Detection and Analysis of Self-Disclosure in Online News Commentaries

Prasanna Umar
Pennsylvania State University
University Park, Pennsylvania
pxu3@psu.edu

Anna Squicciarini
Pennsylvania State University
University Park, Pennsylvania
acs20@psu.edu

Sarah Rajtmajer
Pennsylvania State University
University Park, Pennsylvania
smr48@psu.edu

ABSTRACT

Online users engage in self-disclosure - revealing personal information to others - in pursuit of social rewards. However, there are associated costs of disclosure to users' privacy. User profiling techniques support the use of contributed content for a number of purposes, e.g., micro-targeting advertisements. In this paper, we study self-disclosure as it occurs in newspaper comment forums. We explore a longitudinal dataset of about 60,000 comments on 2202 news articles from four major English news websites. We start with detection of language indicative of various types of self-disclosure, leveraging both syntactic and semantic information present in texts. Specifically, we use dependency parsing for subject, verb, and object extraction from sentences, in conjunction with named entity recognition to extract linguistic indicators of self-disclosure. We then use these indicators to examine the effects of anonymity and topic of discussion on self-disclosure. We find that anonymous users are more likely to self-disclose than identifiable users, and that self-disclosure varies across topics of discussion. Finally, we discuss the implications of our findings for user privacy.

CCS CONCEPTS

• **Security and privacy** → Social aspects of security and privacy.

KEYWORDS

Online self-disclosure, public platforms, privacy, anonymity

ACM Reference Format:
Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2019. Detection and Analysis of Self-Disclosure in Online News Commentaries. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313669

1 INTRODUCTION

Public discourse has been emboldened by the ubiquitous use of social media, including social networking websites, blogs, and online newspaper comment forums. Users can now share information, express opinions, and discuss various topics of interest with a wide audience through these platforms. This engagement often includes *self-disclosure*— the communication of personal information to others [10]. Online social networks and comment forums thrive on user-generated content and, therefore, encourage self-disclosure [33].

Self-disclosure is generally considered conscious behavior, in which users engage in pursuit of strategic goals including social connectedness, validation, self-expression, relational development, identity clarification, and social control [1, 6, 11]. However, there are associated costs of this behavior to users' privacy. Personal information posted online becomes shared knowledge and may be retained or shared downstream, in ways or by parties unintended by its original owner. Understanding self-disclosure is paramount to devising ways of mitigating these risks and achieving data parsimony [23]. We note that while work on self-disclosure in seemingly bounded environments (e.g. social networks [6, 23]) has shed light on users' motivations for and practices of self-disclosure, less is known about user disclosures in public commentaries. In this work, we address these gaps.

Following, we present in-depth analyses of self-disclosure in a longitudinal dataset of about 60,000 comments on 2202 news articles from four major English news websites [5]. Inspired by prior work on online disclosure and anonymity, we focus on two contextual features that we suspect are related to self-disclosure in public forums, namely anonymity and topic of a commentary.

Our contribution is two-fold. First, we provide an automated method for detecting language indicative of various types of self-disclosure, leveraging both syntactic and semantic information present in texts. We identified 9 categories of self-disclosure, ranging from users' personal attributes (e.g. location, sexual orientation) to subjective categories, such as opinions. Second, we use these results to examine the effects of anonymity and topic of discussion on self-disclosure. Our results indicate that users who have greater anonymity are more likely to self-disclose online, and that self-disclosure varies for different topics of discussion. Finally, we discuss the implications of our results on online self-disclosure for privacy. To the best of our knowledge, this is the first study which uses an automated detection method to analyze triggers of oversharing and personal information disclosure outside the classic boundaries of social network domains.

The remainder of the paper is organized as follows. In the next section, we present related work in this area. Next, we provide background and discuss our hypotheses. We then discuss a novel method that we used to detect self-disclosure in online public platforms, followed by validation of the method. The remainder of this paper is dedicated to hypotheses testing and discussion of findings.

2 RELATED WORKS

People engage in self-disclosing behavior consciously for various intrinsic and extrinsic benefits. As disclosing personal information

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313669

online has privacy risks, users try to maintain a balance between self-disclosure costs and benefits. They utilize different media affordances such as anonymity, audience representations, etc. in this pursuit. Users' self-disclosure decisions are therefore dependent on these media affordances. Previous studies have used survey-based functional models incorporating social media affordances to examine self-disclosure characteristics in social networking sites (SNS). A recent study [9] highlighted the differences in disclosures between private SNS and public SNS. The authors showed the roles of audience representations (bounded vs unbounded), and network characteristics (size and diversity) in describing self-disclosure. However, they noted the need for extending studies of self-disclosure to online platforms other than Facebook and Twitter.

Anonymity, as a social media affordance contributing to self-disclosure, has been extensively studied under the term online disinhibition [31]. Multiple studies have noted the positive relationships between anonymity and self-disclosure in computer mediated communication [16, 21]. However, the findings are not consistent for all types of online platforms. For instance, Qian and Scott [27] reported no association between increased visual anonymity with greater self-disclosure. Similarly, Chen et al. [8] showed negative effect of network anonymity on self-disclosure on Sina Weibo platform but reported positive effect of perceived anonymity. These mixed findings suggest that different online media have different affordances of anonymity and therefore, the relationship between anonymity and self-disclosure is dependent on media characteristics. In this work, we study the effect of anonymity on self-disclosure in online news commentaries which afford different levels of anonymity as described in in Section 3.

Studies on self-disclosure have mostly relied on manual coding and analysis of users' contents primarily into broad categories of personal information, personal thoughts and opinions, and personal feelings and emotions [4, 14]. Personal information is related to facts about a person while textual contents that convey thoughts, opinions, feelings and emotions are subjective in nature. Each of these broad categories may include several sub-categories such as biographic information, property, location, family details, etc [7, 14]. In line with existing literature, we adopt similar terminologies and consider two broader categories of self-disclosure namely: objective and subjective (See Table 1). Objective categorization include factual information about a person: birthday/age, race, sexual orientation, affiliation, money, relationships and experiences. Subjective categorization includes categories related to internal states of an individual: interests, opinions and feelings.

Automated detection of self-disclosure has been tackled by recent studies [7, 33]. Caliskan et al. [7] created a supervised machine learning method to detect private information in tweets through the use of privacy ontology, named entity recognition, topic modeling and sentiment analysis. They assigned privacy scores to each user based on the percentage of the userâĂŹs tweets annotated. We used a different approach as we rely on an unsupervised method to detect individual self-disclosure categories in texts. Bak et al. [3] created a supervised method with topic models and SVM to detect personally identifiable information (PII) and personally embarrassing information (PEI). Precision and recall for their supervised method of PII detection were 0.23 and 0.21 while for PEI precision and recall were 0.30 and 0.23 respectively. In comparison, our unsupervised method achieved greater precision and recall across most categories of self-disclosure. Further, Bak et al. [2] applied modified latent Dirichlet allocation (LDA) topic models for semi-supervised classification of Twitter conversations into three self-disclosure levels. They did not categorize texts into individual categories of self-disclosure as we do.

3 BACKGROUND AND HYPOTHESES

We study self-disclosure as it correlates to a few critical contextual dimensions. Here, we discuss our hypotheses, and ground them in relevant literature. The first dimension we consider is *anonymity*.

H1: Anonymous users are more likely to self-disclose than identifiable users in online public commentaries.

Other authors have found that anonymity contributes to increased self-disclosure both on- and offline [21, 25, 29]. In particular, previous studies have suggested that, emboldened by the cover of anonymity, people may feel less restrained and express fewer inhibitions [31]. Many online platforms offer some degree of anonymity to users in service to freedom of expression [30], where the nature of that anonymity varies across platforms. To comment within article commentaries, in most cases, users either need to sign in to a commenting platform (e.g., Disqus [12]), provide an email address (medium/high anonymity if a throw-away account is used [19]) or provide links to an existing social network account (high identifiability), e.g. Facebook.

Recent work indicates [32] there is a predictive relationship between anonymity state and perceived anonymity. Users intuit that they are more identifiable when linking social network accounts than they are when using pseudonyms. Our analyses consider two anonymity states among users, representing pseudonym vs. social-media-linked accounts. *Following, we refer to users logged into Disqus as "anonymous" and users logged in through existing social media accounts as "identifiable".* Our first hypothesis considers differences in self-disclosure between anonymous and identifiable users.

H2: Users' self-disclosure varies across topics.

A second contextual dimension that we suspect is relevant to self-disclosure is the *topic* of a discussion. Public discourse in news websites and opinion forums is topic-oriented. The news article or the opinion column usually determines the initial topic of discussion. Because individuals relate to content based on prior experiences and feelings, they respond differently to different topics [33] where self-disclosure is a subset of that response. Hence, we hypothesize that users self-disclose differently across topics.

4 DETECTION OF SELF-DISCLOSURE

We anchor our approach for detection of traces of self-disclosure in an opinion extraction technique [17], wherein opinion, opinion-holder and the opinion topic are extracted. However, our goal is not limited to detection of subjective language (opinion). We intend to detect objective language pertaining to self-disclosure as well. To do so, we leverage both the semantic and the syntactic resources in a sentence. The intuition is that a self-disclosing sentence has self-reference[1] as a subject or the object of reference is the self.

[1]We do not consider any indirect forms of self-disclosure with no explicit self-reference in the text and we do not detect the degree of self-disclosure

Table 1: Description of categories of self-disclosure used as instructions for labelling survey

Language	Categories	Description
Objective	Birthday/Age	Sharing one's own birthday information or references to own age.
	Race	Sharing one's own race such as being black, white, Hispanic, etc
	Sexual Orientation	Sharing one's sexual orientation and identity such as being straight or LGBT. Includes marital status
	Location	Sharing one's own location such as town, city, states, proximity to a landmark, etc
	Affiliation	Sharing one's nationality, religion, political affiliation, loyalties to groups and brands of a certain nationality, etc
	Money	Sharing one's own financial worth, monetary values of property, plans/goals related to money, etc
	Relationships	Sharing information about the family composition such as having children, brothers, sisters, etc
	Experience	Sharing past experiences of events, habits, work-life, etc. Includes positive or negative experiences and recollections of any past events or memories
Subjective	Interests	Sharing one's own hobbies and interests, including pastimes, favorites, tastes in music, movies, and books. Includes disclosures about pets, as well.
	Feelings	Expressions of deep personal feelings, including humiliation, desires, anxiety, depression, fears, pain, and beliefs which most people would likely disclose only to a friend or family
	Opinion	Discussing one's own opinions, attitudes, and beliefs about current and/or historical events that one is NOT relating to personal experience. Includes views on government, trends, specific events in entertainment/sports, religion, etc.
-	Other	Personal information about the author is revealed but can't be categorized in any categories
-	None	No information about the author revealed

Presence of first person pronouns is generally considered as linguistic markers for self-disclosing texts [15]. Moreover, several studies have used presence of first person pronouns as a prominent feature in determining self-disclosure in online contents [2, 11, 33].

The categories of self-disclosure as shown in Table 1 are distinguished based on related named entities present in the sentence in the vicinity of related verbs. For instance, anyone disclosing where he/she lives uses sentences with verbs related to location like live, stay, etc together with the place where they live. A typical example would be the sentence "I live in Pennsylvania". In this scenario, the author of this text includes the place of living (a location entity) and the verb "live" (related to location) with reference to the self (use of "I"). Hence, our method identifies self-disclosure categories in text through the knowledge of subject, verb, object and named entities. The overall method of detection and categorization of self-disclosure consists of four phases: (1) dictionary construction, (2) subject-verb-object triplet detection, (3) named entity recognition and (4) rule based matching.

4.1 Category Specific Dictionary Construction

We constructed a vocabulary of verbs often associated with the different categories of self-disclosure. The dictionary consists of verbs used in a labeled dataset from a study of self-disclosure in Airbnb profiles [22]. The dataset consisted of 5248 annotated sentences from 1234 user profiles. Each sentence was categorized into one or many of the eight categories. We extracted the frequent root verbs from sentences categorized into categories of work or education, origin or residence, travel, interests and tastes, and relationships to construct dictionaries of verbs associated with our categories. The dictionaries also consist of additional verbs manually added to aid in the categorization. We constructed multiple dictionaries of verbs: subjective (36), location (21), affiliation (20), , money (12), birthday/age (3), and race and sexual orientation (1). We also created additional dictionaries of attributes for categories of relationships, race, and sexual orientation. Specifically, names of different racial groups, types of sexual orientation and types of social and family relationships were compiled to create these individual dictionaries. We used these dictionaries as entities for the aforementioned three categories.

4.2 Subject, Verb and Object Extraction

We used the Python package Spacy based subject verb object extraction[2]. The implementation was adapted and modified to suit our goals. The subject, verb and object extractor takes text as input, pre-processes the text, splits it into individual sentences and further splits a sentence into individual clauses if present. Text pre-processing includes expanding the contractions such as "I'm" to "I am". Each clause in a sentence passes into the dependency parser and all the verbs are extracted. The dependencies or the syntactic relationships between other tokens in the clause with the verbs is used to determine the subject and objects in the text. Our method was aware of the passive voice and negation in sentences.

4.3 Named Entity Recognition

We attribute the presence of "real-world object" called a named entity like a location, nationality, etc in a sentence as a distinguishing feature among different categories of self-disclosure. We used a model[3] trained on OneNotes corpus, which can detect about 18 different types of entities[4]. Entities of DATE and CARDINAL (number) were used for birthday/age category. Location related entities such

[2]https://github.com/NSchrading/intro-spacy-nlp/blob/master/subject_object_extraction.py
[3]https://github.com/explosion/spacy-models/releases//tag/en_core_web_lg-2.0.0
[4]Details about entities: https://spacy.io/api/annotation

as GPE, FAC and LOC served as distinguishing entities for location category while NORP (Nationalities or religious or political groups) distinguished affiliation category in our categorization scheme.

4.4 Rule Based Matching

Overall, the self-disclosure categorization scheme remained similar for all the objective categories. Self-reference with presence of specific category related verb and appropriate named entities signaled self-disclosure in text. As the proximity of named entities in the sentences vary according to sentence construction, we added a proximity window of up to five words on either side of verb to effectively detect specific categories. It is useful in removing false positives as in case of sentence like *I have countless arguments with seemingly educated people in many countries on why Singapore works.* Here, all the attributes for location disclosure such as matched verb "have" with location entity "Singapore" and subject "I" are present but the proximity check prevents it's categorization into location category. Similarly, we added a proximity window of three words between subject and verb. It should also be noted that we do not include any named entities in the categorization scheme for subjective categories like personal interests, opinions and feelings. Such categories are solely based on presence of first person pronouns and verbs often used in expression of subjectivity.

5 VALIDATION OF METHOD

In this section, we describe process we used to validate our method of detecting and categorizing self-disclosure from user comments.

5.1 Dataset

Our work uses a dataset of user comments on news websites obtained from authors of the paper [5]. It consists of 309319 comments on 52260 news articles crawled from 10 selected news websites over a three month period. We constructed a smaller subset of the data by considering four news websites from majority native English speaking nations: ABC News, CNBC, The Huffington Post and Techcrunch. The time period of the contents was between March and August 2015. We cleaned the data by removing duplicates and comments with no readable text. In total, there are about 59249 comments from 22132 users (14219 identifiable and 7913 anonymous). As described in the earlier section, we differentiate anonymity states based on the account people use to comment on news websites. Commentors on The Huffington Post and Techcrunch used Facebook profile as their identity on these news websites. The users of CNBC and ABC News used Disqus accounts. Accordingly, we consider the users who commented on The Huffington Post and Techcrunch as identifiable and the rest as anonymous. The data consisted of 2202 news articles: The Huffington Post (1136), Techcrunch (119), CNBC (421) and ABC News (526). The anonymous users had higher average number of comments (4, SD = 10.85) than identifiable users (1.94, SD = 2.82).

In order to validate our method, we categorized all the comments using the method described earlier in the paper. Because the data consisted of numerous comments with incomplete sentences and imperative sentences (with no subject), we sampled the dataset after categorization. If we sampled before categorization, the sample size required to include samples of all the relevant categories of self-disclosure would be large. Hence, we performed stratified random sampling only after categorization.

The categorization scheme used here generated multi-label classification for each comment as any comment may contain different types of self-disclosure (including no self-disclosure category). We converted multi-label data into multi-class before sampling. We sampled 200 comments from each class and any class with number of samples less than 200 included all the examples of that class in the overall sample. In this way, we represented all types of examples in the final sample of 3516 comments.

5.2 Labeling

We used the crowd-sourced platform Amazon Mechanical Turk to obtain labels for our sample data. The labeling task was conducted under IRB protocol "STUDY00010405" approved by the Penssylvania State University's Institutional Review Board (IRB). Detailed instructions were given to the labelers about different categories of self-disclosure. Moreover, we included an example of categorization. Each comment was labeled by three different workers into atleast one of the categories in Table 1. To ensure the quality of labels and to ensure that the recruited workers read instructions carefully, the workers were required to answer an attention check question similar to the instructional manipulation check in the paper [26]. We rejected about 8.2% of the responses from the crowd-sourced workers failing the attention check.

We used Gwet's AC1 [13] as a reliability metric for the labeled data. Categorization scheme that classifies a comment into one or many categories of self-disclosure was treated as a group of binary classifiers: one binary classifier for presence/absence of a particular category of self-disclosure. We dropped the category of "organization" because of poor reliability. This category was related to the disclosures of one's workplace, school, membership in any organization, following of sports teams and ownership of properties of a certain brand. Overall, we observed good agreement among labelers: birthday/age (96.5%), race (97.7%), sexual orientation (97.7%), location (85.8%), affiliation (83.5%), money (89.3%), relationships (89.2%), experience (62.9%), interests (93.6%), feelings (81.4%), other (74.1%) and none (70.5%). Also, considering a single label for self-disclosure presence or absence in the text, the reliability score was 70.5%. For each comment, final labels (one or many) were obtained through majority voting. 9.76% of total number of comments did not have a consensus among the labelers.

5.3 Performance of Method

We evaluated the categorization scheme by comparing the crowd-sourced labels against the labels from the algorithm. Performance metrics were calculated only for 3174 comments which had final labels obtained through consensus among multiple labelers. We consider a comment to be self-disclosing in nature if it contains any one of the categories of self-disclosure (except Other) and if None, it is considered as not self-disclosing in nature. Overall, the precision, recall and F1 scores were 98, 89 and 93 respectively. Individual categories' classification performance metrics are summarized in Table 2 with a combined subjective category. We merged the categories of interests, opinions and feelings as a combined subjective category

because unlike objective facts about a person, the subjectivity in interests, opinion and feelings are difficult to distinguish and detect. As pointed out in [24, 34], "because subjectivity is a feature of a person's mind, it is not open to objective observation or verification". Hence, distinguishing between individual categories of subjective language is a task not in purview of our work.

Table 2: Performance metrics for categorization

Category (Support)	Precision	Recall	F1-Score
Birthday/Age (37)	19	43	26
Race (21)	46	62	53
Sexual Orientation (13)	50	62	55
Location (188)	26	70	38
Affiliation (106)	27	42	33
Money (150)	50	23	31
Relationships (206)	36	82	50
Experience (464)	29	50	37
None (97)	15	28	19
Subjective: interests, opinions and feelings (2383)	76	73	74
Overall self-disclosure	98	89	93

6 RESULTS ON HYPOTHESIS TESTING

We performed multiple tests to examine two specific hypotheses as described in detail in section 3. In this section, we present the statistical results of our hypotheses tests.

6.1 H1: Anonymity and Self-Disclosure

We tested our first hypothesis, namely, that anonymous users are more likely to self-disclose than identifiable users. We considered only those comments in a commentary which did not have any self-disclosing comments within five[5] preceding comments, in order to remove any peer effects (peer influence). We aggregated comments by user, obtaining histories of comments for 12,936 users—5218 anonymous and 7718 identifiable. We labeled users who self-disclosed in at least one comment *self-disclosing*. A Chi-squared test revealed that self-disclosure was significantly associated with anonymity [$\chi^2(1) = 59.538$, $p < 0.001$]. Binary logistic regression showed that anonymous users were more likely (1.366 times, $p < 0.001$) to self-disclose than identifiable users, adding support to H1.

Beyond absolute presence/absence of self-disclosure, we were also interested to see if there was a difference in frequency of self-disclosure between users of different anonymity states. Because users with higher number of comments have higher opportunities of self-disclosure, we suspected to see a difference in proportion of self-disclosure between two groups. Hence, we balanced the dataset for the average number of comments (2.91) for both user groups of different anonymity states. Specifically, we created a subset with total of 9468 users and comprising of 4734 identifiable and 4734 anonymous users. Stratified random sampling across strata of different number of comments was done to create this balanced data. We conducted a Mann-Whitney U test to test the differences between the two independent groups of anonymity states in terms of proportion of self-disclosure. The mean rank of proportion of self-disclosure was higher for anonymous users than identifiable users [U=11632245, $p < 0.001$], confirming our hypothesis.

In the next section, we further explore these findings on the effects of anonymity on self-disclosure by accounting for topic of conversation.

[5]We used five comments only because users are more likely to read only few preceding comments either the most popular ones or the most recent ones [20]

Table 3: Top five topics from topic model with frequent keywords

Topic	Keywords
Crime & Police Shootings	police, cop, people, kill, officer, shoot, death
Economy & Finance	money, tax, pay, people, government, wage, year
Security & Terrorism	people, war, kill, american, isis, fight, country
Technology	apple, watch, good, phone, time, car, people
Politics	people, time, hillary, country, vote, good, party

6.2 H2: Topics and Self-Disclosure

In order to validate our hypothesis about the effect of topics on self-disclosure, we first pre-processed our dataset for topical extraction. We performed topical modeling of users' comments using a Python wrapper[6] for LDA mallet. To get an optimal number of topics, multiple topic models were created with number of topics varying from 1 to 200. We used coherence score [28] as a measure of quality and interpretability of the topic models. Subsequent analysis revealed the best topic model with 20 topics and coherence score of 0.358. In this topic model, topics on related discussions were merged. For instance, two topics related to politics were merged into one. In this way, we merged similar topics to obtain 15 clear and distinct topics. Examples of some topics are shown in Table 3. Each news article with the associated comments were assigned the most dominant topic from topic model.

We used Kruskal-Wallis H test [18] to examine if users disclose about themselves differently for different topics. Kruskal Wallis H test is a non-parametric version of one way ANOVA and therefore, it is applicable to test our data (non-normal) for differences in proportion of self-disclosing users across 15 topics. Results showed that there was a statistically significant difference in proportion of self-disclosing users between the different topics of discussion [$\chi^2(16) = 29.928$, $p < 0.05$], confirming our hypothesis. Statistical significant differences were also observed for topics in proportion of users disclosing birthday/age, location, affiliation, relationships, money and experiences categories (each $p < 0.001$). We discuss the implications of these results in section 7. We refined the analysis to examine the effect of topic of discussion on the relationship between anonymity state and disclosure behavior. Multiple Mann-Whitney U tests were performed to test if the proportion of self-disclosure

[6]https://radimrehurek.com/gensim/models/wrappers/ldamallet.html

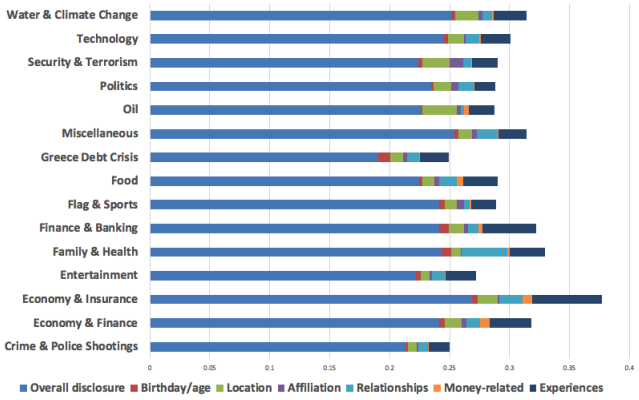

Figure 1: Average proportions of self-disclosing users per category across topics

for users differs within topics for different anonymity states. No significant differences were found after Bonferroni correction. Hence, the results show that the topic of discussion has no effect on the relationship between anonymity states and self-disclosing behavior.

7 DISCUSSION

In this study, we were able to detect language patterns of self-disclosure with differentiation into pertinent categories at reasonable accuracy levels. Because past studies on automatic detection of self-disclosure focused on supervised classification of levels of disclosure rather than unsupervised categorization of self-disclosure language [2, 7, 33], we cannot make like to like comparisons of our method's performance. Inherently, the task of detecting nuances of self-disclosure language is complex, even for humans as revealed by the agreement statistics between human annotators (See Section 5.2).

We used our categorization scheme to label users' comments in the experimental dataset. In general, *our findings confirmed our first hypothesis: users who have higher anonymity state are more likely to self-disclose than identifiable users.* The effect of anonymity as such to encourage or increase self-disclosure has been observed in social networking sites [8, 21] and our results show it is consistent even in online news commentaries.

Regarding the second contextual factor, topic of discussion, we found that proportion of self-disclosing users varied for different topics. Self-disclosing users' proportions across many individual categories also varied across topics (See Figure 1). Specifically, we observed discussions related to economy and insurance had the highest proportion of self-disclosing users. The two most relevant articles, as assigned by topic model to this topic, were centered around health-care policy. We observed many users shared their experiences and financial information, e.g. regarding health insurance, and note that this is reflected in the high proportion of users disclosing experiences and money-related information. Disclosures of specific categories in the context of certain topics are intuitive. For instance, highest proportion of users disclosed their relationships in the context of discussions about family and health. While we did not pursue specific hypotheses related to patterns of disclosure across topics, the findings here warrant further research in this regard. These results suggest the importance of context provided by the topic of discussion [33] of self-disclosure in SNS.

Our study on self-disclosure behavior has several implications. First, the results show that the online dis-inhibition actively drives responses from users on online news commentaries. Users utilize the affordances of anonymity provided by the online platform to exercise self-expression, especially subjective opinions and feelings. Hence, there are design implications for platforms to afford some levels of anonymity to maintain users' intimate participation in expression of opinions and feelings.

Second, the use of automated methods like ours to detect personal information in public comments highlights potentially hidden privacy threats. Users may be aware of the risks of disclosing too much in single comments (even with relative anonymity) and yet may be oblivious to the use of longitudinal aggregation of disclosed information. With close to 60,000 public comments over about 22k users, aggregated disclosures for each user revealed multiple categories of personal information. The ratio of number of anonymous users to number of identifiable users disclosing personal information increased with increase in number of categories disclosed. We found that higher number of users (48) disclosed four or more categories of personal information as compared to identifiable users (34). For example, multiple users disclose information about birthday/age and location. This can lead to unique identification, even if the users sign themselves with a truly random id. If such detection methods are carried out at larger scales in publicly available data (like public comments), we believe user profiles with sensitive information can be created and even anonymous users may be identified. Beyond loss of anonymity, aggregated personal information can be used by adversaries for malicious purposes.

8 CONCLUSIONS

In this study, we presented a novel method of detecting self-disclosure within commenting platforms online. We then examined two contextual factors contributing towards self-disclosure in online news commentaries: anonymity and topic of discussion. Using a dataset of user comments on online news commentaries, we tested and confirmed our hypotheses regarding these contextual factors. Specifically, our findings showed that anonymity elicits self-disclosure in online public comments. Additionally, context as specified by topic of discussion results in more likelihood of self-disclosure. Lastly, we provide implications of our results on online privacy. Self-disclosure behavior exhibited over time is perilous to user's privacy.

We see this work as a first step toward a more comprehensive study on online self-disclosure and the context within which it occurs. Our approach on automated self-disclosure detection is limited by the category-specific dictionaries. Also, albeit comprehensive, our taxonomy on self-disclosure categories is not exhaustive; there are several other categories of personal information revealed by online users. Further research is warranted in these areas.

ACKNOWLEDGEMENTS

We thank J. Barua, D. Patel and V. Goyal for the data used in this study. This work was supported by the National Science Foundation under grant 1453080.

REFERENCES

[1] Olga Abramova, Amina Wagner, Hanna Krasnova, and Peter Buxmann. 2017. *Understanding Self-Disclosure on Social Networking Sites-A Literature Review.* Technical Report. Darmstadt Technical University, Department of Business Administration, Economics and Law, Institute for Business Studies (BWL).

[2] JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 1986–1996.

[3] Jin Yeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in Twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2.* Association for Computational Linguistics, 60–64.

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

[5] Jayendra Barua, Dhaval Patel, and Vikram Goyal. 2015. TiDE: Template-Independent Discourse Data Extraction. In *International Conference on Big Data Analytics and Knowledge Discovery.* Springer, 149–162.

[6] Natalya N Bazarova and Yoon Hyung Choi. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication* 64, 4 (2014), 635–657.

[7] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. 2014. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society.* ACM, 35–46.

[8] Xi Chen, Gang Li, YunDi Hu, and Yujie Li. 2016. How anonymity influence self-disclosure tendency on sina weibo: An empirical study. *The Anthropologist* 26, 3 (2016), 217–226.

[9] Yoon Hyung Choi and Natalya N Bazarova. 2015. Self-disclosure characteristics and motivations in social media: Extending the functional model to multiple social network sites. *Human Communication Research* 41, 4 (2015), 480–500.

[10] Paul C Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin* 79, 2 (1973), 73.

[11] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity.. In *ICWSM.*

[12] Disqus. 2018. Disqus Login Page. https://disqus.com/

[13] Kilem Gwet. 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment* 1, 6 (2002), 1–6.

[14] Erin E Hollenbaugh and Marcia K Everett. 2013. The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication* 18, 3 (2013), 283–302.

[15] David J Houghton and Adam N Joinson. 2012. Linguistic markers of secrets and sensitive self-disclosure in twitter. In *System Science (HICSS), 2012 45th Hawaii International Conference on.* IEEE, 3480–3489.

[16] 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 (2001), 177–192.

[17] Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text.* Association for Computational Linguistics, 1–8.

[18] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.

[19] Alex Leavitt. 2015. This is a throwaway account: Temporary technical identities and perceptions of anonymity in a massive online community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.* ACM, 317–327.

[20] Cong Liao, Anna Squicciarini, Christopher Griffin, and Sarah Rajtmajer. 2016. A hybrid epidemic model for deindividuation and antinormative behavior in online social networks. *Social Network Analysis and Mining* 6, 1 (2016), 13.

[21] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, 3857–3869.

[22] Xiao Ma, Jeffrey T Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles.. In *CSCW.* 2397–2409.

[23] Philipp K Masur and Michael Scharkow. 2016. Disclosure management on social network sites: Individual privacy perceptions and user-directed privacy strategies. *Social Media+ Society* 2, 1 (2016).

[24] Myriam D Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing* 5, 2 (2014), 101–111.

[25] Melanie Nguyen, Yu Sun Bin, and Andrew Campbell. 2012. Comparing online and offline self-disclosure: A systematic review. *Cyberpsychology, Behavior, and Social Networking* 15, 2 (2012), 103–111.

[26] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872.

[27] Hua Qian and Craig R Scott. 2007. Anonymity and self-disclosure on weblogs. *Journal of Computer-Mediated Communication* 12, 4 (2007), 1428–1451.

[28] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining.* ACM, 399–408.

[29] Zick Rubin. 1975. Disclosing oneself to a stranger: Reciprocity and its limits. *Journal of Experimental Social Psychology* 11, 3 (1975), 233–260.

[30] Arthur D Santana. 2014. Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice* 8, 1 (2014), 18–33.

[31] John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7, 3 (2004), 321–326.

[32] Michail Tsikerdekis. 2013. The effects of perceived anonymity and anonymity states on conformity and groupthink in online communities: AW ikipedia study. *Journal of the American Society for Information Science and Technology* 64, 5 (2013), 1001–1015.

[33] Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* ACM, 74–85.

[34] Janyce M Wiebe. 1990. Identifying subjective characters in narrative. In *Proceedings of the 13th conference on Computational linguistics-Volume 2.* Association for Computational Linguistics, 401–406.