UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Software Engineering Curriculum

Kasper Kaljuste

CircularCheck: A Tool for Detecting Circular Reporting

Master's Thesis (30 ECTS)

Supervisor(s): Uku Kangur, MSc

Tartu 2025

CircularCheck: A Tool for Detecting Circular Reporting

Abstract:
In the modern information landscape, the speed with which news is spread has reached unprecedented levels. This poses significant challenges in ensuring the accuracy and independence of information. Circular reporting is a situation where a piece of information appears to come from multiple independent sources, but in reality comes from only one source. Such practices can be intentional or accidental and contribute to the spread of false information by creating an illusion of corroboration. While circular reporting has been studied in intelligence and scientific literature, its detection in journalism, particularly in a small media ecosystem like Estonia, has received little attention. This thesis addresses the problem of detecting circular reporting in Estonian online news media. We present a system that detects circular reporting by building reference hierarchies and comparing article content across ERR, Delfi, and Postimees. Here we show that using a combination of link-based and text-based methods, it is possible to flag suspicious reference patterns for manual validation. The results show that 47 positive cases were detected by link analysis and 4 by text similarity. Self-referencing structures were the most reliable. These results reveal that although circular reporting is not widespread, it does occur and can be identified with relatively simple heuristics. The system does not attempt to verify the truthfulness of the information but instead focuses on tracing the propagation of references. This allows researchers and journalists to better assess the credibility and independence of sources. In a broader context, the results offer a framework that can be adapted to other media ecosystems and help improve media transparency.

Keywords: Circular Reporting, Estonian News Media, Media Transparency, Text Similarity, Reference Hierarchy, Information Propagation

CERCS: P170 Computer science, numerical analysis, systems, control

CircularCheck: Tööriist ringviitamise tuvastamiseks

Lühikokkuvõte:

Tänapäevases infokeskkonnas levib teave kiiremini kui kunagi varem. See tekitab raskusi allikate usaldusväärsuse ja sõltumatuse tagamisel. Ringviitamine on olukord, kus info näib pärinevat mitmest sõltumatust allikast, kuigi tegelikult pärineb see ainult ühest. Selline olukord võib tekkida tahtlikult või kogemata ning aitab kaasa valeinfo levikule, luues näilise kooskõlastatuse mulje. Kuigi ringviitamist on uuritud luure, Vikipeedia ja teadusartiklite kontekstis, on selle tuvastamine ajakirjanduses, eriti väikestes meediaruumides nagu Eesti, jäänud tähelepanuta. Käesolev magistritöö tegeleb ringviitamise tuvastamise probleemiga Eesti veebimeedias. Töös esitletud süsteem tuvastab ringviitamist, ehitades artiklite viitehierarhiaid ja võrreldes artiklite sisu ERR-i, Delfi ja Postimehe andmetel. Näitame, et linkide ja tekstipõhiste meetodite kombinatsioon võimaldab tuvastada kahtlasi viiteid, mis vajavad käsitsi valideerimist. Tulemused näitavad, et lingianalüüs tuvastas 47 positiivset juhtu ja tekstisarnasuse meetod 4. Kõige täpsemalt tuvastas süsteem eneseviitamise struktuure. Need tulemused näitavad, et kuigi ringviitamine ei ole väga levinud, esineb seda siiski ning seda on võimalik tuvastada lihtsate heuristikatega. Süsteem ei püüa hinnata info tõesust, vaid keskendub selle leviku jälgimisele. See võimaldab ajakirjanikel ja uurijatel paremini hinnata allikate sõltumatust ja usaldusväärsust. Laiemas kontekstis pakuvad tulemused raamistikku, mida saab kohandada ka teiste meediakeskkondade jaoks, et suurendada meedia läbipaistvust.

Võtmesõnad: Ringviitamine, Eesti uudismeedia, Meedia läbipaistvus, Tekstisarnasus, Viidete hierarhia, Informatsiooni levimine

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Contents

1 Introduction

In the modern information landscape, the speed with which news is spread has reached unprecedented levels. This poses significant challenges in ensuring the accuracy on independence of information.

One of the most popular mechanisms for spreading online disinformation is circular reporting. Circular reporting is a situation where a piece of information appears to come from multiple independent sources, but in reality comes from only one source [1]. Circular reporting can happen in various forms:

- **Citation Loop:** When one article cites other articles, but those articles all refer to the same source.
- **Self-referential Loop:** When an article cites itself or references another article that ultimately links back to the same source.

Such practices, intentional or accidental, contribute to the spread of fake news. Fake news is false information published by a news outlet [2]. Regardless of intent, the result is the same: spreading false information. It erodes trust in media and influences public opinion. While informed individuals may identify and disregard false information, the more considerable concern is the uninformed majority, who might accept the amplified false narrative as truth. Detection and mitigation of citation loops aim to reduce this group's exposure to disinformation. Addressing these loops is critical for restoring trust in media and protecting public discourse from manipulation.

It is important to note that detecting citation loops does not inherently involve identifying misinformation. The detection process focuses on tracing information propagation rather than verifying its truthfulness. However, citation loops can contribute to the spread of fake news in three key ways:

- **Propagating unverified information**: If the original source is unreliable or contains falsehoods, the loop amplifies misinformation across multiple outlets. Studies on hierarchical propagation networks have shown how the structure and patterns of information spread can exacerbate the reach of misinformation, highlighting the need for mechanisms that trace these loops [3].

- **Creating an illusion of independent corroboration**: Even when the original information is factual, the presence of multiple references pointing to the same source can mislead audiences into perceiving a broader consensus than actually exists.
- **Self-citation loops**: Self-citation practices within journalistic or academic work can contribute to circular reporting. By referencing one's prior work, a closed loop of citations is created, which may inflate credibility while failing to provide independent verification [4].

By addressing these issues, the tool (see Appendix I. CircularCheck Tool Repository) developed in this thesis aims to increase transparency in media reporting and enable journalists and researchers to evaluate cited information's credibility better. The methodology is designed to achieve the following goals:

- **Detect Citation Loops**: Accurately identify instances of circular reporting in Estonian news media.
- **Analyze Patterns**: Uncover broader patterns of information propagation and circularity.
- **Facilitate Further Research**: Provide a reproducible framework for detecting circular reporting in other contexts.

The tool acts as a human-in-the-loop system where identified citation loops can be reviewed to analyse their significance. While the tool cannot independently verify the accuracy of the information, it highlights potential instances of circular reporting for manual evaluation.

The remainder of this thesis is structured as follows. Section 2 provides background on citation loops and their relevance. Section 3 reviews related work in online news media references, fake news detection and citation patterns. Section 4 outlines the methodology, including data collection, reference analysis, and similarity detection. Section 5 presents the results and includes a detailed discussion and limitations. Section 6 concludes the thesis and summarises key findings.

2 Background

Circular reporting is not just a problem in journalism. It also shows up in other areas like Wikipedia and intelligence work. In all of these cases, the issue is the same: the same information gets repeated through different sources, making it look like multiple people have confirmed it, even though it all comes from one place.

Sometimes this happens by accident, other times on purpose. In fast-paced news reporting, it can happen because journalists rely on each other without checking the original source. In intelligence, it can be used as a tactic to make false information seem credible. On Wikipedia, the problem comes up when someone adds a claim, a news site copies it, and then someone cites that same news site back on Wikipedia.

The next sections show how circular reporting appears in Wikipedia, intelligence, and journalism.

2.1 Wikipedia

Wikipedia, one of the most frequently referenced online knowledge repositories, often serves as both a source and a recipient of information in the media ecosystem. The term "citogenesis" describes a feedback loop where misinformation added to Wikipedia is cited by external sources, which are then used to validate the original Wikipedia entry [5]. This phenomenon exemplifies how circular reporting can perpetuate inaccuracies, even in trusted platforms.

Citogenesis can occur in a wide range of contexts. For example, an unverified claim about a historical event might be added to Wikipedia, subsequently referenced by news articles or academic works, and then cited back in the same Wikipedia entry as a credible source. This cycle can mislead readers and researchers alike, demonstrating the ripple effects of poorly sourced information. Examples of citogenesis have been observed in scientific discoveries, historical narratives, and cultural phenomena [6].

2.2 Intelligence

Circular reporting is also prevalent in intelligence gathering and disinformation campaigns. Adversaries often exploit media ecosystems to fabricate the appearance of

consensus, leveraging the same information through multiple outlets to sway public opinion or influence policy decisions.

An example is Operation INFEKTION, a Cold War disinformation campaign by the Soviet KGB [7]. This campaign aimed to spread the false claim that the HIV/AIDS virus was created by the United States as part of a biological weapons program. The disinformation was strategically introduced in a small pro-Soviet newspaper in India, then picked up by other outlets around the world, including those in Western countries. The false narrative gained credibility as it circulated, eventually being cited by multiple sources that appeared independent but ultimately traced back to the same origin. This example highlights how circular reporting can be weaponized to manipulate public perception and create geopolitical discord.

2.3 Journalism

In journalism, the problem of circular reporting is exacerbated by the pressure to deliver news quickly. Major outlets like Estonia's ERR, Delfi, and Postimees are not immune to this phenomenon. The tight interlinkage between these outlets, combined with the urgency of modern reporting, creates fertile ground for citation loops. Circular reporting here not only diminishes transparency but also risks spreading unchecked disinformation across large audiences.

3 Related work

This section provides an overview of existing research relevant to overall journalistic standards, the role of hyperlinks in online news media, detecting citation loops in research publications and the detection of fake news. While these studies provide context, the focus of this thesis aims instead to address the specific problem of detecting circular reporting in Estonian media platforms.

3.1 Journalistic Standards

Journalism is often described as a profession with a set of shared values that guide how news should be gathered, written, and published. Deuze (2005) outlines five such values that make up journalism's professional ideology: public service, objectivity, autonomy, immediacy, and ethics [8]. These are not always followed strictly in practice, but they still provide a useful framework for understanding what is expected from journalists, and how real-world pressures might push them away from those ideals.

A later study by Henkel et al. (2020) applied four of these five values (excluding immediacy) to compare journalists working in online, offline, and mixed settings across Europe [9]. They found that journalists across these platforms generally reported similar professional principles, but with some differences. Online journalists, especially those working only for digital-native outlets, were more likely to justify publishing unverified information and showed less interest in the traditional "watchdog" role. Interestingly, this trend was reversed in several Eastern European countries, where online journalists placed more importance on accountability and public service, possibly as a response to political pressure in legacy media.

The push toward faster news cycles and more competitive environments also plays a role in weakening journalistic standards. Vasterman (2005) introduced the concept of "media hype" to describe how certain stories grow into self-reinforcing news waves without much new information being added [10]. These waves often lead to repetitive coverage, less fact-checking, and more emphasis on emotional or sensational angles. Media hype doesn't just make stories louder but it also contributes to uniformity in reporting and makes it easier for unverified claims to spread quickly.

All of this ties directly into the problem of circular reporting. When journalistic

standards slip, especially around ethics, immediacy, or public service, it becomes easier for news outlets to repeat each other's content without verifying the original source. This is one of the main ways circular reporting can happen in online media. Understanding where and how these standards break down helps clarify why circular reporting is difficult to detect, and why it matters.

3.2 Hyperlink Usage in Online News

De Maeyer (2014) analyzed how hyperlinks are used to display sources in online news articles [11]. The study showed that most hyperlinks do not point to original sources but instead link to internal pages, homepages, or unrelated material. Only a small portion of links directly support the claims being made. The paper also proposes a classification of hyperlink functions and notes variation across different newsrooms.

Spitz and Gertz (2015) extracted a sparse citation network from hyperlinks between online news articles [12]. Their results showed that most articles contain few links, but a smaller subset forms interconnected clusters. These structures resemble citation patterns in scientific literature and reflect how stories are reused and propagated across outlets. The authors suggest that methods from scientometrics can be applied to study such citation patterns in online news.

Cui and Liu (2017) conducted a content analysis of three types of online news media: legacy, explanatory, and citizen journalism [13]. They categorised link usage into sourcing (used to support claims), contextualising (to provide background), and interpreting (to add framing). Contextualising was the most frequent. Legacy media used more sourcing links, while interpretive links were more common in citizen journalism.

Alsuliman et al. (2022) proposed a methodology that utilises URLS to collect news articles from trusted sources and compare them with posts from social media platforms [14]. Web scraping techniques gather links to news articles, which are matched with social media posts using cosine similarity, TF-IDF, and word appearance methods. The study highlights that URLs serve as the backbone for linking social media content to verified news sources, enabling a structured way to trace and evaluate the credibility of claims. Cosine similarity was identified as the most effective matching technique, achieving 80% accuracy in retrieving relevant articles.

3.3 Citation Loops in Research Publications

Fan et al. (2022) used a workflow with human input and machine learning to detect informal references in research publications [15]. Named Entity Recognition (NER) and iterative feedback were employed to identify non-standard citation patterns, combining automated and manual methods for accuracy. The approach improved the ability to detect informal data mentions, creating more complete bibliographies of data-related literature and providing a basis for tracking data reuse trends and scholarly communication networks.

Bu et al. (2020) examined loops in academic citation networks, focusing on physics and computer science [16]. Strongly Connected Components (SCCs) were used to detect and measure loops, with findings showing that most loops result from self-citations. The study quantified how these loops influence the structure of citation networks, such as identifying recurring citation relationships and how they can affect metrics like co-citation frequencies and bibliographic coupling strength.

3.4 Fake News Detection

Thota et al. (2018) used a deep learning approach for fake news detection, focusing on analyzing the stance between headlines and articles [17]. The authors used neural networks, including dense architectures and pre-trained embeddings, to classify stances such as 'agree,' 'disagree,' 'discuss,' or 'unrelated,' achieving 94.21% accuracy on the Fake News Challenge dataset.

Another method utilizes network-based patterns in social media. Zhou and Zafarani (2019) demonstrated that fake news spreads farther, involves more spreaders, and forms denser networks than true news [18]. These patterns, observed at node, triad, and community levels, offer interpretable features for detection.

4 Methodology

This chapter outlines the methodology for identifying circular reporting in major Estonian online news outlets. The methodology is divided into three main components: data gathering, URL matching, and text matching.

Data gathering involves collecting a set of articles to identify potential instances of circular reporting. URL matching refers to the process of retrieving embedded references from these articles, constructing a reference hierarchy, and detecting any duplicate references. Text matching involves retrieving the content of the articles and comparing the text similarity of the referenced articles. Both methods aim to detect possible cases of circular reporting. All positive detections undergo manual validation to confirm their accuracy.

4.1 Data Gathering

A dataset is collected from three major news outlets in Estonia: Delfi, Postimees, and ERR. These outlets are selected because they are among the largest and most widely read in the country. In our approach, we consider four distinct methods for collecting article URLs. Each method leverages different aspects of website design and functionality.

Using Sitemaps Sitemaps are XML files provided by websites to list their URLs for search engines. They are directly accessible and updated automatically, and they mainly include recent articles.

Using Search Functionality The search functionality allows articles to be filtered by criteria such as date, category, and keywords. By querying the search endpoints, it is possible to retrieve a set of article URLs—including those from historical archives.

Hardcoded URL Requests Hardcoded URL requests generate article URLs based on a predetermined pattern (e.g., incorporating an article ID into a fixed URL format). In practice, many of the generated URLs do not lead to valid pages due to the unpredictable nature of the article IDs.

Using Sections Some websites provide categorized lists of articles through sections, which often include basic metadata such as titles and publication dates. This structured approach enables the retrieval of article URLs by navigating these pre-organized sections.

4.1.1 Criteria

The evaluation is based on the following:

- **Historical Coverage:** Assesses the time over which articles can be retrieved.
- **Complexity:** Measures the efficiency of the URL collection process. If each successful request returns k articles, then ideally, the number of requests needed to collect n articles is $\frac{n}{k}$. In practice, however, some requests may fail to return any articles; let x denote the number of these unsuccessful requests. Thus, a higher x increases the total number of requests required, indicating a slower method.
- **Number of URLs:** Evaluates the total number of article URLs that the method can collect. A higher number suggests more comprehensive coverage.
- **Ease of Access to Metadata:** Evaluates how easily supplementary information, such as publication dates, titles, and other metadata, can be accessed alongside article URLs. Methods that provide well-structured and readily available metadata are preferred. Although this thesis utilizes only the article content and embedded references, the availability of additional metadata could facilitate more extensive future research.
- **`robots.txt` Restrictions:** Considers the impact of the site's `robots.txt` file on the URL collection process. Endpoints in the file are disallowed because using them would break the platform's terms of service.

4.1.2 Platform-Specific Analysis

ERR For ERR, our evaluation seen in Table 1 shows that while the sitemap method offers moderate historical coverage (approximately 6 months) and a large batch size ($\frac{n}{1000}$), the structured and complete data retrieval is best achieved using the Search Functionality. This method provides complete historical coverage with a high success

rate, returning full metadata including Article ID, Heading, Dates, and Category ID. Although the complexity metric of $\frac{n}{50}$ indicates a moderate number of requests, the consistency and comprehensiveness of the results justify its selection.

Table 1. Evaluation of URL collection methods for ERR.

Method	**Historical Coverage**	**Complexity**	**Number of URLs**	**Ease of Access to Metadata**
Sitemap	≈ 6 months	$\frac{n}{1000}$	10 000	Article URL, Last Modification Date
Search Functionality	Complete	$\frac{n}{50}$	Complete	Article ID, Heading, Dates, Category ID
Hardcoded URL Requests	Complete	$n + x$	Complete	Not Applicable
Sections	≈ 4 days	$\frac{n}{50}$	50 per section	Article URL, Date, Category, Lead

Delfi As seen on Table 2 the sitemap method for Delfi is limited in historical reach (approximately 1 month) and yield (only 1,000 URLs), while search functionality is not applicable since it is disallowed in the `robots.txt` file. Hardcoded URL Requests achieve complete coverage but do not provide metadata. Moreover, a significant drawback is the high number x of failed requests, which occurs because the article IDs embedded in the URLs are random, resulting in many unsuccessful attempts to retrieve valid pages. Consequently, the sections method is the most effective: it covers a broad historical period (approximately 1–2 years) and can yield up to 10,000 URLs per section, with metadata including title, dates, and category. The complexity, measured as $\frac{n}{41}$, is competitive given the high volume of URLs collected.

Table 2. Evaluation of URL collection methods for Delfi.

Method	**Historical Coverage**	**Complexity**	**Number of URLs**	**Ease of Access to Metadata**
Sitemap	$\approx$ 1 month	$\frac{n}{50}$	1 000	Article URL, Dates, Title, Keywords
Search Functionality	Not applicable	Not applicable	Not applicable	Not applicable
Hardcoded URL Requests	Complete	$n + x$	Complete	Not applicable
Sections	$\approx$ 1–2 years	$\frac{n}{41}$	Up to 10 000 per section	Article ID, Heading, Dates, Category

Postimees Postimees presents a similar challenge as Delfi. As seen on Table 3 The sitemap method has limited yield (only 100 URLs) despite a reasonable complexity ($\frac{n}{100}$), and search functionality is not applicable. Hardcoded URL Requests offer complete coverage, but again, do not provide metadata, and a high number of failed requests makes the method slow. The sections method is chosen for Postimees as it provides extensive historical coverage (approximately 1–2 years) and a large number of URLs (up to 10,000 per section), along with readily accessible metadata (title, dates, category). The overall completeness makes it the preferred approach.

Table 3. Evaluation of URL collection methods for Postimees.

Method	**Historical Coverage**	**Complexity**	**Number of URLs**	**Ease of Access to Metadata**
Sitemap	$\approx$ 1 day	$\frac{n}{100}$	100	URL, Dates, Title, Keywords
Search Functionality	Not applicable	Not applicable	Not applicable	Not applicable
Hardcoded Requests	Complete	$n + x$	Complete	Not applicable
Sections	$\approx$ 1–2 years	$\frac{n}{100}$	Up to 10 000 per section	ID, Heading, Dates, Category

Overview The comparison of URL collection methods chosen between platforms is summarised in Table 4 below. For ERR, the search functionality method is preferred due to its complete historical coverage and robust metadata, despite a moderate request rate ($\frac{n}{50}$). Both Delfi and Postimees benefit most from the sections method, which has a complexity of ($\frac{n}{41}$ for Delfi and $\frac{n}{100}$ for Postimees), it delivers extensive historical coverage (approximately 1–2 years) and a significantly larger volume of URLs per section along with appropriate metadata. It is important to note that the historical coverage across platforms does not need to be entirely exhaustive, as our analysis of circular reporting benefits from overlapping data. For example, if an article from a specific time range is already collected on one platform, there is a higher likelihood that another article from a similar period will reference it, reducing the need to make additional requests for similar information.

Table 4. Overall comparison of URL collection methods across platforms.

Attribute	**ERR**	**Delfi**	**Postimees**
Method	Search Functionality	Section Crawling	Section Crawling
Historical Coverage	Complete	$\approx$ 1–2 years	$\approx$ 1–2 years
Complexity	$\frac{n}{50}$	$\frac{n}{41}$	$\frac{n}{100}$
Number of URLs	Complete	Up to 10 000 per section	Up to 10 000 per section
Ease of Access to Metadata	Article ID, Heading, Dates, Category	Article ID, Heading, Dates, Category	Article ID, Heading, Dates, Category

The total number of unique articles collected was 283,880, with 88,030 from Postimees, 95,850 from Delfi, and 100,000 from ERR, as shown in Table 5. The category distributions vary because each outlet uses its own internal classification system, and the APIs work differently. For ERR, only the latest articles were collected using the available search functionality, which also affects the distribution. The goal was to gather a roughly similar number of articles from each outlet, but the exact breakdown was shaped by these technical limitations.

Table 5. Number of unique articles collected by outlet and section.

Section	Postimees	Delfi	ERR
Estonia	10000	10000	27496
World	10000	10000	13312
Economy	10000		4724
Culture	10000	10000	10557
Sports	10000		27444
Opinion	9999	10000	3143
Science	10000	10000	560
Education	4943		94
Tech (Digi)		10000	265
Health			748
Environment			1287
History			223
Nature			482
COVID-19	10000	5556	
War	10000		
Football		10000	
Basketball		10000	
Crime		10000	
Technology (Tehnoloogia)		801	
Other			9665
Total	**88030**	**95850**	**100000**

With the articles dataset complete, the following section describes the first method for circular reporting detection.

4.2 URL Matching

URL matching analyses the reference hierarchies to detect and classify cases of circular reporting. The method is applied to the full reference hierarchy constructed for each original article.

In this hierarchy, level refers to the number of reference steps from the original article.

First-level references are articles directly cited by the original article. Second-level references are articles cited through a first-level reference.

The length of a reference path is defined as the number of articles in the path, including the original article. A first-level reference has a path length of 2 (original article $\rightarrow$ referenced article), while a second-level reference has a path length of 3 (original article $\rightarrow$ first-level reference $\rightarrow$ second-level reference).

4.2.1 Constructing Reference Hierarchy

Reference extraction maps the flow of citations between articles, constructing a hierarchical dataset. This involves tracing references from the original articles to direct citations and then to second-level references.

Our pipeline ended up checking sources up to 2nd level depth (where the seed source is considered as the 0th level). We did this on the assumption that a journalist is expected to fact-check their cited sources [19], and going beyond that is out of scope with regard to their journalistic duties. Thus, in a two-level system, the journalist is expected to check their references and, at most, to check the references of their references. Only articles with fewer than six references are retained in the hierarchy, as higher reference counts are typically associated with blog-style content, monthly summaries, or editorial overviews. Including articles with a high number of references would introduce significant noise into the analysis, as such articles are less likely to represent individual reporting events and more often serve as aggregated content. Since this thesis focuses on identifying circular reporting in traditional online news articles, these aggregated or editorial pieces are excluded from the hierarchy. Additionally, only articles published in Estonian were included in the dataset, as the scope of the thesis is limited to Estonian-language journalism. The following steps describe the process of building the reference hierarchy.

1. **Extracting First-Level References:** For each original article, all outbound references to ERR, Delfi, or Postimees are extracted from the article. These are treated as first-level references and associated with the originating article.

2. **Fetching First-Level Articles:** Each first-level reference is fetched and parsed using outlet-specific content extraction logic to obtain the outbound references.

3. **Fetching Second-Level Articles:** The same process is repeated for each reference found within first-level articles, forming the second level of the hierarchy.

4. **Building the Citation Hierarchy:** A nested dictionary is constructed in memory, where each article URL maps to its content and a dictionary of referenced child articles. This recursive structure represents the full reference chain up to the defined depth.

5. **Storing the Citation Data:** Once the hierarchy is complete, it is serialized to disk as a JSON file. This format explicitly preserves the structure of citation relationships and enables further analysis.

4.2.2 Duplicate Reference Detection

The implemented system constructs a full reference hierarchy for each original article, based on the extracted citations from Estonian news articles. For each article, the hierarchy captures first-level references (direct citations) and second-level references (citations within the directly cited articles) recursively in a nested structure.

The duplicate reference detection step analyzes the constructed hierarchies to identify potential cases of circular reporting. For each original article, the hierarchy is filtered and retained if a duplicate article URL appears along different paths within the reference tree.

This method is based on the heuristic assumption that duplicate references in the citation hierarchy may indicate circular reporting. Duplicate references can occur when multiple paths point to the same article, as illustrated in Figure 1, or when an article directly or indirectly cites itself, as shown in Figure 2. Detecting such duplicates highlights structures where information may have been recycled or propagated through multiple sources.

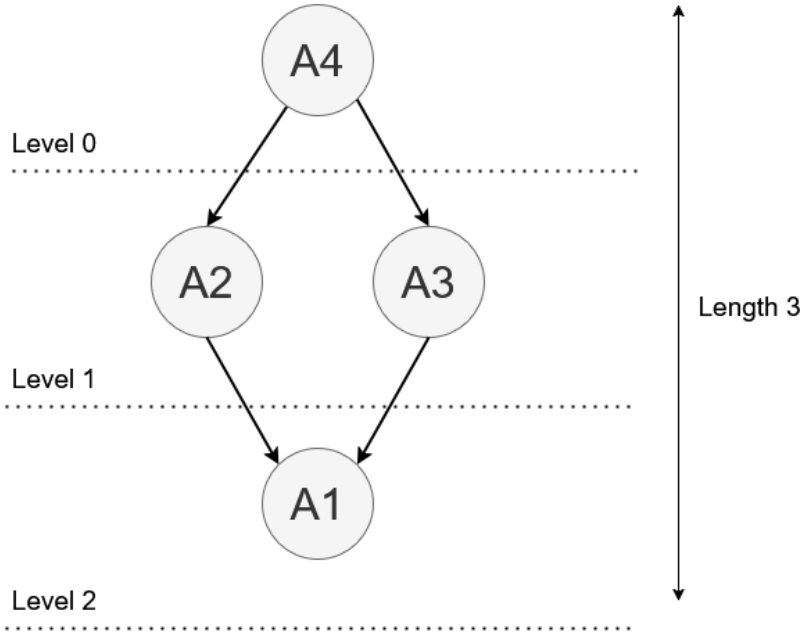

Figure 1. Basic form of circular reporting with four articles.

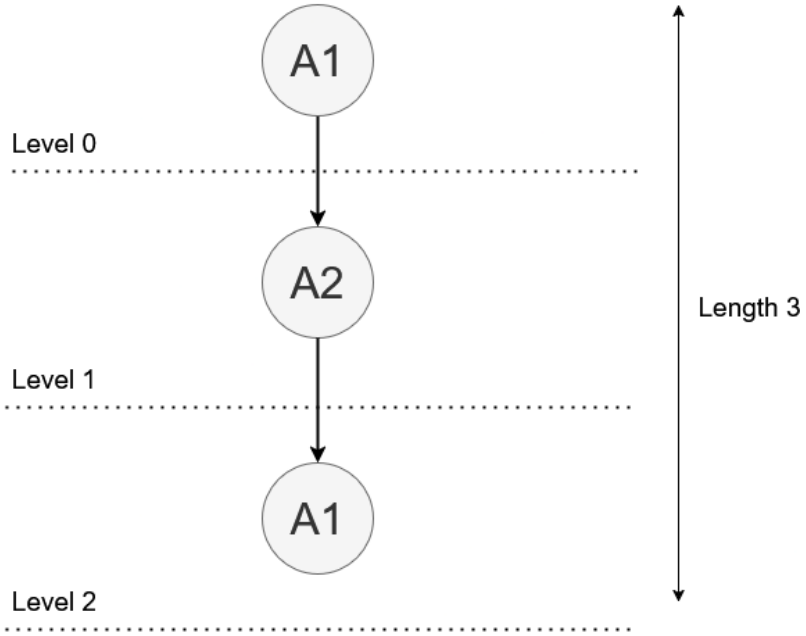

Figure 2. Self-referencing with three articles.

Because circular reporting can also occur without explicit hyperlinks, especially when information is paraphrased or copied, the next section introduces a second detection method based on textual similarity between articles.

4.3 Text Matching

Textual similarity analysis is used to identify cases of content reuse that are not detected through reference-based methods. While duplicate references may indicate circular reporting, not all information reuse involves explicit citation. Journalists may rephrase, paraphrase, or partially quote the same source without linking to it. As a result, articles can exhibit high content overlap without sharing any direct references.

To detect such indirect cases of circular reporting, textual similarity is computed between all pairs of referenced articles in each reference hierarchy (i.e., between all first- and second-level references of a given article). This requires extracting the full text content of each article across all platforms.

Content Extraction. Article content is extracted using outlet-specific logic due to structural differences across ERR, Delfi, and Postimees. For HTML-based content extraction, the `requests`[1] and `BeautifulSoup`[2] libraries are used to load and parse the HTML responses. For API-based content extraction, responses are parsed directly. In both cases, the body text and embedded references are extracted and stored in a structured format. This extraction is performed at all levels of the reference hierarchy, including original articles, first-level references, and second-level references (only content is saved).

TF-IDF Vectorization. Term frequency–inverse document frequency (TF-IDF) is a standard method in information retrieval for numerically representing text. It assigns weights to terms based on their frequency in a document relative to their frequency in the corpus. Manning et al. (2008) describe its role in the vector space model, where documents are treated as weighted term vectors [20].

This thesis uses the implementation provided by the `scikit-learn` library [21], which computes the TF-IDF score for a term t in document d as shown in Equation 1.

[1] `https://pypi.org/project/requests/`
[2] `https://pypi.org/project/beautifulsoup4/`

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \cdot \left(\log \left(\frac{1 + N}{1 + \text{df}(t)} \right) + 1 \right) \tag{1}$$

where:

- $\text{tf}(t, d)$ is the raw term frequency in the document,
- $\text{df}(t)$ is the number of documents containing term t,
- N is the total number of documents.

The resulting vectors are L2-normalized before similarity comparisons.

Cosine Similarity. Cosine similarity is used to compare TF-IDF vectors by computing the angle between them. Manning et al. (2008) describe its use in combination with TF-IDF for document similarity [20]. Sitikhu et al. (2019) found it to have the highest interpretability among semantic similarity methods [22]. Singh and Singh (2021) showed that it achieved the best accuracy, recall, and F1-score for identifying similar news content [23]. This thesis uses the implementation provided by `scikit-learn` [24].

The cosine similarity between two vectors A and B is computed as shown in Equation 2.

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \tag{2}$$

Thresholding and Evaluation. To evaluate how similarity thresholds affect detection, article pairs are grouped into threshold ranges, and the number of pairs retained above each threshold is counted. As the threshold increases, fewer but more reliable matches are preserved. This analysis helps distinguish between incidental topical overlap and strong cases of reused content.

As illustrated in Figure 3, even when articles do not cite the same source, a high similarity score may indicate indirect circular reporting. These cases suggest information

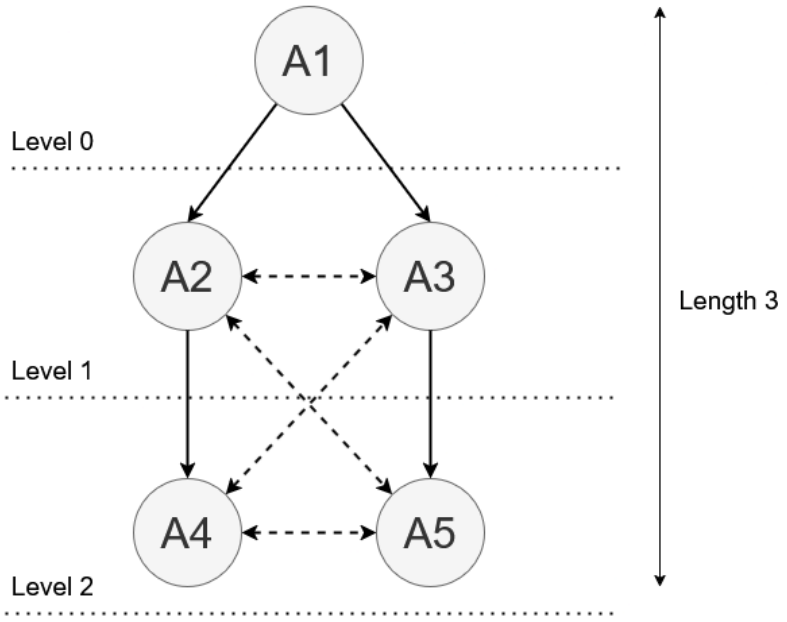

Figure 3. Circular reporting where a dashed line indicates articles rephrasing or copying content.

propagation that bypasses direct reference, such as quoting the same third-party report or copying phrasing from a common origin.

For qualitative analysis, ten pairs of articles were sampled from each similarity range (when available) and manually labelled into four categories:

- **Dissimilar** – Articles are unrelated. Different topics, different content.
- **Somewhat Similar** – Articles are loosely related. Same general topic, but not the same event or angle.
- **Similar (Subtopic)** – Articles are on the same subtopic, but cover different developments, timeframes, or perspectives.
- **Similar** – Articles describe the same event or statement, often with overlapping phrasing or near-duplicate text.

The classification results indicate a correlation between cosine similarity and article pair similarity. Lower thresholds (below 0.2) correspond mostly to dissimilar or weakly

related content. The midrange thresholds (0.2–0.5) contain primarily articles with topical overlap. Higher thresholds (above 0.5) indicate cases of content reuse, with near-identical phrasing or duplication cases.

Figure 4 presents the classification distribution and the number of articles retained per threshold.

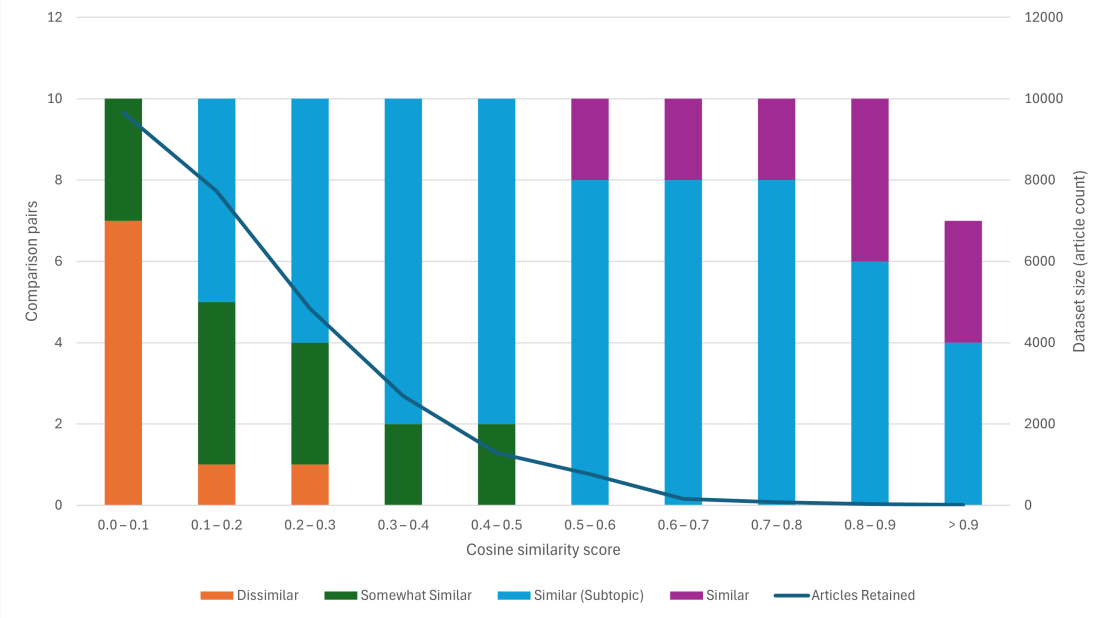

Figure 4. Distribution of Article Pair Classifications and Retained Articles per Similarity Threshold.

4.4 Validation

Validation is performed by manually reviewing all cases flagged as potential circular reporting by the system.

For URL matching, all identified duplicate references were compiled into a CSV file containing the original article, the duplicate article, and the corresponding reference path lengths. These path lengths indicate the positions of the duplicates within the citation hierarchy relative to the original article.

For text matching, all article triplets where the cosine similarity between the two referenced articles exceeds 0.5 are included in the analysis. These are recorded in a CSV file containing the original article, the first referenced article, the second referenced

article, and their computed similarity score. This structure allows manual inspection of article pairs likely to share overlapping content within a given reference hierarchy.

Each flagged case is manually classified as a true positive (correct detection of circular reporting) or a false positive (incorrect detection). This classification depends on context and involves some subjectivity, since it relies on how articles reference one another.

Only precision is reported, as shown in Equation 3, since the goal is to measure how accurately the system identifies relevant circular reporting cases.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Here, TP stands for true positives and FP for false positives. Recall is not evaluated because manually checking the entire dataset, including unflagged cases, is not feasible. Since the system is intended to surface likely cases for human review, precision is the most relevant metric.

Limitations are acknowledged, as not all detected loops necessarily indicate poor reporting practices. Manual validation allows distinguishing between problematic and acceptable forms of information reuse.

5 Results

This section presents and analyses the results of detected circular reporting cases, based on manual classification of whether the articles identified by either method are true or false positives. Table 6 gives an overview of the results by method.

Table 6. Manual Classification Results for Both Detection Methods.

Method	**True Positives**	**False Positives**	**Total**
URL Matching	47	1268	1315
Text Similarity	4	376	380

True positives and false positives by method are analysed further In the following subsections.

5.1 URL Matching Results

Out of 1315 detected duplicate reference cases, 47 were manually confirmed as positive cases of circular reporting, resulting in a precision of approximately 3.6%.

The structure of the detected duplicates was analysed by examining the levels at which duplicate references appeared within the citation hierarchy. Each reference chain was normalised by sorting the path lengths at which the duplicate article appeared (e.g., a duplicate found at the second and third levels was represented as `2, 3`). This normalisation ensured that the order of appearance did not affect the classification. A variety of structural patterns were identified:

- **Simple two-level duplicates (2, 3):** Duplicates found at two distinct path lengths in the reference chain.
- **Same-level duplicates (2, 2 or 3, 3):** Duplicates appearing multiple times at the same depth.
- **Self-referencing structures (1, 2):** The original article appeared again within its own references.

- **Complex multi-level structures (e.g., 2, 2, 3, 3):** Duplicates found at multiple overlapping depths.

Table 7 summarizes the distribution of each observed structure and the corresponding precision in identifying circular reporting.

Table 7. Distribution and Precision of Circular Reporting by Duplicate Structure Type

Duplicate Structure Type	**Total**	**Circular (True)**	**Non-Circular (False)**	**Precision**
1, 2	14	10	4	0.714
1, 3	100	0	100	0.000
1, 3, 3	3	0	3	0.000
2, 2	16	5	11	0.313
2, 2, 3	2	0	2	0.000
2, 2, 3, 3	1	1	0	1.000
2, 3	983	28	955	0.028
2, 3, 3	80	1	79	0.013
2, 3, 3, 3	9	0	9	0.000
3, 3	104	2	102	0.019
3, 3, 3	3	0	3	0.000
Total	**1315**	**47**	**1268**	**0.036**

The analysis revealed that self-referencing cases (where the original article cited itself) were confirmed as true positives with the highest precision.

For non-self-referencing duplicates, some cases where an article referenced the same external article multiple times were caused by variations in URL formatting. Although the URLs initially appeared different, canonicalization revealed that they pointed to the same final article. These duplicate references typically resulted from normal linking practices, such as inline citations combined with "read more" sections.

During methodology development, duplicate references in general were considered important indicators of possible circular reporting and were included for manual validation. However, in a specific subset of cases — where the same external article was referenced multiple times from the same originating article — it was initially assumed that such duplicates would not indicate circular reporting. These cases typically arose

from normal linking practices, such as an inline reference and "read more" link at the end of the article.

Nevertheless, due to differences in URL formatting, several duplicate references originating from the same article appeared separately in the hierarchy. This turned out to be beneficial, as five true positive cases of circular reporting were detected in structures where the reference levels were 2 and 2. Although the links appeared different initially, they canonicalized to the same article.

Therefore, including all duplicate references, even if they canonicalize to the same article, can be justified, especially in smaller datasets where manual review is feasible. Manual validation remains necessary to distinguish legitimate linking from actual information recycling.

5.2 Text-matching Results

To evaluate how textual similarity between referenced articles affects detection precision, article triplets (original, reference 1, reference 2) with a cosine similarity above 0.5 were reviewed manually. Out of 380 such high-similarity cases, only 4 were confirmed as true positives, resulting in an overall precision of approximately 1%. The similarity values for the four true positive cases were approximately: 0.71, 0.86, 0.9 and 1.0. Given the very small number of true positives, no strong conclusions can be drawn. However, increasing the threshold value from 0.5 was observed to improve precision, as illustrated in Figure 5.

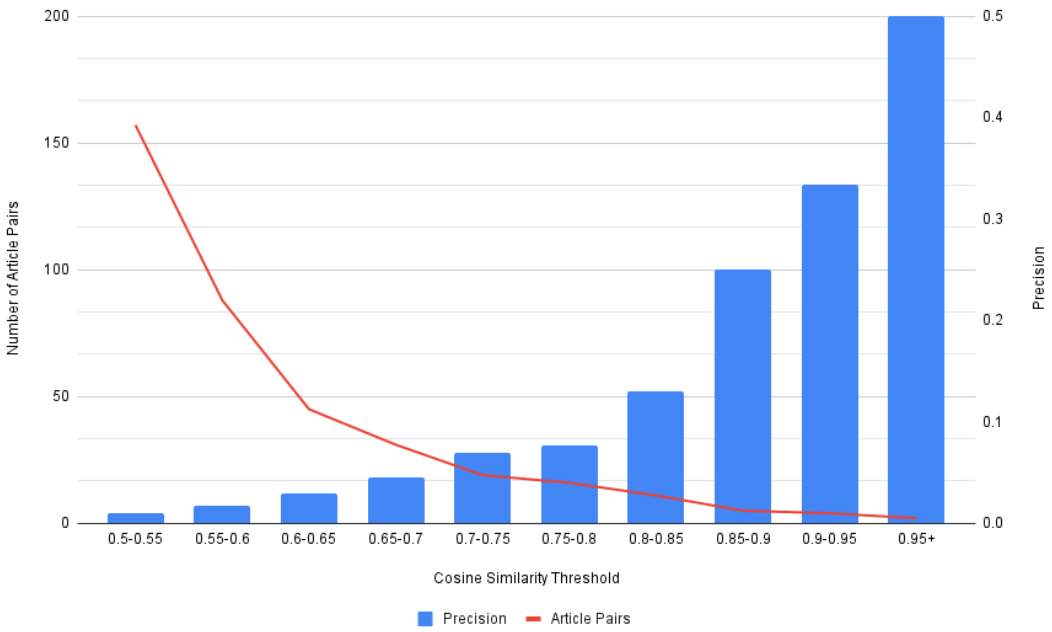

Figure 5. Precision of Circular Reporting Detection and Article Pairs by Cosine Similarity Threshold.

Similarly to detecting circular reporting by using duplicate links in the hierarchy, textual similarity alone cannot be used reliably without manual verification.

5.3 Explorative Analysis of Estonian Circular Reporting

Within our experiments, we identified 51 cases where circular reporting happened. The duplicate reference detection method identified 47 true positive cases of circular reporting, while the text similarity method identified 4 cases. Out of the combined 51 true positive cases, 22 involved original articles published by Postimees, 8 by Delfi and 20 by ERR. There were 277 negative cases from Postimees (226 unique articles), 205 from Delfi (170 unique articles), and 1161 from ERR (831 unique articles). ERR contributed substantially more articles to the analysis, approximately 3 to 4 times more than either Postimees or Delfi. This is likely due to the broader section coverage when using ERR's search functionality, which included more articles from Sports and Estonia sections. One circular reporting case was detected by both text matching and URL matching methods. Additionally, one article was classified as a true positive twice, caused by two separate

duplicate references. Therefore, the overall true positive unique article count was 49.

Of the 49 unique articles identified as true positive circular reporting cases, the majority belonged to the sport section, followed by Estonian internal news. Figure 6 shows the distribution by platform and section.

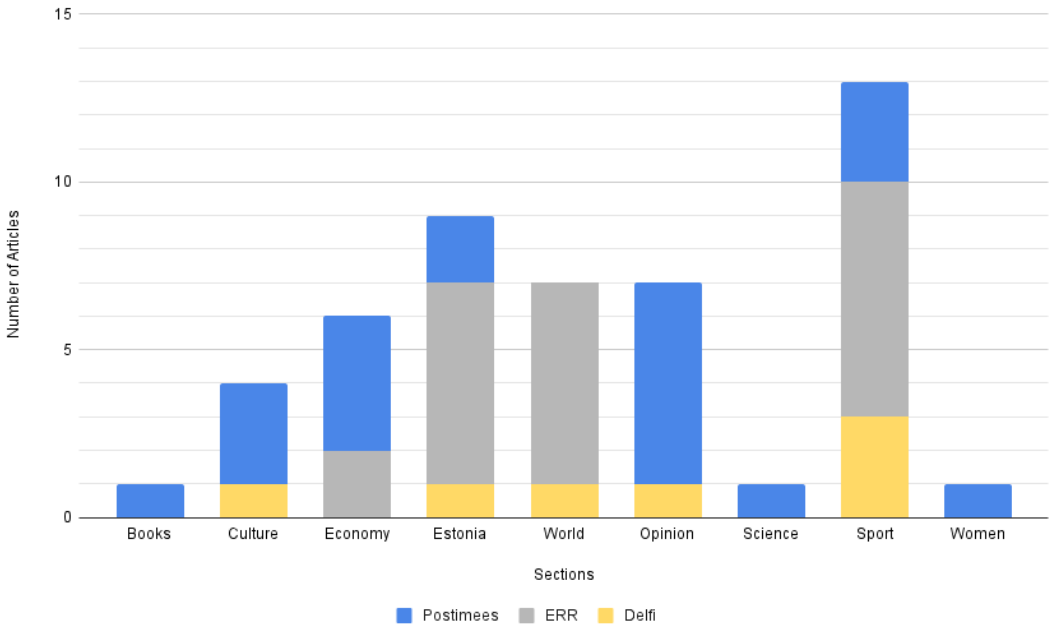

Figure 6. True Positive Circular Reporting Articles by Section and Platform.

To compare the characteristics of articles marked as circular reporting with those that were not, we looked at two basic metrics: text length (in characters) and number of direct references.

On average, the text length of true positive articles was slightly shorter than that of false positives. The average number of direct references followed a similar trend. These results are shown in Figure 7, which compares both metrics for true and false positive articles overall.

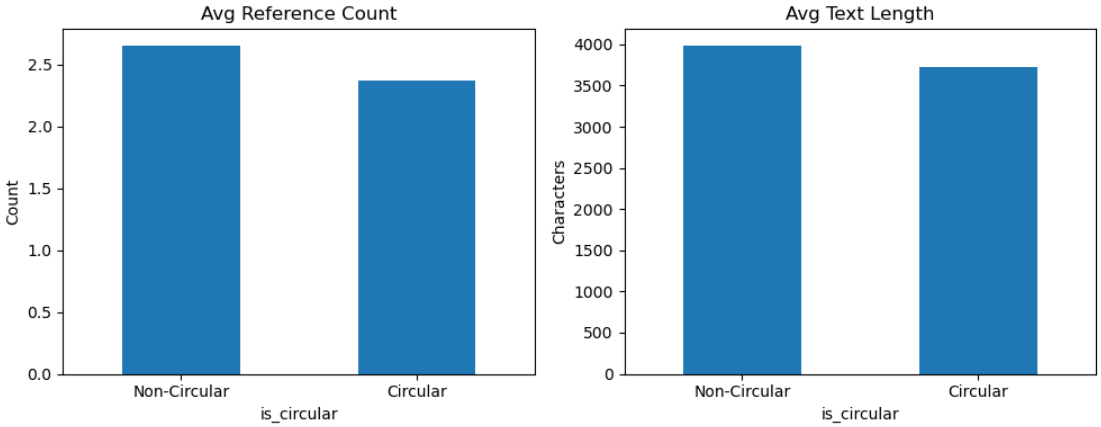

Figure 7. Average Article Text Length and Direct Reference Counts Between True and False Positives.

We also broke the analysis down by platform to check if any platform-specific patterns stood out. As shown in Figure 8, for ERR, the average text lengths between circular and non-circular articles were very similar. In Postimees and Delfi, circular reporting articles tended to be shorter, though the sample of confirmed cases is small and this difference should not be overinterpreted. One consistent observation is that Delfi articles were, on average, almost twice as long as articles from the other two outlets. This applies to both circular and non-circular cases and may be related to how content is structured on Delfi's platform.

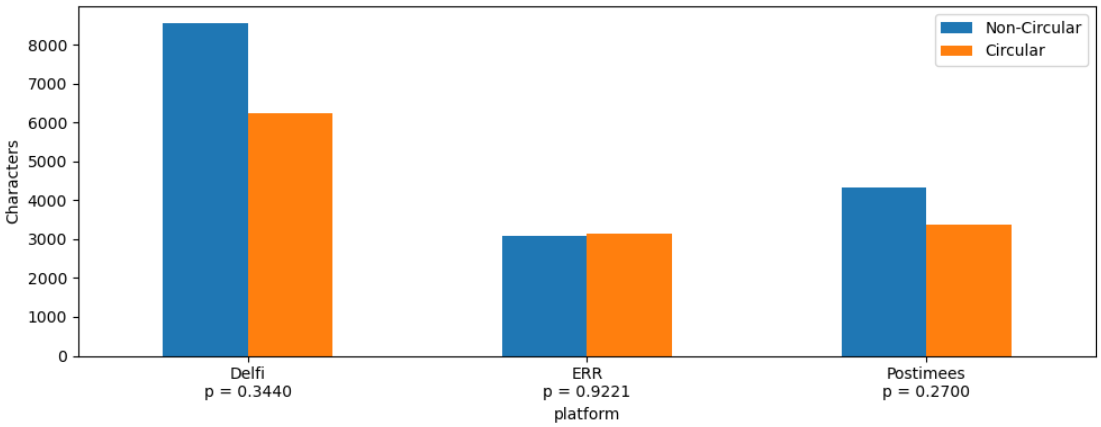

Figure 8. Average Article Text Length per Platform, Grouped by Circularity Label.

As for direct references, Figure 9 shows that circular reporting articles had slightly

fewer links than non-circular ones across all three platforms. However, the differences are small and not statistically significant. For this thesis, one-way ANOVA (analysis of variance) was used to assess whether the differences in article reference count and text length between circular and non-circular articles were statistically significant across platforms [25]. As shown in Figure 8 and Figure 9, the returned p-values were all greater than 0.05, suggesting that the observed differences likely occurred by chance and are not statistically meaningful. Overall, the figures suggest that circular reporting cases do not differ meaningfully in article length or linking behaviour compared to false positives.

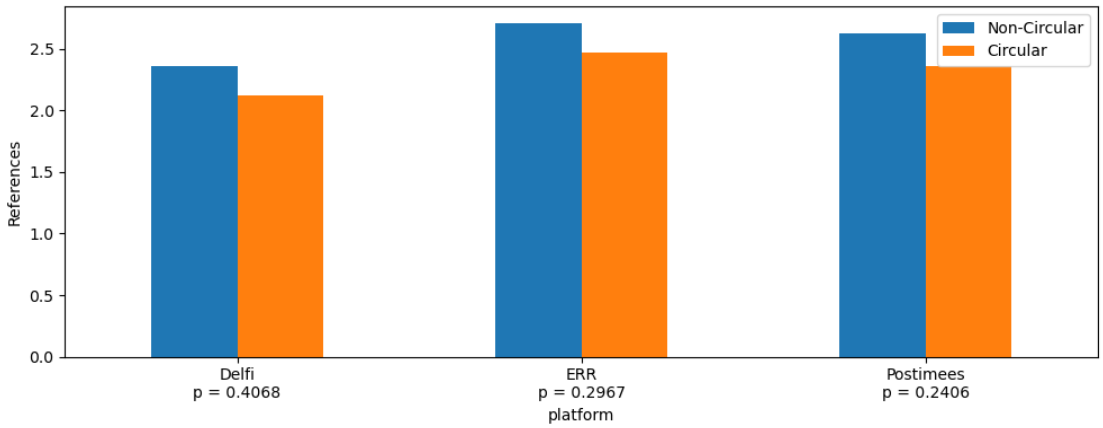

Figure 9. Average Direct Reference Count per Platform, Grouped by Circularity Label.

These comparisons provide context for how circular reporting cases relate to article structure and linking behaviour. The next section discusses how well the detection methods performed and interprets the results.

5.4 Discussion of Results

The duplicate reference detection approach demonstrated a higher precision compared to the text similarity method. The most reliable indicator was the presence of a self-referencing structure, where the original article cited itself within the reference hierarchy. In such cases, the precision was 71%.

For text similarity-based detection, raising the minimum similarity threshold to at least 0.7 helps reduce false positives and manual validation effort. However, setting the minimum similarity threshold higher may exclude some true positives.

In a practical deployment scenario, where the detection system would operate continuously (processing new articles periodically), the volume of detected candidates would likely be significantly smaller than in this retrospective batch analysis. As a result, manual validation would become a more manageable and less time-consuming task.

Overall, the precision for both URL matching (3.6%) and text matching (1.0%) was very low. This can be explained by how articles are often structured. In many cases, a third article references both a second and a first article to provide chronological or contextual background. If the second article was published earlier and also references the first, the system identifies the first as a duplicate reference (2, 3 duplicate structure type). This was especially common in URL matching results, though not exclusive to them. Similar sequences of events described across articles in the same context can also produce high textual similarity.

These cases are not truly circular reporting when the references are appropriate and transparent. They do not create the illusion of multiple independent sources, but instead reflect a single evolving news story that is updated or expanded over time. In such cases, the references serve to provide context rather than to falsely reinforce credibility. It is more a case of moving news than static misinformation.

5.5 Limitations

One limitation encountered was related to canonicalization inconsistencies. When an original article referenced itself, differences in URL prefixes (such as section names like "sport") caused the system to treat the original and referenced URLs as distinct, leading to unnecessary re-fetching of the same article. Similarly, duplicate links arose when different URLs were used to reference the same external article, typically due to minor

formatting differences. Although canonicalization ultimately identified these cases as duplicates, the articles were still fetched separately during hierarchy construction.

Rate limiting also imposed a practical constraint, as each article had to be fetched individually to extract its content, making large-scale collection significantly slower.

Both detection methods had low precision, with the text similarity approach performing even worse than the duplicate reference method. As a result, extensive manual validation was required to assess whether the detected cases truly constituted circular reporting. Raising the similarity threshold or limiting detection to self-references would improve precision but at the cost of missing some valid cases. A deeper contextual understanding was necessary to distinguish circular reporting from acceptable citation practices.

Some cases of circular reporting may still go undetected, especially if the content is reused without explicit linking or in a heavily paraphrased form. The detection logic relies on extracted references and measurable textual similarity, so patterns outside those constraints are not captured.

Manual classification was done only by the authors of the work, thus the results may vary a bit depending on interpretation of classes.

Additionally, certain types of content such as collection articles and blog-style pieces were excluded from the analysis, as they often follow a different editorial structure and could distort detection results.

6 Conclusions

This thesis addressed the problem of detecting circular reporting in Estonian online news media, where the same information is recycled across articles, potentially misleading readers regarding the independence and credibility of sources. While prior research has explored fake news detection and citation loops in scientific literature, there has been limited focus on the specific phenomenon of circular reporting in journalism, especially in the context of small media ecosystems such as Estonia's.

To tackle this issue, we proposed a two-stage detection framework. First, a reference hierarchy was constructed for each article using extracted citations from ERR, Delfi, and Postimees. Within this hierarchy, duplicate references were identified and flagged as potential circular reporting cases. Second, a textual similarity analysis using cosine similarity on TF-IDF vectors was applied to detect cases where referenced articles reused content without direct citation, thereby revealing implicit content recycling.

The duplicate reference method detected 1315 candidate cases, of which 47 were confirmed as true positives through manual validation. The textual similarity method detected 380 candidate cases (similarity higher than 0.5), of which 4 were validated as true positives. Among these, the highest precision was observed in self-referencing structures (71%), and higher textual similarity thresholds (over 0.7) were found to reduce false positives. Overall, 49 articles with circular reporting were identified.

Several directions remain open for future work. The format in which URLs are referenced within articles warrants further investigation to assess whether the use of mixed URL formats within a single article is associated with poor citation practices or circular reporting. The precision of textual similarity analysis could be improved by incorporating semantic embedding models such as BERT, potentially capturing deeper rephrasings of reused content. Expanding the reference hierarchy beyond two levels may uncover longer citation loops, although such depth extends beyond the typical scope of journalistic verification. Additionally, deploying the detection system in a live setting could enable real-time identification of circular reporting and contribute to ongoing media accountability.

References

[1] Diego Saez-Trumper. Online disinformation and the role of wikipedia, 2019.

[2] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5), September 2020.

[3] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, volume 14, pages 626–637, May 2020.

[4] David A. Pendlebury Martin Szomszor and Jonathan Adams. How much is too much? the difference between research influence and self-citation excess. *Scientometrics*, 123(2):1119–1147, 2020.

[5] Wikipedia. Circular reporting — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Circular_reporting`, 2024. [Online; accessed 17-November-2024].

[6] Stephen Harrison. The internet's dizzying citogenesis problem. *Slate*, March 2019.

[7] Thomas Boghardt. Soviet bloc intelligence and its aids disinformation campaign. *Studies in intelligence*, 53(4):1–24, 2009.

[8] Mark Deuze. What is journalism? professional identity and ideology of journalists reconsidered. *Journalism*, 6(4):442–464, 2005.

[9] Imke Henkel, Neil Thurman, Judith Möller, and Damian Trilling and. Do online, offline, and multiplatform journalists differ in their professional principles and practices? findings from a multinational study. *Journalism Studies*, 21(10):1363–1383, 2020.

[10] Peter L.M. Vasterman. Media-hype: Self-reinforcing news waves, journalistic standards and the construction of social problems. *European Journal of Communication*, 20(4):508–530, 2005.

[11] Juliette De Maeyer. Citation needed. *Journalism Practice*, 8(5):532–541, 2014.

[12] Andreas Spitz and Michael Gertz. Breaking the news: Extracting the sparse citation network backbone of online news articles. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, page 274–279, New York, NY, USA, 2015. Association for Computing Machinery.

[13] Xi Cui and Yu Liu. How does online news curate linked sources? a content analysis of three online news media. *Journalism*, 18(7):852–870, 2017.

[14] Fahad Alsuliman, Siddhartha Bhattacharyya, Khaled Slhoub, Nasheen Nur, and Candice Normalee Chambers. Social media vs. news platforms: A cross-analysis for fake news detection using web scraping and nlp. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, page 190–196, New York, NY, USA, 2022. Association for Computing Machinery.

[15] Lizhou Fan, Sara Lafia, David Bleckley, Elizabeth Moss, Andrea Thomer, and Libby Hemphill. Librarian-in-the-loop: A natural language processing paradigm for detecting informal mentions of research data in academic literature, 2022.

[16] Yi Bu, Yong Huang, and Wei Lu. Loops in publication citation networks. *Journal of Information Science*, 46(6):837–848, 2020.

[17] Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, and Nibrat Lohia. Fake news detection: A deep learning approach. *SMU Data Science Review*, 1(3), 2018.

[18] Xinyi Zhou and Reza Zafarani. Network-based fake news detection: A pattern-driven approach. *SIGKDD Explor. Newsl.*, 21(2):48–60, November 2019.

[19] Morten Hertzum. How do journalists seek information from sources? a systematic review. *Information Processing Management*, 59(6):103087, 2022.

[20] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[21] scikit-learn developers. Tf-idf term weighting — scikit-learn documentation. `https://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting`, 2025. Accessed 2025-05-01.

[22] Pinky Sitikhu, Kritish Pahi, Pujan Thapa, and Subarna Shakya. A comparison of semantic similarity methods for maximum human interpretability. 11 2019.

[23] Ritika Singh and Satwinder Singh. Text similarity measures in news articles by vector space model using nlp. *Journal of The Institution of Engineers (India): Series B*, 102(2):329–338, 2021.

[24] scikit-learn developers. Cosine similarity — scikit-learn documentation. `https://scikit-learn.org/stable/modules/metrics.html#cosine-similarity`, 2025. Accessed 2025-05-01.

[25] Tae Kyun Kim. Understanding one-way anova using conceptual figures. *Korean Journal of Anesthesiology*, 70(1):22–26, February 2017.

Appendix

I. CircularCheck Tool Repository

The source code and pipeline used in this thesis are available at: `https://github.com/KasperKaljuste/CircularCheck-Est`.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Kasper Kaljuste**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   CircularCheck: A Tool for Detecting Circular Reporting,
   *(*title of thesis)

   supervised by Uku Kangur.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Kasper Kaljuste
15/05/2025