Friedrich-Schiller-Universität Jena
Institut für Informatik
Studiengang Informatik, B.Sc.

Automatically Estimating the Trustworthiness of Wikipedia Articles

Bachelorarbeit

Luca-Philipp Grumbach

1. Gutachter: Prof. Dr. Matthias Hagen
2. Gutachter: Jan Heinrich Merker, M.Sc.

Datum der Abgabe: 21. Februar 2025

Zusammenfassung

Wikipedia ist eine der meistgenutzen Informationsquellen im Internet und umfasst mehrere Millionen von Artikeln über eine Vielzahl an Themen. Diese Artikel werden von freiwilligen und teilweise anonymen Autoren erstellt und bearbeitet. Als Folge dieser offenen Struktur ist es jedoch möglich, Wikipedia zu missbrauchen um zum Beispiel Falschinformation und Propaganda zu verbreiten. Um die manuelle Überprüfung der enormen Anzahl von Artikeln zu unterstützen, widmen wir uns in dieser Arbeit einem automatischen Verfahren zur Einschätzung der Vertrauenswürdigkeit von Wikipedia-Artikeln.

Wir präsentieren ein Modell, welches auf Basis von manuell annotierten Wikipedia-Artikeln, die Vertrauenswürdigkeit von externen Quellen einschätzen soll. Dazu analysieren wir wie oft eine externe Quelle in Wikipedia-Artikeln referenziert wurde, in denen entweder ein Problem mit der Zuverlässigkeit festgestellt oder ein zuvor festgestelltes Problem gelöst wurde. Aus der Häufigkeit der jeweiligen Vorkommen sollen Rückschlüsse auf einen positiven oder negativen Einfluss der Quelle auf die Vertrauenswürdigkeit von neuen Wikipedia-Artikeln gezogen werden. Unser Ziel ist es, basierend auf den externen Quellen die ein Wikipedia-Artikel zitiert, vorherzusagen ob in diesem Artikel ein Problem mit der Zuverlässigkeit vorliegt oder nicht. Erste Experimente zeigen jedoch, dass unser Modell die Vertrauenswürdigkeit von Wikipedia-Artikeln noch nicht zuverlässig einschätzen kann. Als Gründe für unsere Resultate identifizieren wir sowohl Defizite in den zugrundeliegenden Daten für die Einschätzung der Vertrauenswürdigkeit von externen Quellen, als auch in der vereinfachten Modellarchitektur. Abschließend führen wir eine Diskussion über mögliche Verbesserungen und zukünftige Forschungsrichtungen.

Zusammenfassung

Wikipedia has emerged as one of the most used sources of information on the internet, with millions of articles spanning a wide range of topics. Its collaborative nature, where content is contributed by volunteers worldwide, allows for rapid updates but also creates the possibility of misuse, for example by spreading misinformation and propaganda. In order to support the manual review of the vast number of articles, we explore a possible method for automatically estimating the trustworthiness of Wikipedia articles.

We present a model to assess the trustworthiness of external sources based on manually annotated Wikipedia articles. To do so, we analyze how often an external source was referenced in Wikipedia articles in which either a problem with reliability was identified or a previously identified problem was solved. From the frequency of the respective occurrences, we aim to draw conclusions about a positive or negative influence of the source on the trustworthiness of new Wikipedia articles. For this, we use the external sources referenced in a Wikipedia article to predict whether the article contains a reliability issue or not. First experiments show that our model is not able to reliably assess the trustworthiness of Wikipedia articles yet. As reasons for our results, we identify shortcomings in the underlying data for assessing the trustworthiness of external sources as well as in our simplified model architecture. Finally, we discuss possible improvements and future research directions.

Inhaltsverzeichnis

Kapitel 1

Introduction

Wikipedia has emerged as one of the most used sources of information on the internet,[1] with millions of articles spanning a wide range of topics.[2] Its collaborative nature, where content is contributed by volunteers worldwide, allows for rapid updates but also creates the possibility of misuse, for example by spreading misinformation and propaganda.

As a platform commonly used for education, work, and personal decision-making, as shown by Lemmerich et al. [2018], ensuring the reliability of Wikipedia articles is crucial. While Kräenbring et al. [2014] have concluded that Wikipedia can be as accurate as traditional encyclopedias, concerns about its reliability persist. In the medical field for example, Azer [2014] and Phillips et al. [2014] have highlighted inaccuracies in Wikipedia articles, emphasizing the need for caution when using it as a source of information.

Although Wikipedia employs mechanisms such as editorial oversight, citation requirements, and vandalism detection,[3] the vast amount of articles renders manual verification intensive. One way of supporting this process is through automated tools, that can assist in evaluating article quality and trustworthiness. Research in this field has primarily focused on author reputation, article quality, and edit history. Notable works include the papers of Adler et al. [2008], Moturu and Liu [2009], and Suzuki and Yoshikawa [2012], who have developed models to predict article trustworthiness based on various features automatically.

Estimating the trustworthiness of Wikipedia articles based on their external sources has not been studied extensively. We believe this aspect deserves further consideration because Wikipedia articles strongly depend on references to external sources. This is because Wikipedia itself is not a source of original

[1] `https://www.semrush.com/website/top/`

[2] `https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia`

[3] `https://en.wikipedia.org/wiki/Wikipedia:Editorial_oversight_and_control`

thought but rather a collection of existing knowledge.[4] Consequently, we assume that the trustworthiness of a Wikipedia article is heavily influenced by the trustworthiness of the sources it references. To examine this further, we propose a model that first estimates the trustworthiness of external sources using manually annotated problematic articles and then leverages this information to estimate the trustworthiness of new Wikipedia articles.

1.1 Structure of Wikipedia

Before we go into the details of our approach, it is important to understand the structure of Wikipedia. At the top level, Wikipedia consists of separate projects with their own communities, policies, and guidelines for each language. In this thesis, we focus on the English Wikipedia, the largest and most active project.[5] Wikipedia is organized into pages, each having a unique ID, a title, and a namespace. The namespace groups pages by their type, with each namespace serving a specific purpose. Some examples include articles, categories, templates, user pages, and talk pages.[6] In our analysis, we only estimate the trustworthiness of articles, as they are the main encyclopedia entries providing detailed information on specific topics.[7]

Articles consist of revisions, with each revision representing a snapshot at a particular point in time. When a user makes an edit, the updated content is published as a new revision with a unique and monotonically increasing ID. This is also the case for reverts, which are revisions that undo the changes made in a previous revision.

The second type of page that we intensively use in our analysis are templates. These are pages that are designed to be included in other pages, providing a way to reuse content across multiple articles. They are often used with customizable input and for various purposes, such as navigation, formatting or to display messages for users.[8] In order to include a template in an article, editors place the template's name enclosed by double curly braces within the article text, for example: `{{Template name}}`. When the article is rendered, the template name is replaced with the actual content.[9]

[4] https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not
[5] https://en.wikipedia.org/wiki/English_Wikipedia
[6] https://en.wikipedia.org/wiki/Wikipedia:Namespace
[7] https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article
[8] https://en.wikipedia.org/wiki/Help:A_quick_guide_to_templates
[9] https://en.wikipedia.org/wiki/Help:Template

1.2 Use of External Sources on Wikipedia

External sources are citations of books, articles, websites, and other materials that provide information on a topic while not being part of Wikipedia itself. In the context of Wikipedia, they are officially categorized in **External Links** and **References**.

External Links are hyperlinks to websites that provide additional information on a topic. These hyperlinks should be placed in the *External Links* section, which is located near the bottom of the article. External Links are purely optional and must not be used to verify any content. However, users may follow external links for further reading or meaningful, relevant content that is not deemed suitable for inclusion in the article itself.[10]

References on the other hand are citations to external sources that verify the information presented in an article. References are placed within the text itself and immediately after the statement which they support. On the rendered page, references are displayed as superscript numbers, which link to the full citation in the *References* section at the bottom of the article. This section is created automatically by Wikipedia and contains a list of all references used throughout the article. The references are listed in the order they appear in the article, with each reference having a unique number that corresponds to the superscript number in the text. This system allows readers to easily verify the information presented in the article by checking the corresponding source.[11]

Given that external links are optional and not used for verification, we will only consider references in our analysis. Accordingly, whenever we refer to *external sources*, we mean sources cited by references. If external links are intended, we will explicitly state so.

1.3 Trustworthiness and Reliability

Trustworthiness has various definitions and interpretations, especially in the world of computer science and information systems, as discussed by Viljanen Viljanen [2005]. In order to perform objective estimations, we will use this section to ensure a common understanding first.

In the context of Wikipedia, trustworthiness is closely linked to reliability, where a reliable article is one that provides accurate and unbiased information. To uphold this standard, Wikipedia's content policies require all information to be verifiable, meaning it must be supported by reliable sources.[12] This is also

[10] `https://en.wikipedia.org/wiki/Wikipedia:External_links`
[11] `https://en.wikipedia.org/wiki/Help:External_links_and_references`
[12] `https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies`

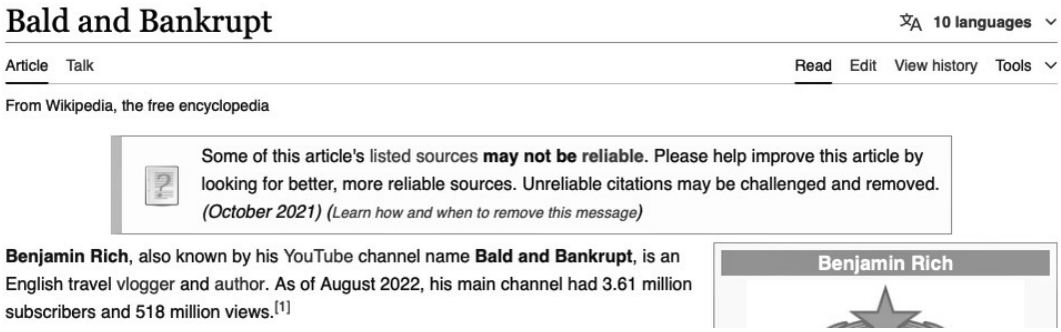

Abbildung 1.1: Example of a Wikipedia article with the *Unreliable sources* template.

reflected in the citation guidelines, which recommend citing reputable sources to ensure the reliability of the information presented.[13] One effective method for assessing article reliability has been described by Wong et al. [2021], which involves filtering revisions to identify the use of templates maintained by the WikiProject Reliability.

WikiProjects are groups of editors, working together with the goal of improving Wikipedia.[14] The WikiProject Reliability aims to improve article reliability by ensuring adherence to Wikipedia's content policies and encouraging robust sourcing practices.[15] To achieve this, the project maintains a collection of templates that editors can use to flag articles with reliability issues. These templates are designed to highlight specific problems, such as dubious statements or references to unreliable sources.[16] Note that in all further parts of this work, we refer to the templates maintained by the WikiProject Reliability as *reliability templates*. When an article contains a reliability template, a warning is shown at the top of the article, providing a clear signal to readers and editors that the article may have issues. An example of the *Unreliable sources* template is shown in Figure 1.1, where a text box at the top of the article can be seen.

Our definition of trustworthiness uses these reliability templates as indicators of editorial assessments. The *addition* of a reliability template to an article is seen as a negative signal, indicating that the article contains an issue regarding its reliability. We say that articles containing reliability issues are less trustworthy. Conversely, the *removal* of a reliability template is seen as a positive signal, indicating that the article has been improved and is now more

[13] `https://en.wikipedia.org/wiki/Wikipedia:Citing_sources`
[14] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject`
[15] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Reliability`
[16] `https://en.wikipedia.org/wiki/Template:Citation_and_verifiability_article_maintenance_templates`

reliable. We say that articles where reliability issues have been resolved are more trustworthy. Although the presence of such a template is a strong signal for reliability issues, its absence does not necessarily imply that an article is trustworthy. Articles containing reliability issues may not have been flagged yet, or the issue may have been overlooked. To address this, we focus on revisions where templates were deliberately added and then removed by editors, as this reflects conscious human evaluation. Additionally, for the purpose of this thesis, we analyze reliability templates independently, as evaluating their interconnections lies beyond the scope of this work.

For our analysis, it is reasonable to prioritize templates related to reliability issues of referenced sources, since our approach is fundamentally based on evaluating the connection of sources with these templates. Although the WikiProject Reliability offers a wide range of source-related templates,[17] we are interested in the commonly used templates to ensure a sufficient number of data points for our analysis. The more often a template is used, the more unique external sources we can analyze and the more information we can gather about the trustworthiness of these sources.

For that reason, we narrowed down our selection to two common reliability templates concerning issues with the reliability of referenced sources. Specifically, we focus on the templates **Unreliable sources** and **Dubious**. The *Unreliable sources* template is used to flag articles where some of the referenced sources are assumed to be of questionable reliability, recommending to look for better and more reliable sources.[18] The *Dubious* template is used to mark statements that are doubtful or questionable and require verification. It may also be used to raise questions on the veracity, accuracy, or methodology employed by a given source.[19]

1.4 Outline of the Approach

The first step of our approach is to extract data from a Wikipedia dump, parsing revisions to identify the presence of the *Unreliable sources* and the *Dubious* templates. Note that all steps are performed separately for each template. We select revisions where a template was added and then search for the first subsequent revision where that template is no longer present. This process results in revision pairs, that capture the states of articles when they were considered problematic and the following state when the reliability issue

[17] `https://en.wikipedia.org/wiki/Wikipedia:Template_index/Cleanup/Verifiability_and_sources`

[18] `https://en.wikipedia.org/wiki/Template:Unreliable_sources`

[19] `https://en.wikipedia.org/wiki/Template:Dubious`

was resolved. For both revisions in a pair, we extract the referenced external sources and store them for further analysis.

In the second step, we estimate the trustworthiness of referenced sources. We analyze whether a source is more likely to be present during the addition or during the removal of a reliability template. In practice, this is done by counting how often each source is used in a revision where a template was added and how often in a revision where a template was removed. We are left with two counts for each source, which we use to compute the respective probabilities by dividing each count by the sum of both counts. These probabilities serve as a means of trustworthiness, indicating how likely a source is to be associated with the addition or removal of a reliability template. We say that a source is more trustworthy if it is more likely to be present when a reliability template is removed, as this indicates that the source could be associated with an improvement in reliability. Conversely, we say that a source is less trustworthy if it is more likely to be present when a reliability template is added, as this indicates that the source could be associated with a reliability issue.

In the third step, we estimate the trustworthiness of Wikipedia articles. Given a revision of an article, we look up the sources referenced in the revision in our trustworthiness estimates and for each source retrieve the probabilities of being present during the addition or the removal of a reliability template. Next, we average the probabilities of all referenced sources. We are left with the probability of the revision being associated with the addition of a reliability template and the probability of the revision being associated with the removal of a reliability template. Similarly to the trustworthiness estimation of referenced sources, we say that a revision is more trustworthy if it is more likely to be associated with the removal of a reliability template, and less trustworthy if it is more likely to be associated with the addition of a reliability template.

Kapitel 2

Related Work

Substantial research has been conducted on the automatic estimation of trustworthiness in Wikipedia articles, with most approaches relying on a combination of textual and user-related data. One example is the research done by Moturu and Liu [2009], who define trust as a combination of quality and credibility. Quality is derived from content-related metrics, such as the proportion of paragraphs with citations and the overall size of the article. Credibility on the other hand is determined by user behavior, including editing patterns and the article's development history.

Another example is the work of Adler et al. [2008], who propose a mechanism that assigns a value of trust to each word based on the word's survival ratio and the reputation of the editor who contributed the word. The survival ratio describes how many revisions of an article a word has endured without being altered or removed. Words that survive more revisions are seen as more trustworthy. The reputation of the editor is based on the survival ratio of the words they contribute. Editors who consistently add words that remain in an article over time see their reputation increase, while those whose contributions are frequently edited or removed experience a decline in reputation.

The combination of survival ratio and editor reputation has been explored further by Suzuki and Yoshikawa [2012], who propose a method designed to be more resistant to vandalism. Their approach evaluates text quality based on both the survival ratio of words and the reputation of the editors. The reputation of an editor is calculated as the average quality of the text they contribute. Initially, this creates a challenge since text quality and the editor's reputation are interdependent. To address this, the process begins with a fixed value for editor reputation, which is then used to calculate text quality. In subsequent steps, the values for text quality and editor reputation are updated iteratively until they converge.

An example of a more recent and innovative approach is the research of

Wong et al. [2021]. Similarly to our work, their objective is to predict the presence of reliability templates. To achieve this, they compile a dataset consisting of articles where a selected template was added, paired with the corresponding revisions where it was later removed. For each pair, they also extract various features, such as the number of words, images, citations, and external references. Using these features, they then train machine learning models to predict whether the template would be present or not. However, their results fall short of expectations, with the best-performing model achieving an accuracy of 62%. They conclude that while the metadata features they used offer some predictive value, the task is inherently challenging and demands further research.

Kapitel 3

Revision Extraction

Before estimating the trustworthiness of Wikipedia articles and their referenced sources, we need to obtain data to work with. In this chapter, we will describe the necessary steps such as choosing reliability templates, parsing a Wikipedia dump, and extracting relevant revisions with their referenced sources.

3.1 Selection of Reliability Templates

When selecting reliability templates, the first consideration is the relation of the template to referenced sources. Since our approach is based on associating external sources with additions and removals of reliability templates, we seek templates that are strongly related to the trustworthiness of referenced sources. We assume that source-related templates have a stronger connection to the used references than other templates that question the structure or tonality of the article. However, templates such as **Unreferenced** or **Citation needed** are not considered, as they only criticize the quantity of references and not directly the trustworthiness of the referenced sources.

The second consideration is how commonly the template is used. While the WikiProject Reliability maintains a variety of templates related to the reliability of the referenced sources, for most of them we extracted only a few hundred revision pairs. Some templates that we excluded due to extracting less than 200 revision pairs are listed in Table 3.1. We excluded these templates because we assume that our model benefits from larger amounts of data to learn from to estimate the trustworthiness of new Wikipedia articles more accurately. The more often a template is used, the more revision pairs we can extract. This in turn can allow us to estimate the trustworthiness of a wider range of external sources, which is crucial for the reliability of our model. With each additional revision pair, we might not only identify new unique sources but also increase the number of data points used to estimate the trustworthiness of

Template	Description	Number of revision pairs
Better sources needed	Used to flag articles that need better or more reliable citations.[a]	21
Circular	Used if a referenced source previously got its information from Wikipedia.[b]	128
Independent sources	Used to flag articles that rely on sources too close to the subject, thus likely being biased.[c]	65
No reliable sources	Used to flag articles where all of the referenced sources are considered unreliable.[d]	9
User-generated	Used to flag articles where multiple referenced sources are user-generated content, meaning the content was written and published by random members of the public.[e]	86

Tabelle 3.1: Description of the excluded reliability templates.

[a] `https://en.wikipedia.org/wiki/Template:Better_sources_needed`
[b] `https://en.wikipedia.org/wiki/Template:Circular`
[c] `https://en.wikipedia.org/wiki/Template:Independent_sources`
[d] `https://en.wikipedia.org/wiki/Template:No_reliable_sources`
[e] `https://en.wikipedia.org/wiki/Template:User-generated`

already extracted external sources. If a source is not referenced at all during the addition or removal of a reliability template, we cannot make any assumptions about the trustworthiness of the source, as the article that references it might not have been checked for reliability issues by editors yet. We believe that any referenced source in a Wikipedia article for which we have no trustworthiness estimate increases the uncertainty of the article's trustworthiness estimate. As a result, we aim to retrieve as many revision pairs as possible to estimate the trustworthiness of as many external sources as possible.

The templates we chose to work with are **Unreliable sources** and **Dubious**. Both are widely used with over 12,000 and 43,000 revision pairs respectively. The *Unreliable sources* template is used to flag articles where some of the referenced sources may be unreliable,[1] while the *Dubious* template is used to mark statements or alleged facts that seem dubious despite being sourced.

[1] `https://en.wikipedia.org/wiki/Template:Unreliable_sources`

The *Dubious* template may also be used to question the accuracy or methodology used by a given source.[2] Since both templates challenge the reliability of the referenced sources, we consider them suitable for our purposes.

3.2 Parsing Wikipedia Dumps

To gain as much information on reliability templates and external sources as possible, we use the full revision history of Wikipedia articles. The Wikimedia Foundation, which is the organization behind Wikipedia, provides regular data dumps of the entire Wikipedia database.[3] A full dump contains all revisions of all pages, which expand to multiple terabytes of text when decompressed. The full dump holds an extensive amount of data, including all templates and references that have ever been used in Wikipedia articles. Fortunately, this data is provided in multistream XML format, which allows for efficient parallel processing.

In our case, we already have access to previously downloaded Wikipedia dumps on a computer cluster. Being the most recent, we work with the Wikipedia dump from September 2022 in all our experiments. The bzip2[4] compressed XML data is dispersed among 770 multistream files which sum up to around 1.3 terabytes in size. To handle this amount of data, we use Apache Spark, a distributed computing framework that enables efficient processing of large datasets.[5] We leverage Apache Spark's capabilities using Scala and multiple jobs that run on a Kubernetes cluster. Each job is a piece of code, designed to perform a specific task, such as extracting revisions from a Wikipedia dump or estimating the trustworthiness of an article's referenced external sources. This allows us to scale our processing power according to the size of the data we are working with in each step.

The job responsible for parsing the Wikipedia dump is the first one we run. It streams Wikipedia's XML data and decompresses it on the fly, filtering out all Wikipedia pages that are not main articles. Redirects are also ignored, as they do not contain any information themselves and automatically navigate the user to another page. Next, we use both XML parsers and regular expressions to filter revisions for the presence of a selected reliability template. Further information on the template such as the date or an editor's note is ignored. Much like Wong et al. [2021] has done, we extract revisions where a reliability template was added and pair it with the first subsequent revision in which

[2] `https://en.wikipedia.org/wiki/Template:Dubious`
[3] `https://dumps.wikimedia.org`
[4] `https://sourceware.org/bzip2/`
[5] `https://spark.apache.org`

that template is no longer present. This results in a balanced dataset, where for each revision in which a reliability issue was detected, we have a corresponding revision in which the issue was resolved. The revisions are later saved as pairs, so we can easily compare the state during the addition of the template and the state during the removal of the template. However, most of the content of a revision is not of interest to us, since we are only concerned with the external sources that are referenced in the article. Therefore, for each revision in a pair, we only store the ID for later identification and the referenced external sources.

Using pattern matching mechanisms, which look for reference tags and citation templates we can find the majority of the referenced external sources of any Wikipedia revision. It does not suffice to simply scan a revision for URLs or ISBNs since we intend to avoid external links (see 1.2). Next, we clean the extracted sources by removing unnecessary characters such as extra braces or commata, which are occasionally present, presumably due to syntax errors by editors. To further ensure that we can compare sources across revisions, we also have to normalize URLs by only storing their registrable domains. Simply put, we save the highest-level domain that is controlled by a single entity. For example, the registrable domain of `https://en.wikipedia.org/wiki` is `wikipedia.org`. This is done by using the public suffix list, which is a collection of all top-level domains and their subdomains.[6] Normalized URLs allow for an evaluation on a broader scale instead of considering specific links. We assume that if a domain such as a news website is trustworthy, then all published articles from that domain are likely to be trustworthy as well, even if they come from different authors or even different subdomains. For books, this is much simpler, as we can just store the ISBN to identify books even when their citations are formatted differently. Lastly, all data is stored in Parquet[7] files, which is a columnar storage format that is optimized for big data processing.

3.3 Train and Test Sets

To analyze the effectiveness of our model, we reserve some data to compare the trustworthiness estimates of unknown articles against ground truth values. This requires splitting the dataset of revision pairs into training and test sets: we estimate the trustworthiness of external sources using the larger training set and then make predictions for unseen revisions in the smaller test set. For each revision in the test set, we have known labels indicating whether a reliability template was added or removed. These labels are unknown to our model because they were not used for trustworthiness estimations of external sources.

[6] `https://publicsuffix.org`
[7] `https://parquet.apache.org`

Template	Training Set	Test Set
Unreliable sources	9,942	2,485
Dubious	34,708	8,677

Tabelle 3.2: Number of revision pairs in the training and test sets for each reliability template.

Our goal is to predict these labels based on the external sources referenced by the revisions of the test set and to evaluate the effectiveness of our model by comparing the predictions to the actual labels.

To split our dataset, we use an 80/20 ratio, meaning 80% of the revision pairs are used for training and 20% for testing. This provides a good balance between having sufficient data to train on while still having a substantial test set. Typically, the data that will be used for training and the data that will be used for testing is selected randomly from the whole dataset. In our case, we aim to replicate a more realistic scenario. To achieve this, we sort the revision pairs in our dataset by the revision ID of the first revision in each pair and use the first 80% for training. Since revision IDs generally increase over time, this approach allows us to model a real-world situation where the model is trained on past, already evaluated Wikipedia revisions and tested on more recent, unevaluated revisions. To provide some information on the size of the datasets we used, we list the number of revision pairs in the training and test sets for each template in Table 3.2.

Kapitel 4

Trustworthiness Estimation of External Sources

After extracting revision pairs for a given reliability template and determining a test split, we estimate the trustworthiness of referenced sources. The goal of this step is to assess the likelihood that a source is associated with the addition or the removal of a reliability template. Note that similarly to the other steps of our approach, the estimation of source trustworthiness is done for one reliability template at a time.

4.1 Trustworthiness Estimation Process

We begin by loading all revision pairs from the training set and iterate over the referenced external sources of both revisions in the pair. To explain which referenced sources we consider in our analysis and how we estimate their trustworthiness, we will use the following example. Let $T_{Addition}$ be a revision where a reliability template was added and $T_{Removal}$ the first subsequent revision of the same Wikipedia article where the template was removed. For each external source, we determine whether it was referenced in $T_{Addition}$, $T_{Removal}$, or both. While doing so, we can distinguish between the three scenarios S_1, S_2, and S_3, as displayed in Figure 4.1.

S_1 describes an external source referenced in the revision where the reliability template is added, but not in the revision where the template is later removed. We say that this source is likely to be associated with the *addition* of the reliability template and is by our definition of trustworthiness less trustworthy. S_2 describes an external source referenced in the revision where the reliability template is removed, but not in the revision where the template was added. We say that this source is likely to be associated with the *removal* of the reliability template and is by our definition of trustworthiness more trustwor-

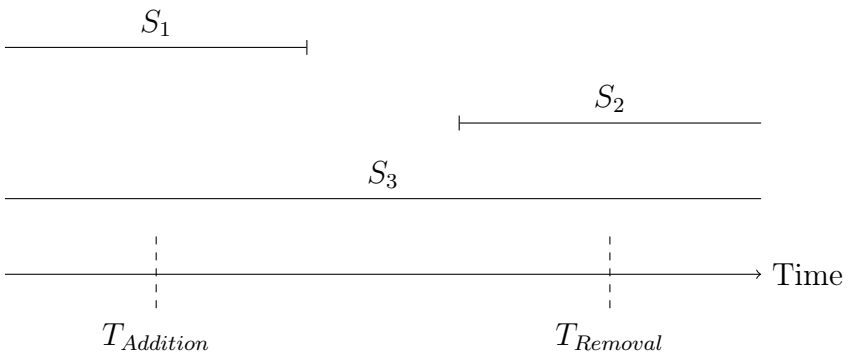

Abbildung 4.1: Timeline illustrating the time spans for sources S_1, S_2, S_3 relative to the events $T_{Addition}$ and $T_{Removal}$.

thy. S_3 describes an external source referenced in both the revision where the reliability template is added and the revision where the template is removed. Sources such as S_3 are ignored in our analysis. We assume that S_3 could not have been responsible for the addition of the template, as it was still present when the template was removed. Similarly, we assume that S_3 could not have been responsible for the removal of the template, as it was already present when the template was added. Furthermore, our program cannot distinguish whether a reliability template applies to the entire article, a specific section, or even a single statement. While one might assume that any referenced source that survived the removal of a reliability template is not problematic, this remains speculative. In larger articles with numerous references, it is likely that sources referenced in sections unrelated to the template were not checked by the editor who added or removed a reliability template. By considering only referenced sources that were added to or removed from the article, we aim to capture those sources that are most relevant to the template change, as we know that they were actively considered by an editor.

To perform the trustworthiness estimation of referenced external sources across all revision pairs, we count how often each source appears in scenario S_1 and how often each source appears in scenario S_2. This results in two counts for each source, for which we compute probabilities by dividing each count by the sum of both counts. These probabilities indicate the likelihood of a source being associated with the addition or the removal of a reliability template. The limitation of this approach is that we cannot distinguish whether a source was referenced in many revisions but mostly stayed unaffected by a reliability template or if a source was referenced rarely but often associated with a reliability template. In the first case, the source might have been commonly referenced but played no crucial role in the addition or the removal of a reliability template. There is no trivial method of using this source for predicting

the presence of reliability templates, as it seems to have no strong connection to the addition or the removal of a reliability template. In the second case, the source might have been referenced rarely but mostly in connection with a reliability template. Here we should be cautious of noise such as vandalism, wrong evaluation by an editor, or simply because the reference to the source was added or removed for reasons unrelated to the reliability template. This is because when relying on a few data points, the trustworthiness estimate of a source is easily skewed by outliers and noise. We think that for a simplified approach like ours, the distinction between these two cases is not necessary. In both cases, the source is not a reliable indicator for the addition or removal of a reliability template and should be treated with caution.

Lastly, we store the registrable domain or ISBN of each external source, along with the computed probabilities of being associated with a template addition or a template removal and the source's number of occurrences. The number of occurrences describes how often a source was referenced during the addition (S_1) or the removal (S_2) of a reliability template but not in both (S_3). Storing the number of occurrences ensures that we can later filter out sources that have rarely been associated with a template addition or removal, as they might be less reliable indicators for the presence of a reliability template. It also opens up the possibility of weighting sources by their occurrences when estimating the trustworthiness of Wikipedia articles, which we will discuss in Chapter 5.

4.2 Results

For the *Unreliable sources* template, we computed trustworthiness estimates for 22,278 unique sources. To better understand our model's results, we manually verify the results of the top 10 sources (by occurrence), as seen in Table 4.1. Unless specified otherwise, the manual evaluations are taken from Wikipedia's list of frequently discussed sources by Wikipedia contributors [2025]. Here we observe that all sources generally considered reliable have a higher probability of being referenced during the removal of the *Unreliable sources* template. Furthermore, the sources generally considered unreliable have a higher probability of being referenced during the addition of the *Unreliable sources* template. This is a strong indicator that our model is working as intended. For a brief manual evaluation of the top sources, please refer to Table 4.2. Note that in the presented tables, $P(T_{Addition})$ and $P(T_{Removal})$ are the probabilities of a source being present during the addition or the removal of a specified template, respectively.

Looking at the trustworthiness estimates of the same 10 sources using the *Dubious* template, we see different results (see Table 4.3). Although all sources

that are generally considered reliable are more often associated with a template removal, the probabilities are not always as clear-cut as with the *Unreliable sources* template. For example `nytimes.com` has no strong association with either the addition or the removal of the *Dubious* template. Sources that are generally considered unreliable on the other hand are not always more likely to be present during template additions. Examples of this are `youtube.com`, `imdb.com`, and `wordpress.com`, which are all present more often during the removal of the *Dubious* template than its addition. This trend may be explained by the fact that while they are generally unreliable due to user-generated content, they may also host credible material in certain contexts. Authoritative sources might publish high-quality content on YouTube or WordPress, and IMDb includes structured, verified data on film credits and production details.[1] The presence of these sources in *Dubious* template removals might suggest that Wikipedia editors sometimes accept authoritative content from these platforms to replace other questionable sources. Another explanation is that the *Dubious* template itself leaves room for interpretation, as it may be used in a variety of cases concerning specific statements or alleged facts that are sourced but seem dubious. When an author misinterprets a source or uses it out of context, the source itself might not be questionable, but the template may still be used.[2]

This highlights the complexity of estimating the trustworthiness of sources based on their association with reliability templates. While the *Unreliable sources* template allows for a clearer distinction between reliable and unreliable sources, the evaluation of sources becomes more nuanced when templates such as *Dubious* allow for a wider range of interpretations. This is an important aspect to consider when interpreting the results of our trustworthiness estimation, as the computed estimates should always be seen in the context of the specific reliability template they were computed for.

[1] `https://help.imdb.com/article/imdb/general-information/where-does-the-information-on-imdb-come-from`

[2] `https://en.wikipedia.org/wiki/Template:Dubious`

Source	$P(T_{Addition})$	$P(T_{Removal})$	**Occurrences**
archive.org	0.1515	0.8485	931
google.com	0.2627	0.7373	571
youtube.com	0.5674	0.4326	527
imdb.com	0.6893	0.3107	338
nytimes.com	0.2862	0.7138	269
facebook.com	0.6653	0.3347	245
twitter.com	0.6000	0.4000	215
wikipedia.org	0.7011	0.2989	184
theguardian.com	0.1548	0.8452	168
wordpress.com	0.5897	0.4103	156

Tabelle 4.1: Top 10 sources with their probabilities and occurrences for the *Unreliable sources* template

Source	**Manual Evaluation**
archive.org	Mostly reliable and factual. It hosts books, papers, and other documents without containing original thoughts.[a]
google.com	Difficult to evaluate directly as Google itself is not typically a direct source of information. Possible subdomains such as Google Books and Google Scholar are generally reliable.
youtube.com	Generally unreliable. The videos are mostly anonymous, self-published, and unverifiable.
imdb.com	Unreliable due to user-generated content.
nytimes.com	Generally reliable.
facebook.com	Generally unreliable due to self-published content.
twitter.com	Generally unreliable due to self-published content.
wikipedia.org	Unreliable due to self-published content.
theguardian.com	Generally reliable.
wordpress.com	Generally unreliable due to self-published content.

Tabelle 4.2: Manual evaluation of the top 10 sources for the *Unreliable sources* template

[a] `https://mediabiasfactcheck.com/internet-archive-bias/`

Source	$P(T_{Addition})$	$P(T_{Removal})$	**Occurrences**
archive.org	0.2124	0.7876	1803
google.com	0.3282	0.6718	1868
youtube.com	0.3898	0.6102	449
imdb.com	0.4471	0.5529	208
nytimes.com	0.4882	0.5118	805
facebook.com	0.5441	0.4559	68
twitter.com	0.5077	0.4923	65
wikipedia.org	0.5885	0.4115	243
theguardian.com	0.1526	0.8474	308
wordpress.com	0.4170	0.5830	223

Tabelle 4.3: 10 selected sources with their probabilities and occurrences for the *Dubious* template

Kapitel 5

Trustworthiness Estimation of Articles

After estimating the trustworthiness of external sources, we estimate the trustworthiness of Wikipedia articles. Specifically, we predict for any revision in our test set whether a template was added or removed. To make these predictions, we use the previously computed probabilities of the external sources being associated with a reliability template addition or a reliability template removal.

5.1 Template Prediction Process

At first, we load all revision pairs of the test set and iterate through the referenced external sources of each revision. For each source, we look up the previously computed probabilities of being associated with a template addition or a template removal. If a source is not found, it has not been used in the training data. These sources are referred to as unknown sources, to which we assign default probabilities of 0.5 for being associated with a template addition or a template removal. The probabilities of 0.5 indicate that the source is neutral and does not provide any information on the addition or removal of a reliability template. Furthermore, we set the source's occurrences to zero. By still including the source instead of ignoring it completely, we can decide in the next step whether to include the source in our computations, allowing for experimentation with different approaches.

After loading the probabilities for all external sources referenced by a revision from a data file, we first combine the probabilities of the sources being associated with a template addition and then combine the probabilities of the sources being associated with a template removal. We do this to estimate the probability of the revision being associated with a template addition and to estimate the probability of the revision being associated with a template remo-

val. The consolidation of multiple probabilities can be done in different ways. We have chosen to use a weighted average, where the weight of a source is determined by its number of occurrences. The weight can be interpreted as a measure of confidence that the estimated trustworthiness of a source represents the true trustworthiness. In statistical analysis, the amount by which an estimate might deviate from the true value is called the *margin of error*. Therefore, the smaller the margin of error for a trustworthiness estimate, the more confident we are that the estimate is accurate. The margin of error is determined by the number of data points, where more data points result in a smaller margin of error. Based on this concept, we assume that the more often a source occurs in our training data, the more accurate the trustworthiness estimate of the source is. If a source's trustworthiness estimate is based on only a few data points, we say that it is more likely to be inaccurate and consequently should have less influence on the trustworthiness estimate of the article that references the source. However, the relationship between the number of data points and the margin of error is not linear but rather follows a curve. This is because the margin of error is inversely proportional to the square root of the number of data points. As summarized by Hunter [2025], this means that there is a point of diminishing returns, where the margin of error decreases only slightly with each additional data point. In our case, when a source rarely occurs during template additions or template removals, the confidence in the trustworthiness estimate of that source increases significantly with each additional occurrence. However, after we have reached a certain confidence in a source's trustworthiness estimate, the confidence further increases only slightly with each additional occurrence. To represent this relationship in our model, we use a function that starts steep and flattens out as the number of occurrences increases. Specifically, we use the function shown in Figure 5.1. It can be seen that the curve limits the weight to a maximum of 100, meaning that all weights are normalized to the range [0, 100]. A weight of 0 indicates that we have no confidence in the trustworthiness estimate of the source, while a weight of 100 indicates that we are very confident in the trustworthiness estimate of the source. Note that although both the weight limit and the steepness of the function are tunable parameters, we use fixed values due to time constraints.

Using the weights, we calculate the final probabilities of a template addition and a template removal for the whole revision by taking the weighted average of the probabilities of the referenced external sources. To do this, we first normalize the weights by dividing each weight by the total weight sum. We then multiply each probability by the corresponding normalized weight and sum up the results to calculate the probability of a template addition and the probability of a template removal. Note that the two probabilities are complementary, as they sum up to 1.0.

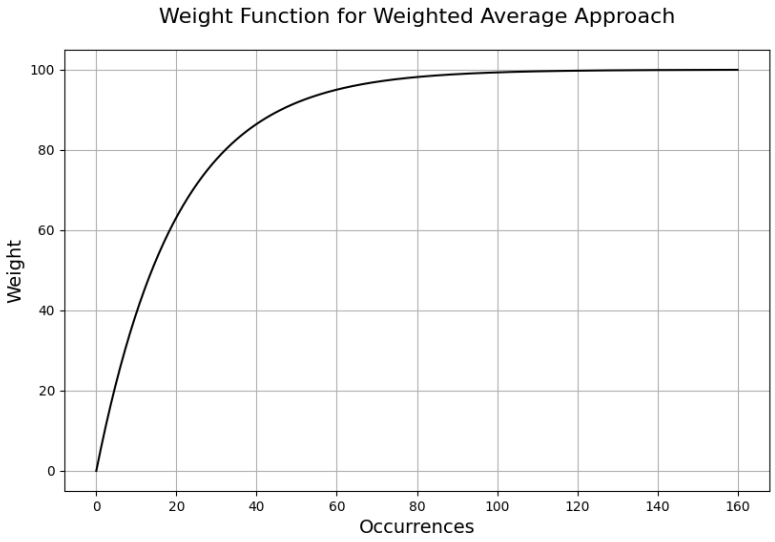

Abbildung 5.1: Plot of the weight function: $weight = 100 - 100 * e^{-0.05 * occurrences}$

Our test set contains only revisions that can be assigned to one of two classes. The first class consists of revisions where a reliability template was added, while the second class contains revisions where a reliability template was removed. The distinction between the two classes is signaled by the label T_{Added} which is set to 1.0 if a template was added and 0.0 if not, and the label $T_{Removed}$ which is set to 1.0 if a template was removed and 0.0 if not. To predict if a template was added in a revision, we compare the computed probability of a template addition with a threshold. If the probability is greater than the threshold, we predict the T_{Added} label to be 1.0, otherwise we predict it to be 0.0. The same procedure is applied for the $T_{Removed}$ label, where we predict if a template was removed in a revision. Lastly, we evaluate the effectiveness of our predictions by comparing the predicted labels with the actual labels.

5.2 Results

To present the effectiveness of our model, we use *Receiver Operating Characteristic* (ROC) curves. A ROC curve is a graphical representation of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The area under the curve (AUC) is a measure of how well the model can distinguish between classes. An AUC of 1.0 indicates a perfect model, while an AUC of 0.5 indicates a model that is no better than random guessing. In the graphs we present, the random classifier is represented by the dotted diagonal line. The closer the ROC curve is to the upper left corner, the better

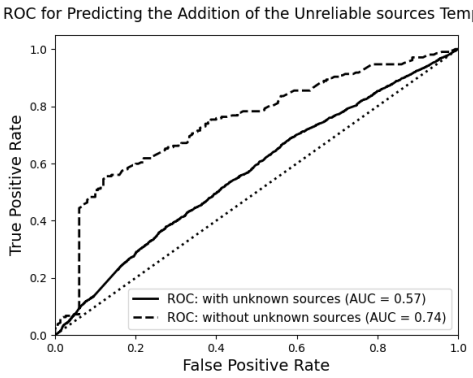

Abbildung 5.2: ROC curves for the prediction of an *Unreliable sources* template addition

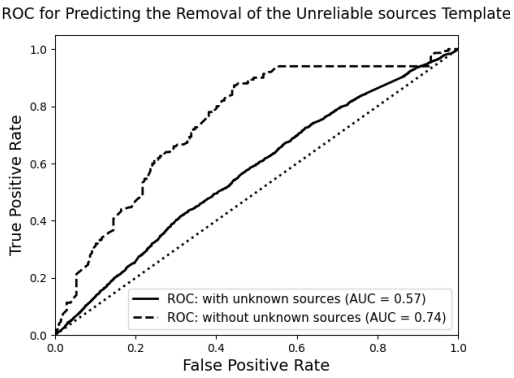

Abbildung 5.3: ROC curves for the prediction of an *Unreliable sources* template removal

the model can distinguish between the two classes.

5.2.1 Template: Unreliable Sources

Figure 5.2 shows the ROC curves for the prediction of whether an *Unreliable sources* template was added in a revision. The solid curve describes our standard model, where unknown sources are assigned default probabilities of 0.5. This approach performs only marginally better than random guessing (diagonal dotted line), which is also reflected in the AUC of 0.57. We can see that for most thresholds, the number of correctly identified revisions where a template was added is only slightly higher than the number of falsely identified revisions where a template was not added.

Figure 5.3 shows the ROC curves for the prediction of whether an *Unreliable sources* template was removed in a revision. Here too, the solid curve describes our standard model, where unknown sources are assigned default probabilities of 0.5. Note that the solid curve is flipped along the diagonal, as we are now predicting the removal of a template instead of the addition. This is expected, as our test set only contains two classes where for each revision, the binary labels T_{Added} and $T_{Removed}$ are always the opposite of each other. Furthermore, the probabilities of a template addition and a template removal are complementary, as they sum up to 1.0. Consequently, we can observe that the number of correctly identified revisions where a template was removed is only slightly higher than the number of falsely identified revisions where a template was not removed.

One possible reason for the poor performance of our model is that the majority of revisions in our test set have unknown sources. This means that

the model has to rely on default probabilities for a large amount of the referenced sources because we have no computed trustworthiness estimate for these sources based on the training data. To examine this further, we conduct a second experiment: We predict labels only for a subset of the test set, where we have computed a trustworthiness estimate for all referenced external sources. The ROC curves for this subset are also shown in Figure 5.2 and Figure 5.3 as dashed lines. We observe that the model performs significantly better when all sources have a trustworthiness estimate: the curve reaching higher into the upper left corner in Figure 5.2 indicates that the model correctly identified more revisions where a template was added as in the first experiment. In Figure 5.3, the curve reaching higher into the upper left corner indicates that the model correctly identified more revisions where a template was removed as in the first experiment. This is also reflected in the AUC of 0.74, which is a significant improvement over the AUC of 0.57 when using default probabilities for unknown sources. However, we should mention that of the original 4,970 revisions in the test set, only 357 revisions were used in this experiment, as they were the only ones where we had computed trustworthiness estimates for all sources.

5.2.2 Template: Dubious

Figure 5.4 shows the ROC curves for the prediction of whether a *Dubious* template was added in a revision. Figure 5.5 shows the ROC curves for the prediction of whether a *Dubious* template was removed in a revision. Similarly to the *Unreliable sources* template, the ROC curve of the standard model for the removal of the *Dubious* template is flipped along the diagonal when compared to the ROC for the addition of the *Dubious* template. The AUC in both cases of the standard model is 0.52, which is about as good as random guessing. This indicates that our standard model cannot distinguish between revisions where a *Dubious* template was added or removed.

As with the *Unreliable sources* template, we also conduct a second experiment where we predict labels only for a subset of the test set, where we have computed a trustworthiness estimate for all referenced external sources. This time, only 702 of the original 17,354 revisions in the test set remained. The ROC curves for this subset are shown in Figure 5.4 and Figure 5.5 respectively, and are portrayed as dotted lines. We observe that for most thresholds, the model performs significantly better when all sources have a trustworthiness estimate. However, for some thresholds, the model performs worse than a random classifier. We believe that this is due to the small number of revisions used in the test set of this experiment, which makes the results less reliable and more prone to random fluctuations and outliers. The AUC of 0.60 indicates

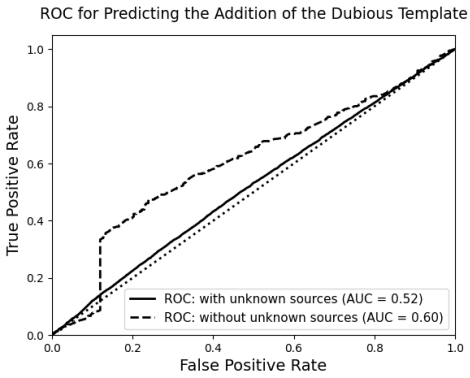

Abbildung 5.4: ROC curves for the prediction of a *Dubious* template addition

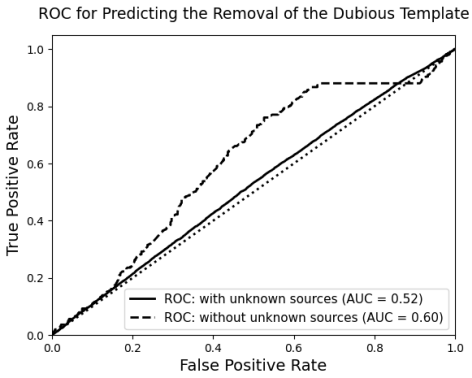

Abbildung 5.5: ROC curves for the prediction of a *Dubious* template removal

that the model has a slight advantage over random guessing, but is still not able to reliably distinguish between revisions where a *Dubious* template was added or removed.

Kapitel 6

Discussion

In this chapter, we discuss the results of our experiments and analyze the limitations of our approach. We explore potential solutions to the identified issues and discuss possible real-world applications of our model.

6.1 Error Analysis

Error analysis is crucial, particularly when results are not as expected, as it helps to identify the limitations of the current approach and guides future improvements. This section explores the factors contributing to the shortcomings of our approach and discusses potential solutions.

6.1.1 Model Constraints

In our current model, we only consider the template-added and template-removed states of a Wikipedia article. This means that to estimate the trustworthiness of external sources, our model solely relies on the data that is retrieved when a template is added and later removed. Therefore, any additions or removals of referenced sources before a template was added and after a template was removed are ignored. While this allows for a much simpler approach, it brings significant limitations. To analyze the impact of this constraint, we will consider the scenarios shown in Figure 6.1. It models a timeline of a Wikipedia article with the event of a template addition $T_{Addition}$ and the event $T_{Removal}$ of the removal of that template. On the timeline there are three sources S_4, S_5, and S_6 which are referenced in the article at different times. S_4 is referenced before the template is added, S_5 is referenced while the template is present and S_6 is referenced after the template is removed. Note that none of these sources are being referenced at the time of the template's addition or removal, which is why they are not included in our source extraction process and are

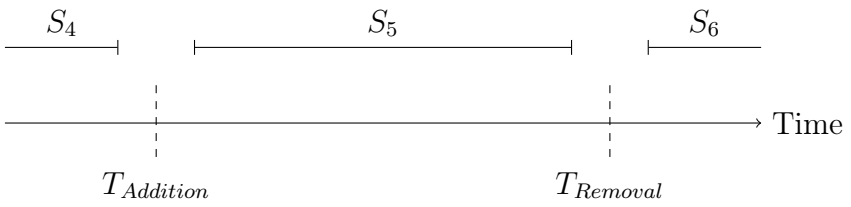

Abbildung 6.1: Timeline illustrating the time spans for sources S_4, S_5, S_6 relative to the events $T_{Addition}$ and $T_{Removal}$.

not considered in our model. While we assume that it is safe to ignore S_4 as there is no indication that it was the reason for the template addition, S_5 and S_6 might contain valuable information that we miss out on. S_5 describes any source, which is referenced while the template was present but removed before the template was removed. This could be a source that was tested for reliability but was found not reliable enough to keep in the article. One could argue that this source should be considered when estimating the trustworthiness of external sources, as it could be marked to be of questionable reliability. An even more critical scenario however is the addition of a source after $T_{Removal}$, such as S_6. An editor might remove complete sections of a Wikipedia article, prioritizing displaying less information over the risk of spreading unreliable or misleading content. In this case, the template is removed and the article is later rebuilt using more reliable sources, which we do not capture in our approach. Although it is uncertain how often this scenario occurs, we are missing out on valuable data that could improve the trustworthiness estimates of external sources.

To solve the mentioned limitations, one would need to consider a model that takes more than just the template-added and template-removed states into account. This would include scanning revisions in between the template addition and the template removal, as well as scanning consequent revisions after the template was removed. Additional data could be retrieved, such as external sources that were tested while the template was present or possibly reliable sources that were added immediately after the template was removed to rebuild the article. This approach could pave the way for a more comprehensive trustworthiness estimation of external sources.

6.1.2 Complexities in Template Identification

Another factor contributing to the model's limitations is the complexity of identifying templates. One reason why we only consider pairs of revisions where a reliability template was added and later removed is that we do not know which parts of a Wikipedia article are challenged by a reliability template.

When we filter for the presence of reliability templates, we assume that the templates refer to the entire article and we therefore extract all referenced external sources from the article. By then comparing the referenced sources of the article when the template was added and when the template was removed, we identify only the external sources that were changed, with the assumption that these sources are generally related to the template's addition or removal. In reality, however, this is not the case. The *Unreliable sources* template for example is used to mark entire articles, but the template directly states that only *some* of the referenced sources in the article might be unreliable. The *Dubious* template on the other hand is not used to mark an entire article as unreliable, but rather to highlight specific sections that need improvement. In both cases, any references that are added to or removed from sections that are not questioned by the template are falsely brought into association with the template. Specifically, references that are removed from sections unrelated to the template before the template's removal are falsely marked as unreliable, while references added to sections unrelated to the template after the template's addition are falsely marked as reliable. These issues are not only present when considering the *Unreliable sources* and *Dubious* templates, but also for many other reliability templates. Our original assumption was that this noise would be negligible when using a large enough dataset, but we have found that in reality, our data is very susceptible to noise, as shown in Section 6.1.3.

Identifying the exact parts of an article that are challenged by a reliability template is not a trivial task. This is because templates are not always used in a consistent manner, which makes it difficult to accurately identify the context of a template using automatic methods. The *Dubious* template for example uses a `reason` parameter, such as `{{Dubious|reason=What the problem is}}` to specify the reason why the information is considered dubious.[1] This allows editors to mark specific parts of the article that are considered dubious, rather than the entire article. For our automatic template filter mechanism, this is an issue, as it is a complex task to understand the context of the `reason` parameter and to identify the references connected to it. Furthermore, for some templates, the context is not necessarily specified as a comment, but rather through another version of the template which focuses on a specific section. For example, instead of the *Unreliable sources* template which refers to entire articles, the *Unreliable sources section* template marks specific sections that contain possibly unreliable sources.[2] In our experiments filtering for section-specific templates, we encountered difficulties in accurately extracting the referenced external sources related to the template. We assume this is because section-specific templates and headlines are not always used consistently. Even slight

[1] `https://en.wikipedia.org/wiki/Template:Dubious`

[2] `https://en.wikipedia.org/wiki/Template:Unreliable_sources_section`

variations in formatting or placement can strongly increase the complexity of our pattern-matching process.

To improve the effectiveness of our model, one would need to develop a more sophisticated approach to identify templates in revisions. This could involve a more complex template filter mechanism that can identify the context of a template, such as the `reason` parameter of the *Dubious* template. Additionally, the filter mechanism would need to identify section-specific templates. Using these templates, one could also consider revisions where a template was added but not yet removed, to identify specific references that the template was added for, instead of considering all references of an article. Furthermore, we assume that it would be possible to reduce noise in the extracted referenced sources by only considering the references marked by the template. We think that this could allow for extracting more accurate information on referenced sources and their association with reliability templates, which could improve the trustworthiness estimates of external sources.

6.1.3 Data Deficiencies

The data used for training and evaluating our model is a crucial factor that influences the model's effectiveness. In our datasets, we have identified significant deficiencies that help explain the issues with our model's ability to accurately predict whether a template was added or removed in a revision.

One of the main deficiencies is that despite having estimated the trustworthiness of thousands of external domains and ISBNs, the majority of revisions in the test set contain references to unknown sources, by which we refer to sources that did not occur in the training data and consequently have no computed trustworthiness estimate. Specifically, around 87.75% of the revisions in the *Unreliable sources* test set contain a reference to at least one unknown source. Among these 87.75% of revisions, on average 49% of the referenced external sources are unknown. For the *Dubious* test set, 92.28% of the revisions reference at least one unknown source, with on average 44% of the referenced external sources being unknown. Essentially, this means that a large portion of the data the model relies on for predictions are simply default probabilities of 0.5 for unknown sources being associated with a template addition or a template removal. We assume that because so much of the data lacks predictive value, the model's predictions are largely random, resulting in performance only marginally better than a random classifier. We have outlined the impact of this limitation when evaluating the model's effectiveness in Section 5.2.

The second important deficiency is that the trustworthiness estimate for the majority of sources is based on very few data points. When storing the trustworthiness estimates of external sources, we also store the number of

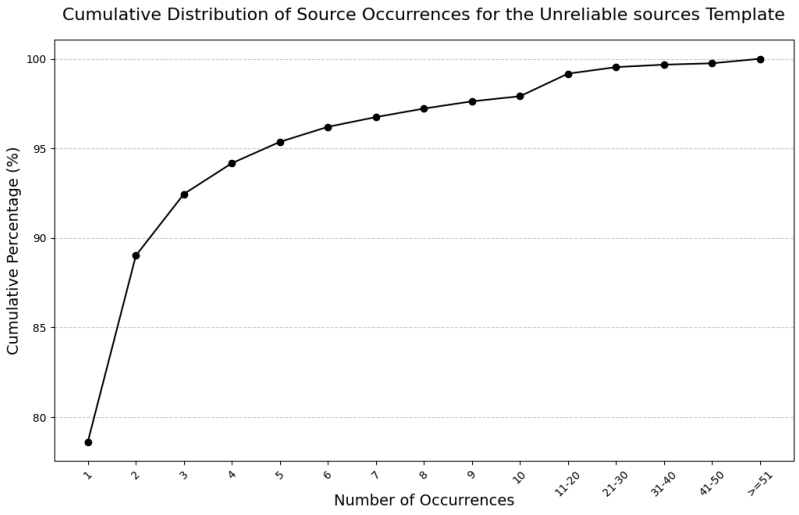

Abbildung 6.2: Cumulative distribution function of occurrences (number of revision pairs used to compute the trustworthiness estimate) of external sources for the *Unreliable sources* template.

revision pairs that were used to compute the trustworthiness estimate for each source, also referred to as the *source occurences*. In Figure 6.2 we plot the cumulative distribution function of the occurrences of the external sources that were extracted from revision pairs where the *Unreliable sources* template was added and later removed. We can see that almost 80% of the external sources have a trustworthiness estimate based on one revision pair and around 92.5% of the referenced external sources have a trustworthiness estimate based on a maximum of 3 revision pairs. This is a significant concern because it indicates that the trustworthiness estimates of the majority of the sources are based on very little data. Noise which is introduced by issues with template identification, vandalism, or simply wrongful template usage, can therefore have a significant impact on the trustworthiness estimates of most sources. We assume that this is a major factor contributing to the model's inability to accurately predict whether a template was added or removed in a revision, even when we only made predictions for revisions containing known sources.

Overall, we conclude that the majority of the trustworthiness estimates for the external sources are based on too little data to be reliable. Furthermore, the high number of unknown sources in the test set exaggerates this issue, as the model cannot make accurate predictions using these sources. To combat these issues, one would need a more extensive dataset to increase the number of known sources and the number of revisions used to compute the trustworthiness estimates of the external sources. This could be done using a more sophisticated

model that can combine the data from multiple reliability templates to extract a higher number of revision pairs. Alternatively, one could consider a model that can compute trustworthiness estimates of external sources with improved accuracy for a single template. This could be achieved by reducing noise due to the usage of section-specific templates, as discussed in Section 6.1.2.

6.2 Possible Real World Applications

We believe that our model has the potential to be used in real-world applications. Specifically the trustworthiness estimation of external sources using the *Unreliable sources* template showed promising results (see Section 4.2). Combined with the computationally efficient nature of our model, we believe that it could be used in a variety of applications. The only computationally heavy workload is parsing Wikipedia data and filtering revisions containing specific templates. While this task requires significant storage and parallelization capabilities, it suffices to do this occasionally. It is unnecessary to parse the entire Wikipedia database each time a trustworthiness estimate for an external source is needed. We assume that this task could be done on a monthly or even yearly basis, to ensure that a significant number of new revisions containing reliability templates can be extracted. Once all revision pairs for a reliability template have been extracted, all subsequent steps are principally lightweight and can be performed on standard hardware. Additionally, apart from the original Wikipedia dump, our model is very memory-friendly. After the trustworthiness estimation process for external sources is complete, the resulting CSV file containing information on external sources is only a few megabytes in size.

This opens up many possibilities for real-world applications. To provide a specific example, our model could be used inside a browser extension that scans any open Wikipedia article for its referenced external sources and looks up their previously computed trustworthiness estimates. The browser extension could then highlight sections of the text that are supported by sources that were strongly associated with additions or removals of the *Unreliable sources* template in the past. This would allow users to quickly assess whether they can rely on the information provided in the article or whether they should be cautious.

While the idea of such a browser extension is appealing, it is important to note that the model's performance is not yet sufficient for such an application. As discussed in Section 6.1.3, the trustworthiness estimates of most external sources are based on too little data to be reliable. If the model were to be used in a real-world application, it would be crucial to reliably estimate the

trustworthiness of a wide range of external sources. Furthermore, identifying the exact parts of an article that are supported by an external source is not a trivial task, as discussed in Section 6.1.2. If the browser extension were to highlight sections of the text that are supported by particularly reliable or particularly unreliable external sources, it would need to accurately identify the statements backed by these sources.

6.3 Conclusion and Future Work

In our approach, we first created a dataset of Wikipedia revision pairs containing revisions where a reliability template was added and the first subsequent revision where that template was removed. We then extracted the referenced external sources from these revisions and computed trustworthiness estimates for each source. This was done by computing probabilities of the sources being associated with the addition or the removal of a reliability template. While we could manually verify the trustworthiness estimates of a small subset of sources for the *Unreliable sources* template, the manual evaluation of the same sources for the *Dubious* template required speculative reasoning, highlighting that the trustworthiness estimates for templates with a wider range of use cases are more difficult to interpret manually. Future work could involve the analysis of the trustworthiness estimates of external sources for a wider range of reliability templates, to better understand the potential of our model.

Using the computed trustworthiness estimates of the external sources, we then predicted whether a reliability template was added or removed for a selected test set of revisions. We found that the model's performance was only marginally better than a random classifier, which we mostly attribute to the high number of unknown sources in the test set and the low number of data points used to compute the trustworthiness estimates of the majority of the external sources. Future work could involve the development of a more complex model, combining the data from multiple reliability templates to increase the number of known sources and data points for each source. Alternatively, we propose that to compute more accurate trustworthiness estimates of external sources for a single template, one could try to reduce noise in the model by using section-specific templates to identify the references challenged by the reliability template more accurately.

Literaturverzeichnis

B. Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. Assigning trust to wikipedia content. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581286. doi: 10.1145/1822258.1822293. URL https://doi.org/10.1145/1822258.1822293.

Samy A. Azer. Evaluation of gastroenterology and hepatology articles on wikipedia: Are they suitable as learning resources for medical students? *European Journal of Gastroenterology & Hepatology*, 26(2):155–163, February 2014. doi: 10.1097/MEG.0000000000000003.

Pamela Hunter. Margin of error and confidence levels made simple. *Retrieved January 21st*, 2025.

Jona Kräenbring, Tika Monzon Penza, Joanna Gutmann, Susanne Muehlich, Oliver Zolk, Leszek Wojnowski, Renke Maas, Stefan Engelhardt, and Antonio Sarikas. Accuracy and completeness of drug information in wikipedia: a comparison with standard textbooks of pharmacology. *PLOS ONE*, 9(9):e106930, 2014. doi: 10.1371/journal.pone.0106930. URL https://doi.org/10.1371/journal.pone.0106930. Epub 2014 Sep 24.

Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. Why the world reads wikipedia: Beyond english speakers, 2018. URL https://arxiv.org/abs/1812.00474.

Sai T. Moturu and Huan Liu. Evaluating the trustworthiness of wikipedia articles through quality and credibility. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605587301. doi: 10.1145/1641309.1641349. URL https://doi.org/10.1145/1641309.1641349.

Jennifer Phillips, Connie Lam, and Lisa Palmisano. Analysis of the accuracy and readability of herbal supplement information on wikipedia. *Journal of the American Pharmacists Association*, 54(4):406–414, July–August 2014. doi: 10.1331/JAPhA.2014.13181.

Yu Suzuki and Masatoshi Yoshikawa. Mutual evaluation of editors and texts for assessing quality of wikipedia articles. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450316057. doi: 10.1145/2462932.2462956. URL `https://doi.org/10.1145/2462932.2462956`.

Lea Viljanen. Towards an ontology of trust. In Sokratis Katsikas, Javier López, and Günther Pernul, editors, *Trust, Privacy, and Security in Digital Business*, pages 175–184, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31796-8.

Wikipedia contributors. Wikipedia:reliable sources/perennial sources, 2025. URL `https://en.wikipedia.org/w/index.php?title=Wikipedia:Reliable_sources/Perennial_sources&oldid=1275862533`. [Online; accessed 19-February-2025].

KayYen Wong, Miriam Redi, and Diego Saez-Trumper. Wiki-reliability: A large scale dataset for content reliability on wikipedia. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2437–2442. ACM, July 2021. doi: 10.1145/3404835.3463253. URL `http://dx.doi.org/10.1145/3404835.3463253`.

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Jena, 21. Februar 2025

. . . . . .
Luca-Philipp Grumbach