Integration of Social Media

Bachelor's thesis:

Michael Johansson

2016-06-05

LUND UNIVERSITY
Campus Helsingborg

LTH School of Engineering at Campus Helsingborg

Department of Computer Science

LTH School of Engineering
Lund University
Box 882
SE-251 08 Helsingborg
Sweden

LTH Ingenjörshögskolan vid Campus Helsingborg
Lunds universitet
Box 882
251 08 Helsingborg

Printed in Sweden
Lunds universitet
Lund 2016

Abstract

SAAB Training & Simulation develops simulators for military and emergency services. To enhance background information in these, they wish to add information from social media from various events, such as acts of terrorism. The first step was to study a few selected social media and see how data is accessed and if they can be integrated into a database, preferably the same for all social media. This step also includes an overview into what type of information can be found on each social media. The second step was actually integrating a social media to a database, as a proof of concept.

The studied social media are Facebook, Twitter, Instagram and Google+.

Some legal aspects were also studied, which have implications on storing information about individuals that can be traced back to the person in question.

The result was a conclusive study on the selected social media, with information relating to their APIs, what information can be found on respective social media and also an analysis of the possibilities they offer in terms of implementations.

The practical implementation was constructed to fetch posts from Twitter's Public Sample Stream API, and store these on a WISE database. This was done to prove the concept of storing data from social media on a database.

It was found that most social media have similar public APIs, and mostly deliver data in the form of JSON-strings. No real obstacles that could hinder integration of social media were found.

Keywords: social media, data mining, facebook, twitter, intagram, google+

Sammanfattning

SAAB Training & Simulation utvecklar simulatorer för militären och räddningstjänsten. För att förbättra bakgrundsinformationen i dessa finns ett önskemål om att hämta in information från sociala medier relaterat till vissa händelser, t.ex. terrordåd. Det första steget är att studera ett antal sociala medier och undersöka hur data hämtas samt huruvida informationen kan sparas i en databas. Företrädesvis skulle samtliga sociala medier integreras till en gemensam databas. Detta steg inkluderar också en undersökning av vilken typ av information som kan hittas på respektive socialt media. Det andra steget är att genomföra en praktisk integrering av ett socialt media, för att bevisa att konceptet fungerar.

De utvalda sociala medierna är Facebook, Twitter, Instagram och Google+.

Vissa legala aspekter studeras också, eftersom dessa har en påverkan när information som kan spåras till en specifik individ hanteras.

Resultatet består av en studie som inkluderar information om de utvalda sociala medierna, deras APIer, vilken typ av information som finns på respektive socialt media samt en genomgång av möjliga implementationer.

Den praktiska implementationen hämtar poster från Twitters Public Sample Stream API, och lagrar dessa i en WISE databas. Detta gjordes som sagt för att bevisa konceptet med att lagra data från sociala medier i en databas.

Resultatet av studien är att flera sociala medier har liknande publika APIer, och levererar företrädesvis data i form av JSON-strängar. Inga reella hinder för integration av sociala medier i stort hittades.

Nyckelord: social media, data mining, facebook, twitter, intagram, google+

Foreword

The aim of this report might have been determined right from the outset, but not the actual content and layout. This was determined only after a long discussion with SAAB Training & Simulation regarding the possibilities of data mining social media. Various limitations was agreed upon to keep the workload at a manageable level. Because the subject of data mining social media is truly massive. Every time the subject was discussed with someone new, more interesting ideas was put forward. At times it seemed like the only limiting factor when it came to possible application, was a person's imagination.

Throughout this report it will become clear that not everything is possible when dealing with social media, but most of the possibilities are still there. It only takes some imagination and ingenuity to realize them.

Acknowledgements

- Mikael Eriksson, and the rest of SAAB Training & Simulation Helsingborg, for all the support, advice and friendly working environment throughout the project.
- Christin Lindholm and Christian Nyberg for all the help and guidance in planning and delivering this report.

List of contents

1 Introduction

1.1 Background and context

SAAB Training & Simulation (T&S) is a subdivision of SAAB, a producer of products, services and solutions for military defense and civil security. I 2015 SAAB had over 14500 employees and sold products for over 27 billion SEK. SAAB T&S mainly produces simulators and training environments for training personnel in the military and emergency services. There is a desire from SAAB T&S to improve realism in background information used in simulations by incorporating information from social media, such as Twitter and Facebook.

Recent acts of terror in Europe and elsewhere have not only increased the pressure on intelligence agencies to improve their ability to foresee and prevent acts of terror, but also the pressure on emergency services to handle the extreme situations that arise from a successful act of terrorism. As was seen after the Paris attack on the 13th of November 2015, many that were in close proximity to the events shared their experience on social media. Journalists and citizens recorded what was happening around them as well as their own reactions to it. This information is stored on social media, and have been used by news agencies to retell the events to people all over the world. Through data mining and data analysis it can also be used to create more accurate training scenarios, aimed at improving response to similar events.

1.2 Aim

An application for social media data mining is not presently available to Saab T&S, and the goal of this thesis is to begin the work of remedying this. After discussions with SAAB T&S it was decided that a comprehensive pre-study was necessary, in which four social media were studied closely.

One implementation of retrieving data from a social media to a database was also desired, as a proof of concept. The exact specifications of this implementation depends on the result of the pre-study, but the social media that will be integrates should be one of the social medias what are a part of the pre-study.

1.3 Problem description

The study needs to answer the following questions:

1. What type of information can be found on social media?
2. How is information collected from different social media?
3. Can information from different social media be compiled in a shared database?
4. Construct a practical implementation that integrated a social media to a database.

These questions are the focus of the study. The first two points deal with the availability of information, which includes both what kind of information is available and how it is accessed. The third point deals with whether different social media can be integrated to one shared database. If this is possible it would mean that more information could be analyzed using the same tools, improving efficiency and making the gathered information more comprehensive. The fourth point, while not being a question, will serve as a proof of the concept of integrating social media to a database.

The social medias selected for the study are Facebook, Twitter, Instagram and Google+. These were selected due to their popularity.

1.4 Delimitations

The pre-study is limited to the social medias mentioned above, namely Facebook, Twitter, Instagram and Google+.

The practical integration covers the integration of one social media to a database, without any data analysis. This implementation should still be open to further development.

1.5 Report structure

This report will begin with an overview of the selected social media and their related APIs, followed by an analysis of social media data mining, results of the study, a description of the practical implementation, and finally a conclusion.

2 Method

2.1 Work process

The overall process of the project can be seen in Figure 1.

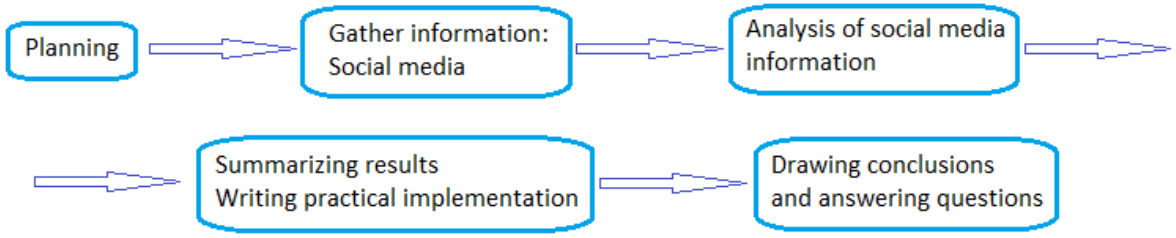

Figure 1. Project progress divided into its basic steps.

When starting the work on this report the first step was planning the scope. This was mostly done during the initial meeting with SAAB T&S, where it was decided that the report should cover four social media and answer the questions asked in section 1.3. The four selected social media was also decided at this meeting, along with the delimitations. It was also decided that most of the work would be done in the offices of SAAB T&S, which also made it easier to stay in contact with the supervisor at SAAB. The primary way of staying contact with the supervisor at Lund University was through email. Since the work load, the overall process and progress seemed clear and easy to overlook, it was later decided that no special project model was needed. A basic timetable was however made to keep track of deadlines. The report was to be written continuously during the project.

The second step was finding information about the selected social media, but also about social media in general. The official documentation was studied for information about respective social media, and general information was gathered from journals and organizations that was considered trustworthy, such as OECD. Focus was put on keeping information about APIs short and understandable, as the detailed information would be too much and affect readability. Further reading into the APIs details are better done in the actual documentation.

During the second step, it became clear that there were other considerations to make when dealing with data mining of social media, such as legal aspect and platform policies. These were not part of the original problem description but was added in the analysis phase, due to the effect they may have on actual implementations. In this step the studied social medias were analysed for similarities.

The fourth step was to summarize the results from the information that had be gathered so far, and constructing the practical implementation. After deciding what API, database and language to use, a preliminary UML diagram of the

implementation was drawn. Motivations for these decisions can be found in section 2.2. The code was written with the preliminary UML as a template, with the only exception being that one class was implemented as a static class. The corrected UML diagram can be seen in Figure 2, in section 5.1.

The last step was drawing conclusions from the results and answer the original questions. The fact that the information gathering was done with the main questions in mind was reflected in the results, making drawing conclusions easier. The questions were answered according to what had be found during the project.

2.2 Tools

LUBsearch[1] was used as the primary search tool for non-company related and more independent information sources. Company websites were used for information about respective company, and also for API documentation.

WISE was selected to be used as the database of the practical implementation, due to it being developed by SAAB T&S. This also narrowed down the selection regarding programming language. Included in the WISE SDK are two driver-wizards, one for C# and one for C++. C# was selected due to personal preference, and Visual Studio was used as the development tool.

[1] http://lubsearch.lub.lu.se/

2.3 Source criticism

When fetching information from impartial sources, such as a company's own website, some information may be inflated or angled to the benefit of the company. For example, information about monthly uses may not be accurate. This includes the references to Facebook Newsroom, About us: Instagram, and About Twitter.

Other information, such as technical information and documentation can be trusted to be correct, otherwise it would have a negative effect on their reliability trustworthiness as a business partner. In this report these are the API documentation pages: Facebook for developers, Google developers, Instagram Developer Documentation and Twitter for developers.

News articles and conference documents have not been peer reviewed, and may therefore have a certain amount of errors. They are however meant for a larger crowd, which should dissuade authors from grave and obvious errors. On this basis they are trusted in this report, since no crucial detailed information will be fetched from these sources. These sources are: (Barrie, 2015), (Dewan & Kumaraguru, 2014), (Hu, et al., 2014), (Kassim, 2012) and (Lee, et al., 2013).

Peer reviewed journals can be trusted to be correct, due to the nature of peer review. In this report these references are: (Humphreys, et al., 2013), (Kaplan & Haenlein, 2010) and (Yoon, et al., 2013).

One report from OECD is used as a reference for definitions of User Generated Content (UGC). OECD is a well-respected international organization and have no self-interest in this subject. They are therefore considered credible.

In this report there is also a reference to a specific Swedish law, and information about this law is collected directly from the Swedish Parliament (Sveriges Riksdag) website. This website is considered as trusted in this report, as it is the official online publication of the actual legal text.

3 Social media

The history of social media can be said to be as old as Internet itself since the Internet was setup as a bulletin board system, where users could exchange information and data. This is not unlike some services available today, under the label of social media.

But what defines social media? An actual definition is hard to find, but User Created Content (UCC) or User Generated Content (UGC) are core features of it (Kaplan & Haenlein, 2010). The Organization for Economic Co-operation and Development (OECD) defines UCC by three requirements (OECD, 2007):

1. Publication Requirement – content must be publicly accessible, or at least accessible to a select group of individuals
2. Creative Effort – a certain amount of creative effort was put into creating the content or adapting existing work in order to construct a new one.
3. Creation outside of Professional routines and practices – the content is not created within an institutional or market context.

These requirements will disqualify services such as mail and bilateral messaging services (1), copying content and republishing others' content (2) and content made outside the scope of the average user (3). The last point is becoming increasingly diluted as users are aiming for monetization of their content (OECD, 2007). Organizations are also increasing their presence on social media as part of marketing strategies and other efforts.

Social media can roughly be divided into six categories: collaborative projects, blogs, content communities, social networking sites, virtual game worlds and virtual social worlds (Kaplan & Haenlein, 2010). Implementations of these have their own unique focuses and possibilities. This report will however not cover examples of all categories, nor possibilities of each category. The basis on which social medias were selected for this report was popularity, not category. It is however important to remember what type of social media is discussed in the coming sections, and how it can be compared to other types. Facebook and Google+ are social networks, Twitter is a type of blog and Instagram is a content community.

3.1 Facebook

Facebook is a social networking site launched by Mark Zuckerberg in February 2004, and since then the service has grown continuously. Today the service has more than 1.5 billion active monthly users (Facebook, 2016).

Registered users can among other things add personal information to their profile, add other users as “Friends”, post status updates, join group of other users as well as create and join “Events”. Users can also post pictures, videos and links to websites, both on their own page and on pages belonging to other users, groups and events. Posts made on a page belonging to a user, group or event are presented as a timeline, where other status updates are also presented. There are also the possibility of showing your reaction to a post using Reactions, which are a set of icons and smileys that represents different emotions. This is an expanded version of the previous “Like” button.

There are also the possibility to search for users, groups and also hashtags.

Companies and organizations can create “Pages” on Facebook. These have the same basic functionalities as users and their profile pages.

3.1.1 API

The API for Facebook is extensive, providing functionalities and assistance for creating new applications and adding Facebook functionalities to existing products (Facebook, 2016). Facebook can be used as a login tool for other websites and applications, authenticating users using their Facebook profiles. Other functions that can be added to existing products are the social interactions, such as the buttons to “Like”, “Share” and “Send” buttons. These will make the content appear on the users Facebook page, in the same way as if the user had used the corresponding buttons on Facebook itself. The API also gives support for monitoring and analyzing details about how users use an application and how to monetize an application.

3.1.1.1 Graph API

The primary way of getting data to or from the Facebook platform is called Graph API, which is a low-level HTTP-based API. It allows among other things data queries, posting of new content, uploading of photos and management of existing objects on Facebook.

The basic architecture is that of a graph with nodes, edges and fields. The nodes are the things on Facebook, such as users, photos, pages or comments. Edges are the connections between the nodes, such as a photo’s comments. Fields are the information about the nodes, such as name of a group or birthday of a user.

3.1.1.2 Facebook Login and Access Tokens

All content on Facebook is generally not freely available. Privacy and containment of information are core features of the Facebook API. To gain access to Facebook content, users can be asked to login with their Facebook account. Facebook Login is a tool in the API developed to handle this situation. When logging in the user gives permissions to the application to access, change and/or post information. The user will effectively be logged onto Facebook, using another application. There are more than 30 different permissions that can be granted, leaving a lot of options for customizing application specific permissions. There is however a limitation. If an application requests permissions other than “public_profile”, “email” or “user_friends”, described in Table 1, the application must be submitted to Facebook for review.

Facebook will then check whether the application fulfills their requirements. One thing that is tested is whether the application is requesting more permissions than it needs or not.

Table 1. Facebook permissions that does not require review by Facebook.

Permission	**Description**
public_profile	Provides access to a subset of information from the user’s public profile, including among other thing user id, full name, first name, last name, age and gender.
email	Provides access to the users primary email address.
user_friends	Provides access to a list of the user’s friends that also uses the current application.

The Facebook Login generates Access Tokens (AT). These are text strings that are generated form a Facebook user’s login. The AT is a text string that contains information about what user it belongs to, what permissions have been granted and how long the AT is valid. The AT is not permanent, but gives a temporary access to the Facebook API and are used to make requests to Facebook systems.

There are four different ATs available:

- User Access Tokens: Obtained from a user login dialog, and is the most commonly used. A User AT is required every time an application wants to read, write or modify a user's data.
- App Access Tokens: Generated using a pre-agreed secret that is shared by the application and Facebook. App ATs are used to read and modify app-wide settings.
- Page Access Tokens: Similar to User AT, but grants permission to read, write and modify data belonging to a Facebook Page. Page ATs are obtained from by first obtaining a User AT from an administrator of the Page and ask for the "manage_pages" permission. When this has been done, the Page AT can be retrieved using the Graph API.
- Client Token: A rarely used token that can be used to identify applications.

3.1.1.3 Public Feed API

The Facebook Public Feed API is a special API only made available to a select few selected media broadcasters. Once the connection is established it provides users with a feed of public posts as they are as they are posted. Therefore it does not rely on HTTP requests like the Graph API.

3.1.1.4 Pages API

As the name suggests, the Pages API is specially tailored for managing Facebook Pages. Available functions include posting as the Facebook Page, create new tabs on the Facebook Page and Facebook Page Insights.

Apart from Page Insight, all other function are also found in the Graph API and works in a similar manner with minor changes to the syntax.

3.1.1.5 Other APIs

Facebook have many more APIs and tools available to developers, but they will not be covered in this report.

A few noteworthy examples are the Marketing AP and Facebook Analytics for Apps. The former is made for automation of advertisement on the Facebook platform, and the latter is made for measuring how customers are using the developed application.

3.1.2 Content type

Facebook is very rich in information about individuals as well as events. The amount of personal information is only limited by what the user wants to share, and may include, but not limited to, name, age, gender and employment history. Through interactions with Facebook Pages, such as likes and shares,

information about views on companies and politics can be ascertained. User's comments and uploaded videos and pictures can also contain valuable information.

By looking at users' attendance to Events information such as activities and interests can be found, as well as information about real life trends.

The topic of posted material can vary greatly, and may be anything from fiction to news of latest scientific discovery. The only limit here is the imagination and agendas of the users.

In addition to all this is the information in the metadata, with information such as timestamps and in some cases geographical tags.

3.1.3 Possible applications and accessibility

The most limiting factor is the matter of access to information being tied to what a user can see. The more accessible APIs require an Access Token to be added to requests, which are tied to a specific user and determines what content is accessible.

Another limiting factor is the matter of what fields are available without having to send in the application to Facebook for review. As displayed in Table 1, only basic information about the user, his/her email and friends are available without review.

When looking at possible applications using the accessible Facebook APIs, it is important to keep the point-of-view situation imposed by the Access Tokens in mind. Most applications will be centered on the profile behind that Access Token and things related to that profile. If the profile is administrating a company's Facebook Page, there are a few additional options. One is an application that collects information from the posts that have been posted on the Facebook Page by other users. Basic data collection from the profiles behind the posts are also possible.

A more refined implementation could include analysis the posts and check if they contain any foul language and, if so, remove them from the Facebook Page.

3.2 Twitter

Twitter was founded in March 2006 and launched in July the same year. The service allows registered users to send and read 140-character messages, called "tweets". Due to the short length of the messages and the fact that they do not necessarily need to be addressed to a specific receiver, it is often referred to as a microblog. It is possible to answer to, repost tweets ("retweeting") and like tweets. Other users can be tagged in tweets using their account name and topics can be tagged using #, e.g. #Paris.

Today the service has 320 million monthly users and 1 billion unique visitors every month (Twitter, 2016), among which several prominent companies, heads of states and other celebrities can be found.

3.2.1 API

Twitter offers a lot of developer tools to help build and customize new applications (Twitter, 2016). Among the tools available are Fabric, which is a platform used to develop mobile applications, and MoPub, which is an advertisement tool.

There are also a few options to customizing websites to add twitter functionalities to them. Most notably, it is possible to add buttons representing Tweet and Follow to a website. Pressing a Tweet button will share the content in question on Twitter, and Follow will add the producer of the content to the user's list of followed Twitter accounts. It is also possible to embed single tweets or whole timelines on websites, and summarize webpages in "Cards" every time a user links to the webpage in a tweet. These Cards maybe summaries of the content, large images, mobile app download links or multimedia content such as audio/video.

There are several different parts of the Twitter API, all with the capability to interact with the Twitter API in some manner. Broadly they can be divided into REST APIs and Streaming APIs.

3.2.1.1 REST APIs

Most calls made by the public REST APIs have a Rate Limit, specifying how many calls may be made during a time frame. The length of a time frame is 15 minutes.

3.2.1.1.1 Public REST API

The basic Public API is a REST interface that returns JSON responses. Twitter users and applications are identified using OAuth, an authorization framework. When identifying using a user account, application-user authentication, requests to the Twitter platform are signed with the applications identity and the user's access token. The alternative, application-only authentication, signs requests on behalf of the application only. This form of authentication however is not

supported by all API methods, and for example post-requests will not be allowed nor will access to Streaming endpoints be accepted.

The Search API is a part of the REST API and allows queries against popular tweets published in the past 7 days. The focus of this API is not to supply a comprehensive set of all tweets in the time period, but rather a more relevant and up-to-date set.

3.2.1.1.2 Collections API

Collections are editable groups of tweets, which can be hand-selected by a user or managed through the Collections API. These collections are public and all have their own page on twitter.com, each with a public URL, making it easier to share with others. Collections can also be embedded on websites and mobile applications.

3.2.1.1.3 TON API

Twitter Object Nest (TON) API allows for uploading of media and other assets to the Twitter platform, using either a resumable or a non-resumable process. More operations are supported according to the documentation, but not specified nor described. To gain access to these, implementers are asked to consult their Platform Relations representative or their Ads API Partner Engineer.

3.2.1.1.4 Curator API

The Curator Broadcaster API targets broadcasters, and supplying them with their Curator-created streams for digital displays, such as on-air graphics. It is a private API that requires special permission to access, and is only made available to TV broadcasters.

3.2.1.2 Streaming APIs

The streaming APIs provide access to Twitter's stream of Tweet data, and will result in pushed messages for each published Tweet and other events. There are a few different alternatives for streaming Twitter APIs.

The Public Stream alternative offers access to the public data flowing through Twitter. These are suitable for following specific users and data mining of the Twitter stream in real-time.

The User Streams alternative provides a stream of data that is specific to a specific and authenticated user. These are restricted in how many simultaneous User Streams connections a Twitter account can have. Should these limits be exceeded, an HTTP 420 error code will be returned, signaling that the account have been logging in too often.

Site Stream is similar to User Streams, but allows for multiple users to stream at the same time. An additional wrapper is added to indicate which user the

message is intended to. At the time of writing the Site Stream is in closed beta and may therefore change before being released for public use.

3.2.1.3 Enterprise APIs

Twitter's enterprise API platform is called Gnip, and offers some additional functionalities. Like the public API there are a real-time/streaming API and a search API designed to search through past Tweets. There is also what is called Insights API, which focuses on information about audiences and their perception about specific Twitter content.

3.2.1.3.1 Premium real-time APIs

The difference is that there are more possibilities and functionalities are more comprehensive. For example, the real-time API offers functionalities such as decahose, which delivers 1/10 of all Tweets. The object of this function is to provide a statistically representative stream of Tweets that is more manageable than the whole Twitter stream. The main real-time API, called PowerTrack, also have some additional functionalities compared to the Public Stream API. Redundant stream for important applications, data recovery to prevent data loss during disconnects and language detection are just a few of the additional functionalities. Queries are filtered using preset specifying what manner of content is desirable.

Another part of the real-time API is the Data Collector, which enables implementers to connect to up to six public APIs, from various social networks. The Data Collector also offers exclusion of duplicates and normalization of formats, allowing for standardization of collected data regardless of its source.

3.2.1.3.2 Premium historical APIs

The Historical APIs provides the functionalities needed for finding already published data. The 30-Day Search API provides just what it says, access to the last 30 days of public Twitter data, along with additional functionalities such as filter and format normalization to name a few.

Historical PowerTrack expands upon this and offers access to the full archive of public Twitter data, thus making it possible to find every single public Tweet that have ever been made. An additional product, Full Archive Search, also offers this functionality but operates using a RESTful interface.

3.2.2 Content type

The contents of tweets vary greatly and cover personal as well as professional topics (Humphreys, et al., 2013). The use of hashtags provides an insight into the topics and relevance of the tweet. Further analysis is needed to determine the exact content of the tweet, such as keyword oriented analysis (Yoon, et al., 2013). Such analysis can give information not only about the topic, but also regarding the sentiment of the tweet. Real world events such as elections,

disasters and deaths of famous individuals are also common themes in tweets (Niles, 2009) (Dewan & Kumaraguru, 2014). Twitter was also used extensively during the Arab Spring revolts in the Middle East to coordinate protests (Kassim, 2012), giving plenty of information about events that have yet to pass.

In addition, all tweets have information such as retweets and likes. There are also related metadata such as timestamps and in some rare cases geographical locations. A study conducted 2012 found that 99.1% of tweets did not contain a geotag (Lee, et al., 2013).

3.2.3 Possible applications and accessibility

The access to the streaming APIs opens a lot of possibilities for real-time information. Shifting opinions running up to elections could be covered and developments of crises and disasters could be monitored. With the User Stream API it is also possible to monitor an owned Twitter account to automatically gather information or provide automated responses.

When searching among previously posted tweets there is a limitation in the public REST API, namely the fact that only tweets from the past seven days are available. This can be circumvented by applying for access to the Gnip APIs, thus opening for long term searches and comparisons.

3.3 Instagram

Instagram was launched in 2010 by Kevin Systrom and Mike Krieger, as a content community, and is focused around sharing photos and videos up to 15 seconds long over mobile devices. There is also a website, but currently it is limited to viewing existing content and cannot be used for uploading.

It is possible to follow other users and therefore receive notifications when they post new content. One thing to note is that following a user does not automatically mean that they will follow you. Every follow-request is strictly one way. Users can also use the private setting to hide their content from all users that are not one of their followers. Any new followers in this case must ask the private user for permission before they are confirmed as a follower of the private user.

Content can be tagged with hashtags, which can be used for categorizing new content and searching after other user's content. Another core feature is filters and (visual) settings, which can be used for modifying content before upload. In 2015 functionalities for sending photos and videos directly to other users were added.

At the time of writing Instagram have about 300 million users and 60 million photos are added every day (Instagram, 2016).

3.3.1 API

Every application developed for the Instagram platform is required to be submitted to Instagram for review, and specify what the application requests access to (Instagram, 2016). These permissions will then be tied into the client id that will be associated with the application. The client id must be provided in the query sent to the Instagram platform when a user logs onto the application in order to get an access token relating to that user.

The API itself is built on the REST architecture, with seven endpoints available through https for queries. These endpoints are Users, Relationships, Media, Comments, Likes, Tags and Locations, and all contain operations relating to their individual context.

The number of requests that can be sent are limited per access token through rate limits. There are global rate limits for API calls to any endpoint, and some endpoints also have individual rate limits. These rate limits are controlled on a one hour sliding window.

3.3.2 Content type

Instagram is as previously stated focused around publishing photos and videos. While descriptions and comments of media provides a platform for user-to-user communication, it is not a central aspect as in Twitter for example.

Research have shown that photos uploaded by ordinary users (not companies, organizations and brands) can mostly be categorized as selfies, friends, activities, gadgets, captioned photo, food, fashion or pet (Hu, et al., 2014).

Companies and organizations may have their own accounts and profiles, and use these for marketing or other social media strategies.

The main information in Instagram can be said to be stored in the photos and videos, such as the object depicted and the context surrounding the object. Image analysis (pref. automated) is required extract quantitative information that can be stored in for example a database.

3.3.3 Possible applications and accessibility

There are some limitations in the Platform Policy for the Instagram API that effect information gathering and storage. The summary of this policy highlights a few of there, and can be found in Table 2.

Table 2. Instagram Platform Policy summary (Instagram, 2016).

1. Instagram users own their media. It's your responsibility to make sure that you respect that right.
2. You cannot use "insta", "gram" or "Instagram" in your company or product name.
3. You cannot replicate the core user experience of the Instagram apps or web site. For example, do not build a media viewer.
4. You cannot use the API Platform to crawl or store users' media without their express consent.
5. Do not abuse the API Platform, automate requests, or encourage unauthentic behavior. This will get your access turned off.

Points 1, 4 and 5 especially limits the possibilities to automate information gathering on a larger scale. Point 1 limits the use of the media, and specifically the possibility to store it in a separate database and Point 4 limits the possibilities of searching through content on Instagram. Point 5 specifically makes automation of requests off limits. Automatic requests and categorization of responses are the ideal way of handling information gathering on a larger scale.

Because of the Platform Policy, possible applications are somewhat limited. Banning automated requests and crawling users' media without consent prevents automatic gathering of information. Manual operations are then the only option left, and can then include information gathering based on published user information and descriptions and comments on published media.

3.4 Google+

Google+ is a social networking service based around interests, and was released in 2011 by Google Inc.

As with most social media each user have a customizable profile page. Friends are organized in circles, and for each posted piece of content it is possible to set who is allowed to see it, either by circle or by individual user. Users can use the equivalent of liking a post, called "+1", if the content is to their liking. Content may be pure text posts, photos, videos or links, and hashtags may be used to help others find the content.

A central part of the content on Google+ are Collections. These are just as the name suggests collections of posted material that can be edited and modified by the user. These Collections have settings regarding who to share the Collection with, just like all other content.

Companies and organizations can also create and maintain Pages, which act like a Profile.

User can also create Communities, which works similar to the Groups on Facebook, where users can share content and interact with one another.

There is a chat function called Hangouts, which also have a separate downloadable app for mobile devices. There is also the function called "What's Hot", which gives presents trending posts and posts that are recommended by Google+.

The current amount of active users are somewhat unclear. This is because everyone with a google account, including all Gmail users, also automatically gets a Google+ profile. There are no official statistic regarding the activity of the users at the time of writing. Some sources claim that there are about 2 billion Google+ accounts, but only 9% have any publicly-posted content (Barrie, 2015).

3.4.1 API

The Google API in its entirety is extensive and include products and services for all Google products (Google, 2016). This section will therefore focus only on the parts directly relating to the Google+ Platform, namely plugins and the REST API. There are another API, Google+ Pages API, which is limited to companies that are registered partners with Google and therefore allowed access to the API.

3.4.1.1 Web API and plugins

Plugins for websites are added primarily using the Web API, which is a JavaScript API. These plugins includes buttons for "+1", Follow and Share. Other plugins include the Badge, which is a link to the user's profile, embedded Google+ posts on websites and finally snippets, which are a customizable settings for what shared content should look like.

3.4.1.2 REST API

The main programming interface to the Google+ API is the REST API, which as the name suggests follows a RESTful design pattern. At the time of writing the Google+ API only provides read-only access. All calls require either an OAuth token or an API key. OAuth token is acquired when a user logs in and an API key is given an application when it is registered with Google. API keys does not give access to any account information, and are intended for accessing data that have been published as public and data owned by a Google service.

3.4.2 Content type

Content on Google+ is broadly similar to that of Facebook, which is natural since they offer similar services.

Personal information about the user, its views and event participations tells a lot about the user, its interests and habits. Company maintained Pages hold similar information relating to the company, and interactions from other users also gives an indication of public opinion of said company.

There are no specific limits as to topics of Collections, Pages and Posts, other than legal ones.

Meta data such as timestamps and in some cases geographical locations are also available from all published posts.

3.4.3 Possible applications and accessibility

Availability of information is similar to that of Facebook as well. Requests are sent to the Google Platform on behalf of a registered and authenticated user. The most notable difference in data accessibility is the ability to share content to specific Circles, which makes the access to other users' posts a little more restrictive.

The most significant difference however lies in the access to the API itself. At the time of writing, the Google+ only provides read-only access. Applications can therefore not affect what is on Google+, only read what is published.

The possible applications tied to the Google+ are for obvious reasons similar to those of Facebook mentioned in section 3.1.3, with some modifications. The central part of all requests sent to the Google platform is the access token, and therefore sets limit for what information can be accessed. When using OAuth tokens, things are just like with Facebook, the user's permissions are the limits. The foul language removal previously mentioned in section 3.1.3, cannot be implemented under the current read-only access to the API. An implementation that detects foul language can however be constructed. Operators of the Page can then be notified, and the post can then be removed manually using the ordinary Google+ website. Other applications that analyzes texts in posts accessible to the user in question are also possible.

When using API keys there are more options regarding collection and analysis of public data. Data with other publicity settings will not be accessible with this access method. It does however allow, for example, gathering of public opinion expressed in posts made by users.

4 Analysis of social media data mining

A few points have emerged in the previous chapters that have a great impact on the effectiveness of data mining social media, and there are of course more aspects to consider. Here we will look more closely into three such aspects parts; what is available, what is allowed under the law and what is allowed by the policies of the social media. A short summary of possible implementations can also be found in chapter 4.4.

4.1 Availability

Among the most notable things to consider is the availability. With the exception of Instagram, all studied social media have premium or enterprise API's. These are not free and require signed deals or partnerships before being granted access. One such example is the Public Feed API in the Facebook Platform, which grants access to public posts without the limitations of an access token tied to a specific user. The success of data mining depends on the goal of the mining and what data is available. In some cases it might be crucial to have the extended access provided by a partnership deal, in others it might not provide any additional value.

Some of the APIs also have limits on what functions can be used without registering the application in development with the social media in question. As mentioned in section 3.3.1, Instagram does not allow any access to the API without registration and submission the application under development to Instagram. Another implementation of this sort of limitation can be found in section 3.1.1.2, which covers the fact that only a few fields in the Facebook API are publically available without registration.

4.2 Legal aspects

When dealing with information posted on social media, developers can come into contact with detailed information about users. There are some legal aspects to consider when dealing with this kind of information, which may have an impact on how social medias are integrated and how data is handled.

In Sweden this is regulated by Personuppgiftslagen 1998:204 (Sveriges Riksdag, 2016). This law defines personal information as any information that can be related directly or indirectly to a physical person. It also covers when it is allowed to collect information and how it is to be handled. To summarize its content, all personal information that is collected automatically must be reported to a supervising authority. Storage of especially sensitive information such as religion, ethnicity, political view, etc., is generally forbidden except when the individual have given explicit permission or clearly made the information public themselves.

First thing to note here is that this law concerns those that store the information, and not the developers of the tools used. As such this law has little effect on the development of a system that can be used to store personal information, other than inform potential customers.

Secondly, if no information of relevance is stored or can be linked to a physical person, this law does not apply. Information can still be gathered and used, but not stored. Alternatively, information can be stored but without information that can help to identify the person the information relates to. Additionally, even if there is support in the system for information storage which is not in use, this law does not apply.

Finally, if an application focuses on information made public by the user of a social media, the ban on sensitive personal information is lifted. This alone however does not lift the requirement to report to the supervising authority. Any application that is to store personal information will have to carefully consider the implications of the law and how it affects their specific application.

4.3 Platform policies and user agreements

Finally there are some the matter of what the social media themselves allow when it comes to data handling. All studied social media have a user agreement or platform policy of some sort that any user must comply with. Failure to do so generally leads to one type of ban or another.

As an example, we can look at the Instagram platform policy, discussed in section 3.3.3. Without going into the details of every paragraph it is clear that there are some obstacles when it comes to usage of the API and the data that is obtainable. The bans on crawling users' media and automating requests are especially problematic if the goal is to build an integrated database based on content.

4.4 Summary of possible implementations

Given all that have been discussed this far the possible applications relating to social media data mining have a few similarities, at least for the public APIs. Most notably is the lack of actual public searchable data. Most of the studied API only provide searchable data from the viewpoint of an authenticated user of the social media. The few alternatives that do offer data that is not bound to an authenticated used do not include a search function. One example of this is Twitter's Sample Stream API. Public data from all over the world is being accessed, but nothing in the API allows for any type of selections.

Limitations on how far back in time searches can go is also common.

Large scale data mining of public data therefore becomes difficult, since it will always be point of view based and limited in amount of available data.

Applications using these public APIs will therefore mostly be orientated around data mining data accessible to an authenticated user. Such as analyzing posts and reacting to them depending on the contents, like in the foul language remover mentioned in chapter 3.1.3. The implementation mentioned in the introduction, collecting posts relating to a specific event such as a terrorist attack, will face some difficulties. Not only will there be a problem with accessing the relevant data relating of what data the authenticated user can access, but also with how long it is accessible.

When looking at the non-public APIs, aiming at enterprises and broadcasters, there are many more options. More information is accessible and for a longer time. Wider searches that are not bound to any single user can be made, and therefore increase the effectiveness of the data mining. In the example dealing with data relating to terrorist attacks, it will be possible to access data from every terror attack there have been in recent years. Data from reactions worldwide can also be accessed and used. The limiting factor with these APIs is mostly related to what business deals can be made and what resources are available to the developers.

5 Results of social media integration

As previously stated the social media that are a part of this study are of different categories, as stated in section 3, and they were also developed separately. Despite that, there are similarities that are worth noting.

Firstly of all, it is worth noting that all the studied media have a version of a REST API, through which all main parts of the media is accessible. Some, like Facebook, allow manipulation of the content on the media, while others like Google+ is locked to read-only.

Some media, such as Twitter, also offer streaming APIs which works be establishing a connection to an endpoint. After successful connection, data will continuously be streamed to the application until it is disconnected.

Secondly, all of the studied media will respond to a successful request, at least to the REST APIs, by returning a JSON object. This means that there are a good basis for integrating different media to the same database.

Third thing to note is the amount of information a single post can supply with a simple request. For example, a successful HTTP request for a Facebook post will return a JSON object with 43 data members. Some of these, such as author of the post, will in turn be representations of other object with additional fields. References to other social media APIs tells a similar story. Naturally, all fields will not be useful for every application. It is therefore advisable to only integrate the fields that are interesting in the context of the application in question, and leave others until there is a use for them. By integrating fields from a media on a need only basis, it is also less likely that sensitive information about users will be stored unnecessarily. This will solve a few potential legal problems for users of developed applications.

5.1 Practical implementation

As a proof of the concept of storing information from social media, one functioning implementation was made. The chosen API was Twitter's Sample Stream API which supplies a small representative portion of tweets as they are published, without any filtering. The reasoning behind this choice was that it is one of few APIs offering streams, which means that when fetching posts with this API the posts will never be the same as previous runs. This ensures that posts are of a wide variety of posts and languages. This API does not implement any search functions, making it all but impossible to find any specific information among the tweets. But what it does supply is a steady stream of posts, which is needed to prove that information from social media can be stored in a database. With this API there will be no waiting for relevant information to be published, eliminating the problem of not having accessible data to store in the database.

WISE was chosen as the database, since it is made and maintained by SAAB T&S. It is an integration platform with a set of databases at the center. Each application using a database have their own database which is in turn connected to a backbone database. This backbone database is the same for all applications connected to the database. This modular structure could be useful for further development of social media integration. When adding more social medias to the implementations, the option will be there to add them as a separate application in the database. Other applications with different functions could also be added, such as data analysis software.

An UML overview of the constructed implementation can be seen in Figure 2 below.

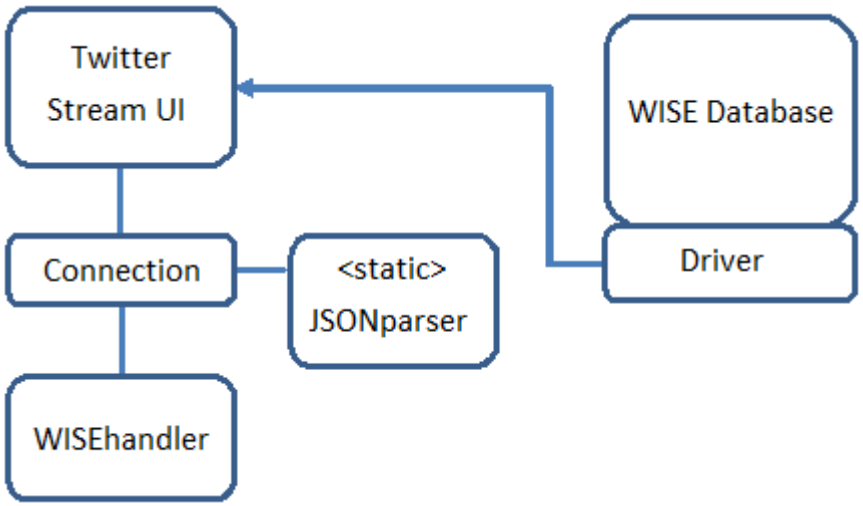

Figure 2.UML overview of the practical implementation.

When starting the database a separate UI will be started from where the user can enter how many posts are to be read from the streaming endpoint. Without a limit on the number of posts to add to the database, the stream will in a short time fill the database with unnecessarily large amounts of entries.

The UI uses the Connection class to make HTTP requests to the Twitter platform. The Connection class will connect to the platform, and maintain the stream until the number of requested posts have been integrated into the database. The stream produces strings of JSON data, which is converted to .NET objects using the static JSONparser-class and then sent to the database using an object of WISEhandler-class.

From all the fields in the tweet JSON-object only a select few are integrated. These fields can be found in Table 3, on the following page.

Table 3. Tweet fields integrated into the WISE database.

Field name	**Description**
id	ID number identifying the tweet
created_at	Time of creation
text	The text in the tweet
retweet_count	Number of times the tweet have been retweeted
favorite_count	Number of times the tweet have been marked as favorite by users
user: id	ID number of the creating profile

The user is a field containing all information relating to the profile that created the tweet, and from these fields the id number is saved to the database.

This small implementation shows that it is possible to integrate a social media to a database, and only integrate selected parts for ease of computation.

6 Conclusion

With the results in mind the questions asked in section 1.3 can now be answered:

1. What type of information can be found on social media?

As described in the subchapters of each social media under section 3, social media are very diverse in content. The actual content is limited only to what individual users are willing to share on social media. User information, opinions and habits, along with real life experiences and events are just a few topics found on social media.

2. How is information collected from different social media?

As mentioned in section 5, the most common way of accessing information in the studied social media are through REST APIs and HTTP requests. Some premium APIs uses tools developed by the company in question. There are also a few examples of streaming APIs.

3. Can information from different social media be compiled in a shared database?

Considering that all but a few of the studied APIs returns information as JSON object, it is very much possible to store information in a shared database. The practical implementation described in section 5.1 starts by turning JSON-strings into .Net object, from which data can easily modified fit most databases.

4. Construct a practical implementation that integrated a social media to a database.

This point was concluded as described in section 5.1, and was successful in retrieving posts from Twitter's Sample Stream API and storing them in a WISE database.

The overall conclusion is that it is possible to integrate different social media into a shared database. This conclusion is based on the similarities how data is requested and the format of received data. But there are some considerations to make in terms of what information is needed and whether or not a premium API is required. There are also a number of things to consider regarding what information may be stored and how, which is regulated by both legislation and platform policies. Some platform policies may also exclude certain applications.

In the context of collecting posts from social media that relates to acts of terrorism, a premium API may be required to access information relating to older events. Should a there be lacking data due to missing geotags in posts, one workaround could involve searching for users living in the relevant area and accessing their posts and tweets from the time period in question. Other than that, no serious problems have been found.

7 References

Barrie, J., 2015. *Buisness Insider UK.* [Online] Available at: http://uk.businessinsider.com/google-active-users-2015-1 [Accessed 1 March 2016].

Dewan, P. & Kumaraguru, P., 2014. *It Doesn't Break Just on Twitter. Characterizing Facebook content During Real World Events,* s.l.: http://arxiv.org/abs/1405.4820v1.

Facebook, 2016. [Online] Available at: http://newsroom.fb.com/company-info/ [Accessed 23 February 2016].

Facebook, 2016. *Facebook for developers.* [Online] Available at: https://developers.facebook.com/ [Accessed 16 May 2016].

Google, 2016. *Google developers.* [Online] Available at: https://developers.google.com/+/ [Accessed 16 May 2016].

Humphreys, L., Gill, P., Krishnamurthy, B. & Newbury1, E., 2013. Historicizing New Media: A Content Analysis. *Journal of Communication,* Volume 63, pp. 413-431.

Hu, Y., Manikonda, L. & Kambhampati, S., 2014. *What We Instagram: First Analysis if Instagram Photo Content and User Types.* s.l., s.n.

Instagram, 2016. *About Us: Instagram.* [Online] Available at: https://www.instagram.com/about/us/ [Accessed 23 February 2016].

Instagram, 2016. *Instagram Developer Documentation.* [Online] Available at: https://www.instagram.com/developer/ [Accessed 25 February 2016].

Kaplan, A. M. & Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons,* Volume 53, pp. 59-68.

Kassim, S., 2012. *Twitter Revolution: How the Arab Spring Was Helped By Social Media.* [Online] Available at: http://mic.com/articles/10642/twitter-revolution-how-the-arab-spring-was-helped-by-social-media#.i7rM34zfC [Accessed 23 February 2016].

Lee, K., Raghu, G. & Mudhakar, S., 2013. *Spatio-Temporal Provenance: Identifying Location Information from Unstructured Text.* San Diego, UCDavis.

Niles, R., 2009. Michael Jackson’s death and its lessons for online journalists covering breaking news.. *Online Journalism Review (http://www.ojr.org)*, 25 June.

OECD, 2007. *Participative Web: User-Created Content*, s.l.: Organisation for Economic Co-operation and Development.

Sveriges Riksdag, 2016. *Svensk författningssamling*. [Online] Available at: https://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/Personuppgiftslag-1998204_sfs-1998-204/
[Accessed 7 March 2016].

Twitter, 2016. [Online] Available at: https://about.twitter.com/company
[Accessed 23 February 2016].

Twitter, 2016. *Twitter for developers.* [Online] Available at: https://dev.twitter.com/
[Accessed 16 May 2016].

Yoon, S., Elhadad, N. & Suzanne Bakken, 2013. A Practical Approach for Content Mining of Tweets. *American Journal of Preventive Medicine*, 45(1), pp. 122-129.