p<0.01) using paired t-test with 5-fold CV.

Thread/Issue	# Posts	# Users	% InTol.	$E[\theta^E_{a,d,DisAg}]$	τ	*p-value*
Repeal Healthcare	1823	33	39.9	0.57	0.65	0.02
Europe's Collapse	1824	33	42.5	0.61	0.61	0.01
Obama Euphoria	1244	26	30.7	0.66	0.71	0.01
Socialism	831	49	44.8	0.69	0.48	0.03
Abortion	1232	58	48.4	0.78	0.37	0.01

Table 6: Tipping points of disagreements for intolerance (τ) of different issues. $E[\theta^E_{a,d,DisAg}]$: the expected disagreement over all posts in each issue/thread, # Posts: the total number of posts, # Users: the total number of users/authors, % Intol: % of intolerant users in each thread, τ: the estimated tipping point, and p-value: computed from two-tailed Fisher's exact test.

mer is aimed at arriving at a consensus or solution, while the latter leads to polarization and intolerance (Sunstein, 2002). It was also shown in (Dahlgren, 2005) that sustained disagreement often takes a transition towards destructive disagreement and is likely to lead to intolerance. Similar phenomena was also identified in psychology literature (Critchley, 1964). In such cases, the participants often stubbornly stick to an extreme attitude, which eventually results in intolerance and defeats the very purpose of deliberative discussion.

An intriguing research question is: What is the relationship between disagreement and intolerance? The question is interesting from both the communication and psycholinguistic perspectives. The best of our knowledge, this is the first attempt towards seeking an answer. We work in the context of five issues/threads in real-life online debates. To derive quantitative and definite conclusions, it is required to perform the following tasks:

- For each issue, empirically investigate *in expectation* the *tipping point* of disagreement beyond which a user tends to be intolerant.
- Further, investigate the *confidence* on the estimated tipping point (i.e., what is the likelihood that the estimated tipping point is statistically significant instead of chance alone).

We formalize the above tasks in the Bayesian setting. Recall from Table 2 of §4, that $\theta^E_{a,Ag}$ (respectively, $\theta^E_{a,DisAg}$) are the estimates of agreeing and disagreeing nature of an author and $\theta^E_{a,Ag} + \theta^E_{a,DisAg} = 1$. Let $TP(\tau)$ denote the event that *in expectation* a threshold value of $0 < \tau < 1$ serves as a tipping point of disagreement beyond which intolerance is exhibited. Note that we emphasize the term "in expectation" (taken over all authors). We do not mean that every author whose disagreement, $\theta^E_{a,DisAg} > \tau$, is intolerant. The empirical likelihood of $TP(\tau)$ can be expressed by the following probability expression:

$$\mathcal{L}(TP(\tau)) = E\left[P\left(\theta^E_{a,DisAg} > \tau | a = I\right) - P\left(\theta^E_{a,DisAg} > \tau | a = T\right)\right] \quad (8)$$

The events $a = I$ and $a = T$ denote that author a is intolerant and tolerant respectively. The expectation is taken over authors. Showing that τ indeed serves as the tipping point of disagreement to exhibit intolerance corresponds is to rejecting the null hypothesis that the probabilities in (8) are equal. We employ a Fisher's exact test to test significance and report confidence measures (using p-values) for the tipping point thresholds. The results are shown in Table 6.

The threshold τ is computed using the entropy method in (Fayyad and Irani, 1993) as follows: We first fit our previously learned model (using the data in Table 1 (a)) to the new threads in Table 6 and its users and posts to obtain the estimates of

$\theta^E_{a,DisAg}$ and other latent variables for feature generation. The learned classifier in §6.1 is used to predict the nature of users (tolerant vs. intolerant) in the new threads[7]. Then, for each user we have his predicted deliberative (social) psyche (Tolerant vs. Intolerant) and also his overall disagreeing nature exhibited in that thread (the posterior on $\theta^E_{a,DisAg} \in [0, 1]$). For a thread, tolerant and intolerant users (data points) span the range [0, 1] attaining different values for $\theta^E_{a,DisAg}$. Each candidate tipping point of disagreement, $0 \leq \tau' \leq 1$ results in a binary partition of the range with each partition containing some proportion of tolerant and intolerant users. We compute the entropy of the partition for every candidate tipping point in the range [0, 1]. The final tipping point threshold, τ is chosen such that it minimizes the partition entropy based on the binary cut-point method in (Fayyad and Irani, 1993).

Since we perform a thread level analysis, the results in Table 6 are thread/issue specific. We note the following from Table 6:

1. Across all threads/issues, we find that the expected disagreement over all posts, d, $E[\theta^E_{a,d,DisAg}] > 0.5$ showing that in discussions of the reported issues, disagreement predominates.
2. $E[\theta^E_{a,d,DisAg}]$ also gives an estimate of overall *heat* in the issue being discussed. We find sensitive issues like *abortion* and *socialism* being more heated than *healthcare*, *Obama*, etc.
3. The percentage of intolerant users increases with the expected overall disagreement in the issue except for the issue *Obama euphoria*.
4. The estimated tipping point of disagreement to exhibit intolerance, τ happens to vary inversely with the expected disagreement, $E[\theta^E_{a,d,DisAg}]$ except the issue *Obama euphoria*. This reflects that as overall disagreement in the issue increases, the tipping point of intolerance decreases, i.e., due to high discussion heat, people are likely to turn intolerant even with relatively small amount of disagreement. This finding dovetails with prior studies in psychology (Rokeach and Fruchter, 1956) that heated discussions are likely to reduce thresholds of reception leading to dogmatism, egotism, and intolerance. Table 6 shows that for moderately heated issues (*healthcare*, *Europe's collapse*), in expectation, author's disagreement $\theta^E_{a,DisAg}$ should exceed 61-65% to exhibit intolerance. However, for sensitive issues, we find that the tipping point is much lower, *abortion*: 37%; *socialism*: 48%.
5. The issue *Obama Euphoria* is an exception to other issues' trends. Even though in expectation, it has $E[\theta^E_{a,d,DisAg}]$ = 66% overall disagreement, the percentage of intolerant users remains the lowest (30%) and the tipping point attains a highest value (τ = 0.71), showing more tolerance on the issue. A plausible reason could be that Obama is somewhat more liked and hence attracts less intolerance from users[8].
6. The *p*-values of the estimated tipping points, τ across all issues are statistically significant at 98-99% confidence levels.

7 Conclusion

This work performed a deep analysis of the socio-opsychological and psycholinguistic phenomenon of tolerance in online discussions, which is an important concept in the field of communications. A novel framework is proposed, which is capable of characterizing and classifying tolerance in online discussions. Further, a novel technique was also proposed to quantitatively evaluate the interplay of tolerance and disagreement. Our empirical results using real-life online discussions render key insights into the psycholinguistic process of tolerance and dovetail with existing theories in psychology and communications. To the best of our knowledge, this is the first such quantitative study. In our future work, we want to further this research and study the role of diversity of opinions in the context of tolerance and its relation to polarization.

Acknowledgments

This work was supported in part by a grant from National Science Foundation (NSF) under grant no. IIS-1111092.

[7] Although this prediction may not be perfect, it can be regarded as considerably reliable to study the trend of tolerance across different issues as our classifier (in §6.1) attains a high (83%) classification accuracy using the full feature set. As judging all users across all threads would require reading about 7000 posts, for confirmation, we randomly sampled 30 authors across various threads for labeling by our judges. 28 out of 30 predictions produced by the classifier correlated with the judges' labels, which should be sufficiently accurate for our analysis.

[8] This observation may be linked to the political phenomenon of "democratic citizenship through exposure to diverse perspectives" (Mutz, 2006) where it was shown that exposure to heterogeneous opinions (i.e., greater disagreement), often enhances tolerance.

References

Abu-Jbara, A., Dasigi, P., Diab, M. and Dragomir Radev. 2012. Subgroup detection in ideological discussions. *ACL*.

Agrawal, R. Rajagopalan, S. Srikant, R. Xu. Y. 2003. Mining newsgroups using networks arising from social behavior. *WWW*.

Bansal, M., Cardie, C., and Lee, L. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. *In COLING*.

Blei, D., A. Ng, and M. Jordan. 2003. Latent Dirichlet Allocation. *In JMLR*.

Boyer, K.; Grafsgaard, J.; Ha, E. Y.; Phillips, R.; and Lester, J. 2011. An affect-enriched dialogue act classification model for task-oriented dialogue. *In ACL*.

Burfoot, C., S. Bird, and T. Baldwin. 2011. Collective Classification of Congressional Floor-Debate Transcripts. *In ACL*.

Cappella, J. N., Price, V., & Nir, L. 2002. Argument repertoire as a reliable and valid measure of opinion quality: electronic dialogue during campaign 2000. Political Communication. *Political Communication*.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. 2013. Leveraging Multi-Domain Prior Knowledge in Topic Models. *In IJCAI*.

Chung, C. K., & Pennebaker, J. W. 2007. Revealing people's thinking in natural language: Using an automated meaning extraction method in open–ended self–descriptions,. *J. of Research in Personality*.

Choi, Y. and Cardie, C. 2010. Hierarchical sequential learning for extracting opinions and their attributes. *In ACL*.

Critchley, M. 1964. The neurology of psychotic speech. *The British Journal of Psychiatry*.

Crocker, D. A. 2005. Tolerance and Deliberative Democracy. *UMD Technical Report*.

Dahlgren, P. 2002. In search of the talkative public: Media, deliberative democracy and civic culture. *Javnost/The Public*.

Dahlgren, Peter. 2005. The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation. *Political Communication*.

Escobar, O. 2012. Public Dialogue and Deliberation: A communication perspective for publicengagement practitioners. *Handbook and Technical Report*.

Fayyad, U., & Irani, K. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. *In UAI*.

Fishkin, J. 1991. Democracy and deliberation. *New Haven, CT: Yale University Press*.

Flor, M., & Hadar, U. 2005. The production of metaphoric expressions in spontaneous speech: A controlled-setting experiment. *Metaphor and Symbol*.

Galley, M., K. McKeown, J. Hirschberg, E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. *In ACL*.

Gastil, J. 2005. Communication as Deliberation: A Non-Deliberative Polemic on Communication Theory. *Univ. of Washington, Technical Report*.

Gastil, J., & Dillard, J. P. 1999. Increasing political sophistication through public deliberation. *Political Communication*.

Gastil, John. 2007. Political communication and deliberation. *Sage Publications*.

Griffiths, T. and Steyvers, M. 2004. Finding scientific topics. *In PNAS*.

Gutmann, A., & Thompson, D. F. 1996. *Democracy and disagreement.* Harvard University Press.

Habermas. 1984. The theory of communicative action: Reason and rationalization of society. *(T. McCarthy, Trans. Vol. 1). Boston, MA: Beacon Press*.

Hillard, D., Ostendorf, M., and Shriberg, E. 2003. Detection of Agreement vs. Disagreement in Meetings: Training with Unlabeled Data. *HLT-NAACL*.

Hansen, G. J., & Hyunjung, K. 2011. Is the media biased against me? A meta-analysis of the hostile media effect research. *Communication Research Reports, 28, 169-179*.

Hassan, A. and Radev, D. 2010. Identifying text polarity using random walks.*In ACL*.

Hofmann, T. 1999. Probabilistic latent semantic analysis. *In UAI*.

Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. *In SIGKDD*.

Joachims, T. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

Kim, S. and Hovy, E. 2007. Crystal: Analyzing predictive opinions on the web. *In EMNLP-CoNLL*.

Landis, J. R. and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lin, W. H., and Hauptmann, A. 2006. Are these documents written from different perspectives?: a test of different perspectives based on statistical distribution divergence. *In ACL*.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publisher, USA.

Luskin, R. C., Fishkin, J. S., & Iyengar, S. 2004. Considered Opinions on U.S. Foreign Policy: Face-to-Face versus Online Deliberative Polling. *International Communication Association, New Orleans, LA*.

Mayfield, E. and Rose, C. P. 2011. Recognizing Authority in Dialogue with an Integer Linear Programming Constrained Model. *In ACL*.

Moxey, L. M., & Sanford, A. J. 2000. Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology*.

Morbini, F. and Sagae, K. 2011. Joint Identification and Segmentation of Domain-Specific Dialogue Acts for Conversational Dialogue Systems. *In ACL*.

Murakami, A., and Raymond, R. 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. In *COLING*.

Mukherjee, A. and Liu, B. 2013. Discovering User Interactions in Ideological Discussions. *In ACL*.

Mukherjee, A. and Liu, B. 2012a. Mining Contentions from Discussions and Debates. *In KDD*.

Mukherjee, A. and Liu, B. 2012b. Modeling review Comments. *In ACL*.

Mukherjee, A. and Liu, B. 2012c. Aspect Extraction through Semi-Supervised Modeling. *In ACL*.

Mukherjee, A. and Liu, B. 2012d. Analysis of Linguistic Style Accommodation in Online Debates. *In COLING*.

Mutz, D. 2006. Hearing the Other Side: Deliberative Versus Participatory Democracy. Cambridge: Cambridge University Press, 2006.

Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. 2007. The development and psychometric properties of LIWC2007. *LIWC.Net.*

Popescu, A. and Etzioni, O. 2005. Extracting product features and opinions from reviews. *In EMNLP*.

Price, V., Cappella, J. N., & Nir, L. 2002. Does disagreement contribute to more deliberative opinion? *Political Communication*.

Rokeach, M., & Fruchter, B. 1956. A factorial study of dogmatism and related concepts. *The Journal of Abnormal and Social Psychology*.

Ryfe, D. M. (2005). Does deliberative democracy work? *Annual review of political science*.

Slavin, M. O., & Kriegman, D. 1992. *The adaptive design of the human psyche: Psychoanalysis, evolutionary biology, and the therapeutic process.* Guilford Press.

Somasundaran, S., J. Wiebe. 2009. Recognizing stances in online debates. *In ACL-IJCNLP*.

Stromer-Galley, J. 2005. Conceptualizing and Measuring Coherence in Online Chat. *Annual Meeting of the International Communication Association*.

Sunstein, C. R. 2002. The law of group polarization. *Journal of political philosophy*.

Thomas, M., B. Pang and L. Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *In EMNLP*.

Wang, L., Lui, M., Kim, S. N., Nivre, J., and Baldwin, T. 2011. Predicting thread discourse structure over technical web forums. *In EMNLP*.

Wiebe, J. 2000. Learning subjective adjectives from corpora. *In Proc. of National Conference on AI*.

Yessenalina, A., Yue, A., Cardie, C. 2010. Multi-level structured models for document-level sentiment classification. *In EMNLP*.

Zhao, X., J. Jiang, H. Yan, and X. Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. *In EMNLP*.

Zingo, M. T. (1998). Sex/gender Outsiders, Hate Speech, and Freedom of Expression: Can They Say that about Me? *Praeger Publishers*.