Bachelor Thesis

Artificial Intelligence

Radboud University

Understanding image-set classifiers for future evaluation of adversarial profiles to gain control of our own privacy.

Author:
Stijn Boosman
s1018438
stijn.boosman@student.ru.nl

First supervisor:
Martha Larson
iCIS, Faculty of Science,
Radboud University
m.larson@cs.ru.nl

Second Reader:
Zhuoran Liu
iCIS, Faculty of Science,
Radboud University
z.liu@cs.ru.nl

June 18, 2021

Abstract

An abundance of images can be found on social media platforms nowadays . These images, uploaded by their users, can give us sensitive insights and information about the person behind it using machine learning techniques. In this work, we propose a framework aiding us in investigating the reaction of different set-based image classifiers when controlling two aspects of picture sets, its dimensions and distribution. The extensive framework allows custom creation of user profiles according to rules, pre-trained models on eleven of MS COCO's super-categories and two different implementations of set-based image classifiers. The ultimate goal is to understand the workings of such methods that can conceivably be used by malicious actors wanting to infer privacy-sensitive information from pictures. That way we can deduce useful information from these findings helping future research to craft adversarial techniques to help minimize privacy infringement.

Contents

Chapter 1

Introduction

Every single day, millions of people use social media to upload pictures, stay in contact with friends, share parts of their life, entertain themselves and much more. Relative to many other fields, social media and the internet are still very young and unexplored. This stands in huge contrast to its widespread use and worldwide adaptation over the last 20 to 30 years. As a result, incredible amounts of data are being uploaded daily by the many users who utilize it. The debate about online privacy has never been bigger than it is today. This paper aims to address some of the concerns avid social media users might already or will have in the future. In today's digital world, users need to take care of their own privacy and many want to have a say in what others should and should not be able to know about them. Laws like the European Data Protection Regulation (GDPR) are already a good step in the right direction, where governments realize that the individual user needs to be protected, especially in cyberspace. Yet, it might still be some steps away until proper regulations are in place, especially when one realizes the internet is a rapidly emerging and ever-changing technology, something policy makers have been struggling to catch up with. This is why we would like to contribute to the research and development of techniques that allow individual users to gain back control over their own data.

To elaborate, the rapid introduction of artificial intelligence allowed computers to infer information from huge amounts of data, something that was previously impossible. Research has shown that a substantial amount of insight can be gained from profile pictures and other aspects on a social media profile by applying machine learning techniques [10, 11]. It seems that some companies have caught onto this and have started exploiting it for targeted advertising where they infer and label users based on online activity [1]. With these automatic profilers, sensitive things can be inferred about a user, from seemingly insensitive data in ways they might have never explicitly gave permission. We believe this to be a great problem and a blatant privacy infringement. Think about the discovery of a person's sexual identity, sexual preference, religion, political interest, personality, addiction, mental health and much more just from a few pictures online [1]. Apart from the fact that users likely do not want a third party to know many of these sensitive matters about them, such information can be exploited to prey

on the vulnerabilities of individuals. The most extreme examples conceivably being: exploitation of addicts, blackmail or persecution due to sexual preference or identity and micro-targeting for political cases, furthering the divide we already see today. To prevent such things from happening, something has to be done.

Where this paper differs in contrast to previous research with machine learning and social media, is that we investigated how automatic profilers respond to sets of images rather than single images (Figure 1.1). A lot of research so far focused on how single-image profilers can actively label individuals with sensitive labels. With set-based classifiers, we have better control over the image-level characteristics and results are more controllable and observable. This is in part why we believe this to be a superior method to single image based classification approaches, especially as research on this topic matures over time. By empirically testing and manipulating different inputs, we can gain understanding about the workings, strengths and weaknesses of different set-based classifiers. Gaining this understanding is vital for developing effective privacy protection tools, like adversarial examples and privacy pivots.

We state the following research question: **How do different set-based image classifiers react to varying sets of images?**

To simplify answering this question, we split it up into two sub-questions that were examined one by one:

1. *What dimensions in sets of pictures, if any, are most relevant for different classifiers to base a prediction on?*

2. *How does the distribution of images play a role in set classification?*

To be able to answer these questions, we investigated how two different classifiers (Majority Voting [11] and Weighted Feature Fusion [17]), encapsulated within a framework, react to different sets of images.

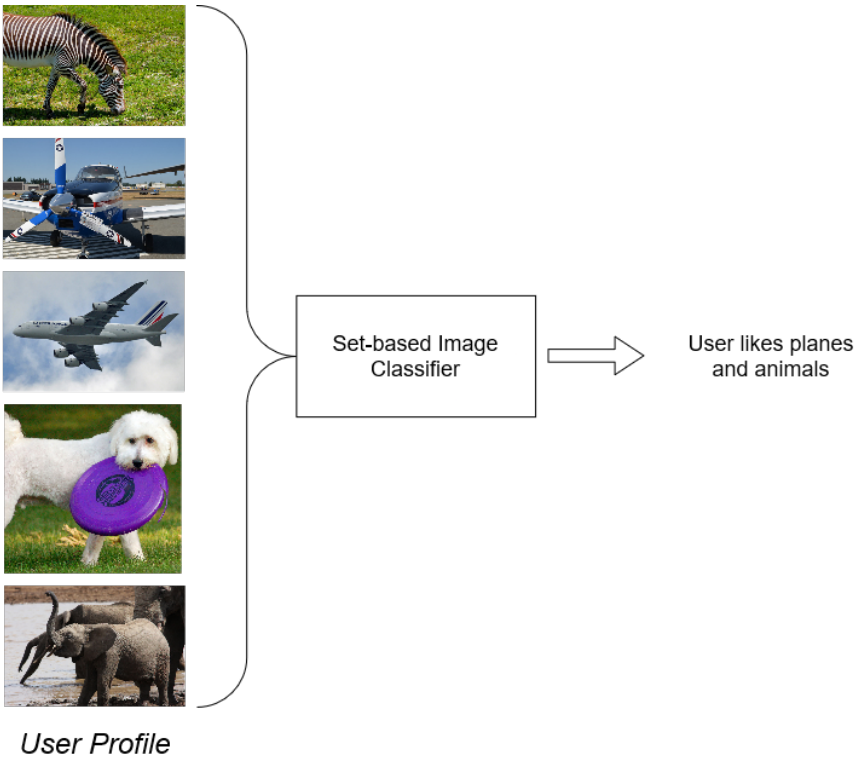

Figure 1.1: A simple illustration of set-based Image Classification

The data set the images are drawn from, is the Microsoft COCO data set [4]. It is the gold standard benchmark for use with object detection models, that is, for models that can recognize multiple objects and their location in a scene. Even though the little research in the field of image-set classification was conducted with image recognition, we, in contrast, chose object detection for various reasons. We believe it to be a more accurate and interesting representation of how social media profiles might look like. Images of our daily lives are often of rather complex scenes, with not just perfectly framed objects in the middle as the main subject of said pictures. Additionally, we thought that with an image recognition approach, a lot of information will be lost. From a malicious agent's perspective that would like to uncover sensitive information about a user, they would want to use all the information possible if it means more accurate inferences.

By using carefully crafted, artificial users sampled from the COCO data set [4], we can try to gain an understanding on how the differing aspects impact the performance of our models and classifiers. Ideally, our artificial users should reflect social media picture sharing behaviour. However, since social media is so new, there is very little research to base ecological validity on, so we try to make educated guesses on how profiles could look like. It is very likely, that the variance between real profiles is much larger than anything we will be able to capture in our artificial users. This means that it is a lot harder for

algorithms to make very confident predictions, as the evidence is weaker. Therefore, we believe that if we make ideal profiles and can successfully inhibit inference of their labels by using adversarial addition, we can expect that this will translate to real world examples where inferring in the wild might be more difficult. However, to be completely certain of this, further research is required.

The main contributions of this research include the creation of an extensive framework in python, which allows highly adjustable creation of user profiles and automating a lot of tedious tasks. It interfaces best on any data set that is annotated with the COCO JSON format, the standard format for the COCO data set itself, and follows a clear three step pipeline outlined in Figure 3.1. Additionally, we provide a user-labelled data set with pictures sampled from the COCO data set using the framework. It is the data set used in training, testing and validating the models, as well as compiling all the results of this work and can be used in future research.

Using the framework and the data, we hope to shed light on the workings of tools used by potential malicious actors, allowing the creation of adversarial examples and other protection mechanisms to mislead classifiers so users gain control back over their own privacy.

In the following chapters, we will explain the concepts needed to understand the content of this paper, some related research in the field, which illustrates the problem space more, the approach we used in answering our research question, as well as results. Finally, we end with a discussion about our findings before concluding this paper.

Chapter 2

Related Work & Preliminaries

The following section outlines previous research and concepts describing the problem space and creating common ground for understanding the content in this work.

An area of research that is fairly novel and of great importance to this paper, is the use of image-set classification. To elaborate, image-set classifiers base their prediction on sets of images rather than single images (see Fig. 1.1)[2]. While traditionally used for analyzing video footage [15, 3], this paper uses it for sets of individual images, assumed to be independent of each other. We assign such a set of images a single label, for example, we can assign personality traits to social media users that have a collection of pictures on their profile, as was done in the PsychoFlickr dataset by Segalin et al. [11]. Using these profile sets, they were able to inference personality traits based on multiple pictures of a user. This ties closely into research investigating how labels can be assigned to individuals that are often of a sensitive nature [1, 10, 14].

These have shown remarkable success and have taught us that machine learning is a powerful tool that, like any powerful tool, can be exploited for good and bad. As a response, other research emerged aiming to obfuscate labelling and thus gain control back of our privacy in cyberspace, for instance TrackMeNot [6] and BlurMe [16]. Especially notable are efforts in the field of computer vision that aim to protect people from being recognized by intrusive systems [7, 8].

In the paper by Liu et al. [5], they propose a privacy pivot for disrupting predictions of set-based classifiers. Such adversarial techniques are of importance in this paper as well, although it is not the main aim of this work to propose new techniques, but rather provide a framework and results using our framework, that beget novel understanding in this field. This will allow future research to make informed decisions when creating adversarial techniques. While classical examples of adversarial examples, like adding noise to an image of a dog, so that the classifier will interpret the image as an ostrich [13], have existed for a while now, this paper focuses on set-based approaches.

In contrast to previous research with set-based image classifiers, this work utilizes object detection data as opposed to image classification data. Image classification data sets are defined as having one label per image, whereas object detection data sets char-

acteristically have multiple labels per image, and often some annotation indicating the position of these objects in a scene by means of a mask or bounding box. We believe that incorporating such detailed information of images, allows much easier control over image aspects when empirically investigating different inputs. These inputs should be manipulated freely and on demand and object detection data sets contain much more information in their labels allowing for such precise crafting of data sets. The object detection data set utilized in this work is the MS COCO data set [4].

In this paper we mainly utilized two different set classifiers: Majority Voting [11] and Weighted Feature Fusion [17], referred to as MV and WFF. Their implementation heavily relies on the way it was used in the paper by Liu et al. [5], and has been modified to fit our needs for this research.

Majority Voting, as the name implies, works on a fairly simple majority basis. If a certain category is most prominent in the examined set of pictures, it will gravitate to classify the set as such; the most represented category wins the vote.

Weighted Feature Fusion on the other hand works quite different. Depending on a computed discriminability value, it gives certain image features more or less importance. A centroid based on the training data is computed for every COCO super-category, which leaves us with a positive and a negative centroid per category. Using the Cosine similarity metric, we can calculate the similarity of new predictions to the positive and negative centroids of a certain category. We can then ultimately calculate the discriminability value based on the similarities to either class, which will determine the final predictions [5, 17].

Chapter 3

Experimental Setup

3.1 Approach

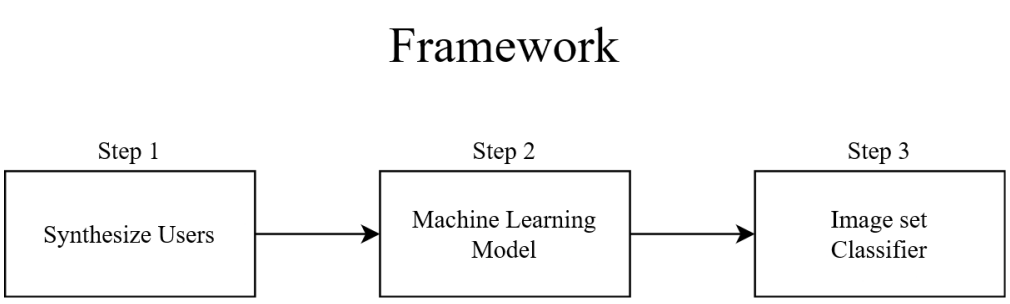

Figure 3.1: Diagram of our framework

The following section explains the general approach of this work.

To be able to answer the main research question: "How do different set-based image classifiers react to varying sets of images?", we split up the problem into smaller ones for easier examination.

1. *What dimensions in sets of pictures, if any, are most relevant for different classifiers to base a prediction on?*

2. *How does the distribution of images play a role in set classification?*

To be able to answer these questions, we define a framework (see Figure 3.1) consisting of three main steps: The first step is to synthesize users along certain criteria we can set ourselves. In the second step, the individual images of these artificial users are passed to an object detection model that can return predictions. The last step of the pipeline is an image set classifier that will predict the label for a user, based on their respective images. The two classifiers used in this paper were mainly featured in the privacy pivot paper [5]. We will call them Majority Voting (MV) and Weighted Feature Fusion (WFF) as coined in the privacy pivot paper. Their respective implementations

for this research can be found on our GitHub[1]. In theory, the framework supports any correct implementation of an image-set classifier. These two algorithms were chosen to examine their differences and for demonstrating the framework.

Our motivation behind creating artificial user profiles, meaning we end up with "fake" data sets, is as follows. First of all, at the time of writing this paper there were no good user-labelled data sets out there that incorporate object detection. This meant, we had to create our own methodology to create such data. Additionally, we believe that real life data sets are much harder to fit, especially for set-classifiers, so along that reasoning, if we create a "perfect" data set along a simple rule with very distinct evidence between classes such that our models have high accuracies, any adversarial method that worsens predictions on such "perfect" data sets will translate just as well to data sets that are noisier and have larger variability, i.e. much harder to fit. If methods exist that disrupt already very robust classifiers, then less robust classifiers will likely be thrown off even more, increasing the adversarial effect. Having this baseline of robust models trained on artificial data, which our research hopes to provide, opens up the possibility of testing along a large amount of highly customizable aspects, which will ultimately help research create effective adversarial methods that will later on translate to real-life data sets.

3.2 Specific implementation

3.2.1 Synthesizing users

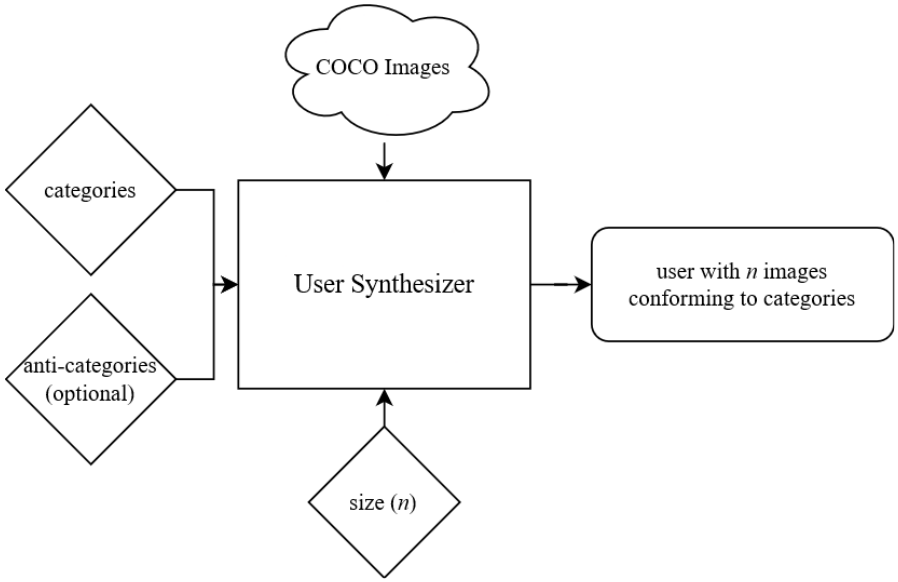

Figure 3.2: Overview of how users are created

[1]https://github.com/HcKide/ImageSetClassificationFramework

Since the COCO dataset is just a collection of individual images and has no notion of user-level annotation, we were forced to create users from the original data. A high-level abstraction of this can be see in Figure 3.2. One large limitation of this research is, without a doubt, that all users and user labels used are entirely artificial. On the other hand, it leaves room to experiment with different aspects, as we are not bound by the restrictions of a real dataset. This makes it perfect for the empirical investigation on the reaction of image-set classifiers.

The way we labelled our users is by utilizing the already pre-made super-categories COCO provides. There are 12 different super-categories in total that hierarchically cover all 91 normal categories. For example, the super-category animal would include dogs, cats, elephants and so on; the super-category food includes pizza, hot dogs and others; et cetera. It has to be noted that the super-category 'person' was completely ignored. This category is present in over half of all images in the training set, which overshadows every other super-category significantly. Therefore, the class was omitted entirely and ignored as a label and no model was trained for it, nor is it treated as a valid user label. We are left with 11 different super-categories.

When it comes to varying our input data, we can narrow it down to two main aspects that we would like to manipulate and control:

- Dimensions within the set of pictures.
- Distribution of the images.

Manipulating the dimensions in sets of pictures can be rather complex. Distribution, on the other hand, is more straightforward, which is why we investigated that first.

One of the approaches we took was to create a kind of 'data recipe' that would allow us to have a consistent baseline and allow reproducible results in the future.

To illustrate, say we would like to investigate the user label 'electronic'. By varying the amount of 'electronic' labelled images for a single user, we can investigate what kind of effect it will have on our outcome. Therefore, multiple different users were created, all with slightly varying distributions of target pictures. For example, one user where out of the 20 pictures, two are considered 'electronic', or in other words 10 percent. Subsequently, we have another user with 25 percent, 50 percent and so on. See Figure 3.3.

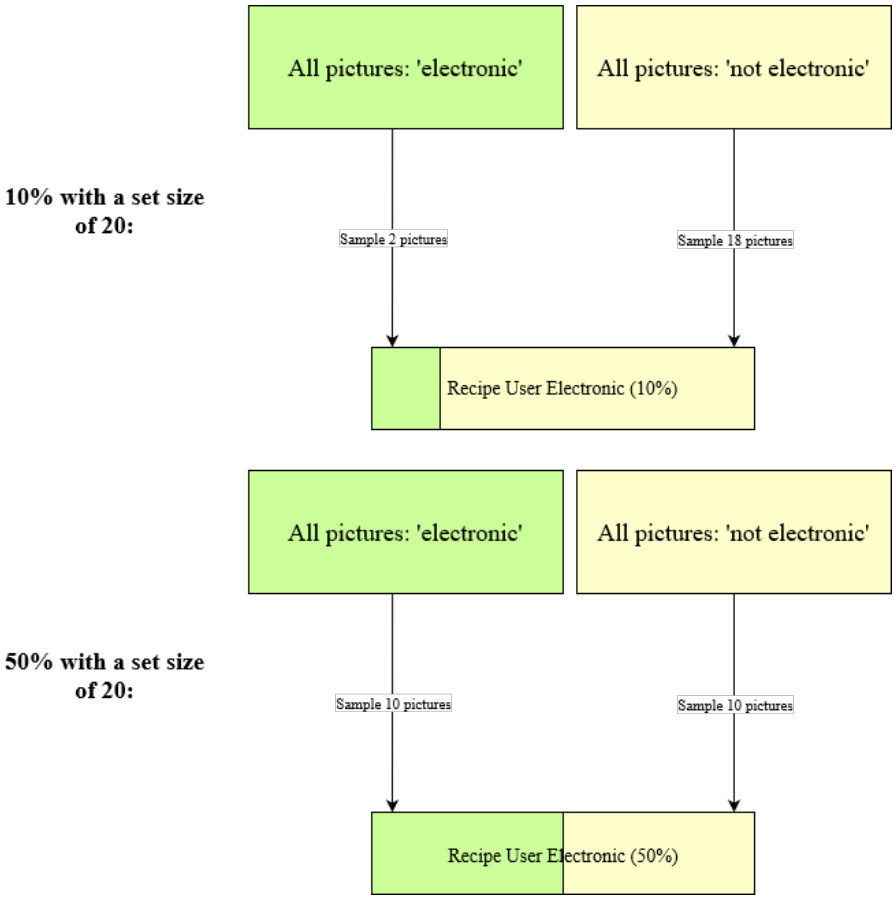

Figure 3.3: Illustration of the way the data recipe is used.

The way this works is that the code randomly samples a set amount of pictures per user from the original picture pool. It only takes pictures that fall into the criteria of whichever user we want to create. That is, if we would like to create a user with the label 'electronic', it would randomly sample all pictures with the super-category 'electronic'. This structure resembles the one in the privacy pivot paper where, if a user has the label high Openness, its individual pictures will receive the same label, despite their own content being different [5].

To get our final results, we define a few aspects for the users. We decided to investigate the 'sports' super-category and the 'furniture' super-category because of their VGG single image accuracies (see Figure 4.1). With 'sports' having the highest overall accuracy and 'furniture' the lowest out of all 11 super-categories. For the first test set we define that a user with 10 percent or more images which are predicted as 'sports', is considered to have the user label 'sports'. For the second test set we defined 60 percent as the threshold and lastly a 90 percent threshold. All users are uniformly distributed, that is, all users have exactly 10 percent of their pictures labelled as 'sports' or 60 percent and 90 percent respectively. For a more realistic distribution we could investigate drawing

the amount of sports content per user from a normal distribution which is something we might want to investigate in further research.

Moving on to the second aspect we would like to control and observe, we find that dimensions in pictures are harder to define. Additionally, even though we are synthesizing our own data, we are still more restricted than with the distribution of pictures. Namely, we are bound by the content of the individual pictures themselves as we do not have the liberty to go out of our way and stitch together completely new pictures that we would like to include. As such, we took the liberty to only look at some surface dimensions, primarily the objects in it. This was arguably a lot easier to control than more fine-grained details found in pictures as we can just use the original COCO image labelling. We decided to continue to use our data recipe as well, so we could directly compare results.

Since the sub-category 'snowboard' is the least represented category when looking at all the 'sports' sub-categories (see Figure 3.4), it might be interesting to see how predictions change. We compared sets of 'sports' users, with content that is randomly made up of all 'sports' sub-categories, with 'sports' users that only have 'snowboard' pictures.

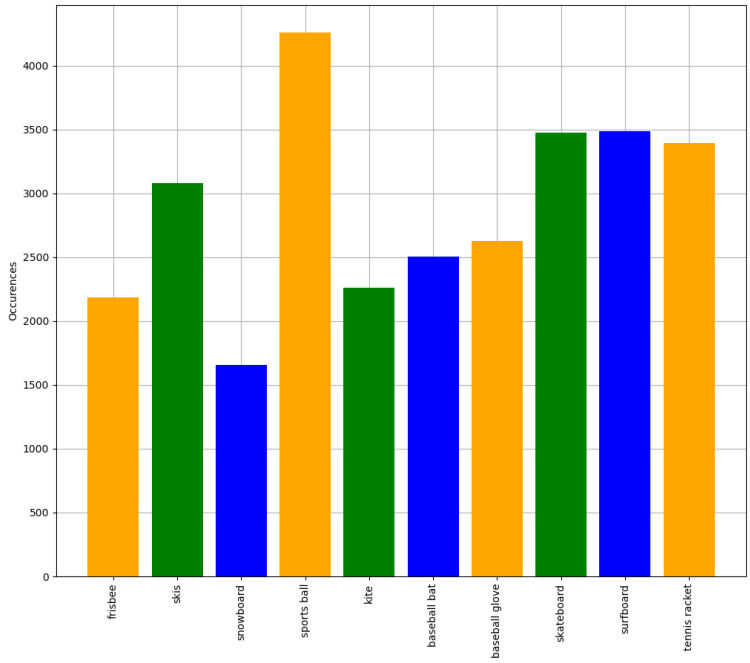

Figure 3.4: Occurences of the sub-categories of sports

We also did the same with the 'furniture' super-category (Figure 3.5), where we

picked 'toilet' as the target category.

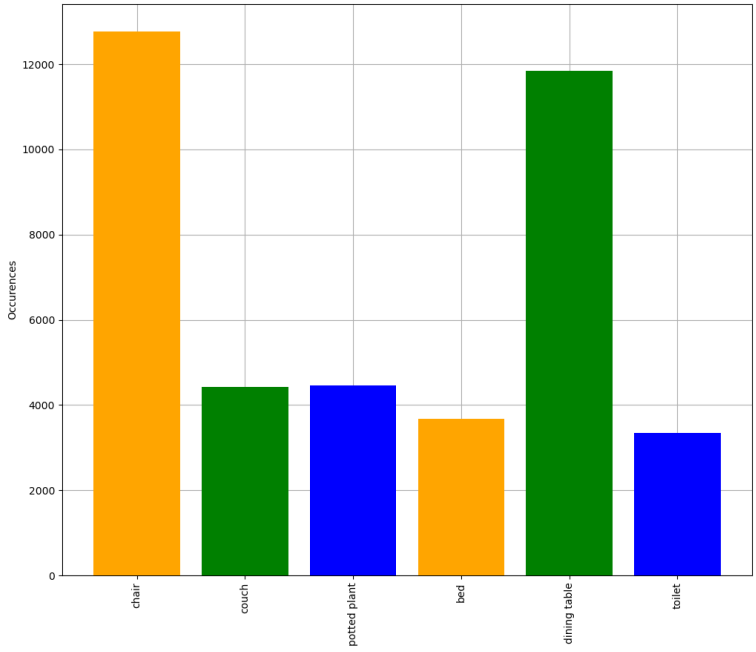

Figure 3.5: Occurences of the sub-categories of furniture

3.2.2 Machine Learning Model

The second step of the framework relies on a machine learning model trained on our data (see Figure 3.1). Initially, we wanted to use an object detection model and thus worked with an implementation of the "Faster R-CNN" model [9]. It is a powerful model trained on the COCO dataset available in the torchvision library. Unfortunately, this implementation did not work for this research, so we changed our approach. This was mostly due to time restrictions. Instead of having one model predicting all super-categories, we simplified it to a pure binary aspect. That is, for every super-category we would like to predict, we train one model to output a positive or negative match. For this we decided to use a VGG 16-layer model [12]. Each model was pre-trained on ImageNet and fine-tuned by us to output two classes, i.e. binary. Since 'person' was omitted, we are left with 11 different models, each specialized to output a binary prediction for its respective super-category.

The models were trained for 24 epochs on a university computer cluster with an Nvidia V100 GPU. We recommend to consult the GitHub page for more implementation details.

Every network has different training data, that way it uses balanced data by having

an equal representation of both the negative and positive class. This is easily achieved because we control each and every aspect of user creation. As an illustration, if we train a network for the label 'animal', we want half of the training users to be labelled 'animal' and the other half 'not-animal', i.e. the remaining ten classes. Therefore, different training data per category was created. However, the images are all sampled from the same training pool, originating from the original COCO training set (see Figure 3.6). This allows clear separation of training and testing data when sampling.

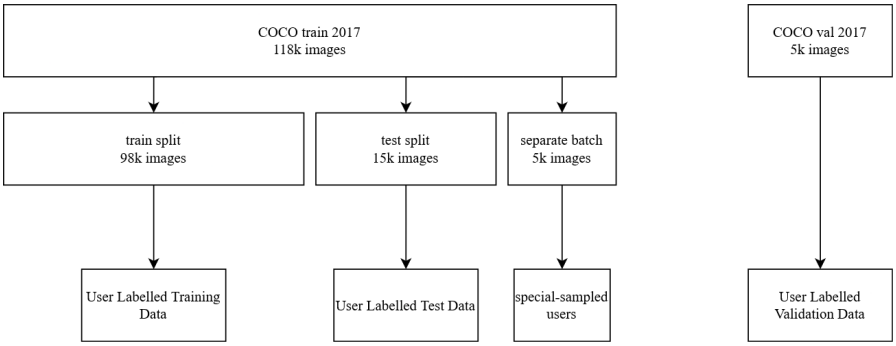

Figure 3.6: Diagram showing a high-level representation of how we went about creating the data needed to train, validate and test the machine learning model.

The training data consists of 8 users per non-target category with 80 images per user. One of the 11 classes is our target, therefore we are left with 10 non-target categories. This results in $6,400 (= 8 \cdot 80 \cdot 10)$ images labelled as negative and the same number as positive. In total, every network is trained with $12,800$ images with balanced classes, 80 positive users and 80 negative users.

3.2.3 Set Classifiers

As explained in Chapter 2, this paper uses two different set-based classifiers called Majority Voting (MV)[11] and Weighted Feature Fusion (WFF)[17]. We can give them inputs and compare their outputs to see their differences. In theory, the framework is set up to support any correct implementation of a set-based classifier which makes it a useful tool in future research if there is a need for a different set-based approach.

On top of standard Majority Voting, to allow for setting an arbitrary threshold, we coin the term 'Threshold Voting', which acts like an extension of Majority Voting. Similar to Majority Voting, if the amount of content satisfies the threshold we set, it will predict true. To illustrate, if we decide that a user X has a label Y with only 30 percent of it's pictures conforming to label Y, then we can set the threshold for Threshold Voting to 0.3, or for any other arbitrary amount between 0 and 1 respectively. Standard Majority Voting can be seen as using a threshold of 0.5.

Chapter 4

Experimental Results

4.1 VGG Network Accuracies

Network Category	**Accuracy**
Outdoor	82.4%
Food	87.9%
Indoor	75.1%
Appliance	90.8%
Sports	93.1%
Animal	90.9%
Vehicle	86.4%
Furniture	74.5%
Accessory	76.6%
Electronic	85.1%
Kitchen	79.9%

Table 4.1: The individual VGG 16-layer network accuracies for its corresponding test data.

The above table shows the accuracies we get when testing our individual machine learning models. That is, it does not look at set-based classifying but only at single-image level and we record how often it predicts the correct label per image. The entire test set consisted of 800 images, half positive, half negative. We observe very good accuracies, especially for some classes like 'sports' and 'animal'.

4.2 Sports & Furniture Results

In the following sections we report and discuss some of our findings in using the data recipe and looking at the differences in regards to distribution within a set. Since both Weighted Feature Fusion and Threshold Voting (the extension of Majority Voting) require a threshold, we calculated the optimal threshold for the results on the validation set. It is optimized to pick the highest threshold that yields the highest accuracy. The validation set was only comprised of positive users. We refer to Threshold Voting as TV in our tables.

We hypothesize, that for the 'sports' super-category, the optimal thresholds for TV, should be around the actual percentage count of target content. To elaborate, we define a user to be labelled positively at 60% or more 'sports' labelled images, therefore expect the optimal threshold to be 0.6 or slightly lower. For 'furniture' on the other hand, given that the single-image classification accuracy is significantly lower (see Figure 4.1), we expect the the thresholds to be lower than its actual percentage as it will not recognize all 'furniture' labelled pictures within a set. Additionally, we expected the two methods TV and WFF to achieve roughly the same accuracies when optimizing thresholds. The way WFF thresholds work is different than in Threshold Voting so it is not surprising to see different values.

4.2.1 Sports Users

The depicted table shows the accuracy at each percentage level, as well as the respective method and threshold used to get the results. Every user possesses 20 images where for the negative users, none of the images have 'sports' content in it; yet for the positive users, all of them have exactly 10, 60 or 90% (depending which percentage we want to investigate) of images that are labelled 'sports'. So we are left with 3 different test sets, one for each percentage. We record the accuracy in respect to the target category, for each image-set classification method. As mentioned, the thresholds depicted are the ones optimized for accuracy on a positive validation set. By comparing the differences in percentages with each other, as well as the differences between super-categories, we can gain insight about what role distribution and dimension play in classification.

Percentage	Accuracy	Method	Threshold
10%	82%	TV	0.1
10%	87%	WFF	0.7
60%	100%	TV	0.5
60%	91%	WFF	0.9
90%	100%	TV	0.7
90%	86%	WFF	0.9

Table 4.2: Table with accuracies, methods and corresponding thresholds. Thresholds were picked for maximum accuracy.

As somewhat expected, the thresholds for the instances of TV, are around the same as the percentage (0.1 for 10%, 0.5 for 60% etc.). The instances where it is slightly lower is likely due to the fact that the model incorrectly classifies a few target images as non-target. This creates the illusion that there is less evidence needed for a set of pictures to be labelled as the target category, whereas in reality it is higher.

We also note that WFF accuracies are in general lower than it's respective TV counterpart, except for 10%. This is likely due to the fact, that WFF usually excels at dealing with sets of images containing low confidence pictures, whereas TV treats them all the same. It seems that the artificial profiles made for 60 and 90%, have less low confidence pictures as we would otherwise see WFF performing better than TV.

4.2.2 Furniture Users

In this section, we apply the same method as previously but our target is defined as 'furniture'. As we can see in Table 4.1, 'furniture' scores the lowest when it comes to individual image accuracy. This made it an interesting target to examine in contrast to the highest individual classifier, 'sports'.

Percentage	Accuracy	Method	Threshold
10%	52%	TV	0.1
10%	54%	WFF	0.2
60%	97%	TV	0.3
60%	87%	WFF	0.4
90%	100%	TV	0.5
90%	96%	WFF	0.6

Table 4.3: Table with accuracies, methods and corresponding thresholds. Thresholds were picked for maximum accuracy.

We see that in contrast to Table 4.2, the thresholds for TV are significantly lower for 60% and 90%, likely due to even heavier classification errors. The 10% accuracies here are significantly lower than in Table 4.2. Similar as in the previous table, the WFF accuracy for 10% is slightly higher than its TV counterpart, giving some weak evidence that WFF performs better at lower percentage levels of evidence within a user profile. However, when looking at higher percentages levels like 60% and 90%, TV clearly outperforms WFF once again.

4.3 Snowboard & Toilet Results

In the following sections we report and discuss some of our findings in using the data recipe and looking at the differences mainly in regards to dimensions within a set. We only look at the surface level dimension of objects in a picture, which is easiest to control

and we are not limited by the content of pictures itself. We hypothesize that looking at the least represented object of a super-category, will worsen predictions to some degree. This is because at training time, the VGG networks has seen this instance the least and might have trouble correctly classifying it. We predict 'toilet' will have a larger impact than just using 'snowboard', as the predicting power of the 'furniture' super-category to which 'toilet' belongs to is already weaker than that of 'sports'.

4.3.1 Snowboard Object Results

We apply the 'data recipe' and set our target to 'sports', but at sampling time we specifically indicate that we only want the object 'snowboard' in every picture that is considered 'sports'. That is, we now explicitly control the object dimension in sets. To illustrate, if we investigate a profile with the user label 'sports', a certain percentage of pictures within the set are 'sports' images. However, in this instance all those 'sports' images only have snowboards on them and no other sports related items. Since 'snowboard' is the least represented sub-category of the super-category 'sports' (see Figure 3.4), we would like to see whether it makes a difference on performance as the model likely hasn't seen many 'snowboard' instances during training.

Percentage	**Accuracy**	**Method**	**Threshold**
10%	74%	TV	0.1
10%	84%	WFF	0.8
60%	98%	TV	0.5
60%	87%	WFF	0.9
90%	100%	TV	0.7
90%	91%	WFF	0.9

Table 4.4: Table with accuracies, methods and corresponding thresholds. Thresholds were picked for maximum accuracy.

Comparing the results with Table 4.2, we see a slight decrease in accuracy overall, except for 90% WFF, while the thresholds remain largely the same. This indicates that, as predicted, the classifier makes more mistakes and therefore performs worse.

4.3.2 Toilet Object Results

Here we apply the tactic for 'toilet', as it is picked because it is the least representative sub-category out of all 'furniture' sub-categories (see Figure 3.5).

Percentage	**Accuracy**	**Method**	**Threshold**
10%	56%	TV	0.1
10%	59%	WFF	0.2
60%	100%	TV	0.4
60%	93%	WFF	0.5
90%	100%	TV	0.6
90%	95%	WFF	0.6

Table 4.5: Table with accuracies, methods and corresponding thresholds. Thresholds were picked for maximum accuracy.

Comparing this table to Table 4.3, accuracies are still fairly consistent but only because the constraints are lowered because it uses less strict thresholds. Meaning it still underperforms but perhaps not as much of as expected. It is possible that the absolute effect is smaller than when comparing 'sports' with 'snowboard', as 'snowboard' is one of 10 sub-categories of 'sports', whereas 'toilet' is one of six (see Figure 3.4 & 3.5). Meaning relatively, it has seen less 'snowboard' pictures than it has seen 'toilet' pictures at training time.

Chapter 5

Discussion

5.1 Distribution and Dimension Analysis

Revisiting both of our sub-questions:

1. *What dimensions in sets of pictures, if any, are most relevant for different classifiers to base a prediction on?*

2. *How does the distribution of images play a role in set classification?*

We can start to draw some conclusions from the results of this research. When it comes to distribution, we can conclude that Weighted Feature Fusion performs better than Threshold Voting in low evidence cases, as can be seen in Tables 4.2 to 4.5, where WFF consistently scores a few percent higher at 10% when optimizing for accuracy. Whereas in high evidence cases such as 60% and 90% Threshold Voting clearly outperforms.

When looking at dimensional differences, it is abundantly clear that using a super-category with lower classification accuracy performs worse than one with high classification accuracy (see Tables 4.2 & 4.3). This holds for both accuracy at lower percentages and approximated threshold especially for Threshold Voting. When examining the same target category, but isolating its least represented sub-category such as 'snowboard' for 'sports' and 'toilet' for 'furniture', accuracy and threshold approximation does indeed decline, making less represented sub-categories a viable solution for adversarial effects. We saw the same effect of WFF being better at lower percentages, and remarkably in Table 4.4 at 90% be better than Table 4.2.

5.2 Limitations

The following section talks about the limitations of this research. We initially wanted to create profiles of images along certain parameters that can be found on social media. However, we ran into issues since there is very little research about behaviour of posting on social media. We believe this to be because social media is still very young and

new, research has not had much time yet to emerge, additionally, predicting accurate human behaviour is likely not easy. Therefore, the creation of artificial users is entirely arbitrary along a certain set of rules. This means that ecological validity of the resulting user dataset cannot be guaranteed. However, our motivation behind this was explained in Section 3.1

Furthermore, we believe our initial approach with an actual object detection model, is still superior if done properly. Since there is a direct link between prediction and the raw object detection data, it should in theory have better inferencing capabilities about the specific objects in the scene. Yet to train a model that converges well, it will likely need very clear and distinct training data. While VGG networks work great for this research, they look at the entire image, including all the potential useless background information, which is less efficient than only using specific objects in scenes, which is what we get with object detection models. Therefore, it leaves room for future research to be conducted in this direction.

Additionally, for the changed aspects, we only look at the super-categories 'sports' and 'furniture' due to their single-image accuracies (see Figure 4.1). For a more definitive conclusion, applying the same methodology to additional super-categories will likely support the conclusions of this paper further. Additionally, the computed thresholds that were optimized for accuracy, were computed on validation data comprising of only positive users. Due to time constraints, it was not possible to compute them over again for both positive and negative users. However, the results still show good accuracies, showing that the thresholds still work reasonably well.

Lastly, this work only looked at the surface dimension of objects, whereas looking at different dimensions might give more insight about the workings of set-based classifiers.

5.3 Outlook

In this section we describe the outlook for future research, where we can improve upon this work and be aware of the afore mentioned limitations. As this field of research is fairly novel, there are a lot of directions future research can go in. To be able to craft good adversarial examples, it can be very beneficial to dedicate even more time and resources into understanding different types of image-set classifiers. The framework allows different implementations of such and it would certainly benefit the maturity of this research topic if more image-set classifiers are integrated and examined. In addition to that, creating users and artificial data can be even further optimized, for example by drawing recipe users from a normal distribution rather than a uniform one, as well as trying to adhere to more social media content, especially if other research emerges that somewhat successfully describes picture sharing behaviour on social media.

Lastly, something that we noticed in this research but did no extensive analysis for, is the fact that there is a strong correlation of appearance between certain categories. Future work could investigate the presence of two rare and opposing categories appearing in pictures together, which can throw off the machine learning models leading to worse predictions.

Chapter 6

Conclusions

In conclusion, we have found evidence that using less represented categories in low amounts, has an adversarial effect on classification. By examining distributional differences we can observe how Weighted Feature Fusion and Threshold Voting measure up against each other under differing circumstances, showing that WFF is certainly more robust in low-evidence cases, but TV better with high evidence cases. Furthermore, manipulating object dimensions in sets of pictures to be of low represented sub-categories, shows having an adversarial effect on image classification. It seems both WFF and TV suffer from it, by either lowering constraints in terms of thresholds or flat out having worse accuracy. This is especially prevalent when using a less accurate underlying machine learning model, such as could be seen for the super-category 'furniture'. Additionally, we have an extensive framework allowing highly customizable data set creation, evaluation and data manipulation along with an artificial data set originally sampled from MS COCO. It provides a glimpse into the strengths and weaknesses of two image-set classification techniques, Majority Voting and Weighted Feature Fusion which could hypothetically be used by malicious actors.

Finally, while this research does not solve the immediate problem of effective privacy protection in cyberspace, it is intended to lay some small stepping stones for future developments. With work conducted in this paper, we hope to gain more understanding which is essential for developing effective adversarial techniques. Using these techniques, we can help users stay safe online by maintaining control of their own privacy and not let the attackers gain the upper hand.

Bibliography

[1] José González Cabañas, Ángel Cuevas, Aritz Arrate, and Rubén Cuevas, *Does facebook use sensitive data for advertising purposes?*, Commun. ACM 64, 1 (January 2021), 2020, pp. 62–69.

[2] Munawar Hayat, Mohammed Bennamoun, and Senjian An, *Deep reconstruction models for image set classification*, vol. 37, IEEE, 2014, pp. 713–727.

[3] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla, *Discriminative learning and recognition of image set classes using canonical correlations*, vol. 29, 2007, pp. 1005–1018.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, *Microsoft coco: Common objects in context*, European Conference on Computer Vision (Cham) (David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, eds.), Springer International Publishing, 2014, pp. 740–755.

[5] Zhuoran Liu, Zhengyu Zhao, and Martha Larson, *Pivoting image-based profiles toward privacy: Inhibiting malicious profiling with adversarial additions*, 29th Conference on User Modelling, Adaptation and Personalization, to appear, 2021.

[6] Helen F. Nissenbaum and Howe Daniel, *Trackmenot: Resisting surveillance in web search*, Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society, Eds. I. Kerr, C. Lucock, and V. Steeves, Oxford: Oxford University Press, 2009.

[7] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele, *Faceless person recognition: Privacy implications in social media*, European Conference on Computer Vision, Springer, 2016, pp. 19–35.

[8] Seong Joon Oh, Mario Fritz, and Bernt Schiele, *Adversarial image perturbation for privacy protection–a game theory perspective*, Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1482–1491.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, Advances in Neural Infor-

mation Processing Systems (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[10] Cristina Segalin, Fabio Celli, Luca Polonio, Michal Kosinski, David Stillwell, Nicu Sebe, Marco Cristani, and Bruno Lepri, *What your facebook profile picture reveals about your personality*, Association for Computing Machinery, 2017, pp. 460–468.

[11] Cristina Segalin, Dong Seon Cheng, and Marco Cristani, *Social profiling through image understanding: Personality inference using convolutional neural networks*, Computer Vision and Image Understanding, Volume 156, 2017, pp. 34–50.

[12] Karen Simonyan and Andrew Zisserman, *Very deep convolutional networks for large-scale image recognition*, 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings (Yoshua Bengio and Yann LeCun, eds.), 2015.

[13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, *Intriguing properties of neural networks*, 2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings (Yoshua Bengio and Yann LeCun, eds.), 2014.

[14] Alessandro Vinciarelli and Gelareh Mohammadi, *A survey of personality computing*, vol. 5, 2014, pp. 273–291.

[15] Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai, *Covariance discriminative learning: A natural and efficient approach to image set classification*, 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2496–2503.

[16] Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, and Nina Taft, *Blurme: Inferring and obfuscating user gender based on ratings*, Proceedings of the sixth ACM conference on Recommender systems, 2012, pp. 195–202.

[17] Manyuan Zhang, Guanglu Song, Hang Zhou, and Yu Liu, *Discriminability distillation in group representation learning*, European Conference on Computer Vision, 2020.

Chapter 7

Appendix

Link to GitHub: https://github.com/HcKide/ImageSetClassificationFramework

7.1 Confusion Matrices

Confusion matrices of the categories as seen in Chapter 4. One matrix per percentage and method used, Threshold Voting (TV) and Weighted Feature Fusion (WFF).

7.1.1 Sports Users

		Prediction		
		Positive	Negative	
Actual Class	Positive	46	4	50
	Negative	14	36	50
		60	40	

Table 7.1: Target: 'sports' for 10%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	41	9	50
	Negative	4	46	50
		45	55	

Table 7.2: Target: 'sports' for 10%, WFF

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	0	50	50
		50	50	

Table 7.3: Target: 'sports' for 60%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	49	1	50
	Negative	8	42	50
		57	43	

Table 7.4: Target: 'sports' for 60%, WFF

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	0	50	50
		50	50	

Table 7.5: Target: 'sports' for 90%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	48	2	50
	Negative	12	38	50
		60	40	

Table 7.6: Target: 'sports' for 90%, WFF

7.1.2 Furniture Users

		Prediction		
		Positive	Negative	
Actual Class	Positive	48	2	50
	Negative	46	2	50
		96	4	

Table 7.7: Target: 'furniture' for 10%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	22	28	50
	Negative	18	32	50
		40	60	

Table 7.8: Target: 'furniture' for 10%, WFF

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	3	47	50
		53	47	

Table 7.9: Target: 'furniture' for 60%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	47	3	50
	Negative	10	40	50
		57	43	

Table 7.10: Target: 'furniture' for 60%, WFF

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	0	50	50
		50	50	

Table 7.11: Target: 'furniture' for 90%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	4	46	50
		54	46	

Table 7.12: Target: 'furniture' for 90%, WFF

7.1.3 Snowboard Users

		Prediction		
		Positive	Negative	
Actual Class	Positive	46	4	50
	Negative	22	28	50
		68	32	

Table 7.13: Target: 'snowboard' for 10%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	44	6	50
	Negative	10	40	50
		54	46	

Table 7.14: Target: 'snowboard' for 10%, WFF

		Prediction		
		Positive	Negative	
Actual Class	Positive	48	2	50
	Negative	0	50	50
		48	52	

Table 7.15: Target: 'snowboard' for 60%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	13	37	50
		63	37	

Table 7.16: Target: 'snowboard' for 60%, WFF

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	0	50	50
		50	50	

Table 7.17: Target: 'snowboard' for 90%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	9	41	50
		59	41	

Table 7.18: Target: 'snowboard' for 90%, WFF

7.1.4 Toilet Users

		Prediction		
		Positive	Negative	
Actual Class	Positive	49	1	50
	Negative	43	7	50
		92	8	

Table 7.19: Target: 'toilet' for 10%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	21	29	50
	Negative	12	38	50
		33	67	

Table 7.20: Target: 'toilet' for 10%, WFF

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	0	50	50
		50	50	

Table 7.21: Target: 'toilet' for 60%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	49	1	50
	Negative	6	44	50
		55	45	

Table 7.22: Target: 'toilet' for 60%, WFF

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	0	50	50
		50	50	

Table 7.23: Target: 'toilet' for 90%, TV

		Prediction		
		Positive	Negative	
Actual Class	Positive	50	0	50
	Negative	5	45	50
		55	45	

Table 7.24: Target: 'toilet' for 90%, WFF