UPPSALA
UNIVERSITET

Självständigt arbete i informationsteknologi
June 15, 2019

Surveillance Using Facial Recognition and Social Media Data

Pontus Eriksson
Carl Nordström
Alexander Troshin

Civilingenjörsprogrammet i informationsteknologi

Master Programme in Computer and Information Engineering

UPPSALA
UNIVERSITET

Institutionen för informationsteknologi

Besöksadress:
ITC, Polacksbacken
Lägerhyddsvägen 2

Postadress:
Box 337
751 05 Uppsala

Hemsida:
http://www.it.uu.se

Abstract

Surveillance Using Facial Recognition and Social Media Data

Pontus Eriksson
Carl Nordström
Alexander Troshin

People share more and more on social media, aware that they are being surveilled but unaware of the scope and the ways that their data is processed. Large amounts of resources are dedicated to performing the surveillance, both in terms of labor and computation. This project explores the scope of data collection and processing by showing that it is possible to gather, process, and store data from the social media platforms Twitter and Reddit in real-time using only a personal computer. The focus was to use facial recognition to find specific individuals in the stream of data, but the data collected can be used for other purposes.
We have also explored the ethical concerns regarding both the collection and processing of such data.

Extern handledare: Alexander Okonski
Handledare: Mats Daniels, Virginia Grande Castro och Björn Victor
Examinator: Björn Victor

Sammanfattning

Människor delar mer och mer på social medier medvetna om att de blir övervakade, men omedvetna om i vilken utsträckning och på vilka sätt datan är processad. Idag används mycket resurser för att urföra dessa uppgifter. Med det här projektet visar vi att det är möjligt att samla in, processa och spara data från sociala medierna Reddit och Twitter i realtid genom att enbart använda en persondator. Vårat fokus har varit att använda ansiksigenkänning för att identifiera specifika individer från en dataström, men datan kan användas för andra syften. Vi har också kollat på de etiska dilemman som dyker upp i samband med insamling och processning av sådan data.

Contents

1 Introduction **1**

2 Background **2**

2.1 Surveillance 2

2.2 Social Media 2

2.3 Biometric Identification 4

2.4 Who Can be Tracked 4

2.5 General Data Protection Regulation 5

3 Ethics **5**

4 Purpose, Aims, and Motivation **6**

5 Related Work **7**

5.1 Surveillance Using Live Feed Cameras 7

5.2 Facial Recognition Using Social Media Data 8

6 System Structure **9**

6.1 Data Collection 9

6.2 Processing 10

6.2.1 Ingest 10

6.2.2 Facial Recognition 10

6.3 Monitoring 10

7 Approach **10**

7.1 Data Collection 11

7.1.1 Collecting Data for the Pipeline 11
7.1.2 Choosing Social Media 12
7.2 Faking Data 12
7.3 Processing 13
7.3.1 Facial Recognition 13
7.3.2 Database 14
7.4 Scalability 14
7.4.1 Modularization 14
7.5 Queue 16

8 Requirements and Evaluation Methods **16**
8.1 Data Collection 16
8.2 Data Processing 17

9 Evaluation Results **17**
9.1 Reddit 17
9.2 Twitter 17
9.3 Processing 18

10 Data Collection from Social Media **18**
10.1 Reddit 18
10.2 Twitter 18

11 Processing **19**
11.1 Queues 19
11.2 Ingest 19
11.3 Facial Recognition 19

11.3.1 Serializing 20
11.3.2 Classification 20
11.4 Database 20

12 Result and Discussion **20**

13 Conclusions **21**

14 Future Work **22**

1 Introduction

We increasingly use social media to share experiences with the world using text, images, and video. Simultaneously countries and organizations are deploying systems to collect data on their citizens and users [20]. Surveillance is not a new concept though. Police and security services have long used surveillance techniques to conduct investigations. State actors such as China and private businesses such as Facebook now have access to an unprecedented amount of data [9, 4]. They also have access to computer systems that can sift through and analyze it at speeds no human could achieve [19]. Our aim with this project is to show and raise awareness that this kind of processing and collection is not limited to large organizations and states, but also possible for those with fewer resources.

A report by Pew Research Center claims that there has been a tenfold increase in social media usage from 2005 to 2015 [30]. More precisely, 7% of all American adults used social media in 2005, and by 2015 the number had reached 65% [30]. With more of the population moving towards using social media, it could prove to be a valuable source for surveillance data.

Social media platforms such as Twitter, Facebook, Snapchat, Reddit, and Instagram expose user data to the Internet, both through API's but also through the interface [32, 8]. Users of these services are often confused by the privacy policies describing the data collected even though people care about their digital footprint [31].

This project aims to show that it is possible to gather, process, and store data from the social media platforms in real-time using only a personal computer. It will focus on processing the data using facial recognition to find specific individuals in it.

It is essential to consider whether it is ethical or legal to collect personally identifiable data without consent. Even though the project only aims to show that it is possible, the cause does not necessarily justify the means. Therefore our system has been cut in half. Our system still downloads data, but it will not save it anywhere. In place of the data from social media we use a data source emulating Reddit. This way, no "live" data will be collected nor saved.

The project resulted in us being able to collect and process a data stream comparable to the one from Reddit in real-time. We did not evaluate the accuracy of the facial recognition component.

2 Background

This section will begin with a brief description of surveillance. Then, an overview of current social media platforms, biometric identification, who can be tracked using facial recognition on social media, and lastly the General Data Protection Regulation.

2.1 Surveillance

It is important to define the term surveillance. One definition is that surveillance is the observation of personal details for purposes such as influence, management, protection, or direction [23]. It is important to stress that it encompasses more than just physical surveillance such as using cameras but rather general data collection. This definition is what we will refer to in this report.

It has been possible to surveil people even before computers and cameras. Early on surveillance was done physically. This kind of surveillance requires significant resources since everything must be done manually by people following the ones being surveilled [43].

The first known use of video surveillance was in 1942. This technology made it possible to surveil a larger amount of people [22]. The video surveillance systems had at the beginning of their usage no functionality to record videos, so manual watching of them was required. Some years later, it became possible to record videos, and in 1990 the surveillance systems were able to show video footage from multiple cameras on one screen.

In recent years China has become an example of how it is possible to surveil a large population using modern technology. An article published in NPR [34] details how facial recognition systems paired with surveillance cameras are deployed to monitor citizens. Using this monitoring system they collect information about their citizens. The information is then combined into one "score" - a so-called social credit score. The score is then used to determine access to everything from internet speed to if one is allowed to keep a dog [24].

2.2 Social Media

Social media usage has increased rapidly in the last couple of years. In the last year alone there has been a 9% increase in active social media users worldwide, to a total of 3,5 billion active social media users which amounts to 45% of the world's population

[12]. Information interesting for surveillance on social media are personal details such as geolocation, posts, images etc. [23]. Following is an overview of the data made available by five currently large social media platforms. This data is accessible either through application programming interfaces (APIs) that give programmatic access to data, or through web-scraping, a method where a computer acts as a user to collect data.

- **Facebook**

  Facebook, as a platform, has a focus on communication with friends. A user can add individuals to their friend list and then share text, images, and video that then appears in their friends' news feed. It is also possible to chat, play games, etc. By default, everything shared is only visible to friends. Facebook also has a requirement that users use their real identities.

  Facebook provides an API for developers to access data, though each individual must explicitly allow access to their data before it can be accessed through the API [8]. The British firm Cambridge Analytica accessed the data of 80 million Facebook users through an app [39].

- **Instagram**

  Instagram has a focus on sharing photos and videos with followers; a user can follow any number of other users of the platform. When a photo or image is shared, it is sent to the users' followers.

  Unlike Facebook, Instagram does not provide an API where data can be collected, though it is possible to scrape public data from user profiles on their website [15].

- **Snapchat**

  Snapchat is comparable to SMS. It can be used to send ephemeral images, videos and chat messages to friends that by default expire after they are viewed. It also has a feature named "Snap Map" where users can publicly post images and videos on a world map.

  Similarly to Instagram, Snapchat does not provide an API, though the Snap Map can be scraped to get the content and the location data associated with it [36].

- **Reddit**

  Reddit is a platform where users can submit text, links, images, and video to so called subreddits where they are publicly displayed. Users are also encouraged to use disposable identities.

  They expose an API which facilitates the collection of all user-generated data, both comments and posts. This data consists of the author, a title, a timestamp, and the text, video, link, or image posted by the user [32].

- **Twitter**

  Twitter is a social platform where people create blogs. On these blogs, people can publish short text segments, pictures, or videos called tweets that are publicly available by default.

  It is possible to get a sample of tweets published in real-time from their streaming API. There are full versions of the stream available as well, though these require an agreement with Twitter. It means that API access is limited. Twitter also includes a clause that forbids the usage of their API for surveillance [40].

2.3 Biometric Identification

Facial recognition software makes it possible for a device to find and identify people using images of their faces. By computing a unique identity based on facial features, it is possible to check if two photos of a face have a matching face in them. [35].

There are other ways to identify a person e.g., iris scanning and fingerprint scanning, of which the latter is frequent in applications such as smartphones [45]. The advantage of facial recognition over other methods of biometric identification is that the subject can be mostly unaware of it and that the source data can be found publicly. It is challenging to hide the collection of fingerprints, such as when applying for a biometric ID or passport since the subject needs to place their hand on a scanner. Modern techniques such as in-display fingerprint scanners can partially hide the data collection but are still obvious.

2.4 Who Can be Tracked

By using social media, a large number of people can be tracked. According to Hootsuite, there were 3.5 billion social media users 2019 [12]. Those 3.5 billion users are not posting images regularly; some might not even post images, which means that there are people that cannot be tracked directly using facial recognition. First, the number of social media users is growing [37], thus increasing the number of people that can be tracked. Second, as the number of users grows, more people can be tracked indirectly. Indirect tracking can be achieved using images where the faces of not only the subject appear, such as in a crowd or group photo, creating a network effect - more users of social media will increase the reach of the tracking.

2.5 General Data Protection Regulation

Since our work is done in the European Union the General Data Protection Regulation (GDPR) is of the upmost importance. The GDPR sets out a framework of rules and guidelines on the collection and processing of personally identifiable data [7]. A key part of the regulation is protecting the privacy and fundamental rights and freedoms of citizens whilst still balancing it with legitimate business, research and governmental interests [6].

Special attention was paid so that personally identifiable data was not processed. Thus the project was divided into two separate parts. The first part is where we show that it is possible to gather data from social media whilst not storing nor processing it. In the second, the processing part, we used fake data.

3 Ethics

Surveillance and the processing of personally identifiable data (PII), especially biometric data, is an often debated topic [31]. There are reasons both for and against processing PII. On the one hand, businesses and governments have legitimate interests such as upholding contracts. On the other hand, individuals have certain rights and freedoms. Both of these codified in regulations such as the GDPR [7]. As noted in previous sections, countries such as China argue that surveillance is necessary to uphold peace, security, and the rule of law. At the same time, human rights organizations say that this processing can be used to violate fundamental human rights, and in the case of China it is used to violate human rights [20, 14]. It is apparent that there are diverse viewpoints on this topic, and this section is aimed to give a brief introduction to them [31].

We begin with government surveillance. A common theme is that such surveillance facilitates upholding the rule of law. As noted previously, China has been subject to critique for its comprehensive surveillance policies, especially in the province of Xinjiang; however, these kinds of policies are not limited to them. Other countries use surveillance as well though to a lesser extent. One such example is Australia, where a new bill was passed giving government agencies the right to demand access to encrypted communications [16]. Other examples are the UK with the Investigatory Powers Act., or the Patriot Act in the US, giving similar powers to their respective governments [16, 17].

The UN sustainable development goals align closely with the justification given for such powers. Goal 16 promotes peace, justice, and strong institutions. In particular, goal 16.A - strengthening relevant national institutions to prevent violence and combat

terrorism and crime [42]. For example, the Patriot Act mentioned previously is a direct response to terrorism [17].

Governments are not the only ones using their surveillance powers, so do private businesses. Google, Facebook, and other platforms collect data from their users, store it, and then process it. There are multiple reasons for this. In the early days of Google, they exclusively used the data for service improvement, feeding it back into their system so that search results were as relevant as possible [46]. While Facebook only used the data as a part of their service. Nowadays, these platforms collect more data than ever and use it for more than just service improvement. Google, Facebook, Twitter, and others provide marketing platforms that use the data collected to serve "relevant" ads or sell products based on this data such as the Twitter Trends API [46, 41]. In other words, the data and its byproducts are bought and sold in a kind of data market driving revenues and profits for the businesses with data [46].

It seems that the justifications are reasonable. Businesses must recoup costs and users accept their terms, and if surveillance helps in protecting society against terror and crime, it surely is reasonable. However, this does not seem to be the case. Studies by Pew Research Center found that two-thirds of Americans think that current laws are inadequate in protecting their privacy and that only 50% felt confident that they understood how and for what their data is used [31].

In the supreme court case "Warden v. Hayden" Justice Douglas, W. argues that privacy involves the choice of the individual to disclose information [38]. This notion of individual choice and free will is one of the basic tenets of liberalism and in turn, commonly democracy [2].

In Shoshana Zuboff's latest work she argues that it is impossible to grasp the issues when one is unable to comprehend and describe them and claims that privacy is redistributed from the individual to other entities, specifically businesses, and in our case even further [46]. Before the issues of data collection can be addressed individuals must understand the implications of data collection. For this there is a need to find ways of educating people. This aligns with the aims of our project. To give insight into the data that can be collected from social media platforms and show one way of processing this data.

4 Purpose, Aims, and Motivation

The aim is to show that it is possible to gather, process, and store data from the social media platforms Twitter and Reddit in real-time using only a personal computer, though still scalable to several computers. The processing of the real time stream of data was

done using facial recognition to find specific individuals. The real-time stream should also be saved so that it can be used for future purposes, however in this project we used fake data for the real-time stream as described in section 7.2. The implementation of these aims must be done in accordance to the specification in Section 2.5 about the General Data Protection Regulation necessitating a split between the data collection and processing so that personally identifiable data is not stored nor processed.

Facial recognition and machine learning have become increasingly more important subjects and can be used in the interest of people but also as forces of evil. By showing the possibility of connecting and applying disparate technologies using social media data for tracking movements and interactions the hope is to raise awareness to information security and to the fact that the resources needed to complete such a project are low. There are already providers of such services working with government agencies without being in the public eye; one such company is IntelCenter [25].

An important issue is the ethical concerns that the possibility of tracking people without their knowledge brings. Legislators in the European Union introduced the GDPR to promote the responsible use of personally identifiable data [7]. Although regulations do not stop nefarious actors which again is why this work is important - to bring these issues into the public discourse.

At the same time, tracking individuals linked to crime can bring society closer to achieving the UN sustainability goals, which is more closely discussed in Section 3. Helping criminal investigations and working as a deterrent such a tracking system can be a force of good as long as it is not used for such things as racial or religious profiling [20].

5 Related Work

This section will describe three different projects. Starting with more general uses of camera surveillance, and ending with a project showing how it is possible to find specific individuals on social media.

5.1 Surveillance Using Live Feed Cameras

There are multiple technologies for tracking people using live feed cameras. These technologies are especially interesting for the police force as locating and identifying people is one of their main tasks. It can be criminals or missing people. There has been previous work looking at this [33]. The solution involved mounting a camera on

police officers. The mounted camera sends a live video feed to a server. On the server there is a database with criminals and missing people and the video will continuously be compared to the database until a match is found.

There is also a possibility of using surveillance cameras for the live video feed. The main advantage of this is that there are already a lot of cameras installed throughout different cities. Live facial recognition on video surveillance feeds is already being trialled in the UK. The technology is provided by a Japanese software company called NeoFace technology [26]. Our system is different since it does not use live feed video but instead pictures from social media.

Also there was recently an article in the New York Times about a surveillance system using facial recognition [5]. In the article, they described how they used a camera that was open for public use to do facial detection in Bryant Park. To determine which faces in the video stream belonged to which person they used publicly available photos of people working near the area. Doing this they were able to find and track people, the objective of the work was to raise awareness of this issue. They emphasize how the government could do the same thing as they did but at a larger scale. Our work aims to do something similar, but instead of using live cameras we will use images from social media.

5.2 Facial Recognition Using Social Media Data

Recognizing people on social media using facial recognition has been done before. A program that does this is the open source tool Social Mapper [44]. This tool helps the user find an individual's different social media accounts without having to look them up manually. If you have the name and a picture of a person you want to collect social media accounts from it is straightforward, just provide the tool with this information. Social mapper will then compare the photo you gave it with pictures on the most common social media applications using facial recognition. After this comparison, it will list the social media accounts of this person.

Social mapper relates to our program since both of our programs use facial recognition to find people on social media. Other than that our system is fairly different e.g., we use pictures from live data streams while they use already posted photos for facial recognition. The usage of our system is also different. It collects information about people while social mapper provides us with different accounts.

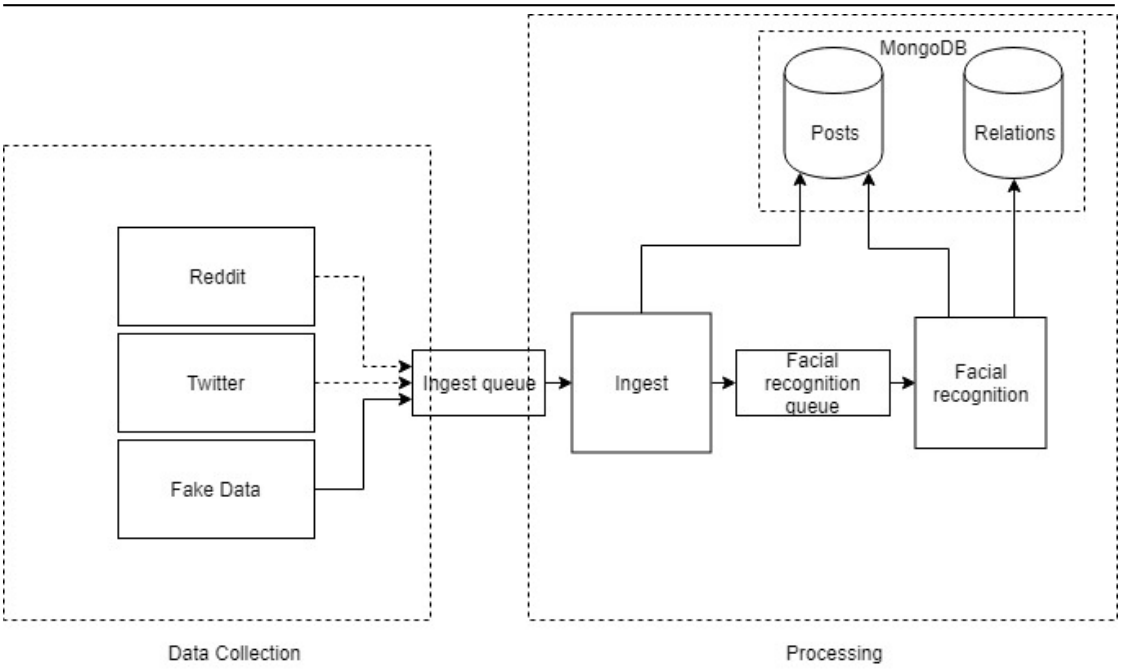

Figure 1 System overview of the data collection and processing

6 System Structure

Our system is divided into two parts, seen in Figure 1. The first part is the data collection part which gathers data from social media, specifically from Twitter and Reddit. The second part is the processing of fake data using facial recognition.

6.1 Data Collection

The first part of the system is data collection, which is divided into two modules - one collecting Reddit data and one collecting Twitter data. The data from these applications is collected as it is posted, creating a live stream. From the data, the system collects image URLs, usernames, geolocations, and additional text.

Instead of processing personally identifiable data, a fake data source is used. This data source emulates data collection from Twitter and Reddit. The rate at which fake data is sent is configured to correspond to the rate at which posts are published on the chosen social media platform. The data is then sent to the ingesting queue as a standardized JSON object, see the JSON object example in Figure 6.

6.2 Processing

In the processing part of our system we process fake data sent from the data collection.

6.2.1 Ingest

In this part of the system we have a ingest queue that fetches JSON objects with data from our fake data stream. We add the received data to our database of posts. The ID and the image URL of the post is then sent to the facial recognition queue.

6.2.2 Facial Recognition

From the facial recognition queue JSON objects are received. We check if an object contains an image URL, and if it does we download the image. If our system successfully downloads the image, it checks if there are faces in it. In the case that we do not find a face the image is thrown out. If our system detects any faces, they are compared with our dataset of stored faces to find ones that are similar to them. If a match is found the names of the people in the image are sent to our relations database.

6.3 Monitoring

Queue monitoring is required to gain insights into performance issues, specifically concerning queue length. The message broker RabbitMQ that is used for queues in the system has a built-in management website which shows basic metrics such as messages in and out of each queue. Though we opted for Prometheus, a time-series database, to save metrics since RabbitMQ has support for exporting metrics to it and Prometheus exposes a query API. We also used Grafana to create a visualization of the pipeline state using the data from Prometheus. Grafana displays the current backlog and graphs of the messages in and out of the ingest and facial recognition queues. A picture of the setup is found in Figure 2.

7 Approach

In this section, it is described how the pipeline was designed. We start by discussing data collection. Here we cover how our data is collected, what social media we are

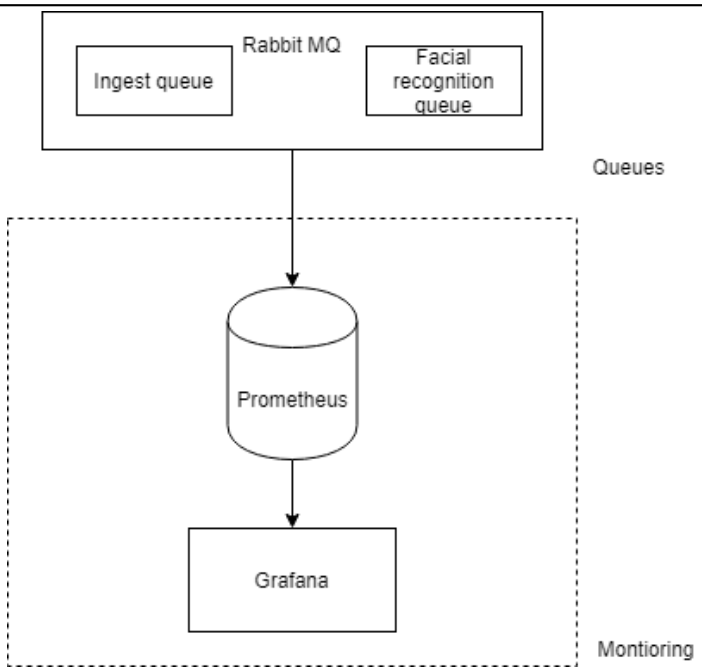

Figure 2 System overview of the queue setup and the monitoring

collecting from and lastly how we handle the data. After that section we move on to the processing of the data. In the processing section we cover two things, how we chose our facial recognition system and our database choices. When processing has been discussed we move on to system architecture and scalability. Why we made it scalable, how we made it scalable and the design choices related to scalability are covered here. Lastly we have a section about the logging of our system, there we discuss why it is important and what tools we used.

7.1 Data Collection

In this section, we discuss the data collection part of our system.

7.1.1 Collecting Data for the Pipeline

There are two main ways to collect data from social media, through web scraping and APIs.

Web scraping is the use of computational programs to extract information from web

pages by processing them, e.g. social media sites [11]. The author describes that this is useful when social media do not provide any official access to their data or when seeking data other than what is officially provided, although doing so may breach the platforms terms of service [11]. Since using web scraping is notWe did not choose to employ web scraping in our project since we wanted to comply with the data providers.

Another alternative is receiving data directly from the social media site through APIs [11]. It is possible to buy full datasets from platforms such as Twitter [11]. This is however something that someone with limited resources would not do. Some companies offer public APIs with samples of data for free [11]. Since public APIs provided by the social media companies are open to use we decided to use public APIs in this project.

7.1.2 Choosing Social Media

Facebook: As mentioned in Section 2.2, Facebook requires permission to be given by each user to collect their data, at least to collect their data from the official API. We decided that collecting consent from users is out of scope for this project.

Snapchat: As discussed in Section 2.2 Snapchat has a feature called the Snap Map. By scraping the site that it is hosted on it is possible to retrieve publicly posted pictures and videos with location data attached. Since we chose not to use web scraping, Snapchat was not used as a data source.

Reddit: In Section 2.2 it is mentioned how Reddit has an API. The API provides a way for us to stream everything posted on Reddit live. Since accessing their API is allowed and free of charge we decided to use Reddit as one of our data sources.

Twitter: has a streaming API that provides a small sample of all tweets. It is not known how the tweets are sampled [11]. This stream is provided free of charge but it is possible to pay to receive an even larger stream [41].

To conclude Reddit and Twitter were chosen as the data sources for our project. The other options could have been viable but due to company policy and the scope of work required to collect the data they where not chosen.

7.2 Faking Data

Since we did not want to process real social media data due to ethical and legal concerns we decided to generate fake data. The fake data consists of images from the Labeled Faces in the Wild dataset and stock photos without faces in them [13]. We decided to

have different types of photos to resemble social media content better. The fake data is sent to our system the same way that data from Reddit and Twitter would have been sent.

7.3 Processing

In this section we discuss the design choices for our facial recognition system and the database.

7.3.1 Facial Recognition

The goal with the facial recognition system is to receive a picture and by examining the faces in it determine their identity. To do this, a data set with pictures of known people was needed to compare the received image to, and a system for the actual comparison.

Creating a large data set of labeled faces manually is a long and gruesome task. A labeled face is an image that contains a face where the owner of the face is known. It is however not necessary to do manual markings. There are datasets that can be found online that contain labeled faces. One of those is Labeled Faces in the Wild [13]. The images in the dataset are in natural conditions with varying poses, lighting, traits, accessories, occlusions, and background. Since the pictures from social media are taken in varying conditions we decided that it is appropriate.

Comparing faces with the dataset can be done in multiple ways. Since the aim of the project is to show that it is possible and relatively simple, we decided to use the face_recognition library written in Python hosted by Adam Geitgey [10]. The library implements an API to DLIB facial recognition functions [18]. DLIB is a library containing machine learning algorithms written in C++. The face_recognition library allows us to convert images with faces in them to multiple vectors with one for each face. If two vectors are close to each other, they are likely from faces of the same person. After converting the entire labeled dataset into vectors, the next step is classifying the faces in the received images. This was resolved by using the k-nearest neighbor algorithm (KNN). Using KNN, it is possible to tell for each of the faces in the received image if it is similar to a face in the dataset.

7.3.2 Database

To store the collected posts together with the facial vectors and also the names of the people discovered in the photos a database was needed. We decided to base our choice of a database on a paper by Parker et al. [28]. The authors compare the NoSQL database MongoDB to an SQL database. In the paper, it is discussed how MongoDB is a good choice for a database without a rigid schema. It also mentions how MongoDB has a higher performance than SQL databases on insertions. Since there might be a need to change the database schema in the future, we decided to use MongoDB due to it not needing a rigid schema. Another contributing factor to our choice is that MongoDb is faster at insertion, and since one of our goals is to collect data in real-time having a faster insertion helps.

7.4 Scalability

One of the aims of the project is that the system is scalable. In this section the modularization of the system and how this facilitates scalability will be described in detail.

7.4.1 Modularization

A requirement is that the system has to be scalable, i.e. there can be more than one instance of each module, on more than one server. Further, the modules have different hardware requirements, e.g. facial recognition benefits from access to a GPU. Thus a way to configure and deploy the system was required. Using virtual machines or containerization can solve these problems. The differences between the two solutions are illustrated in figures 3 and 4. Mainly, virtual machines (VMs) run a guest operating system (OS) on top of the host operating system unlike containers that run directly on the host reducing the resource footprint [27].

Containerization solutions such as the Docker platform provides tools to package applications and deploy them [27]. We opted for containerization using Docker. This was done since our system does not require perfect isolation. VMs also require more resources due to the guest OS which is contrary to the aim of using as little as possible. In addition, Docker is commonly used as a containerization platform, thus it has an established support platform and user base [27].

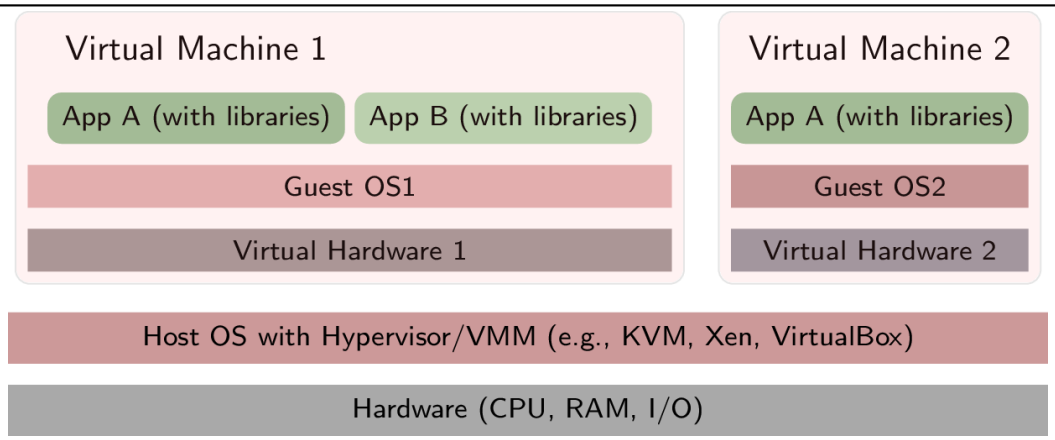

Figure 3 Virtualization overview [21]

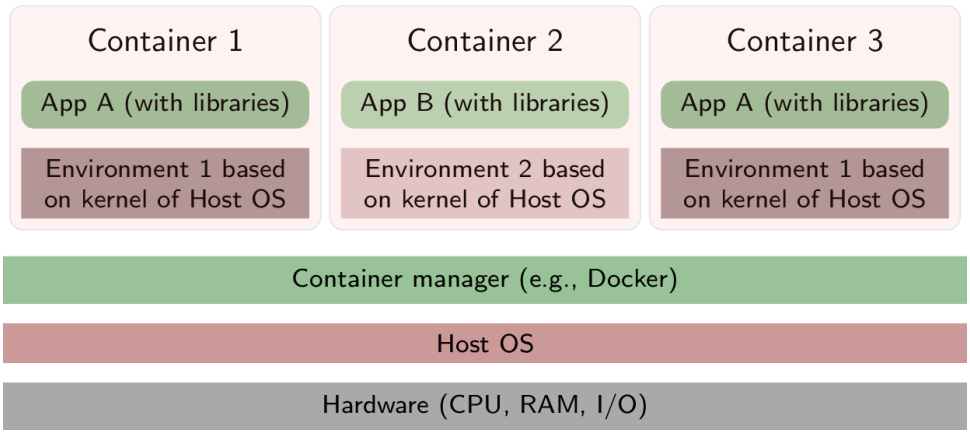

Figure 4 Container overview [21]

7.5 Queue

To manage the data flow between containers we employ queues. Queues allows each container to consume at their own rate. They also allow the data collectors to push data as fast as possible. This is great since the data rate may not be constant, thus ensuring that we can build up a backlog during spikes of data collection and consume it when it goes down.

We opted to use first in first out (FIFO) queues in this project, since we wanted to process in real-time. The queues where implemented using RabbitMQ.

8 Requirements and Evaluation Methods

There are two distinct parts in the system - data collection and data processing described in Section 6. Data collection as outlined in 7.1 cannot be used in conjunction with the data processing pipeline due to the ethical and legal concerns described in the Section 2.5 and 3, thus its evaluation must be done separately from the processing. Even though the assessment is done separately the resource usage of both parts needs to be accounted for so that both could run simultaneously. The evaluation methods used are detailed in the sections that follow. The tests are done on a machine with 8 CPU threads at 5GHz without GPU acceleration. This machine is

8.1 Data Collection

Data collection is limited by two resources - CPU and network bandwidth. Since the data is collected from external services on the internet bandwidth limits the rate at which messages can be ingested. The CPU is a limiting factor due to the processing required to parse and adapt the data to the requirements of our system. Thus the resource usage is dependent on the rate at which data is collected, faster data collection requires more processing and more network bandwidth. The data collection also shares resources with the processing part of the system. This sharing has to be considered and balanced when evaluating the system. If the data collection is faster than the data processing the queues will fill up causing messages to be backlogged.

The data collection will be evaluated by how much resources that is required to collect data from Twitter and Reddit. It will also be evaluated by if it is possible to collect Reddit data while simultaneously processing data with the data rate of Reddit.

8.2 Data Processing

When evaluating data processing we have to consider two things, accuracy and speed. We have little control over the accuracy of the facial recognition system since it is outside the scope of this project. Studying the results also requires manual control of accuracy. For data on the accuracy of the library used refer to [10]. Speed is evaluated by how fast messages can be processed.

As noted in the Section 8.1 speed is dependent on the available resources and has to be balanced with the data collection. One way to increase the processing speed is to add more servers hence parallelizing the workload.

A particular requirement of the system was to handle the data rate of Reddit using the resources detailed in the beginning of this section. It was tested using fake data. The faked data consisted of images from the Labeled Faces in the Wild (LFW) dataset [13] and images without people.

9 Evaluation Results

In this section the results of our evaluation are presented.

9.1 Reddit

The rate at which posts from the Reddit API are received is approximately 0.9 posts per second. We only collect the posts that contain an image. This uses approximately 1 % CPU and 26MB RAM.

9.2 Twitter

The rate at which the Twitter sample API sends tweets is approximately 10 tweets per second. Out of a sample of 1 million tweets, 53% contain images. The resource usage of this module is approximately 3% CPU and 70MB of RAM when streaming data.

9.3 Processing

The system was tested using a fake data stream similar to the one of Reddit on a machine with the specifications mentioned in Section 8. The bottleneck of the processing is the facial recognition module. With one instance of the facial recognition module consuming from the queue, the queue kept growing. If we scale the module to two instances our system could process all the data from Reddit in real-time.

10 Data Collection from Social Media

In this section the implementation of the data collection is described.

10.1 Reddit

To collect data from Reddit we used the python library PRAW. PRAW is a Reddit API wrapper from which we could collect an image URL, a title, a writer, and a text for the posts. We collected data from the subreddit /r/all in a real-time stream. For the live streaming of data we used the streaming functionality that PRAW provides. When the data is collected a JSON object with the data is created and sent to the Ingest queue.

10.2 Twitter

The Twitter API has a so called "streaming" API meaning that a long lived HTTP connection is used to continuously send new data. The API sends a number of message types including tweets and status messages. We implemented the module in Elixir using a supervisor and two processes.

The first process handles the connection to the streaming API buffering incomplete messages, parsing complete messages and sending them to be preprocessed. A typical Twitter message is JSON formatted as specified in the Twitter API documentation, an example response can be found in Figure 5 [41, 1]. It begins with an integer indicating the size including the newline at the end of the message, followed by the JSON object, and ends with "\r\n". When an entire message has been read it is parsed excluding the integer specifying its size and the newline in the end.

Parsed messages are then consumed by the preprocessor. It modifies message structure to fit our system requirements and further parses and converts timestamps and the list of

images. A more detailed specification of the message structure used in our system can be found in Figure 6. When ready the message is sent to the ingest queue of our system.

11 Processing

In this part we describe the implementation of the processing part. We cover how the facial recognition works and how data is saved.

11.1 Queues

To facilitate communication between containers we used RabbitMQ. RabbitMQ is a message broker, in the system it provides queues between modules. The queues store data until a consumer such as the facial recognition is ready to receive new data. This is also used to enable scaling where several instances of one module can fetch messages from the same queue. An example message sent in the queues can be found in Figure 6.

11.2 Ingest

The first step in the processing is the preprocessing. This is where all the data collected from social media goes. It is responsible for creating the initial entry in the database for each social media post that is collected. To implement this we decided to use the programming language Python and as described earlier the main traffic in the system is managed by RabbitMQ. To communicate with rabbit we needed a implementation of the communication protocol AMQP. We resolved this by using the pika library. To communicate with the database we used the pymongo library. Pymongo allows us to connect to the database and make changes.

11.3 Facial Recognition

The facial recognition part of our system is built in Python. It consumes images that is going to be processed from a queue, it also writes the results to the same database as the ingest part does. The facial recognition part does two things in the processing, serializing faces to an easy to compare format and comparing them to earlier processed images to find a match.

11.3.1 Serializing

To do the serializing of the images we used the face_recognition lirbrary. The library maps a face into a 128 dimensional vector space. Serializing faces is the most compute heavy part of our system and it is possible to do it in two ways, using GPU or CPU. When starting the face recognition container it is possible to decide which to use by switching a flag from true to false.

11.3.2 Classification

The classification solution that we implemented is divided into two parts, processing a labeled dataset of images and classification. Classification is the process of identifying a face from a database of known encodings. Preprocessing involves serializing the labeled images as described in Section 11.3.1 and saving the encodings to a database. To classify a new face its encoding is compared to the database created in the preprocessing step using the k-nearest neighbor algorithm which finds the closest match of a known face by distance. The distance can be calculated since the face encodings are vectors. Scikit learn was used to implement the k-nearest neighbor classification [3].

11.4 Database

As discussed in the method we decided to use a NoSQL database called MongoDB instead of a SQL database. The database is hosted in a Mongo container, it is made up of two collections. The first collection contains all the posts that we have processed and the associated serialized facial data. The second collection contains all the instances of recorded matches, and also an id linking each entry to a post in the first collection.

12 Result and Discussion

In this project we have explored the possibilities of using facial recognition on social media posts in real-time. We did this as a first exploration to show what is possible to extract from the data published on social media. This was done by creating a system that collects data from social media as it is being posted and processing it in real-time. We decided to divide our system in to two parts, a data collection part and a processing part. This was done due to ethical, legal and contractual concerns connected to developer

agreements with Twitter. Instead we measured the data rates of the two platforms and created a data source that fakes data.

We have briefly explored the possibilities of streaming live data published from social media. We built proof of concepts for collecting data from Twitter and Reddit. From Reddit it is possible to collect 100% of the posts. From Twitter we managed to collect only a sample of tweets due to the API restrictions. It is possible to buy access to all of the data, but since the goal was to do this with low resources it was not a possibility for this project. It would be interesting to explore more social media platforms but due to API policy reasons we chose not to proceed with some of them, specifically Facebook, Snapchat and Instagram.

In the processing part we built a system that shows that it is possible to recognize people in images from a real-time stream. We do this by comparing the person to a labeled dataset. Our goal was to be able to handle everything posted in real-time on Reddit. The test described in Section 9.3 showed that we were able to handle the stream, which means that we reached or goal. However that does not mean that we would be able to process all social medias in real-time, but at least the entirety of Reddit.

One possible improvement of the implementation would be if we after storing the facial data also added it to the data set that were searched through to find a match for the incoming images. Since we collect the name and images associated to it in the data collection part we could also have labeled the faces before adding them so that we could recognize new people and not just the ones we had in the initial data set. We made the decision that this was not the most important part of our system.

We show with our system that it is possible to recognize a person by comparing the collected image with our database of stored faces. We were also able to save the facial data that was created from the photos.

13 Conclusions

Our project implements a facial recognition sytem that uses real-time social media data as the source. Two data collection modules to gather data in real-time from social media were implemented. The facial recognition allows us to locate certain people from real-time data and also the buildup of a database of facial data that can be used later.

The goal was to show that it is possible to collect and process image data from social media for surveillance purposes. We have demonstrated the possibility of collecting facial data both from Reddit and Twitter, two prominent social media platforms. We

have also shown one method of how this data can be used. By showing how easy our system was to implement we hope that we will raise awareness about the possibilities to surveil people using social media.

14 Future Work

This project has shown that it is possible to collect personal information from social media. With this information, we believe it is possible to tell even more about people, e.g. peoples location and political alignment. The system also collects the text from tweets and Reddit submissions, which potentially includes information about locations, friends, interests etc. Natural language processing can be used to extract this data to build more extensive profiles.

Only Reddit and Twitter are implemented as data sources in the system though it is possible to collect data from more social media platforms. Both Snapchat and Instagram are promising. Snapchat has a service called the Snap Map. It allows users to post photos and videos to a publicly available map, thus providing both image and location data that can be scraped. Instagram does not have publicly available APIs, though it is possible to scrape data from there since it is accessible without authentication.

Currently, the system is only able to use a database containing labeled images such as Labeled Faces in the Wild. Another future improvement is to add self-training so that the database is updated with new faces from the streamed data. It would allow the system to improve the facial recognition accuracy continuously.

References

[1] T. Bray, "The JavaScript Object Notation (JSON) Data Interchange Format," Internet Requests for Comments, RFC Editor, RFC 8259, December 2017, retrieved 2019-04-28. [Online]. Available: https://tools.ietf.org/html/rfc8259

[2] G. W. Brown, *The concise Oxford dictionary of politics and international relations: edited by Garrett Wallace Brown, Iain McLean, and Alistair McMillan*, 4th ed. Oxford, United Kingdom: Oxford University Press, 2018.

[3] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly,

B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[4] C. Cadwalladr and E. Graham-Harrison. (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. Retrieved 2019-05-06. [Online]. Available: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

[5] S. Chinoy. (2019, April) We built an 'unbelievable' (but legal) facial recognition machine. Retrieved 2019-04-17. [Online]. Available: https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-new-york-city.html

[6] EU. EU Charter of Fundamental Rights. European Union. Retrieved 2019-04-21. [Online]. Available: https://fra.europa.eu/en/charterpedia/title/ii-freedoms

[7] EU. (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council. European Union. Retrieved 2019-04-08. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

[8] (2019) Facebook for developers. Facebook. Retrieved 2019-05-06. [Online]. Available: https://developers.facebook.com/docs/graph-api

[9] R. Gallagher. (2016) Documents reveal secretive U.K. surveillance policies. Retrieved 2019-05-06. [Online]. Available: https://theintercept.com/2016/04/20/uk-surveillance-bulk-datasets-gchq/

[10] A. Geitgey. (2017) Face recognition. Retrieved 2019-04-17. [Online]. Available: https://github.com/ageitgey/face_recognition

[11] S. Halford, M. Weal, R. Tinati, L. Carr, and C. Pope, "Understanding the production and circulation of social media data: Towards methodological principles and praxis," *New Media & Society*, vol. 20, no. 9, pp. 3341–3358, 2018. [Online]. Available: https://doi.org/10.1177/1461444817748953

[12] (2019) Digital 2019. Hootsuite Inc. Retrieved 2019-05-06. [Online]. Available: https://p.widencdn.net/kqy7ii/Digital2019-Report-en

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[14] (2019) UN: China Responds to Rights Review with Threats. Human Rights Watch. Retrieved 2019-05-20. [Online]. Available: https://www.hrw.org/news/2019/04/01/un-china-responds-rights-review-threats

[15] (2019) API endpoints. Instagram. Retrieved 2019-05-06. [Online]. Available: https://www.instagram.com/developer/endpoints/

[16] P. Karp. (2018) Australia's war on encryption: the sweeping new powers rushed into law. Retrieved 2019-05-20. [Online]. Available: https://www.theguardian.com/technology/2018/dec/08/australias-war-on-encryption-the-sweeping-new-powers-rushed-into-law

[17] O. S. Kerr, "Internet surveillance law after the USA Patriot Act: The big brother that isn't," *Nw. UL Rev.*, vol. 97, p. 607, 2002.

[18] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[19] A. Kofman. (2018) Interpol rolls out international voice identification database using samples from 192 law enforcement agencies. Retrieved 2019-05-06. [Online]. Available: https://theintercept.com/2018/06/25/interpol-voice-identification-database/

[20] L. Kuo. (2019) Chinese surveillance company tracking 2.5m xinjiang residents. Retrieved 2019-04-07. [Online]. Available: https://www.theguardian.com/world/2019/feb/18/chinese-surveillance-company-tracking-25m-xinjiang-residents

[21] J. Lechtenbörger. (2018) Figures. Retrieved 2019-05-20. [Online]. Available: https://gitlab.com/oer/figures

[22] A. Longdin. (2014) The history of CCTV – from 1942 to present. Retrieved 2019-05-09. [Online]. Available: https://www.pcr-online.biz/2014/09/02/the-history-of-cctv-from-1942-to-present/

[23] D. Lyon, *Surveillance Studies: An Overview*. Cambridge CB2 1UR, UK: Polity Press, 2007.

[24] A. Ma. (2018) China has started ranking citizens with a creepy 'social credit' system — here's what you can do wrong, and the embarrassing, demeaning ways they can punish you. Retrieved 2019-04-08. [Online]. Available: https://nordic.businessinsider.com/china-social-credit-system-punishments-and-rewards-explained-2018-4

[25] J. McLaughlin. (2016) Private intellligence firm proposes "google" for tracking terrorists' faces. Retrieved 2019-04-07. [Online]. Available: https://theintercept.com/2016/11/04/private-intellligence-firm-proposes-google-for-tracking-terrorists-faces/

[26] Metropolition police. (2019) Live face recognition trials. Metropolition police. Retrieved 2019-04-21. [Online]. Available: https://www.met.police.uk/live-facial-recognition-trial/

[27] C. Pahl, "Containerization and the PAAS cloud," *IEEE Cloud Computing*, vol. 2, no. 3, pp. 24–31, 2015.

[28] Z. Parker, S. Poe, and S. V. Vrbsky, "Comparing NoSQL MongoDB to an SQL DB," in *Proceedings of the 51st ACM Southeast Conference*, ser. ACMSE '13. New York, NY, USA: ACM, 2013, pp. 5:1–5:6. [Online]. Available: http://doi.acm.org.ezproxy.its.uu.se/10.1145/2498328.2500047

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] A. Perrin. (2015) Social media usage: 2005-2015. Retrieved 2019-04-20. [Online]. Available: https://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

[31] (2016) The state of privacy in post-Snowden America. Pew Research Center. Retrieved 2019-05-20. [Online]. Available: https://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

[32] (2019) Reddit API documentation. Reddit. Retrieved 2019-05-06. [Online]. Available: https://www.reddit.com/dev/api

[33] M. Sajjada, M. Nasira, K. Muhammad, S. Khana, Z. Jana, A. Kumar, S. Mohamed, E. Sung, and W. Baik. (2017) Raspberry pi assisted face recognition framework for enhanced law-enforcement services in smart cities. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X17309512

[34] R. Schmitz. (2018) Facial recognition in china is big business as local governments boost surveillance. Retrieved 2019-05-20. [Online]. Available: https://www.npr.org/sections/parallels/2018/04/03/598012923/facial-recognition-in-china-is-big-business-as-local-governments-boost-surveilla

[35] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[36] (2019) Snapchat: Snap map. Snapchat. Retrieved 2019-05-06. [Online]. Available: https://map.snapchat.com/@59.816600,17.623500,12.00z

[37] Statista. (2017) Number of social media users worldwide from 2010 to 2021. Retrieved 2019-04-17. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[38] (1967) Warden v. Hayden. Supreme Court of the United States. Retrieved 2019-05-20. [Online]. Available: https://www.law.cornell.edu/supremecourt/text/387/294#pg_323

[39] (2018) Facebook scandal: I am being used as scapegoat – academic who mined data. The Guardian. Retrieved 2019-05-20. [Online]. Available: https://www.theguardian.com/uk-news/2018/mar/21/facebook-row-i-am-being-used-as-scapegoat-says-academic-aleksandr-kogan-cambridge-analytica

[40] (2019) Twitter API documentation. Twitter. Retrieved 2019-05-16. [Online]. Available: https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

[41] Twitter, Inc. and Twitter International Company. (2019) Twitter Developer Platform. Retrieved 2019-04-28. [Online]. Available: https://developer.twitter.com/

[42] UN. (2018) Sustainability goals 2030. United Nations. Retrieved 2019-04-07. [Online]. Available: https://www.un.org/sustainabledevelopment/inequality/

[43] (1977) Surveillance of premises, vehicles and persons - new agent training. US Department of the Treasury, Bureau of Alcohol, Tobacco and Firearms. Retrieved 2019-05-20. [Online]. Available: https://www.ncjrs.gov/pdffiles1/Digitization/60130NCJRS.pdf

[44] J. Wilkin. (2019) Social mapper. Retrieved 2019-04-07. [Online]. Available: https://github.com/Greenwolf/social_mapper

[45] J. Woodward, C. Horn, J. Gatune, and A. Thomas. (2003) Biometrics: A look at facial recognition. Retrieved 2019-04-17. [Online]. Available: https://apps.dtic.mil/docs/citations/ADA414520

[46] S. Zuboff, *The age of surveillance capitalism: the fight for the future at the new frontier of power*. London: Profile Books, 2019.

```
xxxx
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "1\/ Today we\u2019re sharing our vision for
    the future of the Twitter API platform!
    \nhttps:\/\/t.co\/XweGngmxlP",
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter
       Platform news, updates & events."
  },
  "entities": {
    "hashtags": [
    ],
    "urls": [
      {
        "url": "https:\/\/t.co\/XweGngmxlP",
        "unwound": {
          "url": "https:\/\/cards.twitter.com\/cards
              \/18ce53wgo4h\/3xo1c",
          "title": "Building the Future of Twitter"
        }
      }
    ]
  }
}
\r\n
xxxx
{
.
.
.
```

Figure 5 Example of Twitter streaming API response

```
{
    _id: "5cd55bd4ee42e7b4f86407d6",
    source: "fake",
    source_id: "XXXXXXXXXXXX1",
    timestamp: 1557486548,
    sensitive: false,
    user: [
        {
            id: "XXXXXXXXXXXX1",
            readable: false,
            type: "id"
        },
        {
            id: "Arnold Schwarzenegger",
            readable: true,
            type: "name"
        }
    ],
    text: [
        {
            text: "Lorem ipsum color domet"
        }
    ],
    photos: [
        {
            url: "https://upload.wikimedia.org/wikipedia/
                commons/thumb/b/be/Arnold_Schwarzenegger
                _-_2019_%2833730956438%29_%28cropped%29
                .jpg/800px-Arnold_Schwarzenegger_-_2019
                _%2833730956438%29_%28cropped%29.jpg"
        }
    ]
}
```

Figure 6 Example of message structure used in our system