Designing and developing an application for online hate speech detection on social media

Ludwig Hedlund and Simon Åberg

DIVISION OF PRODUCT DEVELOPMENT | DEPARTMENT OF DESIGN SCIENCES
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY

2021

MASTER THESIS

CIVIL
RIGHTS
DEFENDERS

Designing and developing an application for online hate speech detection on social media

Ludwig Hedlund and Simon Åberg

LUND
UNIVERSITY

Designing and developing an application for online hate speech detection on social media

Published by
Department of Design Sciences
Faculty of Engineering LTH, Lund University
P.O. Box 118, SE-221 00 Lund, Sweden

Subject: Innovation Engineering (INTM01)
Division: Product Development
Supervisor: Lars Bengtsson
Co-supervisor: Erik Helldén, Civil Rights Defenders
Examiner: Glenn Johansson

Abstract

Social media platforms have a significant impact on the world. Through connecting people over the globe, social media has resulted in multiple positive effects. However, the impact is not only beneficial. Social media platforms have also caused polarization and criminal uprisings, leading to events such as the Rohingya genocide in Burma. By tracking the discourse on social media channels, it is possible to get insights into societal movements and warnings about what is to come. However, the massive number of conversations that occur on social media channels can be hard to analyze. Especially for NGOs, which commonly has strained resources and budgets.

This study aims to research how tools can be developed to help NGOs and similar organizations in their work towards finding threats from online activity, as well as analyzing the general discourse in a subject. The contribution of this research is two-fold: (1) The paper investigates how the NGO Civil Rights Defenders is currently working with analytics of social media and how their work can be helped. This is done by interviews with Civil Rights Defenders, and the results are presented in two user scenarios and use cases. (2) A tool is developed for helping Civil Rights Defenders improve their work within social media analysis. This is done through a literature review of adjacent research and the current state of technology, as well as using insights from interviews. The collected information is then used to build a fully working web application, targeted mainly towards NGOs and their work in fighting hate speech online.

The developed tool was tested, and the results were satisfactory. The testers were able to use the tool without much, if any, confusion, and the tool provided help in their workflow.

Keywords: AI, Big Data, Hate Speech, NGO, Machine Learning, Social Media, Twitter

Sammanfattning

Sociala medieplattformar har stor inverkan på världen. Genom att ansluta människor över hela världen, har sociala medier resulterat i flera positiva effekter. Påverkan är dock inte enbart bra. Sociala medieplattformar har även orsakat polarisering och kriminella uppror, exempelvis folkmordet mot Rohingyer i Burma. Genom att spåra diskursen på sociala mediekanaler är det möjligt att få inblick i samhällsrörelser och varningar om vad som kommer. Den enorma mängden konversationer som sker på sociala mediekanaler kan dock vara svår att analysera. Speciellt för icke-statliga organisationer som ofta har lite resurser och små budgetar.

Denna studie syftar till att undersöka hur verktyg kan utvecklas för att hjälpa icke-statliga organisationer och liknande organisationer i deras arbete för att hitta hot från onlineaktivitet samt analysera den allmänna diskursen kring ett ämne. Bidraget från denna forskning är dubbelt: (1) Rapporten undersöker hur den icke-statliga organisationen Civil Rights Defenders för närvarande arbetar med analys av sociala medier och hur deras arbete kan underlättas. Detta görs genom intervjuer med Civil Rights Defenders, och resultaten presenteras genom två användar-scenarion och användar-personas. (2) Ett verktyg utvecklas för att hjälpa Civil Rights Defenders att förbättra sitt arbete med sociala medieanalyser. Detta görs genom en litteraturstudie av liknande forskning och aktuella tekniker, samt med insikter från intervjuer. Den samlade informationen används sedan för att bygga en fullt fungerande webbapplikation, särskilt riktad mot icke-statliga organisationer och deras arbete för att bekämpa hate speech på internet.

Det utvecklade verktyget testades, och resultaten var tillfredsställande. Testarna kunde använda verktyget med lite, om någon, förvirring och verktyget hjälpte deras arbetsflöde.

Nyckelord: AI, Big Data, Hate Speech, NGO, Machine Learning, Social Media, Twitter

Acknowledgments

This thesis has been conducted with close collaboration with representatives from Civil Rights Defenders and the Faculty of Innovation Engineering at LTH.

First, we would especially thank our supervisor Erik Helldén at Civil Rights Defenders for his valuable insights and guidance during the project. We also want to thank Marcin De Kaminski, and the rest of the employees at Civil Rights Defenders for their time committed and the positive feedback we have received during the process.

Second, we want to show our gratitude to our supervisor Lars Bengtsson at LTH for believing and supporting us during the process.

Finally, we would like to thank Twitter and Tisane Labs for their generous trial periods and insights, which enabled us to develop this proof of concept, as well as the field experts interviewed for showing their interest in our project.

Lund, May 2021

Ludwig Hedlund and Simon Åberg

Table of contents

Introduction

1.1 Background

Social media platforms have formed the beginning of the 21st century. The largest platform, Facebook, has at the end of 2019 over 2 billion monthly active users, while Twitter has more than 300 million monthly active users (Facebook, 2021a)(Twitter, 2021a). Social media platforms are a powerful tool to spread ideas and discussions. These discussions have often resulted in good changes on a societal level, such as the "black lives matter "-movement or the "me-too"-campaign. However, as the tools allow for such fast spread in the networks, harmful ideas spread too. Hate speech is a growing issue on social media platforms. Recent outcomes from hate speech resulted in the Rohingya genocide in Burma, synagogue shootings in Pittsburgh, and anti-muslim mobs in Sri Lanka. (Mathew et al., 2019)

Hate speech does not have any international legal definition, and the characterization of what is "hateful" is disputed. In this thesis, we will follow The UN's definition of hate speech: "Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor." (UN, 2019)

The European Union (EU) has been working with these kinds of questions for several years. The primary and arguably the most influential European instrument is the Code of Conduct countering illegal hate speech online (CoC). Some of its positive features are that it is not a governmental speech restriction but rather a voluntary self-regulation of social media platforms and that it is reactive and only removes content that is flagged as hateful by the internet community. Some of its drawbacks are that it is untransparent and leaves the information about the volume and impact of hate speech on the social media platforms and that the regulations could make people move to less monitored platforms to express their racist and xenophobic opinions. One could argue that it does not fight the underlying root of hate speech. (Bayer & Bard, 2020)

During the Covid-19 pandemic, the United Nations have seen a rise of hate speech and xenophobia on social media, often against migrants and refugees falsely spreading the virus. Negatively affecting the process made to reach the UN's sustainable development goals of the agenda 2030. (DGC, 2020) To counter this effect, NGOs, governments, and academia need to work together with social media platforms to develop policies and different technologies to tackle online hate speech, especially in a global pandemic, such as Covid-19. (Lee & Li, 2020)

1.2 Civil Rights Defenders

Civil Rights Defenders (CRD) is an international non-profit human rights organization that supports local human rights defenders in capacity building, training, funding, and security for human rights defenders at risk. In 2019, CRD had 65 employees worldwide, mainly from a social science and law background. They have the last two years had an income of around 90 million SEK, funding is primarily from public grants (65%), as well as foundations and organizations (8%), companies (21%), and the general public (6%). CRD is both politically and religiously independent and operates in Sweden as well as in countries where human rights are the weakest. For example, they are working to prevent the democratic debate from shrinking and helping journalists and other opinion-makers carry out their work without receiving threats and hate speech. This is done by supporting local human rights defenders and pressuring governments and global organizations such as the UN and the EU. (CRD, 2019)

CRD is continuously working with innovation regarding security for human rights defenders. Some of their more successful technical projects are the Defenders Database and the Natalia project. The Defenders Database is a database to store documents and evidence of human rights violations digitally. The database was initially launched in Cuba in 2019 but has since then grown to several countries where human rights violations are increasing. The Natalia Project is the world's first alarm and position system for human rights defenders at risk. The project was launched in 2013 and is today providing increased security for 170 individuals. In case of an attack, the alarm will go off, and CRD can coordinate help for the individual in danger. (CRD, 2019)

Both EU and CRD has seen a causal relationship between individual targeted hate speech on social media and physical attacks against human rights journalist and opinion-makers. For example, the murder of Walter Lübcke, a German politician who worked for a more open refugee politics (Bayer & Bard, 2020). CRD's employees are today actively monitoring social media but are only searching the platforms by the platforms' own search functionality, which is not that efficient for their CRD's work. CRD, therefore, needs a tool to more efficiently be able to detect and monitor hate speech on social media. (De Kaminski, 2021)

1.3 Problem definition

With the amount of daily political discussions globally carried out on social media, a tail of hate speech and oppression is followed on the same platforms. Governments and organizations have therefore seen a need for tools to analyze the discourse on social media. (Bay, 2021)

In the EU's (CoC), they recommend that the EU support and provide funding to NGOs working to combat hate speech online and supporting its victims. They also recommend that NGOs should be entitled to notify the police about hate speech online which the police should be obligated to start an investigation. (Bayer & Bard, 2020)

CRD is already working with hate speech online and should therefore be a perfect fit for this kind of funding if the EU goes through with their recommendations. This would make CRD able to further invest in the workforce and tools to monitor and analyze social media platforms. (De Kaminski, 2021)

This master thesis will conduct an exploratory study to investigate how one could build an application to analyze social media platforms. The application developed during the project will be a proof of concept to showcase the possibilities of such a tool.

1.4 Objective

The research objective is to research how to develop an application that fulfils the NGOs needs when analyzing a social media platform and develop it. This leads to the following research question:

1.4.1 **Research question**

> ***How can one develop an application to support NGOs work in conducting hate speech analysis on social media platforms?***

1.5 Limitations

The scope of the thesis has been limited to analyze Twitter. There are multiple reasons why Twitter was chosen. The main one was their helpfulness in sharing their data for academic reasons. Another reason for choosing Twitter was the focus on political topics in the discussions compared to similar platforms such as Facebook or Instagram. (Pelletier et al., 2020)

This project was also under a strict time constraint which made the authors restrain the application only to analyze Twitter and use third-party providers for various services such as text analysis and user-account analysis.

1.6 Contributions

This project contributes to the work against online hate speech with a platform that can visualize tweets from Twitter and detect and analyze hate speech. Compared to today's way of working with this kind of data, the application enables less technical people to gain insight into the quantitative side of tweets from Twitter.

The project also contributes to the open-source community by being a foundation on which future applications can be built upon.

1.7 Outline

The outline of the remaining thesis is as follows:

Chapter 2: Background

This chapter presents the theoretical background of the project. The chapter starts with introducing the social media landscape and continues with a description of the technologies used in the project. Lastly, the 2030 goals are presented.

Chapter 3: Methodology

This chapter describes the work methodology that is used in the research. More specifically, the literature study, interviews, problem investigation, development, and evaluation are presented.

Chapter 4: Investigation phase

This chapter presents the literature review, which leads to two scenarios together with personas being created. The scenarios and personas are then analysed, and the chapter finishes with investigation conclusions and application requirements.

Chapter 5: Development phase

This chapter describes the development phase, from initial mockups to beta evaluation. The initial mockups were developed into an alpha version of the tool, which was then evaluated through unstructured interviews. The results from the alpha evaluation were used to create mockups of a beta version, which was developed and evaluated to create a base for the final tool.

Chapter 6: Evaluation phase

This chapter describes the final product evaluation, which was based on interviews with CRD employees.

Chapter 7: Discussion

This chapter presents a final discussion about the project. The result is discussed in comparison to the research question. Limitations and ethical aspects of the projects are discussed, as well as how the project contributed to the UN Sustainable Development Goals. The chapter finishes with a section about how the results from this project could be improved with future contributions.

Chapter 8: Conclusion

This section presents the conclusions and summary of the project.

2 Background

This chapter presents the theoretical background of the project. The chapter starts with introducing the social media landscape and continues with a description of the technologies used in the project. Lastly, the 2030 goals are presented.

2.1 Differences between social media platforms

Facebook is the biggest platform and is primarily used for entertainment purposes, at least in the west. Twitter is the platform users tend to go-to when it comes to information. Twitter is also the only platform that is completely open. Many of the discussions on Facebook or Instagram are closed to public access, while the discussions on Twitter are open to anyone. (Cullinane, 2015)(Facebook, 2021b)

In addition to the openness of Twitter from a user perspective, Twitter has opened its data up for academic research through various tools that make it possible to analyze all data on Twitter. Facebook, on the other hand, does not have any tool available for public or academic access. (Twitter, 2021a) (Facebook, 2020)

In work against platform manipulation and hate speech, the platforms definitions of hate speech are the foundation. Facebook's definition of hate speech is "a direct attack towards people with characteristics: ethnic group, national background, disability, religious beliefs, caste, sexual orientation, gender, gender identity and serious illness. We define attacks as violent or dehumanizing statements, harmful stereotypes, statements of inferiority, expressions of disgust, disgust or dismissal, swearing and calls for exclusion or segregation." (Facebook, 2021c). Twitters rules against hate speech are "You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.". (Twitter, 2021d)

During the Brexit voting, many automated bot accounts were detected on Twitter. Evidence points towards that Russia were involved in managing the bot activity, which were on the leave side of Brexit. The bots were spreading divisive messages, which were increasing the polarization around the event. Twitter had 5.4 million spam reports during the period January to June 2020, a rise of 16% since the period before. Twitter has their own algorithms to counter automated activity and believes that external tools for finding bot accounts are too limited to be used. (Roth, Y. & Pickles, N., 2020) (Twitter, 2021e) (Williams et al., 2020)

In a report by the European Commission, where they investigated how different social media platforms acted on user-reported content, the result was that Facebook removed 82.4% of the reported content, while Twitter removed 43.5% of the reported content. There was also a measure of how quickly the platforms responded to reports. Facebook assessed notifications in less than 24h in 92.6% of the cases, and Twitter assessed 88.3% of cases within 24h. In the CoC, it is stated that the target is for companies to "review a majority of all notifications within 24h", which all major platforms succeed with. (Jourová, 2019)

When it comes to the detection of hate speech, Facebook released numbers in 2019. The platform had detected 7 million posts with hate speech, 80% was automatically detected, and 20% was detected by users and then removed by moderators. The AI that detects hate speech works with bigger languages. In an interview, Facebook says that its AI cannot detect hate speech in smaller languages, such as Assamese, because they do not have a large enough dataset to train the algorithm (Perrigo, B., 2019). Twitter continuously releases reports on their activities towards enforcing their rules. In 2019, the total amount of content that Twitter removed was about 5 million. In the same period, they also suspended 1.5 million accounts. There is no information about how Twitter detects hate speech (Twitter, 2021f).

2.2 Overview of used technology

2.2.1 **Application Programming Interfaces (APIs)**

APIs help develop complex functions more efficiently by abstracting functionality and giving the developer an easier way to use the functions. Instead of writing low-level code which connects to a database or changes a graphical interface, APIs create a quicker and more human-readable way of developing. When developing, many different types of APIs are used without thinking about it. For example, there are APIs for manipulating how a website is rendered in the browser to change the user interface, fetch data from a server, or manipulate data in certain ways. We will focus on APIs used to fetch data and, more specifically, third-party APIs that allow for fetching data from a third-party source. (MDN contributors, 2021)

2.2.2 **Natural Language Processing (NLP)**

Natural language processing is an area containing technologies that strive to help machines understand human language. NLP is a branch of Artificial Intelligence (AI), which can help software use human language as input. Many applications currently use NLP, such as Apple's intelligent assistant Siri or text translation application Google Translate. Many of the most common applications of NLP can be divided into the following areas:

- Speech recognition
- Natural language understanding
- Natural language generation

This paper will only focus on NLP regarding text analysis, which more specifically involves sentiment analysis, named entity recognition, and part of speech tagging.

Sentiment analysis subtracts subjective values from a text, most commonly if a text is "positive" or "negative", but could also be, for example, "hate" or "love". Named entity recognition is the technology that attempts to extract entities from a text. For example, it could extract the tag "organization" from "UN". Part of speech tagging is tagging a text or part of a text with a label. (IBM, 2021)

What the NLP model detects in the text and how it detects it depends on the purpose of the model and how it is working. NLP models can be constructed in different ways. Most common are lexicon base models and models based on trained neural networks.

- Neural networks are a way of training a computer algorithm similar to how humans learn. By giving the algorithm text data as input, which is pre-classified, the algorithm can learn how to classify other texts.
- Lexicon-based NLP models are based on a dictionary of words and rules. By having a large set of words and rules, the model can extract words and apply the pre-defined rules on the words to give the output. If, for example, new slang for hate speech appears like the "Chinese virus" during the Covid-19 pandemic, the pre-defined rules need to change, and the models need to be updated. (Donges, 2021)(Karlgren, 2021)

2.2.3 **Graph Databases**

A database is structured data organized in a collection. Most databases are structured in rows and columns, similar to a spreadsheet. The data can be accessed, deleted, updated, and created. A database can have many different tables of rows and columns, for example, a table for users and another table for tweets. To link them, there can be a pointer from the tweet table to the user who wrote the tweet.

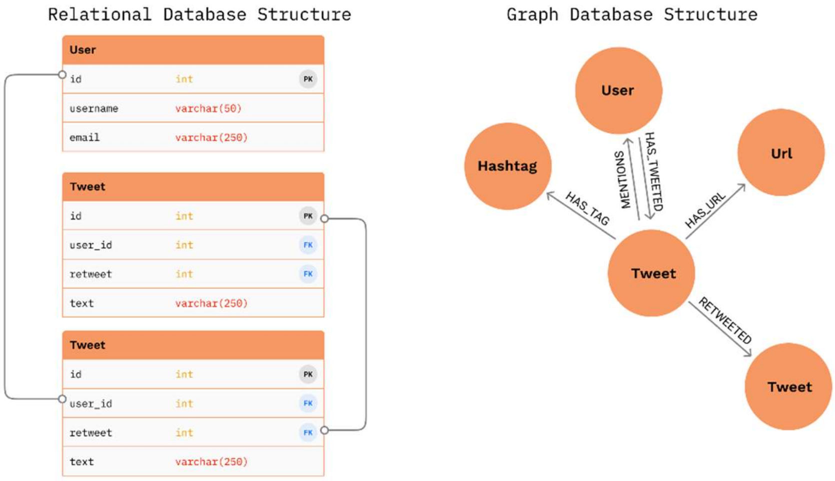


Figure 1: An overview of a graph database structure. (Created by authors)

A graph database is a database that, opposed to most traditional databases, treats the relation between data, tables in traditional databases, equally as important as the data itself. When creating a Twitter database, there would be users and tweets. In a traditional, relational database, the relationship between them would have to be defined in the node itself. For example, a tweet has an author field, which can be seen in the relational database structure in figure 1. In a graph database, the user and tweets are connected with relationships that hold these properties instead. The tweet could point to the user, and the relationship would have the value “mentions”, or a tweet could point to a tweet, and the relationship would have the value “retweeted”. (Neo4j, 2021b)

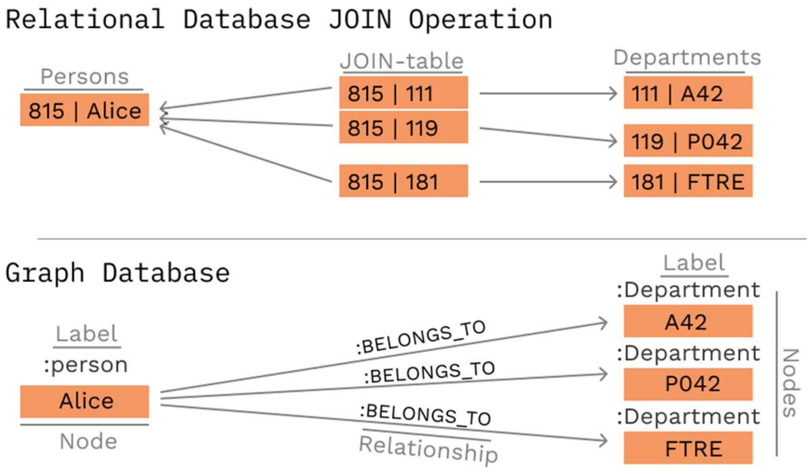


Figure 2: Relational database JOIN operation vs Graph database - only a single hop from Alice is required to see her departments. (Created by authors)

Another aspect of graph databases versus relational databases is the performance in complex queries. In a relational database, when a query is referencing a table from another table, a JOIN operation has to be done. The JOIN operation between tables is computationally heavy. On the other hand, in a graph database, a reference from one node to another does only need to see which nodes have the desired relationship and belong to each other, see figure 2. (Neo4j, 2021b)

2.3 The 2030 Agenda

The United Nations (UN) sustainable development goals are a call to action for all countries in the world to develop in a sustainable way both for all people living on the planet and the environment. The 2030 agenda was set in 2015, and progress has been made in many places, but to achieve the goals 2030, action to meet the goal needs to speed up. (UNSD, 2020)

In this thesis, two of the goals, 10 and 16, are especially applicable. Goal 10 concentrates on reducing inequalities within and among countries, and goal 16 focuses on promoting peaceful and inclusive societies whatever one's ethnicity, sexual orientation, or faith. Fighting discrimination works towards both goals. Discrimination can often lead to inequalities within countries and make society less inclusive and even dangerous for groups of people. Detecting and monitoring discrimination, for example, hate speech on social media, could be a way to reduce inequalities and work towards safer and more inclusive societies. (UNSD, 2020)

With the Covid-19 pandemic, the inequalities that are addressed in agenda 2030 in the world have been further exposed. For example, the most vulnerable people are often the ones whom Covid-19 is hitting the hardest, amplifying inequalities and can lead to increased discrimination in the form of example Covid-19 related hate speech (UN, 2020a) (UN, 2020b) (DGC, 2020). However, when social norms and policies are disrupted in a time of crisis, it could be possible to take action and rebuild the world towards reaching the UN's sustainable development goals 2030 (*The Sustainable Development Goals: Our Framework for COVID-19 Recovery*, 2021).

3 Methodology

This chapter describes the work methodology that is used in the research. More specifically the literature study, interviews, problem investigation, development, and evaluation are presented.

3.1 Research design

When developing a new product or service, design thinking is a human-centered process for problem-solving. The approach focuses on the human needs behind every decision and can decrease the risk when creating new products or services. (IDEO, 2020) When developing software, different methodologies can be used depending on the task and its requirements. For this project, an agile development process was chosen because of the size of the project and its need for iterative feedback during the development process. The development process consisted of three different prototypes to gain continuous feedback during the development of the new application. The work process can be seen in figure 3 below.

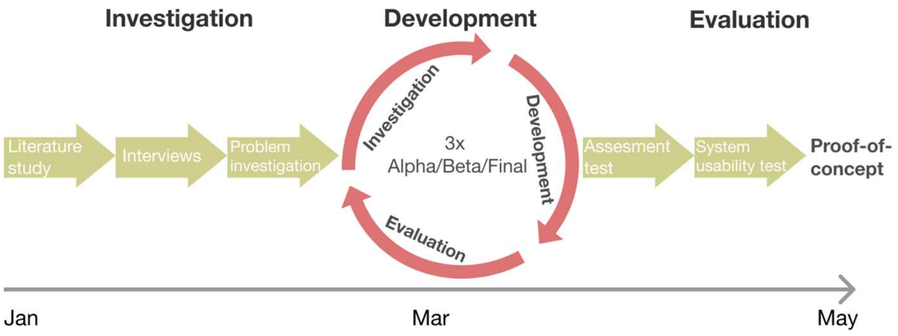


Figure 3: Research design. (Created by authors)

3.1.1 **Agile development**

Agile development is commonly used in modern software development. Two of its fundamental characteristics are:

1. The process of different specifications is minimized. For example, the user requirement specification is kept only to define the most important characteristics of the system.
2. The product is developed in a series of versions. Between the versions, stakeholders evaluate the product and give feedback for changes to future versions.

Agile development works best for small to medium projects where the customer/user is committed to being involved in the development process to provide continuous feedback during the process. (Sommerville, 2011)

3.1.2 **Literature study**

A literature study has been conducted to give theoretical background to the subject of hate speech on social media and techniques to analyze hate speech.

The study has been done through different sources and databases such as Google Scholar, LUBsearch (Lund University's own search engine), as well as suggestions from our supervisors (Bengtsson, 2021) (Helldén, 2021) and expert interviews (De Kaminski, 2021)(Karlgren, 2021)(Petterson, 2021)(Kullenberg, 2021a)(Bay, 2021). Keywords used in the search include: "Hate speech social media", "Hate speech Twitter", "Sentiment analysis social media", "Sentiment analysis Twitter", "Text analysis social media", "Text analysis Twitter", "Hate speech Covid-19", "Graph database Twitter", "Graph database sentiment analysis", and "Graph database text analysis".

3.1.3 **Interview study**

3.1.3.1 Interviewee selection

When selecting interviewees, the aim was:

1. To gather comprehensive background information to what existing research and efforts there are to counter and/or analyze hate speech.
2. To gather information about how existing products or services, if any, are used and what value they bring.
3. To gather information about possible user personas and scenarios.

To find relevant interview objects, the Civil Rights Defenders network was used to a great extent.

3.1.3.2 Interview structure

Interviews were conducted with a mix of an unstructured and semi-structured approach, depending on the aim of the interview. When the purpose of the interview was to gather background information, the interviews had an unstructured approach. An unstructured approach works well when the intention of the interview is for the researcher to develop a better understanding of the interviewee's perspective and point of view when developing a new application. In unstructured interviews, the interviewee leads the conversation, and a skilful interviewer mainly listen and asks broad and open questions but still control the interviewee to stick to relevant topics. (Zhang & Wildermuth, 2009)

When the interview aimed to do a product assessment, the interview followed a semi-structured approach. This is appropriate when the interview intends to explore certain opinions among the respondents, and the respondents have different backgrounds, such as professional or educational, meaning that a structured approach is not applicable. The semi-structured approach gives the interviewer freedom to ask clarification questions and explore interesting issues raised by the respondent. (Barriball & While, 1994) See Appendix A for the semi-structured interview.

3.1.4 **Problem investigation**

To be able to develop a useful product, an idea of the users' needs is essential. By creating a good foundation with insights into the users, a set of requirements can be created to help the development direction. By combining insights from literature and interviews with possible users, the risk of developing an inadequate service is reduced. (Brown, 2020)

The problem investigation process, or in software development terms, the Requirements engineering process, consist of four main activities:

1. Feasibility study
2. Requirement gathering
3. Requirement specification
4. Requirement validation

In agile development, which is the method used in this thesis project, the requirement engineering process is an iterative process conducted through the whole development process. The final requirements specify what the system should be able to do but not how. Requirements can be structured into domain-level requirements and design-level requirements. Domain-level requirements describe the user's tasks and are easy to test. Design-level requirements are “soft” requirements, focusing on requirements that are not as easily tested, such as the design of a user interface.

The first step in the process is a feasibility study, which should be cheap and quick and inform if the overall problem is achievable and worthwhile for the user in terms of functionality compared to existing software as well as possible to develop and maintain in an economic sense. The feasibility study is conducted by investigating existing software, literature reviews of intended technologies, and interviews with experts both in intended technologies and already existing software.

The requirement gathering consists of tasks such as investigating similar products, interviews with potential users, and interviews with experts in the intended field. To understand and gather desirable requirements, building scenarios and personas could be effective when discussing with stakeholders. Scenarios are a user-centric approach to help create and test valid requirements focused on the correct issues. Scenarios structured in user stories describe what a user should be able to do and what requirements there are to solve the task. Personas describe the target user and are a way of finding the user’s constraints. For example, data scientists have different constraints than journalists.

Prototyping is also a valuable way to gather information about specific requirements better. This is common in an iterative development process where the prototypes can both validate existing requirements as well as helping with understanding and finding new missing requirements. After the requirements gathering, it is important to specify them, so all stakeholders have a common view and similar expectations on the end-product. The fourth and last activity is to validate the specified requirements. An important process to secure that the product that is going to be built fits the customers’ needs. (Sommerville, 2011)

3.1.5 **Development**

As mentioned earlier, the development process in this project consisted of three bigger iterations which lead to an alpha release, beta release, and the final release. Each iteration consisted of an investigation phase, a development phase, and finally, an evaluation phase to gain continuous feedback during the process to find new requirements and validate already existing requirements. The intended functionalities were implemented out of the requirement specification, starting with the overall software architecture and more prominent components such as the backend, frontend, and database structure. When the structure was set, smaller components were implemented to satisfy the requirements successfully. (Sommerville, 2011)

3.1.6 **Evaluation**

Continuous evaluation when designing a new product is of great importance. Both to gain new ideas but also to validate that the product will be usable for its user. In this thesis, different evaluation methods will be used during different stages of the process. For the alpha and beta versions, exploratory tests will be conducted through unstructured interviews with stakeholders and field experts. The final result will primarily consist of usability testing in the form of an assessment test with intended end-users. (Sommerville, 2011)

3.1.6.1 Exploratory test

Exploratory tests' primary objective is to examine the early design concepts in the form of effectiveness. It strives to build a deeper understanding of the user's needs and evaluate if the product will be able to achieve those needs. When conducting exploratory tests, the participant and the test moderator have extensive interaction, usually in an informal and unstructured or semi-structured way, to encourage the participant to open up and "think aloud". Instead of focusing on questions that provide answers on how well the product is working, exploratory testing strives to understand why the user performs in a specific way and gain insights into possible improvements. (Rubin & Chisnell, 2008)

3.1.6.2 Assessment test

An assessment test is a type of usability test where the test persons will be performing different tasks. An assessment test aims to investigate how well the product has been implemented by letting the test user perform realistic tasks and identify possible deficiencies in the product. (Rubin & Chisnell, 2008)

The final version of the application will be tested with an assessment test. Its result will be presented as the final result and answers from a usability questionnaire from each participant of the assessment test. See the assessment test procedure in Appendix C.

3.1.6.3 System Usability Scale

The System usability scale (SUS) is a quick and easy scale of ten usability questions. The statements range from different usability aspects and are graded by the test persons on a five-point Likert scale. These results are then transformed into a predefined algorithm presented on a scale between 0-100. A result over 70 is acceptable and usually indicates that the product was well received and user-friendly. (Brooke, 2013) The questionnaire can be found in Appendix C.10.

4 Investigation Phase

This chapter presents the literature review, which leads to two scenarios together with personas being created. The scenarios and personas are then analysed, and the chapter finishes with investigation conclusions and application requirements.

4.1 Literature review

The literature review is ordered by relevant initiatives, relevant services, and third-party technologies.

4.1.1 **Relevant initiatives**

4.1.1.1 HateLab

HateLab describes themselves as a global hub for data and insight into hate speech and crime. By using, data science methods, including AI, they can measure and counter the problem of hate, both online and offline. The project is funded by UK Research and Innovation, as well as the US Department of Justice. (HateLab, 2021)

The organization is developing a platform for aggregating trends over time. HateLab has not released the service publicly yet, but it is piloted by National Online Crime Hub and has received £1,726,841 in funding. (HateLab, 2019)

They are one of the world’s leading initiatives in work against hate speech online and have released several research articles provided by their platform.

Their articles contribute to data and theory-driven research about hate speech online. They continuously address the importance of social media as a part of the formula when investigating hate speech and discrimination against minorities. They have done multiple researches about hate speech triggers and have shown that offline attacks such as the Christchurch, New Zealand extreme-right wing attack 2019 triggered hate speech across the United Kingdom. (Williams et al., 2020)

4.1.1.2 OsoMe

The observatory on social media (OsoMe) is an initiative by Indiana University, run as a joint project by the universities technological and media faculties. Their mission is to build tools and research the diffusion of misinformation, uncover vulnerabilities of the media ecosystem and develop methods for increasing the resilience of citizens and democratic systems against manipulation.

They have developed a range of tools that helps them grow the understanding and analyze information diffusion, detect misinformation and evaluate the trustworthiness of new influentials. Some of these tools are Botometer, which is a tool that can be used to check how "bot-like" a Twitter user is, as well as Hoaxy, which can visualize the spread of information on Twitter. (Osome, 2021a) (Osome, 2021b)

Both Hoaxy and Botometer have free APIs. The Botometer API is used in our application and is described in the "technologies" section of the literature review. (RapidAPI, 2021a) (RapidAPI, 2021b)

4.1.1.3 Jigsaw

Jigsaw is a unit within Google that investigates online threats and builds services that counter the threats. The initiative was started as Google Ideas in 2010 but changed its name to Jigsaw in 2016 (Schmidt, 2016). Jigsaw has developed tools, conducted research, and done initiatives against online threats. For example, Jigsaw has built the Perspective API, a freely available service for analyzing texts for toxic content. The API uses machine learning and natural language processing to classify a text with a toxicity score and is available for free with a rate limit of 1 request per second (Perspective API, 2021). Another example is Redirect Method, an initiative for counter-terrorism that uses targeted digital ads to confront online radicalization. (Jigsaw, 2021)

4.1.1.4 Detecting and Monitoring Hate Speech in Twitter

In this paper, the authors present HaterNet, an intelligent system currently being used by the Spanish National Office Against Hate Crimes of the Spanish State Secretariat for Security that identifies and monitors the evolution of hate speech on Twitter. The system can monitor and visualize hate speech on social media using network analysis techniques. In the report, the authors have also investigated different methods of analyzing text on social media using NLP algorithms. The conclusion was that a combination between two machine learning algorithms, MLP and LSTM, based on neural networks gave the best result. (Pereira-Kohatsu et al., 2019)

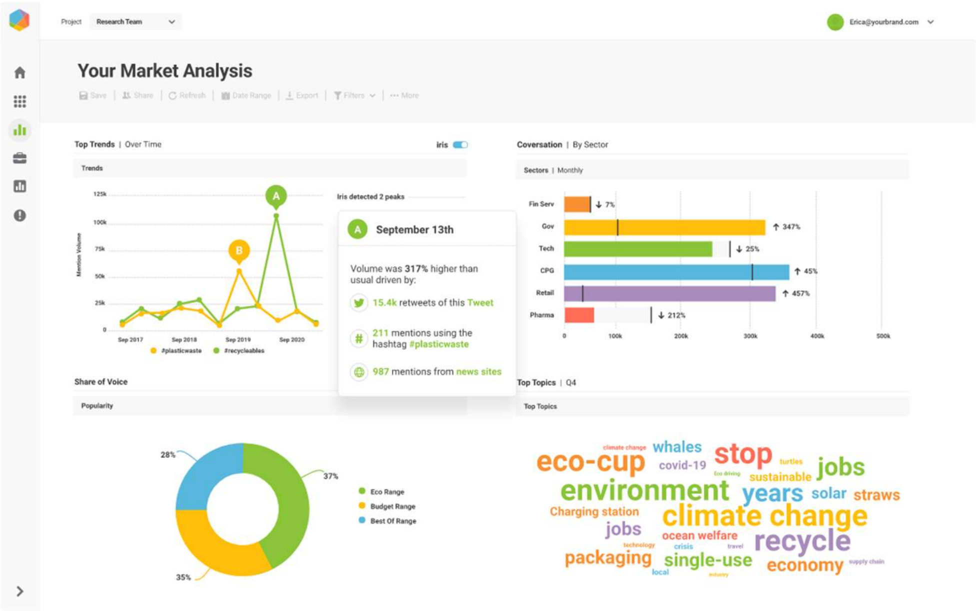
4.1.1.5 PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis.

In this paper, the authors built the system PoliTwi, an application made to detect emerging political topics on Twitter. The method of finding emerging topics was analyzing how many times a hashtag appeared in a period compared to a previuos period. A hashtag with significantly higher occurance rate in the later period would be classified as emerging. They tested the system during the 2013 parliamentary election in Germany, and found that the system could find trending topics sooner than Google's own Google Trends application. This also indicates that trends emerge sooner on Twitter than on Google. Secondly, they show how these topics can be used to extend existing analysis models with the new topic data. (Rill et al., 2014)

4.1.2 **Relevant services**

4.1.2.1 Brandwatch

Brandwatch is one of the biggest services when it comes to social media analytics, and finding consumer insights from online activity. The service can analyze multiple social media platforms, such as Twitter, Facebook, and Instagram, among others. Brandwatch can find posts, comments, and conversations on keywords and can also analyze the data through an AI that can give automatic insights. The service is targeted towards market analysis but can be used for social analysis. The Swedish Defence Research Agency (FOI) is, for example, using them for some of their online analysis for broad research (Bay, 2021). In 2019, the cost was 1000 USD per month for usage, Brandwatch does not provide pricing details (Marvin, 2019). An overview of Brandwatch's UI can bee seen below in figure 4.



A September 13th

Volume was 317% higher than usual driven by:

15.4k retweets of this Tweet

211 mentions using the hashtag #plasticwaste

987 mentions from news sites

Figure 4: Overview of Brandwatch UI. (Brandwatch, 2021)

4.1.2.2 CrowdTangle

CrowdTangle is a tool by Facebook for analyzing Facebook activity. They offer user interfaces where it is possible to search and filter for activity, as well as an API. With the service, it is possible to follow content across Facebook's social media platforms, Facebook and Instagram, as well as the external platform Reddit. Crowdtangle has a dashboard where it is possible to analyze millions of social media accounts in real-time and find trending topics.

In their own words, people who use CrowdTangle are journalists that search social media for content relevant to their reporting, fact-checkers identifying posts that contain misinformation, researchers analyzing thousands of accounts over time and reports how information spreads. CrowdTangle is free to use but is only available for selected Facebook publishing partners. (CrowdTangle, 2021a)(CrowdTangle, 2021b)

4.1.3 **Third party technologies**

4.1.3.1 Twitter API

In this project, Twitter's API has been used to get Twitter data. The API gives the possibility to fetch millions of Twitter posts and users. During this project, Twitter initiated a free academic access program, which opens up features that were only available for enterprise customers before. To be eligible for the program, the researcher has to be at least a master's student. Some of the features in the academic access plan are:

- Query historical tweets all the way back to the first tweet in 2007. Before the academic access plan, it was only possible to find tweets from the last seven days.
- 10 million tweets per month instead of the limit of 100 000 tweets on the regular plan.
- Tweet stream, which allows for subscribing to specific keywords/hashtags and getting new tweets in real-time.

When querying for tweets, multiple parameters can be added to filter the search. Some of them are: user, keyword, date range, exclude user, and exclude keyword. (Twitter, 2021g)

The data that is received in response is the following:

Tweet data:

`attachments`, `author_id`, `context_annotations`, `conversation_id`, `created_at`, `entities`, `geo`, `id`, `in_reply_to_user_id`, `lang`, `public_metrics`, `possibly_sensitive`, `referenced_tweets`, `reply_settings`, `source`, `text`, `withheld`

User data:

`created_at`, `description`, `entities`, `id`, `location`, `name`, `pinned_tweet_id`, `profile_image_url`, `protected`, `public_metrics`, `url`, `username`, `verified`, `withheld`

(Twitter, 2021g)

The rate limitations of the API are a maximum of 500 tweets per request, 1 request per second, and/or 300 requests per 15 minutes. In other words, it is possible to get a maximum of 150 000 tweets per 15 minutes. There are also restrictions in terms of data distribution. For example, if an application that uses Twitter's data is publicly available, the data has to be in sync with Twitter. If a user deletes a tweet or an account, the account and tweets must be deleted in the application, which uses the API. (Twitter, 2021b) (Pettersson, 2021)

4.1.3.2 Tisane Labs API

Tisane's API was used to analyze Twitter data. Tisane offers a natural language processing service, where a text string can be sent to them for analysis, where they respond with the result from the analysis. Tisane advertises itself as an API made to be used for analyzing the content in online communities. It could, for example, help moderators quickly find posts that need to be checked. The underlying model is a lexicon-based NLP model. Due to company secrets, it was not possible to gather more information. (Tisane Labs, 2021b)

A request to Tisane is made by sending the text which is to be analyzed, together with the language of the text and a few optional settings parameters which can alter how text is analyzed and what data is sent back. (Tisane Labs, 2021b)

The data that can be received from Tisane is the following:

- Abusive content, an array of detected instances that may violate the terms of use, or in our case, be tweets that have hate speech or other dangerous content.
- Sentiment expressions, an array of detected sentiment expressions. Includes data of why the sentiment expression was triggered on a word and if the sentiment is positive or negative.
- Entities, an array of named entities detected in the text, such as a company or a country.
- Topics, an array of topics, which includes detected subjects, domains, or themes in the text.

(Tisane Labs, 2021b)

The API limit of Tisane is 400.000 API requests per month, and 120 API requests per minute. The price of Tisane with these limits is 159 USD per month (Tisane Labs, 2021c). In our project, the rate limit of the API was 1200 requests per minute, due to a custom trail.

4.1.3.3 Botometer API

Botometer, developed by OsoMe, analyzes a Twitter user for suspicious activity and gives a score for how similar the user account is to a bot. It uses machine learning algorithms that are trained on tens of thousands of Twitter accounts. It analyzes, among other things, tweets, networks, language, and sentiment to give a score of bot-like appearances.

A request to Botometer is made by sending the user id and screen name to the API endpoint.

The data that can be received from Botometer is the following:

- Twitter user object: Twitter's user object with additional data about the most used language
- Raw bot score object: There are two scores in the object, English and universal language score, both scored between 0 - 1.
- Display scores object: Same as the raw bot score but in a 0 - 5 range instead.
- Complete Automation Probability (CAP): A conditional probability that accounts with a score equal to or grated than this are automated. This score is calculated based on inferred language.

The limit of the free API is a maximum of 500 user checks per day. (RapidAPI, 2021b)

4.1.3.4 Neo4j

Neo4j is one of the most acknowledged graph database management systems available today. It was released in 2007 and has since then been continuously improved. Neo4j stands out with its simplicity and powerful query language Cypher to fetch data from the database, which is inspired by the normal query language SQL, compared to other graph databases. Neo4j also has the most active graph community. (Fernandes & Bernardino, 2018)

Neo4j is providing multiple applications and tools to support engineers when using their database system. They have, for example, their "Graph Data Science Library", which have several prewritten graph-database algorithms and functions to efficiently query data from the database. A cloud database to ease the management of the infrastructure, various visualization tools, and much more. (Neo4j, 2021a)

The pricing of Neo4j differs depending on the solution. They offer fully managed databases, where Neo4j manages everything from hosting to maintenance and scaling. This is starting at 65 USD per month. If managing the database on one's own premises works, they have free alternatives. (Neo4j, 2021c)(Neo4j, 2021d)

4.2 Use cases/Personas

Two applicable use cases/Personas were identified from interviews with Marcin De Kaminski and Erik Helldén at CRD. The use cases and their personas are first defined and then analyzed.

4.2.1 **Defining use cases and personas**

Two use cases were defined from interviews with Civil Rights Defenders. The two scenarios and their respective persona represent historical cases where Civil Rights Defenders would have been helped through easier information gathering and analysis on Twitter. Information such as names, organisations, and countries have been covered due to the sensitivity of Civil Rights Defenders work.

4.2.1.1 Persona scenario one

Scenario one involves Civil Rights Defender's partner organizations and their members. In this section, an introduction to the organization and its members will be presented. "Organization A" works with defending civil rights in "Country". They work mainly by taking legal actions towards their regime. The organization members consist of lawyers and journalists, working both with the legal actions and communicating their actions and reports.

4.2.1.2 Scenario one

A CRD partner organization, organization A, takes a political standpoint in a country run by a totalitarian regime. A short time later, a rise in hate speech and threats online occurred, targeting both the organization and its members. The organization has a hard time analyzing the level of severity of the threats. A while after, the organization's office is attacked by an extremist group. All their equipment is destroyed, leading to members leaving the organization in fear and weeks of limited ability to pursue their political agenda for human rights.

4.2.1.3 Persona scenario two

Scenario two involves a Civil Rights Defenders worker. In this section, an introduction to the Civil Rights Defenders employee is presented. "Person" works at Civil Rights Defenders communications department. The person has worked with writing reports for two years at Civil Rights Defenders and has an education in political science. "Person" is uncomfortable using data analysis tools such as Excel.

4.2.1.4 Scenario two

“Person” is writing an article about the discourse on Twitter surrounding a presidential election in “Country” with dictatorship ruling. The election caused high spikes in activity online, and several political riots occurred. As the main online forum for political discussions in "Country” is Twitter, "Person” search for different keywords to get an overview of the discourse. After the activity have been analyzed through Twitter’s own platform, the article is written. Due to too much data, “Person” could not analyze the discourse well enough to get insights to the article, therefore, it was not published.

4.2.2 **Analysis of use cases and personas**

In this section, the two cases defined in 4.2.1 are analysed. The information was gathered from interviews with Civil Rights Defenders. The scope of this section is to define the core problems from the scenarios and analyse how a Twitter analysis tool could have been used to prevent and/or improve the situation.

4.2.2.1 Scenario one

Scenario one consists of a human rights organization consisting of mainly lawyers and journalists. They have limited knowledge in fetching data from API, NLP, and data visualization. Their core problem is to in real-time detecting and analyzing threats against their organization as well as specific members of their organization. In the scenario, the organization has employees monitoring their Twitter feed manually on the Twitter platform. They could see a few negative direct comments against their organization, and one of their employees received a personal threat in their private messages. Because of the limitation of Twitter’s platform, they could not detect and understand the severity of the discourse on other parts of Twitter where their organization was mentioned in very negative tweets as well as tweets with direct links to the future attack on their office.

Had it been possible to prevent the attack against their office or at least better secure their equipment stored in the office if the spike of hate speech was detected in time?

4.2.2.2 Scenario two

In this scenario, Civil Rights Defenders must get an overview of Twitter activity in a country to write a report about the insights from the online discussions. In addition to analyze the general discourse, it would also be valuable to get insights if there are criminal activities or activity showing signs of civil rights violations. As media might be government-controlled and hence not report on acts that might damage the state, Twitter and other social media can often be a reliable way of getting insights into government activity. Twitter’s own platform is not made for analyzing large amounts of data. As none of the communications team members in Civil Rights

Defenders are technological savvy, programmatic data analysis is not an option either. This is an area where a user-friendly tool could help get an overview of the activity quickly and dig deeper into interesting keywords and/or users.

4.3 Investigation conclusions

For this project, we have concluded to limit our scope to fetch data from Twitter mainly for their open data access and the project's time constraint. To collect data from Facebook, one could implement CrowdTangle. However, it is only available for selected partners and only provides limited datasets, which, therefore, would not be applicable for this project.

Several of the interviewed, such as Jussi Karlberg and Cristopher Kullenberg, are data scientists with much knowledge of similar tools and the data analysis workflow. They are not the intended user. However, their input is important when it comes to technical constraints and desired outcomes of data analysis on Twitter.

The initial idea of the product involved more natural language processing and automation of getting insights. This would be similar to HaterNet, which mostly is a data pipeline that can find hate speech and trends in Twitter data. However, when speaking to Jussi Karlberg, an expert in natural language processing, the result was that it is very difficult, if not impossible, to build a model that can analyze multiple languages and have sufficient performance in today's state of machine learning.

Because of this, the project will go in the direction of Brandwatch; A user-friendly UI that guides the user into finding their own insights in the data, but improving on that with a natural language processing model which can give insights specific to hate speech. If Brandwatch would have had hate speech detection and a lower price, the service would be suitable for solving the tasks in the scenarios described in the use cases section.

Building one's own machine learning model is time-consuming and needs continuous maintenance because of the emerging of new words. For example, during the Covid-19 pandemic, new racist vocabulary such as "Chinese Virus" and "Kun Flu" has arrived. However, if using an NLP-solution provided by a third party such as Tisane, the development and maintenance of the model is done by them. This provides an easy integration into the application, which will stay up to date with changes in languages. Adding Jigsaw's Perspective API, which could give a toxicity score on a tweet, could improve the tweet analytics of the product. However, the rate limit for the standard plan is 1 tweet per second which is too low.

The main takeaway from the interviews in the investigation phase is that the application must be user-friendly for users without much technical background. In both the described user scenarios, the user persona is a non-technical person. There are also many similarities between the scenarios when it comes to the requirements of the application. The user must be able to fetch Twitter activity, both in real-time and historical data. Scenario one has a heavy need of getting activity in real-time, while scenario two could be more about historical data. Twitter's API has a limit of 500 tweets per second, which can be seen as a real-time search.

Both the scenarios would be helped if the application could highlight important tweets. Relevant tweets could be tweets that have hate speech, signs of criminal activity, or political activity. Human language is difficult to analyze, and languages are very different, meaning that tweets in some languages will be labelled more, or less, often than other languages, even though the content is the same. (Karlberg, 2021)

The interview with Christopher Kullenberg gave insight into how a Twitter analytics process could go. A regular workflow would be first to do a broad search and then dig deeper down into the data by filtering on keywords, hashtags, users, and more. (Kullenberg, 2021a)

With this analysis in mind, the application has some key features that should be implemented:

- Easy to use without any programmatic knowledge
- Fetch data from Twitter both in real-time as well as historical data
- Detect and highlight threatening activities on Twitter
- Visualize an overview of tweets from Twitter
- Visualize trends in tweets from Twitter
- Store tweets from Twitter for future legal reasons.

4.4 Requirement specification

Task 1 – Find potentially illegal activites

Purpose: A person should be able to analyze a keyword and find potentially illegal activites.

Precondition: None.

Sub-tasks: Analyse individual tweets, Narrow down a broad search to few keywords.

Variants: Manually look through individual tweets, use an automatic analysis service.

Task 2 – Keyword/user overview

Purpose: A person should be able to get an overview of tweets for a specific keyword or from a specific Twitter account.

Precondition: None.

Sub-tasks: None.

Variants:

- Visualize the number of tweets for the specific search.
- Visualize what languages are used for the tweets.

Task 3 – Manage tweets

Purpose: A person should be able to find and view important tweets.

Precondition: None.

Sub-tasks: Sorting through tweets based on network reach, sentiment, and suspicious activity.

Variants: None.

Task 4 – View activity trends

Purpose: A person should be able to find activity peaks to dig deeper into the activity.

Precondition: None.

Sub-tasks: Change date-range

Variants: View activity for different keywords. View overall activity. View trendline of suspicious activity. View trendline of sentiment.

Task 5 – Export tweets

Purpose: Users should be able to export data to analyze them in another tool to use as evidence in legal cases.

Precondition: None.

Sub-tasks: Filter for the desired tweets.

Variants: None.

5 Development Phase

This chapter describes the development phase, from initial mockups to beta evaluation. The initial mockups were developed into an alpha version of the tool, which was then evaluated through unstructured interviews. The results from the alpha evaluation were used to create mockups of a beta version, which was developed and re-evaluated to create a ground for the final tool.

5.1 The alpha product iteration

This section describes the process of developing and testing the alpha version of the application. The process started by creating a mockup of the application, which was then developed into a working product. Once the alpha version of the application was developed, an investigation was done into how potential users perceived the application.

5.1.1 **Investigation**

The investigation phase for the alpha product was combined with the investigation phase for the whole project. The process of creating the mockup began by ideating what features were important in an alpha version of the application and how those would be implemented. The features that were going to be implemented had to support the requirements made from the literature review and interviews.

From the requirements, some prominent features were identified:

The main view, figure 5, had one user input, where the user could select a country that would search for the country and populate the view. The purpose of the country input was to give the user the possibility to find information regarding a specific country of interest. For example, if there has been a recent event in a country, the users would like to investigate twitter data only regarding that country.

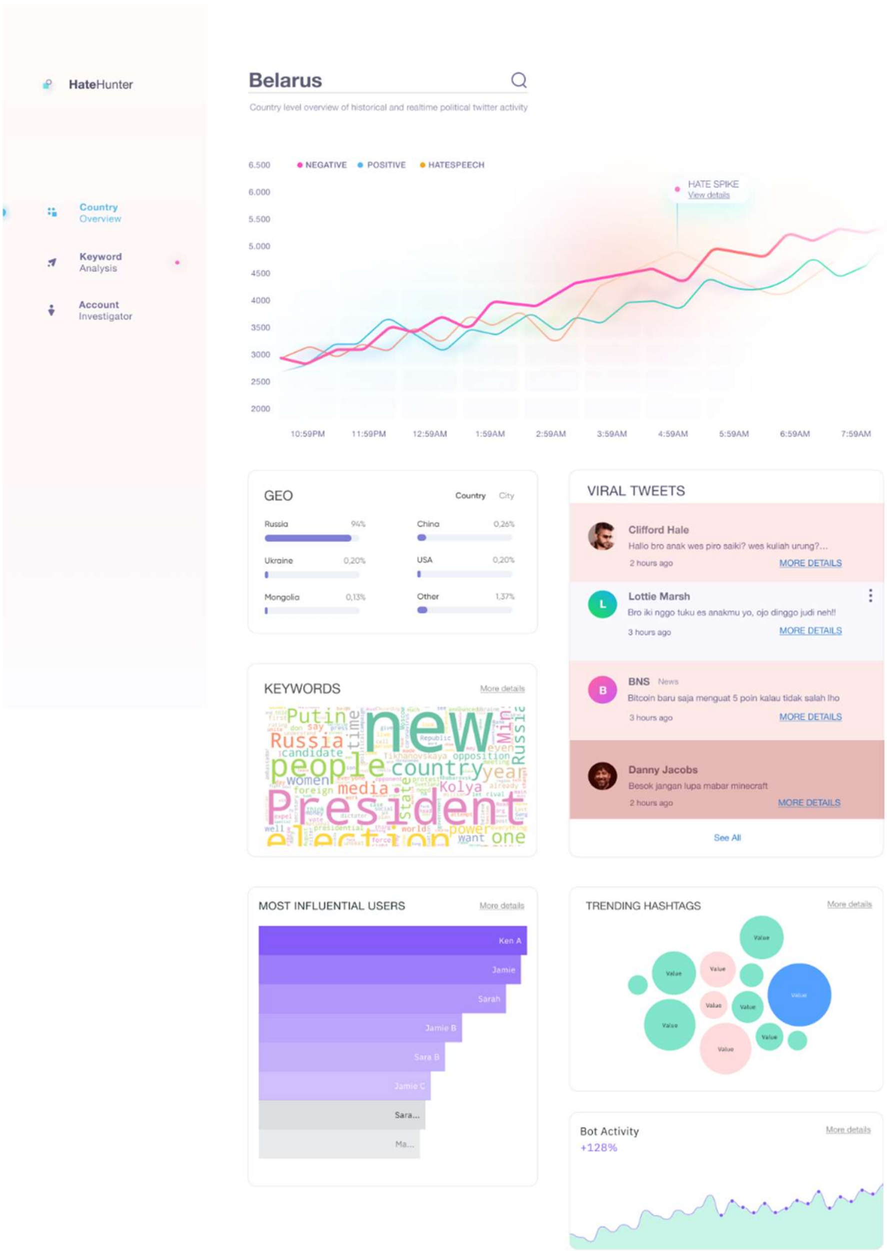
A user analysis page could be added to explore user details deeper. This view would display the overall activity of a Twitter author and an overview of key data. The user analysis page would have similar components as the main view, figure 5, for analyzing a users' Twitter activity.

A keyword search view could be added. This view would have a search filter where a user filter for keywords or hashtags. The view would have similar components as the main view for analyzing the Twitter activity.

A trendline was added to the main view, seen at the top of figure 5, fulfilling requirement 4 - viewing activity trends. The purpose of the trendline was to give the user the possibility to get insights into the overall Twitter activity that matched their country filter. The trend line would have time on the x-axis and the amount of tweets on the y-axis. The trendline would have three lines showing the activity of tweets analyzed as negative, positive, and hate speech. The classification of the tweets would be received from an NLP analysis on each tweet.

The viral tweets component was designed, seen in figure 5's middle-right, to fulfil requirement 3 - manage tweets. The component would show a list of tweets that was matching the country selection. In each tweet, data about the author, date of creation and the tweeted text would be displayed. In addition to the data from Twitter, the tweets would have a background color that would correspond to a sentiment score received from sentiment analysis. The color would go from green, a positive tweet, to red, which would be a negative tweet. The list would be sorted by a "virality"-score, calculated using the database function for network centrality that Neo4j provides.

The bot activity component, seen at the bottom of figure 5, was added to meet requirement 4 - view activity trends. By analyzing bot activity and displaying the result in a trendline, it would be possible to see if there are any peaks in suspicious activity, and from that, dig deeper into the activity.



VIRAL TWEETS

Clifford Hale
Halo bro anak wes piro saiki? wes kuliah urung?...
2 hours ago MORE DETAILS

Lottie Marsh
Bro iki nggo tuku es anakmu yo, ojo dinggo judi neh!!
3 hours ago MORE DETAILS

BNS News
Bitcoin baru saja menguat 5 poin kalau tidak salah lho
3 hours ago MORE DETAILS

Danny Jacobs
Besok jangan lupa mabar minecraft
2 hours ago MORE DETAILS

See All

Figure 5: Mockup for the alpha product.

5.1.2 **Development**

The development phase for the alpha release focused on the underlying technology, such as the database structure, backend architecture, and building a robust data pipeline from the Twitter API. The database structure to handle Twitter data can be seen in figure 6. The abuse node was pre-implemented even though we did not have any text analysis implemented at this stage. In the alpha development phase, neither the country analysis view nor the user analysis view was implemented due to technical complexity. Instead, only the keyword search view was created. The user interface (UI) was not prioritized and did not match the mockup.

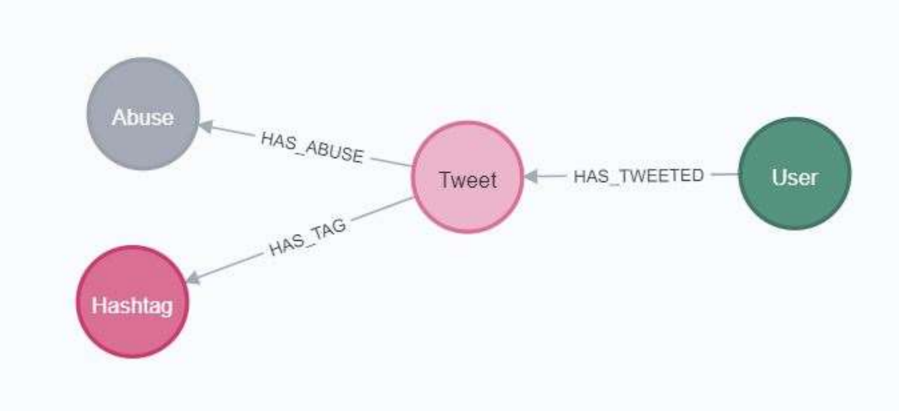


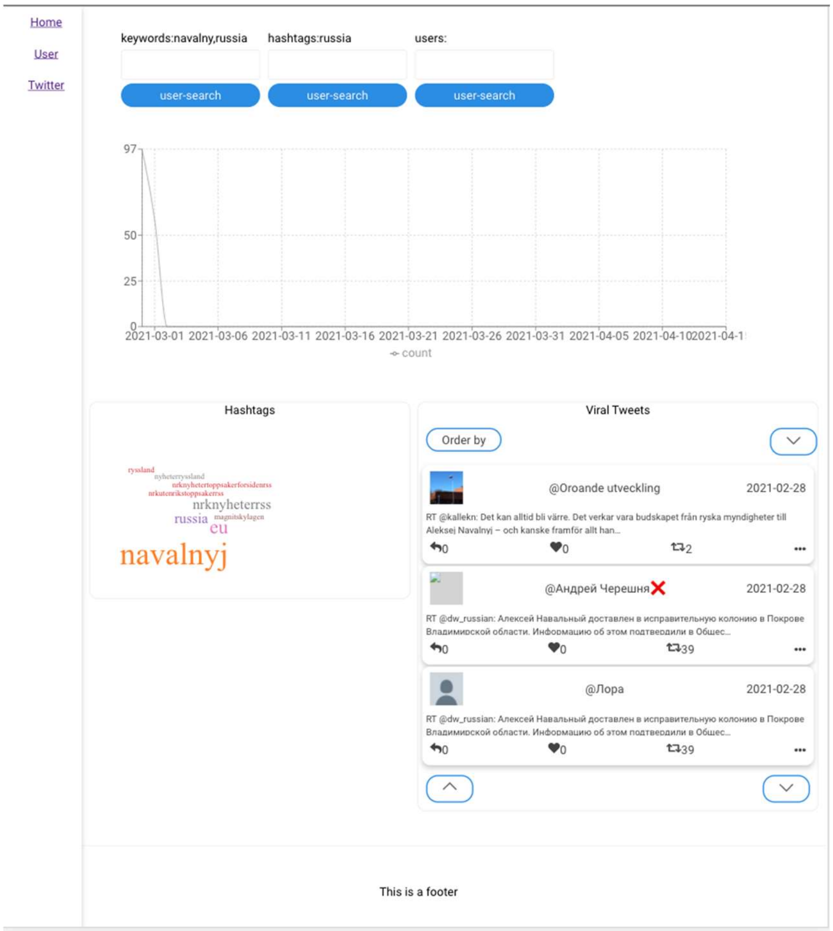
Figure 6: Data structure for the graph database. (Created by authors)

Components that were implemented and can be seen in figure 7:

- Search bar where users could input keywords, hashtags, and/or users.
- Trendline displaying Twitter activity matching the search. The implementation was only displaying overall Twitter activity. No NLP had been added to the tweets to show a line for positive, negative, or tweets that contained hate speech.
- Viral tweets component displaying a list view of tweets was implemented. The UI was changed from the mockup. In addition, no function for calculating tweets network centrality was implemented. Therefore, tweets were ordered by likes or retweets instead.
- Wordcloud of hashtags visualizing which hashtags used in the dataset. The bigger the hashtag, the more time used. When the user is hovering over a given hashtag, the count of that hashtag is shown.

Some requirements such as automatic text analysis to find suspicious activity and sentiment analysis were not fulfilled in the alpha version. To add such functionality, a third-party service would need to be used, as a custom implementation would have been too big and out of scope for the application. In addition, the tool was not able to search in real-time. A separate request had to be made to the backend to add data, which then fetched data. This was a significant insight during the development phase. The pipeline to fetch tweets and add them to the database was too slow to be able to implement a real-time search.

The bot check with Botometer was not added due to the API not being sufficient. The response had low accuracy, and it was too slow with several seconds in response time, as well as the limit of 500 checks per day being too low. The bot-like score was also difficult to interpret and was often incorrect.



@Oroande utveckling 2021-02-28

RT @kallekn: Det kan alltid bli värre. Det verkar vara budskapet från ryska myndigheter till Aleksej Navalnyj – och kanske framför allt han...

↩0 ♥0 ⇄2 •••

@Андрей Черешня❌ 2021-02-28

RT @dw_russian: Алексей Навальный доставлен в исправительную колонию в Покрове Владимирской области. Информацию об этом подтвердили в Общес...

↩0 ♥0 ⇄39 •••

@Лора 2021-02-28

RT @dw_russian: Алексей Навальный доставлен в исправительную колонию в Покрове Владимирской области. Информацию об этом подтвердили в Общес...

↩0 ♥0 ⇄39 •••

This is a footer

Figure 7: Development for the alpha product.

5.1.3 **Evaluation**

When evaluating the alpha product, unstructured interviews were held with Marcin De Kaminski, Security and Innovation Director at CRD, and our supervisor Erik Helldén, Press and communication officer at CRD. The interviewees were set out to test if they could analyze activity regarding the Russian politician Alexei Navalnyj, arrested by the Russian government in January 2021 (Troianovski & Nechepurenko, 2021). The results from the interviews were that they were quickly able to understand the functions in the tool by themselves and that the tool already could help them in their workflow. The most significant pain points were that the search was not real-time, the UI was unpolished, and the tool did not help find suspicious tweets.

Three features were mentioned as future additions to the platform: a user analysis page, a component to find important users, as well as a line chart displaying the top hashtags for every date. This would help with quickly getting insights into the data. A user analysis page would help get insights into specific authors, view their activity and an overview of their data. From the received feedback, we conclude that the Botometer's bot-like score could be implemented on the author page and provide insights without heavily affecting the website's speed. An "important users"-component would make it easier to see if there are any users who are outliers and either view them or exclude them from a search. Line charts displaying hashtags would help to get an overview of the discourse and could help with filtering unwanted tweets.

5.2 The beta product iteration

This section describes the process of investigating, developing, and evaluating the beta version of the application.

5.2.1 **Investigation**

The investigation phase of the beta started by using insights from the evaluation of the alpha, for example, that the UI needed a lot of polishing and that a few components were missing. The features that were to be a part of the beta mockup had to support both the requirements and the insights in the alpha evaluation.

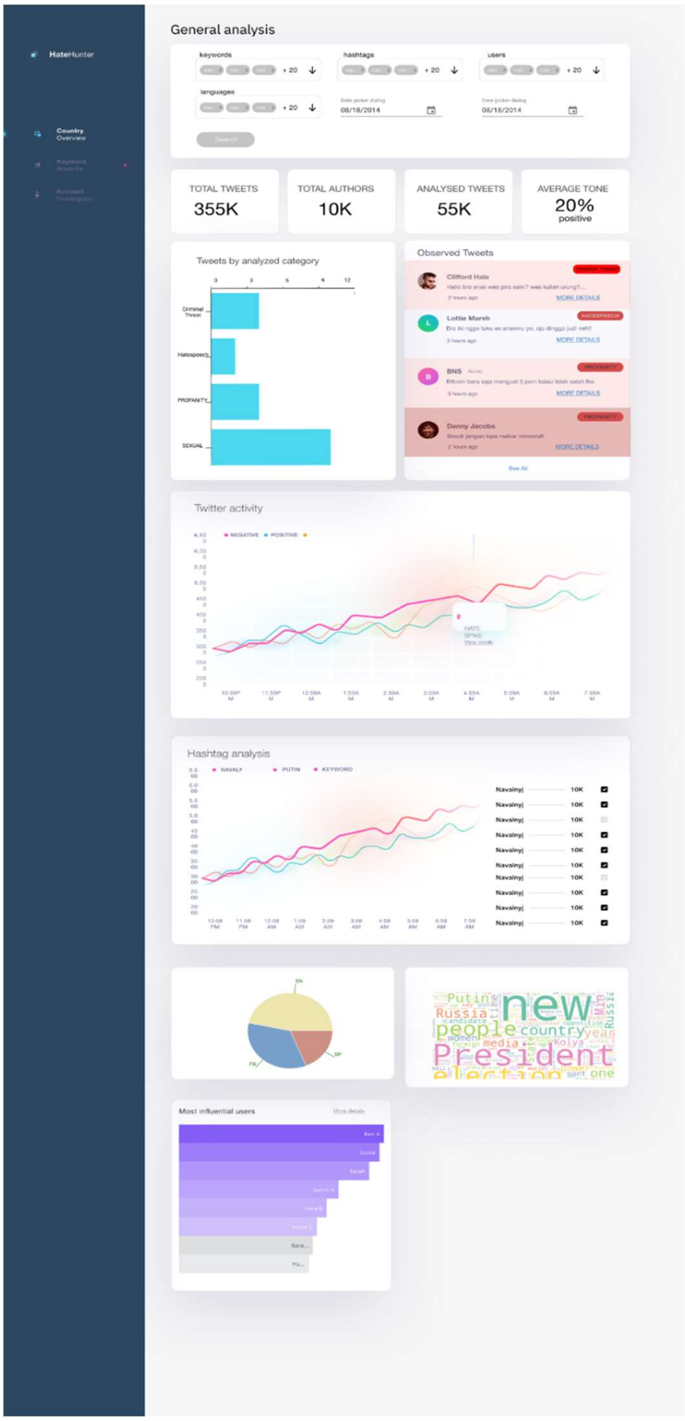
Being able to filter on geo-location is also something many people have asked for. Only 1-2% of tweets on Twitter have a correct geo-location today, which is why this functionality is not yet implemented (Twitter, 2021c). The initial thought was to have a "country"-view, which would be a quick way of getting insights into the political discussions in a country. The idea was to implement this in a similar way as PoliTwi did. They were able to build an algorithm that could automatically find trending topics on Twitter (described further in the investigation phase). However, due to a more complex implementation than expected, as well as Twitters geo-location being insufficient, a "country"-view was not implemented in this project. Instead of a country view, the main view will be a general search view with advanced filtering possibilities.

From the ideation process, a few main components and changes were identified:

The filter menu component was changed, seen at the top of figure 8, to extend the fulfilment of requirement 4 - viewing activity trends. The change added more filters, as well as improved UI. From the alpha development, we learned that it was possible to filter on keywords, hashtags, users, languages, and dates. The mockup incorporated these filters with a more user-friendly design. By filtering on dates, the user would have the possibility to narrow down a search on a specific range of dates where the activity was interesting. By filtering on languages, the user would be able to either search for tweets that the user could understand or filter out languages that are not of interest.

The hashtag trendline component was designed, seen at the bottom of figure 8, to meet the request from the alpha evaluation. The component displays the top 10 hashtags and the activity of the hashtag in a chart. The component would have a list with the top 10 hashtags on the right side, displaying how many times the hashtag had been used in tweets that matched the filtering, and a line chart on the left side would display a trendline for each hashtag. The user would be able to click on a checkbox next to a hashtag to remove that hashtag's trendline from the chart. This would help the user to filter if there were only a few hashtags that were of interest.

A problem with the previous version was that the tool did not help find suspicious tweets. To counter this, a natural language processing service was added. Tisane Labs, which is described further in the investigation phase, was added to the data pipeline. By using Tisane Lab's text analyzing service, it was possible to get a sentiment score on each tweet and thereby label the text with abusive content. The analytics results can be seen in two places: first, the component "Tweets by analyzed category" was added, seen in the middle left of figure 8. The component is displaying how many tweets there are that have been labelled in each category. In addition, resulting labels can be seen in the "Observed tweets" component, seen in the middle-right of figure 8. Each tweet has a colour depending on the sentiment analysis, where red is negative and green is positive. In the tweets, there is also a label if the tweet was labelled as having abusive content.



Clifford Hale
Hallo bro anak wes piro saiki? wes kuliah urung?...
2 hours ago

Lottie Marsh
Bro iki nggo tuku es anakmu yo, ojo dinggo judi neh!!
3 hours ago

BNS News
Bitcoin baru saja menguat 5 poin kalau tidak salah tho
3 hours ago

Danny Jacobs
Besok jangan lupa mabar minecraft
2 hours ago

Navalnyj	10K	☑
Navalnyj	10K	☑
Navalnyj	10K	□
Navalnyj	10K	☑
Navalnyj	10K	☑
Navalnyj	10K	☑
Navalnyj	10K	□
Navalnyj	10K	☑
Navalnyj	10K	☑
Navalnyj	10K	☑

Figure 8: Mockups for the general analysis page for the beta product.

5.2.2 **Development**

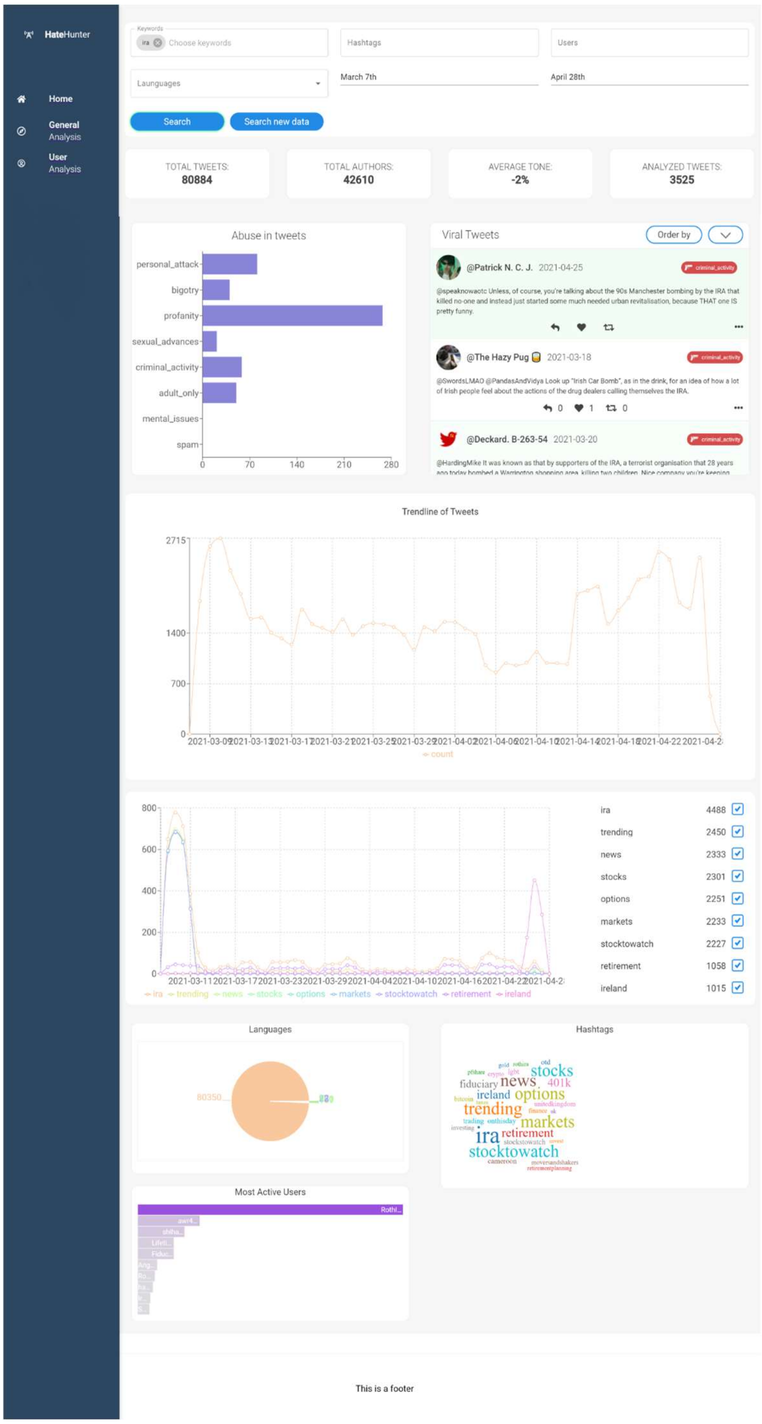
The development phase for the beta release focused on adding natural language processing, polishing the UI and adding components from the mockups. In this development phase, a lot of effort was put into adding natural language processing into the data pipeline, as well as adding the user page. The user page reused many components from the general search view, which saved significant development time.

Components that were added and can be seen in figure 9:

- Data overview of data points, which are total number of tweets, total number of authors, total number of analyzed tweets and average tone which is calculated by doing an average of all tweets sentiment score and range from -100% to +100%.
- Hashtag graph showing a trendline for the top 8 hashtags and the total number of the hashtag.
- Bar chart showing how many tweets that had been labelled by an abuse category, grouped by each abuse category.
- Abuse label on each tweet that had been labelled with abusive content.
- Sentiment score on each tweet, shown by a red-green color scale.
- Most active users chart. Instead of using the page-rank algorithm from neo4j, the chart is using total number of tweets from the user.

Views that were added:

- The user analysis view, consisting of a user overview component, that can be seen on the top of figure 10. It displays the author's name, description, profile picture, number of tweets, followers and following. The bot score from the Botometer API was also added to this component. Underneath the user overview, a component displays all the user's tweets, a chart of abuses found in the user's tweets, and a trendline of the user's tweets.



Keywords: ira

March 7th

April 28th

TOTAL TWEETS: **80884**

TOTAL AUTHORS: **42610**

AVERAGE TONE: **-2%**

ANALYZED TWEETS: **3525**

@Patrick N. C. J. 2021-04-25 criminal_activity

@speaknowaotc Unless, of course, you're talking about the 90s Manchester bombing by the IRA that killed no-one and instead just started some much needed urban revitalisation, because THAT one IS pretty funny.

@The Hazy Pug 🥃 2021-03-18 criminal_activity

@SwordsLMAO @PandasAndVidya Look up "Irish Car Bomb", as in the drink, for an idea of how a lot of Irish people feel about the actions of the drug dealers calling themselves the IRA.

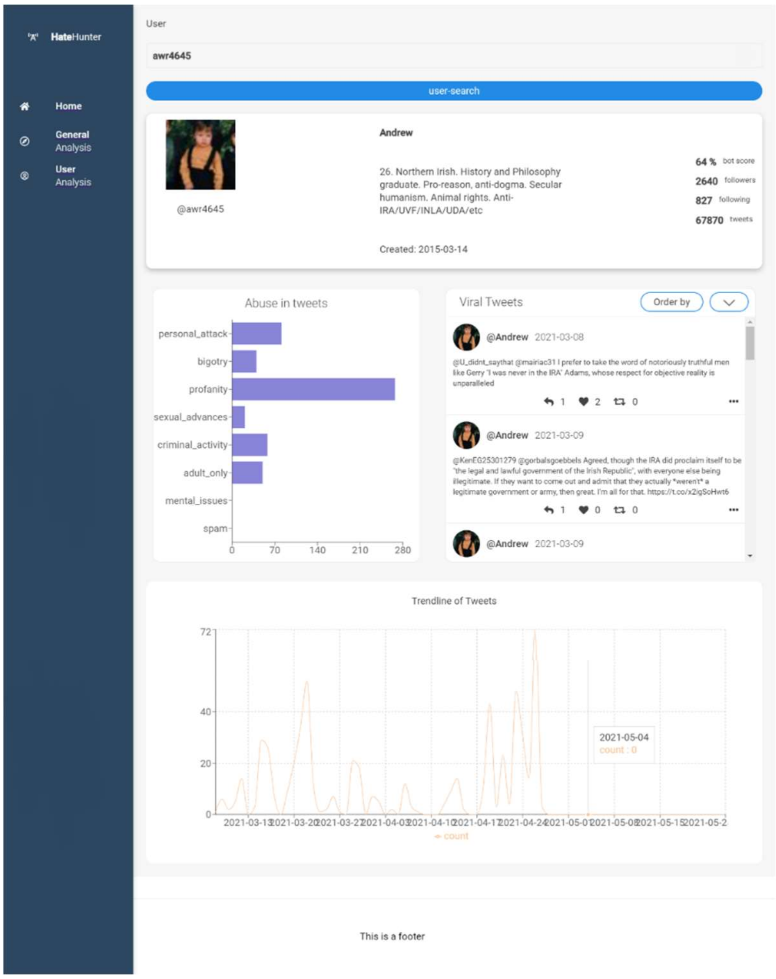
0 1 0

@Deckard. B-263-54 2021-03-20 criminal_activity

@HardingMike It was known as that by supporters of the IRA, a terrorist organisation that 28 years ago today bombed a Warrington shopping area, killing two children. Nice company you're keeping

ira	4488	☑
trending	2450	☑
news	2333	☑
stocks	2301	☑
options	2251	☑
markets	2233	☑
stocktowatch	2227	☑
retirement	1058	☑
ireland	1015	☑

Figure 9: Implementation of the general analysis page for the beta product.



awr4645

Andrew

26. Northern Irish. History and Philosophy graduate. Pro-reason, anti-dogma. Secular humanism. Animal rights. Anti-IRA/UVF/INLA/UDA/etc

@awr4645

64 % bot score
2640 followers
827 following
67870 tweets

Created: 2015-03-14

@Andrew 2021-03-08

@U_didnt_saythat @mairiac31 I prefer to take the word of notoriously truthful men like Gerry 'I was never in the IRA' Adams, whose respect for objective reality is unparalleled

1 2 0

@Andrew 2021-03-09

@KenEG25301279 @gorbalsgoebbels Agreed, though the IRA did proclaim itself to be "the legal and lawful government of the Irish Republic", with everyone else being illegitimate. If they want to come out and admit that they actually *weren't* a legitimate government or army, then great. I'm all for that. https://t.co/x2igSoHwt6

1 0 0

@Andrew 2021-03-09

Figure 10: Implementation of the user analysis page for the beta product.

5.2.3 **Evaluation**

The evaluation of the beta version consists of semi-unstructured interviews following the protocol in Appendix B. The objective of the evaluation is to ensure that the product can solve basic data analysis tasks in an easy and interpretable way. Key functionalities to improve the product are also looked into.

Interviews were held with Christoffer Kullenberg, Senior Lecturer in Theory of Science at the University of Gothenburg, and Simon Lindgren, Professor of Sociology at the University of Umeå and director of DIGSUM, Centre for Digital Social Research, whom both have experience analyzing Twitter data in a social science aspect as well as Marcin De Kaminski and Erik Helldén at CRD (Kullenberg, 2021b)(Lindgren, 2021)(De Kaminski, 2021)(Helldén, 2021). The interviewee will analyze tweets with the keyword “IRA”, a shortening for the paramilitary group Irish Republican Army, 15th of April and 14th of May, 2021, regarding the violent riots occurring in Northern Ireland at the beginning of April 2021 (Hirst, 2021). More background information about the specific use case can be found in Appendix B.6.

When evaluating the beta product, the test persons had overall good experiences when carrying out the tasks in Appendix B.7. The beta product worked well to get an overview of the dataset and was well suited for the usual tasks at the beginning of analyzing tweets from Twitter. The product still lacked some desirable functionality to construct a more profound analysis for an eventual report about the use case, and the preciseness of the text analysis and user-account analysis was questioned. Extended filtering functionality was requested, such as the possibility to exclude keywords and users. The idea behind it regarding labeling tweets as for example, “criminal attack” and “personal attack” was very appreciated.

Regarding the user-friendliness of the application, all of the interviewees found the application easy to use. Kullenberg even though the application was so easy to use that he wanted to incorporate it in his research course where his students would analyze Twitter data. (Kullenberg, 2021b)

In academic research, as well as CRD’s second use case where the analysis will be the backbone for a journalistic report, some functionalities regarding traceability were wanted. The idea of being able to download the dataset was well received even though it was not yet implemented. However, more information about the analysis fundamentals will be needed, for instance, a clearer understanding of how the text analysis works.

For the CRD’s first use case, which focuses on real-time detection of the threat picture against an organization, the product seemed promising but still lacked some desired functionalities. For example, the data pipeline that analyzes tweets still has room for improvement regarding speed, where each tweet takes roughly 20 seconds to fetch and analyze. For CRD's first use case, aspects like traceability and information about underlying technologies are not important.

5.3 The final version iteration

This section describes the process of investigating and developing the final version of the application.

5.3.1 **Investigation**

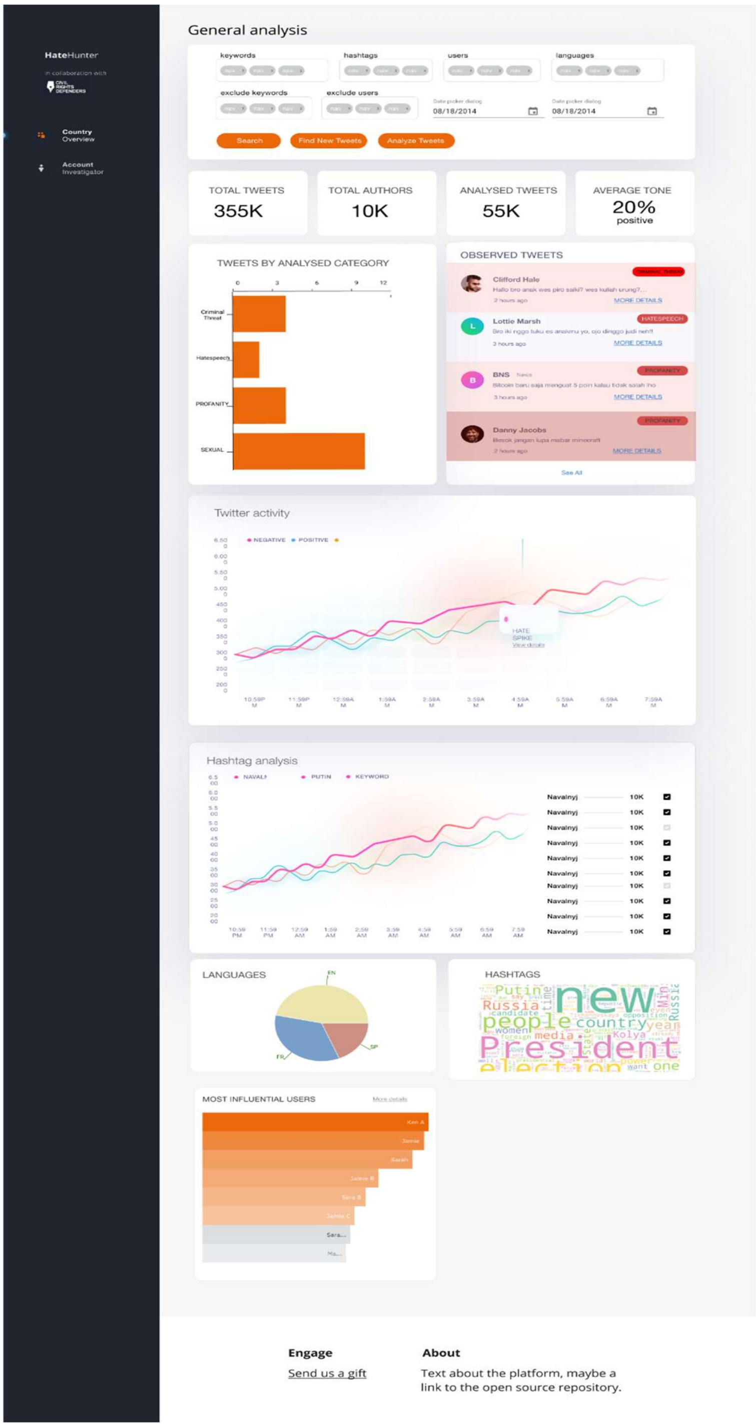
Similar to the investigation phase for the beta product, the investigation phase for the final version started by using previous insights. The main insights from the beta evaluation were that the filter possibilities needed to be extended, as well as more insight into how the natural language processing worked. We discovered during the interviews that a big issue in the application was that it was not possible to do real-time searches. Due to technical constraints of the applications data pipeline, it was not possible to implement real-time searches within the project timeline.

From these insights, these components were created/updated and can be seen in the final mockup in figure 11:

Three search buttons instead of one. By allowing the user to choose between searching the database, updating the data with new data from Twitter, and analyzing the data, the limitations of not having real-time search could be countered.

Exclude filters were added. This gives the user more possibilities around narrowing down a search to display only the most critical data. During our interviews with Christopher Kullenberg and Simon Lindgren, one big improvement point that came up was that the application did not have enough functionality to dig deep into the data and view the important data for a specific use case.

The design was updated to use Civil Rights Defenders branding. Color scale was updated to match Civil Rights Defenders brand colors, and Civil Rights Defenders logo was added to the sidebar and the footer.



General analysis

keywords

hashtags

users

languages

exclude keywords

exclude users

Date picker dialog 08/18/2014

Date picker dialog 08/18/2014

TOTAL TWEETS 355K

TOTAL AUTHORS 10K

ANALYSED TWEETS 55K

AVERAGE TONE 20% positive

OBSERVED TWEETS

Clifford Hale — CRIMINAL THREAT
Hallo bro ansk wes piro saiki? wes kuliah urung?...
2 hours ago MORE DETAILS

Lottie Marsh — HATESPEECH
Bro iki nggo tuku es ansimu yo, ojo dinggo judi neh!!
3 hours ago MORE DETAILS

BNS News — PROFANITY
Bitcoin baru saja menguat 5 poin kalau tidak salah iho
3 hours ago MORE DETAILS

Danny Jacobs — PROFANITY
Besok jangan lupa mabar minecraft
2 hours ago MORE DETAILS

See All

Navalnyj	10K	☑
Navalnyj	10K	☑
Navalnyj	10K	□
Navalnyj	10K	☑
Navalnyj	10K	☑
Navalnyj	10K	☑
Navalnyj	10K	□
Navalnyj	10K	☑
Navalnyj	10K	☑
Navalnyj	10K	☑

Engage

Send us a gift

About

Text about the platform, maybe a link to the open source repository.

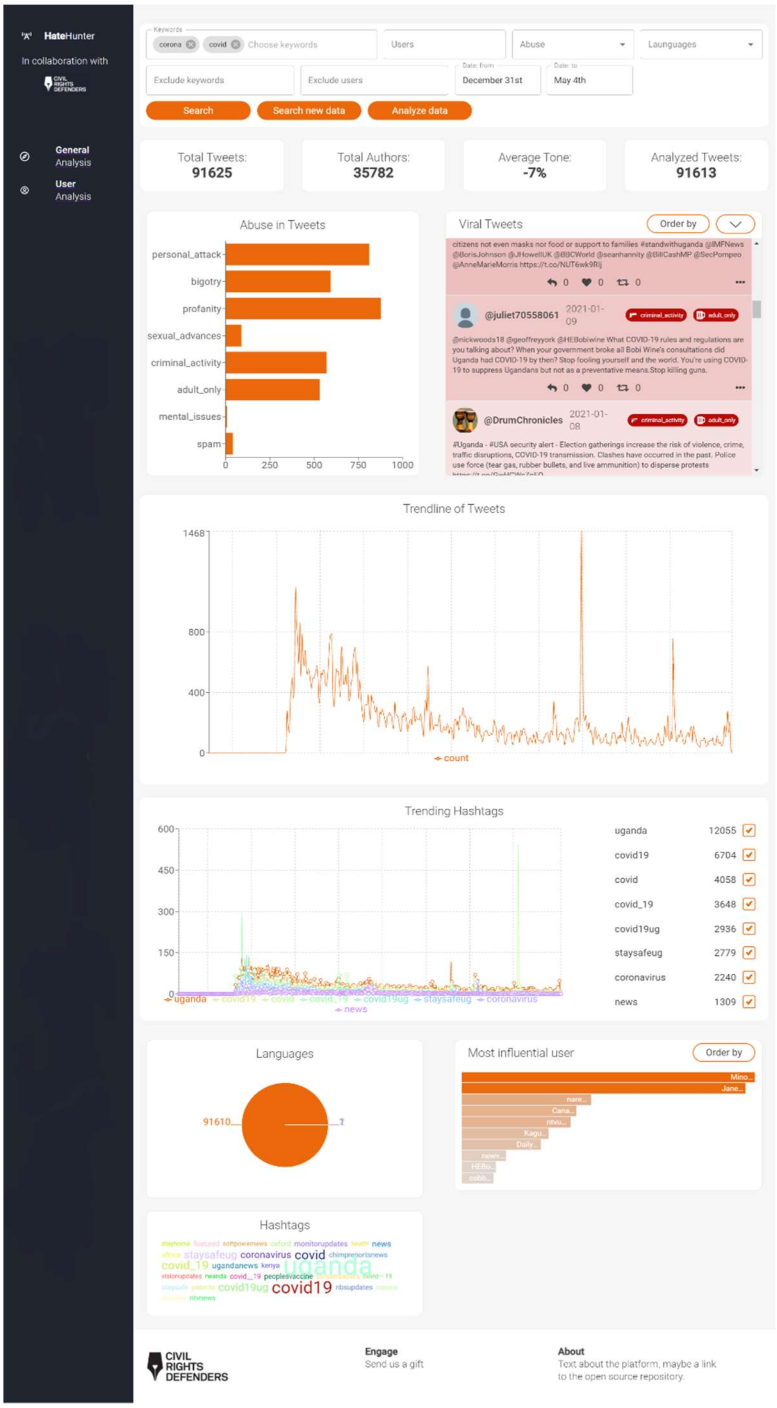
Figure 11: Final mockup.

5.3.2 **Development**

The development phase for the final version added all components from the investigation phase and stability updates for the final user tests. A significant part of the development time was spent on performance updates and bug fixes. The time to fetch and analyze tweets was reduced significantly. In the beta version, fetching 35.000 tweets took an average of 20 hours. In the final version, that time was reduced to less than 10%, and it took an average of 2 hours to fetch 35.000 tweets. However, this is still far from the theoretical maximum of 600.000 tweets per hour, which is Twitters API rate limit, and further optimizations can be done. The average time to analyze a tweet was also reduced from 20 seconds per tweet to about 0.8 seconds per tweet by removing unnecessary analytics from the API, such as hashtag lookup. The response time from Tisane's API was 0.2 seconds, which means that the pipeline of receiving the data to adding it to the database took 0.6 seconds per tweet. This can be optimized further to reach the Tisane Lab' API's limit of 1200 tweets per minute.

Components that were added:

- "Search", "Find new tweets", and "Analyze tweets" buttons were added to the search component, as well as backend functionality supporting the buttons, seen in figure 12.
- Exclude filters were added to the search component, with options for excluding keywords, hashtags, and users. The backend was also updated to support the exclusion, seen in figure 12.
- The design was updated with new colors and the Civil Rights Defenders logo, and a footer at the bottom of the application, seen in both figure 12 and figure 13.
- Improved performance of fetching tweets from Twitter and analyzing.



Date: from: December 31st

Date: to: May 4th

Total Tweets: **91625**

Total Authors: **35782**

Average Tone: **-7%**

Analyzed Tweets: **91613**

citizens not even masks nor food or support to families #standwithuganda @IMFNews @BorisJohnson @JHowellUK @BBCWorld @seanhannity @BillCashMP @SecPompeo @AnneMarieMorris https://t.co/NUT6wk9Rij

0 0 0

@juliet70558061 2021-01-09 criminal_activity adult_only

@nickwoods18 @geoffreyyork @HEBobiwine What COVID-19 rules and regulations are you talking about? When your government broke all Bobi Wine's consultations did Uganda had COVID-19 by then? Stop fooling yourself and the world. You're using COVID-19 to suppress Ugandans but not as a preventative means.Stop killing guns.

0 0 0

@DrumChronicles 2021-01-08 criminal_activity adult_only

#Uganda - #USA security alert - Election gatherings increase the risk of violence, crime, traffic disruptions, COVID-19 transmission. Clashes have occurred in the past. Police use force (tear gas, rubber bullets, and live ammunition) to disperse protests

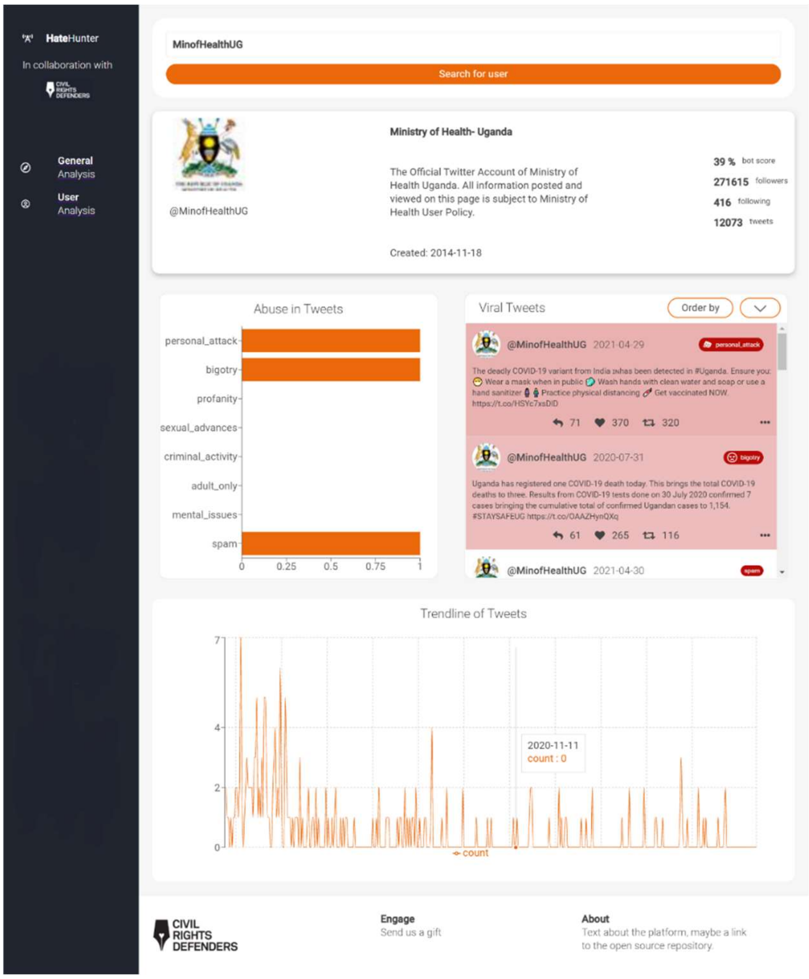
uganda	12055	☑
covid19	6704	☑
covid	4058	☑
covid_19	3648	☑
covid19ug	2936	☑
staysafeug	2779	☑
coronavirus	2240	☑
news	1309	☑

CIVIL RIGHTS DEFENDERS

Engage
Send us a gift

About
Text about the platform, maybe a link to the open source repository.

Figure 12: Implementation of the general analysis page for the final version of the product.



The Official Twitter Account of Ministry of Health Uganda. All information posted and viewed on this page is subject to Ministry of Health User Policy.

The deadly COVID-19 variant from India ₁₈has been detected in #Uganda. Ensure you: 😷 Wear a mask when in public 🧼 Wash hands with clean water and soap or use a hand sanitizer 🧍‍♀️🧍 Practice physical distancing 💉 Get vaccinated NOW. https://t.co/HSYc7xsDID

Uganda has registered one COVID-19 death today. This brings the total COVID-19 deaths to three. Results from COVID-19 tests done on 30 July 2020 confirmed 7 cases bringing the cumulative total of confirmed Ugandan cases to 1,154. #STAYSAFEUG https://t.co/OAAZHynQXq

Text about the platform, maybe a link to the open source repository.

Figure 13: Implementation of the userpage for the final version of the product.

6 Evaluation Phase

This chapter describes the final product evaluation, which was based on interviews with CRD employees.

6.1 Test participants

The assessment test was held with four test participants. The aim was to get participants that matched the user personas described in section 4.2. The roles of the test participants matched the user personas, with roles such as human rights lawyer, security expert and communication specialists being among the four test participants.

6.2 Assessment test

To evaluate the usability of the application and the product/market fit, an assessment test was conducted where the participants were asked to execute a set of tasks. The whole interview format can be seen in Appendix C. As a basis for the test design. The following research questions were set to be answered.

1. How easily can the user get an overview of key metrics?
2. How easily can the user get insights from the text analysis?
3. How easily can the user find twitter activity trends?
4. How easily can the user find trending hashtags?
5. How easily can the user find influential users and analyze them?
6. How easily can the user search for keywords and analyze the activity?

The tests were conducted remotely through Zoom, with meetings that were 20 minutes. Each participant was first given a background to the project. Then an onboarding was given where the application was given a brief introduction.

6.2.1 **Test case**

6.2.1.1 Introduction

At the beginning of 2020, covid-19 started to spread around the world, causing a global pandemic. From the start and continuously over the pandemic, social media has been filled with misinformation. During the election in Uganda in January 2021, Facebook shut down several government-linked accounts proven to be fake or duplicated to manipulate opinions on Facebook by impersonating users and boost specific posts (Wakefield, 2021). With this recent information about government manipulation on social media as well as their new anti-human rights law (Amnesty UK, 2020), investigating how the covid-19 outbreak affects the discussion on Twitter in Uganda is interesting.

To analyze the covid-19 discussion in regards to Uganda on Twitter, English tweets containing "Uganda & Covid" and "Uganda & Corona" were fetched from the 31/12-2019, when World Health Organization (WHO) picked up a media statement of 'viral pneumonia' in Wuhan, People's Republic of China (WHO, 2020), to the 2/5-2021.

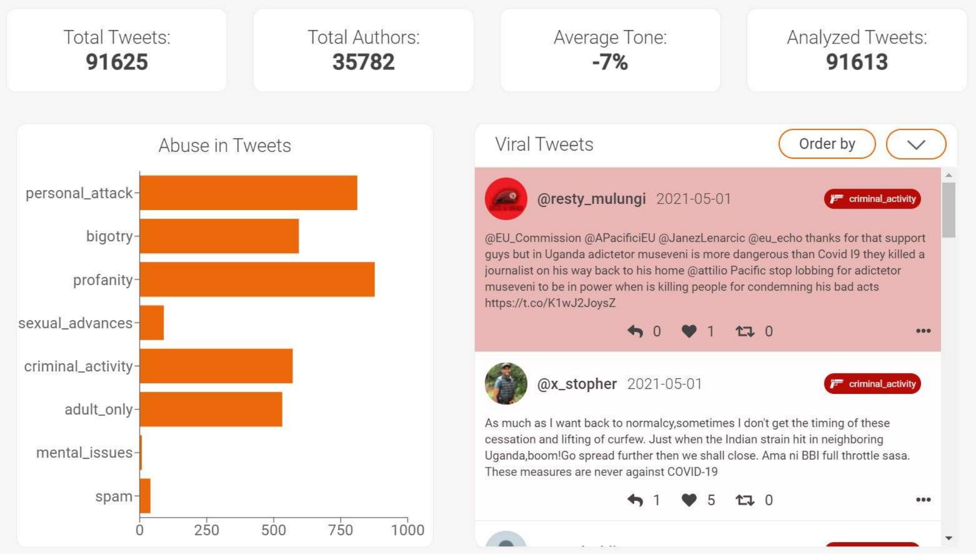
6.2.1.2 Tasks to perform

The participant will execute the following tasks:

1. With the initial search with "Covid" & "Corona" as keywords over the given timeframe, how many tweets does the dataset contain?
2. How many of these tweets were labelled as a criminal activity?
3. Which day had the most tweets?
4. Which were the top 3 hashtags for the specific keyword "Covid" & "Corona" during the given timeframe?
5. Find the most influential user. Does this person have any tweets classified as bigotry?
6. Search for the keyword "Vaccine" and determine the average tone for "Vaccine" in Uganda.
7. With the search for "Vaccine" over the given timeframe, is there any day or period with significantly more tweets? What has generated this spike?

6.2.1.3 Correct answers on the tasks

The correct answers on the assessments test, is shown in figures 14 to 22.



Total Tweets: **91625**

Total Authors: **35782**

Average Tone: **-7%**

Analyzed Tweets: **91613**

Viral Tweets

@resty_mulungi 2021-05-01 criminal_activity

@EU_Commission @APacificiEU @JanezLenarcic @eu_echo thanks for that support guys but in Uganda adictetor museveni is more dangerous than Covid I9 they killed a journalist on his way back to his home @attilio Pacific stop lobbing for adictetor museveni to be in power when is killing people for condemning his bad acts https://t.co/K1wJ2JoysZ

0 1 0

@x_stopher 2021-05-01 criminal_activity

As much as I want back to normalcy,sometimes I don't get the timing of these cessation and lifting of curfew. Just when the Indian strain hit in neighboring Uganda,boom!Go spread further then we shall close. Ama ni BBI full throttle sasa. These measures are never against COVID-19

1 5 0

Figure 14: Overview over the key metrics for the Uganda covid-19 dataset.

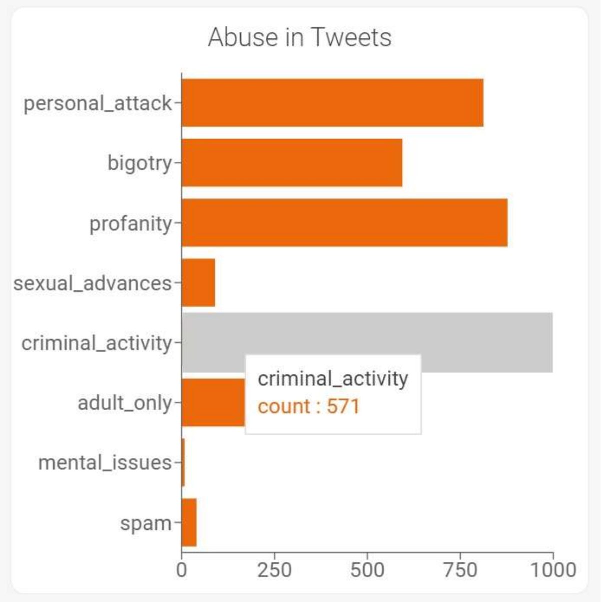


Figure 15: A zoomed-in view the number of tweets labeled with a specific abuse label.

Figure 14 and 15, which provides an overview of the Uganda Covid-19 dataset, can identify the answer for the first and second assessment questions. One can read in figure 14's top-left corner that the total tweets of the dataset are 91625 and in Figure 15 that 571 tweets are labelled as criminal activity when hovering the correct bar in the abuse graph. Figure 14 also provides data of how many authors there are for the tweets and the average tone of the dataset. The average tone ranges from -100% to +100%, where -100% is very negative, and +100% is very positive. The viral tweet component in figure 14 also provides the user to check specific tweets in the dataset manually.

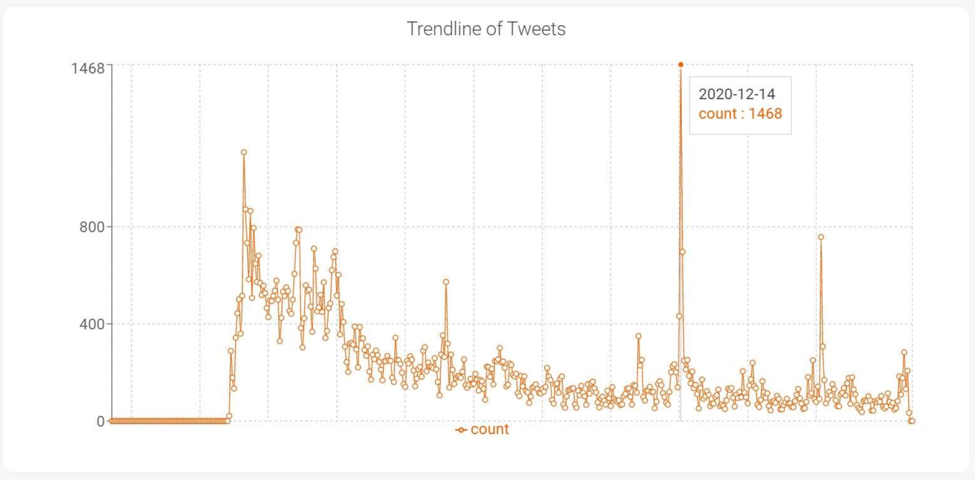
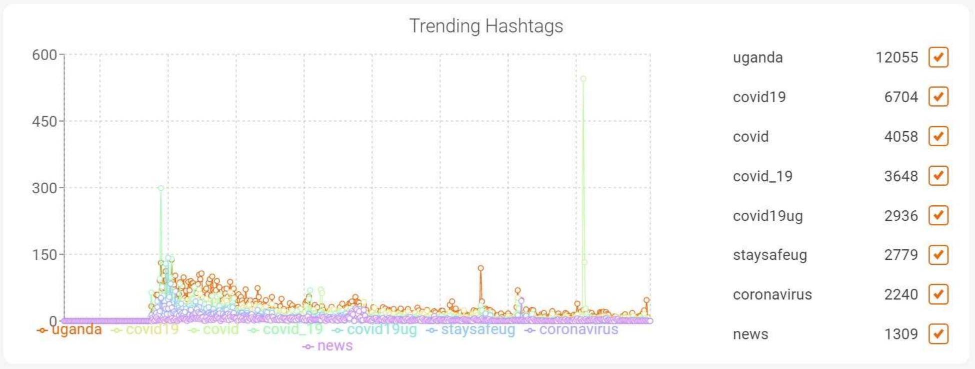


Figure 16: A trendline over the tweets in the dataset.

A trendline of the tweets in the dataset can be seen in figure 16. The result for question 3 is shown as a spike, and by hovering the specific day of the spike, the user can read that on the 14/12-2020, 1468 were tweeted.



uganda	12055	☑
covid19	6704	☑
covid	4058	☑
covid_19	3648	☑
covid19ug	2936	☑
staysafeug	2779	☑
coronavirus	2240	☑
news	1309	☑

Figure 17: Overview of the top 8 hashtags for the given dataset.

The answer for question 4 can be found in figure 17, which shows the trending hashtags for the dataset. The answer is the 3 top hashtags on the right, Uganda with 12055 tweets, Covid19 with 6704, and covid with 4058. By unclicking the checkbox on the right, the given hashtag will be removed from the graph.

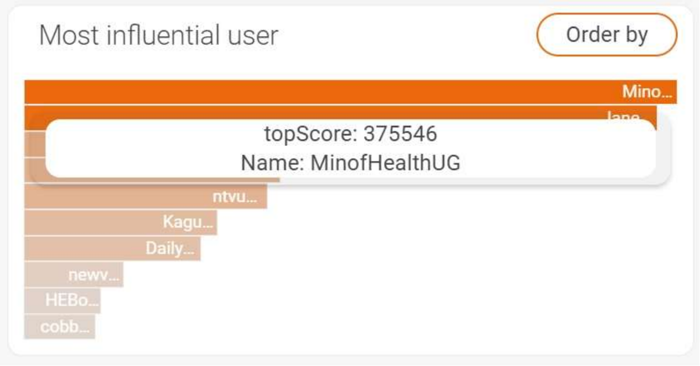
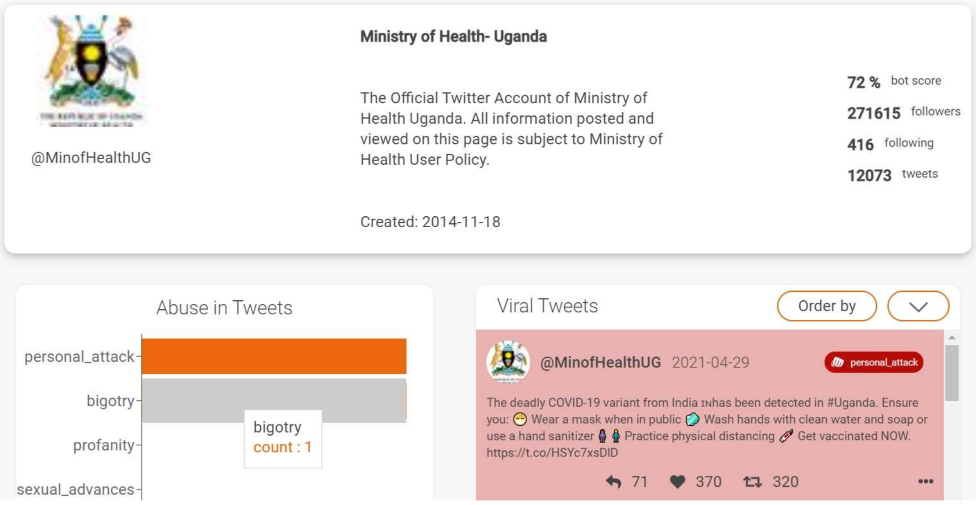


Figure 18: Visualization of the most influential users.


Ministry of Health- Uganda

The Official Twitter Account of Ministry of Health Uganda. All information posted and viewed on this page is subject to Ministry of Health User Policy.

@MinofHealthUG

Created: 2014-11-18

72 % bot score
271615 followers
416 following
12073 tweets

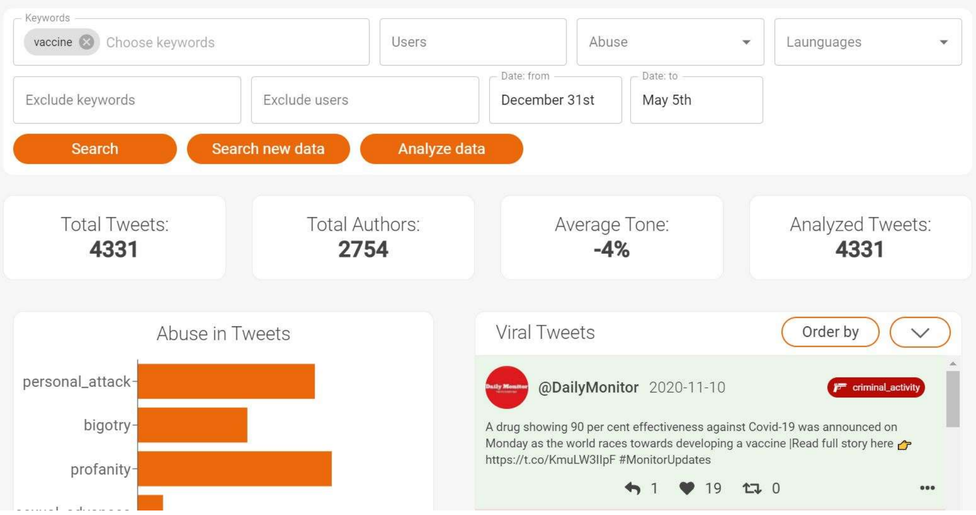
@MinofHealthUG 2021-04-29 personal_attack

The deadly COVID-19 variant from India 🇮🇳has been detected in #Uganda. Ensure you: 😷 Wear a mask when in public 🧼 Wash hands with clean water and soap or use a hand sanitizer 🧴🧴 Practice physical distancing 💉 Get vaccinated NOW. https://t.co/HSYc7xsDID

71 ♥ 370 320

Figure 19: Overview of the Ministry of Health- Uganda userpage, calculated as the most influential user by the previously mentioned algorithm.

Figure 18 and 19 provides the answer to question 5 in the assessment test. This by first finding the most influential user in figure 18. If clicking on the top bar, the user will be redirected to a page that provides an overview for that specific account. This page can be seen in figure 19 and give the user further information about the chosen account. For example, information about how accounts the user follows and get followed by or how bot-like the account's appearance is. This is a score calculated by the Botometer API. One can also find information about how many of the accounts tweets are labelled as bigotry, which can be found on the bottom left corner in the "Abuse in Tweets" graph.



Keywords: vaccine

Date: from: December 31st

Date: to: May 5th

Total Tweets: 4331

Total Authors: 2754

Average Tone: -4%

Analyzed Tweets: 4331

@DailyMonitor 2020-11-10

A drug showing 90 per cent effectiveness against Covid-19 was announced on Monday as the world races towards developing a vaccine |Read full story here https://t.co/KmuLW3IlpF #MonitorUpdates

Figure 20: An overview for the keywords search "vaccine".

Figure 20 is an overview for the new keywords search "vaccine". This figure provides the answer for question 6 in the assessment test by identifying the average tone as -4%. One can see that the average tone for the keyword search vaccine is slightly more positive than the average tone of the whole dataset. However, the overall tweets are still somewhat negative.

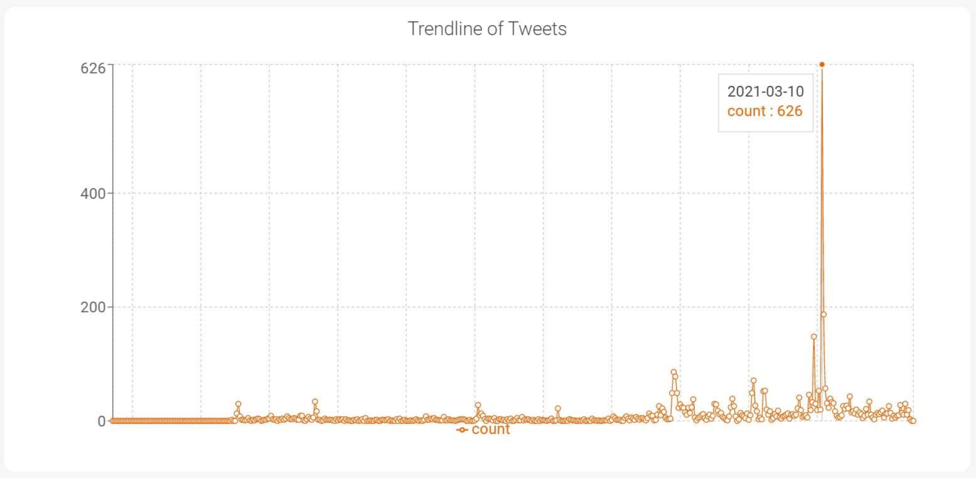


Figure 21: Trendline over the tweets containing the keyword "vaccine" in the given dataset for covid-19 in Uganda.

One can identify a significant spike of tweets in figure 21 on the 10/3-2021. On this day, the number of tweets with the keyword "vaccine" from the original dataset was 626 compared to no other day receiving more than 200 tweets. This spike provides half of the answer to question 7. The second part of question 7's answer can be found in the trending hashtag graph shown in figure 22. Here one can see that new hashtags are trending for the keyword search “vaccine” compared to the original search of “covid” and “corona”. Peoplevaccine, oxford and patents are trending, and when hovering the day 10/3-2021, one can see that those hashtags were especially trending that day which is the answer to the second part of question 7 in the assessment test.

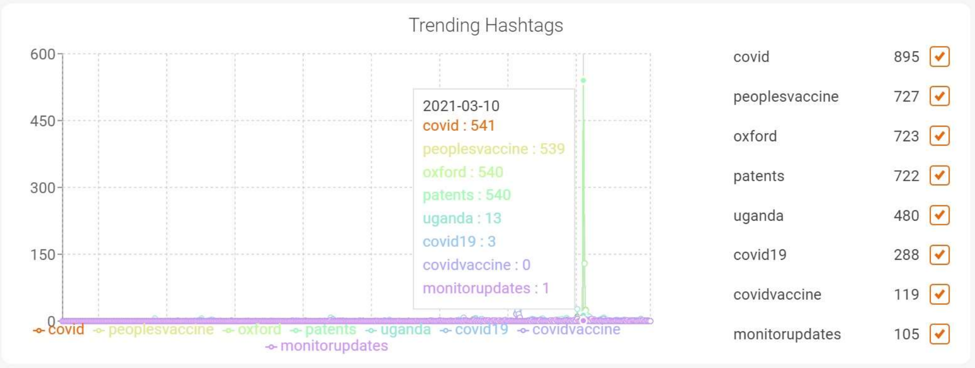


Figure 22: Visualization of the trending hashtags for the keyword search “vaccine”.

6.2.2 **Assessment test results**

6.2.2.1 How easily can the user get an overview of key metrics?

All participants found it easy and user friendly to get an overview of key Twitter activity metrics.

Table 1: The success rate for the task: With the initial search with “Covid” & “Corona” as keywords over the given timeframe, how many tweets does the dataset contain?

Participant	1	2	3	4
Task success	Yes	Yes	Yes	Yes

6.2.2.2 How easily can the user get insights from the text analysis?

Some test participants found it very easy to get insights from the text analysis, while some had minor issues. Two of the participants did not immediately connect the “abuse graph” to the question. When they had found it, one participant did not understand that it was possible to hover over the bar to see the exact count and therefore received a hint to hover the abuse graph to see the exact number.

One participant immediately understood that it was possible to filter abuses in the search menu and filter for “criminal activity” and then see the total number of tweets. This was not the intended solution but a perfect solution to the problem that shows that the filter menu is understandable.

Table 2: The success rate for the task: How many of these tweets were labelled as a criminal activity?

Participant	1	2	3	4
Task success	Yes	With hint	Yes	With hint

6.2.2.3 How easily can the user find twitter activity trends?

The participants were asked to find which day had the most tweets. All of the users quickly completed the task. One of the participants had some issues hovering over the correct date of the peak. The timeline had too many data points in a small space, making it difficult to point to a specific date. The factor of the remote controlling could have affected the result of the task negatively, as the participants lost some amount of accuracy when remote controlling instead of using their computer.

Table 3: The success rate for the task: Which day had the most tweets?

Participant	1	2	3	4
Task success	Yes	Yes	Yes	Yes

6.2.2.4 How easily can the user find trending hashtags?

The participants were asked to find the top 3 hashtags for the current search. All of the users could quickly find the trending hashtags using the hashtag graph. Some users had to scroll around the page for a few seconds before knowing what to look for, but once they had found the hashtag graph components, everyone knew what to do. None of the users used the hashtag word cloud component.

Table 4: The success rate for the task: Which were the top 3 hashtags for the specific keyword “Covid” & “Corona” during the given timeframe?

Participant	1	2	3	4
Task success	Yes	Yes	Yes	Yes

6.2.2.5 How easily can the user find important users and analyze them?

The participants were asked first to find the most influential user and then see if that person had any tweets classified as bigotry. All of the users succeeded with the tasks.

However, two of the participants needed a few hints. Every user could quickly find the component with the most influential users, and all of them could see who the most influential user was in the component. One participant did not immediately understand that it was possible to click on the user to get a more detailed view of the user data. Another participant was confused after being redirected to a new page and thought something had gone wrong.

Once the participants were on the user page, everyone quickly understood how to see if the users had any tweets classified as bigotry.

Table 5: The success rate for the task: Find the most influential user. Does this person have any tweets classified as bigotry?

Participant	1	2	3	4
Task success	Yes	Yes	Yes	Yes

6.2.2.6 How easily can the user search for keywords and analyze the activity?

The participants were asked to do two sets of tasks:

1. Change the search to a new keyword, “Vaccine,” and determine the average tone.
2. With the new search results, see if there were any spikes in the Twitter activity and what had generated the spike.

Everyone could do the first task, but everyone had issues with pressing “enter” after typing a keyword. Some were confused by the three search buttons and did not understand that they had to click “search” or which one of the buttons they had to click.

Table 6: The success rate for the task: Search for the keyword “Vaccine” and determine the average tone for “Vaccine” in Uganda.

Participant	1	2	3	4
Task success	Yes	With hint	With hint	Yes

Everyone could identify the spike in the trendline chart. Most participants could also directly identify that the hashtag graph contained the information needed to see what had generated the spike. However, everyone had issues with hovering over the correct day to get the trending hashtags for the specific day. Identifying the trending hashtags would lead a part of the solution but the intended idea of filtering the search to only receive tweets from the specific day of the spike and manually read a few tweets from the tweet view to understand what generated the spike were not succeeded by anyone. One user tried looking at the tweet view, but without filtering the dates, and therefore figured out that it would be difficult to get an overview of several hundreds of tweets by looking at the tweets one-by-one. All the participants received a hint to filter down the search and manually read a few tweets to understand the spike. The participants who struggled to identify that the hashtag graph could give valuable insights on what caused the spike were given a hint to look there as well.

Table 7: The success rate for the task: With the search for "Vaccine" over the given timeframe, is there any day or period with significantly more tweets? What has generated this spike?

Participant	1	2	3	4
Task success	With hint	With hint	With hint	With hint

6.3 System Usability Scale Score

As mentioned in the background, the SUS score based on the test persons' answer to the questions:

1. I think that I would like to use this application frequently

2. I found the application unnecessarily complex

3. I thought the application was easy to use

4. I think that I would need the support of a technical person to be able to use this application

5. I found various functions in this application were well-integrated

6. I thought there was too much inconsistency in this application

7. I would imagine that most people would learn to use this application very quickly

8. I found the application very cumbersome to use

9. I felt very confident using the application

10. I needed to learn a lot of things before I could get going with this application

In figure 23, one can see each participant's answer to each question. Overall, all participants answer very high on each question. The only question with a bit lower result was the first question, "I think that I would like to use this application frequently ". Question 2 and 3 received the highest score by each of the participants.

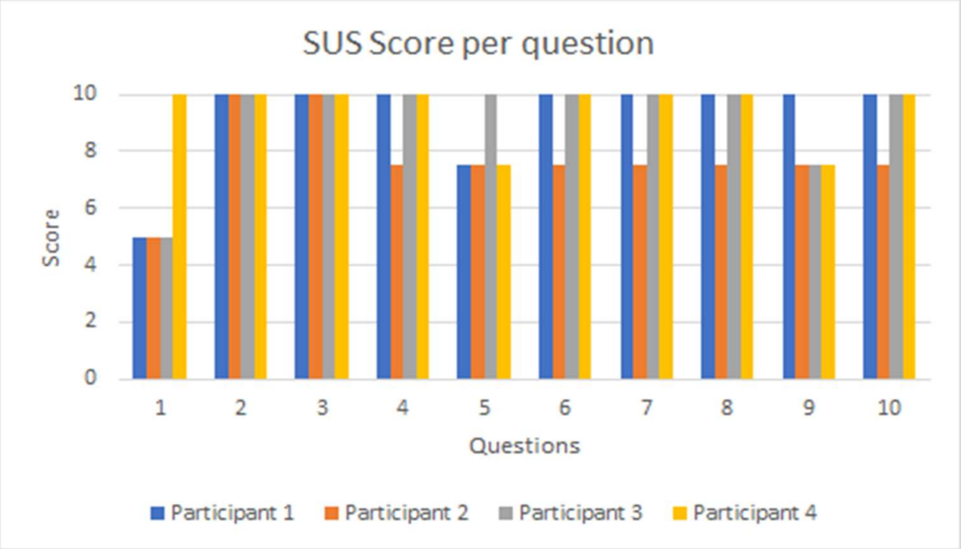


Figure 23: SUS Score per specific question.

The overall SUS score is shown in figure 24. One can see that all participants received a score over the acceptable score of 70, which speaks for the product being user-friendly and well implemented. Three participants scored a value over 90, which advocates that the product was very well received by the potential users. The SUS score is in line with the results from the user tests, where everyone completed all tasks without any or little effort.

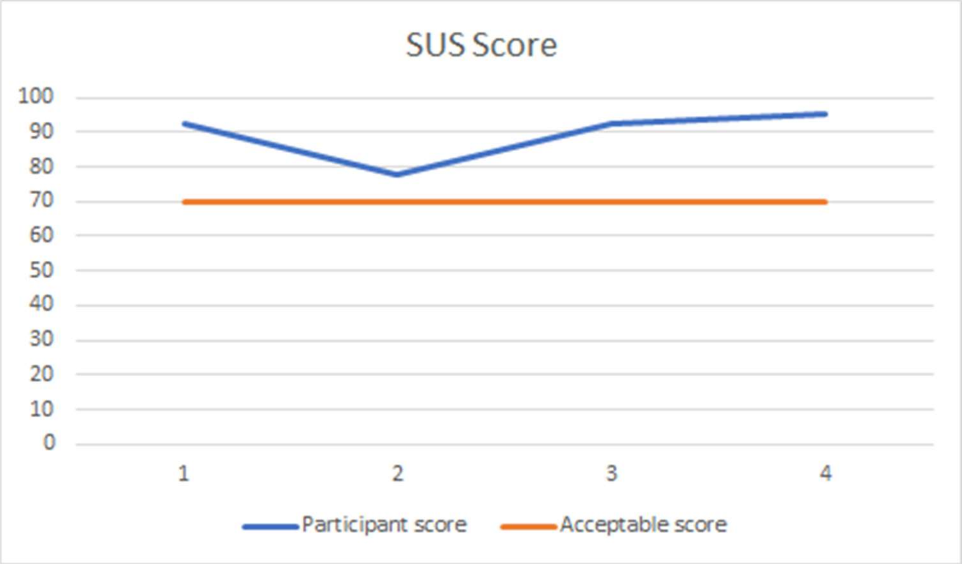


Figure 24: SUS score for each test person. Values over the orange horizontal line at 70 are generally considered acceptable.

7 Discussion

This chapter presents a final discussion about the project. The result is discussed in comparison to the research question. Limitations and ethical aspects of the projects are discussed, as well as how the project contributed to the UN Sustainable Development Goals. The chapter finishes with a section about how the results from this project could be improved with future contributions.

7.1 How can one develop an application to support NGOs work in conducting hate speech analysis on social media platforms? (RQ)

During the investigation phase, we explore how people are working when analyzing social media platforms. We found examples of big and expensive services such as Brandwatch to tech-savvy individuals who conducted their research by writing scripts to fetch, process, and analyze the data by themselves. For an NGO such as CRD we concluded that they needed a user-friendly, trustworthy, and cheap application that could detect hate speech in several languages. We, therefore, decided that a web application would suit the needs of CRD the best.

The use cases in mind when developing this application have been the two mentioned in the investigation phase. The first one is a tool to detect spikes of hate speech against an organization or employees at an organization to foresee a potential physical attack against them. The second one is for a journalist to analyze tweets and get a quantitative overview of the discourse regarding a topic or in a country to have a deep understanding and good facts when writing reports. The use cases have different demands, and there are still possible improvements to achieve an application fully suitable for both use cases. The first use case demands almost real-time speed which the application still struggles with and the second use case demands high traceability which the application lacks when using third-party services such as Tisane Labs for text analysis.

With multiple evaluations during the development process, we are confident that the underlying data analysis is correct. We understand that the application lacks flexibility and that the preciseness of the text analysis could be questioned. The final version of the application was tested with end-users. Here we received excellent results in both the assessment test and the SUS questionnaire. Most of the participants quickly found the answers in all tasks. We identified a few improvements in components regarding usability. The main points were the difficulty to hover the intended day in the trendline graphs and the lack of a quick overview when hovering a specific label in the abuse graph or hovering a specific day in the trendline graph.

The application is developed with inspiration from Brandwatch. They, of course, have much more functionality in the forms of filtering and visualization options of their data. The main difference between other tools and ours is the abuse classification implemented in our product. This classification makes it possible to detect and get a clear overview of potential hate speech in the users' given search. During the assessment test, we also received feedback from CRD's employees who have used other tools and been part of implementing them in other partner organizations and felt that our tool was more straightforward and easier to understand. Therefore, it would demand less onboarding for a new user. Results also strengthen from question 3,4,7, and 10 in the SUS questionnaire and the feedback from the beta evaluation.

All participants' SUS scores were over 70, and three of them were even over 90 which proves that the application was well received by the intended user. From the SUS questionnaire, we could see that question 2, and 3 received the highest possible score by all the participants, two questions that focus on the easiness and interpretability of the product. The first question, *"I think that I would like to use this application frequently",* was the one with the lowest score, with three participants being neutral to the question. This came as no surprise. Analyzing social media is not an everyday task for employees at CRD but rather a recurring task depending on what happens globally.

7.2 Limitations

The application is built according to the requirement specification, and the assessment and SUS result is evaluated from those requirements. According to the result, the intended end-user was satisfied. However, one could question if the requirement specification is good enough and if the application would have the functionality to provide the user with better insights if the investigation phase had been conducted in another way.

During the investigation phase, interviews were held with field experts and managers at CRD. The investigation lacked ethnographic research. Conducting interviews with employees at CRD who used Twitter to attain information had been helpful in that stage of the development process to understand their current issues further.

Two weak parts of the application today is the text analysis model and the bot-like score from Botometer. Tisane Labs currently provides text analysis for 28 different languages, and we have only tested their English and Swedish analysis. The English analysis is working well, but we saw some severe problems with the Swedish analysis. Those issues were directly reported to Tisane Labs, who immediately fixed those, but one could expect more issues to be detected for smaller languages such as Danish or Albanian. The Botometer scores have been very varied and hard to understand and identify what causes a high or low bot-like score. When using the Botometer score, an important point is to keep in mind that it is a score on bot-like appearances and not a score on if the account is an actual bot. For example, a very spammy user who is active at strange times during the day would score high on the bot-like score.

7.3 Ethical aspects

With an application that processes personal data, one always needs to be aware of different laws and understand and respect the person's privacy whose data one is collecting. This application is using Twitter's API to access data from Twitter with research developer access. This demands the application to follow Twitter's policies as well. In our example, where we, for example, are storing their data offline needs to continuously update our data to keep it in the current state of Twitter. In other words, if a Tweet is deleted from Twitter, it needs to be deleted from our database as well to comply with their policy as well as GDPR. (Twitter, 2021b) If the application would further develop to also fetch data from other social media platforms such as Facebook or Instagram, one needs to also comply with their policies.

One could also argue that analyzing and monitoring social media to prevent hate speech and suspend accounts is against the freedom of speech, something that Tisane labs, with their text analysis tool, has received comments on. Their answer argument, which we agree with, is that "better moderation aids advance the freedom of speech".(Tisane Labs, 2021a) There is also a severe risk that stricter moderation on social media platforms can induce individuals who want to engage in hate speech to migrate to other, less-regulated platforms and therefore not solve the underlying problem of hate speech.

A final ethical aspect worth mentioning is that a tool possible to analyze and monitor a social platform can gain valuable insight, which could be used for harmful reasons.

Governments or organizations could use these kinds of platforms to monitor and control people with different opinions.

7.4 The 2030 Agenda

2030 is getting closer and closer each year, and the process to reach the UN sustainable goals is progressing, but not at the desired speed. With the Covid-19 pandemic and natural disaster hitting hard on vulnerable groups and individuals, reaching the goals 2030 is even more challenging. However, with the enormous adverse effects the global Covid-19 pandemic is causing, optimistic minds argue that it enables new initiatives and more aggressive actions thanks to its disruption of social norms and policies. This study aims to contribute to goals 10, which concentrates on reducing inequalities within and among countries and 16, which focuses on promoting peaceful and inclusive societies regardless of ethnicity, sexual orientation, or faith.

The application developed simplifies the process of monitoring and understanding the discourse on a social media platform such as Twitter regarding hate speech and bigotry. Shedding light on the negative discourse on social media will both give people in charge a better base of understanding about this problem as well as enable human rights organizations to pressure social media platforms to take action against accounts that are oppressing or discriminating people based on, for example, their ethnicity, religion, or sexual orientation.

7.5 Future improvements of the application

The application has been developed under a strict time constraint, and key functionalities have been prioritized during the development process. Valuable feedback has been received from both field experts and potential end-users with functionalities that could be implemented or adjusted to improve the application further. Some of the feedback received is, for example, more filtering options, such as keyword co-occurrences, which means that the user would be able to search for tweets that match both keywords one and keyword two instead of fetching all tweets that match either keyword 1 or keyword 2.

Another functionality, which the interviewed field experts asked for, is further functionality to get insights into the influence of users. Both in terms of visualization of how users affect each other as well as functionality on filtering on users, for example, the number of likes or retweets. These functionalities were not implemented because of the time constraint.

In this application, the text analysis used comes from the company Tisane Labs. Their text analysis comes with these predefined labels, such as bigotry, criminal activity, personal attack, etc. This leads to complications both in the effectiveness of the text analysis and the understanding of how the text analysis is carried out to prove traceability, especially when conducting research. Using one's text analysis model would give one better control of the analysis and better explain how the analysis is carried out. People also asked for customization of the text analysis used. This to be able to use different kinds of labels for different kinds of reports.

The final evaluation also proved that some of the components lacked some user-functionalities. For example, the different trendline graphs are hard to correctly hover over when the time frame ranges over several months. Functionality to give insights on specific days would also be appreciated if implemented. To clarify, when hovering over a specific day in the graph, people asked for an overview of what trended that day. Different functionalities which enable the user to click on, for example, a specific day in the trend graph or a specific label in the abuse graph, should lead the user to view tweets for that search term. Finally, the functionality to download the dataset from the application to the user’s own computer is not yet implemented.

As mentioned earlier in this application proof of concept, some security and robustness improvements need to be made to reach production level to deploy this in an organization functionally. The applications codebase is open-source on Github[1], and it will be possible for other people to support the implementation of future functionality.

[1] HateHunter Github repository - https://github.com/luuddan/hate-hunter

8 Conclusion

This paper has proved that developing an application which is able to analyze Twitter is feasible. Developing a service which can deliver a more profound and flexible data analysis while not demanding much technological knowledge by the user is complex and time-consuming. There already exist multiple applications for analyzing social media platforms. Many of these are expensive and have their focus on marketing and brand awareness. The application developed in this paper focus on hate speech and is filling a void for organizations or researchers with minimal budgets.

During the development process, we have received feedback on other possible use cases. When talking to field experts in academia, they have shown an eagerness to use the tool in their work. The experts we have talked to can already fetch and analyze tweets from Twitter by writing their own code. This is time-consuming work but leads to a good result with absolute traceability. Their use case would be to use this application during an initial state of their research to search for interesting research topics and evaluate possible research questions.

These experts from academia are, as well as researching how social media is affecting the social discourse, lecturing in university courses on how to analyze social media. The students at the courses are mainly social science students with minimal experience in data analysis. The experts see great potential to use this application to give the students a user-friendly tool to quickly fetch tweets from Twitter and be introduced to fundamental data analysis.

We believe our application suits the needs of CRD and will be easy to implement in the organization, thanks to the user-friendliness of the application. We, therefore, argue that the thesis RQ is fulfilled. As a final word, we want to point out that monitoring and analyzing social media does not solve the root problem of hate speech. For a long-term improvement, the world needs to work towards the 2030 Agenda and solve the underlying factors that causes hate speech: social insecurity, inequality, and poverty.

9 References

Amnesty UK. 2020. *Uganda's new anti-human rights laws aren't just punishing LGBTI people.* May 18, 2020. Amnesty International. Accessed: May 13, 2021. https://www.amnesty.org.uk/uganda-anti-homosexual-act-gay-law-free-speech

Louise Barriball, K. & While, A., 1994. *Collecting Data using a semi-structured interview: a discussion paper.* Journal of advanced nursing, 19(2), pp.328-335.

Bay, S. 2021. *Researcher at Department of Defence Analysis, Swedish Defence Research Agency.* Interview January 29, 2021.

Bayer, J. & Bard, P. 2020. *Hate speech and hate crime in the EU and the evaluation of online content regulation approaches*. European Parliament. Brussel.

Bengtsson, L. 2021. *Supervisor at LTH.* Personal communication January-June 2021.

Cullinanem, C. 2015. *Data and #GE2015: Public bodies need to prioritise consistent data formats and commit to accessible information.* Impact of Social Sciences. May 7, 2015. Accessed: May 13, 2021. https://blogs.lse.ac.uk/impactofsocialsciences/2015/05/07/data-and-the-general-election/

Brandwatch. 2021. *Product.* Accessed: May 24, 2021. https://www.brandwatch.com/platform/

Brooke, J. 2013. *SUS: a retrospective. Journal of usability studies*, 8(2), pp.29-40.

Brown, S. 2020. *The Innovation Ultimatum: How six strategic technologies will reshape evety business in the 2020s*. 1st ed. New York: John Wiley & Sons Inc.

Civil Rights Defenders (CRD). 2019. *Annual Report 2019*. Accessed: May 13, 2021. https://crd.org/wp-content/uploads/2020/07/CRD-Annual-Report-2019-ENG-webb.pdf

Jourová, V. 2019. *Code of Conduct on countering illegal hate speech online.* European Comission. Accessed: May 13, 2021. https://ec.europa.eu/info/sites/default/files/code_of_conduct_factsheet_7_web.pdf

CrowdTangle. 2021a. *FAQ: General CrowdTangle Questions.* Accessed: May 24, 2021. https://help.crowdtangle.com/en/articles/2541882-faq-general-crowdtangle-questions#:~:text=There%20is%20no%20cost%20for,to%20select%20Facebook%20publishing%20partners.

CrowdTangle. 2021b. *About us.* Accessed: May 24, 2021. https://help.crowdtangle.com/en/articles/4201940-about-us

CrowdTangle. 2021c. *What data is crowdtangle tracking.* Accessed: May 24, 2021. https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking

De Kaminski, M. 2021. *Security and Innovation Director at CRD*, Personal communication January-June 2021.

Department of Global Communications (DGC). 2020. *COVID-19: UN counters pandemic-related hate and xenophobia*. United Nations. Accessed: May 13,2021. https://www.un.org/en/coronavirus/covid-19-un-counters-pandemic-related-hate-and-xenophobia

Donges, N. 2021. *Introduction to NLP*. May, 2021. Builtin. Accessed: May 24, 2021. https://builtin.com/data-science/introduction-nlp

Facebook. 2020. *Facebook Platform Terms*. Facebook inc. Accessed: May 13, 2021. https://developers.facebook.com/terms/

Facebook. 2021a. *Number of monthly active Facebook users worldwide as of 4th quarter 2020 (in millions)*. Statista. Statista Inc.. Accessed: May 13, 2021. https://www-statista-com.ludwig.lub.lu.se/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Facebook. 2021b. *Facebook Data for Independent Research*. Facebook inc. Accessed: May 13, 2021. https://research.fb.com/data/

Facebook. 2021c. *Hate Speech.* Facebook inc. Accessed: May 13, 2021. https://www.facebook.com/communitystandards/hate_speech

Fernandes, D. & Bernardino, J., 2018. *Graph Databases Comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB*. In DATA (pp. 373-380).

HateLab. 2021. *About*. Accessed: May 20, 2021. https://hatelab.net/

HateLab. 2019. *Online Hate Speech Predicts Hate Crimes on the Streets*. Accessed: May 20, 2021. https://hatelab.net/2019/10/14/online-hate-speech-predicts-hate-crimes-on-the-streets/

Helldén, E. 2021. *Supervisor at CRD*. Personal communication January-June 2021.

Hirst, M. 2021. *NI riots: What is behind the violence in Northern Ireland?* April 14, 2021. BBC. Accessed: May 13, 2021. https://www.bbc.com/news/uk-northern-ireland-56664378

IBM. 2021. *Natural Language Processing (NLP).* Accessed: May 24, 2021. https://www.ibm.com/cloud/learn/natural-language-processing

IDEO, 2020. *What is Design Thinking?*. Accessed: May 12, 2021. https://www.ideou.com/blogs/inspiration/what-is-design-thinking

Jigsaw. 20221. *Jigsaw.* Accessed: May 24, 2021. https://jigsaw.google.com/

Kullenberg, C. 2021a. *Investigation interview*, Interview: January 28, 2021.

Kullenberg, C. 2021b. *Beta evaluation interview*, Interview: April 15, 2021.

Lee, R.K.W. & Li, Z. 2020. *Online Xenophobic Behavior Amid the COVID-19 Pandemic: A Commentary. Digital Government: Research and Practice*, 2(1), pp.1-5.

Lindgren, S. 2021. *Beta evaluation interview*, Interview: April 16, 2021.

Marvin, R. 2019. *Brandwatch Analytics Review.* Mar, 2019. Accessed: May 24, 2021. https://uk.pcmag.com/cloud-services/71215/brandwatch-analytics

Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. 2019. *Spread of Hate Speech in Online Social Media*. In Proceedings of the 10th ACM Conference on Web Science (WebSci '19). Association for Computing Machinery, New York, NY, USA, 173–182. DOI:https://doi.org/10.1145/3292522.3326034

MDN contributors. 2021. *Introduction to web APIs*. Accessed: May 25, 2021. https://developer.mozilla.org/en-US/docs/Learn/JavaScript/Client-side_web_APIs/Introduction

Neo4j. 2021a. *Products*. Accessed: May 20, 2021. https://neo4j.com/product/

Neo4j. 2021b. *Concepts: Relational to graph.* Accessed: May 23, 2021. https://neo4j.com/developer/graph-db-vs-rdbms/

Neo4j. 2021c. *Try Neo4j.* Accessed: May 24, 2021. https://neo4j.com/try-neo4j/?ref=product

Neo4j. 2021d. *Neo4j Aura.* Accessed: May 24, 2021. https://neo4j.com/cloud/aura/

Neo4j. n.d. *What is a Graph Database?*. Accessed: May 25, 2021. https://neo4j.com/developer/graph-database/

Osome. 2021a. *About OSoMe.* Accessed: May 24, 2021. https://osome.iu.edu/about/mission

Osome. 2021b. *Misinformation tools.* Accessed: May 24, 2021. https://osome.iu.edu/tools

Pelletier, M.J., Krallman, A., Adams, F.G. and Hancock, T. 2020. *One size doesn't fit all: a uses and gratifications analysis of social media platforms.* Journal of Research in Interactive Marketing.

Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F. and Camacho-Collados, M. 2019. *Detecting and monitoring hate speech in Twitter*. Sensors, 19(21), p.4654.

Perrigo, B. 2019. *Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch*. Time. Accessed: May 20, 2021.

Perspective API. 2021. *Limits & Errors.* Accessed: May 25, 2021. https://developers.perspectiveapi.com/s/about-the-api-limits-and-errors

Petterson, Ylwa. 2021. *Communication & PR manager at Twitter.* Interview February 23, 2021.

RapidAPI. 2021a. *Hoaxy API documentation.* Accessed: May 24, 2021. https://rapidapi.com/truthy/api/hoaxy

RapidAPI. 2021b. *Botometer PRO API documentation.* Accessed: May 24, 2021. https://rapidapi.com/OSoMe/api/botometer-pro

Rill, S., Reinel, D., Scheidt, J. and Zicari, R.V., 2014. *Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis*. Knowledge-Based Systems, 69, pp.24-33.

Roth, Y. & Pickles, N. 2020. *Bot or not? The facts about platform manipulation on Twitter.* May 18, 2020. Twitter inc. Accessed: May 13, 2021. https://blog.twitter.com/en_us/topics/company/2020/bot-or-not.html

Rubin, J. & Chisnell, D., 2008. *How to plan, design, and conduct effective tests. Handbook of usability testing,* 17(2), p.348.

Schmidt, E. 2016. *Google Ideas Becomes JigSaw.* February, 2016. Medium. Accessed: May 24, 2021. https://medium.com/jigsaw/google-ideas-becomes-jigsaw-bcb5bd08c423

Sommerville, I. 2011. *Software engineering 9th Edition*. ISBN-10, 137035152, p.18.

The Sustainable Development Goals: Our Framework for COVID-19 Recovery. 2021. United Nations. Accessed: May 13, 2021. https://www.un.org/sustainabledevelopment/sdgs-framework-for-covid-19-recovery/

Tisane Labs. 2021a. *What you're doing is wrong! It's dictatorship! Against free speech!.* Accessed: May 13, 2021. https://tisane.ai/knowledgebase/what-youre-doing-is-wrong/

Tisane Labs. 2021b. *Text Analysis.* Accessed: May 24, 2021. https://dev.tisane.ai/docs/services/5a3b6668a3511b11cc292655/operations/5a3b7177a3511b11cc29265c

Tisane Labs. 2021c. *Subscription plans.* Accessed: May 24, 2021. https://tisane.ai/subscription-plans/

Troianovski, A. & Nechepurenko, I. 2021. *Navalny Arrested on Return to Moscow in Battle of Wills With Putin.* January 17, 2021. The New York Times. Accessed: May 13, 2021 https://www.nytimes.com/2021/01/17/world/europe/navalny-russia-return.html

Twitter. (2019). *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions).* Statista. Statista Inc.. Accessed: May 13, 2021. https://www-statista-com.ludwig.lub.lu.se/statistics/282087/number-of-monthly-active-twitter-users/

Twitter. 2021a. *Developer terms*. Twitter inc. Accessed: May 13, 2021. https://developer.twitter.com/en/developer-terms

Twitter. 2021b. *Developer terms policy*. Twitter inc. Accessed: May 13, 2021. https://developer.twitter.com/en/developer-terms/policy

Twitter. 2021c. *Tutorial: Tweet geospatial metadata.* Twitter inc. Accessed: May 13, 2021. https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata

Twitter. 2021d. *Hateful conduct policy.* Twitter inc. Accessed: May 13, 2021. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Twitter. 2021e. *Platform manipulation.* Twitter inc. Accessed: May 13, 2021. https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jan-jun

Twitter. 2021f. *Rules enforcement.* Twitter inc. Accessed: May 13, 2021. https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jan-jun

Twitter. 2021g. *Search Tweets.* Twitter inc. Accessed: May 24, 2021. https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all

United Nations Statistics Division (UNSD). 2020. *The Sustainable Development Goals Report*. New York, NY: United Nations.

United Nations (UN). 2020a. *Reduce Inequalites: Why it matters.* United Nations. Accessed: May 13, 2021. https://www.un.org/sustainabledevelopment/wp-content/uploads/2018/01/10_Why-It-Matters-2020.pdf

United Nations (UN). 2020b. *Goal of the month May 2020*. United Nations. Accessed: May 13, 2021. https://www.un.org/sustainabledevelopment/goal-of-the-month-may-2020/

United Nations (UN). 2019. *UNITED NATIONS STRATEGY AND PLAN OF ACTION ON HATE SPEECH.* United Nations. Accessed: May 24, 2021. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

Wakefield, J. 2021. *Uganda elections 2021: Facebook shuts government-linked accounts.* January 11, 2021. BBC. Accessed: May 13, 2021.

Williams, M.L., Burnap, P., Javed, A., Liu, H. and Ozalp, S. 2020. *Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime.* The British Journal of Criminology, 60(1), pp.93-117.

World Health Organization (WHO). 2020. *Listings of WHO's response to COVID-19.* June 29, 2020. World Health Organization. Accessed: May 13, 2021. https://www.who.int/news/item/29-06-2020-covidtimeline2

Zhang, Y. & Wildemuth, B.M., 2009. *Unstructured interviews. Applications of social research methods to questions in information and library science*, pp.222-231.

Appendix A

A.1 Semi-structured interview

- What is the objective with a product as ours?
- Which functions do you think will be the most important?
- What do you think will separate our service from a data analysis tool or manually analyzing Twitter?

Appendix B

Beta test plan

B.1 Test plan

This is a script created for the testing of the beta version of the service, created from the requirements.

B.2 Objective

The purpose of the test is to evaluate the following points.

- Does the service support tasks that are performed during a typical data analysis of Twitter?
- How easily can the user get valuable insights into Twitter activity
- Does anything need to be changed in the intended workflow or interface to support the tasks?

B.3 Selection of participants

The desired test participants are participants with prior experience within civil rights work and/or data analysis of social media.

B.4 Equipment

The interview is conducted over Zoom and the product is screenshared while running on one of the interviewer's local computers.

B.5 Execution

The interview will be conducted by two interviewers who will collaborate with the interviewee in a semi-structured interview format to solve the tasks in B.1.6.

The interview will consist of the following steps:

1. The interviewee will join the Zoom meeting.

2. A short description of the overall project is presented by the interviewers.

3. The interviewer will read the information in B.1.5. about the specific use case.

4. One of the interviewers will be in control of the computer but be guided by the interviewee to solve the tasks described in B.1.6.

5. The interviewee will be asked the questions in B.1.7 to evaluate the product.

B.6 Use case information

At the beginning of April 2021, multiple violent riots occurred in Northern Ireland. Before the interview, all the tweets on Twitter between the dates 15/3-2021 and 14/4-2021 that contained the shortening “IRA”, a paramilitary organization in Northern Ireland active in the riots, were fetched and 10% of the tweets were analyzed.

How would you analyze this dataset in regards to writing a summary of the discourse on Twitter about the “IRA” during this timeframe?

B.7 Task scenarios

1. Search for “IRA” as a keyword

2. Find tweets with criminal activity

3. See an overview of the activity trend

4. Cherry-pick keyword trendlines

5. View a relevant user

B.8 Post test questions

- How did you experience the tasks that were carried out?
- Were they easy or hard to perform?
- Did the tasks feel like tasks that you think a civil rights worker would carry out?
- Would you want the information to be presented another way?
- Do you think that the feed would be suitable for a large number of posts and a long time span? Would anything have to be changed to support this?
- Was there anything that you felt was confusing? In the application or in the tasks?
- Was there anything that you felt was missing in the application?
- Did you feel you were missing any information to be able to use the application?

Appendix C

Final test plan

C.1 Test plan

This is a script created for the testing of the final version of the service, created from the requirements. 20min interview, 5 min introduction with background information, use case information and a short onboarding of the application where all its functionalities are briefly described, 10 min testing, 5 min post-questions.

C.2 Objective

The purpose of the test is to evaluate the following points.

- Are the requirements fulfilled?
- How well does the service suit users without a technical background?
- Will the application be suitable for the work carried out by CRD's employees?
- Does anything need to be changed in the intended workflow or interface to support the tasks?

C.3 Selection of participants

The desired test participants are employees at CRD.

C.4 Equipment

The interview is conducted over Zoom and the product is screenshared but monitored and steered by the interviewee while running on one of the interviewer's local computers.

C.5 Execution

The interviewee will be asked to solve the tasks in C.1.6.

The interview will consist of the following steps:

1. The interviewee will join the Zoom meeting.
2. A short description of the overall project is presented by the interviewers.
3. The interviewer will read the information in C.1.5. about the specific use case.
4. The interviewee will monitor the computer and try to solve the tasks described in C.1.6.
5. The interviewee will be asked the questions in C.1.7 and C.1.8 to evaluate the product.

C.6 Background information

We are engineering students with master's in computer science writing our master thesis at CRD.

We have this spring been building a proof of concept application to monitor and quantitatively analyze tweets from Twitter. Hopefully, this could be useful for you at CRD and give you another tool to get a better overview on the discourse on Twitter.

C.7 Use case information

At the beginning of 2020, corona spread around the world. Due to Corona there has been new hate speech activity on the internet. Before the interview, all the tweets on Twitter between the dates 31/12-2019 and 2/5-2021 that contained the keywords "Corona & Uganda" or "Covid & Uganda" were fetched and 100% of the tweets were analyzed.

This test is a usability test of the application, where you will be asked to perform a few tasks, after an onboarding to the application has been done.

C.8 Task scenarios

1. With the initial search with "Covid" & "Corona" as keywords over the given timeframe, how many tweets does the dataset contain?

2. How many of these tweets were labeled as criminal activity?

3. Which day had the most tweets?

4. Which were the top 3 hashtags for the specific keyword "Covid" & "Corona" during the given timeframe?

5. Find the most influential user, does this person have any tweets classified as bigotry?

6. Search for the keyword "Vaccine" and find out what the average tone for "Vaccine" is in Uganda.

7. With the search for "Vaccine" over the given timeframe, is there any day or period with significantly more tweets? What has generated this spike?

C.9 Post test questions

- How did you experience the tasks that were carried out? Were they easy or hard to perform?
- Did the application feel like something that you think a civil rights worker would use?
- Was there anything that you felt was missing in the application?

C.10 SUS Questionnaire

1. I think that I would like to use this application frequently

2. I found the application unnecessarily complex

3. I thought the application was easy to use

4. I think that I would need the support of a technical person to be able to use this application

5. I found various functions in this application were well-integrated

6. I thought there was too much inconsistency in this application

7. I would imagine that most people would learn to use this application very quickly

8. I found the application very cumbersome to use

9. I felt very confident using the application

10. I needed to learn a lot of things before I could get going with this application