What You Want Is Not What You Get: Predicting Sharing Policies for Text-based Content on Facebook

Arunesh Sinha* Yan Li† Lujo Bauer*

* Carnegie Mellon University
Pittsburgh, PA
{lbauer,aruneshs}@cmu.edu

† Singapore Management University
Singapore
yan.li.2009@smu.edu.sg

ABSTRACT

As the amount of content users publish on social networking sites rises, so do the danger and costs of inadvertently sharing content with an unintended audience. Studies repeatedly show that users frequently misconfigure their policies or misunderstand the privacy features offered by social networks.

A way to mitigate these problems is to develop automated tools to assist users in correctly setting their policy. This paper explores the viability of one such approach: we examine the extent to which machine learning can be used to deduce users' sharing preferences for content posted on Facebook. To generate data on which to evaluate our approach, we conduct an online survey of Facebook users, gathering their Facebook posts and associated policies, as well as their intended privacy policy for a subset of the posts. We use this data to test the efficacy of several algorithms at predicting policies, and the effects on prediction accuracy of varying the features on which they base their predictions. We find that Facebook's default behavior of assigning to a new post the privacy settings of the preceding one correctly assigns policies for only 67% of posts. The best of the prediction algorithms we tested outperforms this baseline for 80% of participants, with an average accuracy of 81%; this equates to a 45% reduction in the number of posts with misconfigured policies. Further, for those participants (66%) whose implemented policy usually matched their intended policy, our approach predicts the correct privacy settings for 94% of posts.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and Protection*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

Keywords

privacy; social network; Facebook; machine learning; natural language processing

AISec'13, November 4, 2013, Berlin, Germany.
ACM 978-1-4503-2488-5/13/11.
http://dx.doi.org/10.1145/2517312.2517317

1. INTRODUCTION

As the amount of content that users publish on social networking sites rises, so do the danger and costs of inadvertently sharing content with an unintended audience. Social networks have tried to meet this challenge by offering increasing flexibility in the sharing settings (i.e., access-control policies) and mechanisms that can be used to restrict the audience of content uploaded by users. These include allowing users to specify both group-based and per-person policies [41], automatically creating groups to assist in specifying sharing settings, and developing interfaces to help users understand which of their content is visible to others [25].

Despite these efforts, studies show that users frequently misconfigure the access-control policy for the content they share (e.g., [23, 17]). These misconfigurations have a variety of undesirable consequences, from causing regret and embarrassment to loss of (real-life) friendships and jobs [40, 10] to affecting international relations [28].

One way to mitigate these problems is to develop automated tools to assist users in correctly setting their policy. A basic step in that direction is Facebook's default behavior of suggesting to users that they protect a new post with exactly the policy they assigned to their previous post[1].

This paper explores the use of a more powerful automated mechanism: we examine the extent to which machine learning can be used to deduce the users' sharing preferences for content uploaded to Facebook. In our approach, a machine-learning algorithm is trained on a user's past posts—where "post" refers to any textual or partially textual content users share—and the policies that were used to protect them. During the creation of a new post, the machine-learning algorithm evaluates the new content and suggests which of the policies used on previous posts is the best fit.

To generate data on which to evaluate the effectiveness of our approach, we conduct an online survey of Facebook users. As part of the survey, we use a Facebook application to gather participants' Facebook posts and the policies under which they were published. We interactively review a random sample of 20 posts with each participant to determine any inconsistencies between the policy that was used to protect each post on Facebook (the *implemented policy*) and the policy that the participant wished to enact (the *intended policy*).

The machine-learning algorithm we focus on, after preliminary experiments with LLDA [32], is the MaxEnt algorithm [5]. The features on which the algorithm bases its predictions are chosen on the basis of best cross-validation performance from the pool of

[1]Facebook suggests the policy that was assigned to the previous post when that post was created, ignoring any changes in policy made after the post was created.

features that we considered, including the text of posts, time of creation, word n-grams, and the presence of URLs and attachments. We evaluate the effectiveness of these algorithms on test data, kept aside specifically for this purpose. In evaluation, we measure accuracy with respect to both the policy predicted to be the best match and the two policies predicted to be most likely; the latter could be useful if prediction was used to generate a list of likely policies for the user to choose from.

We show that the baseline with which we compare—Facebook's default behavior of assigning to a new post the privacy settings of the preceding one—is only somewhat effective. Evaluated against participants' intended policy, it correctly suggests the policy for only 67% of posts. In contrast, we show that our machine learning approach performs significantly better, correctly predicting 81% of the intended policies on average. In predicting the intended policy, we find, unsurprisingly, that the accuracy of our approach strongly correlates to the availability of good training data. For participants who indicated (through the review of sampled posts) that their implemented policies matched their intended policies at least 75% of the time, our approach was almost invariably effective, achieving a prediction accuracy of 98%. We also find that all the post features we consider are helpful for achieving best prediction accuracy, and that accuracy ramps up surprisingly quickly as the amount of training data grows.

We believe these results to be very promising, as they suggest that it is feasible to use machine learning to effectively help users in setting the policy for the content they upload to social networks. Integrating such an approach in a deployed social network would require that the machine-learning algorithms be able to distinguish between good-quality and poor-quality training data, which could be accomplished in a several ways, including by explicitly interacting with users and by examining corrections users make to their initial policies. We find that combining a small amount of good-quality data with some unknown-quality data yields predictions sufficiently accurate to be practically useful; this mitigates the problem of obtaining a large quantity of good-quality data in a real-world application. Used in this way, an automated approach like the one we explore can significantly lower the likelihood that a user's momentary inattentiveness while posting will lead to incorrectly set privacy settings.

To summarize, the contributions of this paper are the following:

- We develop several configurations in which machine-learning algorithms could be used to assist users in setting the policy for their posts, and several metrics for evaluating the effectiveness of these approaches.
- We test the effectiveness the MaxEnt machine-learning algorithm in these configurations, including principled experimentation with different post features for training and prediction, on data collected from Facebook users. The predictions are correct for 81% of intended policies on average, leading to a 45% reduction in the number of posts with incorrect policies when compared to Facebook's default method of suggesting the policy used on the previous post.
- We analyze in detail the circumstances under which our approach does not perform well. We discover that knowledge of intended policy combined with discarding poor-quality data yields prediction accuracies above 95%.

The rest of the paper is organized as follows: Section 2 describes related work. Section 3 presents the methodology used in the user survey and application of the machine learning algorithms. We describe the results from the survey in Section 4.1 and the results from our application of machine learning for prediction in Section 4.2. Finally, we discuss the limitations of our approach in Section 5 and conclude in Section 6.

2. RELATED WORK

We group related work into four categories: work on privacy in online social networks, coping strategies and mechanisms, automated support for policy specification, and natural language processing in the context of social media. We additionally discuss the machine-learning algorithms we use in Section 3.2.

Privacy in online social networks.

Several studies have examined users' concerns related to sharing on online social networks (OSNs). Krasnova et al. used focus groups to find that users had a broad range of worries, ranging from oversharing with friends, relatives, and coworkers to their online data being mined by corporations [19]. Besmer and Lipford examined users' concerns about sharing photos, similarly finding that social network privacy tools do not satisfactorily address users' needs [6]. Johnson et al. discovered that while the overwhelming majority of social-network users is concerned about revealing information to strangers, most users have taken steps to mitigate these concerns (e.g., by using appropriate privacy policies); on the other hand, many users also had specific concerns about sharing content with friends and acquaintances that they were not addressing as effectively [17]. Hu et al. explain how policy conflicts can arise when multiple users have a stake in the privacy policy (e.g., multiple Facebook users that are tagged in a photo) and suggest strategies for resolving such conflicts [15].

Less abstractly, studies have found that users are unsuccessful at using privacy controls and other mechanisms to avoid oversharing. Liu et al. found prevalent use of default privacy settings and a low match between privacy settings and users' expectations [23]. Other researchers found through an in-depth interview study that OSN users often overshare and regret it [40], with specific consequences ranging from temporary embarrassment to broken romantic relationships and lost jobs. The results of our user study (Section 4.1) are consistent with these findings.

Researchers have recently also looked at the effect that privacy concerns have on the continued usage of OSNs, with mixed findings. Tufecki reported finding no correlation between users' OSN sharing habits and their concerns about the privacy of their age, gender, interests, and other similar profile information [38]. Recent evidence, however, suggests that the inability of users to have confidence that their dynamic content (e.g., status updates and posts) will be shared according to their preferences is a major factor in determining the frequency of use of social networks [36].

Coping strategies and mechanisms.

In reaction to these problems, users employ a number of coping strategies beyond the features offered by OSNs. Some users moved away from broadcast content (e.g., status updates and posts) towards private messages [42]. Others maintain multiple online profiles or accounts, using each to interact with a different audience [34]. Finally, deleting friends and posts and removing tags from posts are also increasingly used [24].

Also in an attempt to mitigate these problems, OSNs have been enhanced with features that make it easier for users to set and understand privacy policies. These include Google+'s "circles," Facebook's "Smart Lists," and interfaces that allow a user to understand in detail which of her published content is visible to which other users (e.g., Facebook's "Audience View"). Researchers have also

advanced new tools and approaches, including better visualizations of friend groups and networks [25], and experimented with different policy-creation approaches, such as tag-based policy, in which policies are specified exclusively in terms of tags with which content is labeled [18]. Although all these mechanisms help, none that have been deployed have been reported to significantly mitigate users' concerns and problems with oversharing (e.g., [41]).

Automated support for policy specification.

There is a long history of using machine learning to help detect policy errors or specify policy. An early target of such analysis were firewall policies, for which tools were developed to analyze policies for consistency or the presence of specified properties (e.g., [3, 1]). Similar approaches were used to suggest to firewall administrators policies that match prespecified goals [16]. Other works used rule mining and Bayesian inference to analyze router policies and automatically detect configuration errors (e.g., [21]).

Related techniques, including rule mining, have also been used in other contexts to detect policy errors. Das et al. analyzed file-server access-control policy to detect inconsistencies in the permissions given to otherwise similar users [11]. Bauer et al. examined logs of accesses to physical space and inferred which potential accesses that are not permitted by policy are consistent with observed accesses [4].

Closer to this paper's focus, machine learning has been used to classify images uploaded to content-sharing sites and to suggest sharing policies [35]. In that work, photos are first classified according to image content, and then the resulting categories are further broken down based on descriptive tags that users attach to the photos. Our work pursues a similar goal for text content such as posts and status updates, but the specific algorithms and features used for classification in the two approaches differ.

Closely related to this paper is work by Fang et al., which addresses the problem of allowing friends access to information in a user's OSN profile [13]. This work assumes that there is an underlying privacy preference that needs to be learned. The problem we address is different, as we aim to predict the access control policy that should be applied to status messages, which involves mapping information contained in a status message to the privacy setting.

Natural language processing for social media.

Natural language processing (NLP) has been used extensively in analyzing social media content from microblogs (e.g., Twitter), OSNs, wikis, internet forums, etc. The purposes of such analyses vary, and include opinion and sentiment mining [26], determining the usefulness of product reviews [22], topic extraction [7], and examining the diffusion of information through social networks [2]. Using NLP for analyzing short text such as from microblogs and OSNs is an active area of research [31, 33]. We try two popular NLP approaches: one is a Latent Dirichlet Allocation (LDA) [7] based approach and another a classifier based on the MaxEnt (maximum entropy) [5] principle (see Section 3.2).

3. METHODOLOGY

We next describe the online study by which we collected data for training and testing machine-learning algorithms (Section 3.1), and provide an overview of our approach to predicting policy, including describing the algorithms we used (Section 3.2).

3.1 Study Design

We recruited participants through Craigslist, compensating them $5 for participation in a 20-minute survey. We required participants to be English-speaking adults, to have created at least 60 Facebook posts or status updates over the past four months, and to have used at least two different privacy settings to protect their posts.

After asking participants for demographic data, we asked them about their usage of Facebook. We asked whether they used Facebook for personal or business purposes. Using a seven-point Likert scale, we asked participants to indicate their level of agreement with statements that probed how comfortable they would be with their Facebook posts being publicly visible, whether they post personal information, and whether they believe they use strong privacy rules. We also asked whether they found Facebook's privacy controls easy to use, or confusing, and whether they had ever regretted posting content on a social network. We did not tell participants the purpose of our study, other than it was about "Facebook sharing."

Our participants were next redirected to our Facebook application. The application compiled a list of each participants' posts and their privacy settings, which were then downloaded for analysis. Study participants were given the chance to view their posts before they were downloaded, and to exclude from downloading any particularly private ones; however, no participant excluded any post. Next, 20 posts were drawn at random from the downloaded ones. These were shown to the participant, who was asked to indicate her preferred policy (i.e., privacy settings) for each post, by selecting from among any of the policies the participant had used for any post, or "other," with the option of entering any policy. While eliciting the preferred policy for most posts would have given us more ground truth data about policies, the extra burden on participants would have increased the chance of participants providing incorrect data or dropping out of the study.

Our study design was reviewed and approved by our institutional review board (IRB).

3.2 Predicting Policies

The goal of our work is to predict, based on the content and context of a new Facebook post, the access-control policy that a user will want to apply to it. We envision that a tool that performed such prediction could be integrated with OSNs, and would suggest to users at the time of content creation the policy, or several possible policies, with which to protect the new content. In this paper we focus on determining whether and to what extent such a tool could suggest policies that match users' sharing intentions.

Facebook's privacy policies for posts specify the audience as a set of users. The set of users is specified with the help of predefined sets such as "public" (anyone, including people without Facebook accounts), "friends of friends" (all Facebook users who are friends of the friends of the user under consideration), and friend lists. A friend list is a subset of all the user's friends, and is either specified manually by the user or is automatically generated by Facebook. In addition, the user may specify a "deny" and "allow" list of friends that further restricts or expands the intended audience for posts. In our work, we transform each policy to a string (see Section 3.2.2), which is then considered the label of that post.

The high-level approach we pursue is to train a machine-learning algorithm on posts that are annotated with such labels, and then use the trained algorithm to predict the policy of test posts. Our training and test data are derived from the posts and policies downloaded from study participants' Facebook accounts. Because users strongly differ in the kind of content they publish and the sharing policies they apply to it, we do not aggregate data across users—we make predictions about a participant's posts (test data) using an algorithm trained only on that participant's previous posts (but excluding the posts we test on).

We choose test data for each participant at random from the chronological sequence of posts (or from a subset for which the intended policy is known—see Section 4.2). For each test data point, we build the learning model using training data that appears chronologically before the test data point. We try to maximize *average accuracy*, the average success of prediction of the learned model—a different model for each test data point—on the test data. As is customary for small datasets, we select features by performing multi-fold cross validation in the training phase. We try all combinations of the features that we consider, in addition to the basic feature of words in the post (see Section 4.2.3). We repeat 100 times the process of partitioning the training set of posts into 90% model-building data and 10% evaluation data for each model (set of features) under consideration; the *training* accuracy for a model is the average over all 100 runs on a single participant's data. Finally, the model with the best training accuracy is selected for evaluation over test data.

3.2.1 Machine-learning algorithms

Facebook posts and status messages (we use the two terms interchangeably) almost always contain some user-generated text. Hence, a natural approach for analyzing these is to use tools from natural language processing (NLP).

Latent Dirichlet allocation (LDA) is a popular approach for extracting topics from documents [7]. The topic of a document is a collection of words that represents the theme of the document. Topic extraction refers to predicting the topic of a document using words present in the document. An extension of LDA called *labeled LDA* (LLDA) can be used to infer labels for documents when provided with training data of labeled documents [32]. We treated posts as documents and their policies as labels, and used LLDA to predict the policies for a test set of posts that did not have policies attached. However, we found the performance of LLDA to be poor in our setting. We attribute this to two factors. First, any LDA-based approach requires a substantial amount of training data [30]. The size and number of posts, not just at our disposal for this experiment but in general, is small compared to text documents typically used in LDA-based approaches. Short-text topic extraction is an active area of research [31, 33], but, based on our experiments, has not advanced to the point where it would allow an LDA-based approach to be effective on the short messages typical of our domain of interest. Second, *classifiers* (a class of machine-learning tools) are comparable to LLDA for text classification [32], and can additionally use non-text features (e.g., the time of the post) to aid in classification. Consistently with this, we found LLDA not to be competitive with a classifier-based approach, described next.

A classifier often used in NLP with good results, and which we apply to our problem, is the *MaxEnt classifier* [5]. A classifier attempts to predict the correct label for a given test dataset. A classifier takes as input a labeled training dataset, represents the data as a vector in an n-dimensional space, and then learns a separator that best separates the data by labels. The learned separator is used to predict the labels of the test data points. The n dimensions of the data points are called features. A classifier allows the use of non-text features of posts, such as creation time, the presence of a URL, post length, etc. The MaxEnt classifier is based on the idea of obtaining the separator that maximizes entropy (and hence yields low generalization error, i.e., low error on test data), but constrained to achieving a low empirical error (performing well on training data). The MaxEnt model assumes nothing beyond the information provided by the training data, which enables the classifier to perform well on a randomly chosen test dataset. The MaxEnt model is equivalent to the Multinomial Logistic Regression model, another model for classification in NLP [5]. Since we found the MaxEnt model to perform strictly better than other approaches we tried for our scenario, the results presented in this paper are based on the MaxEnt classifier.

We used the LLDA and MaxEnt tools created by the NLP Group at Stanford (`http://nlp.stanford.edu/software/`). We ran the tools using default settings, except the use of word bi-grams, i.e., pairs of adjacent words, as a feature. Our parsing and result computation code was written in Scala, with some analysis using Microsoft Excel.

3.2.2 Data transformation

Next, we discuss how we transformed the content of participants' posts before using them as input to machine-learning algorithms. As is common in NLP, we make all words lower case, remove punctuation, and filter out *stop words*, which are are common words like "a," "an," "it," and "I" that convey little meaning. We used a standard list of 120 stop words [37]. Filtering out stop words may result in empty posts and hence decreases the amount of training data, but generally increases prediction accuracy. For our dataset, this removed three posts from just one user.

We also perform certain intuitive modifications on the data, after empirically validating through preliminary experiments that these improve the effectiveness of prediction. We replace URLs by only the domain name, e.g., `http://my.com/test` becomes `my.com`. We also transform text, like emoticons and some punctuation, that is otherwise parsed in a semantically incorrect way. We replace emoticons by a corresponding descriptive word, e.g., ":)" and ":-)" are replaced by "happy." We replace question marks ("?") by the word "question," two or more contiguous exclamation marks ("!") by "exclaim," and two or more contiguous periods (".") by "continued." The exact words used for replacement do matter; even though the machine-learning tools do not understand the semantics of words, the user may use those same words in some posts instead of or independently from emoticons. Finally, we filtered out posts of fewer than three words, as these usually do not have enough information to perform inference.

As stated earlier, for each participant, we map each privacy policy for a post to a unique string, which serves as a label for the post. This is done by using symbolic strings for the three sets "public," "friends of friends," and "all friends." Any allow or deny list of friends is taken into account by concatenating the symbolic string with the sorted list of friend IDs specified in the allow or deny list. For any privacy policy specified using friend lists, we use the concatenated, sorted list of IDs of friends in that list as the label, after adding or removing friends from any allow or deny lists.

4. RESULTS

We first present the results of the survey that solicited demographic data and usage habits and sentiment about Facebook (Section 4.1). We then present our results from the use of machine learning for policy prediction (Section 4.2).

4.1 Demographics and Survey

Demographic data.

From July to November 2012, 46 participants took our study and met our requirements. The data for four participants is missing due to data corruption; we present here the self-reported demographic data from the remaining 42.

Our participants ranged in age from 18 to 65, with an average of 29.1 and median 25.5. Thirty five participants reported being female, and 7 reported male. Sixteen participants reported a college

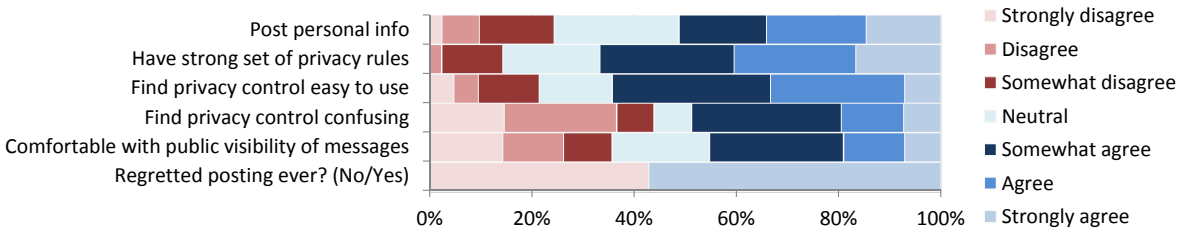

Figure 1: Participants' answers when they were asked to agree or disagree with statements about behavior on and sentiment toward Facebook (only selected questions shown).

degree as their highest achieved degree, 16 reported finishing high-school, and four receiving an advanced degree. Four participants reported majoring or having a degree or job in "computer science, computer engineering, information technology, or a related field;" 36 did not, and two did not provide an answer. One participant was from Singapore, one from Guatemala, one from Russia, and the rest were from the U.S.A.

Facebook usage and sentiment.

Asked whether they found Facebook's privacy controls "confusing," participants answers were almost evenly distributed from "strongly disagree" to "strongly agree" on a 7-point Likert scale, with a small majority expressing some level of agreement. However, presented with the statement that Facebook's privacy controls are "easy to use," 27 participants agreed and only nine disagreed. Somewhat surprising is that participants who self-identified as having good computer skills were almost evenly divided on these questions, suggesting that computer-savvy participants were less likely than other participants to find Facebook's privacy controls easy to use.

Asked whether they have a "strong set of privacy rules" on Facebook, 28 participants replied neutrally or positively and there was no strongly negative reply, indicating that most participants felt they used average or stronger-than-average privacy policies on Facebook. Asked whether they tended to post personal information on Facebook, 21 participants indicated that they did not.

Asked whether they were comfortable with their status message being visible to "everyone on the internet" on a 7-point scale from "not at all comfortable" to "very comfortable," 19 participants expressed various levels of comfort, 15 expressed various levels of discomfort, and 8 were neutral. We found this level of comfort with public sharing surprising, as well as inconsistent with both the policies our participants implemented or wished to implement and the negative experiences they reported with sharing data on Facebook. This disconnect highlights the need for tools that remind participants to set a more restrictive policy for posts for which an overly open policy may lead to later discomfort or regret.

Of our 42 participants, 25 answered positively when asked whether they had "ever posted something on a social network and then regretted doing it," which is more participants than reported being uncomfortable sharing their posts with everyone on the Internet. Reasons that participants reported ranged from friends getting offended to people not thinking enough before posting from a smartphone. One participant stated that he is an attorney and exercised significant discretion in posting messages, and thus never had any mishaps. Another participant believed that expressing her thoughts will not have a negative impact on her life, and hence she does not regret anything she ever posted. Figure 1 depicts answers to select questions on user sentiment.

Table 1: Confidence factor for 42 participants.

Confidence factor	No. of users
0–0.25	7
0.26–0.50	7
0.51–0.75	12
0.76–1.00	16

Not counting automatically generated messages (e.g., "check-ins" from location-tracking applications), our participants created 6150 posts (average 146 per person, std. dev. 89.3, median 117.5) in the four months preceding their participation in our survey. Although participants could exclude specific posts from being submitted to us, no participant excluded any post. We excluded 17 posts from our analysis, because Facebook's API returned an empty privacy policy for them. Participants used between two and 35 policies to protect their posts, with an average of 4.6 (std. dev. 5.4) and median of 3.

Confidence about policies.

As explained in Section 3.1, we randomly selected 20 posts from each study participant and interactively elicited from the participant the policy she desired to implement (the *intended* policy) for each. The purpose of this was to measure how often participants either changed their minds about policies after implementing them or had not implemented their intended policies. We refer to the fraction of the 20 that matched the implemented policy as the *confidence factor*. For example, if 15 of the 20 interactively elicited (intended) policies matched the user's intended policies for those messages, the confidence factor is .75.

Table 1 shows the counts of participants by confidence factor ranges. Fourteen participants changed their mind about or incorrectly implemented the policy for more than 50% of the posts shown to them, and only 16 participants have 75% or more of their posts labeled correctly on Facebook. This phenomenon results in noisy data for the learning tools; in Section 4.2.4 we show its effect on the accuracy of learning.

4.2 Policy Prediction

In this section we present our results on using machine learning to predict policy. We first describe the datasets and the success criteria for evaluating the accuracy of predictions (Section 4.2.1). We then describe the prediction mechanisms that we use as a baseline (Section 4.2.2) and the features used with the MaxEnt classifier (Section 4.2.3). Finally, we discuss the effectiveness of our prediction strategy (Section 4.2.4).

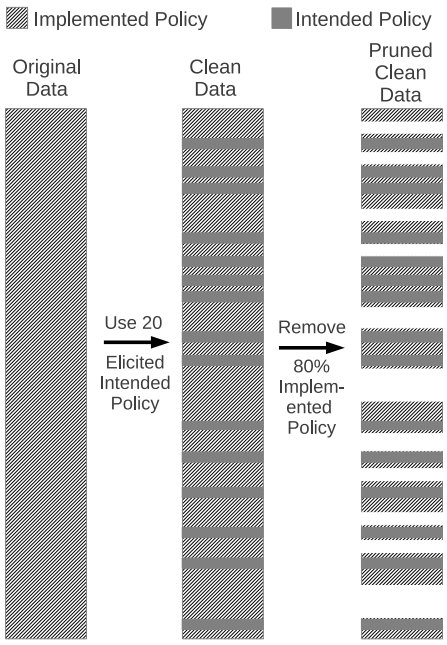

Figure 2: Datasets used in evaluation and their construction.

4.2.1 Datasets and Success Metrics

Datasets.

To investigate the effects of noise (i.e., erroneous policies) in training data on prediction performance, we evaluate on three datasets.

1. *Original:* The *original* dataset is comprised of the posts and policies downloaded from participants' Facebook accounts. The policies in this dataset match what participants implemented in Facebook, even when they did so by mistake or later changed their mind.

2. *Clean:* The *clean* dataset is constructed from the original dataset by replacing the downloaded policies for 20 posts with the corresponding interactively elicited policies. In other words, in the clean dataset 20 potentially erroneous entries have been corrected using the interactively elicited data; remaining entries are the same as in the original dataset.

3. *Pruned clean*: The *pruned clean* dataset is constructed from the clean dataset by discarding the 80% of the unconfirmed, possibly erroneous data, leaving the corrected 20 entries as is in the remaining 20% data.

Figure 2 illustrates the creation and contents of these datasets.

When we test on the clean and pruned clean datasets, we test on data drawn solely from the 20 policies per participant that were manually elicited, as we consider these to be closest to ground truth. We discuss this further in Section 4.2.2.

Success metrics.

We use two metrics to measure the success of a prediction.

1. *Exact*: The prediction is counted as successful when the predicted policy is the same as the policy that the user assigned to the post.

2. *Top-Two*: The prediction is counted as successful when either of the two policies predicted as most likely is the policy that the user assigned to the post. This simulates a scenario in which the prediction algorithm is used to generate two policy suggestions for each message.

The ratio of successful predictions to the total number of predictions is the *accuracy* of the prediction strategy. By definition, the Top-Two metric always outperforms the Exact metric in this respect. Although we report performance according to both metrics, we believe the Exact metric (on which we focus when reporting summary results) is more realistic, since it would likely yield a more usable tool in practice and can be more fairly compared to Facebook's default approach.

4.2.2 Baseline Prediction Strategies

Next, we present the two prediction strategies that we use as baselines. These are straightforward ways to predict policy and involve no machine learning. We consider the following baselines:

1. *Previous Policy*: For the current post, suggest the policy used on the last post.

2. *Most Common*: Always suggest the policy used most commonly by the participant in the four-month span for which we have data.

We report the accuracies of the baseline strategies for two datasets (original and clean), as well as reporting the *intended accuracy*, where predictions are restricted to the 20 posts for which the intended policy is known. The accuracy for each participant for the original and the clean dataset is computed by iterating over all posts, predicting a policy for each post using the baseline approach, and counting the number of successful predictions. The intended accuracy is computed similarly, but the iteration is only over the 20 corrected posts for which the intended policy is known.

Facebook currently implements the Previous Policy approach. This approach performs well on the original dataset (85% average accuracy), slightly worse on the clean dataset (80% average accuracy), but has poorer accuracy when tested against intended policy (67% on average), which is the measure we consider most meaningful, since for that measure all test data points are confirmed to be accurate. The performance of Most Common (79%, 77%, 73%) follows a similar trend. Table 2 shows these results in more detail.

The Top-Two metric clearly results in a greater fraction of policies predicted correctly, since its criteria for correctness is less stringent than that of the Exact metric. We speculate that the high accuracy of Previous Policy (for the original dataset) can be attributed to the fact that many participants have a dominant policy, and often do not change their policy over a set of contiguous posts. Using the same policy for consecutive posts may make sense semantically, e.g., if all the posts in the set are on the same topic. On the other hand, it could also reflect policy errors caused by participants' reluctance to change their policy from the default offered by Facebook. Indeed, the lower accuracy when evaluating against intended policy reveals a mismatch between Facebook's predictions and intended policy for many participants.

We also experimented with another success metric: whether the predicted policy was at least as restrictive as the desired policy. The accuracy with such a success metric was quite high—0.94 with the Previous Policy baseline—which is not unexpected due to the coarse nature of this success metric. The use of the MaxEnt classifier did not yield much better results than 0.94; thus, we do not analyze this success metric further in this paper. Also, while oversharing is clearly a problem, undersharing diminishes the utility of sharing. Hence, we focus mainly on exact matches.

Table 2: Average performance of baseline approaches for 42 participants.

Baseline metric	Avg. accuracy (original dataset)	Avg. accuracy (clean dataset)	Avg. intended accuracy (clean dataset)
Previous Policy	0.85	0.80	0.67
Most Common	0.79	0.77	0.73
Previous Policy (Top-Two)	0.91	0.88	0.74
Most Common (Top-Two)	0.88	0.85	0.85

4.2.3 Features Available for Model Selection

Before describing results in detail, we discuss the features available to the learning algorithms. The subset of these that was used in any specific model (recall that a sequence of models was created for each participant, as described in Section 3.2) is derived by trying all combinations of features and using cross-validation to select the best combination. In addition to the words of a post, which is a feature included in all models, other features were post creation time, the presence of attachments, word n-grams, and the policy of the post immediately preceding the post for which a model was being generated.

The basic feature we used was the words that made up the content of posts. The MaxEnt classifier tool we used automatically converts words into a vector representation, although we first manually remove stop words and perform other low-level transformations (see Section 3.2).

We examined several bucketing strategies to represent the time when a post is created. The most effective strategy, on which we settled, classified posts into three buckets: (1) *office*: between 9AM and 6PM on weekdays, (2) *evenings and weekends*: between 6PM and midnight daily and between 9AM and 6PM on weekends, and (3) *nights*: midnight to 9AM daily. We also experimented with six buckets of four hours each, but with less success.

We encoded as a binary flag whether a post included an attachment such as an image, video, or a URL (which may be different from URLs contained in the text of a post).

We additionally used word 2-grams as a feature, as often the semantics of a sentence is better captured by pair of adjacent words than by each word in isolation. For example, if a post includes the words "Sherlock" and "Holmes," then in addition to each of those words being used as a feature, the bi-gram "Sherlock Holmes" is also used.

Finally, we included the policy of the previous post as a feature, hypothesizing that for some participants this may be a good indicator of the policy for the current post.

Another feature we experimented with, but do not use in the results we report, was the number of words in posts. Including this feature reduced the accuracy on the pruned clean dataset by 6%. We conjecture this is because of the large number (compared to the size of pruned clean dataset) of possible values for this feature; this can cause the learned model to over-fit the training data, consequently reducing testing accuracy.

The MaxEnt tool performs a standard transformation on the words, by mapping each word (or bi-gram) to a dimension to build a multi-dimensional dataset with 0's and 1's. The zeros indicate absence of a word and vice versa. The other features described above are all categorical variables, each adding a dimension to this dataset.

4.2.4 Policy Prediction Accuracy

As discussed in Section 4.2.1, we analyze the performance of policy prediction using the original dataset, the clean dataset (in which 20 of a participants' implemented Facebook policies have been substituted with those interactively elicited from her), and the pruned clean dataset. When analyzing performance on the original dataset, we randomly pick eight policies to use as test data and calculate accuracy as described in Section 3.2; these eight are never used for training. When using the clean and pruned clean datasets, we choose the test data from among the 20 policies elicited from the participant.

Table 3: Average accuracy for baselines and MaxEnt with Exact and Top-Two success metrics, using the clean and pruned clean datasets.

Prediction strategy	Avg. intended accuracy (Exact)	(Top-Two)
Previous Policy	0.67	0.74
Most Common	0.73	0.85
MaxEnt (clean dataset)	0.71	0.94
MaxEnt (pruned clean dataset)	0.81	0.94

Our experiments using the LLDA tool show strictly less accurate performance than using MaxEnt, by a significant amount: on the original dataset the average accuracy of predictions made using LLDA was 0.62 using words only—including other features is not possible with LLDA in a straightforward fashion. Hence, we focus on results obtained by using the MaxEnt classifier.

Accuracy with original dataset.

The accuracy of the MaxEnt classifier on the original dataset is 0.86 for the Exact and 0.94 for the Top-Two metric. When making a single prediction per candidate post (i.e., Exact), the average accuracy of the MaxEnt classifier on the original dataset is only 1% better than the better of the two baselines (Table 2). We believe this is because the trends in the original dataset are largely the result of Facebook's use of the Previous Policy strategy. The MaxEnt tool successfully picks up these trends in the original dataset. However, learning the trend in the original dataset does not help make good predictions of *intended* policy, since the trends are dominated by inaccurate data; we show this next.

Accuracy with clean and pruned clean dataset.

More indicative of real-world effectiveness is prediction accuracy measured with respect to the clean and pruned clean datasets (Table 3). For both these datasets, the test data is chosen from among the 20 posts with correct intended policy. In this setting, the MaxEnt classifier has slightly higher accuracy (71%) on the clean dataset than does the Previous Policy baseline (67%) for the Exact case. The corresponding MaxEnt accuracy (Exact) on the

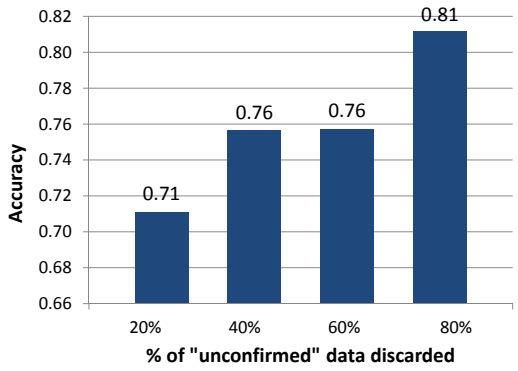

Figure 3: Average intended accuracy of MaxEnt using the Exact metric when different fractions of unconfirmed training data are discarded.

pruned clean dataset (in which some potentially erroneous data is discarded) is substantially better (81%), yielding more than 45% fewer incorrect predictions than the baseline.

We attribute the differences in the performance of MaxEnt across our datasets to the mismatch between a participant's implemented policy and her intended policy. Accuracy on the clean dataset is not much greater than on the original one because the vast majority of training data is unconfirmed, and hence the intended policy is not successfully learned from it. Testing on the pruned clean dataset reveals that when some unconfirmed data is removed, increasing the fraction of good data, prediction accuracy increases.

To further test this hypothesis, we experimented with changing the fraction of clean training data by discarding varying fractions of the unconfirmed data; this makes each training set smaller, but increases the proportion of clean data. As the fraction of clean data in the training set grows, prediction accuracy improves (Figure 3). Decreasing the size of the training set would normally be expected to decrease accuracy; in this case, that is more than made up for by the increase in the quality of the training data as the training set shrinks.

We next further investigate the dependence of prediction accuracy on the quality of the training data.

Accuracy vs. confidence factor.

Recall that a low confidence factor means that the 20 intended policies supplied by participants were frequently *not* the policies that were actually implemented for the corresponding posts on Facebook. An examination of the (clean and pruned clean datasets) results by subsets of participants split by confidence factor yields an interesting observation—for participants with high confidence in their policies, the MaxEnt classifier has substantially better average accuracy than the baseline approaches (Table 4).

This provides further support for the behavior we had previously observed: when trained on good data, MaxEnt predicts policies more accurately than the baseline strategies. More specifically, the participants with a low confidence factor are those whose Facebook policies typically *do not* correspond to their intentions. Since the accuracy of prediction is measured with respect to intended policy (interactively elicited during the study), it is no surprise that training on data that is highly inconsistent with intended policy does not allow accurate prediction of intended policy. From the standpoint of learning, a low confidence factor implies that there is significant

Table 5: Result of excluding features with pruned clean dataset. The last column shows the number of users for whom prediction was most accurate with the set of features described by each row. The first row shows performance with no feature excluded.

Excluded Features				Avg. intended accuracy	# users w/ best acc.†
attachment	n-gram	last policy	time		
				0.81	25
×				0.76	9
×	×			0.79	14
×	×	×		0.75	9
×	×	×	×	0.79	11

† There is overlap in the sets of users that characterize best accuracy of each set of features. Hence, sum of last column > 42.

noise in the labels (policies) present in the training data; machine-learning algorithms perform poorly at classification as the fraction of noise in labels increases [20]. Ignoring the 14 (33%) participants who have confidence factor lower than 0.5, the improvement in accuracy for the remaining 28 (66%) participants is on an average 0.14 and the resulting average accuracy is 0.94. This accuracy is comparable to the $\sim$94% accuracy reported by NLP topic modeling tools like LDA [7]. Additionally, Table 4 also shows that discarding 80% of the unconfirmed data (pruned clean dataset) results in improved predictions for almost all users, including those with a low confidence factor. Notably, the accuracy for users in the lowest confidence class, even though low in absolute terms, was more than double the accuracy of the Facebook baseline (0.25 to 0.52). In other words, although greatest absolute accuracy is achieved for users with a high confidence factor, on average even users with a low confidence factor benefit from our predictions.

Accuracy by feature.

Perhaps surprisingly, we found that the average prediction accuracy was affected only slightly by excluding various features from our models (Table 5)—the majority of the benefit stems from analyzing the text of posts. Interestingly, the decrease in accuracy was not monotonic as features were removed, showing that their interplay is subtle. We believe that this is because individual users benefited differently from the various features. This is supported by the data in the last column of Table 5, which shows the number of users for whom the feature set under consideration performed best. All the features together perform best for the largest number of users (25); smaller sets of feature perform best for different numbers of users, ranging from 9 to 14.

Prediction accuracy over time.

As described in Section 3.2, we train our prediction algorithm to mimic how a real-time prediction tool could be used: only posts that appeared chronologically before an element (post) of the test set are used to train the model used to make predictions for that test post. Figure 4 shows the average (over all participants) accuracy of MaxEnt for predicting each of the eight elements of the test set in chronological order. Accuracy is initially poor, as only $\sim$12.5% of the training data is available when the first prediction is made. Perhaps surprisingly, however, accuracy almost immediately rises to the range where it will remain. This suggests that, in practice, the accuracy of prediction can be improved by not making predictions when too little data is available, but also paints a surprisingly optimistic picture of how little training data is needed to achieve reasonable prediction accuracy. Delaying predictions in this way

Table 4: Prediction accuracy of Previous Policy and MaxEnt (on the clean and pruned clean datasets), using the Exact metric, split by Confidence factor.

Confidence factor	Previous Policy avg. intended acc.	MaxEnt avg. intended acc. and # users with better acc. (clean dataset)		MaxEnt avg. intended acc. and # users with better acc. (pruned clean dataset)	
0–0.25	0.25	0.23	(2 of 7)	0.52	(6 of 7)
0.26–0.50	0.58	0.62	(4 of 7)	0.58	(3 of 7)
0.51–0.75	0.62	0.68	(6 of 12)	0.88	(10 of 12)
0.76–1.00	0.91	0.96	(14 of 16)	0.98	(14 of 16)

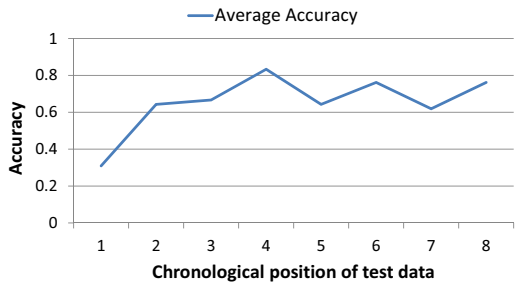

Figure 4: Average intended accuracy of MaxEnt using the Exact metric on the clean dataset, for each element of the test set in chronological order (e.g., "1" indicates the chronologically first element of the test set; "8" the chronologically last).

would hence also result in higher accuracy than we reported earlier in this section.

Correlating data characteristics with accuracy.

To more precisely characterize the apparent differences in the performance of prediction based on the confidence factor, we compute the Pearson correlation coefficient, which is a standard measure for determining linear dependence between two variables. We found the Pearson correlation of the accuracy of MaxEnt classifier (pruned clean dataset) with the confidence factor to be 0.67 ($p < .001$), indicating a strong correlation (correlation values of magnitude greater than 0.5 are held to mean strong correlation in behavioral sciences [9]).

Another factor we considered to better understand the performance of MaxEnt on a per-user basis relates to those privacy policies that were each applied by a participant to fewer than 30 posts (*sparse* policies). The *sparse class factor* is the fraction of all policies used by a participant that are sparse, e.g., for a user with two policies a sparse class factor of 0.5 means that one of the two policies was used for less than 30 posts. We consider this factor because classifiers are known to perform poorly if the amount of training data for a particular label (class) is small and those labels show up in test data [29]. While the correlation of the sparse class factor with the difference between MaxEnt classifier accuracy and baseline accuracy is not significant, the sparse class factor for the three participants for whom prediction accuracy was worst is high (Table 6). Their average sparse class factor is 0.84, indicating that these participants use a dominant policy, and use other policies infrequently enough for this to pose a problem for prediction.

Table 6: Confidence factor and sparse class factor for the three participants for whom MaxEnt performed worst (using the Exact metric, on pruned clean dataset).

Participant ID	Accuracy	Confidence factor	Sparse class factor
2	0.16	0.50	0.80
19	0.25	0.20	0.81
20	0.28	0.15	0.91

The correlation between accuracy and confidence factor, and the high sparse factor of participants for whom prediction performs poorly, demonstrate known behavior of classifier performance with noisy labeling (policies, in our case) and low amounts of data per label: the presence of either factor reduces the accuracy of our predictions with the clean dataset. For existing Facebook posts it is reasonable to expect noisy labeling, as our survey and other results (e.g., [23]) have shown. Also, many participants in our survey used a dominant policy, and hence other policies (labels) were associated with small amount of data. The latter factor is likely to remain a problem for prediction regardless of the accuracy of the training data: if very few posts use a specific policy, predicting that policy is unlikely to be accurate. The former factor, however—the noisiness of the data—is one that could be mitigated in practice if prediction were integrated into Facebook, e.g., by better interfaces for specifying policy, or by periodic, interactive spot-checking of the accuracy of policies. Indeed, when used in practice a prediction mechanism such as the one we explore here would lead to more accurate policies, in turn providing more accurate training data and hence further improved prediction.

5. LIMITATIONS

This section discusses several limitations of our approach, including those related to noisy data, data collection methodology, our focus on text content, and sample size.

Our survey results show that the policies chosen by participants are often not reflective of the intended policy, both impacting the performance of machine-learning algorithms and making it challenging to interpret results. We address the problem caused by this inaccuracy in two ways: (1) by discarding some noisy data, and (2) by evaluating performance on more accurate and less accurate data independently. We use participants' inputs to correct their privacy policy, which relies on the critical assumption that users are able to properly assign privacy policies during the survey.

Since the accuracy of policies attached to posts used for both training and testing is important for precise evaluation, a natural

question is whether true intended policies can be collected for all posts created by our participants. Unfortunately, this was not feasible given our survey methodology: Participants in online surveys are typically less motivated than participants in laboratory or other in-person studies. Hence, our study design had to strike a balance between collecting enough data for analysis and asking few enough questions that participants would not lose focus and start providing incorrect answers. As future work, we are considering alternative data-collection methodologies, including long-running studies in which participants' are periodically asked, over weeks or months, about the posts they have recently created. Another alternative is in-person surveys structured similarly to the online survey described in this paper. Both of these would involve substantial additional effort, and so are out of scope for this paper.

Our use of machine learning focused on the text content of posts. Although we included as features to be used for learning the presence of an attachment or URL, we did not analyze the content of such attachments. Doing so, particularly for images included in posts, would have required the use of different techniques from the ones we utilized. Research that focused solely on images has found them to be amenable to classification that could potentially help in setting privacy policies [35]. This suggests that including both image and text analysis to make policy predictions is a promising direction to explore; we defer it to future work.

Our findings are based on the 42 participants who took our survey, which is a small sample size when compared to the total number of Facebook users, leading to questions about the generalizability of our results. While a larger sample is desirable, methodology and resource constraints made that infeasible. Such limitations are unfortunately common; many Facebook human subjects experiments, including some thematically close to ours, are restricted to samples of 20–60 users (e.g., [25, 12, 27, 39, 14, 8]). Because our use of machine learning utilized data on a per-user basis (i.e., training and testing took place on a single individual's data), a larger sample size would not have directly benefited the efficacy of learning, but it would increase the likelihood that the sample was representative of the Facebook population. Our sample did, however, cover a wide range both in terms of demographics and in terms of the accuracy of users' policies (i.e., noisiness of data); hence, even given the limitations of our sample size, we believe that our findings strongly suggest that applying techniques like the ones we study in this paper could aid vast numbers of social network users in setting their privacy policies.

6. CONCLUSION

Research has shown, and our results confirm, that Facebook users often fail to judiciously set the privacy policies for their posts [23]. We demonstrate that machine learning can accurately predict, using existing data, the policies with which users want to protect their Facebook posts. Using the MaxEnt classifier we predict policies with an average accuracy of 81%, compared to the 67% average accuracy obtained by following the default Facebook strategy of suggesting the policy that was used for the last post. In other words, our application of machine learning to policy prediction reduces the number of posts for which policy is misconfigured by over 45% on average. Moreover, for users whose implemented Facebook privacy policies are consistent with their intentions for more than 50% of their posts, the MaxEnt classifier correctly predicts policies for 94% of new posts. Hence, any practical use of machine learning for an existing system would benefit from taking into account the difference between implemented and intended policies.

Primary challenges in applying machine learning to predicting intended policies seem to be inaccuracies in implemented policies and, for some policies, a low number of posts per policy. We plan to apply other machine-learning techniques to this problem. One is "upsampling," in which data is selectively duplicated to mitigate the problem of low data in a class. Another is using a multiclass SVM classifier, which is also often used in NLP applications.

While our work provides insights about privacy policy management on Facebook, a number of interesting open questions remain. For example, is there more than one "correct" privacy policy for a particular post? It may be that a user would be equally satisfied with any of several privacy policies being used to protect a particular post. Another question is whether categorization of posts using a big knowledge base, e.g., Wikipedia, can improve the performance of machine learning. Wikipedia can be used to extract categories of words, revealing commonalities in meaning that are not obvious from a purely syntactic interpretation of text; this technique has been used previously in NLP applications [43]. These and other questions form fertile ground for future work.

7. ACKNOWLEDGMENTS

This research was supported in part by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office; by Carnegie Mellon CyLab under Army Research Office grants DAAD19-02-1-0389 and W911NF-09-1-0273; by NSF grant CNS-0831407; and by the AFOSR MURI on Collaborative Policies and Assured Information Sharing. A. Sinha was also partially supported by a CMU CIT Bertucci Fellowship.

8. REFERENCES

[1] E. Al-Shaer and H. H. Hamed. Discovery of policy anomalies in distributed firewalls. In *Proc. INFOCOM*, 2004.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on Twitter. In *Proc. ACM International Conference on Web Search and Data Mining*, 2011.

[3] Y. Bartal, A. J. Mayer, K. Nissim, and A. Wool. Firmato: A novel firewall management toolkit. In *Proc. IEEE Symposium on Security and Privacy*, 1999.

[4] L. Bauer, S. Garriss, and M. K. Reiter. Detecting and resolving policy misconfigurations in access-control systems. *ACM TISSEC*, 14(1), 2011.

[5] A. L. Berger, S. D. Pietra, and V. J. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[6] A. Besmer and H. Richter Lipford. Moving beyond untagging: Photo privacy in a tagged world. In *Proc. CHI*, 2010.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] E. Chin, A. P. Felt, V. Sekar, and D. Wagner. Measuring user confidence in smartphone security and privacy. In *Proc. SOUPS*, 2012.

[9] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Psychology Press, 1988.

[10] Daily Mail. Not so sweet 16: Birthday girl goes into hiding after 1,500 Facebook users turn up for her party. `http://www.dailymail.co.uk/news/article-1394536`, 6 June 2011. [accessed 14-Feb-2013].

[11] T. Das, R. Bhagwan, and P. Naldurg. Baaz: A system for detecting access control misconfigurations. In *Proc. USENIX Security Symposium*, 2010.

[12] M. de Sa, V. Navalpakkam, and E. F. Churchill. Mobile advertising: evaluating the effects of animation, user and content relevance. In *Proc. CHI*, 2013.

[13] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *Proc. WWW*, 2010.

[14] E. Hayashi, O. Riva, K. Strauss, A. J. B. Brush, and S. Schechter. Goldilocks and the two mobile devices: going beyond all-or-nothing access to a device's applications. In *Proc. SOUPS*, 2012.

[15] H. Hu, G.-J. Ahn, and J. Jorgensen. Detecting and resolving privacy conflicts for collaborative data sharing in online social networks. In *Proc. ACSAC*, 2011.

[16] T. Jaeger, A. Edwards, and X. Zhang. Policy management using access control spaces. *ACM Transactions on Information and System Security*, 6(3):327–364, 2003.

[17] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and privacy: It's complicated. In *Proc. SOUPS*, 2012.

[18] P. F. Klemperer, Y. Liang, M. L. Mazurek, M. Sleeper, B. Ur, L. Bauer, L. F. Cranor, N. Gupta, and M. K. Reiter. Tag, you can see it! Using tags for access control in photo sharing. In *Proc. CHI*, 2012.

[19] H. Krasnova, O. Günther, S. Spiekermann, and K. Koroleva. Privacy concerns and identity in online social networks. *Identity in the Information Society*, 2:39–63, 2009.

[20] C. P. Lam and D. G. Stork. Evaluating classifiers by means of test data with noisy labels. In *Proc. 18th International Joint Conference on Artificial intelligence*, 2003.

[21] F. Le, S. Lee, T. Wong, H. Kim, and D. Newcomb. Detecting network-wide and router-specific misconfigurations through data mining. *IEEE/ACM Transactions on Networking*, 17(1):66–79, 2009.

[22] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *Proc. EMNLP-CoNLL*, 2007.

[23] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook privacy settings: user expectations vs. reality. In *Proc. IMC*, 2011.

[24] M. Madden. Privacy management on social media sites. `http://pewinternet.org/Reports/2012/Privacy-management-on-social-media.aspx`, Feb. 2012. [accessed 14-Feb-2013].

[25] A. Mazzia, K. LeFevre, and E. Adar. The PViz comprehension tool for social network privacy settings. In *Proc. SOUPS*, 2012.

[26] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008.

[27] T. Paul, M. Stopczynski, D. Puscher, M. Volkamer, and T. Strufe. C4PS - helping Facebookers manage their privacy settings. In *Proc. 4th international conference on Social Informatics*, SocInfo'12, 2012.

[28] E. Phneah. Japan govt used wrong privacy settings in Google Groups. *ZDNet*, 11 July 2013. `http://www.zdnet.com/japan-govt-used-wrong-privacy-settings-in-google-groups-7000017923/` [accessed 20-Jul-2013].

[29] F. Provost. Machine learning from imbalanced data sets 101 (extended abstract). In *Proc. AAAI Workshop on Imbalanced Data Sets*, 2000.

[30] K. Puniyani, J. Eisenstein, S. Cohen, and E. P. Xing. Social links from latent topics in Microblogs. In *Proc. NAACL HLT Workshop on Computational Linguistics in a World of Social Media*, 2010.

[31] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *Proc. ICWSM*, 2010.

[32] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. EMNLP*, 2009.

[33] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of Twitter conversations. In *Proc. HLT-NAACL*, 2010.

[34] M. Skeels and J. Grudin. When social networks cross boundaries: a case study of workplace use of Facebook and Linkedin. In *Proc. GROUP*, 2009.

[35] A. C. Squicciarini, S. Sundareswaran, D. Lin, and J. Wede. A3P: adaptive policy prediction for shared images over popular content sharing sites. In *Proc. Hypertext*, 2011.

[36] J. Staddon, D. Huffaker, L. Brown, and A. Sedley. Are privacy concerns a turn-off? Engagement and privacy in social networks. In *Proc. SOUPS*, 2012.

[37] Text Fixer. Common English Words List. `http://www.textfixer.com/resources/common-english-words.txt`. [accessed Feb-14-2013].

[38] Z. Tufekci. Can you see me now? Audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, 28(1):20–36, 2008.

[39] Y. Wang, P. G. Leon, K. Scott, X. Chen, A. Acquisti, and L. F. Cranor. Privacy nudges for social media: an exploratory Facebook study. In *Proc. WWW*, 2013.

[40] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. "I regretted the minute I pressed share": a qualitative study of regrets on Facebook. In *Proc. SOUPS*, 2011.

[41] J. Watson, A. Besmer, and H. R. Lipford. +Your circles: Sharing behavior on Google+. In *Proc. SOUPS*, 2012.

[42] A. L. Young and A. Quan-Haase. Information revelation and Internet privacy concerns on social network sites: a case study of Facebook. In *Proc. 4th International Conference on Communities and Technologies*, 2009.

[43] T. Zesch and I. Gurevych. Analysis of the Wikipedia category graph for NLP applications. In *Proc. TextGraphs-2 Workshop (NAACL-HLT 2007)*, 2007.