Self-disclosure topic model for classifying and analyzing Twitter conversations

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

Chin-Yew Lin
Microsoft Research
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 strengthens interpersonal relationships and increases social support. Although there are many social science studies of self-disclosure, they are based on manual coding of small datasets and questionnaires. We conduct a computational analysis of self-disclosure with a large dataset of naturally-occurring conversations, a semi-supervised machine learning algorithm, and a computational analysis of the effects of self-disclosure on subsequent conversations. We use a longitudinal dataset of 17 million tweets, all of which occurred in conversations that consist of five or more tweets directly replying to the previous tweet, and from dyads with twenty of more conversations each. We develop self-disclosure topic model (SDTM), a variant of latent Dirichlet allocation (LDA) for automatically classifying the level of self-disclosure for each tweet. We take the results of SDTM and analyze the effects of self-disclosure on subsequent conversations. Our model significantly outperforms several comparable methods on classifying the level of self-disclosure, and the analysis of the longitudinal data using SDTM uncovers significant and positive correlation between self-disclosure and conversation frequency and length.

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). A common instance of self-disclosure is the start of a conversation with an exchange of names and additional self-introductions. Another example of self-disclosure, shown in Figure 1c, where the information disclosed about a family member's serious illness, is much more personal than the exchange of names. 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 conversational behaviors which can serve as proxies for measuring intimacy between two conversational partners.

Twitter conversation data, explained in more detail in section 4.1, enable an extremely large 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 conversation frequency and length patterns on self-disclosure are shown different over time. Our contributions to the research community include the following:

*This work was done when JinYeong Bak was a visiting student at Microsoft Research, Beijing, China.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1986–1996,
October 25-29, 2014, Doha, Qatar.

- We present key features and prior knowledge for identifying self-disclosure level, and show relevance of it with experiment results (Sec. 2).
- 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).
- 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 and conversation behaviors to show that there is significant relationship over time (Sec. 6).

A fabio capello is the manager are u sure its someone else whos playing lol
B common you guys England manager is Roy Hodgson
A noooooo we mean the manager before!
B haha!! the manager before Roy was Fabio yes, Roy became Manager in May after Fabio resigned in February
A ohhhhhhhhhhh we learn something new everyday! Haha

(a) A G level Twitter conversation

A Today's my mother's birthday and she was extremely happy when I informed her I'm applying for Phoenix soon. Happy Bi rthday mom! :D
B HAHA, nice! Tell her I said Happy Birthday and give her a ki ss and hug for me! :3
A that is a bit problematic. My mommy is not here lol
B HAHA! I figured that'd be the case. Well I'm off tomorrow so I guess I'll do it myself tomorrow. XP
A lol She gets home around 6 or 6:30

(b) A M level Twitter conversation

A My mom has just been taken to the hospital by ambulance. Please pray for her. Thank you
B Hugs **A**. Glad your mom is doing better.
A thanks, she is in hospital & is very disoriented.
B My dad was like that when he was in the hospital. Talked to ppl who had been dead for years.
A yeah, she did that too.
it is so scary to see her that way....
B Extra hugs sweetie. I am glad it wasn't a stroke.

(c) A H level Twitter conversation

Figure 1: An example of a Twitter conversation (from annotated dataset) with G, M and H level of self-disclosure.

2 Self-Disclosure

In this section, we look at social science literature for definition of the levels of self-disclosure. Using that definition, we devise an approach to automatically identify the levels of self-disclosure in a large corpus of OSN conversations. We discuss three approaches, first, using first-person pronoun features, and second, extracting seed words and phrases from the Twitter conversation corpus, and third, extracting seed words and phrases from an external corpus of anonymously posted secrets, and we demonstrate the efficacy of those approaches with an annotated corpus.

2.1 Self-disclosure (*SD*) level

To 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 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. Figure 1 shows Twitter conversation examples with G, M and H levels from annotated dataset (see Section 4.2 for a detailed description of the annotated dataset).

2.2 G Level of Self-Disclosure

An obvious clue of self-disclosure is the use of first-person pronouns. For example, phrases such as 'I live' or 'My name is' indicate that the utterance contains personal information. In previous research, the simple method of counting first-person pronouns was used to measure the degree of self-disclosure (Joinson, 2001; Barak and Gluck-Ofri, 2007). Consequently, the absence of a first-person pronoun signals that the utterance belongs in the G level of self-disclosure. We verify this pattern with a dataset of Tweets annotated with G, M, and H levels. We divide the annotated Tweets into two classes, G and M/H. Then we compute mutual information of each unigram, bigram, or trigram feature to see which features are most discriminative. As Table 1 shows, 18 out of 30

Category	Words/Expressions
Unigram	my, I, I'm, I'll, but, was, I've, love, dad, have
Bigram	I love, I was, I have, my dad, go to, my mom, with my, have to, to go, my mum
Trigram	I have a, is going to, to go to, want to go, and I was, going to miss, I love him, I think I, I was like, I wish I

Table 1: High ranked words and expressions by mutual information between G and M/H level in annotated conversations.

Category	Words - SECRET	Words - Annotated
physical appearance	acne, hair, overweight, stomach, chest, hand, scar, thighs, chubby	ankle, face, toe, skin
mental/ physical condition	addicted, bulimia, doctor, illness, alcoholic, disease, drugs, pills	ache, epilepsy, pain, chiropractor, codeine

Table 3: Example words for identifying H level of SD from secret posts (2nd column) and annotated data (3rd column). Categories are hand-labeled.

most highly ranked discriminative features contain a first-person pronoun.

2.3 M Level of Self-Disclosure

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.

2.4 H Level of Self-Disclosure

Utterances with H level express secretive wishes or sensitive information that exposes self or someone close (Barak and Gluck-Ofri, 2007). These are generally kept as secrets. With this intuition, we crawled 26,523 posts from *Six Billion Secrets*[1] site where users post secrets anonymously[2]. We call this external dataset SECRET. Unlike G and M levels, evidence of H level of self-disclosure tends to be topical, such as physical appearance, mental and physical illnesses, and family problems, so we take an approach of fitting a topic model driven by seed words. A similar approach has been successful in sentiment classification (Jo and Oh, 2011; Kim et al., 2013).

A critical component of this approach is the set of seed words with which to drive the discovery of topics that are most indicative of H level self-disclosure. To extract the seed words that express secretive personal information, we compute mutual information (Manning et al., 2008) with SECRET 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.

Since SECRET is quite different from Twitter, we must show that posts in SECRET are semantically similar to the H level Tweets. Rather than directly comparing SECRET posts and Tweets, we use the same method of extracting discriminative word features from the annotated H level Tweets (see Section 4.2). Table 3 shows the seed words extracted from SECRET as well as the annotated Tweets. Because the annotated dataset consists of only 200 conversations, the coverage of the topics seems narrower than the much larger SECRETS, but both datasets show similarities in the topics. This, combined with the results of the model with the two sets of seed words (see Section 5 for the results), shows that SECRETS is an effective and simple-to-obtain substitute for an annotated corpus of H level of self-disclosure.

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

[2] This site is regularly monitored for spam.

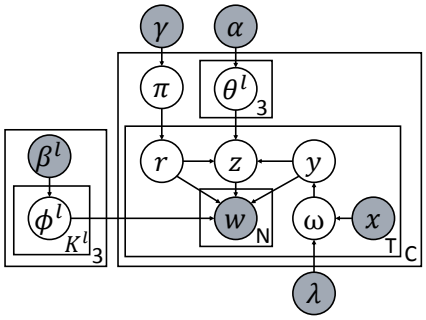

Figure 2: Graphical model of SDTM

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 4: Summary of notations used in SDTM

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

In section 2, we discussed different approaches to identifying each level of self-disclosure, based on social science literature, annotated and unannotated Tweets, and an external corpus of secret posts. In this section, we describe our self-disclosure topic model, based on the widely used latent Dirichlet allocation (Blei et al., 2003), which incorporates those approaches.

Figure 2 illustrates the graphical model of

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 3: Generative process of SDTM.

SDTM and how those approaches are embodied in it. The first approach based on 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 approach of seed words and phrases for levels M and H 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 4 lists the notations used in the model and the generative process, and Figure 3 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).

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. In the graphical model, r is the latent variable that represents this classification, and $\boldsymbol{\pi}$ is the distribution over r. $\boldsymbol{\gamma}$ is a non-informative prior for $\boldsymbol{\pi}$, and $\boldsymbol{\beta}^l$ is an informative prior for each SD level by seed words. For example, we assign a high value for the seed word 'acne' for $\boldsymbol{\beta}^H$, and a low value for 'My name is'. This approach is the same as joint models of topic and sentiment (Jo and Oh, 2011; Kim et al., 2013).

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:

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

4 Data Collection and Annotation

To test our self-disclosure topic model, we use a large dataset of conversations consisting of Tweets over three years such that we can analyze the relationship between self-disclosure behavior and conversation frequency and length 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. We also make sure that each conversation is at least five tweets, and that each dyad has at least twenty conversations.

Users	Dyads	Conv's	Tweets
101,686	61,451	1,956,993	17,178,638

Table 5: Dataset of Twitter conversations. We chose conversations consisting of five or more tweets each. We chose dyads with twenty or more conversations.

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 initialize the set of users by randomly sampling thirteen users who reply to other users in English from the Twitter public streams[3]. Then we crawl each user's public tweets, and look at users who are mentioned in those tweets. It is a breadth-first search in the network defined by users as nodes and edges as conversations. We run this search for dyads until the depth of four, and filter out users who tweet in a non-English language. We use an open source tool for detecting English tweets[4]. 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. Table 5 summarizes the dataset.

4.2 Annotating self-disclosure level

To measure the accuracy of our model, we randomly sample 301 conversations, each with ten or fewer tweets, and ask three judges, fluent in English and graduate students/researchers, 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.

As a result of annotation, there are 122 G level converstaions, 147 M level and 32 H level con-

[3] https://dev.twitter.com/docs/api/streaming

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

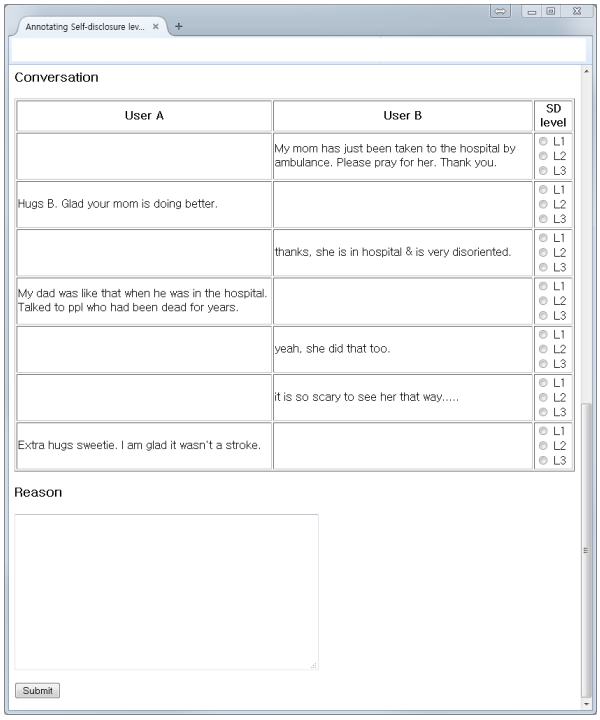

Figure 4: Screenshot of annotation web-based platform. Annotators read a Twitter conversation and annotate self-disclosure level to each tweet.

versations, and inter-rater agreement using Fleiss kappa (Fleiss, 1971) is 0.68, which is substantial agreement result (Landis and Koch, 1977).

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 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. The reason for setting 0.2 as the threshold is that a conversation containing tweets with `H` or `M` level of self-disclosure usually starts with a greeting or a general comment, and contains one or more questions or comments before or after the self-disclosure tweet.

We compare SDTM with the following methods for classifying conversations 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[5]. Used in previous work (Houghton and Joinson, 2012).
- Bag of Words + Bigrams + Trigrams (BOW+): A bag of words, bigram and trigram features. We exclude features that appear only once or twice.
- Seed words and trigrams (SEED): Occurrences of seed words/trigrams from SECRET which are described in section 3.3.
- SDTM with seed words from annotated Tweets (SDTM−): To compare with SDTM below using seed words from SECRET, this uses seed words from the annotated data described in section 2.4.
- ASUM (Jo and Oh, 2011): A joint model of sentiments and topics. We map each SD level to one sentiment and use the same seed words/trigrams from SECRET as in SDTM below. 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).
- First-person pronouns + Seed words/trigrams (FP+SE1): First-person pronouns and seed words/trigrams from SECRET.
- Two stage classifier with First-person pronouns + Seed words/trigrams (FP+SE2): A

[5]personal pronouns, 3rd person singular words, family words, human words, sexual words, etc

Method	*Acc*	G F_1	M F_1	H F_1	*Avg* F_1
LDA	49.2	0.00	0.65	0.05	0.23
MedLDA	43.3	0.41	0.52	0.09	0.34
LIWC	49.2	0.34	0.61	0.18	0.38
BOW+	54.1	0.50	0.59	0.15	0.41
SEED	54.4	0.52	0.60	0.14	0.42
ASUM	56.6	0.32	0.70	0.38	0.47
SDTM−	60.4	0.57	0.70	0.14	0.47
FirstP	63.2	0.63	0.69	0.10	0.47
FP+SE1	61.0	0.61	0.67	0.16	0.48
FP+SE2	60.4	**0.64**	0.69	0.17	0.50
SDTM	**64.5**	0.61	**0.71**	**0.43**	**0.58**

Table 6: 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 macroaveraged 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).

two stage classifier with first-person pronouns and seed words/trigrams from SECRET. In the first stage, the classifier identifies G with first-person pronouns. Then in the second stage, the classifier uses seed words and trigrams to identify M and H levels.

- SDTM: Our model with first-person pronouns and seed words/trigrams from SECRET.

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[6]. BOW+ uses SVM with a radial basis kernel which performs better than all other settings tried including maximum entropy. We split the data randomly into 80/20 for train/test. We run MedLDA, ASUM and SDTM 20 times each and compute the average accuracies and F-measure for each level. We run LDA and MedLDA with various number of topics from 80 to 140, and 120 topics shows best outputs. So we set 120 topics for LDA, MedLDA and ASUM, 60; 40; 40 topics for SDTM K^G, K^M and K^H respectively which is best perform from 40; 40; 40 to 60; 60; 60 topics. We assume that a conversation has few topics and self-disclosure levels, so we set $\alpha = \gamma = 0.1$ (Tang et al., 2014). 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 the same as previous papers (Jo and Oh, 2011; Kim et al., 2013).

As Table 6 shows, SDTM performs better than the other methods for accuracy as well as 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. BOW which is simple word features also does not perform well, showing especially low F-measure for the H level. 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, confirming the effectiveness of a topic modeling approach to this difficult task, but not as well as SDTM. FirstP shows good F-measure for the G level, but the H level F-measure is quite low, even lower than SEED. Combining first-person pronouns and seed words and trigrams (FP+SE1) shows better than each feature alone, and the two stage classifier (FP+SE2) which is a similar approach taken in SDTM shows better results. Finally, SDTM classifies G and M level at a similar accuracy with FirstP, FP+SE1 and FP+SE2, but it significantly improves accuracy for the H level compared to all other methods.

6 Relations of Self-Disclosure and Conversation Behaviors

In this section, we investigate whether there is a relationship between self-disclosure and conversation behaviors over time. Self-disclosure is one way to maintain and improve relationships (Jourard, 1971; Joinson and Paine, 2007). So two people's intimacy changes over time has relationship with self-disclosure in their conversation. However, it is hard to identify intimacy between users in large scale online social network. So we choose conversation behaviors such as conversation frequency and length which can be treated as proxies for measuring intimacy between two people (Emmers-Sommer, 2004; Bak et al., 2012).

[6]It performs better than other classifiers (C4.5, Naive-Bayes, SVM with linear kernel, polynomial kernel and radial basis)

With SDTM, we can automatically classify the SD level of a large number of conversations, so we investigate whether there is a similar relationship between self-disclosure in conversations and subsequent conversation behaviors with the same partner on Twitter.

For comparing conversation behaviors 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 20 days later. And for the $subsequent$ period, we include conversations during the subsequent 10 days. We compute proportions of conversation for each SD level for each dyad in the $initial$ and $subsequent$ periods.

More specifically, we ask the following three questions:

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

2. If a dyad displays high SD level in their conversations at a particular time period, would their subsequent conversations be longer?

3. If a dyad displays high overall SD level, would their conversations increase in length over time more than dyads with lower overall SD level?

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), and run with 28 threads.

Table 7 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.

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.

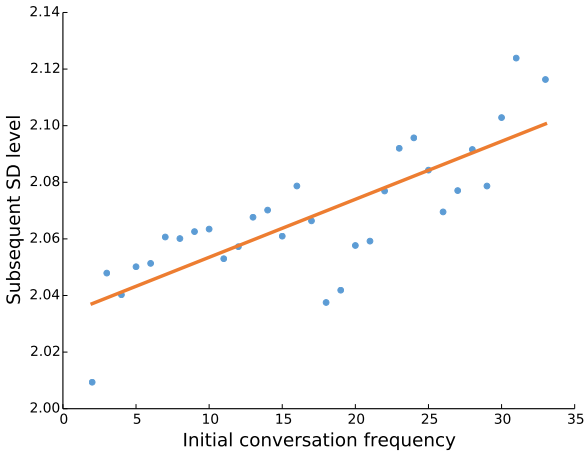

Figure 5: Relationship between initial conversation frequency and subsequent SD level. The solid line is the linear regression line, and the coefficient is 0.0020 with $p < 0.0001$, which shows a significant positive relationship.

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

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 0.0020 with low p-value ($p < 0.0001$). Figure 5 shows the scatter plot. We can see that the slope of the regression line is positive.

6.3 Does high self-disclosure lead to longer conversations?

Now we investigate the effect of the self-disclosure level to conversation length. We run linear regression with the intial SD level score as the independent variable, and the rate of change in conversation length between $initial$ period and $subsequent$ period as the dependent variable. Conversation length is measured by the number of tweets in a conversation.

The result of regression is that the independent variable's coefficient is 0.048 with a low p-value ($p < 0.0001$). Figure 6 shows the scatter plot with the regression line, and we can see that the slope of regression line is positive.

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
make	going	chelsea	link	facebook	come	dick	cough	arms
love	time	liverpool	need	followed	i'll	kick	bad	head
yum	party	won	message	omg	family	face	i've	smirks
hot	election	football	let	right	fun	hoe	need	slowly
cookies	gop	united	sure	saw	friends	lmfao	sore	hair
banana	paul	final	thanks	page	tonight	nigga	flu	face
bread	way	away	my email	timeline	plans	bi	today	chest

Table 7: High ranked topics in each level by comparing KL-divergence with other level's topics

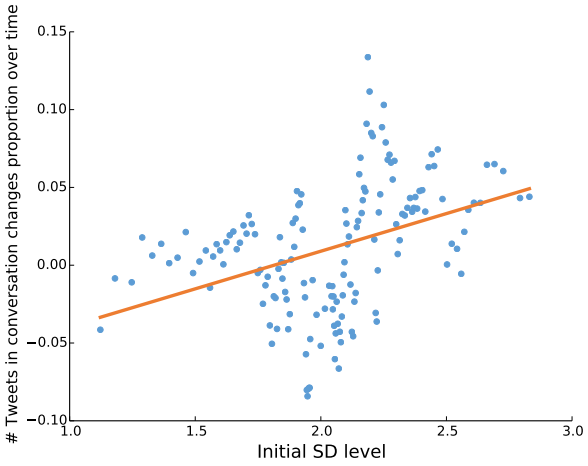

Figure 6: Relationship between initial SD level and conversation length changes over time. The solid line is the linear regression line, and the coefficient is 0.048 with $p < 0.0001$, which shows a significant positive relationship.

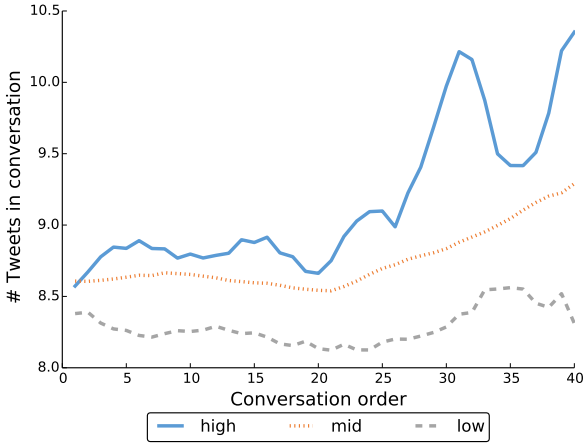

Figure 7: Changes in conversation length over time. We divide dyads into three groups by SD level score as low, medium, and high. Conversation length noticeably increases over time in the medium and high groups, but only slight in the low group.

6.4 Is there a difference in conversation length patterns over time depending on overall SD level?

Now we investigate the conversation length changes over time with three groups, low, medium, and high, by overall SD level. Then we investigate changes in conversation length over time.

Figure 7 shows the results of this investigation. First, conversations are generally lengthier when SD level is high. This phenomenon is also observed in figure 6, but here we can see it as a long-term persistent pattern. Second, conversation length increases consistently and significantly for the high and medium groups, but for the low SD group, there is not a significant increase of conversation length over time.

7 Related Work

Prior work on quantitatively analyzing self-disclosure has relied on user surveys (Ledbetter et

al., 2011; Trepte and Reinecke, 2013) or human annotation (Barak and Gluck-Ofri, 2007; Courtney Walton and Rice, 2013). 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.

Subjectivity which is aspect of expressing opinions (Pang and Lee, 2008; Wiebe et al., 2004) is related with self-disclosure, but they are different dimensions of linguistic behavior. Because there indeed are many high self-disclosure tweets that are subjective, but there are also counter examples in annotated dataset. The tweet "England manager is Roy Hodgson." is low self-disclosure and low subjectivity, "I have barely any hair left." is high self-disclosure but low subjectivity, and "Senator stop lying!" is low self-disclosure but high subjectivity.

8 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 publish the source code of SDTM and the dataset include annotated Twitter conversations and SECRET publicly[7].

We also analyzed the relationship between SD level and conversation behaviors over time. We found that there is a positive correlation between initial SD level and subsequent conversation length. Also, dyads show higher level of SD if they initially display high conversation frequency. Finally, dyads with overall medium and high SD level will have longer conversations over time. These results support previous results in social psychology research with more robust results from a large-scale dataset, and show the effectiveness of computationally analyzing at SD behavior.

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 other linguistic 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. We will explore these directions in our future work.

[7] http://uilab.kaist.ac.kr/research/EMNLP2014

Acknowledgments

We would like to thank Jing Liu and Wayne Xin Zhao for inspiring discussions, and the anonymous reviewers for helpful comments. Alice Oh is supported by ICT R&D program of MSIP/IITP [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*.

S Courtney Walton and Ronald E Rice. 2013. Mediated disclosure on twitter: The roles of gender and identity in boundary impermeability, valence, disclosure, and stage. *Computers in Human Behavior*, 29(4):1465–1474.

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

Tara M Emmers-Sommer. 2004. The effect of communication quality and quantity indicators on intimacy and relational satisfaction. *Journal of Social and Personal Relationships*, 21(3):399–411.

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.

Adam N Joinson. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2):177–192.

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

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

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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.

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*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

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

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198.

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.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

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.