Loose Tweets: An Analysis of PrivacyLeaks on Twitter

Huina Mao, Xin Shuai, Apu Kapadia

WPES 2011

Presented by Luam C. Totti on May 9, 2012
Motivation

- 65 million tweets every day.
- How many contain sensitive information?
- Qualitatives studies show that users ‘regret’ some tweets.
- What kind of information is leaked? Why? When? By whom?
- Could that leaked information harm you?
Contributions

1. Characterization of the types of privacy leaks on Twitter.
2. Proposal and evaluation of automatic classifiers to detect such leaks.
3. Analysis of who leaks information and how.
5. Discussion on how to build defensive systems using the results.
Overview

Figure: Overall Architecture
Overview

- Tweets from Streaming API.
- Around 15% of all public tweets randomly sampled.
- January to September, 2011
- Only tweets with pronouns like I, me, my, me, us, she, ...
- Which gives around 160 million tweets.
- All filters are word/sentence based: holiday, travel, I'm drunk.
Vacation Tweets

- Tweets in which the user reveals vacation plans.
- Are sensitive when they reveal *concrete* vacation plans.
- Keywords used: *vacation*, *holiday*, *travel*, *trip*, *leave for*, *fly to*.
- A total of 575,689 tweets were filtered.
Vacation Tweets

- 1000 randomly sampled tweets were manually labelled as “sensitive” or “non-sensitive”.
- 10.8% were considered as sensitive.
- Among those 108 tweets, some entities were found:
  - Location: 90.7%
  - Time: 55.5%
  - Person: 44.4%
- Can be done automatically using NER techniques.
Classification

How to automatically identify sensitive tweets?

Some words are strong indicatives that the tweet is sensitive.

Like the presence of entities: location, time and person.

Table 1: List of representative words for classifying vacation tweets.

<table>
<thead>
<tr>
<th>Category</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>place and facility</td>
<td>beach, coast, hotel, conference, island, airport, flight</td>
</tr>
<tr>
<td>positive</td>
<td>go, going, gonna, leave, leaving, pack, booked, before, will,</td>
</tr>
<tr>
<td></td>
<td>until, wait, plan, ready, here I come, looking forward</td>
</tr>
<tr>
<td>negative</td>
<td>need, wish, not, no, want, wanna, back, went, may, might, maybe,</td>
</tr>
<tr>
<td></td>
<td>had, recent, was, were, could, should, hope, got, suppose, if, didn’t</td>
</tr>
<tr>
<td>url</td>
<td>http</td>
</tr>
<tr>
<td>hashtag</td>
<td>#</td>
</tr>
</tbody>
</table>
Evaluation

- 600 tweets were manually labeled and used to train classifiers (Naive Bayes and SVM).

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>0.415</td>
<td>0.383</td>
<td>0.28</td>
<td>0.324</td>
</tr>
<tr>
<td>Time</td>
<td>0.605</td>
<td>0.63</td>
<td>0.51</td>
<td>0.563</td>
</tr>
<tr>
<td>Location</td>
<td>0.715</td>
<td>0.717</td>
<td>0.71</td>
<td>0.713</td>
</tr>
<tr>
<td>All words</td>
<td>0.7</td>
<td>0.682</td>
<td>0.75</td>
<td>0.714</td>
</tr>
<tr>
<td>Representative words</td>
<td>0.61</td>
<td>0.567</td>
<td>0.93</td>
<td>0.704</td>
</tr>
<tr>
<td>Location + Representative words</td>
<td><strong>0.785</strong></td>
<td><strong>0.761</strong></td>
<td><strong>0.83</strong></td>
<td><strong>0.794</strong></td>
</tr>
</tbody>
</table>

Which tells that 76% of the alerts supplied to burglars would indicate real vacation periods.

Now the burglars only need to “use existing geo-inferencing algorithms to infer the user’s residence address”
Evaluation

- 600 tweets were manually labeled and used to train classifiers (Naive Bayes and SVM).

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>0.415</td>
<td>0.383</td>
<td>0.28</td>
<td>0.324</td>
</tr>
<tr>
<td>Time</td>
<td>0.605</td>
<td>0.63</td>
<td>0.51</td>
<td>0.563</td>
</tr>
<tr>
<td>Location</td>
<td>0.715</td>
<td>0.717</td>
<td>0.71</td>
<td>0.713</td>
</tr>
<tr>
<td>All words</td>
<td>0.7</td>
<td>0.682</td>
<td>0.75</td>
<td>0.714</td>
</tr>
<tr>
<td>Representative words</td>
<td>0.61</td>
<td>0.567</td>
<td>0.93</td>
<td>0.704</td>
</tr>
<tr>
<td>Location +Representative words</td>
<td>0.785</td>
<td>0.761</td>
<td>0.83</td>
<td>0.794</td>
</tr>
</tbody>
</table>

- Which tells that 76% of the alerts supplied to burglars would indicate real vacation periods.

- Now the burglars only need to “use existing geo-inferencing algorithms to infer the user’s residence address” 🙄
Drunk Tweets

Two research questions arise:

- What private information is revealed during drunk tweeting?
- Are people more likely to reveal private information when they are drunk?

Keywords used in filter: I[’m|m|am]drunk.

21,297 tweets were filtered.
Drunk Tweets

![Graph showing the daily pattern of drunk tweets between April and June 2010. The graph includes a bar chart and an autocorrelation graph. The bar chart shows the drunk tweets plotted against the date (Day/Month), with peaks on 03/04, 10/04, 17/04, 24/04, 01/05, 08/05, 15/05, 22/05, 29/05, 05/06, 12/06, 19/06, and 26/06. The autocorrelation graph displays the autocorrelation values ranging from -1 to 1, with peaks indicating daily patterns. The x-axis represents the date, and the y-axis represents the drunk ratio (z-score).]
Drunk Tweets

- Top 100 users with most drunk tweets were selected.
- Tweets that match the filter are reference points $T_d$.
- Tweets after 3 hours of a $T_d$ are considered drunk tweets.
- Tweets between 6 and 3 hours preceding a $T_d$ are considered sober tweets.
- 645 were drunk tweets and 208 were sober tweets.
- May confirm that drunk people are more talkative. (??)
5.3 Drunk driving classification

We designed a scheme for drunk tweetsm we also annotate those sober tweets as opposed to sober tweets. Based on our coding, approximately 25% of drunk tweets were made jokingly, and so we filtered tweets or nonsensitive tweets. We found that a small fraction of drunk tweets and classified them into sensitive tweets or non-sensitive tweets. We performed content analysis to find out all private topics revealed. Then we manually annotated them into the above topics. We give some excerpted and slightly reworded examples to illustrate the privacy leaks when they are 'tweeting under the influence'.

To demonstrate the threat of automated detection of such tweetsm we designed a classifier to automatically detect those drunk driving tweetsm we considered some other textual features which we found frequently occur in sensitive drunk driving tweetsm indicating the person who is driving and the person who is drunk is the same person. Specifically, we looked at some representative negative words or phrases such as

- I just fell down the stairs
- I drove drunk around the corner
- I just hit my head really bad

We plotted the private topics distribution in drunk tweets. Among all topics, we consider the following topics that related to privacy issues and labeled them into the above topics. During our annotationm we found the following topics that related to privacy issues and labeled them into the above topics. Our basic features are the words just after the key word 'drive' selves, the category of words also plays an important role. We found that in addition to words themn feature as a pattern is found in a tweetm then we mark the pattern for that tweet. Fourthm we employed part-of-speech taggingo We found that the smaller the distance between keywords, the higher probability that the tweet is a sensitive sample. Secondm we considered some other textual features which we found frequently occur in non-sensitive samples. Once such words are found, we marked the negative feature as true for that tweet. We found that the smaller the distance between keywords, the higher probability that the tweet is a sensitive sample. Figure 5: Comparison between drunk and sober tweets by percentage of sensitive topics.

The word just after the key word 'drive' was considered as our basic feature. Other than that, we considered some other textual features which we found frequently occur in sensitive drunk driving tweets. We performed content analysis to find out all private topics revealed. Then we manually annotated them into the above topics. We give some excerpted and slightly reworded examples to demonstrate the privacy leaks when they are 'tweeting under the influence'.

Drunk Tweets Topics

- Expressed Emotions: 22.0%
- Sexuality: 25.0%
- Confessions: 16.7%
- Disrespectful Behaviors: 7.8%
- Illegal Activities: 3.4%
- Bodily Harm: 25.0%
- Sexuality: 25.0%

Figure 4: Distribution of sensitive topics found in drunk tweets and non-sensitive tweets.
Drunk vs Sober

Bodily Harm
Illegal Activities
Confessions
Disrespectful Behaviors
Sexuality
Expressed Emotions

Drunk tweets
Sober tweets

Figure 4: Distribution of sensitive topics found in drunk and sober tweets.

Figure 5: Comparison between drunk and sober tweets by percentage of sensitive topics.

Drunk driving classification
5.3 Drunk driving classification
privacy leaks when they are 'tweeting under the influence'. Thus we observe that users are more susceptible to confessions and labeled them into the above topics. We also annotate those sober tweets as opposed to sober tweets. Based on our coding scheme for drunk tweets, we also annotate those sober tweets. The most serious privacy leak is specifically, most of the ill activities are about drunk driving. To demonstrate the threat of automated detection of such tweets, we designed a classifier to automatically detect those drunk driving tweets. We plotted the private topics distribution in drunk tweets. Among all topics, we consider the following topics that related to privacy issues. During our annotation, we find out all private topics revealed. Then we manually annotate those tweets containing drunk开车 around the corner. I just fell down the stairs and hit my head really bad. Me and my girl broke up. I'm single again. Confessions - revelation of personal information. Disrespectful Behaviors - rants and embarrassing behavior. Sexuality - revelation of sexual orientation or sexual activities and desires. Emotions - expression of love and hate for someone. Speech tagging. We found that in addition to words themselves, the category of words also plays an important role. During manual annotation, we consider some other textual features which we found frequently occur in sensitive drunk driving tweets. indicating that the person who is driving and the person who is drunk is the same person. Specifically, we considered some other textual features which we found frequently occur in non-sensitive samples. Once such words are found, we look at some representative negative words or phrases such as I'm drunk. We found that the smaller the distance between keywords, the higher the probability that the tweet is a sensitive sample. Second, we considered the relative frequency of words. First, we analyzed the relative frequency of all other topics show increased rates in drunk tweets. We plotted the private topics distribution in drunk tweets and observed that except for illegal activities, drunk driving or other illegal activities, are the most mentioned topics in drunk tweeting. we give some excerpted and slightly reworded examples to demonstrate the most serious privacy leak. Specifically, most of the ill activities are about drunk driving.

5.3.1 Feature selection.

• The word just after the key word driven or drove may confirm that people get more talkative when they are drunk.
The “Illegal Activities” topic includes drunk and driving tweets.

Is it possible to automatically identify such tweets?

Only tweets containing both words drunk or (drive,drove) were considered.
Drunk Driving Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>key words distance</td>
<td>position index difference between <em>drunk</em> and <em>drove/drive</em></td>
</tr>
<tr>
<td>negative words or phrases</td>
<td>don’t, not, no, couldn’t, can’t, didn’t, wasn’t, won’t, wouldn’t, if, wish, too drunk to drive</td>
</tr>
<tr>
<td>regular expression pattern</td>
<td>...I/I’m/me...drunk...I/I’m/me...drive/drove... or</td>
</tr>
<tr>
<td></td>
<td>...I/I’m/me...drove/drive...I/I’m/me...drunk...</td>
</tr>
<tr>
<td>words tagging</td>
<td>category of word after <em>drunk</em></td>
</tr>
<tr>
<td></td>
<td>category of word after <em>drove/drive</em></td>
</tr>
</tbody>
</table>
Evaluation

Table 4: Naive Bayes classification results from different combinations of features for drunk tweets.

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>All words feature</td>
<td>0.75</td>
<td>0.797</td>
<td>0.67</td>
<td>0.728</td>
</tr>
<tr>
<td>Textual features</td>
<td>0.695</td>
<td>0.741</td>
<td>0.6</td>
<td>0.663</td>
</tr>
<tr>
<td>Both</td>
<td><strong>0.79</strong></td>
<td><strong>0.845</strong></td>
<td><strong>0.71</strong></td>
<td><strong>0.771</strong></td>
</tr>
</tbody>
</table>

- Meaning that in 85% of the times law enforcement agencies (or your mom) will correctly identify a drunk driving episode.
Disease Tweets

- To what extend people disclose their healthy conditions on Twitter?
- Tweets were filtered using the names of 390 health conditions.
- Each tweet must also contain words like “has”, “had” and “have”.
- 21,508 tweets were found with 45 different diseases.
Diseases Mentions

- Cancer: 61.9%
- Diabetes: 11.0%
- Depression: 12.1%
- AIDS: 6.6%
- Down Syndrome: 4.7%
- Obesity: 2.1%
- HIV: 3.2%
- Others: 12.1%

Figure 6: Distribution of names of diseases in tweets.
Classification

- Once again, which tweets actually reveal sensitive information?
- Is it possible to automatically identify such tweets?
- For this type of leak, a generic classifier was too difficult to conceive, so only the cancer mentions were considered.
- Instead of classifiers, they used a regular expression for making the decision.
Classification

- If \( has/have/had \ldots cancer \) exists and \( dog/cat/kitty/puppy \ and \ doesn’t/doesn’t have/has/had \) doesn’t exist in a tweet, then it is sensitive; otherwise, it is not.

- For evaluation, 200 tweets were annotated and tested.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.782</td>
<td>0.759</td>
<td>0.82</td>
<td>0.788</td>
</tr>
</tbody>
</table>

- Meaning that in 76% of the times an insurance company will detect that a potential client have an existing condition.
Types of Leaks

- We looked into *what* was being leaked.
- What about *how* the leaks happen?
  - *Status leaks*: Normal status updates.
  - *Conversation leaks*: Tweets containing @.
- What about *who* is causing the leaks?
  - *Primary leaks*: Information about self.
  - *Secondary leaks*: Information about someone else.
Loose Tweets: An Analysis of Privacy Leaks on Twitter

Vacation

(а) Vacation tweets.

Figure 8: Comparison between primary and secondary leaks for vacation, drunk, and disease tweets.

Next we investigate the prevalence of sensitive tweets in vacation, drunk, and disease tweets. We randomly sampled three countries for our comparison: the United States, the United Kingdom, and Singapore, which represents different cultures. We chose three countries for our comparison: the United States, the United Kingdom, and Singapore, which represents different cultures. We analyzed tweets originating in these countries from the UK are least likely to contain sensitive information and then chose only the US, UK, and Singapore. We also found that the number of status tweets is very close to that of conversation tweets for all three diseases. There is no difference among the three types of tweets.

Table 5: Percentage of vacation, drunk, and disease tweets for each country.

<table>
<thead>
<tr>
<th>Country</th>
<th>Vacation</th>
<th>Drunk</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>20%</td>
<td>25%</td>
<td>15%</td>
</tr>
<tr>
<td>UK</td>
<td>18%</td>
<td>22%</td>
<td>17%</td>
</tr>
<tr>
<td>SG</td>
<td>22%</td>
<td>24%</td>
<td>16%</td>
</tr>
</tbody>
</table>

The results are shown in Table 5. We see that the percentage of vacation, drunk, and disease tweets for each country is similar. The US and UK have similar ratios, and Singapore has a slightly higher percentage of vacation tweets. For all three countries, we have the similar percentage of vacation, drunk, and disease tweets for each country. The results are shown in Table 5. We see that the percentage of vacation, drunk, and disease tweets for each country is similar. The US and UK have similar ratios, and Singapore has a slightly higher percentage of vacation tweets. For all three countries, we have the similar percentage of vacation, drunk, and disease tweets for each country.
Drunk

(b) Drunk tweets.
### Table 3: Textual features for classifying drunk driving tweets

<table>
<thead>
<tr>
<th>Feature</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>key words distance</td>
<td>position index difference between drunk and drove/drive</td>
</tr>
<tr>
<td>negative words or phrases</td>
<td>don't/ not/ couldn't/ can't/ wasn't/ won't/ wouldn't/ if/ wish/ too drunk to drive</td>
</tr>
<tr>
<td>regular expression pattern</td>
<td>...I/I'm/me...drunk...I/I'm/me...drive/drove... or ...I/I'm/me...drove/drive...I/I'm/me...drunk...</td>
</tr>
<tr>
<td>words tagging</td>
<td>category of word after drunk</td>
</tr>
<tr>
<td></td>
<td>category of word after drove/drive</td>
</tr>
</tbody>
</table>

Precision, recall, and $F_k$ measure are $nluvpi nluwpi nlvp$ and $nluvvi$ respectively. Compared with our best results obtained from vacation and drunk driving classifications, our classification result for cancer disease again demonstrates that alerts to an insurance company, for example, would have utmost precision and demonstrates a real threat.

### 7. DIFFERENT TYPES OF LEAKS

In the previous sections, we analyzed the sensitive content from three private topics: vacation, drinking, and disease, as well as designed binary classifiers to automatically detect sensitive tweets in those categories. In other words, we studied what leaks in sensitive tweets. In this section, we shift our focus to two other questions:

- **how does privacy leak from sensitive tweets?**
- **who is revealed in sensitive tweets?**

Specifically, we have two ways to categorize sensitive tweets. Based on the type of tweet, we have status leaks, i.e., privacy leaks from status update tweets, and conversation leaks, i.e., privacy leaks from conversation-based tweets. Given a tweet, if it starts with @username, then it is a conversation tweet. Otherwise, it is a status tweet. Next, based on who is revealed, we have primary leaks, where the original tweet poster implicates himself/herself, and secondary leaks, where a tweet poster implicates some other person. We can distinguish such primary and secondary leaks through content analysis. An example in our dataset of a secondary leak through a status update is "...My mom 'borrowed' my car and drove it around drunk" and an example of a secondary leak through a conversation tweet is "@[anonymized] I will pray for your mom. I had 2 family members diagnosed with cancer in the past 2 days. My brother and aunt. :-("). Note in the latter example, it may have been unlikely for the author to implicate his/her brother and aunt in a regular status update, but he/she provides this information as part of a conversation. Thus, there are four types of leaks, which we analyze in the categories of outgoing vacation, drunk driving, and three representative diseases: leukemia, diabetes, and HIV.

### Diseases

![Chart showing comparison between primary/secondary and status/conversation leaks in disease tweets.](image)

(a) Cancer
Diseases

<table>
<thead>
<tr>
<th>Feature</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>key words distance</td>
<td>position index difference between drunk and drove/drive</td>
</tr>
<tr>
<td>negative words or phrases</td>
<td>don't, not, couldn't, can't, wasn't, won't, wouldn't, if, wish, too drunk to drive</td>
</tr>
<tr>
<td>regular expression pattern</td>
<td>...I/I'm/me...drunk...I/I'm/me...drive/drove... or ...I/I'm/me...drove/drive...I/I'm/me...drunk...</td>
</tr>
<tr>
<td>words tagging</td>
<td>category of word after drunk</td>
</tr>
<tr>
<td></td>
<td>category of word after drove/drive</td>
</tr>
</tbody>
</table>

Precision, recall, and Fk measure are 0.868, 0.895, and 0.881, respectively. Compared with our best results obtained from vacation and drunk driving classifications, our classification result for cancer disease again demonstrates that alerts to an insurance company, for example, would have a precision and demonstrates a real threat.

7. DIFFERENT TYPES OF LEAKS

In the previous sections, we analyzed the sensitive content from three private topics: vacation, drinking, and disease, as well as designed binary classifiers to automatically detect sensitive tweets in those categories. In other words, we studied what leaks in sensitive tweets. In this section, we shift our focus to two other questions:

- How does privacy leak from sensitive tweets?
- Who is revealed in sensitive tweets?

Specifically, we have two ways to categorize sensitive tweets. Based on the type of tweet, we have status leaks (privacy leaks from status update tweets) and conversation leaks (privacy leaks from conversation-based tweets). Given a tweet, if it starts with @username, then it is a conversation tweet; otherwise, it is a status tweet. Next, based on who is revealed, we have primary leaks (where the original tweet poster implicates himself/herself) and secondary leaks (where a tweet poster implicates some other person). We can distinguish such primary and secondary leaks through content analysis. An example in our dataset of a secondary leak through a status update is "...My mom 'borrowed' my car and drove it around drunk" and an example of a secondary leak through a conversation tweet is "@[anonymized] I will pray for your mom. I had 2 family members diagnosed with cancer in the past 2 days. My brother and aunt. :-(").

Note in the latter example, it may have been unlikely for the author to implicate his/her brother and aunt in a regular status update, but he/she provides this information as part of a conversation. Thus there are four types of leaks, which we analyze in the categories of outgoing vacation, drunk driving, and three representative diseases: leukaemia, diabetes, and HIV.

Figure 7: Comparison between primary/secondary and status/conversation leaks in disease tweets.

(b) Diabetes
Cross-cultural Analysis

- Is this behavior consistent across the world?
- Tweets were separated by location they were posted.

Table 5: Percentage of vacation, drunk, disease tweets across these three countries, US, UK and SG (Singapore).

<table>
<thead>
<tr>
<th>Privacy types</th>
<th>US</th>
<th>UK</th>
<th>SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vacation</td>
<td>0.34</td>
<td>0.40</td>
<td>0.34</td>
</tr>
<tr>
<td>Drunk</td>
<td>0.01</td>
<td>0.01</td>
<td>0.006</td>
</tr>
<tr>
<td>Disease</td>
<td>0.02</td>
<td>0.02</td>
<td>0.008</td>
</tr>
</tbody>
</table>
Figure 10: Fraction of sensitive tweets across countries.
Loose Tweets: An Analysis of Privacy Leaks on Twitter

Diseases

Figure 9: Distribution of diseases over US, UK and Singapore
Conclusions

- Some privacy threats were highlighted in this work.
- Direct negative implications were shown.
- It’s possible to create automated systems to detect such methods.
- Leave the creation of guardian systems as future works.
Observations

- Uses very standard techniques.
- Very extensive manual labeling and annotation.
- Quantitative data missing at some points.
- Analysis rely much on the word filter, but not much is talked about it.
- Classification features choice has some good arguments.
- In general, the analysis are well presented and relevant.
THANK YOU