Self-disclosure topic model for Twitter conversations

JinYeong Bak
Department of Computer Science
KAIST
Daejeon, South Korea
jy.bak@kaist.ac.kr

Chin-Yew Lin
Microsoft Research Asia
Beijing 100080, P.R. China
cyl@microsoft.com

Alice Oh
Department of Computer Science
KAIST
Daejeon, South Korea
alice.oh@kaist.edu

Abstract

Self-disclosure, the act of revealing oneself to others, is an important social behavior that contributes positively to intimacy and social support from others. It is a natural behavior, and social scientists have carried out numerous quantitative analyses of it through manual tagging and survey questionnaires. Recently, the flood of data from online social networks (OSN) offers a practical way to observe and analyze self-disclosure behavior at an unprecedented scale. The challenge with such analysis is that OSN data come with no annotations, and it would be impossible to manually annotate the data for a quantitative analysis of self-disclosure. As a solution, we propose a semi-supervised machine learning approach, using a variant of latent Dirichlet allocation for automatically classifying self-disclosure in a massive dataset of Twitter conversations. For measuring the accuracy of our model, we manually annotate a small subset of our dataset, and we show that our model shows significantly higher accuracy and F-measure than various other methods. With the results our model, we uncover a positive and significant relationship between self-disclosure and online conversation frequency over time.

1 Introduction

Self-disclosure is an important and pervasive social behavior. People disclose personal information about themselves to improve and maintain relationships (Jourard, 1971; Joinson and Paine, 2007). For example, when two people meet for the first time, they disclose their names and interests. One positive outcome of self-disclosure is social support from others (Wills, 1985; Derlega et al., 1993), shown also in online social networks (OSN) such as Twitter (Kim et al., 2012). Receiving social support would then lead the user to be more active on OSN (Steinfield et al., 2008; Trepte and Reinecke, 2013). In this paper, we seek to understand this important social behavior using a large-scale Twitter conversation data, automatically classifying the level of self-disclosure using machine learning and correlating the patterns with subsequent OSN usage.

Twitter conversation data, explained in more detail in section 4.1, enable a significantly larger scale study of naturally-occurring self-disclosure behavior, compared to traditional social science studies. One challenge of such large scale study, though, remains in the lack of labeled ground-truth data of self-disclosure level. That is, naturally-occurring Twitter conversations do not come tagged with the level of self-disclosure in each conversation. To overcome that challenge, we propose a semi-supervised machine learning approach using probabilistic topic modeling. Our self-disclosure topic model (SDTM) assumes that self-disclosure behavior can be modeled using a combination of simple linguistic features (e.g., pronouns) with automatically discovered semantic themes (i.e., topics). For instance, an utterance "I am finally through with this disastrous relationship" uses a first-person pronoun and contains a topic about personal relationships.

In comparison with various other models, SDTM shows the highest accuracy, and the resulting self-disclosure patterns of the users are correlated significantly with their future OSN usage. Our contributions to the research community include the following:

- We present a topic model that explicitly includes the level of self-disclosure in a conversation using linguistic features and the latent semantic topics (Sec. 3).

Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pages 42–49,
Baltimore, Maryland USA, 27 June 2014.

- We collect a large dataset of Twitter conversations over three years and annotate a small subset with self-disclosure level (Sec. 4).
- We compare the classification accuracy of SDTM with other models and show that it performs the best (Sec. 5).
- We correlate the self-disclosure patterns of users and their subsequent OSN usage to show that there is a positive and significant relationship (Sec. 6).

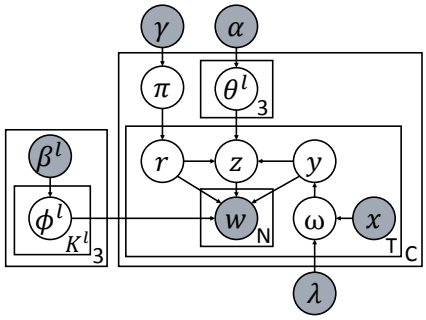

Figure 1: Graphical model of SDTM

2 Background

In this section, we review literature on the relevant aspects of self-disclosure.

Self-disclosure (*SD*) level: To quantitatively analyze self-disclosure, researchers categorize self-disclosure language into three levels: `G` (general) for no disclosure, `M` for medium disclosure, and `H` for high disclosure (Vondracek and Vondracek, 1971; Barak and Gluck-Ofri, 2007). Utterances that contain general (non-sensitive) information about the self or someone close (e.g., a family member) are categorized as `M`. Examples are personal events, past history, or future plans. Utterances about age, occupation and hobbies are also included. Utterances that contain sensitive information about the self or someone close are categorized as `H`. Sensitive information includes personal characteristics, problematic behaviors, physical appearance and wishful ideas. Generally, these are thoughts and information that one would generally keep as secrets to himself. All other utterances, those that do not contain information about the self or someone close are categorized as `G`. Examples include gossip about celebrities or factual discourse about current events.

Classifying self-disclosure level: Prior work on quantitatively analyzing self-disclosure has relied on user surveys (Trepte and Reinecke, 2013; Ledbetter et al., 2011) or human annotation (Barak and Gluck-Ofri, 2007). These methods consume much time and effort, so they are not suitable for large-scale studies. In prior work closest to ours, Bak et al. (2012) showed that a topic model can be used to identify self-disclosure, but that work applies a two-step process in which a basic topic model is first applied to find the topics, and then the topics are post-processed for binary classification of self-disclosure. We improve upon this work by applying a single unified model of topics and self-disclosure for high accuracy in classifying the three levels of self-disclosure.

Self-disclosure and online social network: According to social psychology, when someone discloses about himself, he will receive social support from those around him (Wills, 1985; Derlega et al., 1993), and this pattern of self-disclosure and social support was verified for Twitter conversation data (Kim et al., 2012). Social support is a major motivation for active usage of social networks services (SNS), and there are findings that show self-disclosure on SNS has a positive longitudinal effect on future SNS use (Trepte and Reinecke, 2013; Ledbetter et al., 2011). While these previous studies focused on small, qualitative studies, we conduct a large-scale, machine learning driven study to approach the question of self-disclosure behavior and SNS use.

3 Self-Disclosure Topic Model

This section describes our model, the self-disclosure topic model (SDTM), for classifying self-disclosure level and discovering topics for each self-disclosure level.

3.1 Model

We make two important assumptions based on our observations of the data. First, first-person pronouns (*I, my, me*) are good indicators for medium level of self-disclosure. For example, phrases such as 'I live' or 'My age is' occur in utterances that reveal personal information. Second, there are topics that occur much more frequently at a particular SD level. For instance, topics such as *physical appearance* and *mental health* occur frequently at level `H`, whereas topics such as *birthday* and *hobbies* occur frequently at level `M`.

Figure 1 illustrates the graphical model of SDTM and how these assumptions are embodied

Notation	Description
G; M; H	{general; medium; high} SD level
C; T; N	Number of conversations; tweets; words
K^G; K^M; K^H	Number of topics for {G; M; H}
c; ct	Conversation; tweet in conversation c
y_{ct}	SD level of tweet ct, G or M/H
r_{ct}	SD level of tweet ct, M or H
z_{ct}	Topic of tweet ct
w_{ctn}	n^{th} word in tweet ct
$\boldsymbol{\lambda}$	Learned Maximum entropy parameters
$\boldsymbol{x}_{ct}$	First-person pronouns features
$\boldsymbol{\omega}_{ct}$	Distribution over SD level of tweet ct
$\boldsymbol{\pi}_c$	SD level proportion of conversation c
$\boldsymbol{\theta}_c^G$; $\boldsymbol{\theta}_c^M$; $\boldsymbol{\theta}_c^H$	Topic proportion of {G; M; H} in conversation c
$\boldsymbol{\phi}^G$; $\boldsymbol{\phi}^M$; $\boldsymbol{\phi}^H$	Word distribution of {G; M; H}
$\boldsymbol{\alpha}$; $\boldsymbol{\gamma}$	Dirichlet prior for θ; π
$\boldsymbol{\beta}^G$, $\boldsymbol{\beta}^M$; $\boldsymbol{\beta}^H$	Dirichlet prior for ϕ^G; ϕ^M; ϕ^H
n_{cl}	Number of tweets assigned SD level l in conversation c
n_{ck}^l	Number of tweets assigned SD level l and topic k in conversation c
n_{kv}^l	Number of instances of word v assigned SD level l and topic k
m_{ctkv}	Number of instances of word v assigned topic k in tweet ct

Table 1: Summary of notations used in SDTM.

in it. The first assumption about the first-person pronouns is implemented by the observed variable $\boldsymbol{x}_{ct}$ and the parameters $\boldsymbol{\lambda}$ from a maximum entropy classifier for G vs. M/H level. The second assumption is implemented by the three separate word-topic probability vectors for the three levels of SD: $\boldsymbol{\phi}^l$ which has a Bayesian informative prior $\boldsymbol{\beta}^l$ where $l \in \{G, M, H\}$, the three levels of self-disclosure. Table 1 lists the notations used in the model and the generative process, Figure 2 describes the generative process.

3.2 Classifying G vs M/H levels

Classifying the SD level for each tweet is done in two parts, and the first part classifies G vs. M/H levels with first-person pronouns (*I, my, me*). In the graphical model, y is the latent variable that represents this classification, and $\boldsymbol{\omega}$ is the distribution over y. x is the observation of the first-person pronoun in the tweets, and $\boldsymbol{\lambda}$ are the parameters learned from the maximum entropy classifier. With the annotated Twitter conversation dataset (described in Section 4.2), we experimented with several classifiers (Decision tree, Naive Bayes) and chose the maximum entropy classifier because it performed the best, similar to other joint topic models (Zhao et al., 2010; Mukherjee et al., 2013).

1. For each level $l \in \{\texttt{G}, \texttt{M}, \texttt{H}\}$:
   For each topic $k \in \{1, \ldots, K^l\}$:
   Draw $\boldsymbol{\phi}_k^l \sim Dir(\boldsymbol{\beta}^l)$
2. For each conversation $c \in \{1, \ldots, C\}$:
   (a) Draw $\boldsymbol{\theta}_c^G \sim Dir(\boldsymbol{\alpha})$
   (b) Draw $\boldsymbol{\theta}_c^M \sim Dir(\boldsymbol{\alpha})$
   (c) Draw $\boldsymbol{\theta}_c^H \sim Dir(\boldsymbol{\alpha})$
   (d) Draw $\boldsymbol{\pi}_c \sim Dir(\boldsymbol{\gamma})$
   (e) For each message $t \in \{1, \ldots, T\}$:
      i. Observe first-person pronouns features $\boldsymbol{x}_{ct}$
      ii. Draw $\boldsymbol{\omega}_{ct} \sim MaxEnt(\boldsymbol{x}_{ct}, \boldsymbol{\lambda})$
      iii. Draw $y_{ct} \sim Bernoulli(\boldsymbol{\omega}_{ct})$
      iv. If $y_{ct} = 0$ which is G level:
         A. Draw $z_{ct} \sim Mult(\boldsymbol{\theta}_c^G)$
         B. For each word $n \in \{1, \ldots, N\}$:
            Draw word $w_{ctn} \sim Mult(\boldsymbol{\phi}_{z_{ct}}^G)$
      Else which can be M or H level:
         A. Draw $r_{ct} \sim Mult(\boldsymbol{\pi}_c)$
         B. Draw $z_{ct} \sim Mult(\boldsymbol{\theta}_c^{r_{ct}})$
         C. For each word $n \in \{1, \ldots, N\}$:
            Draw word $w_{ctn} \sim Mult(\boldsymbol{\phi}_{z_{ct}}^{r_{ct}})$

Figure 2: Generative process of SDTM.

3.3 Classifying M vs H levels

The second part of the classification, the M and the H level, is driven by informative priors with seed words and seed trigrams.

Utterances with M level include two types: 1) information related with past events and future plans, and 2) general information about self (Barak and Gluck-Ofri, 2007). For the former, we add as seed trigrams 'I have been' and 'I will'. For the latter, we use seven types of information generally accepted to be personally identifiable information (McCallister, 2010), as listed in the left column of Table 2. To find the appropriate trigrams for those, we take Twitter conversation data (described in Section 4.1) and look for trigrams that begin with 'I' and 'my' and occur more than 200 times. We then check each one to see whether it is related with any of the seven types listed in the table. As a result, we find 57 seed trigrams for M level. Table 2 shows several examples.

Type	Trigram
Name	My name is, My last name
Birthday	My birthday is, My birthday party
Location	I live in, I lived in, I live on
Contact	My email address, My phone number
Occupation	My job is, My new job
Education	My high school, My college is
Family	My dad is, My mom is, My family is

Table 2: Example seed trigrams for identifying M level of SD. There are 51 of these used in SDTM.

Utterances with H level express secretive wishes or sensitive information that exposes self or someone close (Barak and Gluck-Ofri, 2007). These are

Category	Keywords
physical appearance	acne, hair, overweight, stomach, chest, hand, scar, thighs, chubby, head, skinny
mental/physical condition	addicted, bulimia, doctor, illness, alcoholic, disease, drugs, pills, anorexic

Table 3: Example words for identifying H level of SD. Categories are hand-labeled.

generally keep as secrests. With this intuition, we crawled 26,523 secret posts from *Six Billion Secrets* [1] site where users post secrets anonymously.

To extract seed words that might express secretive personal information, we compute mutual information (Manning et al., 2008) with the secret posts and 24,610 randomly selected tweets. We select 1,000 words with high mutual information and filter out stop words. Table 3 shows some of these words. To extract seed trigrams of secretive wishes, we again look for trigrams that start with 'I' or 'my', occur more than 200 times, and select trigrams of wishful thinking, such as 'I want to', and 'I wish I'. In total, there are 88 seed words and 8 seed trigrams for H.

3.4 Inference

For posterior inference of SDTM, we use collapsed Gibbs sampling which integrates out latent random variables $\boldsymbol{\omega}, \boldsymbol{\pi}, \boldsymbol{\theta}$, and $\boldsymbol{\phi}$. Then we only need to compute $\boldsymbol{y}, \boldsymbol{r}$ and $\boldsymbol{z}$ for each tweet. We compute full conditional distribution $p(y_{ct} = j', r_{ct} = l', z_{ct} = k' | \boldsymbol{y}_{-ct}, \boldsymbol{r}_{-ct}, \boldsymbol{z}_{-ct}, \boldsymbol{w}, \boldsymbol{x})$ for tweet ct as follows:

$$p(y_{ct} = 0, z_{ct} = k' | \boldsymbol{y}_{-ct}, \boldsymbol{r}_{-ct}, \boldsymbol{z}_{-ct}, \boldsymbol{w}, \boldsymbol{x}) \propto \frac{exp(\boldsymbol{\lambda}_0 \cdot \boldsymbol{x}_{ct})}{\sum_{j=0}^{1} exp(\boldsymbol{\lambda}_j \cdot \boldsymbol{x}_{ct})} \, g(c, t, l', k')$$

$$p(y_{ct} = 1, r_{ct} = l', z_{ct} = k' | \boldsymbol{y}_{-ct}, \boldsymbol{r}_{-ct}, \boldsymbol{z}_{-ct}, \boldsymbol{w}, \boldsymbol{x}) \propto \frac{exp(\boldsymbol{\lambda}_1 \cdot \boldsymbol{x}_{ct})}{\sum_{j=0}^{1} exp(\boldsymbol{\lambda}_j \cdot \boldsymbol{x}_{ct})} (\gamma_{l'} + n_{cl'}^{(-ct)}) \, g(c, t, l', k')$$

where $\boldsymbol{z}_{-ct}, \boldsymbol{r}_{-ct}, \boldsymbol{y}_{-ct}$ are $\boldsymbol{z}, \boldsymbol{r}, \boldsymbol{y}$ without tweet ct, $m_{ctk'(\cdot)}$ is the marginalized sum over word v of $m_{ctk'v}$ and the function $g(c, t, l', k')$ as follows:

$$g(c, t, l', k') = \frac{\Gamma(\sum_{v=1}^{V} \beta_v^{l'} + n_{k'v}^{l'-(ct)})}{\Gamma(\sum_{v=1}^{V} \beta_v^{l'} + n_{k'v}^{l'-(ct)} + m_{ctk'(\cdot)})} \left(\frac{\alpha_{k'} + n_{ck'}^{l'(-ct)}}{\sum_{k=1}^{K} \alpha_k + n_{ck}^{l'}} \right) \prod_{v=1}^{V} \frac{\Gamma(\beta_v^{l'} + n_{k'v}^{l'-(ct)} + m_{ctk'v})}{\Gamma(\beta_v^{l'} + n_{k'v}^{l'-(ct)})}$$

[1] http://www.sixbillionsecrets.com

4 Data Collection and Annotation

To answer our research questions, we need a large longitudinal dataset of conversations such that we can analyze the relationship between self-disclosure behavior and conversation frequency over time. We chose to crawl Twitter because it offers a practical and large source of conversations (Ritter et al., 2010). Others have also analyzed Twitter conversations for natural language and social media research (Boyd et al., 2010; Danescu-Niculescu-Mizil et al., 2011), but we collect conversations from the same set of dyads over several months for a unique longitudinal dataset.

4.1 Collecting Twitter conversations

We define a Twitter conversation as a chain of tweets where two users are consecutively replying to each other's tweets using the Twitter reply button. We identify dyads of English-tweeting users with at least twenty conversations and collect their tweets. We use an open source tool for detecting English tweets [2], and to protect users' privacy, we replace Twitter userid, usernames and url in tweets with random strings. This dataset consists of 101,686 users, 61,451 dyads, 1,956,993 conversations and 17,178,638 tweets which were posted between August 2007 to July 2013.

4.2 Annotating self-disclosure level

To measure the accuracy of our model, we randomly sample 101 conversations, each with ten or fewer tweets, and ask three judges, fluent in English, to annotate each tweet with the level of self-disclosure. Judges first read and discussed the definitions and examples of self-disclosure level shown in (Barak and Gluck-Ofri, 2007), then they worked separately on a Web-based platform. Inter-rater agreement using Fleiss kappa (Fleiss, 1971) is 0.67.

5 Classification of Self-Disclosure Level

This section describes experiments and results of SDTM as well as several other methods for classification of self-disclosure level.

We first start with the annotated dataset in section 4.2 in which each tweet is annotated with SD level. We then aggregate all of the tweets of a conversation, and we compute the proportions of tweets in each SD level. When the proportion of

[2] https://github.com/shuyo/ldig

tweets at M or H level is equal to or greater than 0.2, we take the level of the larger proportion and assign that level to the conversation. When the proportions of tweets at M or H level are both less than 0.2, we assign G to the SD level.

We compare SDTM with the following methods for classifying tweets for SD level:

- LDA (Blei et al., 2003): A Bayesian topic model. Each conversation is treated as a document. Used in previous work (Bak et al., 2012).
- MedLDA (Zhu et al., 2012): A supervised topic model for document classification. Each conversation is treated as a document and response variable can be mapped to a SD level.
- LIWC (Tausczik and Pennebaker, 2010): Word counts of particular categories. Used in previous work (Houghton and Joinson, 2012).
- Seed words and trigrams (SEED): Occurrence of seed words and trigrams which are described in section 3.3.
- ASUM (Jo and Oh, 2011): A joint model of sentiment and topic using seed words. Each sentiment can be mapped to a SD level. Used in previous work (Bak et al., 2012).
- First-person pronouns (FirstP): Occurrence of first-person pronouns which are described in section 3.2. To identify first-person pronouns, we tagged parts of speech in each tweet with the Twitter POS tagger (Owoputi et al., 2013).

SEED, LIWC, LDA and FirstP cannot be used directly for classification, so we use Maximum entropy model with outputs of each of those models as features. We run MedLDA, ASUM and SDTM 20 times each and compute the average accuracies and F-measure for each level. We set 40 topics for LDA, MedLDA and ASUM, 60; 40; 40 topics for SDTM K^G, K^M and K^H respectively, and set $\alpha = \gamma = 0.1$. To incorporate the seed words and trigrams into ASUM and SDTM, we initialize $\boldsymbol{\beta}^G, \boldsymbol{\beta}^M$ and $\boldsymbol{\beta}^H$ differently. We assign a high value of 2.0 for each seed word and trigram for that level, and a low value of 10^{-6} for each word that is a seed word for another level, and a default value of 0.01 for all other words. This approach is same as other topic model works (Jo and Oh, 2011; Kim et al., 2013).

Method	*Acc*	G F_1	M F_1	H F_1	*Avg* F_1
LDA	49.2	0.000	0.650	0.050	0.233
MedLDA	43.3	0.406	0.516	0.093	0.338
LIWC	49.2	0.341	0.607	0.180	0.376
SEED	52.0	0.412	0.600	0.178	0.397
ASUM	56.6	0.320	0.704	0.375	0.466
FirstP	63.2	**0.630**	0.689	0.095	0.472
SDTM	**64.5**	0.611	**0.706**	**0.431**	**0.583**

Table 4: SD level classification accuracies and F-measures using annotated data. *Acc* is accuracy, and G F_1 is F-measure for classifying the G level. *Avg* F_1 is the average value of G F_1, M F_1 and H F_1. SDTM outperforms all other methods compared. The difference between SDTM and FirstP is statistically significant (p-value < 0.05 for accuracy, < 0.0001 for *Avg* F_1).

As Table 4 shows, SDTM performs better than other methods by accuracy and F-measure. LDA and MedLDA generally show the lowest performance, which is not surprising given these models are quite general and not tuned specifically for this type of semi-supervised classification task. LIWC and SEED perform better than LDA, but these have quite low F-measure for G and H levels. ASUM shows better performance for classifying H level than others, but not for classifying the G level. FirstP shows good F-measure for the G level, but the H level F-measure is quite low, even lower than SEED. Finally, SDTM has similar performance in G and M level with FirstP, but it performs better in H level than others. Classifying the H level well is important because as we will discuss later, the H level has the strongest relationship with longitudinal OSN usage (see Section 6.2), so SDTM is overall the best model for classifying self-disclosure levels.

6 Self-Disclosure and Conversation Frequency

In this section, we investigate whether there is a relationship between self-disclosure and conversation frequency over time. (Trepte and Reinecke, 2013) showed that frequent or high-level of self-disclosure in online social networks (OSN) contributes positively to OSN usage, and vice versa. They showed this through an online survey with

Facebook and StudiVZ users. With SDTM, we can automatically classify self-disclosure level of a large number of conversations, so we investigate whether there is a similar relationship between self-disclosure in conversations and subsequent frequency of conversations with the same partner on Twitter. More specifically, we ask the following two questions:

1. If a dyad displays high SD level in their conversations at a particular time period, would they have more frequent conversations subsequently?

2. If a dyad shows high conversation frequency at a particular time period, would they display higher SD in their subsequent conversations?

6.1 Experiment Setup

We first run SDTM with all of our Twitter conversation data with 150; 120; 120 topics for SDTM K^G, K^M and K^H respectively. The hyper-parameters are the same as in section 5. To handle a large dataset, we employ a distributed algorithm (Newman et al., 2009).

Table 5 shows some of the topics that were prominent in each SD level by KL-divergence. As expected, G level includes general topics such as food, celebrity, soccer and IT devices, M level includes personal communication and birthday, and finally, H level includes sickness and profanity.

For comparing conversation frequencies over time, we divided the conversations into two sets for each dyad. For the *initial* period, we include conversations from the dyad's first conversation to 60 days later. And for the *subsequent* period, we include conversations during the subsequent 30 days.

We compute proportions of conversation for each SD level for each dyad in the *initial* and *subsequent* periods. Also, we define a new measurement, SD level score for a dyad in the period, which is a weighted sum of each conversation with SD levels mapped to 1, 2, and 3, for the levels G, M, and H, respectively.

6.2 Does self-disclosure lead to more frequent conversations?

We investigate the effect of the level self-disclosure on long-term use of OSN. We run linear regression with the intial SD level score as the independent variable, and the rate of change in conversation frequency between *initial* period and *subsequent* period as the dependent variable.

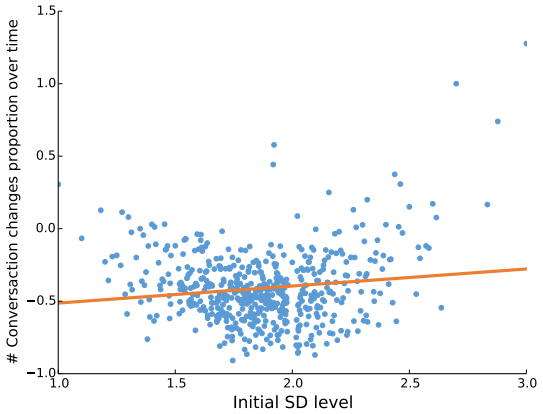

Figure 3: Relationship between initial SD level and conversation frequency changes over time. The solid line is the linear regression line, and the coefficient is 0.118 with $p < 0.001$, which shows a significant positive relationship.

	G level	M level	H level
Coeff (β)	0.094	0.419	0.464
p-value	0.1042	< 0.0001	< 0.0001

Table 6: Relationship between initial SD level proportions and changes in conversation frequency. For M and H levels, there is significant positive relationship ($p < 0.0001$), but for the G level, there is not ($p > 0.1$).

The result of regression is that the independent variable's coefficient is 0.118 with a low p-value ($p < 0.001$). Figure 3 shows the scatter plot with the regression line, and we can see that the slope of regression line is positive.

We also investigate the importance of each SD level for changes in conversation frequency. We run linear regression with initial proportions of each SD level as the independent variable, and the same dependent variable as above. As table 6 shows, there is no significant relationship between the initial proportion of the G level and the changes in conversation frequency ($p > 0.1$). But for the M and H levels, the initial proportions show positive and significant relationships with the subsequent changes to the conversation frequency ($p < 0.0001$). These results show that M and H levels are correlated with changes to the frequency of conversation.

G level			M level			H level		
101	184	176	36	104	82	113	33	19
chocolate	obama	league	send	twitter	going	ass	better	lips
butter	he's	win	email	follow	party	bitch	sick	kisses
good	romney	game	i'll	tumblr	weekend	fuck	feel	love
cake	vote	season	sent	tweet	day	yo	throat	smiles
peanut	right	team	dm	following	night	shit	cold	softly
milk	president	cup	address	account	dinner	fucking	hope	hand
sugar	people	city	know	fb	birthday	lmao	pain	eyes
cream	good	arsenal	check	followers	tomorrow	shut	good	neck

Table 5: High ranked topics in each level by comparing KL-divergence with other level's topics

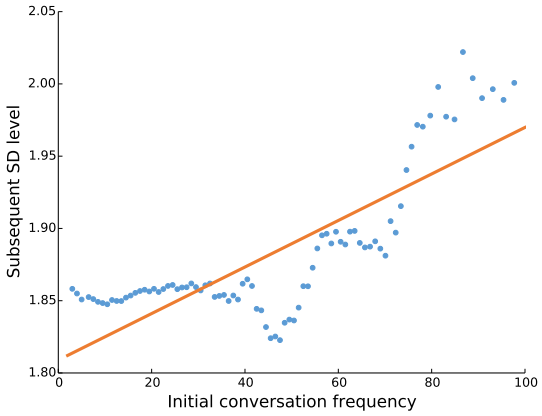

Figure 4: Relationship between initial conversation frequency and subsequent SD level. The solid line is the linear regression line, and the coefficient is 0.0016 with $p < 0.0001$, which shows a significant positive relationship.

6.3 Does high frequency of conversation lead to more self-disclosure?

Now we investigate whether the $initial$ conversation frequency is correlated with the SD level in the $subsequent$ period. We run linear regression with the initial conversation frequency as the independent variable, and SD level in the subsequent period as the dependent variable.

The regression coefficient is 0.0016 with low p-value ($p < 0.0001$). Figure 4 shows the scatter plot. We can see that the slope of the regression line is positive. This result supports previous results in social psychology (Leung, 2002) that frequency of instant chat program ICQ and session time were correlated to depth of SD in message.

7 Conclusion and Future Work

In this paper, we have presented the self-disclosure topic model (SDTM) for discovering topics and classifying SD levels from Twitter conversation data. We devised a set of effective seed words and trigrams, mined from a dataset of secrets. We also annotated Twitter conversations to make a ground-truth dataset for SD level. With annotated data, we showed that SDTM outperforms previous methods in classification accuracy and F-measure.

We also analyzed the relationship between SD level and conversation frequency over time. We found that there is a positive correlation between initial SD level and subsequent conversation frequency. Also, dyads show higher level of SD if they initially display high conversation frequency. These results support previous results in social psychology research with more robust results from a large-scale dataset, and show importance of looking at SD behavior in OSN.

There are several future directions for this research. First, we can improve our modeling for higher accuracy and better interpretability. For instance, SDTM only considers first-person pronouns and topics. Naturally, there are patterns that can be identified by humans but not captured by pronouns and topics. Second, the number of topics for each level is varied, and so we can explore nonparametric topic models (Teh et al., 2006) which infer the number of topics from the data. Third, we can look at the relationship between self-disclosure behavior and general online social network usage beyond conversations.

Acknowledgments

We thank the anonymous reviewers for helpful comments. Alice Oh was supported by the IT R&D Program of MSIP/KEIT. [10041313, UX-oriented Mobile SW Platform]

References

JinYeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in twitter conversations. In *Proceedings of ACL*.

Azy Barak and Orit Gluck-Ofri. 2007. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior*, 10(3):407–417.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of HICSS*.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: Linguistic style accommodation in social media. In *Proceedings of WWW*.

Valerian J. Derlega, Sandra Metts, Sandra Petronio, and Stephen T. Margulis. 1993. *Self-Disclosure*, volume 5 of *SAGE Series on Close Relationships*. SAGE Publications, Inc.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

David J Houghton and Adam N Joinson. 2012. Linguistic markers of secrets and sensitive self-disclosure in twitter. In *Proceedings of HICSS*.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of WSDM*.

Adam N Joinson and Carina B Paine. 2007. Self-disclosure, privacy and the internet. *The Oxford handbook of Internet psychology*, pages 237–252.

Sidney M Jourard. 1971. Self-disclosure: An experimental analysis of the transparent self.

Suin Kim, JinYeong Bak, and Alice Haeyun Oh. 2012. Do you feel what i feel? social aspects of emotions in twitter conversations. In *Proceedings of ICWSM*.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *Proceedings of AAAI*.

Andrew M Ledbetter, Joseph P Mazer, Jocelyn M DeGroot, Kevin R Meyer, Yuping Mao, and Brian Swafford. 2011. Attitudes toward online social connection and self-disclosure as predictors of facebook communication and relational closeness. *Communication Research*, 38(1):27–53.

Louis Leung. 2002. Loneliness, self-disclosure, and icq (" i seek you") use. *CyberPsychology & Behavior*, 5(3):241–251.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

Erika McCallister. 2010. *Guide to protecting the confidentiality of personally identifiable information*. DIANE Publishing.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Sharon Meraz. 2013. Public dialogue: Analysis of tolerance in online discussions. In *Proceedings of ACL*.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of HLT-NAACL*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proceedings of HLT-NAACL*.

Charles Steinfield, Nicole B Ellison, and Cliff Lampe. 2008. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6):434–445.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Sabine Trepte and Leonard Reinecke. 2013. The reciprocal effects of social network site use and the disposition for self-disclosure: A longitudinal study. *Computers in Human Behavior*, 29(3):1102 – 1112.

Sarah I Vondracek and Fred W Vondracek. 1971. The manipulation and measurement of self-disclosure in preadolescents. *Merrill-Palmer Quarterly of Behavior and Development*, 17(1):51–58.

Thomas Ashby Wills. 1985. Supportive functions of interpersonal relationships. *Social support and health*, xvii:61–82.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of EMNLP*.

Jun Zhu, Amr Ahmed, and Eric P Xing. 2012. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research*, 13:2237–2278.