Analysing Privacy Leakage of Life Events on Twitter

Dilara Keküllüoğlu
University of Edinburgh
d.kekulluoglu@ed.ac.uk

Walid Magdy
University of Edinburgh
wmagdy@inf.ed.ac.uk

Kami Vaniea
University of Edinburgh
kvaniea@inf.ed.ac.uk

ABSTRACT

People share a wide variety of information on Twitter, including the events in their lives, without understanding the size of their audience. While some of these events can be considered harmless such as getting a new pet, some of them can be sensitive such as gender-transition experiences. Every interaction increases the visibility of the tweets and even if the original tweet is protected or deleted, public replies to it will stay in the platform. These replies might signal the events in the original tweet which cannot be managed by the event subject. In this paper, we aim to understand the scope of life event disclosures for those with both public and protected (private) accounts. We collected 635k tweets with the phrase "happy for you" over four months. We found that roughly 10% of the tweets collected were celebrating a mentioned user's life event, ranging from marriage to surgery recovery. 8% of these tweets were directed at protected accounts. The majority of mentioned users also interacted with these tweets by liking, retweeting, or replying.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Human-centered computing** → *Social content sharing*; *Social media.*

KEYWORDS

Privacy, Online Social Networks, Twitter, Major Life Events

ACM Reference Format:
Dilara Keküllüoğlu, Walid Magdy, and Kami Vaniea. 2020. Analysing Privacy Leakage of Life Events on Twitter. In *12th ACM Conference on Web Science (WebSci '20), July 6–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394231.3397919

1 INTRODUCTION

Online Social Networks (OSN) are used by billions of users to connect with other people such as friends, colleagues, and people with similar interests. They share their daily lives and life events on social media; such as birthdays, marriages, buying a new car, or acceptance to their dream university. They also use social media as a source of support during difficult life events [4, 31]. While some of these communications are only visible to a user's closed network, they can also be shared publicly for everyone to see. Sharing life events can lead to unintended disclosure especially since OSN users are known to underestimate the audience size of their posts [7]. Some disclosures result in minor discomfort such as colleagues learning that the user is moving house, leaving their job, or getting a divorce. However, some posts can result in serious damage. Posts about surgeries and illnesses can result in insurance premium increases [13] while a post about having gender transition can result in discrimination.

Even when users restrict the audience of their posts, replies can lead to privacy leaks [17]. For example, a funny cat photo with the user's full address visible in the background might be shared, or a post about selling an item including the seller's phone number might be retweeted or sent on by well meaning followers. This is especially problematic in platforms like Twitter where the visibility of posts is only dependent on the tweeter's account type (i.e. public or protected). Hence, all posts from a public account are publicly visible, even if the tweet mentions a protected account. The result is that a public account can easily breach the privacy of a protected account even if the breach is unintentional. Twitter recently added a new feature to hide replies to a tweet but it is still possible for others to access those replies with few clicks [30].

In this paper, we focus on finding happy life events shared on Twitter, a social media platform where users may share posts of up to 280 characters. We selected Twitter partially because its course privacy controls where accounts are either *public* with all their tweets world readable, or *protected* where only an approved set of users can read them. Users are also able to *mention* each other in tweets (e.g. "@username"), which makes it easy to link tweet content to specific accounts. While simple, this privacy model can lead to public accounts having conversations with protected accounts where half the conversation is world readable. For example, a protected account might say "its my birthday!" and the public account might respond "happy birthday @alice!" disclosing the protected account's birthday publicly. Using Twitter also makes it possible to identify users' life events at scale, potentially even for protected accounts.

In particular, we are interested in tweets where the poster says that they are happy for an explicitly mentioned user regarding a life event. We ask the following research questions:

- **RQ1** What kind of life events can be detected in tweets that express happiness for a mentioned user?
- **RQ2** How do users react (e.g. like, retweet or reply) to such tweets when they are mentioned in them?
- **RQ3** How do protected (private) accounts react (e.g. like, retweet) relative to public accounts in these cases?

To answer these questions, we collected 1.4 million tweets/retweets between July and October 2019 containing the exact phrase "happy for you". We removed all posts involving *verified* accounts, as these are held by famous people or organizations and tend to have a large number of Twitter users discussing their life events, which would likely skew the data. We also removed retweets, and tweets that

WebSci '20, July 6–10, 2020, Southampton, United Kingdom

ACM ISBN 978-1-4503-7989-2/20/07…$15.00
https://doi.org/10.1145/3394231.3397919

mention no users, resulting in 635k tweets. We then used LDA [8] to detect topics in the dataset, resulting in 12 identified life events topics; including positive events like having a new baby, marriage, and graduation, as well as sensitive topics such as cancer, surgery, mental health and LGBTQ-related. However, as expected, not all tweets in our corpus corresponded to life events. Out of the original 635k tweets, only 59k belonged to a life event related topic. Looking only at the life event tweets, 51k mention only a single user, providing a clear indication of who is experiencing the life event. Out of these 51k tweets, 4k mention a single protected user, potentially breaching that user's privacy by making their life event public. The majority of protected and public account users reacted to the life event tweet mentioning them.

Our results suggest that it is possible to automatically identify Twitter users' key life events, even if they have a protected account. The outcome has implications for privacy, particularly around the impact of the sharing decisions of connections.

2 RELATED WORK

2.1 Harms

Disclosing events like vacations and illnesses can have unwanted results. Vacations may signal that the tweeter's home is vacant and burglars can use this information (e.g. PleaseRobMe.com). Sharing severe illnesses may result in increased premiums by insurance companies [13]. Mao et al. [26] looked for tweets with vacation, illness and drinking topics by using keyword-based data filtering. They classify these tweets as sensitive or non-sensitive using Naive Bayes with bag-of-word model.

2.2 Networked privacy

OSN users tend to underestimate the size of the audience of their posts. Bernstein et al. [7] found that the users' imagined size of the audience is 27% of the true audience size. This underestimation may lead to oversharing and unexpected privacy leaks. This is especially important because of the collective aspects of the OSNs. Even if a user is privacy-conscious and they can accurately estimate the posts' reach, their networks could still disclose information about them [5]. For example, a user can share their location on an OSN and tag another user, effectively disclosing where they both are. However, most OSN providers do not allow tagged users to remove/modify the post giving them limited control over their privacy. There are some early proof-of-concept research to solve these collaborative privacy situations automatically [16, 21, 22], but they have yet to be adopted by any OSN providers. Some users take drastic measures and deactivate/delete their accounts to protect themselves from this type of tagging [6]. However, this does not prevent creation of shadow profiles where information about them is shared by their networks [12].

Users' age, gender, religion, diet, and personality can be inferred only using the tweets mentioning their account [20]. Also analysing a user's connection network can disclose attributes about them such as age, gender, location, political orientation, and sexual orientation [2, 3, 19, 25, 32].

2.3 Detecting and inferring life events

Detecting life events using social media posts is not a new research area. Researchers have looked at a range of feature sets, types of life events, and approaches in an effort to accurately automatically identify these events from messy social media data.

Simple keyword queries were tried by De Choudhury et al. [9] to find women who were new mothers. They used keywords curated from birth announcements of local newspapers to find accounts of potential new mothers, then used lexicon-driven gender inference to identify women (as opposed to new fathers), with a 83% accuracy rate. Finally, they used crowdsourcing to label the accounts as new mothers or not, to have a high precision dataset.

Other works use keywords to gather a life event themed corpus, crowdsourcing to annotate it, and then use the annotated data to build a model that can automatically associate tweets with a life event. Dickinson et al. [10], focused on five life events psychologists have identified to be the most prominent in peoples' lives: "Starting School", "Falling in Love", "Getting Married", "Having Children", and "Death of a Parent". They were able to use the content features of tweets such as n-grams, mentions, and number of retweets, user, semantic, and interaction features to build an effective classifier using labeled data from crowdsourcing. Similarly, Akbari et al. [1] focused on personal wellness events (diet, exercise, and health). They used keywords to collect tweets from Twitter, manually labeled them, and built a classifier.

Instead of starting with a specific set of life events, some research starts with a very broad corpus and identifies the life events that exist within it. Li et al. [23] collected tweets using the broad keywords "congratulations" and "condolences" and used LDA [8] to find topics in the data set from which they focused on the life event topics. In their approach, they start with tweets identified through keyword matches, then look for any parent tweets, combining the parents and children into one document of only verbs and nouns. They used bootstrapping to find phrases other than "congratulations" and "condolences" used in the tweets such as "have fun" and "my deepest condolences" to expand their dataset. They repeated this process for four times and found 30 different phrases alongside with 42 event types.

Our work also uses broad keywords to gather an interesting corpus and then identify the life events found in it. However, rather than focus on general posts, we attempt to identify life events posted by people other than those experiencing the event. This focus allows us to look not only at public users, but also get a sense of the life events experienced by protected accounts.

3 DETECTING LIFE EVENTS FROM TWEETS

For our study, we collected tweets that contain the phrase "happy for you". Those tweets were then clustered according to the life event discussed in them. We analyzed these tweets to understand whether they are mentioning protected or public users. We describe our data collection, analysis and topic clustering below.

3.1 Tweets Collection

Using the Twitter streaming API [28], we collected tweets that contain the words "happy for you" resulting in all tweets that have

these three words but not necessarily in consecutive order. So, we filtered the tweets to only those that have the exact phrase.

We streamed tweets for four months between July and October 2019. A set of 1.4 million tweets/retweets were collected that contain the phrase "happy for you". Multiple filtering steps were then applied to the collected tweets. Initially, we filtered out all retweets, since we are only interested in the original tweet. In addition, we filtered out tweets that have no explicit mentioned accounts, since we are only interested in the tweets directed to specific users indicating they were about them. We then checked the type of the mentioned accounts within our tweets and removed any tweet mentioning verified accounts which indicates that the event is probably about a famous person, and thus privacy is less of a concern. After all these filtering steps, we had 634,590 tweets mentioning a total of 777K Twitter accounts. We refer to this dataset as "*HFY*" tweets dataset.

We divided the tweets according to the number of accounts mentioned in them and the conversation type of the tweets. We call the tweets that mention only one account *single* tweets, and the ones who mention more than one account is called *multiple* tweets. A tweet that mention another user has one of the following conversation types; a *reply* to an existing tweet, *directed to a user*, or *other*. *Reply* tweets are direct replies to another existing parent tweet; *directed to a user* tweets are not replies but they mention another account at the start of the tweet (e.g. "@username heard about the new addition to the family, I'm happy for you"). *Other* comprises all the tweets that mention at least one user but at the middle or end of the tweet. We use these terms throughout the paper. The majority of the tweets in *HFY* (607,703 tweets, 96%) are replies to another parent tweet.

3.2 Finding Life Events

Since all our tweets have the phrase "happy for you" for a given mentioned user, our next task was to infer the event that the mentioned users are being congratulated for. Manually checking each tweet is not feasible, since our *HFY* dataset contains 635k tweets. Hence, we used Latent Dirichlet Allocation (LDA) topic modeling [8] to cluster tweets into topics. By setting an input n as a suggested number of topics, LDA modeling assumes that each document in the corpus is a mixture of topics and each topic is a mixture of words. It uses a bag-of-words approach where the order of the words are ignored. First, we cleaned the tweets, lemmatized them and used the python implementation of gensim's ldamallet [27] to find the topic models.

Tweet Pre-processing: Given a tweet, we firstly converted it to lower case. We removed URLs, mentioned accounts ("@username"), as well as the # character in hashtags. Then we tokenized the tweets with the NLTK tweet-tokenizer [24]. After this step, we removed the words "happy", "for" and "you", since they were common in all tweets in our collection. We also removed emoji. We used bigram-phraser provided by the gensim to combine words that co-occur concurrently more than 100 times in the dataset. For example "safe travel", "speedy recovery", and "health pregnancy" are some of the bigrams we have in the dataset. Lastly, we lemmatized the tokens and removed the ones that are not nouns or verbs using spaCy [15] following the approach of Li et al. [23]. After all these steps, we stored these lemmatized tweets for use by the LDA model.

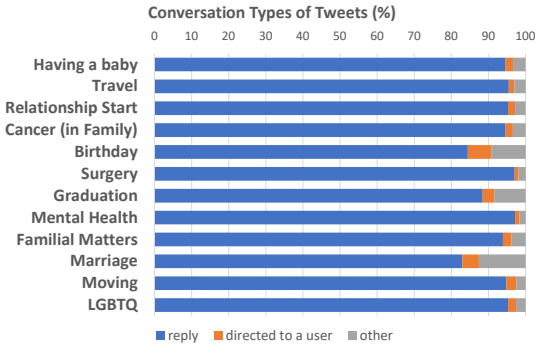

Figure 1: Topics divided by the conversation types; *reply, directed to a user,* and *other.*

Creating the LDA Model: We firstly created a dictionary from the words where each word is represented by an ID. Then we created our corpus with each tweet represented by list of IDs created using the dictionary. We used ldamallet [27] to create our LDA model. A fixed random seed was used to be able to reproduce the model, and we experimented clustering our dataset with different number of topics (clusters) n. We examined the following number of topics n={10, 20, 25, 30, 40, 70, 100, 120}. After looking at the distribution of the topic keywords and themes with each topics number n, we decided to continue with n=100 topics, which based on manual inspection, seems to be the optimal number of topics to produce many clean clusters related to the life events discussed in the tweets.

Topics Selection: For each of the 100 clusters of topics, we extracted the 30 most representative words and 10 most representative tweets, where "most representative" means the highest probability of belonging to that topic. A researcher looked through all the keywords and representative tweets and labeled clusters as involving life events or not, resulting in 22 clusters that involved life events. During the process, 8 clusters were identified that contained a mix of life events because the cluster had formed around an activity, such as prayer, that touches on many different life events. We therefore chose to exclude these clusters. Three of the clusters involved different activities associated with having a baby such as a baby shower, so we combined these three into one "having a baby" cluster. The result was 12 life event clusters, which are shown along with examples in Table 1.

Clustering the Tweets: LDA assumes that each document is a mixture of topics. Some of these topics are more probable in the document and some of them less. We assumed that each tweet only belongs to the most probable topic assigned by the LDA model. This way, we were able to cluster the tweets to topics.

After all these steps, only 58,801 tweets from *HFY* were assigned to the 12 life event topics. 86.3% of these tweets in were *single*, i.e. mentioning only one account. In cases where only one account is

Topic	Keywords	Example tweet
Having a baby	family, congratulation, news, blessing, member, addition, baby, girl, mommy, daddy, pregnancy, delivery, healthy_pregnancy, motherhood, boy, shower, gender_reveal	@username I'm so happy for you! Can't wait for the baby shower!
Travel	enjoy, time, rest, trip, weekend, summer, travel, visit, vacation, relax, holiday, london, japan, flight, safe_travel, korea, europe, germany, chicago, italy	@username so happy for you enjoy your trip [..]
Relationship start	person, relationship, boyfriend, girlfriend, bf, partner, keeper, gf, long_distance	A boyfriend like this is a keeper I'm happy for you sis @username
Cancer (in family)	mom, family, sister, dad, fight, brother, cancer, die, beat, lose, stage, grandma, uncle, cousin, warrior, nephew, aunt, fighter, niece, battle, survivor, monster, treatment, grandpa	@username So happy for you! May God keep you cancer free.
Birthday	day, birthday, today, celebrate, gift, bday, present, belated_birthday	@username CONGRATS AND HAPPY BIRTHDAY IM SO HAPPY FOR YOU
Surgery	hope, continue, pray, stay, recovery, health, take_care, heal, recover, speedy_recovery, rest, improve, surgery	@username [..] hope you have ease in your recovery.
Graduation	congratulation, success, work, achievement, accomplishment, celebrate, future, earn, cheer, graduation	[..] Happy graduation @username
Mental health	deal, pain, struggle, problem, doctor, issue, anxiety, mental_health, fear, therapy, overcome, depression, surgery, brain, stress, relief, med	@username I really hope you overcome your anxiety [..]
Familial matters	parent, kid, son, child, daughter, mother, family, mom, father, dad, bear, wife, raise, age, miracle, birth, husband, awareness, sibling, carry, grandmother, adopt	@username [..] that you have a grandchild too. [..]
Marriage	congratulation, wedding, marry, wife, husband, marriage, invite, day, engage, dress, bride, honor, ring, engagement, hubby, anniversary, party, honeymoon, honour, propose, divorce, fiancé	@username [..] The wedding will be fantastic though! [..]
Moving	move, place, home, house, leave, visit, fall, room, space, settle, city, town, area, apartment, land, pack	@username Everyone needs to leave home at one point. I feel you sis. [..]
LGBTQ-related	speak, people, woman, part, trust, realize, process, lie, power, figure, faith, community, truth, accept, idea, pride, gay, doubt, gender, embrace, tran	Congratulations to my favorite lesbians![..]

Table 1: Life event topics from *HFY-LE* and keywords selected from the 30 most probable words for each topic. Example tweets shown with usernames blinded and some content removed for privacy.

mentioned, we can safely assume that the event is about that user, but for tweets with multiple mentioned users it is impossible to accurately infer who is the event subject. Therefore, we focus on *single* tweets in our analysis. We call this subset *HFY-LE* for Life Event. 8% of *HFY-LE* mention protected accounts.

3.3 Resulting Clusters

Of all the tweets in *HFY-LE*, 47,342 (93%) were in *reply*, 2% were *directed to a user* and remaining were *other*. This shows that nearly all tweets were sent in response to an existing tweet. However, some topics received more tweets as *reply* than others. 97% of tweets with "mental health" and "surgery" topics were *reply* whereas this rate was 83% for "marriage". Tweets with sensitive topics related to health, sexual orientation, and so on were more likely to be replies to existing tweets. On the other hand, commonly celebrated things like marriage, birthday, and graduation are more frequently tweeted as a stand-alone tweet rather than in reply. The rates of the conversation types for each topic is shown in Figure 1.

The largest topic is "having a baby" with 15,289 tweets since it is a combination of three topics from the original clusters. The smallest topic is "LGBTQ-related" with only 2,387 tweets. In Figure 2, we provide the number of tweets from each topic broken by the account type of the mentioned users, we do not display "having a baby" since it is four times the second largest cluster. "Having a baby" has 13,912 tweets mentioning public accounts and 1,377 mentioning protected ones. The topic with the most protected tweets is "marriage" with 9%, whereas "familial matters" tweets are the least common with 6%.

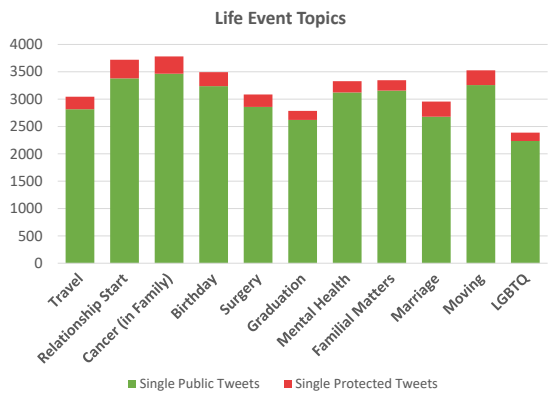

Figure 2: Number of tweets broken by the type of the mentioned user ("having a baby" not shown).

4 REACTIONS FROM MENTIONED USERS

After determining tweet topics, we collected the reactions from mentioned users such as likes, retweets, and replies to understand how users react to having their life-events disclosed by their friends.

4.1 Collection of Reactions

Four months after the last tweet was collected, we extracted the mentioned accounts and gathered the reactions to the tweets in the *HFY-LE*. Firstly, we tried to retrieve the reactions (like, retweet, or reply) to each tweet in *HFY-LE*. Some of these tweets were not available for various reasons, for example the tweet was deleted or the user protected their account.

4.1.1 Collection of Likes/Retweets. For the tweets we could reach, we checked whether the mentioned user liked or retweeted the tweet; however, the provided Twitter API retrieves this data very slowly. Hence, we used Twitter user interface (UI), which shows how many times a tweet is liked/retweeted. We also retrieved the list of who liked/retweeted via the UI. If the mentioned user's screen name is in the list, then we conclude that the user liked/retweeted the tweet. One drawback of this approach is that we can only get 25 people from the list. Hence, if the tweet is popular and has more than 25 likes/retweets, then we cannot decide whether the tweet was liked/retweeted by the mentioned user. However, this is a rare situation that happened in only 229 tweets from the 51k tweets in *HFY-LE*.

Protected accounts' tweets and likes are hidden, so we cannot see if they interacted with tweets. If a protected account likes/retweets a tweet, it will not show in the UI list but will still be counted towards the total likes/retweets. The difference between the like/retweet count and the number of users in the list indicates the number of protected accounts who liked/retweeted the tweet. Thus, we used this information to estimate if a mentioned protected account has likes/retweeted the tweet. While we cannot be sure that the protected account likes/retweeted a tweet is the mentioned one, we believe it is reasonable to assume so. The same drawback mentioned earlier also applies here, if a tweet is liked/retweeted more than 25 times we cannot be sure about the hidden like/retweet counts.

4.1.2 Collection of Replies. Next we collected user replies to the tweet by collecting the timelines of the mentioned users that were not protected and scanning them for replies. Since Twitter API does not have a feature that gives the replies to a tweet, we had to get the timelines of each mentioned user to check whether there is a reply to the tweet mentioned in them in *HFY-LE*. We check every tweet of the mentioned user between the time of the original tweet and the response collection time. The Twitter API only allows us to get the last 3200 tweets from a user's timeline. Hence, for some of the users we were not able to decide whether they replied or not since they tweeted more than 3200 after the original tweet was sent.

We could not apply the same method for protected accounts, since their timeline is inaccessible. Thus, reactions of protected accounts by replying to tweets mentioned them is unfortunately not included in our analysis. Similarly, we could not retrieve tweets from users who were suspended or deleted their accounts or were unreachable for other reasons.

4.2 Analysing Users' Reaction

When collecting likes and retweets, we found that 5,910 (12%) of the tweets could no longer be viewed. However, since we had the ID and the mentions of the tweet, we could still check if there was a reply from the mentioned users. We couldn't reach the tweets from each topic with similar rates; between 10% ("moving") and 13% ("familial matters"). On the other hand, aside from the 4,005 protected accounts, we couldn't reach further 3,084 (7%) mentioned users to collect reactions. These rates are between 4% and 8% for the most of the topics, while the rate for "surgery" is 13%. This might mean that these users delete their profiles more than other mentioned users in other topics.

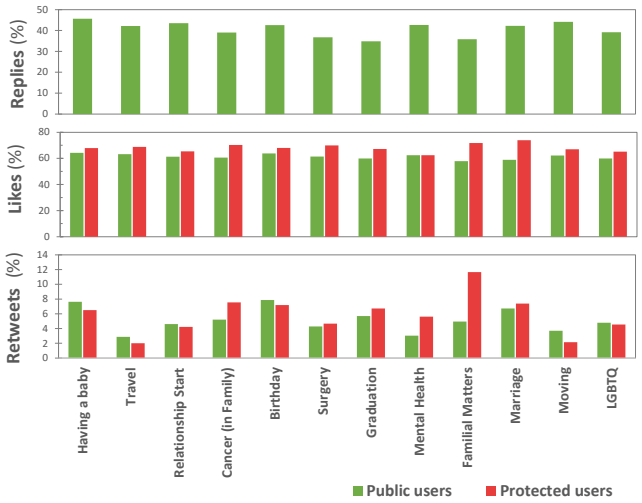

Figure 3: Reactions (reply, like, and retweet) by the mentioned users in the *HFY-LE* tweets.

From the ones we could reach, 24,047 (62%) of the tweets were liked by public mentioned uses. Similarly, 2,221 (6%) of the tweets were retweeted by the mentioned user. On the other hand, 2,319 (68%) of the tweets that mentioned a protected account had hidden likes and 203 (6%) of them had hidden retweets. While we cannot be sure all of these hidden likes/retweets were from the mentioned users, it gives us some idea about the interactions. 11,941 (42%) of the users with public accounts replied to the tweets they were mentioned in. The average time to reply was 5 hours while the longest was just over three months. In total 27,545 of the mentioned users showed at least one type of reaction. 941 of them reacted using all three ways to the tweets (i.e. like, retweet, and reply). 7,256 did not give any reaction.

Figure 3 shows the reactions from the mentioned user for each topic. All of the topics had similar rates of likes from the mentioned user. "Familial Matters", "marriage", and "LGBTQ-related" topics were on the lower end while "having a baby", " birthday", and "travel" were on the higher end. For retweets, "travel" and "mental health" were on the lower end while "marriage" and "having a baby" on the higher end.

In every topic, the rate of likes from the protected users are more than the like rates of the public users. As mentioned, this may be a result of counting every tweet that mentions a protected account with hidden likes as a reaction. Still they show similar patterns with the public tweets.

5 DISCUSSION

The purpose of our study is to understand how users' privacy can be breached from social media posts by people happy about their life-event. To measure this, we used the general phrase "happy for you" to collect potential tweets that might communicate with other users about happy events in their life. We managed to collect a large number of tweets mentioning users, including protected accounts. Using LDA, we inferred the discussed life-event in around 10% of the collected tweets. We managed to identify 12 life-events that we included in our analysis. This result relates to our first research question, where life-events about new born baby, marriage, graduation, and mental health could be inferred from the tweets. Investigating our second and third research questions, it was interesting to find the most users react positively to tweets that disclose their life-events. More surprisingly, we noticed that tweets were liked more often by the protected accounts, who could be assumed to have more privacy concerns but may consider their *protected* status sufficient protection.

5.1 Implications of Findings

The stated purpose of online social media is to help people connect with one another through a shared medium. It is therefore unsurprising that users would use such platforms to share life events since sharing is a common way people build social groups. However, the highly public nature of Twitter also means that information shared is open to a world-wide audience, a fact that may be technically known by users, but hard for them to conceptualize since many people believe that they themselves are not sufficiently interesting for an attacker to bother with [29]. But this view does not necessarily match how groups use data at scale.

The life events we identified are similar to those Janssen and Rubin [18] found when asking adults from Netherlands about the seven most important events that will happen in an ordinary Dutch child's life. "Having children" is the most frequently mentioned life event by Dutch people which is similar to our results where the largest topic cluster was "having a baby". Our other topic clusters also line up well with their results, indicating that our identified "happy for you" life event clusters do align with common important life events, suggesting that such data can be automatically extracted from Twitter.

Life events are big money for companies. Target famously hired analysts to predict which of their customers were pregnant so they could market to them before the birth announcement because new parents tend to purchase large quantities of items [11]. In our clusters we identified multiple tweets from marketers who were targeting people experiencing life events. For example, several wedding planning groups were replying to user tweets where they talk about their engagement or upcoming wedding. For example, a Twitter user posted about getting their venues booked for an upcoming wedding, resulting in the reply: *"@username We're SO happy for you, Kayla! Can we help with planning? [url] Plan with this Free Sample Kit: [url]"*. The threat of life events being automatically identified off of Twitter and used to target users is very real and currently used threat vector.

Life events are not just a privacy or marketing problem. They are also useful for attackers who want to cause harm or steal financial assets. Targeted attacks on valuable people, sometimes called *whaling*, often start with the attacker spending time on company people pages and friending them on social media. The attacker can then use the online trove of personal information as part of social engineering to trick a user or system into providing more valuable access. For example, what some companies consider to be public non-sensitive data is used by other companies for authentication, which means that an attacker can start with seemingly low sensitive data and work there way all the way to remote resetting a Wired journalist's Mac [14].

Personal privacy management is also challenging on Twitter due to the ease of creating *shadow* profiles where information about a user is available via other peoples' tweets. For example, one friend may tweet "happy birthday @ProtectedUser" creating a public shadow record of the protected user's birthday, then another public account tweets "looking forward to our trip" again creating public data. Together these public tweets create a shadow profile of the protected user which they cannot easily control. The design of Twitter also facilitates the creating of shadow profiles through the use of course access control at profile level and the culture of replying and retweeting posts to followers. While only 8% of the life event tweets were mentioning a protected account, that still amounts to 4k users having their life events exposed. Our research also indicates that the types of life events exposed for protected and public accounts are very similar, suggesting that public posters are equally willing to post about a range of life events for both protected and public accounts.

While sharing and long term existence of information is a problem, we also noticed the opposite where tweets vanished and users changed privacy settings between the time when we collected them and when we went back to get reactions. 12% of tweets containing a life event vanished between the collections, with each topic having roughly the same percentage of tweets vanish. More interesting, 6.6% of mentioned accounts were deleted or suspended in that time frame, with 12.8% of accounts mentioned in surgery life events vanishing. This observation suggests that people are taking some actions that protect theirs and others' privacy.

5.2 Limitations

By using Twitter we were able to get a large sample of social media posts to work from, but our data set and analysis still have some important limitations to consider. Our analysis is limited to tweets containing the phrase "happy for you". While this is useful to collect positive life events, it likely has few examples of common negative life events. It also focuses our analysis on a life event that someone else might want to provide positive commentary on.

To cluster the tweets we applied LDA which uses word co-occurrence at the document level to discover topics. However, tweets are very short documents which may inhibit LDA from performing as well as it does with longer documents. LDA also requires us to state the number of topics in advance, which is obviously not a known number. We used the standard approach of selecting several possible topic numbers, running LDA with each setting, and manually checking the coherence of the resulting topics.

We also assume that each tweet has only one topic and therefore assign each only to the most probable topic. While we did read through many tweets during this process, we did not attempt to manually label tweets.

We gathered the reactions from the mentioned users four months after the last tweet was collected. The time delay meant that we could accurately collect reactions, but it also meant that some tweets and accounts vanished. Also in some cases where the mentioned users were very active we were not able to retrieve their older tweets because the Twitter API limits per-user tweet retrieval to 3,200 tweets.

6 CONCLUSION

In this work, we collected 635k tweets containing the phrase "happy for you" that mention at least one user. We used LDA topic modeling to cluster the tweets, resulting in 12 life event topic clusters with 51k of *single* tweets belonging to one of these topics. "Having a baby" was the largest cluster while "LGBT-related" was the smallest. 8% of the tweets mention protected users and the rate of protected user mentions in topics ranged between 6% and 9%.

The majority of tweets received reactions from mentioned users. The most common reaction was liking the tweet, followed by replying. Retweeting was the least common reaction. The rates for likes/retweets/replies were fairly consistent between topics. Protected accounts tended to like the tweets that mention them more often than public accounts with no major variation between topics.

To our knowledge, this is the first work that has focused on using only the tweets that mention users to infer life events about them. This work is important since the mentioned user cannot modify the visibility of the tweet. A user can protect their own tweets but the tweets that mention them can only be controlled by the tweeter. In addition, tweets from public accounts replying to protected account tweets can be seen by anyone and even if the parent tweet is deleted, replies will stay visible.

ACKNOWLEDGMENTS

This work was supported in part by the EPSRC DTA award, funded by the UK Engineering and Physical Sciences Research Council and the University of Edinburgh. We thank everyone associated with the TULiPS Lab and SMASH Group at the University of Edinburgh for helpful discussions and feedback.

REFERENCES

[1] Mohammad Akbari, Xia Hu, Nie Liqiang, and Tat-Seng Chua. 2016. From tweets to wellness: Wellness event detection from twitter streams. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[2] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[3] Abeer AlDayel and Walid Magdy. 2019. Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media. In *The 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing*. ACM.

[4] Nazanin Andalibi and Andrea Forte. 2018. Announcing pregnancy loss on Facebook: A decision-making framework for stigmatized disclosures on identified social network sites. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[5] James P Bagrow, Xipei Liu, and Lewis Mitchell. 2019. Information flow reveals prediction limits in online social activity. *Nature human behaviour* 3, 2 (2019), 122–128.

[6] Eric PS Baumer, Phil Adams, Vera D Khovanskaya, Tony C Liao, Madeline E Smith, Victoria Schwanda Sosik, and Kaiton Williams. 2013. Limiting, leaving, and (re) lapsing: an exploration of facebook non-use practices and experiences. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 3257–3266.

[7] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 21–30.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[9] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Major life changes and behavioral markers in social media: case of childbirth. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1431–1442.

[10] Thomas Dickinson, Miriam Fernandez, Lisa A Thomas, Paul Mulholland, Pam Briggs, and Harith Alani. 2015. Identifying prominent life events on twitter. In *Proceedings of the 8th International Conference on Knowledge Capture*. 1–8.

[11] Charles Duhigg. 2012. How Companies Learn Your Secrets. Retrieved Feb 29, 2020 from https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

[12] David Garcia, Mansi Goel, Amod Kant Agrawal, and Ponnurangam Kumaraguru. 2018. Collective aspects of privacy in the twitter social network. *EPJ Data Science* 7, 1 (2018), 3.

[13] Ki Mae Heussner. 2009. Woman Loses Benefits After Posting Facebook Pics. Retrieved April 4, 2019 from https://abcnews.go.com/Technology/AheadoftheCurve/woman-loses-insurance-benefits-facebook-pics/story?id=9154741

[14] Mat Honan. 2012. How Apple and Amazon Security Flaws Led to My Epic Hacking. Retrieved March 26, 2019 from https://www.wired.com/2012/08/apple-amazon-mat-honan-hacking/

[15] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[16] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. 2013. Multiparty access control for online social networks: model and mechanisms. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (2013), 1614–1627.

[17] Prachi Jain, Paridhi Jain, and Ponnurangam Kumaraguru. 2013. Call me maybe: Understanding nature and risks of sharing mobile numbers on online social networks. In *Proceedings of the first ACM conference on Online social networks*. ACM, 101–106.

[18] Steve MJ Janssen and David C Rubin. 2011. Age effects in cultural life scripts. *Applied Cognitive Psychology* 25, 2 (2011), 291–298.

[19] Carter Jernigan and Behram FT Mistree. 2009. Gaydar: Facebook friendships expose sexual orientation. *First Monday* 14, 10 (2009).

[20] David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Writer profiling without the writer's text. In *International Conference on Social Informatics*. Springer, 537–558.

[21] Dilara Kekulluoglu, Nadin Kokciyan, and Pinar Yolum. 2018. Preserving privacy as social responsibility in online social networks. *ACM Transactions on Internet Technology (TOIT)* 18, 4 (2018), 42.

[22] Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. 2017. An argumentation approach for resolving privacy disputes in online social networks. *ACM Transactions on Internet Technology (TOIT)* 17, 3 (2017), 27.

[23] Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1997–2007.

[24] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1* (Philadelphia, Pennsylvania) *(ETMTNLP '02)*. Association for Computational Linguistics, USA, 63–70. https://doi.org/10.3115/1118108.1118117

[25] Walid Magdy, Yehia Elkhatib, Gareth Tyson, Sagar Joglekar, and Nishanth Sastry. 2017. Fake it till you make it: Fishing for Catfishes. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 497–504.

[26] Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*. ACM, 1–12.

[27] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[28] Twitter. 2019. Twitter API. Retrieved March 27, 2019 from https://developer.twitter.com/

[29] Rick Wash. 2010. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*. 1–16.

[30] Suzanne Xie. 2019. More control over your conversations: now available globally. Retrieved May 13, 2020 from https://blog.twitter.com/en_us/topics/product/2019/more-control-over-your-conversations-globally.html

[31] Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[32] Elena Zheleva and Lise Getoor. 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*. ACM, 531–540.