Twitter and Privacy

What can be mined from a tweet?

Dennis Muller
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
d.w.muller@student.utwente.nl

ABSTRACT
This paper will describe how private social media messages are these days, and how accurate the information that is gathered from this large amount of data is. However, there is no detailed literature review available that covers the different aspects of privacy of social media content. In this paper we focus on Twitter messages by first performing a literature review covering related work - all the research done regarding the privacy of Twitter messages. After this, there will be a section on what information can be used from a Tweet in order to make an user profile and what steps need to be taken in order to extract this information from a Tweet. Lastly, it is safe to say that Twitter profiles are not as private as users believe them to be. A lot can be said about a person based on just a single Tweet of that person and the more Tweets from one person, the more accurate the extracted information is.

Keywords
Twitter, Social Media, Privacy, Personal Information, Data Mining, Twitter Mining.

1. INTRODUCTION
In 2014 there were a total of 1.59 billion social media users in the world [23]. Currently there are 3.1 billion people that have access to Internet. So more than 50% of the people that have access to Internet are using social media [24]. Every day more people are getting access to Internet and every day more people are joining social media. Everyone has seen a profile on a social media website and there is a 50% chance that a person actually has a profile on a social media website. All kinds of companies and researchers are mining data from these social media websites for their own use. However not everyone uses this information for the right purposes. For example, there are many companies out there that are mining data from Twitter. They do this in order to know more about all kinds of users. This makes it possible for them to find out what makes certain users loyal to their own products and how their rivals are doing [9]. Not many users realise how much information they are actually giving away and how they should protect themselves from these big companies that mine their data and use it for their own purposes.

1.1 Problem Statement
With social media becoming a bigger part of our lives, a huge part of our society has a Twitter and/or Facebook account. What many of these users do not realise is that, if they are not protecting their account properly, random data miners in the world can gain access to a huge amount of information of certain users. Data mining is the acquisition of data by using certain tools in order to gain social media messages and store these so that they can be analysed later. Often there is a news article about privacy and how privacy is almost non-existent nowadays. The users are concerned with their privacy but are not really aware of how much information they inadvertently disclose. In this paper we focus on Twitter, and therefore, our central research goal in this research is going to be: How much information can be data mined from a Twitter message? [6].

1.2 Research Question
As stated in the introduction privacy is a huge issue these days and many users are concerned with the information in their profiles. That brings us to the following research question:

- How much “private” information can be gained from Twitter messages and profiles?

This research question can only be answered when the following sub-questions have been answered:

- What information is stored in a Twitter message?
- If an user is not sharing some information, is it still possible to extract this data with a decent degree of accuracy?
- What is the current state of the art and what research has been done so far in extracting user information from Twitter?
- What are the possibilities of extracting further user information from a Tweet?
By answering these research questions it should be clear what information can be gathered from social media and what the potential is to obtain different information.

This paper will start off with a scope of what will be considered in this research, followed by the methodology used in this paper and what the intended results for this research are. After that, there will be a section that covers all relevant literature and related research. The second section will be an explanation of what information in a Tweet can be used to extract private information from an user. Finally the paper will conclude by answering the research questions.

1.3 Scope
In this paper we only focus on Twitter and gaining information from Twitter profiles. The reason for this is that Twitter data is easier to gather and most of the Twitter data is open, so it is easier to check for privacy concerns on Twitter since almost all the data is open and therefore the same for each person. Based on this, one can decide what information can be mined from a Twitter message has to be decided accordingly. Some information is significantly harder to get than others and might not be accurate enough to use in order to get a good profile of the user. If the information is not accurate enough then the information is not used. The goal is to acquire as much information as possible from an user, as long as it is accurate enough.

How much information can be gathered from the profile will depend a lot on how much useful information can be found in the JSON. A JSON is a huge String of data that contains all information that is stored in a single Twitter message, but this also includes all information from the user profile, such as amount of followers and total number of replies to others. There exists some research on what can be obtained from certain fields of the JSON. We cover these in the literature review and later on will be a discussion on how these can be improved or whether they are good enough. Besides that the fields in a twitter message that are not used in literature, but that can contain useful information will also be covered.

1.4 Methodology
First, in order to analyse what data can be obtained from a Twitter message, this paper is going to look into the Twitter API also known as JSON. These JSON messages will be stored in a database that will later be used to obtain information from the JSON. Twitter does support the gathering of parsed JSON data, however this is not considered desirable since some of the information could be wasted. We perform a literature review done by searching through papers and covering all relevant work in order to profile Twitter users. All this literature describes the ways in which the JSON has been used.

After the literature review, there will be a section on what user information can be inferred from the JSON data that is more than what the raw JSON data provides, and the methods previous research suggest. Additionally we explain how certain information can be extracted and inferred even when there are not used or was considered as not very accurate in the previous literature. Finally, there will be a conclusion on what can be extracted from a Tweet and what future research can cover in this research field.

2. LITERATURE REVIEW
The literature review was done in a structured way to ensure coverage of the available literature already done on the subject [8]. In order to do so, we searched Scopus using the search terms: “(Twitter OR tweet) AND (privacy OR “private information” )”, “Twitter AND Demographics”. The results combined gave around 372 results, however a lot of the papers were not relevant for this paper. This can be seen in Appendix B.

The inclusion criteria for a paper to be part of the results were:

- They needed to work with data that was in the JSON.
- An algorithm or method of analyzing the data was discussed in order to infer new information.
- The research used one of the fields present in the Twitter data as a basis for their analysis
- The paper extracted and inferred information that was not present in the raw JSON data.

When these criteria were met, the paper was added to the list of results in this literature review. The number of papers on Scopus found by using this method was close to 10. The next step was to perform backward (considering the papers cited) and forward searches (the papers that cited the paper). By doing so, we ended up with 16 papers. After we used Google Scholar to find some more literature since Scholar is shown to complement Scopus, we added 4 more papers to our literature review. The final process can be seen in Figure 1.

![Figure 1: the process of finding the papers for the literature review](image-url)
2.1 Results of the Literature Review

Table 3: What private information is extracted by which related research and what Tweet Field is used by this.

<table>
<thead>
<tr>
<th>Paper</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Name</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Followers</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Extracted Data

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>E-mail</th>
<th>Personality</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Currently there are loads of ways to extract information from all different kinds of social media. On Facebook there is a function that you can add new people based on an e-mail address. Macro Balduzzi [2] mentions a solid approach on how this query can be abused in order to obtain the e-mail address of users. The approach used to obtain the e-mail address of an user was by obtaining the first and last name of an user. The algorithm is used to determine 24 common e-mail addresses by using the first and last name and then abbreviations such as “_”, “-” and “Firstletter.lastname” etc. To verify this, the e-mail can be sent over the network to check if the e-mail address matches the user found earlier. By using this e-mail, the miner can extract information from all sorts of social networks and extract information from each profile one is interested in. Once an e-mail address has been found it can be used on all websites, since e-mail addresses are unique to a specific user.

Daniele Perito [14] covers a research on how unique and traceable a username can be. They check the likelihood of matching a username from an online service A, onto a username in a different online service B. Twitter uses usernames just like a couple of other social media sites, such as Linked In and Youtube. When a username is found on Twitter and there is not much information known about the person except for his full name, the username could be used to find more information about the person and thus create a more complete user profile. However they find that most usernames are not unique, since users on different websites get to pick their own. A person that joins in the earlier stages of an online service, gets to pick their first name + last name as an username, while a later user with the same name can no longer pick this username as it is already been taken. This causes confusion when one tries to extract more information about a user based on the username as one cannot be completely certain if it is the same person.

Daniele Quercia [15] suggests, “Personality is correlated with a number of a real-world behaviors. It Correlates with music taste: popular music tends to be significantly liked by extroverts, while people with a tendency to be less open to experience to prefer religious music and dislike rock music” [16][Introduction first line]. They base the personality on three factors that are available on Twitter namely: number of profiles the user follows, number of followers and number of times the user has been listed in others reading lists. The paper mentions that the accuracy of the prediction about the personality is as low as 0.88 on a scale from one to five. The information that is extracted with this algorithm can be used in combination with all previous mentioned information to make an even more complete profile picture. Aron Culotta and Nirmal Kumar Ravi [3] also use the followers of a Twitter user to get more information about their user profile. Their research, on predicting the demographics of Twitter Users from Website Traffic Data [3], starts off by using the data from Quantcast.com [20], that is a website that uses cookies to track browser activities of a large panel of respondents. Following the authors get the Twitter accounts of the users found through Quantcast.com. They then executed a Twitter REST API call (followers/ids) used to collect 120 followers of each account to augment their dataset. This information can however be used to identify friends of a Twitter user. That is where vector data comes in to decide the age-brackets, gender, income and ethnicity of a person. The authors use five-fold-cross-validation method and the results are strong with an average correlation coefficient of 0.77. These methods can be used to gain more information about users by using Twitter that should increase the amount of private information that can be said about a person based on a Tweet.

Other researchers have tried to identify demographics of Twitter users [11]. The research was based on a large US population and tries to see how representative this is with known statistics from the US. They use a dataset containing 1.8 billion tweets spread out over 55 million users. The aim is to try and identify an entire population that also includes properties such as socio-economic status, education level and type of employment. The authors determine the race/ethnicity was of a Twitter user by comparing the last name of the Twitter user (the last name is part of the JSON of a tweet) to the data from the U.S. 2000 Census [19]. This was the first step to identify the race/ethnicity of a user and it worked for 71.8% of the users. The next demographic the authors analysed was the gender of a user. The way they did this was using the first name of a Twitter user (again coming from the JSON). And comparing the first name of the user with the 1000 most popular baby names from 1900 till 2009. When doing this it is important to take into account that some names can be used for both male and female. The way this was done was that if the names were less than 95% predictive, the name would be removed from the list. Secondly, many Twitter users use false names and therefore the algorithm would give a wrong outcome when deciding the gender of this user. Also, it is possible that the algorithm misinterprets a name. A second research was done in order to decide the gender of Twitter users. Alexander Panchenko and Andrey Teterin [13] tried to identify the gender based on full names, however they did this with only Russian names. Their approach was different.
from the system that was used on the US names dataset. The first step of their algorithm was to check the ending of each name to see if it ended with an ‘a’. For example Alexander is a male and Alexandra is a female. The second way was to check the character n-grams. These are certain patterns in names that decide if a name is male or female. Lastly, they performed a dictionary attack, which means that they used a dataset with many names (90000 in this case) and calculated the probability that a name is male or female. The accuracy of their gender calculation was 95%, which is a really high percentage. Wendy Liu also did research on how to decide the gender of a user based on their name. [10] They covered three different systems used in other papers. All of these systems used the SVM classifier feature. [7] What they wanted to test was to check if the username added extra accuracy in deciding a gender. A few problems arose here, namely, that usernames could be non-name related things such as: “swagboiboo504”. Kazushi Ikeda also worked with the SVM classifier [5]. He proposed a methodology to determine the status a person by the terms the person often tweeted. The statuses they suggested to check for, were, for example: Single/relationship, Different ages, Hobbies and whether they were still in university, working or were jobless. They had estimated percentages on how different sections would score and what the methods were.

Analyzing text is something that is very useful when it comes to predicting something about a user. The JSON data of a tweet does not contain a lot of personal information, there are a few fields that could be concluded to be personal information (as mentioned before), however if things can be said based on the text, it would mean a single Tweet would say a lot about the user. Luke Sloan [18] researched this matter in his paper. They analyzed the occupation, social class and age of Twitter users. Their first step was to analyze what the user’s occupation was. The way they did this was by looking at the profile message field in the JSON and then looking for the first mention of a job that comes in the NS-SEC [22]. The approach used to check how old a user is, is far from perfect, but definitely a good start. The approach consisted of checking the profile and text for certain keywords that indicated the age of a person, such as: “I am X years old”, “Born in XXXX” and “X years old”, besides looking for the usual DD/MM/YYYY pattern. The result of the research consisted of 1470 cases out of 32032, where the text mentioned age of the user. The prediction rate for this method was not high enough to be consistent in profiling Twitter users. Argamon analyzed [1] 140 million words and studied how they were used by different ages and genders. They looked at different factors and how these were linked to specific gender and age. Nguyen, Dong, et al [12] continues from the results of the previous paper, they tried to look into finding out the age and gender of a person based on an user’s tweets. They use machine learning in order to train their algorithm to recognize patterns about their age. The test set they use has the same distribution as the actual test set, in order to verify the data. Their results are good, with the accuracy on age being as high as 78%. They use error analysis to check the false positives found. After this they compare the performance of the system with that of a manual prediction of humans on the task of inferring the age of users only based on tweets. The conclusion of this was that manual predictions of humans had a lower accuracy percentage than the algorithm that was trained to do so. The last part of the research was to look into how the writing style of a user changes as the user gets older. The authors analyze variables such as capitalizing words, intensifiers, word length and the average tweet length. The only remarks on this research were: that they did not check the current users get older. Meaning that that different generations can have different word usage when they get older. Instead they looked at a different ages and this could also influence how users Tweet as they have grown up differently. Schwartz also works with tweet text in his paper [17]. In his research he concludes that males use certain words more often than females and the same can be said about different age groups. Emoticons, for example, are more often used by females, than by males. He covers many earlier researches that have done research based on text and also mentions that his results contradict most of the findings from literature. The conclusion of all this means that different age groups have different reasoning when it comes to male/female as well, since one of the studies amongst teens. [4]

3. WHAT INFORMATION IS STORED IN A TWITTER MESSAGE?

Appendix A contains the JSON data in a raw Tweet. There are a lot of fields in a tweet that contain information that can be useful or that can be used to track a certain tweet. However for this research there are only a few fields that are going to be considered, since these fields contain private information and are normally not shown in a message and are visible only when one has a deeper look into it, for example by using the JSON data of the tweet. [21]

The first field that contains private information is the “source” field. The source field shows from which device or webpage the tweet has been sent. This means that if there is a set of tweets from different users, one can find out what devices each user used to tweet. This can be used to determine whether users are utilizing iPhones, Samsung’s or using a computer. This research is going to be focused on finding the users that use iPhones and Samsung devices, since this a piece of information is normally not known to other Twitter users and therefore can be considered as private information.

The second field that has the potential to have private information inside it is the field “name” in the user section of the JSON. If an user has filled in their real name and not a nickname or a bogus name (which a fair amount seem to do), then the gender of the person can be determined with an algorithm that checks the first-name of the person and compares this with a name dictionary. However, using this technique there can be a good amount of users whose gender cannot be determined, but these will be set to unknown so that if they have a name it can be determined if it is a male, female, mostly a male, mostly a female or unisex. A matter of complication that one has to keep in mind that there are plenty of names that can belong to a male and female.
The “location” field in the user section of the JSON shows where the user tweeting lives (once again if the field has been filled in correctly). What makes this so important is that this information can be used for targeted marketing. If a company gathers many tweets, it will be simple for example, to decide the number of android users in the Netherlands (as described in the source field). With this the privacy of users is even more compromised, since it will also affect their daily lives if there is targeted marketing in their region or cities. The information can also considered private as normally when just reading a twitter message, the city where the person lives should not be known.

The fourth field that can be potentially used to extract data is the description field. In the description field the user can set any message they want to be shown on their profile. What some commercial organizations would like to know is the birthdate of a person, since this is really valuable to companies in combination with the other fields. If many users provide their birthdate in the description field, then with reasonable certainty, one can determine what the user’s age is, with the help of an algorithm.

Lastly, there is the Geotag field. This is not always filled in every Tweet, but it can be filtered when harvesting Tweets. This is especially handy, when there is more than one tweet available from a person, since the Geotag contains the location the user Tweeted from. The location of the Tweet can say a lot about a person’s whereabouts in a city. For example if every day between 9:00 and 17:00 the person Tweets from the same location, one can determine that the person is working at this location. And if the person is tweeting every morning and evening from the same location, one can determine that this is the place where the person lives. Eventually, if you gather enough Tweets about one person, you can tell what this person does on each day of the week (however, this possibly requires a lot of Tweets from a single person).

4. RESULTS
This paper covered all related research done that dealt with privacy and Twitter. There were multiple researchers that were using methods to extract new information from the JSON data of a Tweet and a lot of these researchers used different kind of fields from the JSON. There are however, some fields that are not used so far, as explained briefly in the previous section. There is a lot of possibilities to extract the private information from a Tweet, however, many of these ways require a lot of data in order to be valid and accurate and this data is hard to come by. To answer the main research question, previous research has shown that a lot of information can be deduced from a twitter message however, the accuracy of these methods are not good enough yet in order to call them good. The methods discussed in the literature review all have their flaws and in order to have attain high accuracy and still mention a lot about an user, they will need to be updated and improved. The gender deciding algorithms are the closest that come to meeting a high accuracy target. Many of these algorithms are open-source, however most of the scripts are still language based and there is a need for a global algorithm that always decides the name independent of the spoken language of the current user. A good step would be to gather the data used by the different researchers and combine this data in order to make a more global working algorithm. The age of a person can be decided most of the time by a machine learning algorithm, however these algorithms are not strong enough yet because of the lack of validated data. Clearly, in order to have a good machine learning algorithm working, there will need to be a huge supply of validated data. The researchers that use the amount of followers and the things Twitter users follow can tell more about the hobbies and personalities of a user. To start off with the personality prediction tool that is used is fairly accurate and can be used to decide the music that a user likes. When it comes to analyzing text however, the accuracy of these researchers are still low and there is a lot of work to be done in order to get text analyzing algorithms to be more accurate.

5. CONCLUSION & FUTURE WORK
As mention in last section, there is a lot of future work to be done about the kind of private information that can be extracted from a Tweet. The potential of the information stored in a Tweet is humongous, but currently the algorithms that have been discussed are not always capable of providing enough accuracy to extract this information from the Tweet. More data collection will most likely sort a lot of these problems, however the improvement of existing algorithms or the making of new ones is definitely important as well. Besides this, there are still fields in the JSON of a Tweet that are not used or not even considered. These fields however do contain a lot of potential information and future research should look into these fields and determine the best way is to use them, in order to extract more information from a user. This provides a challenge that will need to be worked on and hopefully users will realize how much information they are giving away by allowing Twitter to have their information. We hope that soon Twitter will make adjustments to allow users to make their profile more private so that researchers and companies cannot abuse the data gained from Twitter.

6. ACKNOWLEDGMENTS
A special thanks to Chintan Amrit from the University of Twente for supervising me in this paper. His experience on this subject and idea’s made it possible to come to this paper. And also a thanks to fellow students who took the time to give feedback on the paper in order to make it stronger.

7. REFERENCES


[21] https://dev.twitter.com/streaming/reference/get/statuses/sample


APPENDIX A: a Raw Tweet
{
  "created_at": "Wed May 13 14:45:56 +0000 2015",
  "id": 598499422906019841,
  "id_str": "598499422906019841",
  "text": "Ayer vi Midnight y me enamoré, hoy vamos con Dios Cagney y Borgart en The Roaring Twenties.",
  "source": "<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>",
  "truncated": false,
  "in_reply_to_status_id": null,"in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 396463171,
    "id_str": "396463171",
    "name": "Adri Hepburn",
    "screen_name": "adricasanova7",
    "location": "Zihuatanejo",
    "description": "Girl, you'll be a woman soon.// Say My Name.// seriéfilo, cinéfilo, melómano, Atlético y soy Batman. CELTICS.",
    "protected": false,
    "verified": false,
    "followers_count": 558,
    "friends_count": 821,
    "listed_count": 7,
    "favourites_count": 7036,
    "statuses_count": 53961,
    "created_at": "Sun Oct 23 09:29:03 +0000 2011",
    "utc_offset": -7200,
    "time_zone": "Greenland",
    "geo_enabled": true,
    "lang": "es",
    "contributors_enabled": false,
    "is_translator": false,
    "profile_background_color": "022330",
    "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png",
    "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png",
    "profile_background_tile": false,
    "profile_link_color": "3B94D9",
    "profile_sidebar_border_color": "A8C7F7",
    "profile_sidebar_fill_color": "C0DFEC",
    "profile_text_color": "F00D05",
    "profile_use_background_image": true,
    "profile_image_url": "http://pbs.twimg.com/profile_images/591719282679160833/ErTuYLYX_normal.jpg",
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/591719282679160833/ErTuYLYX_normal.jpg",
    "profile_banner_url": "https://pbs.twimg.com/profile_banners/396463171/1428617584",
    "default_profile": false,
    "default_profile_image": false,
    "following": null,
    "follow_request_sent": null,
    "notifications": null
  }
}
"hashtags": [],
"trends": [],
"urls": [],
"user_mentions": [],
"symbols": []
},
"favorited": false,
"retweeted": false,
"possibly_sensitive": false,
"filter_level": "low",
"lang": "es",
"timestamp_ms": "1431528356659"}