Gottfried Wilhelm Leibniz Universität Hannover
Faculty of Electrical Engineering and Computer Science

Scientific Knowledge fit for society - Scoring scientific accuracy in climate change related news articles

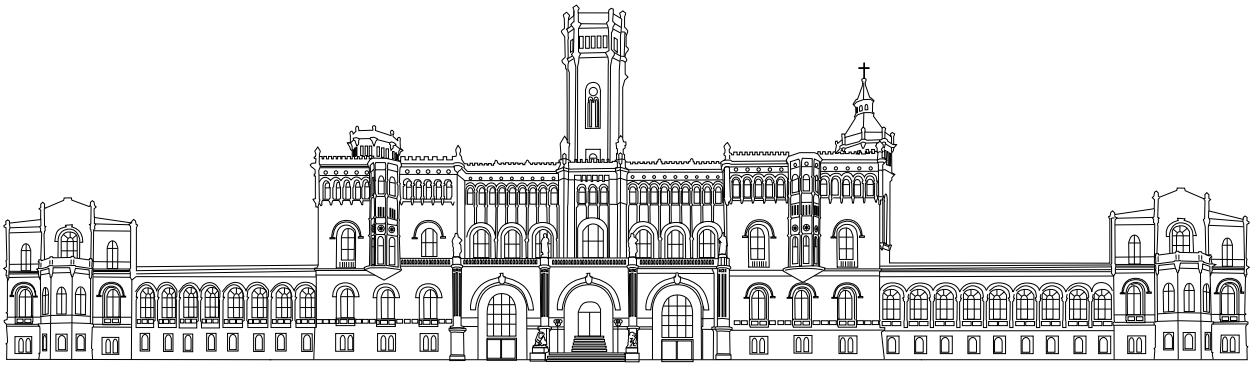
A thesis submitted in fulfillment of the requirements for the
Master of Science in Computer Science

BY

Constantin Sebastian Tremel
Matriculation number: 10014864
E-mail: constantin.tremel@stud.uni-hannover.de

First evaluator: Prof. Dr. Sören Auer
Second evaluator: Dr. Markus Stocker
Supervisor: Tim Wittenborg

30. April 2024



Erratum

Erratum for: Scientific Knowledge fit for society - Scoring scientific accuracy in climate change related news articles

In this master thesis, errors were identified in Chapter 3, Subsection 3.2.2, second paragraph on page 16 and Chapter 5, first sentence on page 31. The errors in the initial publication did not affect the master thesis.

The sentences in Chapter 3 were mistakenly left in and only intended for the writing process. They should have been deleted from the text after being processed. The second paragraph in Subsection 3.2.2 on page 16 contained the following sentences: "*The Washington Post employs a rating system that ranges from one to four Pinocchios, while Politifact utilises a "Truth-o-meter" that ranges from "True" to "Pants on fire" [40, 32].*", followed by the erroneous segment "*Ground News isn't a news source bound by the political leanings of our writers or parent company. Ground News isn't a news aggregator, throwing hundreds of headlines at you that barely skim the surface. Ground News isn't a fact-checker, because we believe in empowerment, not enablement. We are a guide for news readers. Neutral in assessment, efficient in consumption, all in service of empowering you to make educated decisions for yourself.*", followed by "*Additionally, there is Ground News, which is not a fact-checking site but rather evaluates the biases and personal affiliations of authors and news agencies [29].*". The source excerpt, an artifact from the online research [29], was mistakenly left in the text instead of being deleted after being paraphrased into the final sentence. By this Erratum this text segment was removed from the text to present it in its intended form:

"*The Washington Post employs a rating system that ranges from one to four Pinocchios, while Politifact utilises a "Truth-o-meter" that ranges from "True" to "Pants on fire" [40, 32]. Additionally, there is Ground News, which is not a fact-checking site but rather evaluates the biases and personal affiliations of authors and news agencies [29].*"

A similar error occurred in Chapter 5, which started with the following sentence: "*This section presents your implementation.*". Similar to the previous segment, this was only intended for drafting the thesis, not for the final iteration. By this Erratum this sentence was removed from the text.

Declaration of Authorship

I, Constantin Sebastian Tremel, declare that this thesis titled, 'Scientific Knowledge fit for society - Scoring scientific accuracy in climate change related news articles' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Constantin Sebastian Tremel

Signature: _____

Date: _____

“[Wissenschaft] ist aber die Basis einer aufgeklärten, freien Gesellschaft. Und auch wenn wissenschaftliche Evidenz niemals alleine irgendeine gesellschaftliche Debatte entscheiden kann, sollte sie in unseren gesellschaftlichen Debatten der kleinste gemeinsame Nenner sein. Nur, wenn wir auf Basis einer kleinsten gemeinsamen Wirklichkeit streiten, streiten wir nicht nur auf der Stelle, sondern auch vorwärts.”

— Mai Thi Nguyen-Kim (2021)

Acknowledgements

I would like to express my gratitude to my supervisor, Tim W., for his invaluable support, guidance, and insights. His guidance has been instrumental in teaching me the fundamentals of scientific working practices and information processing. Additionally, he has spent considerable time discussing my concerns and helping me to focus on the goal. I am also grateful to my second evaluator, Dr. Markus S., for including my search for knowledge and enabling me to present at the ORKG Team calls. I would also like to express my gratitude to all of the experts from the ORKG Team who took the time to answer my questions and participate in an interview. Finally, I would like to thank my family and friends for their support, patience, and assistance in proofreading my thesis. Their help has been invaluable during the compilation of this thesis.

Abstract

Scientific Knowledge fit for society - Scoring scientific accuracy in climate change related news articles

The quantity of information is increasing exponentially, and there is a vast amount of content viewed on the internet that lacks an indicator as to whether it is scientifically accurate and correct or scientifically inaccurate and incorrect. This thesis proposes the development of an indicator of scientific accuracy in online media. This should help in public debates and help in the detection of misinformation. The thesis presents a baseline score and clear interfaces for further improvement. The necessity for such a score has been validated by a user survey, and the employed methodologies were evaluated and updated through interviews with experts from the ORKG team. Furthermore, an overview of the knowledge required to conduct research in this field and a discussion for future work is provided.

Keywords: scientific accuracy, climate change, validation, knowledge graph, science communication

Zusammenfassung

Die Menge an Informationen nimmt exponentiell zu, und es gibt eine riesige Menge an Inhalten im Internet, für die es keinen Indikator dafür gibt, ob sie wissenschaftlich genau und korrekt oder wissenschaftlich ungenau und falsch sind. In dieser Arbeit wird die Entwicklung eines Indikators für wissenschaftliche Genauigkeit in Online-Medien vorgeschlagen. Dies sollte in öffentlichen Debatten und bei der Aufdeckung von Fehlinformationen helfen. Die Arbeit legt einen Grundstein und schafft klare Schnittstellen für weitere Verbesserungen. Die Notwendigkeit eines solchen Scores wurde durch eine Nutzerbefragung validiert, und die verwendeten Methoden wurden durch Interviews mit Experten aus dem ORKG-Team evaluiert und aktualisiert. Darüber hinaus wird ein Überblick über das für die Forschung in diesem Bereich erforderliche Wissen gegeben und eine Diskussion über künftige Arbeiten geführt.
Keywords: wissenschaftliche Genauigkeit, Klimawandel, Validierung, Wissensgraph, Wissenschaftskommunikation

Contents

List of Figures

List of Tables

List of Listings

Acronyms

AMR Abstract Meaning Representation. 35, *Glossary:* AMR

API Application Programming Interface. 40, *Glossary:* API

AR6 Assesment Report 6. 6, 33, 35, 37, 43, IX, *Glossary:* AR6

EDMO European Digital Media Observatory. 14, 15, *Glossary:* EDMO

EFCSN European Fact-Checking Standards Network. 14, 15, 69, *Glossary:* EFCSN

IFCN International Fact-Checking Network. 14, 16, *Glossary:* AR6

IPCC Intergovernmental Panel on Climate Change. 6, 33, 35, 37, 43, 66, IX, *Glossary:* IPCC

LLM Large language model. 7, 12, 24, 35, 38, 39, 43, 46, 49, 52, 58, 63, *Glossary:* LLM

NER Named Entity Recognition. 35, 52, *Glossary:* NER

NLP Natural Language Processing. 7, 38, 66, *Glossary:* NLP

ORKG Open Research Knowledge Graph. 6, 8, 46, 47, 56, *Glossary:* ORKG

OWL Web Ontology Language. 8, *Glossary:* OWL

RDF Resource Description Framework. 8, 9, 76, *Glossary:* RDF

SPARQL SPARQL Protocol And RDF Query Language. 9, 40, *Glossary:* SPARQL

SRL Semantic Role Labeling. 35, *Glossary:* SRL

URI Universal Resource Identifier. 76

Chapter 1

Introduction

1.1 Motivation

The quantity of information is increasing exponentially. "According to Statista, the total amount of data to be created, captured, copied, and consumed globally in 2023 is 120 zettabytes a number projected to grow to 181 zettabytes by 2025" [20]. To illustrate this in a more relatable format, it can be observed that, on average, every minute of the day, among other things, 43 years of streaming content are watched, 25.1 million hours are spent on the internet and over 280 million messages are sent via email, WhatsApp, X and Instagram [20]. It is evident that the majority of content viewed on the internet lacks an indicator as to whether it is scientifically accurate or correct. A significant proportion of the population places trust in media that employs citations. However, there are instances where content that includes sources is nevertheless erroneous. Not all individuals have the time or expertise to verify every source. Furthermore, the phenomenon of false balance has the potential to influence public opinion by amplifying perceptions of disagreement and uncertainty among experts [41].

Climate change represents a significant threat to the planet, with a rapidly closing window of opportunity to secure a future for all [9]. "The choices and actions implemented in this decade will have long-term impacts, both now and in the future" [9]. Nevertheless, the emissions gap persists even when all planned policies globally are implemented in full, which is not even the current trend [23]. "If current policies are continued, global warming is estimated to be limited to 3°C" [23]. Every increment of global warming has the potential to escalate risks and related losses [9]. It is futile to engage in debates that do not align with the overwhelming consensus within the scientific community and the necessity for action [9, 15].

1.2 Goal

The objective of this thesis is to develop an indicator of scientific accuracy in online media. The evaluation process should be automated as long as the resulting score maintains a certain quality.

1.3 Research Questions

The research questions addressed by this thesis are: **RQ1)** How can natural language processing and knowledge graphs help verify the consistency of secondary literature with scientific findings? **RQ2)** How can scientifically accurate media on climate change be identified?

1.4 Structure

This paper provides an overview of the state of the art in the research areas that were combined to achieve this goal. There is a description of the basic knowledge needed to understand the concept of this paper in Chapter 2 Background. In addition, there is an overview of work by others that addresses similar or sub-problems of the goal of this paper in Chapter 3 Related Work. The Chapter 4 Approach is an overview of the complete concept to achieve the goal, where the interfaces are described. The Chapter 5 Implementation describes working demos for different steps of the approach and also goes through an example use case. Finally, the Chapter 6 Evaluation provides verification from expert interviews and evaluation of demand from a user survey. The Chapter 7 Discussion discusses the benefits, limitations and future work for this thesis. Finally Chapter 8 Conclusion summarises the findings.

Chapter 2

Background

This chapter introduces the topics needed to understand this thesis, of which most are still under active research. The topics are structured into four main categories. At first the differences between the many kinds of information and their sources will be described in Section 2.1 Information source. Section 2.2 Information extraction is about gaining the information from a source. How to prepare that information in a way that further processing is possible will be described in Section 2.3 Knowledge base. Lastly there is Section 2.4 Scoring scientific accuracy about scoring information based on the scientific accuracy by using the ground truth saved in the knowledge base.

2.1 Information source

When talking about Information sources it is important to define some vocabulary. In Figure 2.1 an overview of information concepts is shown [25, 68].

Within this "map of information concepts" information is a term consisting of data in different forms (analogue, digital and binary) serving different functions (primary, secondary, meta, operational and derivative) [68].

"For example, information from an electronic document contains digital data structured according to certain semantic and content standards. While in more complex forms, information can be knowledge existing as internal phenomena to the human mind consisting of analogue (i.e. biological) data" [68].

The definitions of the functional data types are further explained in the Figure 2.2.

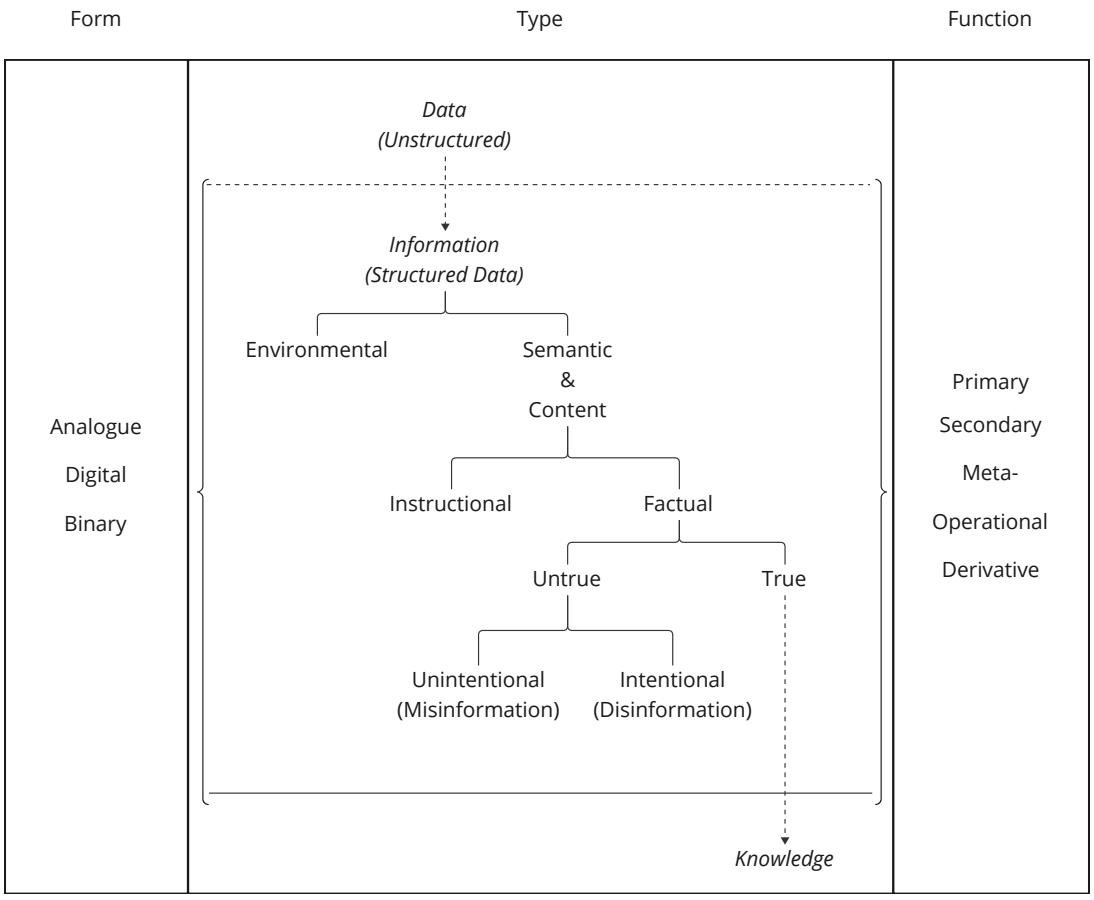


Figure 2.1: "A map of information concepts" reprinted from [25, 68]

Definitions. The most important definitions for this thesis are:

- **Data** is a unstructured representation of circumstances.
- **Information** is structured data.
- **Misinformation** is an unintentional untrue factional information.
- **Disinformation** is an intentional untrue factional information.
- **Knowledge** is a true factional information. Since truth is incomplete, **knowledge** can only be based on the best approximation of the truth.

Functional Type	Descriptions	Examples
Primary data	The principle data stored in a database.	A simple array of numbers in a spread sheet, or string of zeros and ones
Secondary data	The converse of primary data, constituted by their absence.	An engine fails to make any noise thus indicating the secondary information about the flat battery
Metadata	Data that indicates the nature of some other (usually primary) data. They describe properties such as location, format, updating, availability usage restrictions and so forth.	The copyright note on the car's operation manual
Operational data	Data regarding the operations of the whole data systems	Computing logic that instructs the system to act in a certain way a in given condition.
Derivative data	Data that can be extracted from some data whenever the latter are used as indirect sources in search of patterns, clues or inferential evidence about other things than those directly addressed by data themselves.	One's whereabouts can be derived from where he/she use credit cards.

Figure 2.2: "Functional data types" reprinted from [25, 68]

Terms about information that are missing in that overview but also relevant for this thesis are:

- An **opinion** describes a certain view on a topic that is up for discussion. It is formed through individual life experiences. To be meaningful an opinion needs to be supported with knowledge. Opinions are important in areas where the information basis is scarce or in areas where a decision needs to factor in personal feelings and views.
- A **hypothesis** is a reasoned assumption about factual information and needs further research to show whether it is knowledge or misinformation.
- A **diversion** is a factional information that is intentionally not relevant to the topic under discussion.
- **Distracting** is a factional information that is unintentionally not relevant to the topic under discussion.

Information to be used as the ground truth to score popular media needs to be carefully selected. To ensure a high quality the information should be stem from **peer reviewed primary literature** published in an **reputable journal** and be **reproducible** and **traceable**. The issues of peer reviews are discussed in Section 7.3 Ground truth. The greater the number of validations, the greater the probability that the information will be regarded as knowledge. Consequently, if numerous publications and numerous peer review procedures yield identical conclusions, it can be inferred that the information in question is indeed knowledge. The probability of accuracy and the difference between modelled and actual measurements can also be used to determine the quality of the information. A valuable resource to create an overview for which information has been produced by different researches, with which methods and accuracy, is the Open Research Knowledge Graph (ORKG). More information about the ORKG will be presented in Section 2.3 Knowledge base. There are rankings of journals that include criteria like citations to measure the impact and quality of a journal. Some rankings are done by *Science Watch*[1], *Observatory of International Research (OOIR)*[2], *Google*[3] and *Research.com*[4], *Scimago Journal Rank*[5]. In the field of climate change, for instance, in "Science" and "Nature" are highly regarded journals.

With this approach popular media is supposed to be evaluated which consists of secondary literature and tertiary literature. Different types of popular media include newspapers, discussions, social media (videos, podcast, comments).

Intergovernmental Panel on Climate Change. The IPCC releases synthesis of scientific knowledge concerning climate change [9]. The IPCC AR6 is the main source for the ground truth of this approach and has a spread of different information. The AR6 is generally composed of three parts: *current state, projected scenarios, mitigation options/solutions* [9]. In different parts of the report there is a different levels of detail and statements are labeled with confidence levels. The IPCC is a collective of researchers all over the world. The report was written in consultation of governments and should be accepted widely. Also it is a collection of work from a big research group.

[1]Science Watch Ranking https://archive.sciencewatch.com/ana/st/climate/journals/

[2]Observatory of International Research Ranking https://ooir.org/journals.php?field=Multidisciplinary&category=Environmental+Sciences&metric=jif

[3]Google Ranking https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=bio_environmentalsciences

[4]Research.com Ranking https://research.com/journals-rankings/environmental-sciences

[5]Scimago Journal Ranking https://www.scimagojr.com/journalrank.php?category=2306

2.2 Information extraction

Natural language processing (NLP) methods are used to translate natural language into formats that can be understood of machines. For information extraction often entities and their relations are extracted. Two entities and their relation are a triple. Entities mostly are subjects and objects in natural language that can contain information about who did something or what, when, where and how something happened. The translation of natural language into a machine actionable format is not always easy. For example coreferences are understood and used by humans intuitively but need to be resolved for machines. Also linking entities can enrich the information base for machines to understand synonyms, structural information and concepts.

One of the tasks in recognizing relations and entities is Part-of-Speech (POS) tagging. POS tagging still has to deal with challenges including false-positive rates and tagging unknown words [10].

Machado and Ruiz [43] reveal that LLMs can leverage prior knowledge from existing tagged data sets [43].

Information extraction is often trained specifically for certain domains. To handle web-scale corpora Open Information Extraction (Open IE) is implemented unbounded of relations and domain-specific training data [24].

Model types. When working with NLP there are different types of models specialised to deal with certain challenges. There is no silver bullet approach that works well for everything. There are always pros and cons that have to be weighed against each other. Especially transformer are mentioned a lot, a "mathematically precise, intuitive, and clean description of the transformer architecture" is presented by Turner [67].

At the moment Large language model (LLM)s are popular in the field of artificial intelligence where models such as T0, LLama, Palm, GPT-3, GPT-4, or instruction finetuned models, such as ChatGPT demonstrated their exceptional capabilities in various domains, including language translation, summarization, and question answering [7, 42]. As mentioned before LLMs also come with challenges which will be discussed in Section 7.3 Large language models. Another importan model that should be mentioned is BERT which stands for Bidirectional Encoder Representations from Transformers [18]. BERT is a finetuning based representation model that performs well on sentence-level and token-level tasks [18].

2.3 Knowledge base

For knowledge graphs sometimes conflicting definitions have emerged [31]. Here a knowledge graph is viewed as a graph of data from the real world, whose nodes represent entities of interest and whose edges represent relations between these entities. [31]. Resource Description Framework (RDF) triples are comprised of a subject, predicate and object, whereby the subject and object are entities described and the predicate with a relation between these entities. All RDF entities and relations have a unique identifier. In this form information can be saved implicitly through the structure which also enables efficient algorithms that use the graph structure. Often ontologies are used. A knowledge graph comprises two types of entities: classes and instances. Class relationships represent general structural relationships between classes, exemplified by the relationship between a house and a room. Instance relationships, on the other hand, represent concrete instances that can be used to infer general information about the class relationships.

In the context of knowledge graphs, ontology's are employed to model the fundamental relationships that exist within the graph. Instances are used to provide concrete examples of entities and their relationships.

Ontology. An ontology is a formal representation and can be used to model graph data. Ontologies often have a scope of terms in specific domain and can be used to automate entailment. Agreement upon ontologies enables the interoperability of knowledge graphs. Amongst the most popular ontology languages is the Web Ontology Language (OWL). [31]

Open Knowledge Research Graph. In a flood of pseudo-digitized PDF publications around 2.5 million new research contributions every year the ORKG offers a way to change the transfer of knowledge fundamentally by representing scholarly contributions in a structured and semantic way as a knowledge graph [4].

The ORKG offers exploration services such as literature comparisons, contribution services and curation services [39]. Possible intuitive access interfaces include tabular comparison of contributions according to various characteristics, domain-specific (chart) visualizations or answering of natural language questions.[4] ORKG users interact with the front end to create research contribution descriptions in a step by step manner or to directly find similar contributions (and related papers), thus enabling efficient state-of-the-art comparison and literature review [38].

2.4 Scoring scientific accuracy

The basis of the scientific accuracy score on a statement level is information in a knowledge graph. To calculate a score graph analytics are needed. The nature of graphs naturally lends conclusions about nodes and edges based on the topology of the graph [31]. Many techniques from related areas of graph analytics such as graph theory and network analysis are available [31]. Techniques include Centrality, Community detection, Connectivity, Node similarity, Path finding [31]. Especially path finding will be used in this approach.

For the scoring characteristics on a sentence, paragraph or document level additional natural language processing techniques should be considered.

Querying. RDF and RDF* knowledge graphs can be queried with SPARQL Protocol And RDF Query Language (SPARQL) or SPARQL* respectively.

When given to entities that are somehow connected with intermediate nodes it is possible to query the path between them. For this path the path length describes the number of intermediate nodes between them. There can be multiple paths connected the same entities.

The node density is the number of incoming and outgoing relations of an node (entity). This can be split into incoming and outgoing node density.

2.5 Tools

A range of tools has been used in this work. The most important ones are GraphDB, visual studio code and anaconda. All the libraries that were used will be mentioned in the Chapter 5 Implementation.

Chapter 3

Related Work

This chapter provides an overview over what others have already done. There are different sections for publications covering different combinations of categories mentioned in Chapter 2 Background.

Section 3.1 Climate change knowledge graph construction includes full pipeline approaches covering parts of the categories Information source, Information extraction and Knowledge base. The focus of Section 3.2 Fact checking focuses on the category Scoring scientific accuracy. The combination of the related work in Section 3.1 and Section 3.2 has the greatest similarity to the work presented in this thesis. Although there is some degree of similarity, no work was identified that fully covers the approach of this thesis.

The subsequent sections show ways to further explore the categories of Information source, Information extraction and Knowledge base. Description of datasets and sources of scientific knowledge that have been considered as a potential source for assessing scientific accuracy for the specific domain of climate change are presented in Section 7.4 Data sets. Section 3.3 Knowledge base includes work on information extraction and knowledge graph construction.

3.1 Climate change knowledge graph construction

KnowUREnvironment. Islam et al. [35] propose "a knowledge graph for climate change and related environmental issues, extracted from the scientific literature". According to them they extracted 411,860 RDF triples that are evaluated by humans and have a syntactically and factually correctness (81.69% syntactic correctness and 75.85% precision) [35].

Figure 3.1 is taken from their paper and shows an overview of their methods for

extracting triples and selecting only the trusted ones.

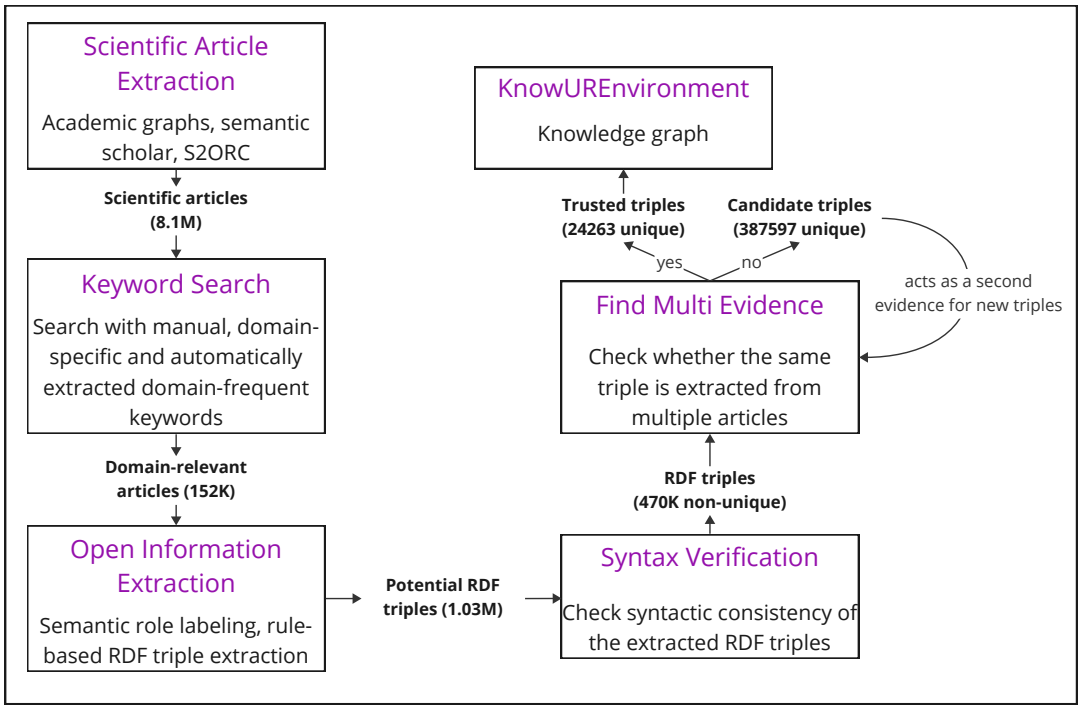


Figure 3.1: "An brief overview of the pipeline employed to construct the KnowUREnvironment knowledge graph" reprinted from [35]

Their process is fully automated and the triples are freely available via an open GitHub repository[1]. This opens the possibility to add additional functionality as proposed in this thesis. Also saving the triples with a reference to the paper Id and sentence number enables to remain reproducible and traceable. Since this is a key interest for information sources in this thesis continuing on their basis could have been saving a lot of work. Unfortunately, most of the triples are not usable for this thesis as they are unhelpful without context.

Their full automation also leads to many triples being lost in the process. They base the reliability of triples on a selection of sources with the same information. This can become critical in at least two cases. Firstly, highly covered topics could lead to

[1]Extracted final triples: https://github.com/saiful1105020/KnowUREnvironment/blob/main/final_tuples_double_evidence.csv

unjustified trust. Secondly, information can be domain relevant and true and still be missed, for example if it is not repeated often, or if a synonym is used. The selection of trustworthy sources should be done by emphasising the origin of the sources (e.g. using only peer-reviewed papers that are reproducible and traceable). Synonym detection would be a worthwhile addition, but its absence could be compensated for by a larger knowledge graph and more sources. The less expensive option should be preferred.

Climate change knowledge graph. Zhang [69] developed "a reproducible automated semantic network approach using natural language processing and open information extraction tools to construct knowledge graphs on climate change news coverage in 5 countries using news articles (N = 19,684) from 1990 to 2020." Although Zhang [69] focuses on a different aspect of climate change news, his pipeline shown in Figure 3.2 requires a similar workflow. He is also working in the domain of climate change related news and extracting triples. In another context, this pipeline can be used extract climate change information in news to score their scientific accuracy. Potentially this pipeline is also usable for the creation of a knowledge graph consisting of scientific publications.

Another aspect that could potentially be useful in this thesis is the usage of multiple knowledge graphs based on different publication dates to allow a time-sensitive observation of climate facts. In the context of climate change this is useful because it is an empirical research field where up-to-date research can change due to assumptions and models being updated. Following a request, the author of this paper was willing to share the source code of this pipeline. However, no code was ultimately shared. Additionally, the author suggested that it would be beneficial to explore LLM technology as a potential avenue for enhancing this approach.

NeuralNERE. As shown in 3.3 this is a proposition of "an end-to-end Neural Named Entity Relationship Extraction model (called NeuralNERE) for climate change knowledge graph (KG) construction, directly from the raw text of relevant news articles" [51].

The proposed pipeline and paper outline concrete steps and describe the problem statement and usefulness of a knowledge graph about climate change. After contacting the authors, there was no response. At the time of writing, they have not released a working tool or source code.

If the pipeline is ever implemented it could be tested whether can improve parts of this thesis approach. Also worth looking at is their published data set, which

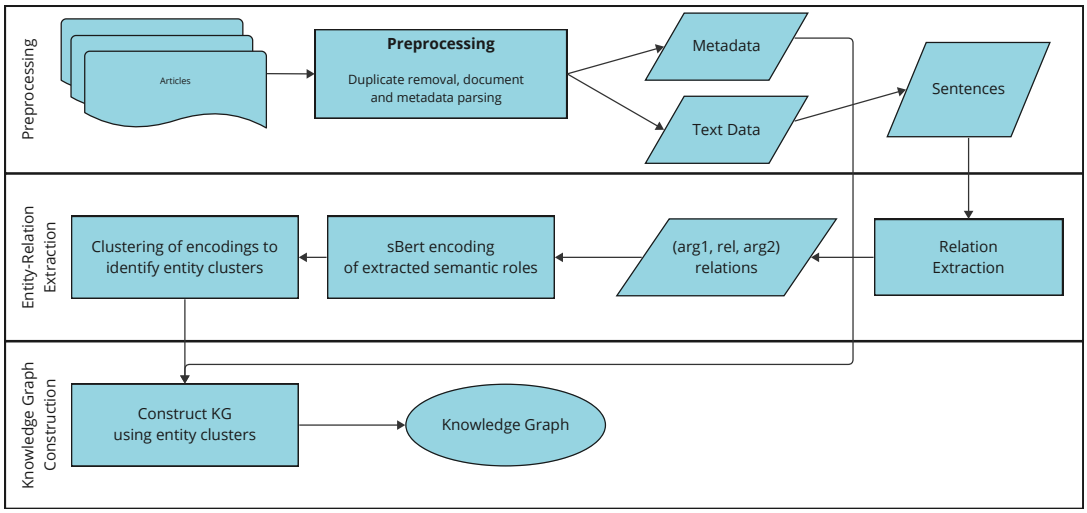


Figure 3.2: "An overview of the knowledge graph construction pipeline employed to construct the climate change knowledge graph" reprinted from [69]

"contains over 11k climate change news articles scraped from the Science Daily website"[51], it is described in Section 7.4 Data sets.

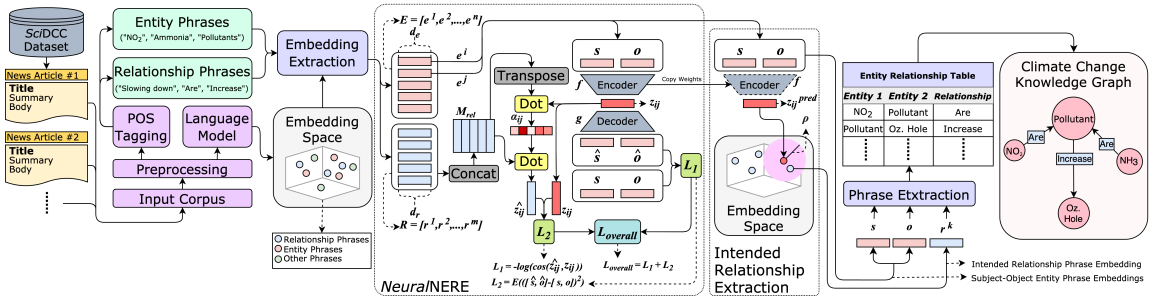


Figure 3.3: NeuralNERE Model Architecture [51]

3.2 Fact checking

Fact checking alone could be a thesis in itself. There are an incredible number of fact-checking sites and they vary greatly in the type of information they check, their grading systems and the level of automation. Ciampaglia et al. say that "fact checking

by expert journalists cannot keep up with the enormous volume of information that is now generated online" [11]. This section begins with an overview of the vast number of fact-checking sites. It then describes some exemplary approaches.

3.2.1 Networks

Some associations and communities are collaborating to address the considerable amount of verification that is required. The International Fact-Checking Network (IFCN), the European Fact-Checking Standards Network (EFCSN) and European Digital Media Observatory (EDMO) are examples of networks that validate hundreds of organisations by upholding certain standards. Some fact checkers are verified by more than one of these networks. A lot of organizations are also connected in IFCN, EFCSN and EDMO. For example *Correctiv*, *Delfi* and *dpa*.

International Fact-Checking Network. The IFCN[2] at Poynter was established in 2015 with the objective of uniting the growing global community of fact-checkers [33]. The IFCN enables fact-checkers through networking and collaboration, as well as through the promotion of training and global events [33]. Two of these events are the International Fact-Checking Day[3] and the World's Largest Annual Conference named "Global Fact"[4] [33]. With the support of partner corporations, the IFCN is able to award grants [14]. In 2021, a grant program was established for climate change-related facts which was funded by facebook with 800,000 $ and is closed now. One notable winner is Science Feedback, which employs a methodology[5] and evaluation criteria that can be used to assess the scientific accuracy of proposed visualisations. Fact checkers can become verified signatories of the IFCN code of principles[6] after applying. The assessment conclusions are publicly available. At the time of writing, there were 155 verified signatories[7] from a multitude of countries.

European Fact-Checking Standards Network. The EFCSN is another non-profit fact checking network with 44 verified members[8] at the time of writing. Their staff, statues and governance body are all accessible in a transparent manner. The

[2]IFCN at Poynter https://www.poynter.org/ifcn/
[3]Fact checking day https://www.poynter.org/event/international-factchecking-day-2024/
[4]Global Fact https://www.poynter.org/event/globalfact-11/
[5]Science Feedback methodology https://science.feedback.org/process/
[6]IFCN code of principles https://ifcncodeofprinciples.poynter.org/
[7]IFCN signatories https://ifcncodeofprinciples.poynter.org/signatories
[8]EFCSN members https://efcsn.com/verified-members/

EFCSN is currently engaged in a number of projects, including Elections24Check and AI@EUElections [53]. Elections24Check is a project in collaboration with Google News Initiative that has developed a fact-checking database focused on the verification of claims related to the 2024 European Elections [53]. Researchers and journalists may access the underlying data via a dedicated form.[9] AI@EUElections is a project that has received support from Meta to conduct training sessions with the objective of equipping fact-checkers across Europe to identify AI-generated and digitally altered content [53].

European Digital Media Observatory. EDMO brings together fact-checkers, media literacy experts, and academic researchers to gain insight into and analyse disinformation [22]. A network spanning 28 countries in the EU and the European Environment Agency (EEA) covers current topics such as climate change, the war in Ukraine, and the 2024 European elections [22]. EDMO has established a platform that guides fact checkers and researchers through a series of detailed steps, from the receipt of potential disinformation leads to the organisation of team members tasked with analysis and finally to the utilisation of a continuously enhanced toolset at the verification stage [21]. EDMO's platform aims to facilitate the 'human in the loop' concept because they are uncertain as to the feasibility of automated verification and fact-checking [21].

3.2.2 Fact Checkers.

The fact-checkers approved by the networks are transparent in their work. They employ similar methods, with the methods of Science Feedback serving as an illustrative example.

1. Selection of an item to assess based on topic, relevance and potential influence. Suggestions are welcome [52].

2. Investigating fact-based assertions and scientific reasoning. Experts and original authors of scientific statements are requested [52].

3. Writing the review giving context to the current state of knowledge in science [52].

[9] Elections24Check https://docs.google.com/forms/d/e/1FAIpQLSdE64dscZd2v9gcFnuUHDM-_oomaxlLPRCt6FczXVsDM1iwQw/viewform

4. Publicizing the review on their website, across their social platforms, and share it with media and scientific partners as well as the journalists and editors of the outlet of the original article [52].

The approval of the IFCN enables organisations to participate in Meta's third-party fact-checking program [60]. However, it should be noted that fact checkers are only permitted to identify instances of misinformation and that Meta takes further steps to prevent the spread of such content [60]. The primary distinctions between fact-checking organisations lie in the topics they cover and the scoring systems they utilise. These organisations tend to have a focus on national news and cover similar topics, for example climate, health and politics. The scoring mechanisms are similar in that there is always a considerable degree of differentiation, with the distinction between true and false not being the sole consideration. Instead, there is a spectrum of varying degrees of truth, including neutral and falsehood [59]. Furthermore, the presence of misleading information or a lack of context is also taken into account [59]. The Washington Post employs a rating system that ranges from one to four Pinocchios, while Politifact utilises a "Truth-o-meter" that ranges from "True" to "Pants on fire" [40, 32]. Additionally, there is Ground News, which is not a fact-checking site but rather evaluates the biases and personal affiliations of authors and news agencies. This is another aspect that should be considered when evaluating the media [29].

3.2.3 Automated

Computational Fact Checking from Knowledge Networks

Ciampaglia et al. propose a way to computationally approximate the the complexities of human fact checking. They are using the "shortest path between concept nodes under properly defined semantic proximity metrics on knowledge graphs" [11]. They worked in the fields of "history, entertainment, geography, and biographical information" and checked information using a knowledge graph extracted from wikipedia infoboxes[11]. It is important to note that Wikipedia infoboxes contain structured information that is easier to extract and could be useful as a starting point. However, in the end, they must be exchanged for sources of primary literature. Additionally, in the domain of climate change, there are entities that require additional structural information in comparison to named entities extracted from Wikipedia infoboxes.

Fact Extraction and Verification (FEVER)

In their work, Thorne et al. introduce a dataset for verification against textual sources, FEVER: Fact Extraction and VERification [66]. This dataset consists

of 185,445 claims extracted from Wikipedia and subsequently verified [66]. The claims are classified as SUPPORTED, REFUTED or NOTENOUGHINFO by annotators [66]. According to the authors, the task is challenging yet feasible. The best performing system achieved an accuracy of 31.87% [66].

Bekoulis, Papagiannopoulou, and Deligiannis call the FEVER dataset the "most well-studied and formally structured dataset on the fact extraction and verification task" [6]. With their work they want to shows issues with existing research, and be a structured guide for new researchers to the field of fact extraction and verification [6].

There also is a version for the domain of climate change named climate-fever publicly available [19]. The methodology of fever is adapted to real life claims collected from the Internet [19]. Even with the expertise of renowned climate scientists Diggelmann et al. were faced with difficulties of the "surprising, subtle complexity of modeling real-world climate-related claims" [19]. For instance, for the type of disputed claims which are absent in the FEVER dataset [19].

"*An article in Science magazine illustrated that a rise in carbon dioxide did not precede a rise in temperatures, but actually lagged behind temperature rises by 200 to 1000 years.*" [19]

Their system provides both supporting and refuting evidence and labelled as such by the annotators [19]. These real-life claims in general and accounting for the specific characteristics of claims related to climate change is complex [19].

Full Fact AI. With Google's support and international experts, they used machine learning to build tools to improve scale fact checking, which are now available for other organisations to use via a paid licence [26]. They aim for a tool to combat the following tasks.

- Know the most important thing to be fact checking each day [26]
- Know when someone repeats something they already know to be false [26]
- Check things in as close to real-time as possible [26]

Their process consists of collecting and monitoring the data, followed by identifying, labelling and matching claims. The collected data can be taken from speech on live TV, online news sites, and social media pages as well as defined by users themselves using a UI. They split down text down to individual sentences which are enriched through a number of steps.

The system distinguishes between a multitude of claim types, including those pertaining to quantities, causal relations, and predictive claims about the future [26]. To assist users in identifying potentially valuable claims, a classifier was developed

using the BERT model and subsequently fine-tuned with the system's own annotated data [26]. Claims are then evaluated to ascertain their alignment with previously fact-checked content [26]. The complexity of claims varies due to the nuances of language employed to describe them [26]. Some claims are more straightforward to model than others due to their specificity and ambiguity [26]. The prediction of a match or no-match for sentences is conducted by another Bert model, which is then expanded by a range of other techniques, including entity analysis [26]. This involves counting the number of instances where both sentences contain the same sample numbers, people, organisations, and so forth [26]. In combination, these stages consistently identify instances of a claim, even if the words used to describe it differ [26]. The fact-checking process is often conducted offline [26].

They still talk about the limitations of automated fact checking and say "Humans aren't going anywhere anytime soon—and nor would we want them to" [26].

3.3 Knowledge base

3.3.1 Knowledge graph construction

SCICERO. Dessí et al. [17] present a knowledge graph generation approach that takes input text from research articles and generates a knowledge graph of research entities. SCICERO employs Deep Learning Transformer models to parse the content of scientific papers in order to extract information and render the written content machine-actionable. SCICERO has been utilised to generate a knowledge graph comprising approximately 10M entities pertaining to Computer Science. [17] The system "has been evaluated on a manually generated gold standard of 3,600 triples that cover three Computer Science subdomains (Information Retrieval, Natural Language Processing, and Machine Learning) obtaining remarkable results [17]". The workflow of SCICERO is shown in Figure 3.4 and available on Github[10].

3.3.2 Ontologies

There are some ontologies that are related to climate change:

Climate Change Timeline (CCTL) Ontology Pileggi and Lamia [50] adopt an ontological approach to construct a knowledge base on climate change-related facts. The resulting ontology is structured as a timeline, which aims to describe the

[10]SCICERO source code https://github.com/danilo-dessi/SKG-pipeline

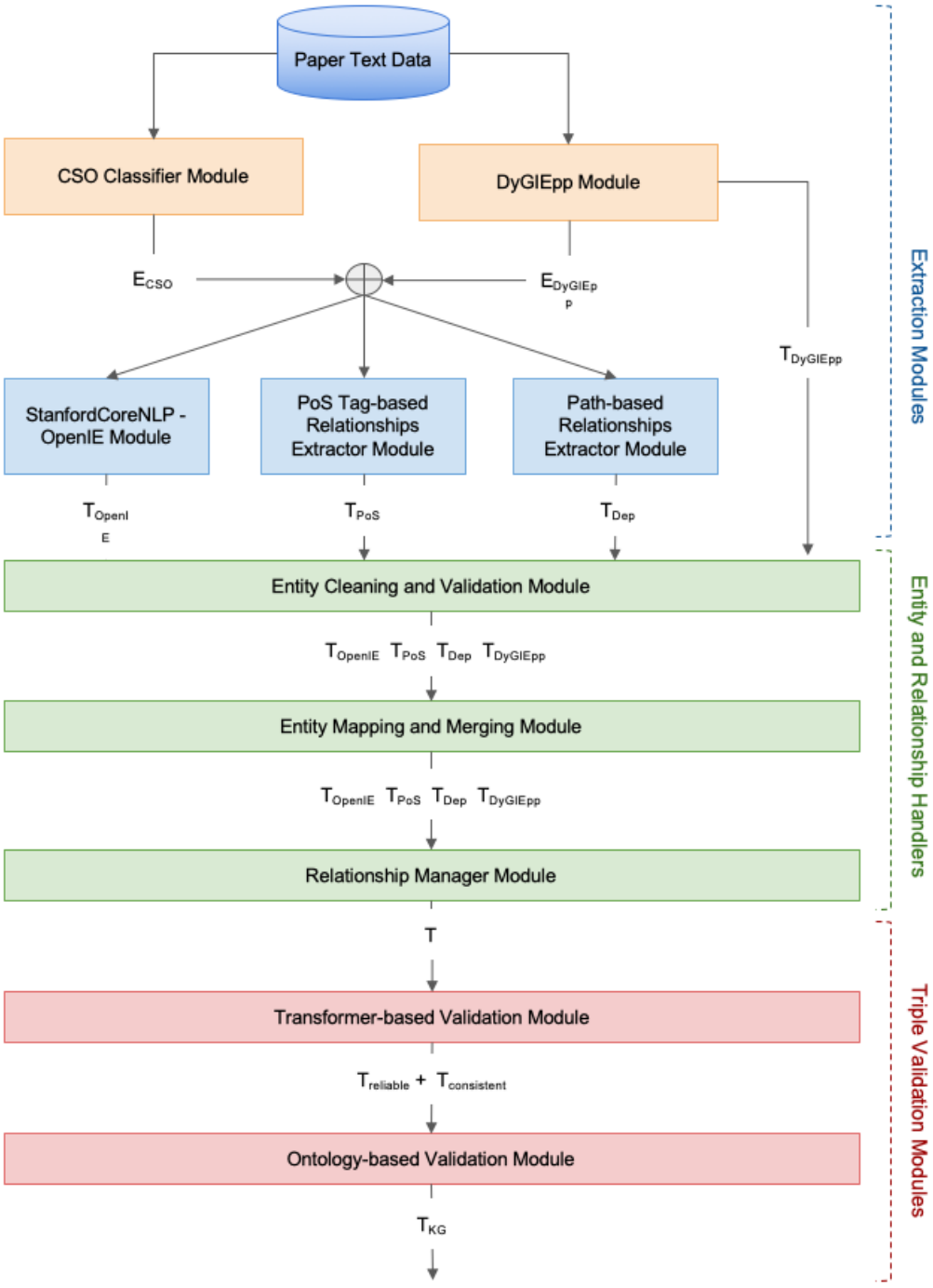


Figure 3.4: "The SCICERO's schema to generate Scientific KGs" [17]

climate change story from multiple perspectives, including scientific, social, political and technological ones. The population of the ontology has a focus on relevant theories, happenings, social and political initiatives.The researchers adopt an ontological approach to construct a knowledge base on climate change-related facts. The resulting ontology is structured as a timeline, which aims to describe the climate change story from multiple perspectives, including scientific, social, political and technological ones. The population of the ontology has a focus on relevant theories, happenings, social and political initiatives. [50] An overview of the ontology is shown in Figure 3.5

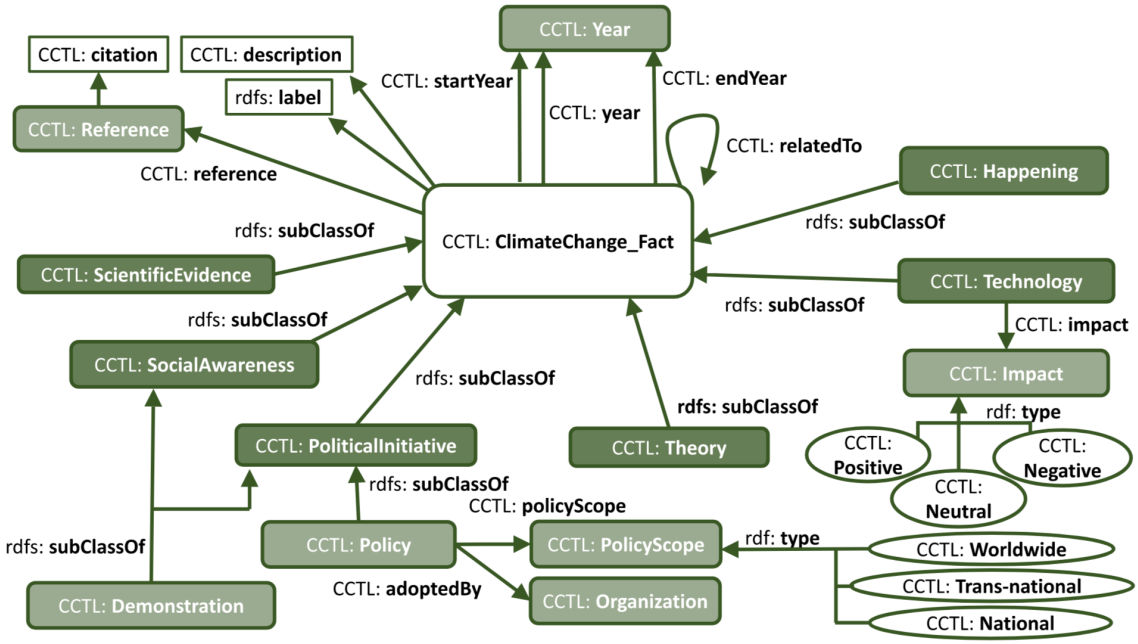


Figure 3.5: CCTL ontology overview [50]

The split into *Happening*, *Technology*, *Theory*, *PoliticalIniative*, *SocialAwareness* and *ScientificEvidence* covers aspects that have an impact on climate change. In order to be implemented in a climate change knowledge graph, the class relations and descriptions of this ontology must be expanded and improved. Furthermore, the instances described in the ontology have statements that could be extracted as triples.

Climate System Ontology (CSO) Davarpanah et al. [16] present an ontology that formally expresses various processes, including non-linear feedbacks and cycles,

which change the compositional, structural, and behavioural characteristics of system components. By reusing top- and mid-level ontologies, they have modelled complex concepts such as the hydrological cycle, forcing, greenhouse effect, feedback, and climate change. [16]

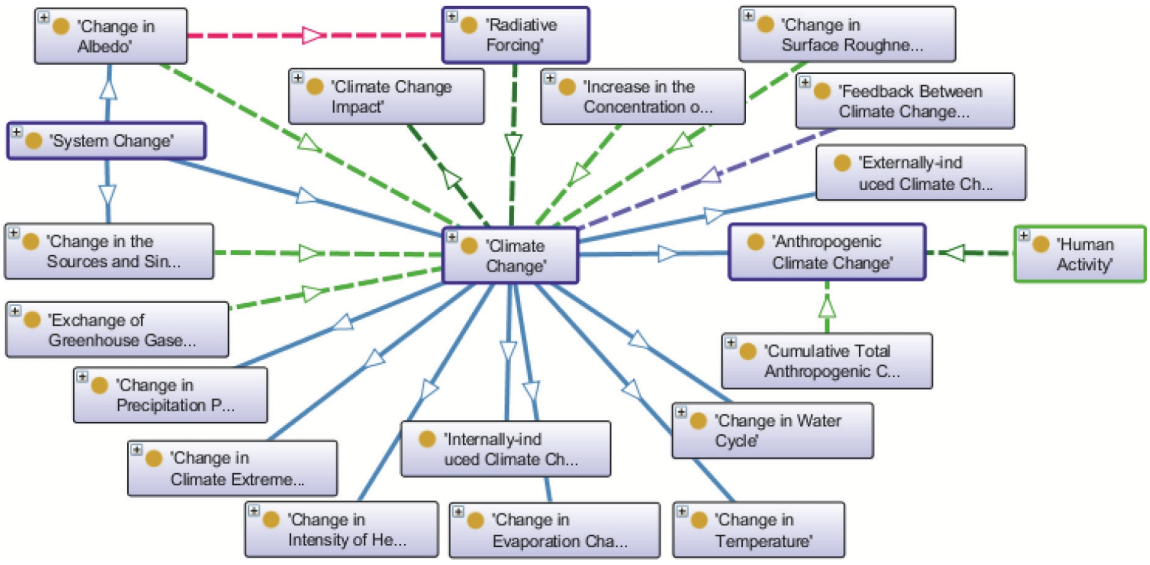


Figure 3.6: CSO model of climate change and relations to other changes [16]

The reuse of top- and mid-level ontologies, such as the widely used Basic Formal Ontology (BFO) and Common Core Ontology (CCO), facilitates collective efforts and prevents the duplication of work [16]. Figure 3.6 illustrates the complex interrelationships between climate change. However, it is evident that the newly added CSO classes lack sufficient detail, necessitating further improvement to ensure their viability.

The Environment Ontology (ENVO). In their study, Buttigieg et al. [8] describe a community-led ontology for specifying a wide range of environments relevant to multiple life science disciplines. The open participation model accommodates all those needing to annotate data using ontology classes. [8] Although this ontology is not primarily focused on climate change, its well-defined structure has the potential to enrich a knowledge graph on climate change.

Chapter 4

Approach

This chapter presents the problem statement and proposed solution. The proposed solution is presented as an overview of the two main processes.

4.1 Problem statement

Public information exchange is not always based on scientific sources. Misinformation can spread rapidly, intentionally or unintentionally. The majority of information lacks clear and transparent indicators of its scientific accuracy, making it difficult to distinguish truth from falsehood. Additionally many journalistic articles suffer from false balance. They present the issue in a balanced way, even though there may be a strong tendency towards one side in the scientific community.

4.2 Proposed solution

A proposed solution for improving the quality of public information exchange based on expert research is to use a scientific accuracy score to evaluate statements in media reports.

This approach involves two types of media with a key characteristic at its core. The initial characteristic is trustworthiness, such as peer-reviewed research. The second aspect to consider is popularity, such as a post that has gone viral and received millions of views. These two types of media undergo two different processes. The first process of building a solid knowledge base begins with the extraction of triples from peer-reviewed scientific literature and constructing a knowledge graph. This process, referred to as 'process trusted media', is illustrated in Figure 4.1. The resulting

knowledge graph consists of dependable information, which can be referred to as ground truth. The second process, shown in Figure 4.6, is called 'process popular media' and takes advantage of the pre-processing and triple extraction methods of the previous process. Each triple is then subjected to a verification check and assigned a score. Finally, a final score is calculated for the entire media based on the scores of all the triples.

4.2.1 Process trusted media

The aim of this process is to preprocess trusted media, extract triples from trusted media, and construct a knowledge graph. An overview of this process is shown in Figure 4.1 and all the steps are explained in more detail in the following sections.

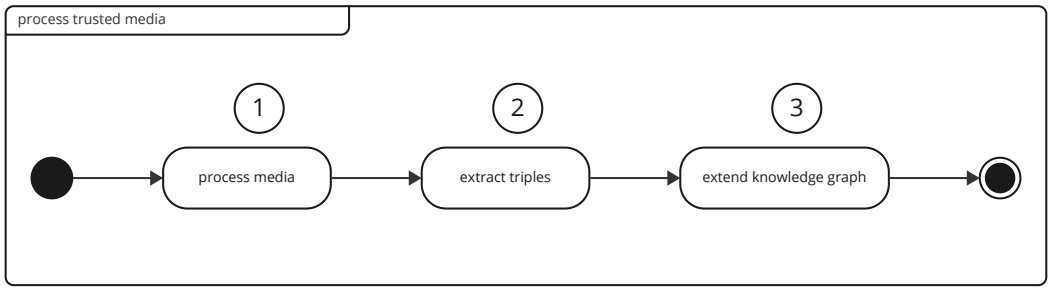


Figure 4.1: Process trusted media overview

Process media

There are often various media types. The purpose of the module is to process the media in a way that allows for independent implementation of further steps, regardless of the media type. Headlines and titles often provide a summary of the information in the medium, making them valuable to save. Additionally, it is beneficial to save an identifier, such as a document ID or URL, for future reference. The next step is to convert the media into text format. In the case of non-textual documents, such as video and audio files, transcription is required. Processing media is shown in Figure 4.2.

As long as the source material is not perfect it will be categorized into tiers. The higher the tier the earlier the source will be used to evaluate the media. Each tier

will build its own knowledge graph which will be able to determine how well the research was done. Tier 1 will be filled with surface level information while the lower tiers will always go into deeper details.

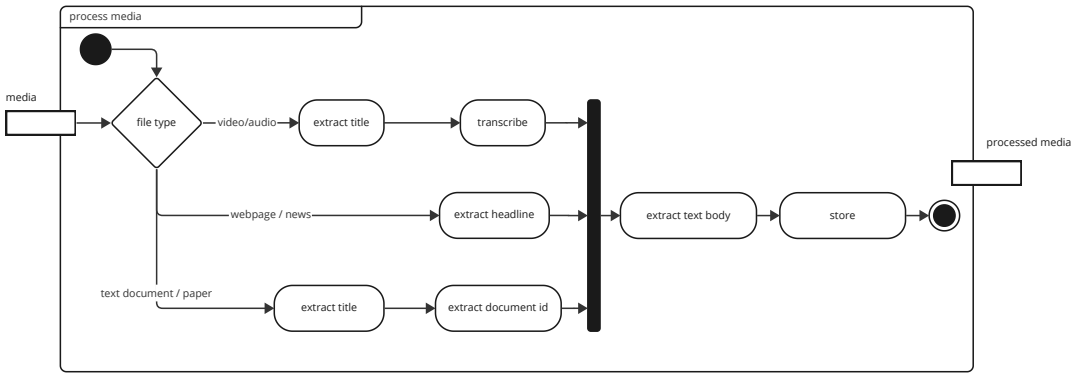


Figure 4.2: Process media pipeline

Extract triples

This pipeline extracts information from plain text. Natural language processing is employed to extract information into a triple structure. Islam et al. [35] proposed a pipeline for "KnowUREnvironment", which provides one possible solution for extracting triples. Following their approach the text is first converted into a spaCy doc, which provides tokenization and prepares it for further language processing. The sentence number is saved from the doc to improve transparency when making evaluations later on. The spaCy doc is then converted into an Abstract Meaning Representation (AMR) graph and triples can be extracted and verified using rule-based methods. If the information is already structured as a triple, the preceding steps can be skipped. All triples must then be aligned and stored. Alignment of the triples is needed to ensure that a different representation of the same statement ends in it being scored worse. The alignment will be done with a straightforward query to an LLM to merge synonymous entities.

During an interim presentation, an expert in information extraction suggested querying an LLM as a substitute for using an AMR graph, as it may yield better results. The revised method is presented in Figure 4.4.

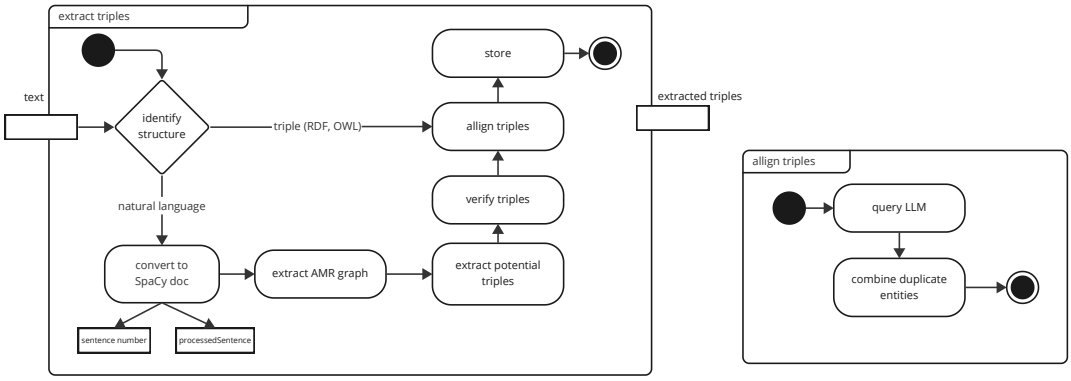


Figure 4.3: Extract triples pipeline

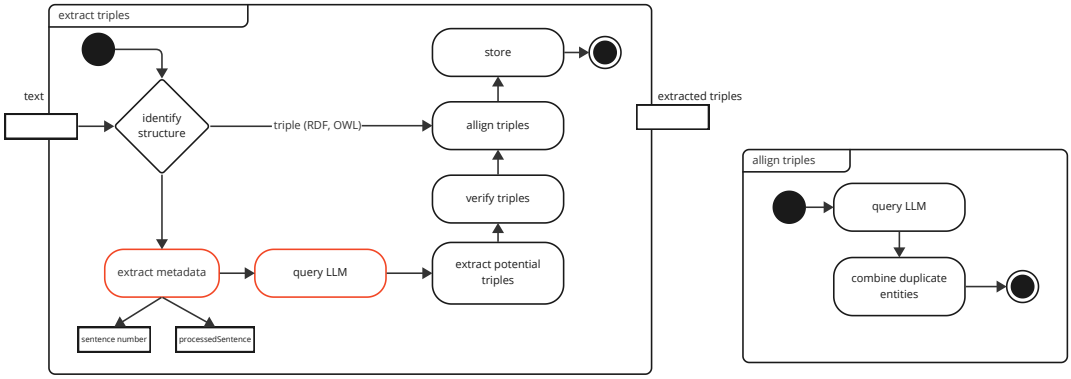


Figure 4.4: Extract triples updated pipeline

Extend knowledge graph

To achieve the ground truth, the final step is to construct and extend the knowledge graph. The triples, along with their metadata, are saved in a triple store such as GraphDB. If metadata is available, such as the publication date or the level of certainty regarding a statement, it is stored using RDF*. The publication date is expected to be available frequently. The IPCC will be the primary source of ground truth in the first version. There the level of confidence is explicitly mentioned. Subsequently, an additional step will be required to assess the confidence in statements

for future versions. The pipeline is shown in Figure 4.5. The available information knowledge graph is referred to as ground truth, which will enable evaluation.

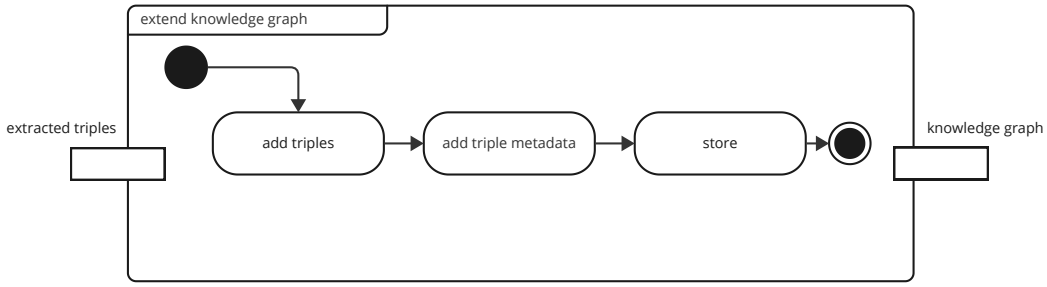


Figure 4.5: Extend knowledge graph pipeline

4.2.2 Process popular media

The main goal of this process is to score the scientific accuracy of different instances of media, using a machine-actionable ground truth. The process includes recurring steps with slight variations in their usage. Figure 4.6 shows the overview of this process. Each step will be explained in detail in the following subsections.

Process media

As stated in Section 4.2.1, the process of acquiring a body of text differs depending on the media type. Videos on social media platforms such as Instagram, TikTok, YouTube, and podcasts are highly popular, and the quality of the transcript plays a significant role in determining the score's quality. When evaluating media, it is important to consider its relevance to the domain and the potential impact of false information.

Quantifying consumption is a straightforward process, but assessing relevance and impact can be subjective and challenging. The number of people consuming the media can be measured through views, listens, subscriptions, or sales. However, obtaining a comprehensive overview of all published media instances and their consumers can be difficult. For instance, in cases where a large group of people may consume the same media, smaller groups may be confined to their own media bubbles and therefore potentially misinformed.

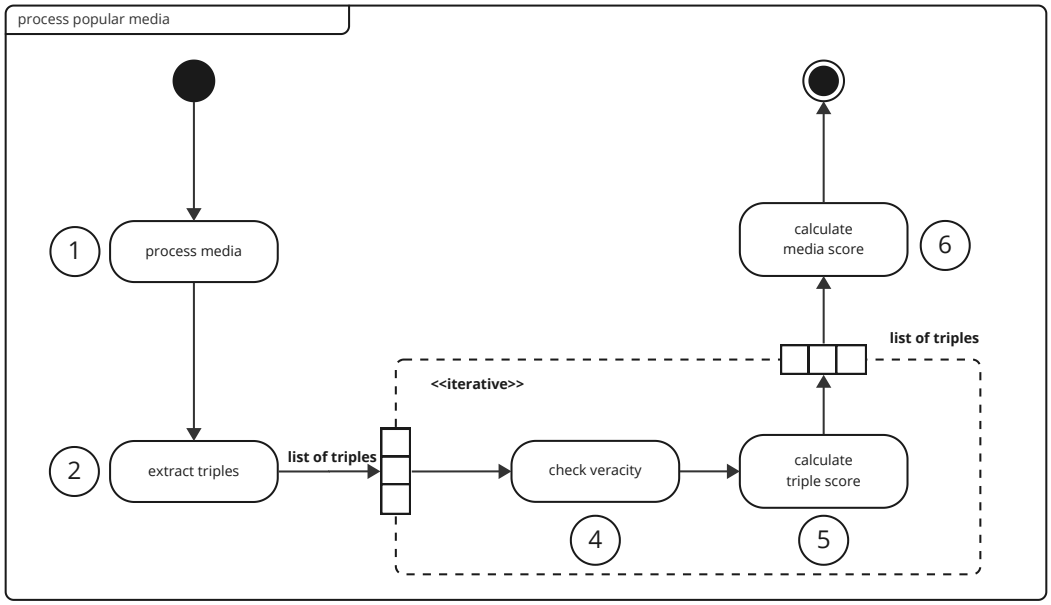


Figure 4.6: Process popular media overview

Extract triples

The process of extracting triples is the same as described in Section 4.2.1 Extract triples. However, the presence of pre-existing structured triples in media articles is exceedingly rare.

Check veracity

This step is performed at the level of individual triples and repeated iteratively for each triple. Future versions should consider checking multiple triples together because the combination of triples may have an impact on the veracity of a statement. The veracity of a statement is the key factor in fact checking. Efficient ways of qualitatively determining the veracity of triples are needed as having a perfect match for each triple would require every entity relationship to be in the ground truth, which is problematic for data volumes of this magnitude. Ciampaglia et al. [11] suggested looking at the shortest path and the density of nodes along that path in "Computational Fact Checking from Knowledge Networks", which provides a possible solution for evaluating the veracity of a triple.

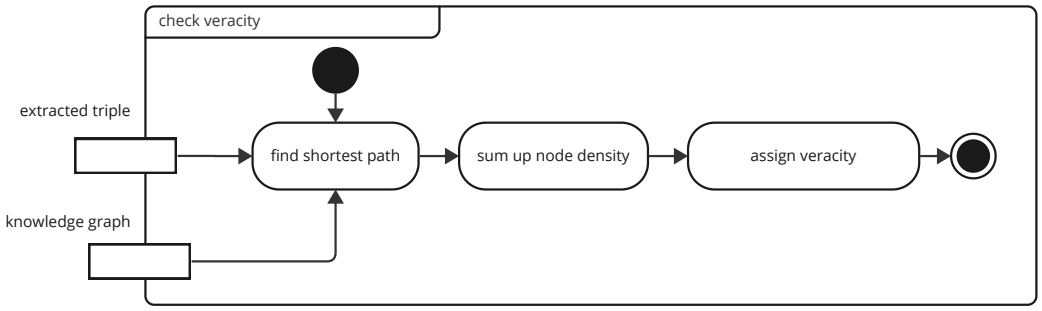


Figure 4.7: Veracity check pipeline

Calculate triple score

In addition to veracity, there are other criteria that are important when evaluating statements for scientific accuracy. Overlooking these factors can have a negative impact on the quality of information. Therefore, they should be reflected in a scientific accuracy score. Below is a list of potential characteristics:

- **veracity:** Determined by the factual correctness of the extracted statements. It is important to only check the veracity of facts, not opinions. The veracity can vary for example:
  - $[climatechange, is, anthropogenic] \Rrightarrow$ highest possible veracity
  - $[species, aren't, impacted] \Rrightarrow$ low veracity

  The **temporal relevance** should also be taken into account, for example:
  - $[2016, is, hottestyear] \Rrightarrow$ low veracity

  A label like "used to be true" might distinguish this example from statements that were never true.

- **confidence:** If a source indicates that a statement requires further investigation, the level of confidence should be reduced. Statements from news articles should be consistent with the level of confidence. Both exaggeration and understatement should result in a lower score.

- **clearness:** Represents the simplicity of a statement. In general, a short sentence will be clearer than one with many subordinate clauses.
- **transparency:** The ease of finding sources that support the statement in question should be considered. Additionally, any conflicts of interest or bias should be disclosed.
- **information depth:** Describes the quality and complexity of the sources employed.
- **objectivity:** Represents whether subjective language is used or not.
- **rationality:** Subjective evaluations that are not clearly marked as such or do not correspond to the facts presented should be penalised.

As most of these criteria are difficult to assess, focus will be put on veracity, clearness and confidence. For confidence, the original sentence is checked for certain vocabulary and for clearness, the sentence length is taken into account. After encoding the triple/statement into vectors, cosine similarity is computed with a comparable vector encoded from ground truth.

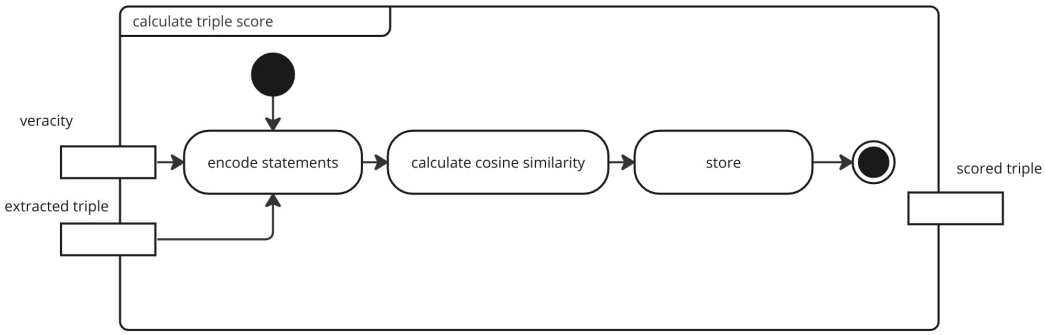


Figure 4.8: Calculate triple score pipeline

Calculate media score

The media score is calculated as the average of all triple scores. Triples without a score due to missing information in the ground truth are excluded from the average. A detailed explanation, including the number of statements found and verified, should be provided.

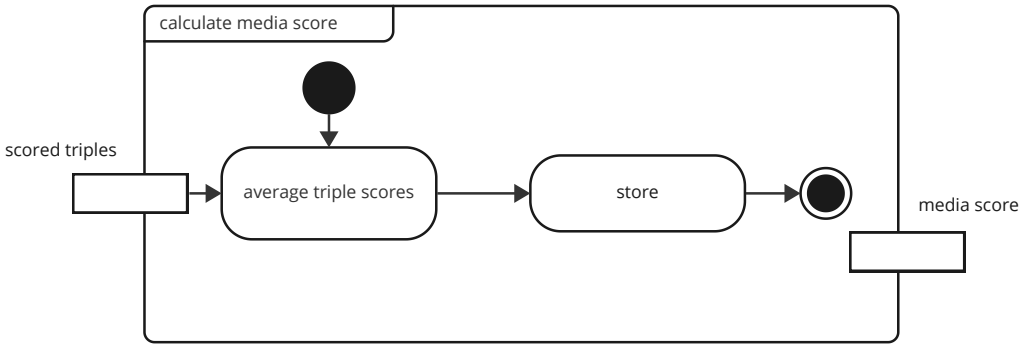


Figure 4.9: Calculate media score pipeline

Transparency regarding the calculation of the score and the sources providing context is crucial for ensuring its reliability.

Chapter 5

Implementation

For this implementation the programming language python in the version 3.10. has been used. This has been implemeneted using anaconda environments and visual studio code as editor. Each step is implemented individually and can be used as such. The python methods can be found on Github[1] where an overview of the whole pipeline is available as well. For more details about the approach read Chapter 4 Approach. The interim results can also be found on Git Hub. This thesis has a great focus on giving an overview and clear interfaces where modules can be changed when someone has more time and expertise to work on them. Every section will show potential modules of which some already have been tested. One will available to have a full pipeline.

5.1 Process media

In the initial iteration of the system, the media is selected manually. However, with further development, it is possible to automate the selection of relevant data by conducting a keyword search. This implementation is compatible with standard HTML websites referenced via URLs, as well as PDF, TXT, MP3, and MP4 files with their file location.

For the extraction of text body and title of HTML documents on websites the python library *beautifulsoup* has been used. The code shown in Listing 5.1 is taken from their extensive documentation[2] where different parser can be selected. The 'html.parser' is included in pythons standard library.

[1]Implementation of this thesis https://github.com/cTremel/Scientific-Knowledge-fit-for-Society/tree/main

[2]Beautifulsoup documentation https://www.crummy.com/software/BeautifulSoup/bs4/doc/

```
import urllib.request

html = urllib.request.urlopen(url).read()
soup = BeautifulSoup(html, "html.parser")
title = soup.title.string
text = soup.get_text()
```

Listing 5.1: Scraping text from html with beautifulsoup

For audio and video files, the transcription tool whisper is employed. This tool is capable of transcribing files in multiple languages. However, when transcribing German audio files, the amount of error was found to be higher. The developers of the tool have stated that the accuracy of the transcription depends on the language being transcribed [55]. As the model and source code are open source and transparently available, whisper does not endanger data protection standards. For the purposes of this thesis, the transcriptions were conducted offline, thus preventing the transmission of data to third parties. The code for the transcription is taken from their documentation[3] and shown in Listing 5.2. The transcription is accessable through *result["text"]* and timestamps can be found in in *result["segments"]*.

```
import whisper

model = whisper.load_model("base")
audio = whisper.load_audio(pathToFile)
result = model.transcribe(audio)
```

Listing 5.2: Transcribing with whisper

The Python library named PDFMiner was employed to extract the text body of PDFs. It is important to note that the process shown in Listing 5.3 only works with PDFs and that only PDF files should be used. The string extracted from the PDF still contains numerous control characters. These can be removed with a simple Python script for subsequent processing. For future versions it is important

[3]whisper documentation https://github.com/openai/whisper

to consider whether the removal of these control characters will result in the loss of context.

```
from pdfminer.high_level import extract_text

text = extract_text("file.pdf")
```

Listing 5.3: Extracting text with pdfminer

The processed media is then saved in a JSON file. This enables an interface between extracting the text from various media sources and subsequently extracting triples from text. Additionally, information regarding the source is also saved within the JSON file for later use. The implementation of saving interim results in JSON files is designed to save time when the process is repeated, particularly in the case of transcription, which can take a considerable amount of time for large files.

This implementation offers a clear interface for integrating modules that automatically collect data sources. Alternative methods for collecting data sources could be implemented through a collaborative effort involving users and experts in selecting sources of interest. It would also be possible to search the web with defined keywords and then prefer sources who match those keywords most often. A combination of both approaches could yield the most optimal results, but is also the most effortful to implement in practice.

5.1.1 Trusted

A variety of reliable sources on climate change are available. A significant proportion of these sources are secondary literature, which already synthesises research. In this way, already combined research is used as a starting point, as it allows for the description of a wider range of topics with limited sources. The IPCC AR6 has been selected for this thesis because it covers past changes, current trends and projections as well as advised mitigation and adaptation options. There is different versions of the AR6 and in this implementation only the the Headline Statements are used as basis for a ground truth. This is a compromise of a manageable amount of information and still cover multiple aspects of topics related to climate. In subsequent additions to this implementation, it is recommended that primary literature be employed, which could include sources from the AR6. A list of other potential sources is provided in Section 7.4 Data sets.

IPCC AR6. At first the full report was supposed to be the source for the ground truth. Since the triple extraction did not perform as well as hoped the process was done automatically for the first part and manually for quality checks and verification. The manual labour limited the input material that can be used for the triple extraction. Therefore instead of the full report the headline statements for policymakers where chosen as input for the ground truth.

5.1.2 Popular

The search for media articles was conducted manually. Identifying articles that include statements that can be evaluated was challenging due to the distinct writing style employed by journalists, which differs from that used in scientific literature. While a writing style tailored to the audience can enhance the accessibility of information for non-experts, it is not necessarily a disadvantage. However, as long as the implementation is not effective at matching statements from different writing styles, not every article can be scored. The selected article is also limited because the actual information within the ground truth is incomplete. When the ground truth includes more information, the scoring of additional articles will be enabled. An article written in a compatible style and with a topic matching information from the ground truth is available on the official NASA website. This article from NASA will be used as an example[4].

To evaluate the quality of the score articles which have been evaluated by someone else can deliver a helpful evaluation. For example the news articles which have been fact checked by organisations described in Section 3.2 Fact checking. Another potential source for news agencies writing articles can be found on wikipedia[5]. The wikipedia community evaluated sources which should be visible in scores.

5.2 Extract triples

The triple extraction process has changed the most during this work. At first the search for existing triples was prioritised and authors were asked whether code can be made available. Avoiding duplication of work for such a difficult task would make a big difference as this approach needs to implement other steps as well. No adequate triples have been found. A pipeline, called the plumber[6], that combines

[4]NASA example articlehttps://science.nasa.gov/climate-change/effects/

[5]Reliability of sources evaluated by wikipedia community https://en.m.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources

[6]The plumber https://github.com/YaserJaradeh/ThePlumber

community created components for information extraction has been build and expanded by Jaradeh et al. [36, 37]. This can be used to extract triples and try out different components for specific subtasks. After asking the author he said that for the domain specific case a new extraction component should be explored and that LLMs have the potential to outperform previous methods even though they come with a lot of challenges as well.

An overview of different methods is shown in Figure 5.1. There are three main approaches which vary in their state of completion. At first Semantic role labeling (SRL) was used to create Abstract Meaning Representation (AMR) graphs. AMR is a framework for semantic dependencies often representing sentences in graph form based on the semantic meaning of its words [5, 28]. Semantic role labels indicate the basic event properties and relations among relevant entities in a sentence to determine essentially "who did what to whom", "when", and "where." [30, 64]. These representations have played a role in optimizing task like translation and machine reading [63].

The second method to extract triples which has been explored is using a domain specific Named Entity Recognition (NER) model. Because of the limited time and scope of this thesis a model to recognize domain specific entities has only been touched theoretically.

The last module that has been considered is based on LLMs. Different models from different providers have been briefly tested while also working on improving the prompt.

5.2.1 AMR

The first method to extract triples was inspired by [35] and uses spaCy and AMR graphs to extract triples. The text is given into a spaCy doc and then transformed into an AMR graph. AMR uses the PENMAN notation, which was originally called Sentence Plan Notation [49] and is based on annotations from large corpora for example "Proposition Bank" [48]. The Proposition Bank takes annotations of the Penn Treebank and adds semantic role labels [48]. The approximately 7 million words of part-of-speech tagged text in Penn Treebank includes among others IBM computer manuals, Wall Street Journal articles and transcribed telephone conversations [65].Figure 5.2 shows an AMR graph representation of the first sentence of the IPCC AR6 Headline Statement A1. The idea to abstract triples from such a graph is to use the labels of the PENMAN notation where generally *arg0* describes entities and *arg1* relationships. AMR also offers additional labels that identify quantities, dates, scales and polarity. Listing 5.4 shows how to create AMR graphs in python

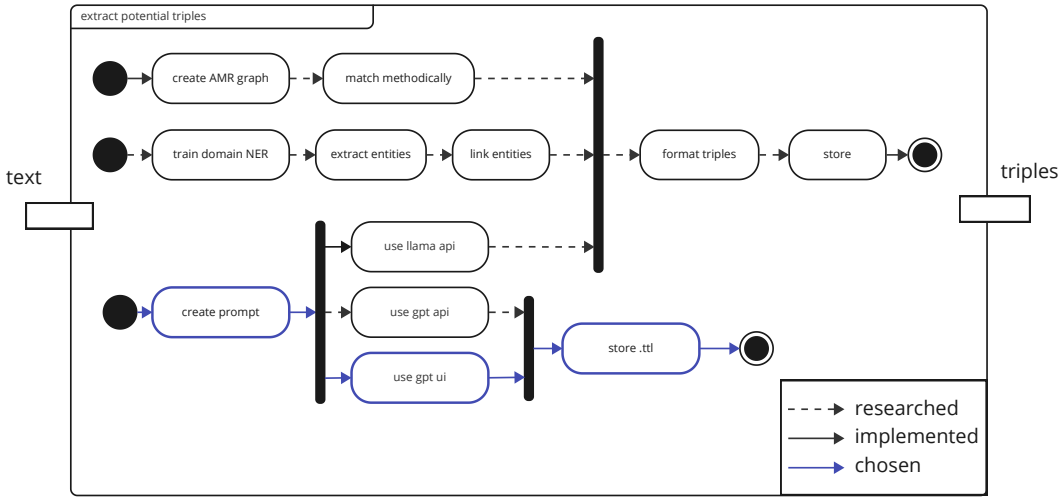


Figure 5.1: Overview of possible modules

and further documentation[7] is available.

```
amrlib.setup_spacy_extension()
nlp = spacy.load('en_core_web_sm')
text = 'Severe weather damage will also increase and intensify.'

doc = nlp(text)
print(len(doc.sents))
graphs = doc._.to_amr()
```

Listing 5.4: Python code to create AMR graphs

[7]Penman graph library https://github.com/goodmami/penman

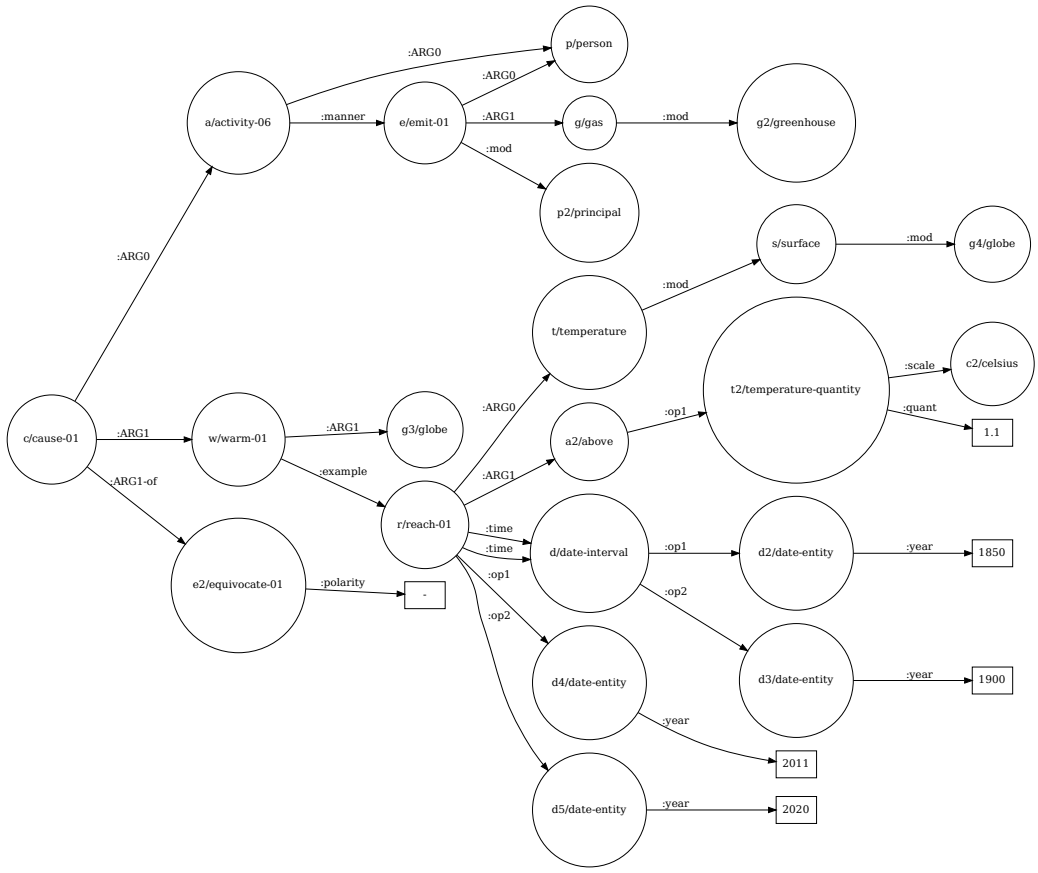


Figure 5.2: AMR Graph of IPCC AR6 headline statement A1

5.2.2 Domain specific NER

During the implementation training a Domain specific NER model was considered. Through the first method spaCy was already present and flair was also considered. The documentation of how to train a model with flair can be found on their website[8]. Flair was convincing because of the easy to follow explanations on how to train a domain specific model with their framework and their state-of-the-art perfomance [2].

[8]flair intro https://flairnlp.github.io/docs/intro

System	# Human Languages	Programming Language	Raw Text Processing	Fully Neural	Pretrained Models	State-of-the-art Performance
CoreNLP	6	Java	✓		✓	
FLAIR	12	Python		✓	✓	✓
spaCy	10	Python	✓		✓	
UDPipe	61	C++	✓		✓	✓
Stanza	66	Python	✓	✓	✓	✓

Table 5.1: "Feature comparison of Stanza against other popular natural language processing toolkits" reprinted from [54]

Much later the findings of Qi et al. [54] and the corresponding overview in Table 5.1 showed that Stanza[9] outperforms other NLP toolkits [54].

Training a domain specific model in combination with the use of an ontology could possibly provide better results, but is cost-intensive and time-consuming. Because training a model and creating an ontology are both tasks that are difficult to do well. Learning ontologys automatically from text is an active research field. Creating an ontology is to much for this thesis. But when working on this task it is recommended to split into sub tasks like Cimiano [12] describes it as an "Ontology learning layer cake" [12]. Another resource in that domain is the review published by Asim et al. [3].

5.2.3 LLM

The third method of extracting triples in this implementation involved the use of LLMs. LLama is freely available for use with the API. However, its results did not as well align with those of GPT3.5, as tested. The UI of GPT3.5, "ChatGPT," was the most effective at fulfilling the steps of extracting and formatting triples for this approach. There is also an API interface, but it comes at a cost and therefore was not used. For this reason, this part of the process was not automated. Automating this process would be technically straightforward by using one of the other models or by paying to use this one. The code needed to query a LLM is shown examplary in Listing 5.5. Given that the pipeline is semi-automated and that time was limited, the decision was made to use this model.

To improve the quality of the output of a LLama prompt, engineering is advised. When continuing to work with LLama, this must be researched further. In this implementation, it was beneficial to create prompts that clearly stated the task in

[9]Stanza https://github.com/stanfordnlp/stanza/

question (information extraction) and the format of the output. Additionally, it was advantageous to label the given input with clear "START" and "END" strings.

The LLM was employed to facilitate the completion of three distinct tasks. The general extraction process involved transforming verbs into their base form and identifying synonyms.

The extracted trusted triples are exported into .ttl format, which GPT3.5 performs when instructed to do so, although the process is not entirely error-free. Such results may be utilised as an interim outcome. If the process of extracting triples is optimised, it is possible to use the more accurate triples for the remainder of the process. For triples that require scoring, the desired output is a list of triples.

```
model = "meta-llama/Llama-2-7b-chat-hf"

t = AutoTokenizer.from_pretrained(model, use_auth_token=True)

llama_pipeline = pipeline(
    "text-generation",  # LLM task
    model=model,
    torch_dtype=torch.float32,
    device_map="auto",
)

def get_llama_response(prompt: str) -> str:
```

Listing 5.5: API to get LLama response

5.3 Extend knowledge graph

5.3.1 Creation

Considering limited time and to avoid duplication of work an already implemented graph store was chosen based on the needed functionality for this pipeline. Not every triple store is able to utilise RDF* and SPARQL*. GraphDB also offers an UI with the possibility of easily uploading files containing triples in various formats. Querying and visualizing the constructed knowledge graph is also included. For

future implementations GraphDB also offers an Application Programming Interface (API) that enable automation of steps which are done via the UI for now.

The *.ttl* and *.ttls* files created in the previous step of the pipeline were uploaded to GraphDB. Statements about statements enable to link sources, add confidence an objectivety scores for statements and also give context of temporal validity. It is also possible to upload other file types like *.owl* to integrate ontologies. In this implementation only small amounts of triples were used thats why it is unclear whether or not GraphDB scales well with larger amount of triples.

5.3.2 Ontologies

Protégé is a tool that helps with creating and investigating ontologies that has been used. Meaningful labels and a structural framework can enrich the information of the ground truth. Also having a clear format and a vocabulary helps in the evaluation process by improving the possibility to compare different statements. Some ontologies include entities which are relevant for this approach but an integration of an ontology in the triple extraction process has not been done.

5.4 Check veracity

To check the veracity queries can be run via a SPARQL API. In this version a check for exact matches of proposed triples against the ground truth is used. But it is also possible to get paths between nodes and have criteria to score it. The properties along the path that are readily available are for example the path length and the degree of nodes on the way. Also the predicates along the path are available. Listing 5.6 shows python code that is able to query the GraphDB knowledge graph. Listing 5.7 and Listing 5.8 show the queries to get a narrow match and the shortest path between to nodes respectively.

5.5 Calculate triple score

The score consists of scores for veracity, clearness, transparency and rationality. In this implementation only veracity is checked by querying the knowledge graph to check for an exact match.

An calculation of a cosine similarity is implemented that can be used for future versions where vectors of different score criteria can be checked against the ground truth.

```python
sparql = SPARQLWrapper(urlToKG)
sparql.setQuery(SPARQLQUERY)
sparql.setReturnFormat(JSON)
results = sparql.query().convert()

for result in results["results"]["bindings"]:
    print(result["label"]["value"])
```

Listing 5.6: SPARQL Wrapper API python

```
PREFIX ex: <http://example.org/>

SELECT ?hasNarrowMatch
WHERE {
OPTIONAL {ex:subject ex:predicate ex:object}
BIND (exists{ex:subject ex:predicate ex:object} AS ?y)
BIND (IF(?y, "true", "false")      AS ?hasNarrowMatch)
}
```

Listing 5.7: SPARQL query: narrow match

5.6 Calculate media score

For this version the media score just sums up the amount of statements which have been found in the knowledge graph. There also is a count of how many statements have been checked. Later the media score will be calculated by averaging all triples scores. In a later versions with multiple scoring criteria those need to be weight differently. The veracity should be the most important quality. More on how to further improve the score is described in Section 7.4 Future work.

5.7 Use Case - "The Effects of Climate Change"

To demonstrate the tool's current state, a sample article from NASA has been scored for scientific accuracy. Out of the 9 extracted statements one could be confirmed. No statement was proven false and the rest could not be evaluated.

```
PREFIX path: <http://www.ontotext.com/path#>
PREFIX ex: <http://example.org/>

SELECT ?pathIndex ?edgeIndex ?edge
WHERE {
    VALUES (?src ?dst) {(ex:subject ex:object)}
    SERVICE path:search { []
        path:findPath
        path:shortestPath;
        path:sourceNode ?src;
        path:destinationNode ?dst;
        path:pathIndex ?pathIndex;
        path:resultBindingIndex ?edgeIndex;
        path:resultBinding ?edge; .
    }
}
```

Listing 5.8: SPARQL query: shortest path

processMedia. The article "The Effects of Climate Change"[10] was selected by hand to show a proof of concept which verifies at least one statement. The article is written rather scientifically which helps mitigate the challenge of different writing styles in the scientific and journalistic community. The text was extracted manually but a python package named beautiful soup[11] can be used to scrape html documents for their text body. Afterwards the different interim results of the triple extraction can be seen in Figure 5.4. The data is shown in grey and goes from text to triple and is refined further.

extractTriples. The steps which are done with help of a LLM are depicted by the used prompts in blue. For one sub step specific domain terms are needed which have been manually collected from the IPCC and are displayed in yellow. The LLM used for this example is GPT3.5.

checkVeracity. Those triples were verified and then checked against the ground truth of the knowledge graph. This knowledge graph consists of *.ttl* and *.ttls* files which have been extracted from the IPCC AR6 as shown in Figure 5.3. The extracted triples are humanly evaluated and corrected if necessary. The veracity check was done by querying the knowledge graph for an exact match of each triple from the article. The triple [**'ex:humanActivity', 'ex:cause', 'ex:climateChange'**] also shown in Figure 5.4 could be verified.

calculateTripleScore. Only the veracity in this case an exact match inside the query of the knowledge graph was calculated. A graphical interface to display the score is not currently available, but in the future it could resemble Figure 5.5.

calculateMediaScore. The media score normally calculates an average over different triple scores. Since only one triple was scored the triple score is also the media score. A hint that this is not reliable is necessary and will be displayed if not the majority of found statements has been evaluated.

[10] NASA Article https://science.nasa.gov/climate-change/effects/
[11] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

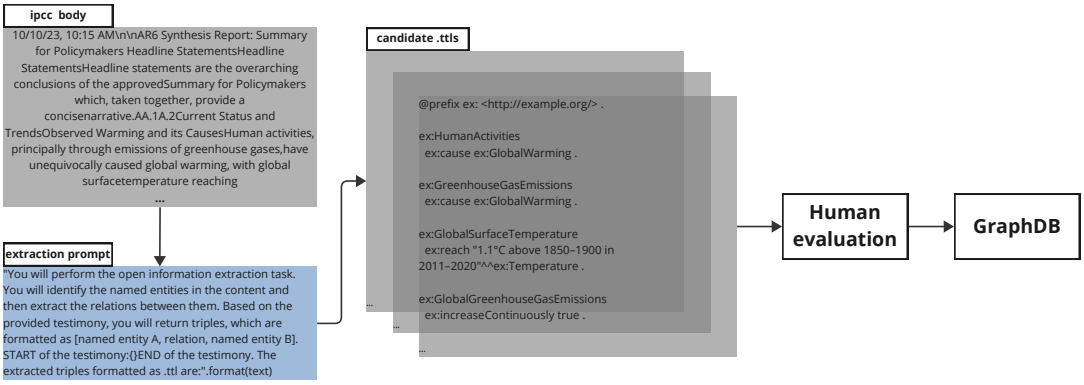


Figure 5.3: Process trusted media example

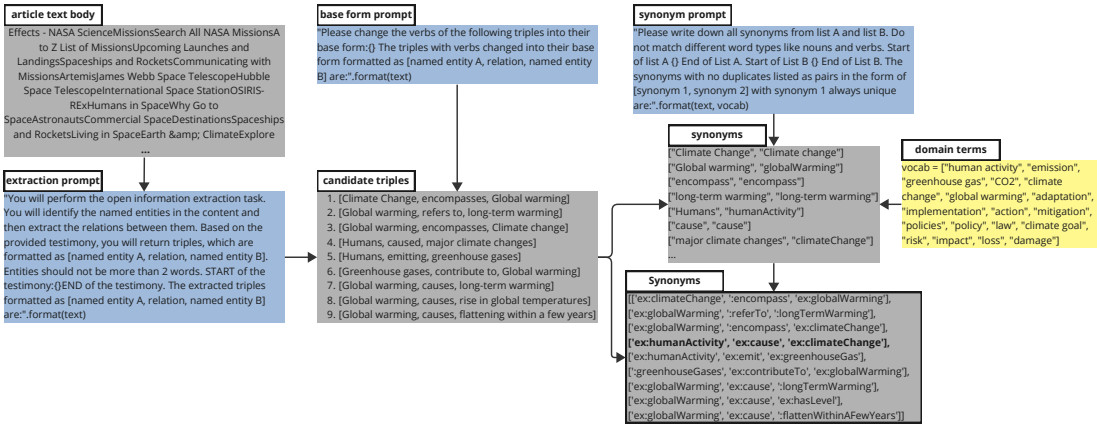


Figure 5.4: Process popular media example

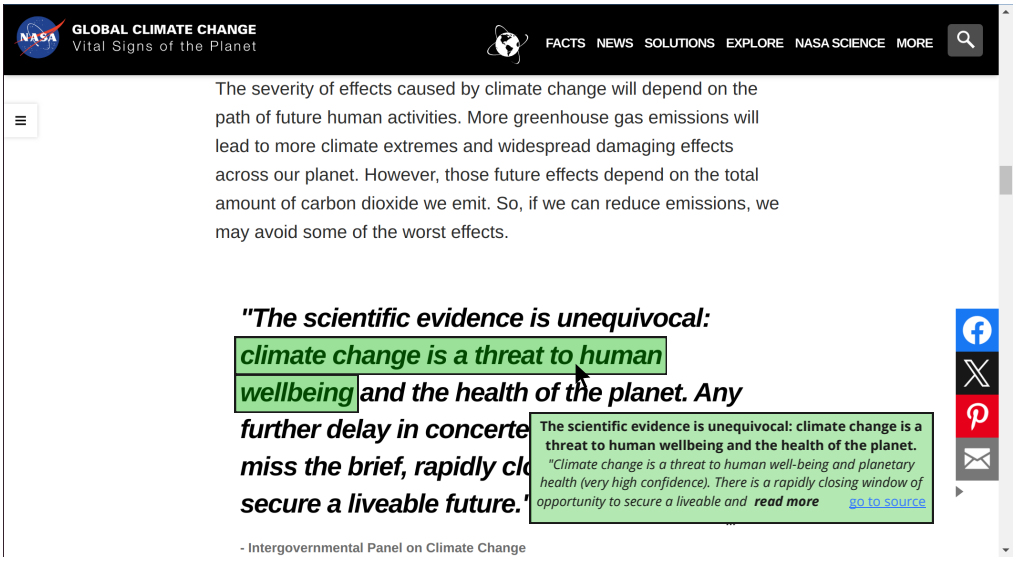


Figure 5.5: Possible appearance of the score

Chapter 6

Evaluation

This thesis is concerned with the use of natural language processing and knowledge graphs to identify scientifically accurate media on climate change, as stated in the Research Questions. Firstly, the design of the evaluation will be described in Section 6.1. Subsequently, the execution of the evaluation from experts in Section 6.2 and users in Section 6.3 will be depicted.

6.1 Design

In order to facilitate the integration of scientific knowledge into public discourse, two primary groups are under consideration. In the field of information extraction and its representation in a machine-actionable format, the researchers from the ORKG possess expertise. When considering public discourse, it is possible to include everyone in order to evaluate the need for a scientific accuracy score.

Experts. In order to ensure that the methods and tools being used are up to date and properly contextualised, experts from the ORKG were invited to participate in separate settings. The ORKG team was selected due to their expertise and the connection was facilitated through a supervisor of this thesis. The expert evaluation is based on a combination of presentations and interviews. In order to gain a general overview, a brief seven-minute online presentation is planned for inclusion within the biweekly meetings of the ORKG team. The presentation will set out the approach, problem statement and pipeline for the solution. The use of LLMs to extract triples, create a knowledge graph and query that knowledge graph will be discussed. Following this, eight minutes will be allocated to answer general and specific questions using a polling tool provided by the videoconferencing tool. Participants will be

invited to respond with either “Yes”, “No”, “Other”. The objective of the polls is to be simple, thereby encouraging a high level of participation. Given the limited time available, a discussion would be impractical. Polls enable the gathering of a significant amount of information from participants in a relatively short period. Once a general trend has been established, there will be an opportunity to delve deeper into specific areas. The majority of questions are designed to evaluate assumptions. Should the responses prove accurate, the approach will be validated. Otherwise, it can be refined through the input of experts. Following the presentation, the most active and topic-relevant experts will be invited for brief interviews. These will be based on their responses during the presentation, and they will be invited to participate in interviews of a duration of about 15 minutes. Following the assessment of the interviews, the findings will be presented once more in the form of a brief presentation in the ORKG team meeting. This will be followed by a poll to legitimate the findings.

Users. The need for a scientific accuracy score will be assessed through a survey. The survey will ascertain whether the problem is perceived by the public and whether a scientific accuracy is considered a viable solution. Furthermore, the current state of the tool will be presented using an example. Finally, the survey will investigate the circumstances and types of media in which the public would utilise such a tool. In order to oversee the representativeness of the survey sample, the age, expertise in suitable fields, and highest degree of respondents will be recorded. The survey will be conducted in a public manner and initially shared through personal contacts.

6.2 Experts

6.2.1 Presentations

28 members of the ORKG team were present for the meeting and two persons had to leave before the presentation which was given at the end of the meeting. The team consists of researchers, curators, developers and PhD Students. Then there is also the lead and co-lead of the team which are the evaluators of this thesis. During the polls not all participants consistently took part. The polls started with general questions and afterwards, the experts were presented with two sets of questions that covered specific aspects of the approach. The questions were framed as statements, asking for agreement or disagreement. The first set focused on triple extraction, while the second set focused on score calculation, specifically the characteristics of entity relations in knowledge graphs.

The four statements referred to for the results presented in Figure 6.1 are:

1. The best way to extract triples is achieved through NER with a domain specific model.
2. Using an LLM to extract triples is a quick, easy and quite good solution.
3. Given a few paths connect two entities A and B inside a trusted knowledge graph:

   The path length is impactful to evaluate a claimed relation between A and B.
4. Given a few paths connect two entities A and B inside a trusted knowledge graph:

   The degree of the nodes on these paths is impactful to evaluate a claimed relation between A and B.

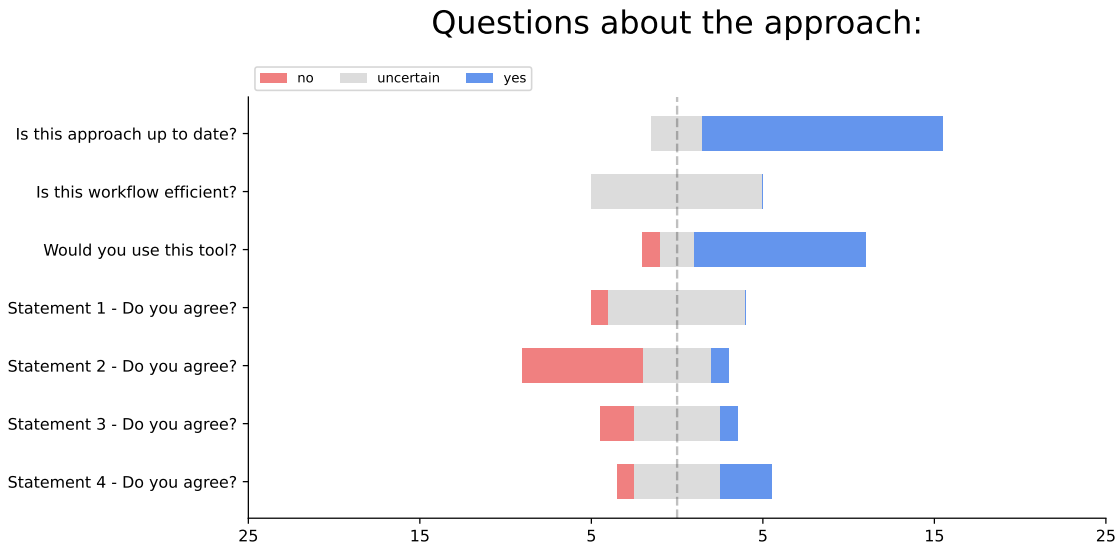


Figure 6.1: Poll results from the first presentation

For the second question **Is this workflow efficient?** one researcher explained that the approach was not described in enough detail to answer such a significant question within such a short presentation. This is also why he selected "Other" in the first question, and many people agreed with this sentiment. Upon further inquiry to question regarding statement one, the researcher who responded negatively explained

that although using NER with a domain-specific model is a good practice for certain aspects of triple extraction, it does not address the complete task. Responses to the second statement about the quality of LLMs were mixed, the reason for this variation was that the term 'quite good' was not well-defined. LLMs have potential, but there are some issues, including reliability and hallucination. Additionally, the compliance of the triples with FAIR principles has been questioned in the usage of LLMs. The importance to consider the context when extracting triples was also mentioned. For the third statement the argument that the path length in a knowledge graph depends heavily on its context was mentioned. For instance, a specific and dense knowledge graph has significantly different path lengths between two closely related entities than a surface-level knowledge graph for the same entities. The problem of inversion was also discussed, which would only increase the path length by one and completely reverse the sentiment.

Additional thoughts. At the end of the presentation, the audience was asked if they had any additional questions, comments, or ideas. It was mentioned that a more challenging issue than climate change, where there is no scientific consensus, could be considered. It was suggested that LLMs are potential climate killers and that this should be mentioned in the thesis. Additionally, adding a time component was suggested. It was also suggested to test the system against domain experts and to try to categorise news articles (e.g. NY Times vs. local articles).

6.2.2 Interviews

Following the presentation, experts were invited to engage in a discussion regarding the questions presented. Ten experts took the time to provide detailed responses to the approach, with the interviews collectively spanning a duration of 4 hours and 42 minutes. The shortest interview was 13 minutes and 39 seconds, while the longest was 43 minutes and 55 seconds. The mean interview length was 28 minutes and 12 seconds, with a standard deviation of approximately 10 minutes and 40 seconds.

During the interviews, the questions were specified and personalised based on the feedback received during the presentation. If multiple people answered the poll in the same way, a similar follow-up question was asked. That is why, following the initial question, there will be various follow-up questions. The responses provided by multiple individuals will be summarised.

General.

Q1	**Is this approach up to date?**
if "yes"	*What makes you think this approach is up to date?*
Expl.	The use of LLMs and assembling information in knowledge graphs are often considered up-to-date approaches.
if "other"	*You answered "Other" what do you mean?*
Expl.	Fact-checking and constructing knowledge graphs are not new practices. Google has manually built knowledge graphs for a long time, and many people are invested in fact-checking. However, when specifically considering the automation of this task, it is considered up-to-date. Another reason to choose "Other" was that some members of the audience were not computer scientists and didn't have enough information to decide whether this approach used the latest models or not.
Q2	**Is this workflow efficient?**
if "other"	*Where do you see potential bottlenecks for this approach?*
Expl.	The triple extraction phase is the core concept here. Therefore, the bottleneck lies in the triple extraction phase. If the entire processing operation is based on triples, then the more robust the triple extraction, the more robust the approach will be. This will result in the extraction of more correct and relevant triples, and possibly the augmentation of those triples. Semantic parsing presents a challenge due to the syntax, particularly when using RDF. The syntax requires a subject, which must be a resource. The challenge in this process is that the syntax of RDF requires a subject that is a resource, while sentences often have subject phrases that contain more than just a noun. Furthermore, sentences may have multiple objects. One issue with RDF is its high expressivity in representing information. This poses a challenge for reproducibility when using the Large Language Model. RDF's expressivity allows for the same statement to be conveyed in an indefinite number of ways in the same sentence. If the information needs to be machine-readable and evaluated by machines, using trusted triples for reasoning, there may be limitations with OWL. While OWL is a powerful tool, it cannot model everything. It is important to be aware of these limitations when using OWL.

Expl.	OWL is based on description logic, which restricts the types of statements that can be made. It is possible to make statements about specific individuals, but statements such as "headaches can be a symptom of COVID-19" cannot be expressed in OWL. To express this relationship, a first-order logic or a similar approach is required. In empirical research, such as climate change, it is common to find evidence that, when taken together, supports a certain cause for a phenomenon. However, it is often not possible to conclude the cause with certainty based on a single piece of data.
Q3	**Would you use this tool?**
if "yes"	*What is a use case for that tool that would interest you most?*
Expl.	The potential effectiveness of a plug-in or browser integration is promising. Providing an indication alone would be helpful. A possible solution is to implement a traffic light system and provide users with the option to access additional information by clicking on it. The system can always identify potential issues and present their pros and cons for consideration. To narrow down the search scope, a semantic search can be used based on the article or term of interest. This can be achieved by combining a keyword search with a semantic search. By using a set of trusted sources, it is possible to determine their relation. This is useful for creating recommendation tools, clustering data sets, and answering questions. Using the tool as a personal assistant can be helpful. It can communicate with a browser, PDF reader, and other sources of information through a trusted graph. With the ability to run in the background and evaluate new information, personal evaluation becomes easier. The potential for using fake news or misinformation in education to raise awareness of the issue is significant. It can be used as a tool to teach students how to critically approach texts. This experience will aid in reading texts with a more critical eye. Utilising specific elements of the workflow within a larger ecosystem could be an intriguing application of this approach.
if "other"	*You answered "Other" what do you mean?*
Expl.	Whether or not someone will want to use this tool depends heavily on its actual performance.
if "no"	*Why would you not want to use it?*

Expl.	It is only through a tool's functionality that it can be considered useful. Consequently, until this aspect has been clarified and demonstrated, the tool cannot be considered useful.

Table 6.1: General questions about the approach

Triple extraction.

Q4	**The best way to extract triples is achieved through NER with a domain specific model. Do you agree?**
if "other" or "no"	*You answered "Other" what do you mean?* *Why don't you agree?*
Expl.	NER doesn't represent the full task of triple extraction. Triple extraction usually consists of some kind of entity and relation recognition. There is also entity disambiguation and entity linking. Even in the domain of climate change, entities are not named entities. These are terms reserved for people, places, etc. Therefore, NER cannot fulfil the whole task of triple extraction, but it can be a good use for entity and relation recognition.
Q5	**Using an LLM to extract triples is a quick, easy and quite good solution. Do you agree?**
if "yes"	*Why do you agree?*
Expl.	There is a consensus that LLMs can be useful in specific domains. However, their effectiveness is debatable and depends on the definition of 'quite good'. Despite their potential, LLMs also present key challenges, such as hallucinations. To mitigate these challenges, it may be helpful to keep LLMs in check with pre-existing methods.
if "other" or "no"	*You answered "Other" what do you mean?* *Why don't you agree?*
Expl.	If the model is not properly guided, it may extract information that is difficult to explain and trace back to its source. It is difficult to trace the origin of this, and it may even alter the entire concept. When searching for specific items or particular types of triples, using the out-of-the-box method may not be the most suitable approach. Perhaps it is advisable to use a few-shot approach instead of a zero-shot approach.

	It is not a straightforward or effortless solution. While it may be an easy starting point, improving the results requires iteration and refinement of the prompt. In some cases, it may even be necessary to engage in a chain of thought reasoning, allowing the model to critique itself and address any issues with the existing output.
Extra	**Do you have additional thoughts about extracting triples?** The best solution to extract triples is uncertain. It is necessary to perform entity recognition and disambiguation, as well as relation extraction and disambiguation. The relation may be explicit or implicit, and it may not be present in the text at all. Class disambiguation and extraction are also likely to be necessary. A combination of NER and LLMs has potential. A mediator is necessary to ensure consistency in moderating statements or types of statements, particularly in cases involving trusted sources and popular media that require verification. The mediator must operate within specific criteria and constraints. When using LLMs, lexical constraints can be applied, such as entity length, number of words in an entity, relation type, and inclusion of classes. Triple extraction requires a query in GPT, as well as a set of patterns and a vocabulary. Currently, there is a significant focus on transformable models and large language models in research, with the aim of exploring ways to use them in creating triples. For an end-to-end system, it is highly likely that a large language model will be used, as this is currently the most common approach. Although it is possible to extract triples from a text using a large language model out of the box, this is not recommended as it often results in poor triple quality and is considered bad practice.

Table 6.2: Questions about triple extraction

Score calculation. Given a few paths connect two entities A and B inside a trusted knowledge graph:

Q6	**The path length is impactful to evaluate a claimed relation between A and B. Do you agree?**
if "yes"	*Why do you agree?*
Expl.	The path is a good starting point. Therefore, path lengths could be one aspect of the scoring mechanism, but should not be the sole factor. Additionally, the meaning and semantics of the relations between nodes should also be considered.
if "other" or "no"	*You answered "Other" what do you mean?* *Why don't you agree?*
Expl.	Losing predicates results in a significant loss of information. Predicates are not simply links between nodes without meaning, as in graph theory. They may represent evidence to support a claim or an article that is cited. Therefore, using path length alone would not accurately represent these connections. Negation is a problem as it minimally increases the path length and changes the sentiment completely. Also a detailed knowledge graph will differ in path length from a surface graph, even when looking at the same two entities.
Q7	**The degree of the nodes on these paths is impactful to evaluate a claimed relation between A and B. Do you agree?**
if "yes"	*Why do you agree?*
Expl.	The definition of the degree should differentiate between in-degree and out-degree. When focusing on supporting claims, incoming data may be more relevant. Incoming nodes could be given more weight than outgoing nodes, similar to how SEO rankings work on websites. The more links that refer back to a website, the more relevant it is considered. Conversely, if there are many outgoing links, it may not be considered as valuable.
if "other" or "no"	*You answered "Other" what do you mean?* *Why don't you agree?*
Expl.	The degree of nodes can provide information, but its relevance depends on the type of search. If the objective is to find related themes, this approach is suitable. However, if the aim is to determine the semantic similarity between two nodes, the degree of nodes on the way may not be significant. Furthermore, each relationship instance in this section describes a relationship between A and B, making the type of relationship crucial.

Extra	**Do you have additional thoughts on features describing triples in a knowledge graph?** Similarity is a broad term that lacks specificity. It is possible to find similarities between any two things, but it is important to consider the significance of these similarities. It is important to be cautious when drawing conclusions from quantified similarities between two randomly paired items. If the similarity is too vague, then the conclusions are also vague. It is important not to ignore the semantics of the relations or predicates in between. Limiting the number of relations may aid in evaluating the path and its connections. This approach allows for the assignment of arbitrary numbers to the relations, making evaluation easier. This technique can be helpful when seeking corroborating evidence or challenging assertions by assigning positive or negative values, respectively. It is also unclear whether this approach would scale effectively. To begin narrowing down predicates, it is helpful to start with an existing set and expand it based on the user's request or use case by assigning weights instead of values. After that, experiments can be conducted to find better ways of encoding the information. This is a good component to include in the scoring process, but it is only one of many factors to consider. For example, path length and the predicates themselves and the in-degree and the out-degree would all be put into a melting pot and then a scoring mechanism would come out of it. Another scoring approach is to use a system that creates word embeddings, such as vector embeddings based on similarity. Graph walks could be another possible scoring mechanism. This could provide context as it is called in that technique. It could be used to find some sort of relation that is not clearly described in the knowledge graph but can be semantically interpreted, and therefore give a check on similarity with an uncertain triple. Graph walks can be utilised to match claims by expanding the connections of the trusted knowledge graph and counting the number of hops required to reach a similar triple to the one being checked.

	The inclusion of classes would be beneficial. A hierarchy of types could aid in matching or contributing to the scoring mechanism. For instance, although different gases may have varying properties, if they are both gases or both greenhouse gases, they can be considered similar. Expanding the matching criteria to include similar matches, rather than just one-to-one matches, could prove useful. A common approach is to use embeddings to compare the distances between different concepts. In this case, triples can be converted to embeddings and cosine similarity can be used to find the distances. This approach provides more meaningful results than simply comparing the length of paths.

Table 6.3: Questions about score calculation

6.2.3 Summary experts

Do confirm the assumptions extracted from the interviews with a group of experts the second presentation was held inside the ORKG meeting. This time twenty six members of the ORKG team were present for the presentation. The presentation was in a hybrid setting with seven people sitting in the conference room. The presentation was given online.

Assumptions. The interviews confirmed the following steps of my approach.

- Please choose every statement you agree with. ⇛

    1. The optimal method for extracting triples is currently unclear. Nine people chose that answer.
    2. LLMs have limitations and should not be used without proper checks to identify non-reproducible or hallucinated triples. Sixteen people chose that answer.
    3. Scientific accuracy checks must take into account the context of statements, which cannot be represented by a single triple. Thirteen people chose that answer.
    4. This tool is more likely to be used as an integrated rather than a standalone tool. Four people chose that answer.

5. As long as "a perfect algorithm to check against the truth" is not achievable: - Having an indication of what is more or less likely to be accurate is already helpful. Eight people chose that answer.

Since every statement could have been voted for by everyone individually, a maximum of 135 votes was possible. Nineteen individuals took part in the votes. However, only 50 votes have been cast. Especially the question about the usage of the tool was hard to answer when the participant did not witness the first presentation.

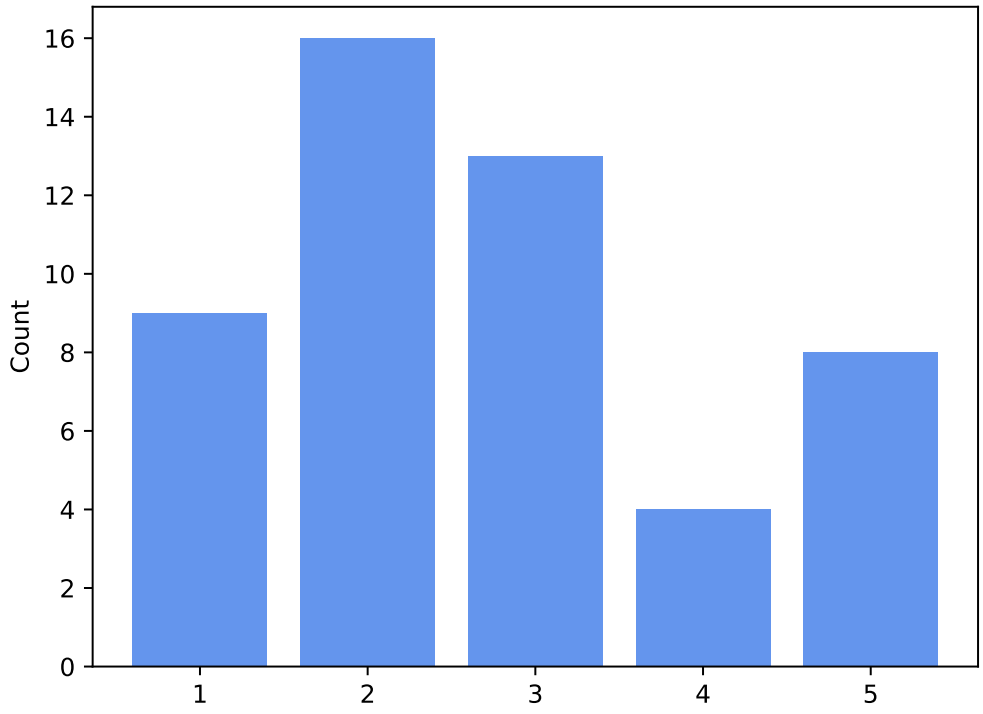


Figure 6.2: Poll results from the assumptions

Core challenges. The following problems were addressed:

- Please choose the two challenges that, when overcome, provide the greatest benefit. ⇛

1. Handling semantic alignment of natural language. Six people chose that answer.
2. Handling LLM hallucinations and lacking reproducibility. Nine people chose that answer.
3. Fully automated triple extraction. No one chose that answer.
4. Keeping the context of statements, especially in empirical research. Eleven people chose that answer.
5. Making statements about statements to show their confidence or time validity. One person chose that answer.

With two votes per person, a maximum of 54 votes was possible. However fifteen individuals took part in the votes and only 27 votes have been cast. This may be due to the relatively short time allotted for the presentation and the fact that not all participants had the opportunity to witness the initial presentation. Nevertheless, there is a clear tendency that future steps should prioritize maintaining the context and ensuring the reproducibility of results, even when working with LLM's. It is also important to note that the objective of fully automating the workflow is not the most pressing issue, provided that other challenges are not addressed.

6.3 Users

The survey was initiated on 5 April 2024 and concluded on 15 April 2024. A total of 65 responses were received, of which 43 were deemed complete. Only the fully completed responses were included in the evaluation and graphics. The participants were distributed as follows: thirty nine were between the ages of 19 and 31, while four were between 55 and 62. Of the participants, ten have completed high school, two have attended a trade school, sixteen have obtained a bachelor's degree, twelve have completed a master's degree, and two have obtained a PhD or higher degree. As illustrated in Figure 6.4, the majority of participants have some experience with science in general. While no one has extensive experience in journalism, natural language processing, or knowledge management.

As illustrated in Figure 6.5, participants perceive misinformation to influence public discourse and be prevalent in certain media formats. No type of media is perceived to be immune to misinformation, although some are perceived to be more susceptible. Participants generally concur that misinformation is a significant issue and challenging to detect.

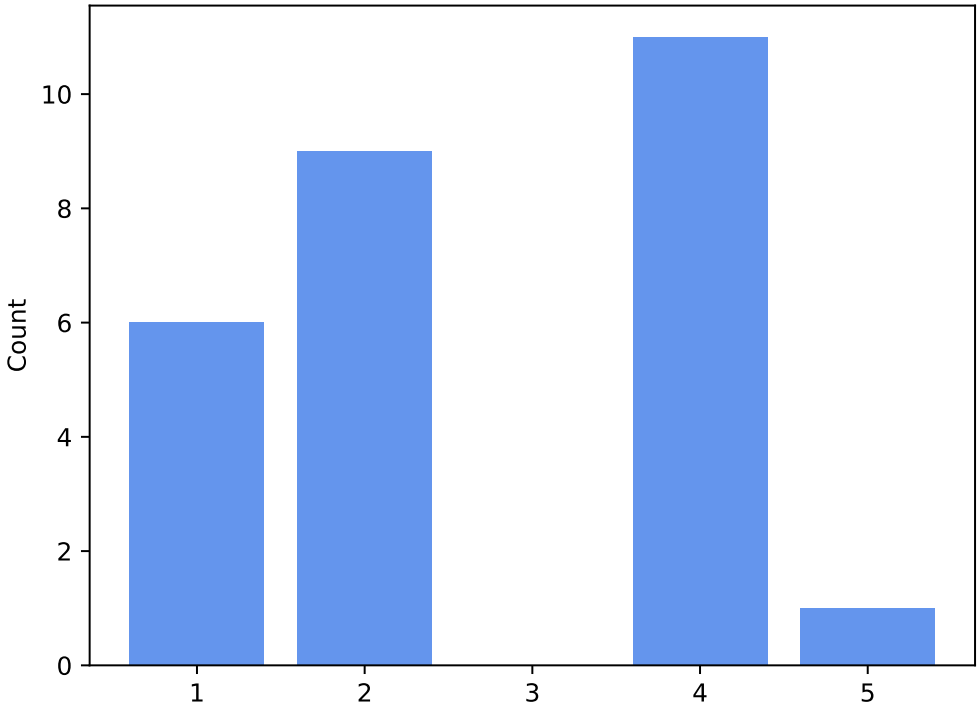


Figure 6.3: Poll results from the core challenges

A comparison of the concept and the current state of the tool, as illustrated in Figure 6.6 and Figure 6.7, respectively, reveals that the concept is one that participants are interested in. The current state is also perceived positively, although to a slightly lesser extent. The concept of the score is perceived as very helpful, as well as innovative and necessary. There is a divergence of opinion as to whether the concept and the current state of the score are confusing. Some participants perceive the current state as transparent, while others do not.

As illustrated in Figure 6.8, the most prevalent manner in which users interact with this tool is through a browser plugin, an app, or a website.

In general, all types of media presented in Figure 6.9 are found to peak interest among users, with the intention of testing them. Only self-written texts and draft laws are slightly less interesting. The highest interest is found in newspaper articles, followed by electoral programmes, political speeches and short viral clips.

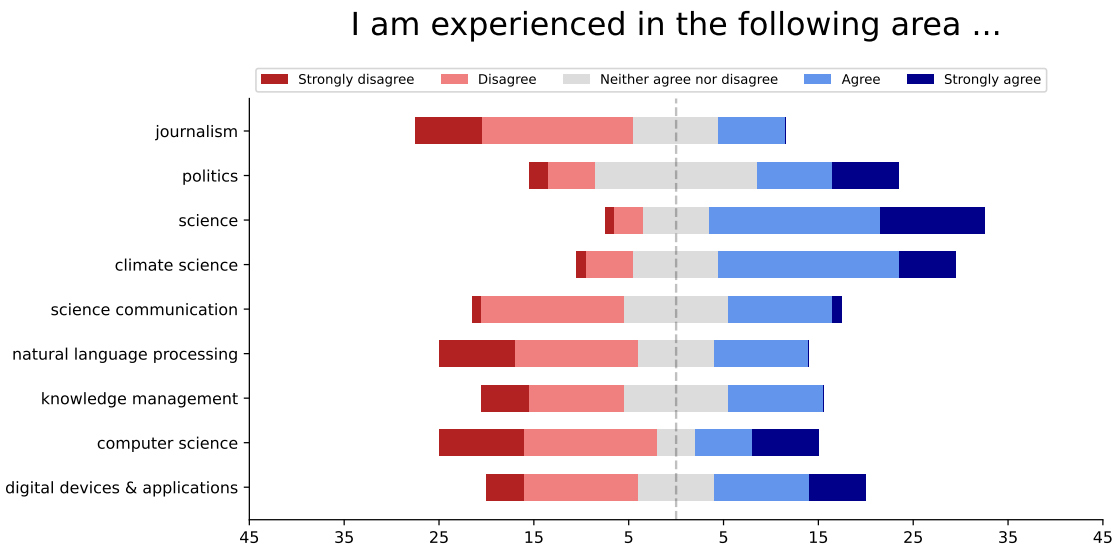


Figure 6.4: Self-reported experience of participants

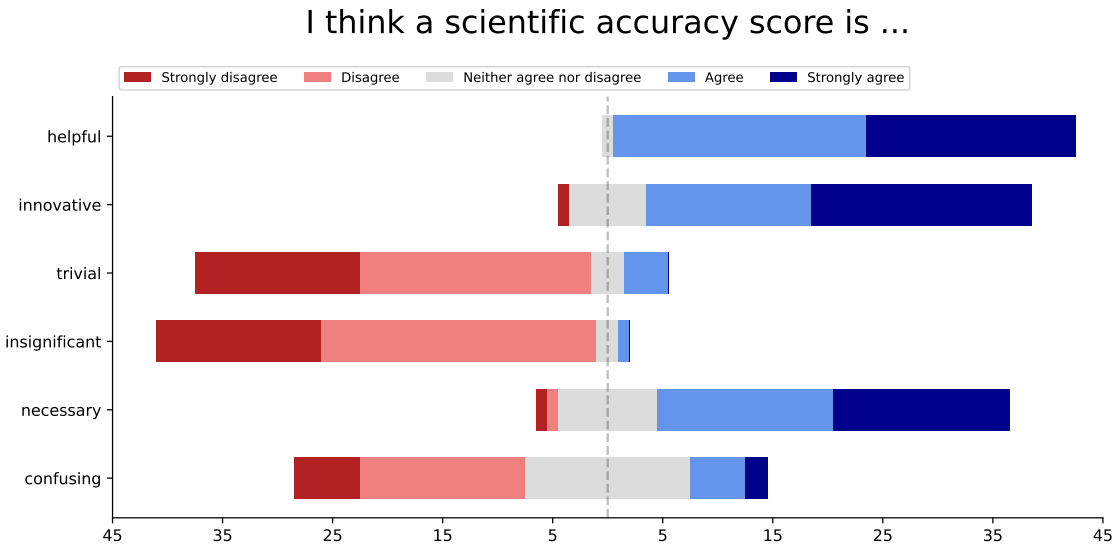


Figure 6.6: Survey results on the concept of scientific accuracy scores

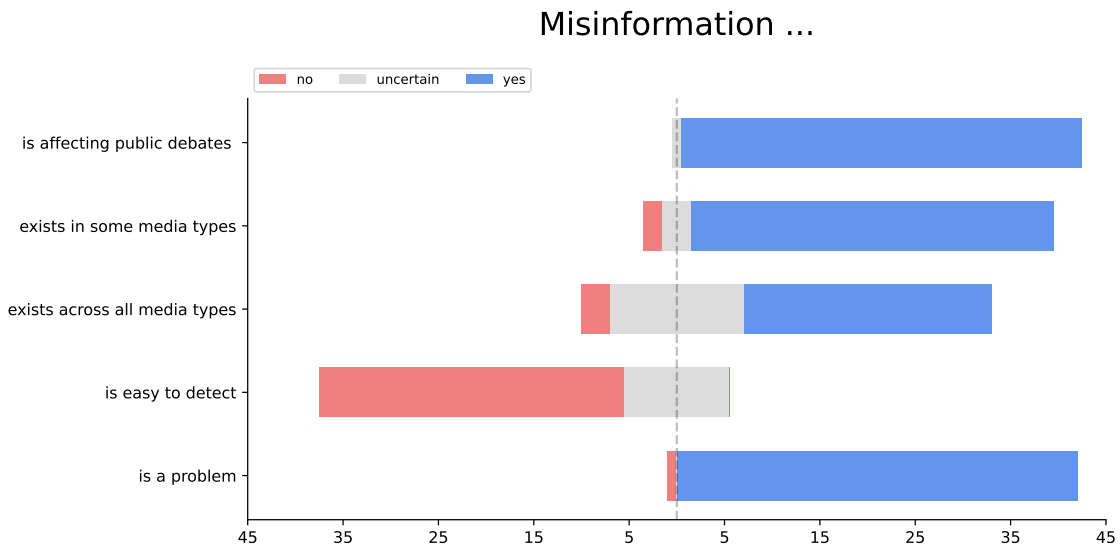


Figure 6.5: Survey results on misinformation statements

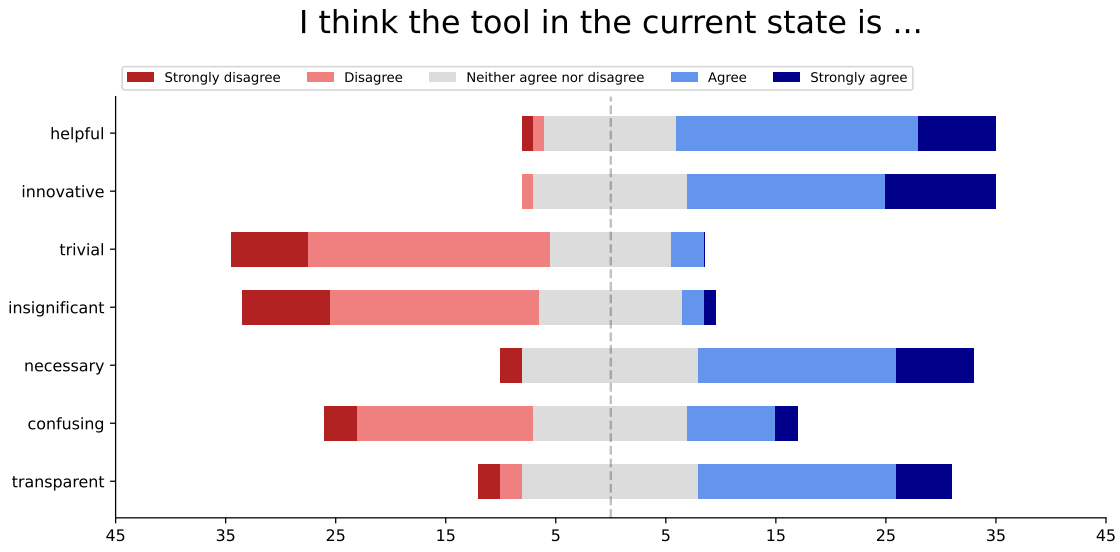


Figure 6.7: Survey results on the current state of scientific accuracy scores

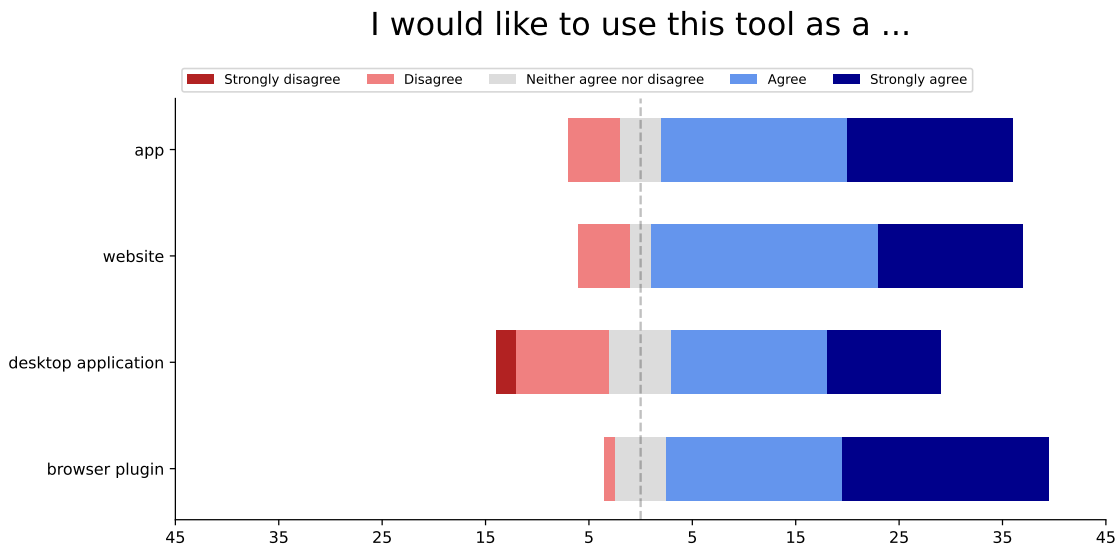


Figure 6.8: Survey results on the representation of the program

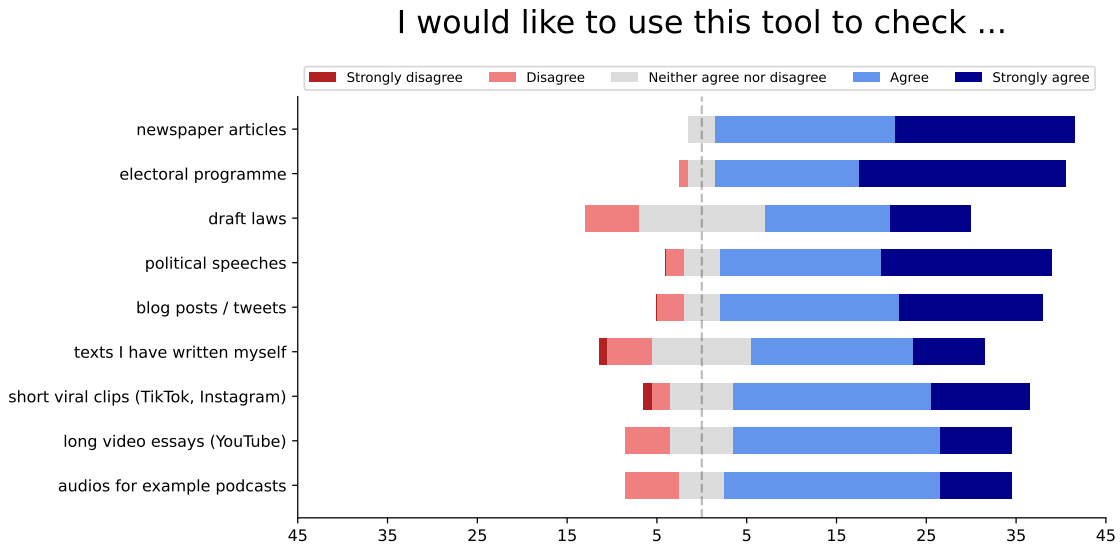


Figure 6.9: Survey results on the use of the program

6.4 Summary

The summary of expert and user evaluation is concluded with a list of confirmed assumptions, limitations and future work.

Confirmed Assumptions.

- the used methods are up to date
- the context of triples is crucial to make a meaningful score
- the best way for extracting triples is uncertain
- the application should be accessible otherwise people will not use it

Limitations.

- Entity Disambiguation
- Semantic interoperability
- RDF expressivity
- reproducebility of LLM outputs
- entity linking with LLMs
- effiency of modeling statements about statements
- for this tool to actually influence the public discourse it needs to be trusted and interpretable
- Because of the small size of the survey this prospective need has to be confirmed in a larger setting.

Future Work. The following additions were mentioned and are now in the future works section.

- Test the tool against a domain expert
- Try to categorize news articles (NY Times vs local article)
- Harder cases than Climate change maybe should be tested

- Implement a time component (articles from the past)
- Full automation would be a much better way to go through large scientific data

Chapter 7

Discussion

This chapter looks back at the research question and discusses advantages, limitations and future work.

7.1 Revisiting the research questions

RQ1. How can natural language processing and knowledge graphs help verify the consistency of secondary literature with scientific findings?

This thesis shows a pipeline that calculates a scientific accuracy score that can be used as indicator to evaluate which media is scientifically accurate. This is done by using a combination of natural language processing and knowledge graphs. The thesis also showcases how this indicator can be improved on various steps of the pipeline and provides an interface that enables further work on specific areas to improve the complete workflow. Next to the implementation there is also an extensive chapter on background and related work that enables people to further work on this topic even when their previous knowledge on these topics is lacking.

RQ2. How can scientifically accurate media on climate change be identified?

This thesis sought to ascertain the types of use cases that the general public desires through the administration of a survey. The findings of that survey can be found in Chapter 6 Evaluation and may be useful in determining which types of media are perceived as trustworthy and which are not.

7.2 Advantages

This thesis has established a foundation for having the need of an indicator that can improve public discourse on scientific topics. Different methods and tools have been researched, put in an overview and implementated as a proof of concept. Furthermore a pipeline with clear interfaces has been created that can be updated at all steps individually.

Domain. The approach is applicable to any domain. However, certain sub-steps, such as triple extraction and NLP methods, demonstrate improvement with domain-specific training. Furthermore, ontology's that enhance the knowledge base are more effective when focused on a particular domain. Therefore, the scope was narrowed by focusing on a specific domain in the hope of achieving a better score. The approach was applied to the domain of climate change, which has a significant number of scientific publications and is an issue that has persisted for a long time without any resolution. Furthermore, it is a domain where a lot of misinformation has been spread.

7.3 Limitations

This thesis represents a proof of concept and an overview of a complex and far-reaching topic. One significant limitation of this thesis was the time constraints and the necessity of working alone. There are numerous points where the current state of the tool is limited but can be improved upon by simply continuing to work on it. These will be described in Future work. In the following paragraphs, some limitations that might limit the approach as a whole will be discussed.

Ground truth. There are social media users who express scepticism about the IPCC. They argue that a considerable number of scientists are not to be trusted. The scientist are perceived as alarmist and motivated by financial gain of research grants. Those who are sceptical of the danger of the climate crisis have cited instances where scientists have been misquoted, either out of context or with false information. The tool can only be as effective as its sources. The survey demonstrated that there are definitely people who would trust a tool like this as long as it is transparent. It would be interesting to know which proportion of the population are sceptical about sources like the IPCC and what their individual reasons are for their scepticism.

On the other hand the IPCC has also been criticised for mitigation and adaptation suggestions that are not being impactful enough.

When chosing the sources it is also important to note that errors can also occur in the peer review of journals. This implies that if one does not replicate every experiment themselves, it is not possible to be 100% certain. Researchers agree that peer review is necessary, but currently sub-optimal [58].

Media sources. In the absence of fully automated scoring, it is necessary to prioritize the order in which media are scored. One possible criterion for this prioritization is to score the most dangerous claims first. Dangerous claims are those whose impact would be most harmful if acted upon. However, it is challenging to ascertain which claims are the most dangerous.

Large language models. In the event that the results yielded by a large language model cannot be replicated and interpreted, this represents a significant challenge for this approach. Furthermore, it is essential to examine the impact of training and utilising large language models, as well as their associated energy consumption.

Availability. There are already numerous fact-checking websites in existence, yet it is questionable whether the majority of users are aware that extensive searches have been conducted. It is evident that fact-checking does not contribute to the advancement of public discourse if the general public is not aware of the checks that have been carried out.

Relevance of information in the domain. One limitation of this approach is that it currently only considers fact statements in the calculation of the score. If the media coverage is distributed equally across all topics, but the scientific community prioritises certain topics, there is a discrepancy. For instance, in the context of climate change, the switch to renewable energy sources represents one important mitigation strategies which is also associated with saving costs in the long run [9]. It is uncertain whether this prioritisation is also reflected in the amounts of media coverage dedicated to that topic. Additionally, within the context of renewable energy, solar and wind are the most prevalent sources. It would be beneficial to ascertain whether the media articles in question have a similar focus and therefore are not distracting.

Facts vs Approaches - what needs to be checked? In the domain of climate change, there is a greater proportion of individuals who acknowledge the existence of anthropogenic climate change, yet there remains a significant number of individuals

who are sceptical about the efficacy of proposed solutions. It is therefore necessary to ascertain the size of this sceptical community in order to inform the prioritisation of resources. In particular, on social media, this group of people is very vocal, and there are news outlets that share misinformation on a regular basis and are fact-checked. The people consuming this news are aware that there are fact-checks evaluating their news source. However, they do not trust those evaluations and some claim that there is no consensus between scientists. This is despite the fact that some of the evidence is based on misquoted publications.

For example, Science Feedback cites authors of papers that have been misused by Bjorn Lomborg in seemingly credible media outlets such as the Wall Street Journal [1] and The Telegraph[2].

7.4 Future work

There is a great deal of work that can be done to expand upon this implementation. At the core of this thesis are information sources. This section will therefore begin with an overview of the various sources of information. These will be divided into two categories: one for reproducible and interpretable information and one for information that is widely available and popular. The remainder of this section will be structured in accordance with the steps of the implementation.

Data sets. Firstly, there are a large number of reliable sources and data sets that can be employed to enhance the ground truth.

- "The Climate Change Performance Index (CCPI)[3] is an instrument to enable transparency in national and international climate politics" [13] The CCPI compares the climate performance of countries which together account for over 90 of global greenhouse gas emissions. [13] This looked like a potentially high quality source of information, but needs further investigation.
- Corporate climate responsibility monitor (CCRM)[4] is a tool which specifically looks at the perspective of companies and industries. Because the industry is

[1] Science Feedback check of Wallstreet Journalhttps://climatefeedback.org/evaluation/bjorn-lomborg-overheated-climate-alarm-wall-street-journal/

[2] Science Feedback check of Telegraphhttps://climatefeedback.org/evaluation/the-telegraph-bjorn-lomborg-in-many-ways-global-warming-will-be-good-thing/

[3] Climate Change Performance Indexhttps://ccpi.org/

[4] Corporate climate responsibility monitor https://www.newclimate.org/resources/publications/corporate-climate-responsibility-monitor-2023

involved in great parts of man made emissions this is an aspect that must be considered.

- skeptical science debunking climate myths[5] is a website of where some scientists have gathered extensive overviews of misinformation[6] spread about climate change. They also offer various data on the consensus of the climate experts.
- Elections24Check[7] is a database which is one of the active projects of the EFCSN and is a database concerned with the 2024 European elections.
- Science Daily Climate Change (SciDCC): The Science Daily Climate Change SciDCC dataset was created by web scraping news articles from the "Earth and Climate" and "Plant and Animals" topics in the environmental science section of the Science Daily (SD) website. The SD news articles are relatively more scientific when compared to other news outlets, which makes SD perfect for extracting scientific-based climate change news. In total, they extracted over 11k news articles from 20 categories relevant to climate change, where each article comprises of a title, summary, and a body. For each category, we were able to extract a maximum of 1k news articles. [51] Before using this data set as ground truth the trustworthiness of Science Daily has to be checked. Otherwise it could become a benchmark for news articles that should show high accuracy scores.
- Lexis Nexis Database: This database provided the news articles for Zhang [69] from Section 3.1. They claim to be the largest online database of international resources with the largest online international database.

Popular information sources. Some people trust in the known, that is why popular source can become trusted. If there is misinformation in a highly trusted source it can be devastating. Misinformation in this sense can also be a evaluated on the relevance of presented topics in comparison to the relevance in the scientific discussion. If there is a diversion from the actual important question it can be equally as bad as consequence. Furthermore if media, such as a newspaper article, references the original source, it is crucial to be able to trace any information back to the primary literature. With regard to journals, it is of paramount importance to ascertain that they have indeed undergone a genuine peer review process.

[5]skeptical science debunking climate myths https://skepticalscience.com/argument.php?f=percentage

[6]Overview of skeptic arguments https://skepticalscience.com/ipcc.php

[7]Elections24Check https://elections24.efcsn.com/about-us

Process media.

- Trusted:
  - using primary literature sources instead of the IPCC
  - include more data sets into the ground truth Section 7.4
  - automated search for sources using keywords
  - retrieving the given sources of media and checking whether or not media is citing them correctly
- Popular:
  - testing approach on coalition agreements, political bills, blog post and social media post
  - testing approach on audio and video
  - testing approach in different languages

Triple extraction.

- enabling larger inputs which requires less errors in a automated setting
- domain specific entity and relation recognition
- handling hallucination of and reproducebility of LLMs
- handling semantic alignment
- extracting more context of triples

Extend knowledge graph.

- handling statements about statements in a large scale efficiently
- verifying a scalable implementation of the knowledge graph
- adding a climate change ontology
- create a domain specific ontology
- implementing a mediator as way of modeling statements the same way to have a better comparability between media and scientific publication

Check veracity.

- better check for a similarity in triples using more characteristics of the knowledge graph e.g. class hierarchies, predicates along the path, in/out degree of nodes

Score triple.

- more score criteria like temporal relevance, clearness, domain relevance, transparency
- give explanation with context from the knowledge graph

Score media.

- creating a visualization / UI
- differentiate the weight of facts in the score calculation
- score domain relevance to avoid diversion
- check bias of articles

The following research and questions should be done and answered in their own papers in ascending order considering the priority with the first being the most important:

1. Creating a large scale knowledge base of up-to-date climate information that can be queried to improve public discussions.
2. Upgrade triple extraction that captures context and is reproducible.
3. How can media with critical impact on public discourse be automatically identified and detected?
4. Creating a browser plugin with an UI to enable user to score media based on scientific accuracy.

Chapter 8

Conclusion

The problem of information overload is well documented, and the lack of an indicator which information is scientifically accurate is a significant issue. This thesis proposes a scientific accuracy score which indicates whether information is based on scientific knowledge or not. This is achieved by giving an overview of the topic, declaring clear interfaces and implementing the semi-automatic prototype. Despite the current score being unsatisfactory in terms of a real-world usage, this implementation has brought together different approaches in the context of climate change. Further differentiation is required between the types of information that can be evaluated. Further expansion is necessary for the coverage of statements that can be checked. Nevertheless, the implementation of a single upgrade to the process results in an immediate improvement in the output. Continued work should focus on the area of information extraction. It is crucial to prioritise the clean extraction of triples over the complete automation of the process. Also, the use of LLMs should be carefully monitored and stopped as soon as the results become inexplicable. The survey indicates that continuing this work is beneficial. A key finding for user acceptance is that the score should be displayed near the original information and have a transparent explanation. Prior to continuing the work, it would be beneficial to engage in an exchange with fact-checking networks. These networks can provide further insights and potentially fund further research. Additionally, it may be worthwhile to explore the potential of full fact AI.

In the context of climate change, the necessity for further work to combat misinformation is of crucial importance in the next century. It is clear that even a relatively minor increase in temperature, such as a fraction of a degree, can have a significant impact on the environment. The scientific community has reached a consensus that calls for the implementation of additional adaptation and mitigation strategies in

the form of policies. The integration of scientific evidence into public discourse can facilitate the translation of scientific knowledge into actionable real-world solutions, which is a way of addressing the challenges posed by climate change.

Glossary

AMR Abstract Meaning Representation is a semantic representation language.. 35

anaconda Anaconda is a tool to create virtual environments. 9, 31

API An Application Programming Interface (API) is a particular set of rules and specifications that a software program can follow to access and make use of the services and resources provided by another particular software program that implements that API. 40

AR6 The sixth assesment report of the IPCC containing three reports written by working groups.. IX

EDMO The European Digital Media Observatory is a european fact checking network. 14

EFCSN European Fact-Checking Standards Network is a european fact checking network. 14

emissions gap The emissions gap is defined as the difference between the estimated global GHG emissions resulting from full implementation of the latest NDCs and those under least-cost pathways aligned with the long-term temperature goal of the Paris Agreement [23]. 1

false balance False balance states opposing statements on a topic in a balanced way even when the scientific evidence on a topic is leaning towards one side heavily. 1

GraphDB GraphDB is a graph database that allows to link diverse data, index it for semantic search and enrich it via text analysis to build big knowledge graphs [45]. 9, 39, 40

ground truth Basis of the evaluation whether information in media should get a high or low score. It should only consist of knowledge as described in Section 2.1 Information source. 6

IFCN International Fact-Checking Network is a international fact checking network. 14

IPCC "The Intergovernmental Panel on Climate Change is the United Nations body for assessing the science related to climate change" [1]. IX

knowledge graph A knowledge graph is a graph of data from the real world, whose nodes represent entities of interest and whose edges represent relations between these entities. [31]. 8, 9, 18, 20

LLM A language model is a probalistic representation of spoken word. A large language model takes millions of trainings data points to fine tune this representation. 7

NER Named Entity Recognition is a NLP method which is a subcategory of entity recognition. Named entities are instances which have a certain name for example specific persons, locations or organizations.. 35

NLP Natural Language Processing is the task of enabling machines to understand human text.. 7

ontology "In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members)." [44]. 38, 66

ORKG "The ORKG makes scientific knowledge human- and machine-actionable and thus enables completely new ways of machine assistance. This will help researchers find relevant contributions to their field and create state-of-the-art comparisons and reviews. With the ORKG, scientists can explore knowledge in entirely new ways and share results even across different disciplines" [46]. 6

OWL "The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of

things, and relations between things. OWL is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit" [47]. 8

peer review "Peer review is the formal quality assurance mechanism whereby scholarly manuscripts (e.g. journal articles, books, grant applications and conference papers) are made subject to the scrutiny of others, whose feedback and judgements are then used to improve works and make final decisions regarding selection (for publication, grant allocation or speaking time)" [58].. 6

primary literature "Primary sources means original studies, based on direct observation, use of statistical records, interviews, or experimental methods, of actual practices or the actual impact of practices or policies. They are authored by researchers, contains original research data, and are usually published in a peer-reviewed journal. Primary literature may also include conference papers, pre-prints, or preliminary reports. Also called empirical research" [27]. 6

RDF "RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. RDF extends the linking structure of the Web to use Universal Resource Identifiers (URIs) to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations." [56]. 8

RDF* "The RDF* extension captures the notion of an embedded triple by enclosing the referenced triple using the strings ¡¡ and ¿¿. The embedded triples, like the blank nodes, may take a subject and object position only, and their meaning is aligned to the semantics of the standard reification, but using a much more efficient serialization syntax." [57]. 9, 39

secondary literature "Secondary literature consists of interpretations and evaluations that are derived from or refer to the primary source literature. Examples

include review articles (such as meta-analysis and systematic reviews) and reference works. Professionals within each discipline take the primary literature and synthesize, generalize, and integrate new research" [27]. 6

spaCy "spaCy is a free open-source library for Natural Language Processing in Python. It features NER, POS tagging, dependency parsing, word vectors and more" [61]. 37

SPARQL "RDF is a directed, labeled graph data format for representing information in the Web. This specification defines the syntax and semantics of the SPARQL query language for RDF. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports aggregation, subqueries, negation, creating values by expressions, extensible value testing, and constraining queries by source RDF graph. The results of SPARQL queries can be result sets or RDF graphs." [62]. 9

SPARQL* To simplify the querying of the embedded triples, SPARQL* extends the query syntax to also query RDF* triples [57]. 9, 39

SRL "Semantic role labeling extracts a high-level representation of meaning from a sentence, labeling e.g. who did what to whom" [63]. 35

tertiary literature "Tertiary literature consists of a distillation and collection of primary and secondary sources such as textbooks, encyclopedia articles, and guidebooks or handbooks. The purpose of tertiary literature is to provide an overview of key research findings and an introduction to principles and practices within the discipline" [27]. 6

whisper "Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web" [34]. 32

Bibliography

[1] *About — IPCC*. URL: https://www.ipcc.ch/about/ (visited on 04/30/2024).

[2] Alan Akbik et al. "FLAIR: An easy-to-use framework for state-of-the-art NLP". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*. 2019, pp. 54–59. URL: https://aclanthology.org/N19-4010/ (visited on 10/30/2023).

[3] Muhammad Nabeel Asim et al. "A survey of ontology learning techniques and applications". In: *Database* 2018 (Jan. 1, 2018), bay101. ISSN: 1758-0463. DOI: 10.1093/database/bay101. URL: https://doi.org/10.1093/database/bay101 (visited on 10/20/2023).

[4] Sören Auer et al. "Improving Access to Scientific Literature with Knowledge Graphs". en. In: *Bibliothek Forschung und Praxis* 44.3 (Dec. 2020). Publisher: De Gruyter, pp. 516–529. ISSN: 1865-7648. DOI: 10.1515/bfp-2020-2042. URL: https://www.degruyter.com/document/doi/10.1515/bfp-2020-2042/html (visited on 04/27/2024).

[5] Laura Banarescu et al. "Abstract Meaning Representation for Sembanking". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. LAW 2013. Ed. by Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 178–186. URL: https://aclanthology.org/W13-2322 (visited on 11/02/2023).

[6] Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. "A Review on Fact Extraction and Verification". In: *ACM Computing Surveys* 55.1 (Nov. 23, 2021), 12:1–12:35. ISSN: 0360-0300. DOI: 10.1145/3485127. URL: https://dl.acm.org/doi/10.1145/3485127 (visited on 10/20/2023).

[7] Jannis Bulian et al. *Assessing Large Language Models on Climate Information*. arXiv.org. Oct. 4, 2023. URL: https://arxiv.org/abs/2310.02932v1 (visited on 11/10/2023).

[8] Pier Luigi Buttigieg et al. "The environment ontology: contextualising biological and biomedical entities". In: *Journal of Biomedical Semantics* 4.1 (Dec. 11, 2013), p. 43. ISSN: 2041-1480. DOI: 10.1186/2041-1480-4-43. URL: https://doi.org/10.1186/2041-1480-4-43 (visited on 11/23/2023).

[9] Katherine Calvin et al. *IPCC, 2023: Full Report [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland*. Edition: First. Intergovernmental Panel on Climate Change (IPCC), July 25, 2023. DOI: 10.59327/IPCC/AR6-9789291691647. URL: https://www.ipcc.ch/report/ar6/syr/ (visited on 09/28/2023).

[10] Alebachew Chiche and Betselot Yitagesu. "Part of speech tagging: a systematic review of deep learning and machine learning approaches". en. In: *Journal of Big Data* 9.1 (Jan. 2022), p. 10. ISSN: 2196-1115. DOI: 10.1186/s40537-022-00561-y. URL: https://doi.org/10.1186/s40537-022-00561-y (visited on 04/28/2024).

[11] Giovanni Luca Ciampaglia et al. "Computational Fact Checking from Knowledge Networks". In: *PLOS ONE* 10.6 (June 17, 2015). Publisher: Public Library of Science, e0128193. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0128193. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128193 (visited on 10/23/2023).

[12] "Ontology Learning from Text". en. In: *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Ed. by Philipp Cimiano. Boston, MA: Springer US, 2006, pp. 19–34. ISBN: 978-0-387-39252-3. DOI: 10.1007/978-0-387-39252-3_3. URL: https://doi.org/10.1007/978-0-387-39252-3_3 (visited on 04/25/2024).

[13] *Climate Change Performance Index (CCPI)*. en-US. Feb. 2024. URL: https://ccpi.org/ (visited on 04/30/2024).

[14] *Climate Misinformation Grant Program (closed)*. en-US. URL: https://www.poynter.org/ifcn/grants-ifcn/climate-misinformation-grant-program/ (visited on 04/23/2024).

[15] John Cook et al. "Quantifying the consensus on anthropogenic global warming in the scientific literature". en. In: *Environmental Research Letters* 8.2 (June 2013), p. 024024. ISSN: 1748-9326. DOI: 10.1088/1748-9326/8/2/024024. URL: https://iopscience.iop.org/article/10.1088/1748-9326/8/2/024024 (visited on 04/23/2024).

[16] Armita Davarpanah et al. "Climate System Ontology: A Formal Specification of the Complex Climate System". In: *Latest Advances and New Visions of Ontology in Information Science*. IntechOpen, May 23, 2023. ISBN: 978-1-80356-918-5. DOI: 10.5772/intechopen.110809. URL: https://www.intechopen.com/chapters/86794 (visited on 11/23/2023).

[17] Danilo Dessí et al. "SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain". In: *Knowledge-Based Systems* 258 (Dec. 22, 2022), p. 109945. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2022.109945. URL: https://www.sciencedirect.com/science/article/pii/S0950705122010383 (visited on 09/28/2023).

[18] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". en. In: ().

[19] Thomas Diggelmann et al. *CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims*. arXiv:2012.00614 [cs] version: 2. Jan. 2021. URL: http://arxiv.org/abs/2012.00614 (visited on 04/23/2024).

[20] *Domo Resource - Data Never Sleeps 11.0*. en. URL: https://www.domo.com/learn/infographic/data-never-sleeps-11 (visited on 04/23/2024).

[21] *EDMO's Collaborative Platform – EDMO*. en-US. URL: https://edmo.eu/resources/edmos-collaborative-platform/ (visited on 04/23/2024).

[22] *EDMOeu – EDMO*. en-US. URL: https://edmo.eu/about-us/edmoeu/ (visited on 04/23/2024).

[23] U. N. Environment. *Emissions Gap Report 2023*. en. Section: publications. Aug. 2023. URL: http://www.unep.org/resources/emissions-gap-report-2023 (visited on 04/23/2024).

[24] Oren Etzioni et al. "Open information extraction from the web". In: *Communications of the ACM* 51.12 (Dec. 1, 2008), pp. 68–74. ISSN: 0001-0782. DOI: 10.1145/1409360.1409378. URL: https://dl.acm.org/doi/10.1145/1409360.1409378 (visited on 10/25/2023).

[25] Luciano Floridi. *Information: A Very Short Introduction*. en. Oxford University Press, Feb. 2010. ISBN: 978-0-19-177734-9. DOI: 10.1093/actrade/9780199551378.001.0001. URL: https://academic.oup.com/book/410 (visited on 04/27/2024).

[26] *Full Fact AI*. en. URL: https://fullfact.org/ai/about/ (visited on 04/23/2024).

[27] Joel Glogowski. *GSU Library Research Guides: Literature Reviews: Types of Literature*. en. URL: https://research.library.gsu.edu/c.php?g=115595&p=1940435 (visited on 04/27/2024).

[28] Michael Wayne Goodman. "Penman: An Open-Source Library and Tool for AMR Graphs". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Asli Celikyilmaz and Tsung-Hsien Wen. Online: Association for Computational Linguistics, July 2020, pp. 312–319. DOI: 10.18653/v1/2020.acl-demos.35. URL: https://aclanthology.org/2020.acl-demos.35 (visited on 11/02/2023).

[29] *Ground News*. en. URL: https://ground.news/ (visited on 04/23/2024).

[30] Luheng He et al. "Deep Semantic Role Labeling: What Works and What's Next". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2017. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 473–483. DOI: 10.18653/v1/P17-1044. URL: https://aclanthology.org/P17-1044 (visited on 11/02/2023).

[31] Aidan Hogan et al. "Knowledge Graphs (extended)". In: *ACM Computing Surveys* 54.4 (May 31, 2022), pp. 1–37. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3447772. arXiv: 2003.02320[cs]. URL: http://arxiv.org/abs/2003.02320 (visited on 10/30/2023).

[32] Angie Drobnic Holan. *The Principles of the Truth-O-Meter: How we fact-check*. en-US. URL: https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/ (visited on 04/23/2024).

[33] *International Fact-Checking Network*. en-US. URL: https://www.poynter.org/ifcn/ (visited on 04/23/2024).

[34] *Introducing Whisper*. en-US. URL: https://openai.com/research/whisper (visited on 04/30/2024).

[35] Saiful Islam et al. "KnowUREnvironment: An Automated Knowledge Graph for Climate Change and Environmental Issues". In: (2022).

[36] Mohamad Yaser Jaradeh et al. *Better Call the Plumber: Orchestrating Dynamic Information Extraction Pipelines*. Feb. 22, 2021. DOI: 10.48550/arXiv.2102.10966. arXiv: 2102.10966[cs]. URL: http://arxiv.org/abs/2102.10966 (visited on 10/30/2023).

[37] Mohamad Yaser Jaradeh et al. "Information extraction pipelines for knowledge graphs". In: *Knowledge and Information Systems* 65.5 (May 1, 2023), pp. 1989–2016. ISSN: 0219-3116. DOI: 10.1007/s10115-022-01826-x. URL: https://doi.org/10.1007/s10115-022-01826-x (visited on 10/30/2023).

[38] Mohamad Yaser Jaradeh et al. “Open Research Knowledge Graph: A System Walkthrough”. en. In: *Digital Libraries for Open Knowledge*. Ed. by Antoine Doucet et al. Cham: Springer International Publishing, 2019, pp. 348–351. ISBN: 978-3-030-30760-8. DOI: 10.1007/978-3-030-30760-8_31.

[39] Mohamad Yaser Jaradeh et al. “Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge”. In: *Proceedings of the 10th International Conference on Knowledge Capture*. K-CAP ’19. New York, NY, USA: Association for Computing Machinery, Sept. 23, 2019, pp. 243–246. ISBN: 978-1-4503-7008-0. DOI: 10.1145/3360901.3364435. URL: https://dl.acm.org/doi/10.1145/3360901.3364435 (visited on 10/27/2023).

[40] Glenn Kessler. “About The Fact Checker”. en-US. In: *Washington Post* (Feb. 2024). ISSN: 0190-8286. URL: https://www.washingtonpost.com/politics/2019/01/07/about-fact-checker/ (visited on 04/23/2024).

[41] Derek J. Koehler. “Can journalistic “false balance” distort public perception of consensus in expert opinion?” In: *Journal of Experimental Psychology: Applied* 22.1 (2016). Place: US Publisher: American Psychological Association, pp. 24–38. ISSN: 1939-2192. DOI: 10.1037/xap0000073.

[42] Mathias Kraus et al. *Enhancing Large Language Models with Climate Resources*. Mar. 31, 2023. DOI: 10.48550/arXiv.2304.00116. arXiv: 2304.00116[cs]. URL: http://arxiv.org/abs/2304.00116 (visited on 11/10/2023).

[43] Mateus Machado and Evandro Ruiz. “Evaluating large language models for the tasks of PoS tagging within the Universal Dependency framework”. In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Ed. by Pablo Gamallo et al. Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, Mar. 2024, pp. 454–460. URL: https://aclanthology.org/2024.propor-1.46 (visited on 04/28/2024).

[44] *Ontology (Computer Science) - definition in Encyclopedia of Database Systems*. URL: http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/ontology-definition-2007.htm (visited on 04/30/2024).

[45] *Ontotext GraphDB*. en-US. URL: https://www.ontotext.com/products/graphdb/ (visited on 04/30/2024).

[46] *Overview - ORKG*. en. URL: https://orkg.org/about/1/Overview (visited on 04/30/2024).

[47] *OWL - Semantic Web Standards*. URL: https://www.w3.org/OWL/ (visited on 04/30/2024).

[48] Martha Palmer, Daniel Gildea, and Paul Kingsbury. “The Proposition Bank: An Annotated Corpus of Semantic Roles”. In: *Computational Linguistics* 31.1 (Mar. 2005), pp. 71–106. ISSN: 0891-2017, 1530-9312. DOI: 10.1162/0891201053630264. URL: https://direct.mit.edu/coli/article/31/1/71-106/1861 (visited on 11/02/2023).

[49] *PENMAN Notation - Penman 1.3.0 documentation*. URL: https://penman.readthedocs.io/en/latest/notation.html#kas1989 (visited on 04/25/2024).

[50] Salvatore Flavio Pileggi and Sawda Alvi Lamia. “Climate Change TimeLine: An Ontology to Tell the Story so Far”. In: *IEEE Access* 8 (2020), pp. 65294–65312. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2985112. URL: https://ieeexplore.ieee.org/document/9055012/ (visited on 11/21/2023).

[51] Mishra Prakamya and Rohan Mittal. "NeuralNERE: Neural Named Entity Relationship Extraction for End-to-End Climate Change Knowledge Graph Construction". In: (June 26, 2021). URL: https://s3.us-east-1.amazonaws.com/climate-change-ai/papers/icml2021/76/paper.pdf.

[52] *Process - How Science Feedback works - Science Feedback*. en. URL: https://science.feedback.org/process/ (visited on 04/23/2024).

[53] *Projects*. en-GB. Apr. 2024. URL: https://efcsn.com/projects/ (visited on 04/23/2024).

[54] Peng Qi et al. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. arXiv:2003.07082 [cs]. Apr. 2020. DOI: 10.48550/arXiv.2003.07082. URL: http://arxiv.org/abs/2003.07082 (visited on 04/26/2024).

[55] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. Dec. 2022. DOI: 10.48550/arXiv.2212.04356. URL: http://arxiv.org/abs/2212.04356 (visited on 04/24/2024).

[56] *RDF - Semantic Web Standards*. URL: https://www.w3.org/RDF/ (visited on 04/28/2024).

[57] *RDF* and SPARQL* — GraphDB Free 9.2 documentation*. URL: https://graphdb.ontotext.com/documentation/9.2/free/devhub/rdf-sparql-star.html (visited on 11/27/2023).

[58] Tony Ross-Hellauer. "What is open peer review? A systematic review". In: *F1000Research* 6 (Aug. 2017), p. 588. ISSN: 2046-1402. DOI: 10.12688/f1000research.11369.2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5437951/ (visited on 04/27/2024).

[59] Benjamin Schubert. *Unsere Faktencheck-Bewertungsskala*. de-DE. Oct. 2018. URL: https://correctiv.org/faktencheck/ueber-uns/2018/10/04/unsere-bewertungsskala/ (visited on 04/23/2024).

[60] *So funktioniert das Facebook-Programm für externe Faktenprüfung*. de. URL: https://www.facebook.com/facebookmedia (visited on 04/23/2024).

[61] *spaCy · Industrial-strength Natural Language Processing in Python*. en. URL: https://spacy.io/ (visited on 04/30/2024).

[62] *SPARQL 1.1 Query Language*. URL: https://www.w3.org/TR/sparql11-query/ (visited on 04/28/2024).

[63] Emma Strubell et al. *Linguistically-Informed Self-Attention for Semantic Role Labeling*. Nov. 12, 2018. arXiv: 1804.08199[cs]. URL: http://arxiv.org/abs/1804.08199 (visited on 11/02/2023).

[64] Zhixing Tan et al. "Deep Semantic Role Labeling With Self-Attention". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 26, 2018). Number: 1. ISSN: 2374-3468. DOI: 10.1609/aaai.v32i1.11928. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11928 (visited on 11/02/2023).

[65] Ann Taylor, Mitchell Marcus, and Beatrice Santorini. "The Penn Treebank: An Overview". en. In: *Treebanks: Building and Using Parsed Corpora*. Ed. by Anne Abeillé. Dordrecht: Springer Netherlands, 2003, pp. 5–22. ISBN: 978-94-010-0201-1. DOI: 10.1007/978-94-010-0201-1_1. URL: https://doi.org/10.1007/978-94-010-0201-1_1 (visited on 04/25/2024).

[66] James Thorne et al. *FEVER: a large-scale dataset for Fact Extraction and VERification*. Dec. 18, 2018. DOI: 10.48550/arXiv.1803.05355. arXiv: 1803.05355[cs]. URL: http://arxiv.org/abs/1803.05355 (visited on 10/20/2023).

[67] Richard E. Turner. *An Introduction to Transformers*. en. arXiv:2304.10557 [cs]. Feb. 2024. URL: http://arxiv.org/abs/2304.10557 (visited on 04/27/2024).

[68] Yifan Xie. "Application of Context Aware Systems to Support Knowledge Work in the Aerospace". en. In: ().

[69] T. Zhang. "What is known about climate change? A knowledge graph approach". University of Oxford, 2021. URL: https://ora.ox.ac.uk/objects/uuid:ca7b0dbb-18ce-4259-8c3a-5d2e89cb8de5 (visited on 10/27/2023).