KEN4256 Projects

Maastricht University
Institute of Data Science

Project

Individual project to demonstrate your creativity and technical virtuosity

(60% of final grade; 10% project proposal, 50% written project report)

Types of Projects

1. **Building knowledge graphs.** choose 2 or more publicly available and open datasets about any topic that you are interested in, that are not already available in RDF. Convert them to RDF, link them to standardized ontologies and/or construct a FAIR ontology, assess KG quality, and demonstrate its utility by answering research-worthy competency questions.
2. **Mining a knowledge graph:** choose one or more existing knowledge graphs (preferably of large size with millions of entities) and apply techniques such as semantic data integration, link prediction, and/or advanced reasoning to derive new information from the graph which is not explicitly present. Ensure there is a proper evaluation of the approach. Try to find insights that are counterintuitive, surprising and add to the body of human knowledge.
3. **Build and mine a knowledge graph.** This project requires a new graph to be constructed in support of KG-specific data mining. The project is focused on this end to end exploration, and therefore may not necessarily have to address every expectation with respect to a study devoted to either building a high quality KG, or mining one or more KGs.

Project Proposal

The project proposal must be **at most** 2 pages (excluding references).

- **Significance:** (What is the problem, and why is it an important problem? Who is impacted by the problem? What is the cost of not solving the problem? What opportunities arise should the problem be addressed?)
- **Related work** (Are there others who tried to do this or something similar before? If not, describe how previous work is different from what you propose and what gap yours will fill in the body of scientific knowledge.)
- **Goal and specific objectives** (What is the overarching goal of the project? What specific objectives will be pursued in the project? Briefly describe how these objectives will be addressed. Indicate how these objectives are relevant to the course learning objectives)
- **Anticipated results** (What do you expect the main outputs of your project to be in terms of new knowledge or insights, new technology or capabilities, challenges addressed, and importance to others)
- **Approach** (How you intend to answer your research questions. What data and technologies are you going to use? Why are they the most suitable to answer your research questions or solve your problem? What elements of the project are risky, and how will you mitigate this risk e.g. what is your plan B? Include an overview diagram to communicate your project architecture, components, and plans)
- **Milestones & deliverables** (Checkpoints and deliverables in the project that can be used to measure progress. These can be in the form of modules for a software package; components or sections of a Knowledge Graph if you are constructing one; subset of a full set of questions that will be answered by certain checkpoints in the project etc.)
- **References:** A section containing proper references for inline citations. Use a citation manager like Endnote, Zotero, Mendeley, or Paperpile to insert and automatically format citations and references.

Project Report (6 pages)

The project report must be **at most** 6 pages (excluding references).

The report must contain the following major sections:

- **Project Title**, date, student name, and student number.
- **Abstract:** A brief summary of the project and findings.
- **Introduction**: An introduction to the project including the problem description and goal and specific objectives.
- **Related Work.**
- **Methodology.**
- **Results**: A description of the results supported by tables and figures. Include references to result materials (e.g. specific files in the project zip file)
- **Discussion**: Your interpretation of the results. In what way does the results support or dispute your hypothesis or shed light on your research questions? What challenges were faced and how were they addressed? What didn't work, and why? What new insights could you find that add to the body of scientific knowledge? What questions are left unsolved? what way could your or another person extend the work in a future study? In what way did this project help fulfill the course learning objectives?
- **Conclusions:** Briefly summarize the approach, findings, and their significance.
- **References**.

Evaluation Rubric

Project proposal is out of maximum of 6 points, and includes Significance, Methodology, and Quality of Writing

Project report is out of a maximum of 10 points, and includes Significance, Methodology, Quality of Results, Quality of Discussion, Quality of Writing

Significance (2 points)

1. The significance of the problem is not described nor supported with literary references [0]
2. The significance of the problem is reasonably well communicated, with some literary references and evidence of its importance [1]
3. The significance of the problem is convincingly communicated with ample references and substantive evidence of importance [2].

Methodology - Feasibility and Innovation (2 points)

1. serious issues with methodology, either too simple or too complex, unclear approach and/or deemed infeasible [0]
2. feasible with some concerns that can be addressed, of moderate ambition and/or complexity [1]
3. ambitious, innovative, well planned, risk mitigation included, clear overview of the approach [2]

Rubric

Quality of Writing (2 points)

1. poorly written, a struggle to understand the work [0]
2. well written, good flow that free of grammatical and spelling mistakes [1]
3. superbly written, inspiring to read [2]

Quality of Results (2 points)

1. the results are poorly described, if at all [0]
2. the results are well described and/or illustrated, but there are several questions regarding the work performed and results obtained [1]
3. the results are clear and very well presented [2]

Quality of Discussion (2 points)

1. The discussion does not properly describe the nature and significance of the results [0]
2. The discussion summarizes findings, and offers a reasonable but faulty analysis of their significance [1]
3. The discussion summarizes findings, correctly interprets these findings, places them in the context of related work, and offers several new avenues for follow up work [2]

Proposal Criteria, elaborated

Relevance
To what extent is the work and methods described relevant to the project as a whole and the course objectives?

Significance
What is the problem being addressed in the project? How **important** is this **problem**? Consider: Who is affected by the problem? How much money do people spend addressing the problem / what is the cost of not solving the problem? What are the benefits / outcomes of solving the problem?

Approach
Is the approach **appropriate** for the problem? Is the approach sufficiently **clear** and **detailed** to enable their reproduction by a peer? Is there a useful diagram to **illustrate** the approach? Are the **limitations** of the approach clearly articulated? Is the amount of **work** involved trivial (using an existing tool) or more challenging (developing a new pipeline and/or method)?

Feasibility. Is the approach feasible? Does it appear overly complicated or is it relatively straightforward? Does it comprise of the data needed to answer the research question? If the project is only on building a KG,does it consist of several independent resources that have been brought together to answer the research question? Is there a plan to assess the quality of the KG? To what extent does the KG address the FAIR principles? Does it use shared vocabularies / ontologies?

Quality of writing

- Clear: avoids unnecessary detail, is free of mistakes in spelling and grammar
- Simple: uses direct language and avoids vague or complicated sentences. Technical terms are properly defined and used only when they are necessary for accuracy
- Organized: text is divided into appropriate sections with clear headings
- Objective: statements and ideas are supported by appropriate evidence. References to scientific literature are provided where necessary
- Length: 2 pages (we should provide a template)

Significance

What is the problem, and why is it an important problem?
Who is impacted by the problem?
What is the cost of not solving the problem?
What opportunities arise should the problem be addressed?

The following includes excerpts from a student proposal, with their permission.

1 Significance

Social media data (SMD) offer insights about many areas, such as public opinion, market trends, and misinformation spreading. Researchers from different fields – including sociology, economics, and computer science – use SMD as the primary data source for qualitative and quantitative experiments. SMD are available at a large-scale and can be obtained in real-time from different platforms. Large-scale availability has benefits: it enables the analysis of many societal issues and offers insight into different communities. However, its indiscriminate use raises many ethical and privacy issues. SMD might compromise users' privacy and safety and cause harms beyond social media. These concerns, combined with strict platform terms and conditions that often prohibit data sharing, heavily contribute to the poor data management practices in social media research.

General introduction

Hemphill, Hedstrom, and Leonard (2021) examine data management practices in the context of SMD research by surveying researchers and analysing published papers. The survey results show a concerning trend: only 46.6% of the respondents make their data available, only 28.8% prepare their data especially for reuse, and only 23.3% do so for replication. Such low percentages show the need for better data sharing practices – especially for ensuring reuse and replication.

Poor data curation in social media research may also lead to other critical issues. As data collected from social media are often texts (e.g. *tweets*, Instagram posts, Youtube comments), Natural Language Processing (NLP) significantly benefits from SMD. With the advent of large language modelling architectures like GPT-3 (Brown et al., 2020), the NLP community has been using massive amounts of data to train models. Emily M Bender et al. (2021) highlight many issues that arise from poor data curation practices in the context of language models. Models trained on large amounts of data might amplify biases in the data, which, in turn, end up implemented as tools that might be harmful to people – and such tools disproportionately affect marginalised communities and people who are already vulnerable. As data is the starting point of trained models, it is fundamental to actively identify and mitigate biases as part of the dataset design process.

This project proposes a metadata schema for SMD to improve the dataset creation and data-sharing practices in social media research, mainly focusing on mitigating ethical and privacy concerns.

1 Significance

Social media data (SMD) offer insights about many areas, such as public opinion, market trends, and misinformation spreading. Researchers from different fields – including sociology, economics, and computer science – use SMD as the primary data source for qualitative and quantitative experiments. SMD are available at a large-scale and can be obtained in real-time from different platforms. Large-scale availability has benefits: it enables the analysis of many societal issues and offers insight into different communities. However, its indiscriminate use raises many ethical and privacy issues. SMD might compromise users' privacy and safety and cause harms beyond social media. These concerns, combined with strict platform terms and conditions that often prohibit data sharing, heavily contribute to the poor data management practices in social media research.

all is not rosy...

Hemphill, Hedstrom, and Leonard (2021) examine data management practices in the context of SMD research by surveying researchers and analysing published papers. The survey results show a concerning trend: only 46.6% of the respondents make their data available, only 28.8% prepare their data especially for reuse, and only 23.3% do so for replication. Such low percentages show the need for better data sharing practices – especially for ensuring reuse and replication.

Poor data curation in social media research may also lead to other critical issues. As data collected from social media are often texts (e.g. *tweets*, Instagram posts, Youtube comments), Natural Language Processing (NLP) significantly benefits from SMD. With the advent of large language modelling architectures like GPT-3 (Brown et al., 2020), the NLP community has been using massive amounts of data to train models. Emily M Bender et al. (2021) highlight many issues that arise from poor data curation practices in the context of language models. Models trained on large amounts of data might amplify biases in the data, which, in turn, end up implemented as tools that might be harmful to people – and such tools disproportionately affect marginalised communities and people who are already vulnerable. As data is the starting point of trained models, it is fundamental to actively identify and mitigate biases as part of the dataset design process.

This project proposes a metadata schema for SMD to improve the dataset creation and data-sharing practices in social media research, mainly focusing on mitigating ethical and privacy concerns.

1 Significance

Social media data (SMD) offer insights about many areas, such as public opinion, market trends, and misinformation spreading. Researchers from different fields – including sociology, economics, and computer science – use SMD as the primary data source for qualitative and quantitative experiments. SMD are available at a large-scale and can be obtained in real-time from different platforms. Large-scale availability has benefits: it enables the analysis of many societal issues and offers insight into different communities. However, its indiscriminate use raises many ethical and privacy issues. SMD might compromise users' privacy and safety and cause harms beyond social media. These concerns, combined with strict platform terms and conditions that often prohibit data sharing, heavily contribute to the poor data management practices in social media research.

Hemphill, Hedstrom, and Leonard (2021) examine data management practices in the context of SMD research by surveying researchers and analysing published papers. The survey results show a concerning trend: only 46.6% of the respondents make their data available, only 28.8% prepare their data especially for reuse, and only 23.3% do so for replication. Such low percentages show the need for better data sharing practices – especially for ensuring reuse and replication.

Poor data curation in social media research may also lead to other critical issues. As data collected from social media are often texts (e.g. *tweets*, Instagram posts, Youtube comments), Natural Language Processing (NLP) significantly benefits from SMD. With the advent of large language modelling architectures like GPT-3 (Brown et al., 2020), the NLP community has been using massive amounts of data to train models. Emily M Bender et al. (2021) highlight many issues that arise from poor data curation practices in the context of language models. Models trained on large amounts of data might amplify biases in the data, which, in turn, end up implemented as tools that might be harmful to people – and such tools disproportionately affect marginalised communities and people who are already vulnerable. As data is the starting point of trained models, it is fundamental to actively identify and mitigate biases as part of the dataset design process.

This project proposes a metadata schema for SMD to improve the dataset creation and data-sharing practices in social media research, mainly focusing on mitigating ethical and privacy concerns.

deeper exploration of the problem, with references to the literature

1 Significance

Social media data (SMD) offer insights about many areas, such as public opinion, market trends, and misinformation spreading. Researchers from different fields – including sociology, economics, and computer science – use SMD as the primary data source for qualitative and quantitative experiments. SMD are available at a large-scale and can be obtained in real-time from different platforms. Large-scale availability has benefits: it enables the analysis of many societal issues and offers insight into different communities. However, its indiscriminate use raises many ethical and privacy issues. SMD might compromise users' privacy and safety and cause harms beyond social media. These concerns, combined with strict platform terms and conditions that often prohibit data sharing, heavily contribute to the poor data management practices in social media research.

Hemphill, Hedstrom, and Leonard (2021) examine data management practices in the context of SMD research by surveying researchers and analysing published papers. The survey results show a concerning trend: only 46.6% of the respondents make their data available, only 28.8% prepare their data especially for reuse, and only 23.3% do so for replication. Such low percentages show the need for better data sharing practices – especially for ensuring reuse and replication.

Poor data curation in social media research may also lead to other critical issues. As data collected from social media are often texts (e.g. *tweets*, Instagram posts, Youtube comments), Natural Language Processing (NLP) significantly benefits from SMD. With the advent of large language modelling architectures like GPT-3 (Brown et al., 2020), the NLP community has been using massive amounts of data to train models. Emily M Bender et al. (2021) highlight many issues that arise from poor data curation practices in the context of language models. Models trained on large amounts of data might amplify biases in the data, which, in turn, end up implemented as tools that might be harmful to people – and such tools disproportionately affect marginalised communities and people who are already vulnerable. As data is the starting point of trained models, it is fundamental to actively identify and mitigate biases as part of the dataset design process.

This project proposes a metadata schema for SMD to improve the dataset creation and data-sharing practices in social media research, mainly focusing on mitigating ethical and privacy concerns.

Rationale for the proposal

Introduction to proposal

Related Work

Are there others who tried to do this or something similar before?
If not, describe how previous work is different from what you propose and what gap yours will fill in the body of scientific knowledge.

2 Related Work

In the context of issues of sharing SMD, Hemphill, Hedstrom, and Leonard (2021) thoroughly analyse current data management practices among social media researchers, focusing mainly on social scientists. They discuss issues that make social media data different from other data regarding archiving. They highlight specific areas which need improvement to create better collections of SMD. Williams, Burnap, and Sloan, 2017 propose an ethical framework inspired by social science research methods for sharing Twitter data. They survey Twitter users' to understand their attitude towards data sharing. Based on the survey results and other variables, they suggest a framework highly focused on user consent – researchers should only share data if they have opt-in user consent. Breuer, Bishop, and Kinder-Kurlanda (2020) analyse different ways of collecting data – including APIs and web-scraping. They discuss the advantages and challenges of seeking public-private partnerships to obtain data and propose strategies for working in this data-sharing setting. In general, these papers identify essential ethical issues and other problems that lead to poor data-sharing practices. However, many of their recommendations (e.g. getting consent from each user) are infeasible, especially in the context of big data. Achieving a wide adoption of good practices requires clear and simple guidelines and standardised tools and methods to facilitate data-sharing.

Emily M. Bender and Friedman (2018) propose *data statements* as a way of designing and documenting NLP datasets. Data statements are documents that include information about most steps of the dataset creation process – including, among others, curation rationale (which texts to include and why), speaker demographic, and annotator demographic. Data statements are unstructured (i.e., written as free text) and do not have standardised representations. Gebru et al. (2020) propose a similar approach to document datasets in general, without the NLP focus. Datasheets include a series of questions that encourage dataset creators to reflect on many aspects of the dataset creation process, such as preprocessing steps, potential uses, data distribution etc. The documents are unstructured. Mitchell et al. (2019) introduce model cards as a framework to represent trained machine learning models. Although they do not propose a description of datasets, model cards are still relevant to this project's context because they consider the "end-user" (a model, in this case) of a dataset. It is important to add this perspective as data is often collected to be used to train models.

addressing elements in the introduction, with inline citations

2 Related Work

In the context of issues of sharing SMD, Hemphill, Hedstrom, and Leonard (2021) thoroughly analyse current data management practices among social media researchers, focusing mainly on social scientists. They discuss issues that make social media data different from other data regarding archiving. They highlight specific areas which need improvement to create better collections of SMD. Williams, Burnap, and Sloan, 2017 propose an ethical framework inspired by social science research methods for sharing Twitter data. They survey Twitter users' to understand their attitude towards data sharing. Based on the survey results and other variables, they suggest a framework highly focused on user consent – researchers should only share data if they have opt-in user consent. Breuer, Bishop, and Kinder-Kurlanda (2020) analyse different ways of collecting data – including APIs and web-scraping. They discuss the advantages and challenges of seeking public-private partnerships to obtain data and propose strategies for working in this data-sharing setting. In general, these papers identify essential ethical issues and other problems that lead to poor data-sharing practices. However, many of their recommendations (e.g. getting consent from each user) are infeasible, especially in the context of big data. Achieving a wide adoption of good practices requires clear and simple guidelines and standardised tools and methods to facilitate data-sharing.

Emily M. Bender and Friedman (2018) propose *data statements* as a way of designing and documenting NLP datasets. Data statements are documents that include information about most steps of the dataset creation process – including, among others, curation rationale (which texts to include and why), speaker demographic, and annotator demographic. Data statements are unstructured (i.e., written as free text) and do not have standardised representations. Gebru et al. (2020) propose a similar approach to document datasets in general, without the NLP focus. Datasheets include a series of questions that encourage dataset creators to reflect on many aspects of the dataset creation process, such as preprocessing steps, potential uses, data distribution etc. The documents are unstructured. Mitchell et al. (2019) introduce model cards as a framework to represent trained machine learning models. Although they do not propose a description of datasets, model cards are still relevant to this project's context because they consider the "end-user" (a model, in this case) of a dataset. It is important to add this perspective as data is often collected to be used to train models.

further elaboration of the findings

2 Related Work

In the context of issues of sharing SMD, Hemphill, Hedstrom, and Leonard (2021) thoroughly analyse current data management practices among social media researchers, focusing mainly on social scientists. They discuss issues that make social media data different from other data regarding archiving. They highlight specific areas which need improvement to create better collections of SMD. Williams, Burnap, and Sloan, 2017 propose an ethical framework inspired by social science research methods for sharing Twitter data. They survey Twitter users' to understand their attitude towards data sharing. Based on the survey results and other variables, they suggest a framework highly focused on user consent – researchers should only share data if they have opt-in user consent. Breuer, Bishop, and Kinder-Kurlanda (2020) analyse different ways of collecting data – including APIs and web-scraping. They discuss the advantages and challenges of seeking public-private partnerships to obtain data and propose strategies for working in this data-sharing setting. In general, these papers identify essential ethical issues and other problems that lead to poor data-sharing practices. However, many of their recommendations (e.g. getting consent from each user) are infeasible, especially in the context of big data. Achieving a wide adoption of good practices requires clear and simple guidelines and standardised tools and methods to facilitate data-sharing.

Emily M. Bender and Friedman (2018) propose *data statements* as a way of designing and documenting NLP datasets. Data statements are documents that include information about most steps of the dataset creation process – including, among others, curation rationale (which texts to include and why), speaker demographic, and annotator demographic. Data statements are unstructured (i.e., written as free text) and do not have standardised representations. Gebru et al. (2020) propose a similar approach to document datasets in general, without the NLP focus. Datasheets include a series of questions that encourage dataset creators to reflect on many aspects of the dataset creation process, such as preprocessing steps, potential uses, data distribution etc. The documents are unstructured. Mitchell et al. (2019) introduce model cards as a framework to represent trained machine learning models. Although they do not propose a description of datasets, model cards are still relevant to this project's context because they consider the "end-user" (a model, in this case) of a dataset. It is important to add this perspective as data is often collected to be used to train models.

introducing innovative elements in the field, which will help to contrast what is being proposed

Although these repositories and schemata are able to represent LRs and datasets, they lack the ethical dimension that the unstructured approaches offer. They also tend to represent only the *final product* of a dataset, failing to include information to aid its design and creation. Based on the gaps I identify here, the project will add novel contributions to the field of SMD creation and sharing.

identify shortcomings, and position the proposed work

Goal and specific objectives

What is the overarching goal of the project?
What specific objectives will be pursued in the project?
Briefly describe how these objectives will be addressed.
Indicate how these objectives are relevant to the course learning objectives

3 Goal and specific objectives

The main goal of this project is to develop a metadata schema to represent social media datasets. The purpose of the schema is to enable ethical dataset creation methodologies and to improve data-sharing practices. Researchers will be able to use the schema to create knowledge graphs (KGs) to represent datasets they are creating or using. The KG creation process, although semi-automatic, will be highly participatory – the schema will encourage researchers to reflect on important issues related to SMD dataset creation. The project's overarching goal is to make researchers engage with the potential ethical problems and harms involved in SMD dataset creation. The proposed metadata schema will allow researchers to make such issues – and the actions they take to mitigate them – explicit.

what will be done

3 Goal and specific objectives

The main goal of this project is to develop a metadata schema to represent social media datasets. The purpose of the schema is to enable ethical dataset creation methodologies and to improve data-sharing practices. Researchers will be able to use the schema to create knowledge graphs (KGs) to represent datasets they are creating or using. The KG creation process, although semi-automatic, will be highly participatory – the schema will encourage researchers to reflect on important issues related to SMD dataset creation. The project's overarching goal is to make researchers engage with the potential ethical problems and harms involved in SMD dataset creation. The proposed metadata schema will allow researchers to make such issues – and the actions they take to mitigate them – explicit.

why it is needed, given what was argued before

3 Goal and specific objectives

The main goal of this project is to develop a metadata schema to represent social media datasets. The purpose of the schema is to enable ethical dataset creation methodologies and to improve data-sharing practices. Researchers will be able to use the schema to create knowledge graphs (KGs) to represent datasets they are creating or using. The KG creation process, although semi-automatic, will be highly participatory – the schema will encourage researchers to reflect on important issues related to SMD dataset creation. The project's overarching goal is to make researchers engage with the potential ethical problems and harms involved in SMD dataset creation. The proposed metadata schema will allow researchers to make such issues – and the actions they take to mitigate them – explicit.

who will benefit?
how will they benefit?

3 Goal and specific objectives

The main goal of this project is to develop a metadata schema to represent social media datasets. The purpose of the schema is to enable ethical dataset creation methodologies and to improve data-sharing practices. Researchers will be able to use the schema to create knowledge graphs (KGs) to represent datasets they are creating or using. The KG creation process, although semi-automatic, will be highly participatory – the schema will encourage researchers to reflect on important issues related to SMD dataset creation. The project's overarching goal is to make researchers engage with the potential ethical problems and harms involved in SMD dataset creation. The proposed metadata schema will allow researchers to make such issues – and the actions they take to mitigate them – explicit.

anticipated impacts

3 Goal and specific objectives

The main goal of this project is to develop a metadata schema to represent social media datasets. The purpose of the schema is to enable ethical dataset creation methodologies and to improve data-sharing practices. Researchers will be able to use the schema to create knowledge graphs (KGs) to represent datasets they are creating or using. The KG creation process, although semi-automatic, will be highly participatory – the schema will encourage researchers to reflect on important issues related to SMD dataset creation. The project's overarching goal is to make researchers engage with the potential ethical problems and harms involved in SMD dataset creation. The proposed metadata schema will allow researchers to make such issues – and the actions they take to mitigate them – explicit.

The schema will include information about the main dataset creation steps: e.g. data sampling methodology, curation rationale, and annotation process. The schema will also include information about the dataset content to improve SMD research reuse and replication. To do so while preserving user privacy, it will only include content information from aggregated data. Therefore, it will not contain instance-level content – i.e., no information about specific users or comments. The content representation will be extracted automatically from the dataset. By representing content, the KG will offer an overview of the data distribution and allow researchers to compare, albeit roughly, data set instances.

the main idea

The main specific objectives of the project are:

- Create an RDF metadata schema that includes information about the main steps of dataset creation – rationale, sampling, annotation, and content.
- Implement a method for extracting content information without compromising user privacy.
- Implement a method to compare dataset instances. Given two KG representations of datasets, the method will analyse the difference in their content and data distribution.
- Implement a method to generate search queries from a KG to be used in a social media platform to obtain a comparable data sample.

tangible objectives

Methodology

How you intend to answer your research questions.
What technologies are you going to use?
Why are they the most suitable to answer your research questions or solve your problem?
What elements of the project are risky, and how will you mitigate this risk e.g. what is your plan B?
Include an overview diagram to communicate your project architecture, components, and plans)

4 Methodology

I will start by developing a metadata framework for social media datasets. The goal of this framework is to set the conceptual foundations for the metadata schema. The framework will include information about a dataset's *creation rationale* and its *content*. To identify the most important concepts, I will critically reflect upon the literature presented in section 2, which includes multidisciplinary SMD research ranging from dataset creation to machine learning applications.

Once the framework is complete, I will develop the metadata schema to create knowledge graphs representing social media datasets. I will use existing ontologies and shared vocabularies as much as possible to ensure reusability and interoperability.

I will evaluate the metadata schema with a case study using a Twitter dataset. I will retrieve the data, create the dataset KG representation using the proposed schema, and evaluate it using two experiments. First, I will generate search queries to sample an equivalent dataset from a social media platform. Then, I will automatically profile the content of a dataset and compare it to another KG instance. The first experiment evaluates the representation power of *creation rationale* concepts. If the KG represents the dataset creation methodology well enough, it is possible to generate search queries that result in comparable data samples. The second experiment aims to evaluate if the KG can represent useful information about a dataset's content without explicitly sharing data. To delimit the scope of the project, I will focus only on Twitter data. For future work, I will extend the schema to include concepts about other platforms – like Youtube, Reddit, and Instagram.

component & how

4 Methodology

I will start by developing a metadata framework for social media datasets. The goal of this framework is to set the conceptual foundations for the metadata schema. The framework will include information about a dataset's *creation rationale* and its *content*. To identify the most important concepts, I will critically reflect upon the literature presented in section 2, which includes multidisciplinary SMD research ranging from dataset creation to machine learning applications.

Once the framework is complete, I will develop the metadata schema to create knowledge graphs representing social media datasets. I will use existing ontologies and shared vocabularies as much as possible to ensure reusability and interoperability.

I will evaluate the metadata schema with a case study using a Twitter dataset. I will retrieve the data, create the dataset KG representation using the proposed schema, and evaluate it using two experiments. First, I will generate search queries to sample an equivalent dataset from a social media platform. Then, I will automatically profile the content of a dataset and compare it to another KG instance. The first experiment evaluates the representation power of *creation rationale* concepts. If the KG represents the dataset creation methodology well enough, it is possible to generate search queries that result in comparable data samples. The second experiment aims to evaluate if the KG can represent useful information about a dataset's content without explicitly sharing data. To delimit the scope of the project, I will focus only on Twitter data. For future work, I will extend the schema to include concepts about other platforms – like Youtube, Reddit, and Instagram.

component &
considerations

4 Methodology

I will start by developing a metadata framework for social media datasets. The goal of this framework is to set the conceptual foundations for the metadata schema. The framework will include information about a dataset's *creation rationale* and its *content*. To identify the most important concepts, I will critically reflect upon the literature presented in section 2, which includes multidisciplinary SMD research ranging from dataset creation to machine learning applications.

Once the framework is complete, I will develop the metadata schema to create knowledge graphs representing social media datasets. I will use existing ontologies and shared vocabularies as much as possible to ensure reusability and interoperability.

I will evaluate the metadata schema with a case study using a Twitter dataset. I will retrieve the data, create the dataset KG representation using the proposed schema, and evaluate it using two experiments. First, I will generate search queries to sample an equivalent dataset from a social media platform. Then, I will automatically profile the content of a dataset and compare it to another KG instance. The first experiment evaluates the representation power of *creation rationale* concepts. If the KG represents the dataset creation methodology well enough, it is possible to generate search queries that result in comparable data samples. The second experiment aims to evaluate if the KG can represent useful information about a dataset's content without explicitly sharing data. To delimit the scope of the project, I will focus only on Twitter data. For future work, I will extend the schema to include concepts about other platforms – like Youtube, Reddit, and Instagram.

evaluation & approach

4 Methodology

I will start by developing a metadata framework for social media datasets. The goal of this framework is to set the conceptual foundations for the metadata schema. The framework will include information about a dataset's *creation rationale* and its *content*. To identify the most important concepts, I will critically reflect upon the literature presented in section 2, which includes multidisciplinary SMD research ranging from dataset creation to machine learning applications.

Once the framework is complete, I will develop the metadata schema to create knowledge graphs representing social media datasets. I will use existing ontologies and shared vocabularies as much as possible to ensure reusability and interoperability.

I will evaluate the metadata schema with a case study using a Twitter dataset. I will retrieve the data, create the dataset KG representation using the proposed schema, and evaluate it using two experiments. First, I will generate search queries to sample an equivalent dataset from a social media platform. Then, I will automatically profile the content of a dataset and compare it to another KG instance. The first experiment evaluates the representation power of *creation rationale* concepts. If the KG represents the dataset creation methodology well enough, it is possible to generate search queries that result in comparable data samples. The second experiment aims to evaluate if the KG can represent useful information about a dataset's content without explicitly sharing data. To delimit the scope of the project, I will focus only on Twitter data. For future work, I will extend the schema to include concepts about other platforms – like Youtube, Reddit, and Instagram.

risk mitigation

MIlestones and Deliverables

Checkpoints and deliverables in the project that can be used to measure progress. These can be in the form of modules for a software package; components or sections of a Knowledge Graph if you are constructing one; subset of a full set of questions that will be answered by certain checkpoints in the project etc.

5 Milestones & Deliverables

This section describes the checkpoints I will use to measure the progress of the project. Each milestone has specific deliverables, which I will share on a public Github repository with thorough documentation.

1. **Metadata Framework Complete (Week 1).** The literature review of related work described in section 2 is complete. The framework which describes the high-level concepts included in the metadata schema is done. The deliverable is a document describing the components of the metadata framework and the reasoning behind its creation.

2. **Metadata Schema Complete (Week 2).** The harmonisation between schemata presented in section 2 and the proposed framework is complete. The RDF Schema, which uses publicly available shared vocabularies as much as possible, is done. The deliverables are the RDF file and a document thoroughly describing how to implement the schema.

3. **Data Profiling Methods Implemented (Week 2).** The methods for extracting information about the dataset content and adding it to the KG are implemented. The deliverable is a publicly available software package that implements these methods.

4. **Case Study Complete (Week 4).** The case study, in which I apply the proposed techniques and create a KG for a real SMD dataset, is complete. The deliverables are the resulting KG, stored in a graph database such as GraphDB[1], and exploratory SPARQL queries showing some applications of the KG.

5. **Documentation Complete (Week 4).** The documentation, including the final report and a detailed description of the case study, is complete. The deliverables are all the documentation files.

component,
timeline,
and deliverable

Anticipated Results

What do you expect the main outputs of your project to be in terms of new knowledge or insights, new technology or capabilities, challenges addressed, and importance to others

6 Anticipated Results

The main output of this project will be a metadata schema for SMD. The schema will encourage researchers to engage with fundamental questions about dataset creation and make their decisions about these questions explicit. This will lead to higher data quality and better data sharing practices. To users of SMD, the project will provide ways of finding existing datasets, understanding their creation rationale clearly, comparing data samples, and sharing insights about their current state. The project will be an initial step towards more ethical, reproducible, and reusable SMD research.

References

A section containing proper references for inline citations. Use a citation manager like Endnote, Mendeley, or Paperpile to insert and automatically format citations and references.

References

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" en. In: p. 14.

Bender, Emily M. and Batya Friedman (2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science". In: *Transactions of the Association for Computational Linguistics* 6, pp. 587–604. DOI: 10.1162/tacl_a_00041.

Breuer, Johannes, Libby Bishop, and Katharina Kinder-Kurlanda (Nov. 2020). "The Practical and Ethical Challenges in Acquiring and Sharing Digital Trace Data: Negotiating Public-Private Partnerships". en. In: *New Media & Society* 22.11, pp. 2058–2080. ISSN: 1461-4448. DOI: 10.1177/1461444820924622.