Universiteit Leiden
Opleiding Informatica

Analyzing privacy awareness of Twitter users through their given location precision

Name: Erik Soelaksana

Date: 21/08/2017

1st supervisor: Frank Takes
2nd supervisor: Walter Kosters

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Analyzing privacy awareness of Twitter users through their given location precision

Erik Soelaksana

Abstract

In this thesis we look to identify characteristics of a user group that may be relatively careless about the dangers of revealing privacy threatening information on a public domain. To do this, we use machine learning to discover patterns in the group of Twitter users that has their location set to their exact coordinates. We explore a dataset of the university of Illinois and from this data we extract more set features. To improve machine learning efficiency we use undersampling and oversampling through SMOTE, after which we apply machine learning through Weka using ZeroR, OneR and REPTree classifiers and evaluate their results.

Contents

Chapter 1

Introduction

This chapter contains an introduction to the thesis, an explanation of the terminology used, personal motivation and an overview of the thesis.

1.1 Twitter and Privacy

In the days of social media and an evergrowing number of people who access the internet daily, Twitter is a media giant whose influence cannot be ignored. With 300 million daily users[1] and 500 million tweets per day[2] the amount of information that passes through the website's domain is staggering. Users post about many things ranging from daily meals to promotions to whatever it is they are currently doing or thinking about. But because users are constantly using the Twitter platform to spread their current thoughts, they also constantly provide the world with an ever more detailed account of their lives. Where they live, where they currently are, what they're doing. And it's not just Twitter who logs these events, there are also many databases already established with a large amount of information from past years and many more databases currently being built using Twitter's own API.

There have been numerous scientists who have looked into what is possible with these data collections, such as determining a user's location accurate to a hundred miles using only their tweets [LWD+12], predicting personality disorders based on the vocabulary of a user [SBBJP12] and much more. There has however been relatively little research done concerning what determines if a user is more or less susceptible to part with information that is potentially privacy-threatening on Twitter. But what information is potentially privacy-threatening? Are active users more likely to part with this information than inactive users or the other way

[1] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
[2] http://www.internetlivestats.com/twitter-statistics/

around? Does it matter how long the user has been registered on Twitter? That is the purpose of this thesis, to try to discover if any of these factors matter significantly, or if none do.

By answering these questions we can provide insight for what type of behaviour can be typical for a user that easily gives away privacy threatening information.

We will try to identify these relations by analyzing a database of around 3 million users, 284 million relations between these users and 50 million tweets.

Possible uses for these results are identifying a target audience for privacy awareness, understanding Twitter user behavior better and possibly identifying factors determining users privacy awareness on a broader scale.

This thesis was written as a Bachelor project at the Leiden Institute of Advanced Computer Science (LIACS) of Universiteit Leiden, and has been supervised by Frank Takes.

1.2 Research Questions

1. Can we find a type of privacy threatening information Twitter users regularly give away?

2. Can we identify the users that reveal this information through machine learning by looking at patterns in their Twitter behaviour?

We believe these to be the base questions. First of all we need to identify what we are looking for, for example, when users use the hashtag 'Vacation' they let readers know that the user is not currently home, or even that they are at a specific place, both scenarios can be taken advantage of by readers with ill intentions.
Then we need to relate this field to other behavioral patterns which we will discover through machine learning algorithms.

1.3 Terminology

The following table contains the Twitter terminology used and an explanation for the terms[3].

[3] https://support.twitter.com/articles/166337

Term	Description
User	An account on Twitter
Profile	A profile displays information a user chooses to share publicly, as well as the Tweets they have posted.
Tweet	A tweet may contain photos, videos, links and up to 140 characters of text.
Follow	Subscribing to a Twitter account is called "following." Users can see tweets of an account they follow as soon as they post something new. Anyone on Twitter can follow or unfollow any other user at any time.
Follower count	How many users follow another user.
Location	The location the user has entered in their Twitter profile

Table 1.1: Twitter Terminology

1.4 Personal Motivation

I find it fascinating to see the amount of information being put on the internet willingly by people all over the world. Although I definitely suspect certain factors to be influential (common sense says that more active users will eventually give out more information) I am very curious to see if there is hard data backing up this suspicion. I was also surprised to see a lack of research focusing on this topic, which makes it all the more interesting to do my own exploration within this subject.

1.5 Thesis Overview

This chapter contains the introduction; Chapter 2 discusses related work; Chapter 3 discusses the search for and preparation of a dataset; Chapter 4 explains what methods were used to gain results; Chapter 5 details the application of the methods; Chapter 6 evaluates our results; Chapter 7 concludes.

Chapter 2

Related Work

Twitter is a very popular platform to perform analysis on due to its wide range of data that literally grows each second. Looking purely at a users' Tweets you can derive their location (Cheng, Caverlee, Lee, 2010), estimate if the user has a certain social disorder (Sumner, Byers, Boochever, Park, 2012), and even just by looking at their location field, which generally contains false information, it is possible to find a general actual location (Hecht, Hong, Suh, Chi, 2011). The ranging field of research surrounding Twitter results in a myriad of datasets being available of which one was fit for this research (Wang, Deng, Wang, Chang, 2012), the reasons why will be expanded upon in chapter 3.

Previous research has been done to identify how much users give away in their tweets (Humphreys, Gill, Krishnamurthy, 2010), which concluded very little personally identifiable information was given away in tweets, but an eighth of their sampled tweets gave away information regarding the location of a user. This, in combination with research showing individuals are not adverse to giving away their location (Barkhuus, Brown, Bell, Hall, Sherwood, Chalmers, 2008), gives us a clear sign that many people do not see location sharing as potentially privacy threatening.

Previous work has also been done to identify what users give away their location on Facebook to the public, that is to say that you do not need to be friends with that user (Kostakos, Venkatanathan, Reynolds, 2011). They concluded that more vocal and active users reveal their location more often to non-friends.

These previous works give a strong incentive to investigate Twitter, to see if we can identify location sharing users.

Chapter 3

Dataset

This chapter details the process of finding and preparing a dataset with sufficient Twitter users and Tweets to analyze.

3.1 Finding a Dataset

The first step is to find or create a large dataset that contains the information that we want to analyze. Twitter's API has a rate limit of 15 requests per 15 minutes. Searching for a user once is one request. If we were to request 300.000 users through Twitter's API we would have to be active for 209 days just to get a list of users, after which we would still have to look for their Tweets and relations. This did not seem feasible, so we had to look for an already established dataset.

There are many datasets available, but few met our requirements. We needed a set that was free of charge, had a large number of users, a large number of tweets and includes as much information per user and tweet (such as retweet count, follower count, tweet count etcetera) as possible. A set from Delft[1] had a very large number of tweets but had relatively few user profiles. A set from Arizona[2] merely had user relations. A set from National Institute of Standards and Technology did have a large amount of information but was created based on an agreement that posed too many limitations for future work. None of these sets met our requirements.

We did eventually find several datasets from Illinois[3], Haewoon[4], Xufei[5] and Caverlee[6] of which Illinois provided the largest amount of user profile data by far, plus it had already filtered for users that have a string

[1] http://www.wis.ewi.tudelft.nl/tweetum/
[2] http://socialcomputing.asu.edu/datasets/Twitter
[3] https://wiki.cites.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012
[4] https://datahub.io/dataset/twitter-social-graph-www2010/resource/f6d608c5-293e-4336-8e06-1cf8a4877a3c
[5] http://dmml.asu.edu/users/xufei/datasets.html
[6] https://archive.org/details/twitter_cikm_2010

in their location field. Sadly this set did not have geolocation stamps for tweets, however considering the fact that in 2014 only about 1.5% [HE14] or even 0.7%[7] of all users actually used these stamps (and an even lower number in 2011) this would have not provided a large amount of extra data and the data available was sufficient for the purposes of this research.

3.2 Dataset properties

The dataset we used was in 2011 by students of the university of Illinois. 100 thousand users were randomly selected as seeds, from there 284 million relations were crawled among 20 million users. Of these, they crawled the profiles of 3 million users who have at least 10 relations in the dataset. Of these, 150 thousand users who had a location set were selected to crawl their public tweets. For each user, at most 500 tweets were crawled[8].

Due to this set up the dataset is biased towards users who are at least slightly active (users have at least 10 followers or friends) and have a non-empty location field. Furthermore, there is a limit of 500 tweets per user, which limits analysis of behavior looking at the amount of tweets sent out.

3.3 Preparing the Dataset

The above dataset came in three parts: the user profile dataset, the relations dataset and the tweet dataset. These needed to be reformatted so we cross reference them and to be able to work with them more easily.

3.3.1 Profiles

To analyze the profiles dataset we needed to have a uniform profiles file. Because Twitter allows for a very large amount of varying characters (including tabs, while the file is tab separated) in certain fields (location, tweets) we had to convert these fields into fields that we can uniformly import with a common delimiter. We moved through the entire profile database to remove unwanted tabs, allowing the set to be imported using tabs as delimiters.

[7] http://altmetrics.org/altmetrics15/haustein/
[8] https://wiki.illinois.edu//wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012

Two sample lines of the dataset:

ID	Username	Friends	Followers	Status Count	Favorites	Account Age	Location name
100008949	esttrellitta	264	44	6853	0	28 Dec 2009 18:01:42 GMT	El Paso, Tx.
132241271	yesimichelle	210	39	458	0	12 Apr 2010 17:50:25 GMT	Texas

Table 3.1: Sample lines from the Twitter dataset

3.3.2 Tweets

The tweets dataset consisted of around 150 thousand plain text files, one for each user as identified in section 3.2. Each text file contained at most 500 tweets of that user in the following format per tweet:

```
***
Type:status
Origin: listen live #debtceiling #politics : http://www.whyy.org/91FM/live.html
Text: listen live  :
URL: http://www.whyy.org/91FM/live.html
ID: 96944336150867968
Time: Fri Jul 29 09:05:05 CDT 2011
RetCount: 0
Favorite: false
MentionedEntities:
Hashtags: debtceiling politics
***
```

Fieldname	Description
Type	Always status (indicating a status update on Twitter)
Origin	Full tweet
Text	Text in the tweet without special handles such as mentions and hashtags
URL	Used to be a direct URL to the tweet, however these are no longer functional
ID	Tweet ID
Time	Time and date of posting
RetCount	How many times the tweet has been retweeted
Favorite	Whether the Tweet has been favorited by other users
MentionedEntities	Other users mentioned in this tweet
Hashtags	Hashtags used in this tweet

Table 3.2: Tweet field description

Due to the large total combined size of these tweet files we did not combine them into one large file but converted these files to thirty large files, each containing of around 5000 users, into a traditional comma-separated values format, allowing us to use it for analysis. There were however issues when appending these files due to some lines in certain files deviating from the standard format. These were relatively few instances and only caused minor issues in most of the thirty large files, however in two of these it became very time

intensive to manually check and correct the files, so we chose to not use these two files. This still left us with more than 125 thousand users.

3.3.3 Network

The network set was simply one large (5.5GB) file containing two columns, Follower and Friend. The follower is the user following another ID, who we call the Friend. The only issue encountered here was the large filesize, but we circumvented this issue by loading the file in chunks.

Follower	Friend
12	260009730
12	17568791

Table 3.3: Sample lines from the network dataset

3.4 Generating new fields

We wanted to combine these sets to give us a meaningful total set of values we can use for data mining. In the following sections we will explain what fields we generated and how we generated these fields.

3.4.1 Location Determination

We first started by assigning groups to users based on how accurately they had set up their location field on their profile, so we had to find a way of identifying what kind of information the users had entered. There are several APIs available to reference services such as Google Maps or OpenMaps, but most of those have a finite number of requests you can make per day without spending a large sum of money. In the case of Google's API it's 2,500 requests per day, calculated as the sum of client-side and server-side queries[9] and the website for OpenMaps explicitly states "No heavy uses (an absolute maximum of 1 request per second)"[10]. With a dataset of 3 million users it would take far too long to identify each user's location. Yahoo currently offers an API that allows verification of locations and seemed to be perfect, but sadly it was also restrained by a limit of 2000 requests per day[11]. However, before 2014 they simply had a downloadable large dataset in comma-seperated values format. Considering the fact that our Twitter dataset was created in 2011, the 2012 version of Yahoo's Geoplanet set seemed a good fit.

[9] https://developers.google.com/maps/documentation/geocoding/usage-limits
[10] https://operations.osmfoundation.org/policies/nominatim/
[11] https://developer.yahoo.com/geo/geoplanet/

If we did have access to unlimited requests from one of the limited APIs, we would have had access to a more up to date database which contains more information about locations such as population of the location, which would allow us to more accurately determine the intended location, which in turn could have increased the accuracy of our research.

WOE ID	ISO	Location Name	Language	Placetype	Parent ID
1940542	KW	Al Ifhaihil	ARA	Town	55943081
1940539	KW	الفروانية	ARA	Town	55943082
23424909	NL	Nederland	DUT	Country	1

Table 3.4: Sample lines from the main Yahoo dataset

WOE ID	Name	Name Type	Language
1940542	Fahaheel	V	ARA
23424909	Niderlandy	Q	POL
23424909	Koninkrijk der Nederlanden	V	DUT

Table 3.5: Sample lines from the Yahoo Alias dataset

Fieldname	Description
WOE ID	Unique ID per location
ISO	ISO code assigned as per ISO 3166-1
Location Name	Name of the location, field with which we compare location in the Twitter set
Language	Language of the name in the Location Name field
Placetype	Category of place
Parent ID	ID of location of which this location is part of
Name Type	P is a preferred English name Q is a preferred name (in other languages) V is a well-known (but unofficial) variant for the place (e.g. ”New York City” for New York) S is either a synonym or a colloquial name for the place (e.g. ”Big Apple” for New York), or a version of the name which is stripped of accent characters. A is an abbreviation or code for the place (e.g. ”NYC” for New York)

Table 3.6: Yahoo Field Description [LWD+12]

We have two Yahoo datasets to use, the set with the main name of a location in its original language (Table 3.4) and a set with different names for the same location (Table 3.5). To be able to use these sets with our profiles set we had to change the format of our profile location field in order to cross reference them with the Yahoo sets. Due to the nature of a user-input field it was not possible to perfectly interpret the field, as many users do not have a single name in their location field like the Yahoo set, but rather a combination of terms such as ”Silver Spring, MD” or ”Chicago, Florida, NYC”, or they have coordinates in their location field such as ”iPhone: 39.053871, 95.674576”. To be able to use the location field, we chose to scan all fields for the following qualities and took action as follows:

1. If a field contains a pattern matching 1–3 digits, dot, 0 or more digits, comma, 1–3 digits, dot, 0 or more digits we recognize it as a coordinates field and change the users’ placetype to ’Coordinates’.

2. If a field contains a series of characters, then a comma, then any combination of characters, we discard

all characters starting at and including the comma. This gives us one term to work with.

We then compared Yahoo's name field from the first list to the Location field in the Twitter set and merged the two sets on every line with a match which allowed us to categorize the users into what kind of location they had set. However, when a user has set their location to 'San Francisco', although common sense would dictate they are referring to the state in the United States, this can refer to many others places all over the world, such as a town in Argentina or Colombia. This means we had to choose which type of location was most likely to be the location the user originally intended. We decided to choose the location based on the following parameters:

1. Location most likely to be correct due to general size of location type (e.g. it is more probable that the user is referring to the state Texas rather than the town Texas)

2. Location name being in English (because this is the most used language on Twitter[12])

3. Likelihood of a placetype being used as a locationname (for example, it is less probable that the user is using a land feature to describe his or her location rather than a town)

This resulted in the following order of which location from the Yahoo set we used:

Priority	Location Description
1	Continent
2	Country and name is in English
3	Country
4	US State and name is in English
5	US State
6	State
7	US County and name is in English
8	County
9	Town and name is in English
10	Suburb and name is in English
11	Town
12	Suburb
13	Point of Interest
14	Local Administration
15	Island
16	Colloquial name
17	Estate
18	Historical Town
19	Historical County
20	Land Feature
21	Supername

Table 3.7: Placetype Description

We decided that it would be more interesting to specifically look at the 125 thousand users who's tweets were collected because we simply have more information about these users. This resulted in Table 3.9:

[12] https://www.statista.com/statistics/267129/most-used-languages-on-twitter/

Placetype	Description
Continent	One of the major land masses on the earth.
Country	One of the countries and dependent territories defined by the ISO 3166-1 standard.
State	Main subdivision of a country, for example in the Netherlands these are the provinces.
County	Smaller subdivision, in the Netherlands an example is Gemeente Den Haag.
Town	One of the major populated places within a country. This category includes incorporated cities and towns, major unincorporated towns and villages.
POI	Points of Interests, such as the Statue of Liberty.
Suburb	One of the subdivisions within a town. This category includes suburbs, neighborhoods, wards.
LocalAdmin	Small group with local authority in a country, such as a Gemeinde in Germany (no relevant example in the Netherlands).
Island	Islands such as Schiermonnikoog Island.
Estate	Only categorised for USA and South Africa, an example in USA is Laguna Heights.
Colloquial	Examples are New England, French Riviera, 関西地方(Kansai Region), South East England, Pacific States, and Chubu Region.
HistoricalTown	Towns that no longer exist, only available for US.
HistoricalCounty	Counties that no longer exist.
Landfeature	Notable landmarks such as Nationaal Park de Maasduinen.
Supername	A place that refers to a region consisting of multiple countries or an historical country that has been dissolved into current countries. Examples include Scandinavia, Latin America, USSR, Yugoslavia, Western Europe, and Central America.

Table 3.8: Placetype Description [LWD+12]

Placetype	Number of users
Town	59221
County	41238
Coords	17362
State	8937
Country	797
Suburb	192
Colloquial	53
POI	12
HistoricalTown	1

Table 3.9: Placetypes

For each user, we added a column categorizing users into several groups of how accurate their location field was. We defined these as shown in Table 3.10. Note that we discarded the colloquial users because the type of location their colloquial name referred to was too heterogeneous to categorize.

1	Coordinates
2	Town, Suburb, Historical Town, Point of Interest
3	County
4	State or Country

Table 3.10: Categories of place accuracy

3.4.2 Tweet Statistics

From each users' tweets we extracted three fields, the amount of times their tweets had been retweeted, the amount of times they mentioned another user in a tweet using the @ handle, and the total amount of tweets of that user contained in our dataset. We added these fields and an average of retweets per tweet and average mentions per tweet to the main user profile set.

Because we were interested in how many times users gave away information about them being away from home, we then looked at all hashtags used by users and scanned whether the hashtags contained the strings 'travel', 'trip', 'holiday' or 'vacation'. For each user, we counted the amount of times this occured.

3.4.3 Relation Statistics

For each user, we looked at all users they follow and determined the average of the fields in Table 3.11 of all those users:

1	Retweets total
2	User mentions count total
3	Follower count total
4	Friend count total
5	Tweet count total
6	Year account created
7	Location accuracy as per Table 3.10
8	Average retweet per tweet
9	Average user mentions per tweet
10	Average times their friends used a vacation hashtag
11	Tweets favorited count total
12	Whether the user had their coordinates set as their location

Table 3.11: Fields averaged of relations

Then we looked at all users they follow and determined the sum of the fields in Table 3.12 of all accounts that they follow.

1	Retweets total
2	User mentions count total
3	Follower count total
4	Friend Count Total
5	Tweet count total
6	Total times their friends used a Vacation hashtag
7	Total times tweets of their friends were favorited

Table 3.12: Fields summed of relations

We then generated the same fields for all users following a specific user. This gave us two sets to work with,

one with information about followers, and one with information about the friends. Both of these had 26 attributes. We also combined these for a large total set with 45 attributes.

Chapter 4

Methodology

In this chapter we explain the methods used to obtain the data used in our analysis and evaluation.

4.1 Data Mining Methods

We used Weka[1] with instructions from Data Mining [WFH11] as a machine learning tool to identify patterns in our dataset. Once our dataset was prepared, we loaded it into Weka and used machine learning to discover patterns in user behaviour when they have coordinates set as their location field. We shall from now on refer to this group as the positive set.

Our dataset was very imbalanced, with 17362 users having their coordinates in their location field and 111261 users not having their coordinates in their location field. This resulted in skewed results, so we oversampled our positive user set by 200%, resulting in 52086 positive users. We used the Synthetic Minority Over-Sampling Technique, SMOTE [CBOHPK02], to generate these entries. We then also undersampled our negative set so there was a balance in 52086 users with and 52086 users without their coordinates as location.

For each stage (raw set, with SMOTE, with SMOTE and undersampling) we ran the dataset through three different types of machine learning classifiers, ZeroR for a baseline percentage to compare against, OneR to check if one field is heavily influential, and REPTree to obtain more advanced results.

[1] http://www.cs.waikato.ac.nz/ml/weka/

4.1.1 Classifiers

We will provide a simple explanation of the classifiers used in our experiments.

Majority Class (ZeroR)

The first classifier we used was the Weka implementation of Majority Class called ZeroR. It is the simplest classification method, only checking the absolute percentage in a binary problem[2]. As an example, refer to Table 4.1.

Outlook	Temperature	Humidity	Windy	Play Tennis
Rainy	Hot	Low	False	No
Rainy	Cold	High	True	No
Sunny	Hot	Low	True	Yes
Sunny	Cold	Low	True	Yes
Rainy	Hot	Low	False	Yes

Table 4.1: Sample tennis dataset

When trying to predict what weather will result in tennis being played, ZeroR will simply look at the 'Play Tennis' field and predict that because in the sample set 3/5 instances were 'Yes', 60% of the time tennis will be played. It will not take into account any other fields, just look at an absolute percentage of 'Yes' instances. This is a very simple classifier and is only used as a base measure in our experiments.

Decision Tree (OneR)

The second classifier was the Weka implementation of a decision tree, called OneR. It is a relatively simple classifier which checks what field is the best predictor[3]. It looks at every field and generates a rule. It then looks at what rule is most accurate, and chooses it as result rule. If we again look at table 4.1, when 'Play Tennis' = 'Yes', there is a 66% chance 'Outlook' = 'Sunny' . When 'Temperature' = 'Hot', there is a 66% chance 'Play Tennis' = 'Yes'. When 'Humidity' = 'Low', there is a 75% chance 'Play Tennis' = 'Yes'. When 'Windy' = 'True', there is a 66% chance 'Play Tennis' = 'Yes'. Because 'Humidity' = 'Low' gives the highest percentage of correct positive results, OneR returns the rule:

```
IF Humidity == Low THEN Play Tennis = Yes
```

[2] http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm
[3] http://www.saedsayad.com/oner.htm

Decision Tree (REPTree)

The third classifer we used was the Weka implementation of a decision tree, called REPTree. REPTree is a fast decision tree algorithm. It applies regression tree logic and generates multiple trees in altered iterations. From these iterations, it chooses the best tree [Dev15]. Furthermore, this tree allows us to define maximum depth, which prevents overfitting and overwhelmingly large trees being generated, which is an issue with other classifiers such as J48.

4.1.2 Filters

Because there was a large data imbalance in our set, with the majority class making up 86.5% and the minority class making up 13.5% of the set, we had to address this in order to make machine learning more reliable and effective. If there is a large majority class, machine learning will have a bias towards this class, creating unreliable results This section will give a short explanation on what the filters we used to combat the data imbalance.

SMOTE

SMOTE is a widely adopted technique that addresses class imbalances. It oversamples a minority class in order to balance a set. This is done by taking a minority class sample, taking a set of neighbours of this sample, and introducing synthetic examples [CBOHPK02]. We used this to triple the number of minority class samples.

SpreadSubsample

SpreadSubsample simply produces a random subsample of a majority set. We used this to perfectly balance our minority and majority sets.

Chapter 5

Analysis

This chapter explains in what way we analyze the data.

5.1 Machine Learning Results

We ran our datasets with the features from Table 3.12 and Table 3.11 through Weka. We did this with three different filters settings (Table 5.1) and three different classifiers (Table 5.2). Each combination was run ten times and the average result was taken as final result. Note that the filters changed the amount of instances available, as noted in the table. Because of the way SMOTE works, synthetic samples had values with decimals, which is not a problem except for the "Account Year" attribute. Because users join in 2008, not 2008.3, we chose to round up the join year of all synthetic samples.

Filter:	Settings:	Total instances:
None	None	128623
SMOTE	Increase minority class by 200%	163347
SMOTE and SpreadSubSample	Increase minority class by 200% and reduce majority class to match minority class	104172

Table 5.1: Filters and settings

Classifier:	Settings:
ZeroR	Default
OneR	Minimum instances per node of 100, 500, 1000, 2500 and 5000
REPTree	Maxdepth of 4, 6, 8, and 10, and minimum instances per leaf of 100

Table 5.2: Classifiers and settings

5.1.1 ZeroR

All results for the ZeroR classifier can be found in Table A.1. These results serve as a baseline comparison for the other classifiers and are simply the total percentage of the majority class.

5.1.2 OneR

The results for OneR can be found in Table A.2. For the unfiltered sets, it cannot find a good rule and simply assumes the same as ZeroR, it takes the majority class and always predicts that outcome. For the filtered sets, however, we see a notable increase in correctly classified instances. For the bucketsizes of 100 and 500 OneR constantly overfits, giving meaningless models. But when we further inspect the models created with the larger bucketsizes they nearly all value the total amount of tweets sent by a user as the most important field, predicting that users that have tweeted less than around 1500 tweets total will be a negative instance and those with more than 1500 tweets will be a positive instance.

5.1.3 REPTree

The results for REPTree can be found in Table A.3. The initial results looked good for REPTree, having significantly better results than ZeroR and in many cases than OneR as well. To understand these results, we had to view the trees REPTree generated. After generating these trees we found that a $MaxDepth = 6$ was the best type of tree to analyze, as $MaxDepth = 4$ gave too little information and $MaxDepth = 8$ or 10 resulted in large trees with small leaves and overfitting. We detail our findings per dataset in the following sections. Note that by scoring high on a field, we mean that either the user has a higher number that implies being an active user (such as a high tweet count) or that the field is scoring high on accuracy in case of location accuracy. When talking about a group's negative counterpart, we are talking about a False group with the same characteristics except one.

Friends

The trees for the Friends dataset can be found in Figure A.1, Figure A.2 and Figure A.3. An explanation of the terms used in these trees can be found in Table A.4.

Unfiltered The unfiltered tree only has one path that leads to a positive node, and that node only has 447 instances of which 173 were incorrectly evaluated, out of around 11575 instances that should be evaluated as true. This is such a small group that it renders the tree with little value to determine anything about the

positive set, but we can extract information by looking at the negative instances. First of all, the total amount of tweets a user has sent is a good way to filter users, leading to 57% of all users being split off immediately with only 8% being incorrectly classified. This leaves the test set with 37,347 users. The next split occurs at the average amount their friends tweet, splitting off another 29,670, or 35%, users with 17% being incorrectly classified. Because this only leaves 7,677 users, the rest of the nodes are far less interesting.

This set implies that users who do not tweet a lot, or whose friends do not tweet a lot, are unlikely to have coordinates set as their location.

SMOTE The tree for the SMOTE filtered set also uses a users' total tweet count first, splitting off 38% with 15% incorrectly evaluated. Beyond this split there are no major single decisions used to sort the tree. There were however three paths leading to a True node.

The first (6452 instances) has a total tweet count of at least 976, mentions other users relatively frequently (≥ 11), their friends have a high average tweet count (≥ 12889) and their friends have a high location accuracy set in their profiles (< 2). These users score high in every field they get evaluated upon.

The second group (6251 instances) is nearly the same, but has its friends have a lower location accuracy (≥ 2.0), has been registered since 2009 or later and their total tweet count is at least 3385. This implies that users with few friends with accurate locations are still likely to have their coordinates set if they tweet a lot.

The third group (9569 instances) is largely the same as the second group with a minimum of 3969 tweets, but their friends' average tweet count is less (< 12889). This is the largest group, which may imply that friends do not influence a user as much as the users' own behaviour.

SMOTE and SpreadSubsample Again, a users' total tweet count is being used as the first split, taking 23% of the entire test set with an error of 22%. In this tree there are five paths leading to True.

The first group (10357 instances) is the similar to the first group in the SMOTE filtered set, many tweets (≥ 3471), relatively high amount of other users mentioned(≥ 14.0) and the average amount of tweets per friend is very high (≥ 11788). Although all of these users are being classified as True, it is notable that if their friends have an average location precision of less than 2, the amount of incorrectly classified instances decreases from 21% to 13%, implying that when friends have their location set accurately, a user is more likely to set their coordinates (the highest accuracy) in their location. This is again a group that scores high on all fields.

The second group (4440 instances) has slightly fewer tweets (< 3472) and has been a member of Twitter since before 2010 or earlier. Again, although this group is entirely classified as True, it can be more accurately

predicted when the location accuracy is higher. Without looking at the location accuracy, it has a 38% misclassification rate. With a location accuracy below 2.5 it has a 34% error rate.

The third group (14959 instances) is largely the same as the second group, but its friends tweet less ($<$ 11788), it has at least 1730 tweets and it has been a member since 2009 or later. This is the largest positive group in this tree.

The fourth group (7344 instances) is similar to the third, but have been members since 2008 or earlier and interestingly have fewer followers ($<$ 5582) than the group with the same characteristics but no coordinates set as location.

The final group (9354 instances) tweets less than the others ($<$ 1730), has fewer followers than their negative counterpart ($<$ 1622), but has been registered since 2010 or earlier.

Followers

The trees for the Friends dataset can be found in Figure A.4, Figure A.5 and Figure A.6.

Unfiltered This tree has some characteristics similar to the tree generated by the friends set. It has only one True node with a small number of instances and it has an identical first splits targeting the total tweet count of a user($<$ 1705). Followers having their coordinates set as location splits off 30152 (36%) users, which is around the same amount the second split in the friends set does. This means that unfiltered, REPTree can not gain much more information from the Friends or Followers set.

SMOTE The tree generated by the SMOTE set again splits off directly at the amount of tweets the user has sent. This takes 25% with a 15% error rate. There are no large splits afterwards. There are four positive groups.

The first group (8329 instances) has many tweets ($\geq$ 5303), more than 25% of their followers have coordinates set as their location, and have been members since 2008. This group is more accurate if their accounts have been created in 2009 or later, giving a error rate of 26% instead of 28%. This group scores high on all attributes and is the biggest positive group of this tree.

The second group (2950 instances) has fewer tweets ($<$ 5303), mentions relatively many other users in their tweets ($\geq$ 7.14) and has at most 9048 total followers of followers.

The third group (286 instances) is largely the same, but has more followers of followers ($\geq$ 9048), and less than 27% of their followers have their coordinates set as their location.

The last group (3252 instances) has many tweets ($\geq$ 3873), relatively high count of user mentions ($\geq$ 11.24), has been a member since 2009 or later and more than 17% of its followers have their coordinates set as their location.

SMOTE and SpreadSubsample Again, this tree splits off at the total count of user tweets (783). However, the split towards fewer tweets has a small subgroup of true instances. This is a very small node though, accounting for 275 out of the 20 thousand users in this split, and only 0.4% of the entire set. Because of the small size of this node, it is unlikely to contain significant data. Other than this group, there are 4 positive groups.

The first (12229 instances) has a relatively high tweet count ($\geq$ 783), mentions users relatively often ($\geq$ 11.1) and over 25% of its followers have their coordinates set as their location. This is one of the two largest groups in this set and scores high on all attributes.

The second group (13917 instances) has fewer followers with coordinates set ($<$ 25%), a tweet count of at least 2419 and has been registered since 2009. This is the second of the two largest groups in this set.

The third group (6260 instances) is similar to the second group, but has been registered since 2008 or earlier and has at most 5007 followers.

The fourth group (5241 instances) has fewer tweets than groups two or three ($<$ 2419), at most 1607 followers, but slightly more followers with their coordinates set as their location than their negative counterpart.

Friends and Followers

The trees for the Friends and Followers dataset can be found in Figure A.7, Figure A.8 and Figure A.9.

Unfiltered This tree continues the trend by splitting off at the tweet count of the user, classifying 57% as False with 8% incorrectly classified. The next split occurs at the average amount of followers with coordinates set as their location, classifying another 36% as False with 17% incorrect. This set does however have four positive groups, significantly more than the previous sets.

The first group (999 instances) scores high on all fields, with many tweets ($\geq$ 10832), friends that have an average high tweet count ($\geq$ 13369) and over 42% of their followers have their coordinates set as location.

The second group (687 instances) has fewer tweets ($\geq$ 1705 and $<$ 10832) than the first, at least 43% of their followers have their coordinates set as location and these users have been registered since 2009 or earlier.

The third and fourth group have a relatively high tweet count ($\geq$ 1705), have at least 26% followers with coordinates set as location, have been registered since 2009 or earlier and then split off into two True groups. For group three (176 instances), if their friends have a relatively inaccurate location ($\geq$ 1.61), the users' followers of friends total count must be high ($\geq$ 1699043). For group four (190 instances), if their friends have a relatively accurate location ($<$ 1.61), the users' friends of friends total count must not be high ($<$ 2389).

SMOTE As always, the tree splits off at tweet count first. This takes 38% of the test set with 21% incorrect. Like with the unfiltered set, this has a small True subgroup with a very small count. It is slightly larger relatively than the True subgroup in the unfiltered set. Its only unique identifier compared to the rest of the False instances is that it has between 0% and 0.07% of followers with their coordinates set as location. Other than this, it has four positive groups.

The first (7605 instances) has a high tweet count ($\geq$ 4880), over 25% followers with their coordinates set as location, have been registered since 2009 or later, and mention users often ($\geq$ 39.5). This is the largest group, scores high on all attributes and most notably this group mentions users very often compared to previous results.

The second group (1022 instances) is similar to the first, except it does not matter how often they mention other users and they have been registered in 2008.

The third group (3180 instances) has a relatively high tweet count ($\geq$ 1789) but a lower count than the first two groups, over 25% followers with their coordinates set as location, mentions other users relatively often ($\geq$ 9.5) and have been registered since 2009 or later.

The fourth group (6604 instances) has a high tweet count ($\geq$ 948), fewer than 25% followers with their coordinates set as location, mentions other users relatively often ($\geq$ 11), has been registered since 2009, has at least some followers with their coordinates as location and their followers mention other users at most 2407 times.

SMOTE and SpreadSubsample Finally this tree also splits at the total tweet count. It sets 28% as False, with 23% incorrectly classified. This tree also has four positive groups.

The first group (12447 instances) has a relatively high total tweet count ($\geq$ 784), mentions other users relatively often ($\geq$ 10.0) and more than 25% of their followers have coordinates set as their location. If a user was registered in 2009 or later and whose friends have an average total tweet count of at least 11856, we can reduce errors in classification to 13%, as opposed to 22% for the larger set.

The second group (17509 instances) has a relatively high tweet count ($\geq$ 1669), mentions other users relatively often ($\geq$ 10.0) and has been registered since 2009 or later. This is the largest group in this tree.

The third group (7544 instances) is similar to the second, except it has been registered since 2008 or earlier, and it has at most 4584 followers.

The fourth group (2076 instances) has fewer tweets than groups two and three ($\geq$ 784 and $<$ 1669), mentions other users relatively often ($\geq$ 10.0), their followers mention other users at most 2370 times and their followers have their coordinates as location more often than their negative counterparts.

Chapter 6

Evaluation

In this chapter we evaluate the results obtained in Chapter 5

6.1 Result Evaluation

One clear trend is that users that scored high in the fields of total tweet count, user mentions and average friend tweets tend to have a high chance of being part of the positive set. This would imply that users that use multiple features of Twitter would be more likely to have their coordinates as their location. The location accuracy of your friends or followers does not seem to have a strong effect on its own, but together with the abovementioned attributes it can help sharpen the percentage of correctly classified instances.

We have created a table out of the values important for positive groups in Table 6.1 and go slightly more in depth per attribute below the table.

Set:	Filter:	Group:	Total Tweet Count:	User Mentions:	Friends AVG Tweets:	Friends/Followers AVG Location Accuracy:	Instances:	Followers:	Friends/Followers with Coordinates:	Followers of Followers:	Year:	Followers of Friends:	Friends of Friends:	Follower User Mentions:
Friends	None	None												
Friends	SMOTE	1	≥ 976	≥ 11.0		< 2	6452							
Friends	SMOTE	2	≥ 3385	≥ 11.0	≥ 12889	≥ 2	6251				2009+			
Friends	SMOTE	3	≥ 3969	≥ 11.0	< 12889	≥ 2	9569							
Friends	SMOTE+Sub	1	≥ 3471	≥ 14.0	≥ 11788		10357							
Friends	SMOTE+Sub	2	$\geq 600 < 3472$	≥ 14.0	≥ 11788		4440							
Friends	SMOTE+Sub	3	$\geq 1730 < 3472$	≥ 14.0	< 11788		14959				2009+			
Friends	SMOTE+Sub	4	$\geq 600 < 3472$	≥ 14.0	≥ 11788		7344	< 5582			2008-			
Friends	SMOTE+Sub	5	$\geq 600 < 1730$	≥ 14.0	≥ 11788		9354	< 1622			2010-			
Followers	None	None												
Followers	SMOTE	1	≥ 5303				8329		$\geq 25\%$		2008+			
Followers	SMOTE	2	$\geq 970 < 5303$	≥ 7.14			2950		$\geq 25\%$	< 9048				
Followers	SMOTE	3	$\geq 970 < 5303$	≥ 7.14			286		$< 27\%$	≥ 9048				
Followers	SMOTE	4	≥ 3873	≥ 11.2			3252		$< 17\%$	≥ 9048	2009+			
Followers	SMOTE+Sub	1	≥ 783	≥ 11.1			12229		$\geq 25\%$					
Followers	SMOTE+Sub	2	≥ 2419	≥ 11.1			13917		$< 25\%$		2009+			
Followers	SMOTE+Sub	3	≥ 2419	≥ 11.1			6260	< 5007	$< 25\%$		2009+			
Followers	SMOTE+Sub	4	$\geq 783 < 2419$	≥ 11.1			5241	< 1607	$< 25\%$		2008-			
Both	None	1	≥ 10832		≥ 13369		999		Fol: $\geq 42\%$					
Both	None	2	$\geq 1705 < 10832$		≥ 13369		687		Fol: $\geq 43\%$		2009-			
Both	None	3	≥ 1705		≥ 13369	≥ 1.61	176		Fol: $\geq 26\%$		2009-	≥ 1699043		
Both	None	4	≥ 1705		≥ 13369	< 1.61	190		Fol: $\geq 26\%$		2009-		< 2389	
Both	SMOTE	1	≥ 4880	≥ 39.5			7605		Fol: $\geq 25\%$		2009+			
Both	SMOTE	2	≥ 4880				1022		Fol: $\geq 25\%$		2008			
Both	SMOTE	3	$\geq 1789 < 4880$	≥ 9.5			3180		Fol: $\geq 25\%$		2009+			
Both	SMOTE	4	≥ 948	≥ 11.0			6604		Fol: $< 25\%$		2009+			< 2407
Both	SMOTE+Sub	1	≥ 784	≥ 10.0			12447				2009+			
Both	SMOTE+Sub	2	≥ 1669	≥ 10.0			17509				2009+			
Both	SMOTE+Sub	3	≥ 1669	≥ 10.0			7544	< 4584			2008-			
Both	SMOTE+Sub	4	$\geq 784 < 1669$	≥ 10.0			2076				2008-			< 2370

Table 6.1: Positive groups characteristics

Total tweet count was the heaviest factor, if users did not have many tweets ($\geq$ 600 in our set), it was not part of the positive set, no exceptions. User mentions was the second most present factor, implying that users who mention other users often have their coordinates set as location more often. In many groups the average friend tweet was high, suggesting that users with more active friends will be more likely to be part of the positive group.

It is notable that having more followers has a negative effect on the chances of giving away coordinates, which could be because users with a large following shy away from revealing their exact location. Followers mentioning more users also seems to have a negative effect, so perhaps a more active follower base makes a user more wary of their private information. Friends or followers with coordinates set as their location was a split that was used often, but with no discernible trends. Interestingly enough, this parameter was always used with followers and never with friends. Although the rest of the attributes were not used often, all noted in Table 6.1 were used at least once, as opposed to the other 15 or 34 attributes.

Chapter 7

Conclusions

This thesis started with the main question ”Can we identify the users that reveal privacy threatening information through machine learning by looking at patterns in their Twitter behaviour?”, to which we can answer 'more often than not'. Although we did not find a perfect pattern to identify users, we did find several factors that are often present in these type of users' behaviour. The largest factors are whether users tweet often, mention other users often and have friends that tweet often. Furthermore, more active followers seem to result in a smaller likelihood of the user having coordinates as their location.

There were two major groups we discovered in our research. In all REPTree trees we found a group that scored high on all attributes they were tested upon, a high tweet count, friends with high tweet counts, followers with a high precision of location set, and so on. This implies that active Twitter users who use many features that Twitter provides are more likely to set their coordinates as their location. The second group was a group that often had a medium amount of tweets and notably they had less followers than a specific number or followers that were less active. This implies that users who have a smaller network are quicker to reveal their coordinates. An example user would be a user that uses Twitter purely with a personal network, not minding that his or her friends see the coordinates set.

A major obstacle encountered during this thesis were defining a users' location accuracy due to the heterogeneous way the location field was filled, following no general format, which required us to consider many different factors before we could make any statements about users. Furthermore, because we had a large amount of users, many of the actions we performed took several hours to complete, which meant that changing a small setting, due to either an error or a decision being changed, could mean many more hours of waiting and analyzing, requiring precise work to prevent this.

7.1 Future Work

For future work, it would be interesting to see what could be found on a more up to date set, as the average population is being exposed more and more to internet services that allow you to share your location, it could well be that there is a larger sample of coordinates sharing users now then there was in 2012. Furthermore, more accurate location recognition and more location data would be able to field more attributes, such as how large a town is that a user has specified as his location. It is also possible to focus less on coordinates as a location field and more on the general accuracy of a users' location, finding trends in groups of users that a town, suburb, country, etc. as their location field.

Bibliography

[CBOHPK02] N. Chawla, K. Bowyer, L. O. Hall, and W P. Kegelmeyer. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research 16*, 16:321–357, 2002.

[Dev15] C. Lakshmi Devasena. Comparative Analysis of Random Forest, REP Tree and J48 Classifiers for Credit Risk Prediction. In *ICCCMIT*, 2015.

[HE14] Jeff J. Hemsley and J. Eckert. Examining the Role of "Place" in Twitter Networks Through the Lens of Contentious Politics. In *HICSS*, 2014.

[LWD+12] R. Li, S. Wang, H. Deng, R. Wang, and K. Chen-Chuan Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *KDD*, 2012.

[SBBJP12] C. Sumner, A. Byers, R. Boochever, and G. J Park. Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In *ICMLA*, 2012.

[WFH11] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining, Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.

Appendix A

Set:	Filter:	Percentage of correctly classified instances:
Friends	None	86.50 %
Friends	SMOTE	68.11 %
Friends	SMOTE and SpreadSubsample	50 %
Followers	None	86.50 %
Followers	SMOTE	68.11 %
Followers	SMOTE and SpreadSubsample	50 %
Friends and Followers	None	86.50 %
Friends and Followers	SMOTE	68.11 %
Friends and Followers	SMOTE and SpreadSubsample	50%

Table A.1: ZeroR results

Set:	Filter:	Instances in bucket:	Percentage of correctly classified instances:	Difference with ZeroR:
Followers	None	100	86.50%	0.00 %
Followers	None	500	86.50%	0.00 %
Followers	None	1000	86.50%	0.00 %
Followers	None	2500	86.50%	0.00 %
Followers	None	5000	86.50%	0.00 %
Followers	SMOTE	100	81.22%	13.10 %
Followers	SMOTE	500	79.83%	11.71 %
Followers	SMOTE	1000	79.87%	11.76 %
Followers	SMOTE	2500	78.19%	10.08 %
Followers	SMOTE	5000	78.19%	10.08 %
Followers	SMOTE and SpreadSubsample	100	71.65%	21.65 %
Followers	SMOTE and SpreadSubsample	500	70.51%	20.51 %
Followers	SMOTE and SpreadSubsample	1000	69.63%	19.63 %
Followers	SMOTE and SpreadSubsample	2500	68.32%	18.32 %
Followers	SMOTE and SpreadSubsample	5000	65.75%	15.76 %
Friends	None	100	86.52%	0.02 %
Friends	None	500	86.50%	0.00 %
Friends	None	1000	86.50%	0.00 %
Friends	None	2500	86.50%	0.00 %
Friends	None	5000	86.50%	0.00 %
Friends	SMOTE	100	80.25%	12.14 %
Friends	SMOTE	500	80.10%	11.99 %
Friends	SMOTE	1000	80.10%	11.99 %
Friends	SMOTE	2500	78.46%	10.35 %
Friends	SMOTE	5000	78.46%	10.35 %
Friends	SMOTE and SpreadSubsample	100	70.58%	20.59 %
Friends	SMOTE and SpreadSubsample	500	68.64%	18.64 %
Friends	SMOTE and SpreadSubsample	1000	68.64%	18.64 %
Friends	SMOTE and SpreadSubsample	2500	67.83%	17.83 %
Friends	SMOTE and SpreadSubsample	5000	66.06%	16.06 %
Friends and Followers	None	100	86.52%	0.02 %
Friends and Followers	None	500	86.50%	0.00 %
Friends and Followers	None	1000	86.50%	0.00 %
Friends and Followers	None	2500	86.50%	0.00 %
Friends and Followers	None	5000	86.50%	0.00 %
Friends and Followers	SMOTE	100	81.36%	13.24 %
Friends and Followers	SMOTE	500	79.87%	11.76 %
Friends and Followers	SMOTE	1000	79.86%	11.75 %
Friends and Followers	SMOTE	2500	78.17%	10.05 %
Friends and Followers	SMOTE	5000	77.44%	9.32 %
Friends and Followers	SMOTE and SpreadSubsample	100	71.91%	21.91 %
Friends and Followers	SMOTE and SpreadSubsample	500	70.64%	20.64 %
Friends and Followers	SMOTE and SpreadSubsample	1000	69.85%	19.85 %
Friends and Followers	SMOTE and SpreadSubsample	2500	68.10%	18.10 %
Friends and Followers	SMOTE and SpreadSubsample	5000	65.63%	15.63%

Table A.2: OneR results

Set:	Filter:	MaxDepth:	Percentage of correctly classified instances:	Difference with ZeroR:
Friends	None	4	86.57%	0.07%
Friends	None	6	86.63%	0.13%
Friends	None	8	86.65%	0.14%
Friends	None	10	86.64%	0.14%
Friends	SMOTE	4	72.57%	4.46%
Friends	SMOTE	6	78.05%	9.93%
Friends	SMOTE	8	81.35%	13.23%
Friends	SMOTE	10	82.35%	14.24%
Friends	SMOTE and SpreadSubSample	4	69.70%	19.70%
Friends	SMOTE and SpreadSubSample	6	72.48%	22.48%
Friends	SMOTE and SpreadSubSample	8	75.12%	25.12%
Friends	SMOTE and SpreadSubSample	10	75.70%	25.70%
Followers	None	4	86.53%	0.03%
Followers	None	6	86.57%	0.07%
Followers	None	8	86.58%	0.08%
Followers	None	10	86.58%	0.08%
Followers	SMOTE	4	74.33%	6.22%
Followers	SMOTE	6	75.80%	7.69%
Followers	SMOTE	8	80.93%	12.82%
Followers	SMOTE	10	82.28%	14.16%
Followers	SMOTE and SpreadSubSample	4	68.49%	18.49%
Followers	SMOTE and SpreadSubSample	6	71.70%	21.70%
Followers	SMOTE and SpreadSubSample	8	75.32%	25.32%
Followers	SMOTE and SpreadSubSample	10	76.54%	26.55%
Friends and Followers	None	4	86.56%	0.06%
Friends and Followers	None	6	86.63%	0.13%
Friends and Followers	None	8	86.65%	0.15%
Friends and Followers	None	10	86.66%	0.16%
Friends and Followers	SMOTE	4	74.47%	6.36%
Friends and Followers	SMOTE	6	75.64%	7.53%
Friends and Followers	SMOTE	8	81.38%	13.26%
Friends and Followers	SMOTE	10	82.17%	14.05%
Friends and Followers	SMOTE and SpreadSubSample	4	69.04%	19.05%
Friends and Followers	SMOTE and SpreadSubSample	6	72.02%	22.02%
Friends and Followers	SMOTE and SpreadSubSample	8	75.55%	25.56%
Friends and Followers	SMOTE and SpreadSubSample	10	76.50%	26.50%

Table A.3: REPTree results

Term:	Explanation:
Status_count	Total tweet count of the user
MentionedEntityCount	Total times a user mentions another user in their tweets
Account_Year	The year the user has joined Twitter
Followers	Total follower count of the user
Friends	Total friend count of the user
(Friends or Followers)MeanStatus_count	Average total tweet count of the user's friends or followers
(Friends or Followers)SumRetweets	Total amount of times a users' friends or followers have been retweeted
(Friends or Followers)SumFriendCount	Total amount of friends of all friends or followers of a user
(Friends or Followers)MeanFollowerCount	Average follower count of a users' friends or followers
(Friends or Followers)SumMentionedEntityCount	Total times all friends or followers of a user mentions other users
(Friends or Followers)MeanLocation	The average location precision as defined in Table 3.10 of a users' friends or followers
(Friends or Followers)MeanCoords_Col	The percentage of their friends or followers with their coordinates set as location
LocationPrio	The average accuracy score of their friends or followers, depending on set

Table A.4: REPTree Terms explanation

Figure A.1: Dataset = Friends, Filter = None

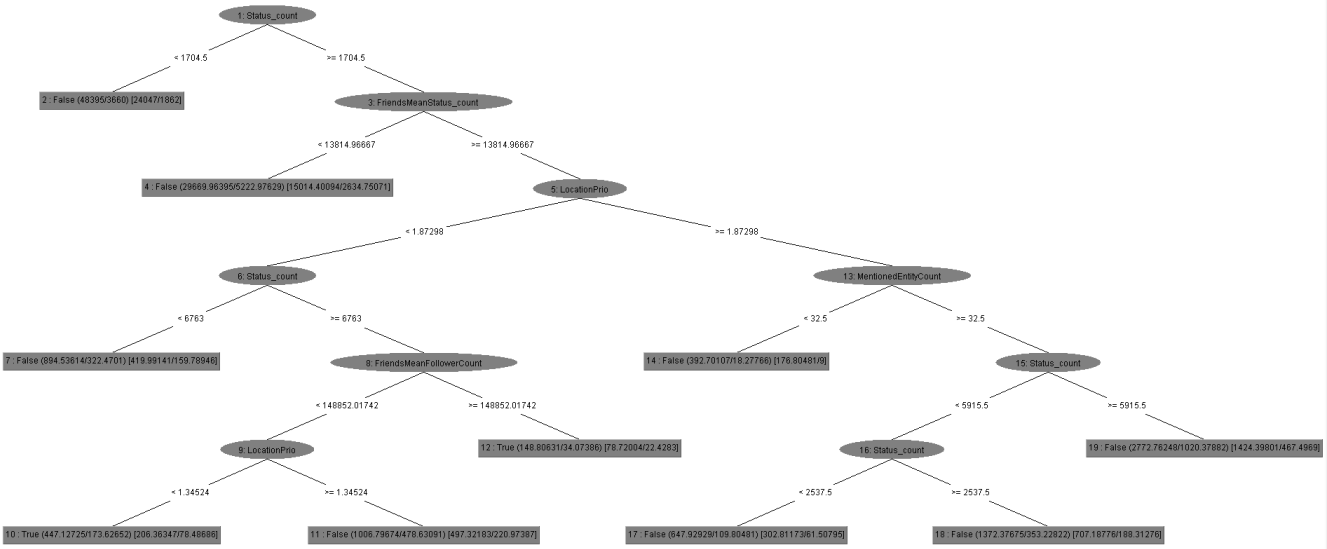

Figure A.2: Dataset = Friends, Filter = SMOTE

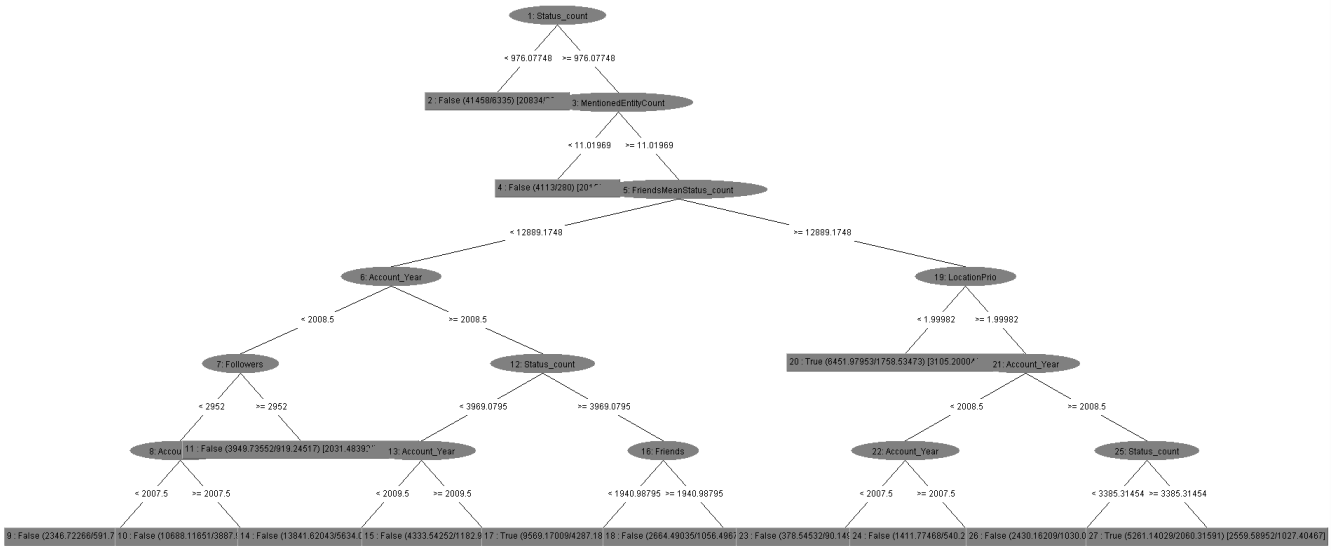

Figure A.3: Dataset = Friends, Filter = SMOTE and SpreadSubsample

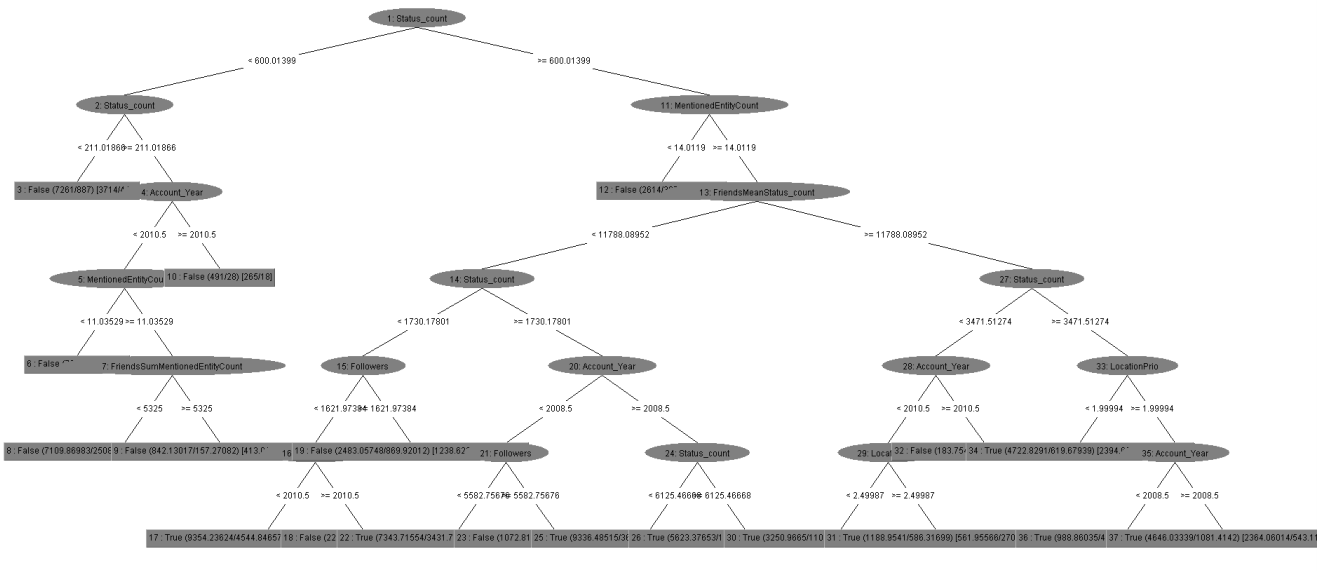

Figure A.4: Dataset = Followers, Filter = None

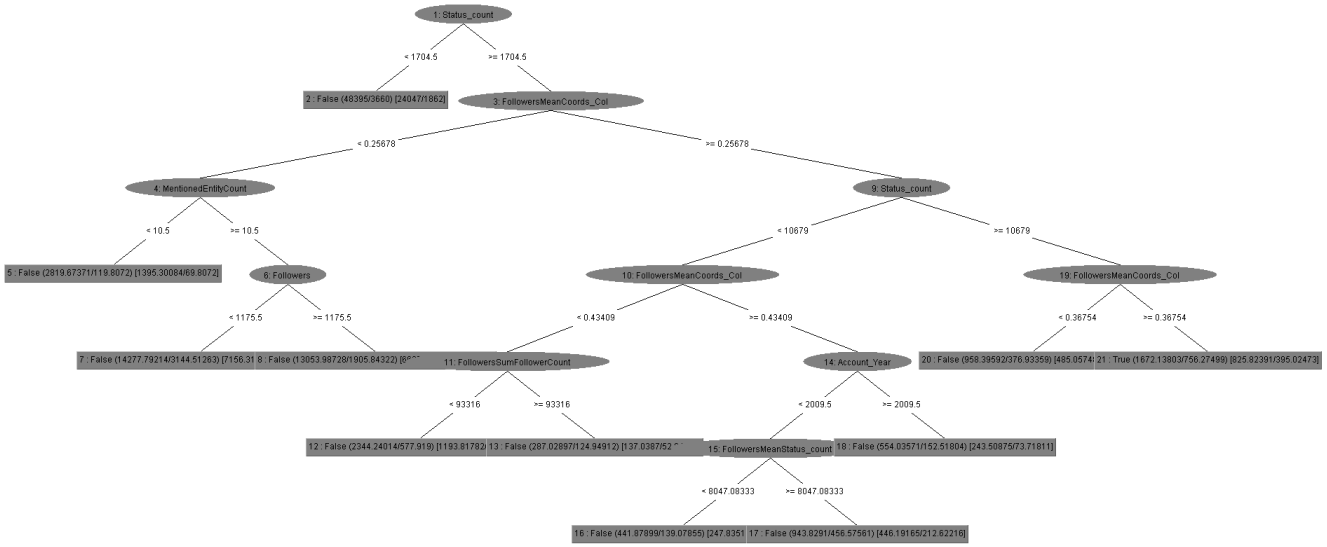

Figure A.5: Dataset = Followers, Filter = SMOTE

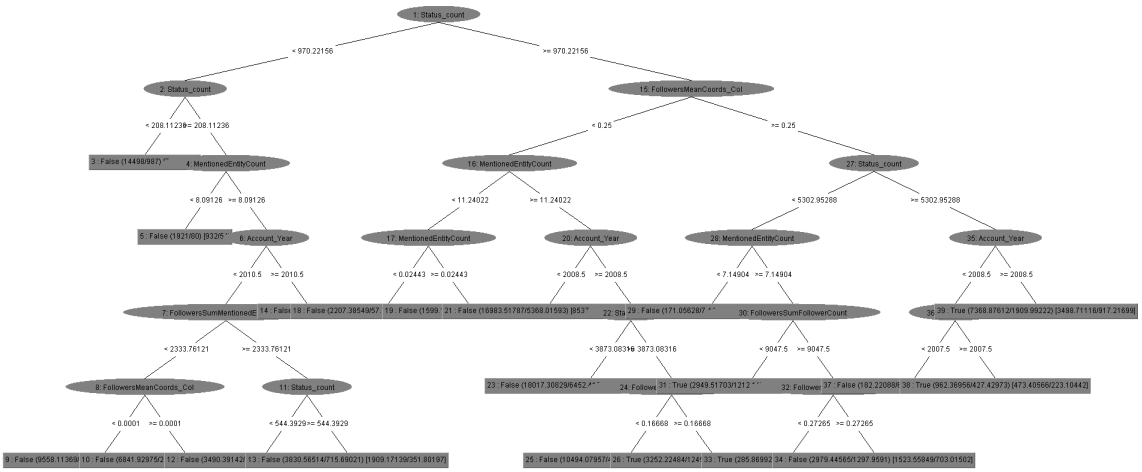

Figure A.6: Dataset = Followers, Filter = SMOTE and SpreadSubsample

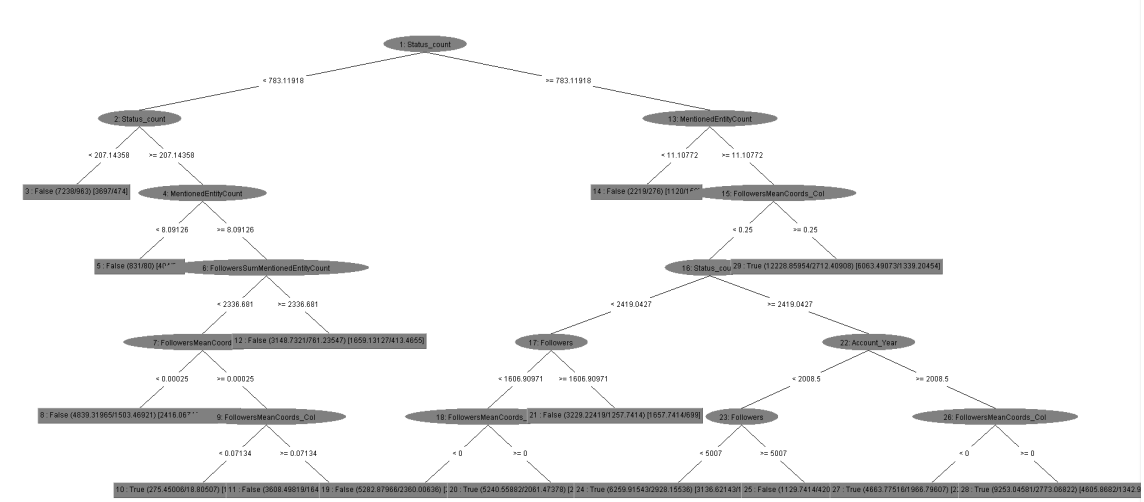

Figure A.7: Dataset = Friends and Followers, Filter = None

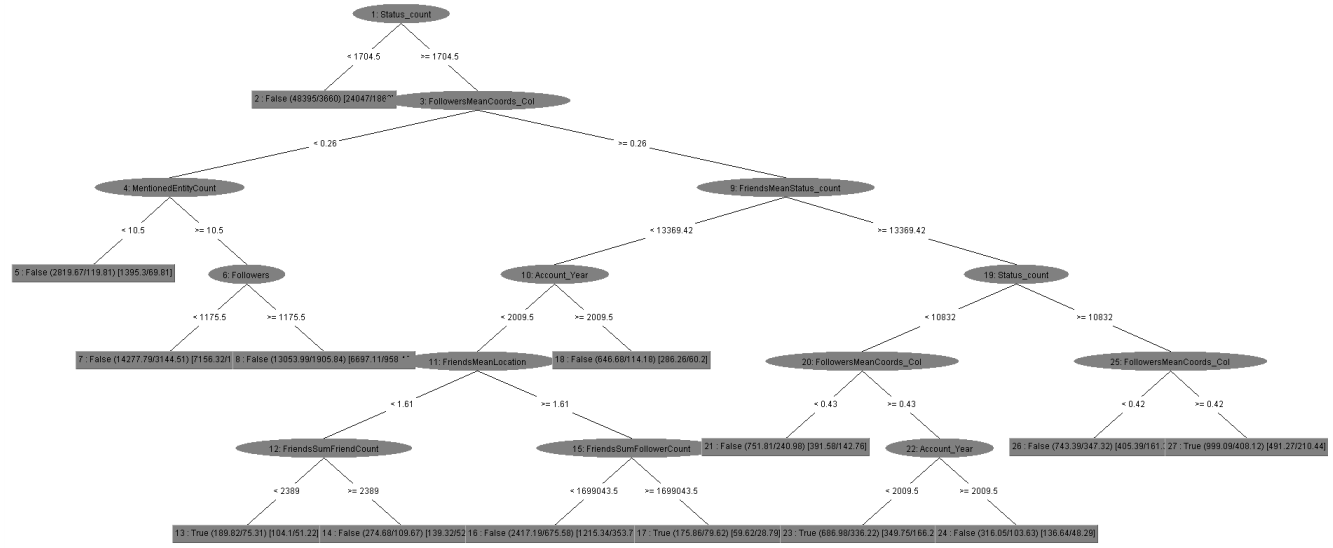

Figure A.8: Dataset = Friends and Followers, Filter = SMOTE

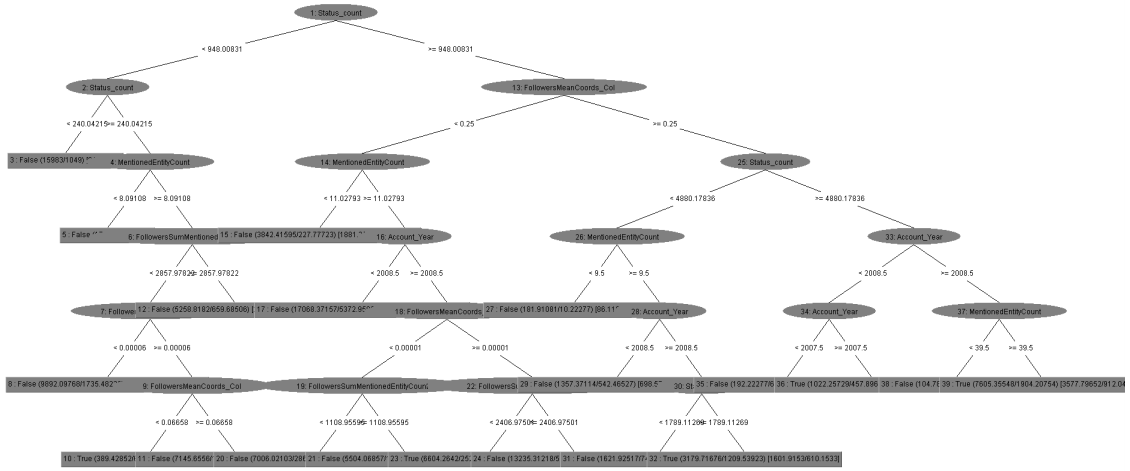

Figure A.9: Dataset = Friends and Followers, Filter = SMOTE and SpreadSubsample