View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by ZHAW digitalcollection

future internet

MDPI

Article

Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities

Albert Weichselbraun [1,*], Philipp Kuntschik [1], Vincenzo Francolino [1], Mirco Saner [2], Urs Dahinden [1] and Vinzenz Wyss [2]

[1] Institute for Information Research, University of Applied Sciences of the Grisons, 7000 Chur, Switzerland; philipp.kuntschik@fhgr.ch (P.K.); vincenzo.francolino@fhgr.ch (V.F.); urs.dahinden@fhgr.ch (U.D.)

[2] IAM Institute of Applied Media Studies, Zurich University of Applied Sciences, 8400 Winterthur, Switzerland; mirco.saner@zhaw.ch (M.S.); vinzenz.wyss@zhaw.ch (V.W.)

* Correspondence: albert.weichselbraun@fhgr.ch; Tel.: +41-81-286-3727

Abstract: Recent developments in the fields of computer science, such as advances in the areas of big data, knowledge extraction, and deep learning, have triggered the application of data-driven research methods to disciplines such as the social sciences and humanities. This article presents a collaborative, interdisciplinary process for adapting data-driven research to research questions within other disciplines, which considers the methodological background required to obtain a significant impact on the target discipline and guides the systematic collection and formalization of domain knowledge, as well as the selection of appropriate data sources and methods for analyzing, visualizing, and interpreting the results. Finally, we present a case study that applies the described process to the domain of communication science by creating approaches that aid domain experts in locating, tracking, analyzing, and, finally, better understanding the dynamics of media criticism. The study clearly demonstrates the potential of the presented method, but also shows that data-driven research approaches require a tighter integration with the methodological framework of the target discipline to really provide a significant impact on the target discipline.

Keywords: Big Data; Web Intelligence; media analytics; social sciences; humanities; linked open data; adaptation process; interdisciplinary research; media criticism

check for updates

Citation: Weichselbraun, A.; Kuntschik, P.; Francolino, V.; Saner, M.; Dahinden, U.; Wyss, V. Adapting Data-Driven Research to the Fields of Social Sciences and the Humanities. *Future Internet* **2021**, *13*, 59. https://doi.org/10.3390/fi13030059

Academic Editor: Carlos Filipe Da Silva Portela

Received: 30 January 2021
Accepted: 21 February 2021
Published: 26 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent advances in areas such as Big Data and deep learning have paved the way for the development of Web Intelligence systems that are capable of performing knowledge extraction and data analytics tasks on large and dynamic web and social media corpora in real time. Driven by this success, methods from computer science have expanded to disciplines within the social sciences and humanities, such as business, communication science, economics, healthcare, and even religion. Some researchers have raised concerns about whether the current approach of transferring methods to other fields without considering the target field's theoretical framework, research background, and concepts is really a good strategy for unfolding the full potential of data-driven research approaches [1]. In fact, there are areas within the field of natural language processing that have been doing particularly well, by building upon existing frameworks from other disciplines. In sentiment analysis, for example, many of the most influential researchers draw upon models from psychology and neuroscience [2,3]. The well-known Hourglass of Emotions model and its revisited version [2], for example, blend concepts from psychology, affective neuroscience, and computer science.

Bartlett et al. [4] note that it is telling that computer scientists are rarely called to embrace traditional sociological thought, and they contest the idea that computer scientists should be legitimate interpreters of social phenomena, even if they have been analyzed with data-driven methods. Researchers such as Connolly argue that computer scientists

should receive a more comprehensive training in social sciences [1] to make them better suited for contributing to these fields. Given the wide area of academic disciplines that are considered as social sciences, such a strategy seems challenging and barely actionable in the short term.

This paper, in contrast, proposes a collaborative process that aims at creating a shared understanding between computer scientists and the researchers in the target domain, provides multiple feedback loops to ensure that knowledge extraction and data analytics tasks are well aligned with the underlying theoretical frameworks, and yields results that significantly impact the target domain. The process guides the research design, the collection and formalization of domain knowledge (e.g., as linked open data), data acquisition, data selection, knowledge extraction, and data analytics. It also actively promotes a close collaboration between researchers from different fields to unfold the full potential of their joint research endeavors.

The rest of this paper is structured as follows: Section 2 discusses related research in the fields of Big Data and Web Intelligence. We then provide an overview of the process used for adapting data-driven research methods to other fields. Afterwards, we elaborate on its application to the field of communication science, provide a short discussion of the relevant research framework and background (Section 4), and demonstrate how the process impacts tasks such as collecting domain knowledge (Section 5) and the processes of data acquisition and selection (Section 6), as well as how it affects the choice of knowledge extraction and data analytics techniques (Section 7). Section 8 finally demonstrates the potential of the constructed data analytics platform based on a use case that analyzes the media coverage of the New Year's Eve sexual assaults in Cologne in 2015. The paper closes with the presentation of conclusions in Section 9.

2. Related Work

Big Data and Web Intelligence provide powerful methods for analyzing web and social media content. The potential and capabilities of these data-driven approaches have been successfully demonstrated in many domains ,such as political science [5], environmental communication [6,7], financial market analysis [8,9], healthcare [10], and marketing [11].

Ranganath et al. [5] drew upon social movement theories from political science to design a quantitative framework for studying how advocates push their political agendas on Twitter. They used two datasets for analyzing message and propagation strategies, as well as the community structures adopted by these advocates. Chung and Zeng [12] used network and sentiment analysis on Twitter to investigate the discussion on U.S. immigration and border security. The authors uncovered major phases in the Twitter coverage, identified opinion leaders and influential users, and investigated the differences in sentiment, emotion, and network characteristics between these phases.

In the environmental communication domain, Khatua et al. [7] studied the perception of nuclear energy in tweets covering the 2017 Nobel Peace Prize won by the campaign to abolish nuclear weapons and the 2011 Fukushima nuclear disaster. Scharl et al. [6] presented visual tools and analytics to support environmental communication in the Media Watch on Climate Change (https://www.ecoresearch.net/climate accessed on 21 February 2021), the Climate Resilience Toolkit (https://toolkit.climate.gov/ accessed on 21 February 2021), and the NOAA Media Watch [13]. All three platforms aggregate and analyze the coverage of environmental topics in different outlets, including news media, Fortune 1000 companies, and social media, such as Twitter, Facebook, Google+, and YouTube. The article discusses (i) the implemented metrics and visualizations for measuring communication success, (ii) monitoring the efficiency and impact of newly published environmental information, programs to engage target groups in interactive events, and the distribution of content through partners and news media, and (iii) tracking communication goals.

However, even in traditional mediums, such as television, where metrics by Nielsen Media Research (https://www.nielsen.com accessed on 21 February 2021) are well-established

standards for audience ratings, Web Intelligence is gaining in importance, since it provides techniques for assessing audience engagement rather than the impact and reach of television programs [14]. Wakamiya et al. [15] and Napoli [14] discussed the advantages of performing complementary analyses of social network activities related to television programs, and Scharl et al. [16] investigated the emotion in online coverage of HBO's Game of Thrones in Anglo-American news media, Twitter, Facebook, Google+, and YouTube.

Li et al. [17] drew upon company-specific news articles to study their impact on the movements of stock markets. They concluded that public sentiments voiced in these articles cause fluctuations in the market, although their impact depends on the company as well as the article content. Xing et al. [18] presented an analysis of common mistakes and error patterns within sentiment analysis methods used in the financial domain and provided suggestions on how to counter them. They also provided a comprehensive study of data-driven approaches used for financial forecasting in [9].

Kim et al. [19] investigated the coverage of the Ebola Virus on Twitter and in news media. They created topic and entity networks, computed per-topic sentiment scores, and analyzed the temporal evolution of these networks. Yang et al. [10] developed a recommender system for patients interested in information on diabetes. Their approach was developed on Weibo.com, which is the largest microblogging site in China, and suggested new content based on the users' interests and their attitude towards a topic by considering features extracted from the users' tweets.

The application domain plays an important role in choosing data sources, analytics, and visualizations. Marketing, for instance, often focuses on the discussion of products and product features in online word-of-mouth channels by applying techniques such as opinion mining and conjoint analysis. Xiao et al. [11] extracted consumer preferences from product reviews and used an economic preference measurement model to derive and prioritize customer requirements at a product level based on this information.

A common theme of the work presented above is the expansion of data-driven approaches to other research fields, particularly to social sciences. As outlined in the introduction, such approaches have been highly successful, but have also raised concerns regarding their efficiency and legitimacy in terms of impact on the target domain [1,4]. The presented paper aims at addressing these concerns by proposing a collaborative process that promotes a shared understanding of the research framework and an alignment of hypotheses and goals between the disciplines, as outlined in the next section.

3. Method

Computer scientists that adapt data-driven research methods to other fields often have only a limited understanding of the theoretical framework that guides research within the target domains. Although such knowledge might not be strictly necessary for applying data science to new fields, it seems sensible to suggest that aligning data-driven research with the concepts, research questions, and methodological framework of the target domain will improve the efficiency, effectiveness, and impact of the research outcomes within the target discipline.

This section introduces an iterative process that supports this alignment by promoting a collaborative, interdisciplinary approach that leverages expert knowledge. We have successfully applied this process to a number of interdisciplinary research projects covering domains such as business ethics, communication science, investment, and pharmaceutical drug development.

Figure 1 illustrates the proposed process, which consists of five main tasks that trigger corresponding feedback loops used to align the research design, goals, hypotheses, and data-driven research methods with the target domain and to consequently improve the quality and impact of the created artifacts. The following subsections elaborate on these tasks in greater detail, and are followed by a comprehensive discussion of how this process has been applied to the creation of the Swiss Media Criticism portal in Sections 4–8.

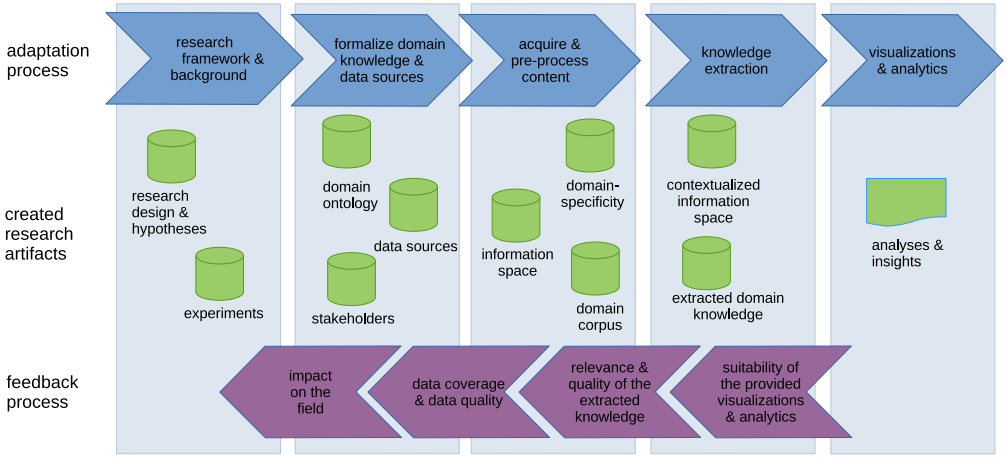

Figure 1. Adapting data-driven research to a new target domain.

3.1. Adaptation to the Research Framework and Background of the Target Domain

The first step aims at creating a joint understanding between computer scientists and researchers from the target domain. It involves introducing the background, research framework, and hypotheses to computer scientists, and communicating the available methods, their potential, and their limitations to the colleagues from the target domain. This stage should also consider how existing concepts from the target domain might support downstream data acquisition, knowledge extraction, and analysis. Finally, the research design, hypotheses, and experiments that help in either confirming or rejecting these hypotheses are created in this step.

An illustrative example of the impact of this first adaptation step is given in Section 4; the definition of media criticism within the communication science literature, which requires a particular stakeholder to voice criticism, had a significant impact on the approach used to detect media-critical content in online news and social media. The domain's research framework has also been instrumental in determining:

1. The required domain knowledge on stakeholders and sources (Section 5), which, in turn, influenced:
2. The sources and source groupings considered by data acquisition components (Section 6) and
3. The types and kinds of entities supported in the named entity-linking process (Section 7.1);
4. Whether a domain-specific affective model or standard sentiment should be used in the knowledge extraction pipeline (Section 7.2);
5. The approach used for data analytics, i.e., investing considerable effort into building the Swiss Media Criticism portal and analytics dashboard, which supports real-time tracking of emerging issues in addition to historical analyses (Section 8).

3.2. Formalization of Domain Knowledge and Selection of Relevant Data Sources

Afterwards, domain experts formalize domain knowledge and collect relevant data sources. The importance of this step cannot be overstated, especially since all later analyses will draw upon data retrieved from sources specified in the formalization step. The domain experts, therefore, need to clearly state which sources are relevant and useful to the analyzed domain and are required for answering the research questions. At the end of this task, they provide:

- Domain knowledge, such as ontological knowledge on entities (e.g., relevant stakeholders, locations, events, etc.) and their relations to each other, which is well suited for supporting data acquisition and knowledge extraction processes. Approaches for computing the domain specificity (Section 6), named entity linking (Section 7.1), and sentiment analysis (Section 7.2) also benefit heavily from domain knowledge.

- Data sources, such as (i) links to relevant web resources (e.g., news media sites), (ii) search terms, and (iii) social media accounts of major stakeholders.

Section 5 outlines the tasks necessary for collecting and formalizing the relevant domain knowledge in greater detail.

3.3. Acquiring and Preprocessing Textual Data

The content acquisition process leverages the specified data sources and domain knowledge for content acquisition and filtering. The later step is of particular importance for content sources, such as news media, that cover a broad selection of topics and, consequently, contain both relevant and irrelevant content.

Criteria and concepts from the research framework (e.g., the concept of media criticism) and formalized knowledge (e.g., stakeholders, domain concepts, etc.) from the previous step are instrumental in ensuring that only relevant documents are included in the information space on which analytics will be performed. Depending on the chosen approach, domain experts might provide (i) regular expressions for white- and blacklists (i.e., text patterns that identify relevant and irrelevant content) or (ii) gold-standard documents for training supervised machine learning components and deep learning, or (iii) might ask computer scientists to apply more advanced techniques that combine these approaches, such as ensemble methods.

Experts then browse this corpus in order to evaluate the acquired documents in terms of (i) relevance, (ii) coverage, and (iii) quality (i.e., whether the content is complete and free of noise, such as navigation elements). The feedback on the data quality triggers (i) adaptations of the domain-specificity component, which aims at improving the document relevance and/or coverage (if relevant documents have been filtered), (ii) the inclusion of additional sources to address missing document sources, and (iii) the optimization of the boilerplate removal to improve the document quality.

This iterative feedback process is also a good point to reflect on the necessity of the selected sources. Within the presented use case, for example, we observed that especially web sites of small media outlets tended to violate web standards and caused problems with the boilerplate removal. This observation, combined with the insight that these sites do not contribute a relevant amount of media-critical content and, therefore, have no real impact on the validation of the research hypotheses, led to the decision to remove them from the list of data sources, yielding a better overall content quality.

3.4. Knowledge Extraction

Once the data acquisition and preprocessing pipeline has been established, knowledge extraction processes (Section 7) draw upon methods such as (i) named entity linking to identify persons, organizations, and locations relevant to the use case, (ii) sentiment analysis to determine the polarity (i.e., positive versus negative coverage) of the retrieved documents, topics, and stakeholders, and (iii) keyword analysis to extract topics and concepts covered in these documents. The contextual information obtained from the knowledge extraction pipeline yields annotations that form the contextualized information space [20], which is then used for further analyses. Domain experts provide feedback on the annotation quality to aid improvements of the underlying methods, as well as the modeling of the relevant stakeholders.

3.5. Visualizations and Analytics

In the last steps, computer scientists draw upon data analytics and visualizations to obtain the results required for verifying or rejecting research hypotheses, to conduct experiments, and to generate insights that are relevant to the target domain. The outcome of this step is not limited to one-time analyses, but may also comprise the creation of expert systems, such as Web Intelligence dashboards, that equip researchers with powerful tools for continuously monitoring relevant web and social media coverage to gain additional insights into stakeholders, new trends, and factors that drive these developments. Finally,

as the case study in Section 8 demonstrates, domain experts are indispensable for interpreting insights generated by data-driven methods and for relating them to the target domain's theoretical framework.

4. Use Case—Analyzing Media Criticism

The following sections demonstrate the application of the introduced process to research in the field of communication science that focuses on media criticism.

As suggested in Section 3, we start by outlining the importance of the topic and the relevant research background within the target discipline. Afterwards, we elaborate on how this background is used to guide the collection of domain knowledge (Section 5), acquire relevant content (Section 6), adapt knowledge extraction methods to the given use case (Section 7), and, finally, create an expert system that is tailored towards performing analyses that help in answering research questions within the domain's theoretical framework.

4.1. Research Background

Mass media play a pivotal role for democratic societies. However, a number of recent national and international debates (e.g., on the media coverage of the 2020 US election and the COVID-19 pandemic) have shown that the performance of mass media is highly contested and trust in journalism is on the decline. A study that was published by Gallup in September 2020 (https://news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx accessed on 21 February 2021) shows that only 9% of U.S. citizens have a great deal of confidence in media reporting, while more than 60% of the respondents described their confidence in news media as "not very much" or "none". The performance of the media is also highly contested in Germany, where the term *Lügenpresse* (lying press) was elected as the worst German word of 2014 (https://www.unwortdesjahres.net/fileadmin/unwort/download/pressemitteilung_unwort2014.pdf accessed on 21 February 2021) and is still used by some ideological groups.

Measuring and understanding the issues, dynamics, and impact of media criticism is a challenging task that can tremendously benefit from systematic studies that cover media criticism from major stakeholders, such as (i) mass media, (ii) media-critical agents, and (iii) social media across different outlets and media. An analysis of the daily output produced by these stakeholders makes it clear that it is no longer feasible to manually collect and analyze these document streams by hand and that a collaborative research design that integrates methods from computer sciences is required.

4.2. Research Framework

Journalism possesses substantial definatory power due to its selection of fragments of reality and the resulting staging of it. However, journalistic descriptions of the reality are at least partly subjective and depend on a number of long-term factors, such as the journalist's socialization, know-how, and self-conception, but also on short-term circumstances, such as the available time for production, events occurring shortly before or after, and the accessible resources (experts, pictures, etc.).

Moreover, citizens depend on trustworthy journalism, which can be achieved by periodic information about current processes in the sector, journalistic work routines, or dominant pressures [21]. Substantial media criticism also empowers the public to overcome its role as an exclusive consumer and enables it to acquire the role of a media-literate agent and citizen who shoulders responsibility for the media system's status quo and quality [22]. Therefore, society cannot go without a continuous, public, and critical debate about journalistic performances [23]. We understand media criticism as optimistic–constructive or negative, public, and reflexive observation, description, and evaluation of media process routines, concerning all relevant participants, referring to accepted rules and standards [21,24]. Crucial is an explicit assessment of a relevant media issue concerning products, agents, actions, or procedures [21,25].

Up to now, academic research lacks a systematic inventory and quantitative empirical basis of editorial-based media criticism, as well as of agents and institutions that practice public media criticism, aside from [26,27]. What is the yearly output of media-critical stakeholders? On which issues do they focus? What resonance do existing agents encounter in mass media? Moreover, little is known from a scientific point of view about the issues, dynamics, and impact of media criticism debates. What seems obvious is that circumstances for editorial-based media critics are deteriorating due to ongoing concentration processes within the media system [28]. Furthermore, on the basis of content analyses, Wyss, Schanne, and Stoffel [27] hint at the often episodic character of media criticism and deplore the absence of explicit evaluations and statements that focus on structural deficits. It remains uncertain whether other agents are able to fill this gap with their own qualitative and systematic coverage. Although the number of critical agents and institutions has increased since digital channels have spread, at least some of them are highly dynamic and show variable lifespans, such as media blogs [29–31]. The societal relevance of media criticism along with recent developments, such as the declining trust in journalistic stakeholders, the structural changes in the media industry, and the neglect of media criticism by communication research [32,33], emphasizes the importance of a systematic scientific analysis of this topic.

4.3. Conclusions

The discussions of research questions, research framework, and the available data analytics capabilities yielded the following insights that directly influenced the project's research design: The dynamics of media criticism are of particular interest to communication science. We, therefore, decided to develop an expert system that acquires, identifies, and monitors media-critical coverage in real time, enabling historical analyses as well as active tracking of current issues. Due to the importance of research questions that focus on the output and impact of different media-critical actors, we have defined three groups from which media-critical coverage will be collected (Section 5.3). The data acquisition component will also draw upon the discipline's definition of media criticism for determining whether an article is relevant for the analysis (Section 6). Finally, the knowledge extraction components (Section 7) will perform named entity linking to identify major stakeholders, phrase detection to automatically obtain associations that provide clues on important topics and the framing behavior of agents, and sentiment analysis to gather insights on whether issues are perceived as positive or negative. Based on these design decisions, the created expert system will allow an efficient and effective analysis of relevant issues, stakeholders, their output, and the impact of their criticism.

5. Collecting Domain Knowledge and Data Sources

5.1. Domain Knowledge on Stakeholders

The research framework emphasizes the importance of gathering qualitative and quantitative insights into (i) the output of media-critical stakeholders and (ii) the issues, dynamics, and impacts of media-critical debates. Consequently, domain experts compiled stakeholder lists that contain media entities (persons and organizations) of all four Swiss language regions, as well as key organizations and persons of foreign countries. Information on international stakeholders was provided because Swiss German media-critical coverage can potentially also refer to entities outside of Switzerland. In addition to active organizations, historical entities were part of the list, as well as existing agents that have not produced any publication output in the last few years. The domain experts also provided background information on entities, such as abbreviations, addresses, and key people. In total, these efforts yielded the object lists outlined in Table 1.

Table 1. Categories of the media-critical entity lists and the corresponding numbers of entities.

Entity Category	Entities
Swiss Stakeholders	1209
Media-critical agents	65
Mass media (print, radio, TV, online)	524
Publishing houses and print offices	85
Publishers, CEOs, and editors in chief	344
Media events and media awards	21
Syndicates and associations	11
Media scholars and (commercial and non-profit) research organizations	50
Foundations and media schools	22
Media politicians and federal councillors	13
Media lawyers	26
Media journalists	48
Foreign stakeholders	39
Selected organizations and persons from foreign countries	39
Miscellaneous	462
Information programs of the Swiss Broadcasting Corporation (SRG)	64
General media-critical keywords (media-relevant terms)	398

In addition, the experts assembled a general keyword list of about four hundred concepts that might indicate media criticism, which contains terms such as journalism, mass media, newspaper, broadcaster, editorial, circulation, tabloid press, practical training, audience rating, or gatekeeper. These terms were manually extracted from a set of media-critical articles that were used by the domain-specificity components (Section 6) and from literature covering communication and media studies. This keyword list was used by the mentioned domain-specificity components as another indicator for media criticism.

5.2. Formalizing Domain Knowledge

The domain experts provided the collected domain knowledge on media-critical stakeholders in a tabular form, specifying information such as the stakeholder's name, possible abbreviations, Twitter accounts, homepages, organization, and type. These tables were converted into the Resource Description Framework (RDF) linked data format, which is more easily interpretable by automated processes and serves as input for the named entity linking component (Section 7.1).

Maali et al. [34] introduced the Publishing Pipeline for Linked Government Data, which utilizes Open Refine (https://openrefine.org/ accessed on 21 February 2021) together with an RDF extension (https://github.com/stkenny/grefine-rdf-extension accessed on 21 February 2021) to convert tabular data into RDF. We simplified this pipeline to contain only the steps "Data clean-up" and "Transformation into RDF", and subsequently adopted it for the targeted lightweight RDF graph used further along in the process.

The main advantages of the outlined procedure are (i) its simplicity and (ii) that the configuration used for the conversion pipeline can be exported, edited, and reused. As such, it is only necessary to create this pipeline once, since it is possible to just reapply it on an updated dataset.

5.3. Selecting Relevant Data Sources

Gaining a comprehensive picture of publicly performed online media criticism requires analyzing the publication output of opinion-leading media-critical agents (press releases, news features, blogs, project reports) and their response rate in relevant mass media and specialized publications focusing on media. In this context, "publicly performed" means mass media online distribution, as well as publications that are accessible on organizational web sites, available for everyone, and, hence, usable cross-systemically [21]. Based

on these insights, the domain experts organized sources of media-critical content into three source categories, as outlined in Table 2: mass media, professional public, and social media.

Table 2. Sources per source type for media-critical content.

Source Type	Sources
Mass media	185
Professional public	170
Social Media	
Media organizations	180
Journalists and media stakeholders	740
Total	1275

1. A total of 185 mass media sources were identified by the experts, comprising web pages of radio stations, TV stations, and printed media with an edition of at least 15,000 copies, as well as online only media. In addition to editions, crucial factors concerning printed media were timeliness, universality, and periodicity. This category also contained well-established TV and radio programs (i.e., programs that have been broadcasted for at least five years).
2. The professional public category gathered articles from 100 Swiss German media-critical agents participating in the public discourse related to media criticism and the corresponding press releases. It included a heterogeneous range of organizations that are either part of the media system (intra-media agents) or belong to another societal system (extra-media agents) [27]. The resulting list contains 170 agents with approximately 100 harvestable URLs.
3. Based on the domain experts' assessment, our research considered two different types of social media accounts: The first one collected tweets from mass media sources, and the second one collected tweets from media- and journalism-related persons. For inclusion in the Twitter sample, profiles needed to fulfill the following three criteria: (i) at least 100 followers, (ii) a relation to Swiss media criticism, and (iii) the majority of the tweets must be written in German. The system monitored 180 Twitter accounts of mass media and 740 accounts of Swiss journalists and media-related persons.

A minor drop-out on the web sources could be determined due to the lack of relevant or up-to-date content, or because content was secured by a paywall or offered exclusively as a podcast, ePaper, or another non-trivial data format. In the case of mass media, program preview sites and audio-only content were removed as well. In the case of paywalls, a subscription for important outlets was organized.

6. Data Acquisition and Preprocessing

Based on the information provided in the previous step, our content acquisition pipeline drew upon web crawlers and the Twitter web API to retrieve potentially relevant content. Since between 50 and 60% of a typical web page consists of noise, such as navigation menus, links to related documents, advertisements, and copyright notes, the content acquisition processes also deploy boilerplate removal [35] to identify and remove noise elements as well as overview pages (i.e., summarization pages and entry points).

Afterwards, a domain-specificity classifier ensures that only relevant documents are included in the corpus (or information space) by filtering irrelevant documents. The information space used for analyzing media criticism, therefore, will contain mostly media-critical documents rather than arbitrary media coverage. The following sections describe the evolution of the domain-specificity component within the Swiss Media Criticism project.

6.1. Keyword-Based Approach

The first version of the content acquisition pipeline drew upon black- and whitelist items to determine the domain specificity of documents. Keyword-based queries are very

common in social sciences and have the advantage of being well accepted within this discipline. Nevertheless, they provide low performance in terms of recall, since media criticism can be voiced in manifold ways and settings, ranging from sport events like the ski accident of Michael Schuhmacher in 2013, to tragic disasters in aviation like the Germanwings accident in 2015, to political statements like the Böhmermann affair in 2016, and to public health and policy issues, such as the reporting on the COVID-19 crisis. Knowledge-aware text classification systems address this problem by considering both the document's vocabulary and background knowledge, such as the presence of media-critical entities and terms in the classification process.

6.2. Knowledge-Aware Text Classification

The next iteration of the domain-specificity component drew upon the formalized domain knowledge to calculate the probabilistic affiliation of a new and unknown text with a given set of categories. In the presented project, two categories—media criticism and not media criticism—were used, corresponding to relevant and irrelevant content.

To gather the needed domain-specificity information, domain experts (i.e., communication scientists) collected a gold-standard corpus of both media-critical and non-media-critical documents. Special attention was paid to finding document pairs dealing with the same topic, where one document contained media criticism while the other one did not, as this helped to minimize the risk of topic-specific bias. In addition, the domain experts aimed to include a wide range of topics, such as sports, direct democracy, affirmative actions, disabilities, the Holocaust, offshore leaks, the Pope, Islam, climate change, the energy transition, and airplane accidents, in the gold standard. All selected documents were published in a Swiss medium after 2010. The media-critical documents in the sample only considered texts for which media criticism was clearly identified as the main topic and was not just peripherally mentioned. The final gold standard comprised 503 media-critical documents and 643 corresponding non-media-critical counterparts, which were used for training the classifier.

This iteration used the Naïve Bayes classification algorithm due to its simplicity and explainability, which allowed the determination of the reasons for a correct (or incorrect) classification result. Being able to observe and correct the process if necessary helped in building the domain expert's trust in the classifier, and this was therefore identified as a crucial step in creating the system.

The overall probability of an unknown text was derived from the probabilities of the contained terms: First, the document ws converted into a "bag of words" representation that also considered n-grams and skip-grams. A stopword filter and a frequency-based filter, which remove terms that have been used less than three times in all collected documents, were applied to speed up the computation time. The prior probability for each element of this bag of words was received from the knowledge base calculated previously. The overall probability of an article being media critical given its included terms $P(M|t)$ was calculated with the cross-product of its terms' prior probability $P(M)$ and the likelihood $P(t|M)$ that they were contained divided by its evidence $P(t)$.

$$P(M|t) = \frac{priorprobability \times likelihood}{evidence} = \frac{P(M)P(t|M)}{P(t)} \quad (1)$$

In this scenario, prior probability describes the general probability of an unknown text containing media criticism according to the training data, while likelihood refers to the probability of an unknown text's term being related to a media-critical text. Meanwhile, evidence means the probability of an unknown text's terms being contained in any of the trained categories (in this cas,e media criticism M and not media criticism $\neg M$).

$$evidence = P(t) = P(M)P(t|M) + P(\neg M)P(t|\neg M) \quad (2)$$

We further introduced ensemble heuristics, where the classification functions as a baseline, and further knowledge sources, such as (i) source whitelists, (ii) black- and whitelists, and (iii) named entity annotation and text patterns, function as regulators. Documents that contain no or only a few keywords received a penalty on the estimated probability of containing media criticism. The ensemble methods applied in this iteration achieved results with a recall of over 71%, which is already considerably better than the keyword-based approach.

6.3. Deep Learning for Text Classification

The final iteration of the domain-specificity component drew upon a transformer architecture to perform the classification process. As outlined in Figure 2, the classifier used a Bidirectional Encoder Representations from Transformers (BERT) language model [36] with twelve transformer layers and a maximum sequence length of 512 tokens to create a contextualized representation of the input document, in which each token was represented by a 768-dimensional vector. Text classification tasks usually only consider the BERT symbol for classification output (`[CLS]`), which starts each document and provides a 768-dimensional vector representation of its content. Since combining multiple hidden layers has been shown to improve accuracy [36], we pooled the output of the last four hidden layers and fed it through two linear layers, which then provided the final classification result.

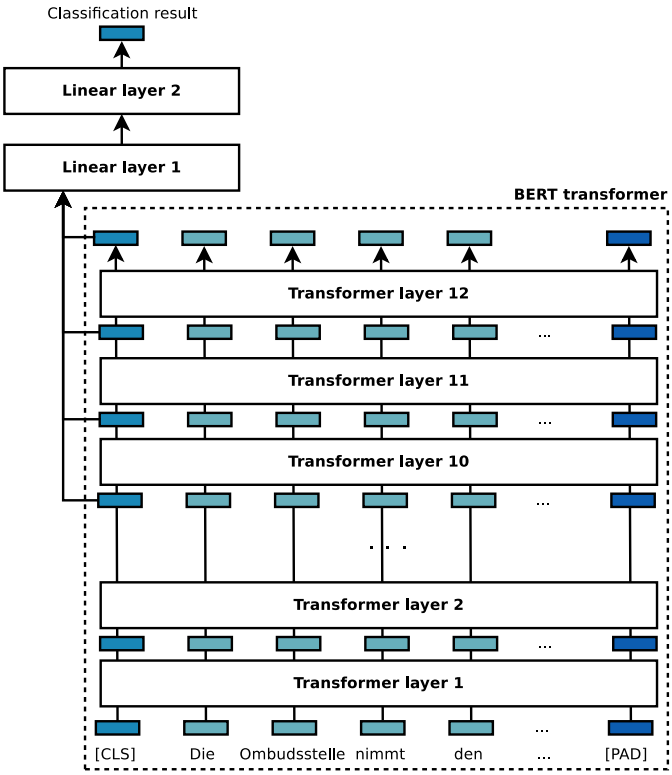

Figure 2. Architecture of the deep learning classifier.

We used the Hugging Face transformer library (https://huggingface.co accessed on 21 February 2021) in conjunction with PyTorch (https://pytorch.org accessed on 21 February 2021) to implement the classifier and evaluated it using the following two pre-trained transformer models:

- German BERT base (https://huggingface.co/bert-base-german-cased accessed on 21 February 2021): A case-sensitive BERT transformer that has been trained on over 10 GB of textual data comprising the German Wikipedia dump (6 GB), the OpenLegalData dump (2.4 GB), and 3.6 GB of news articles.

- Multilingual BERT (https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment accessed on 21 February 2021): A case-insensitive, multilingual BERT model that has been fine-tuned for sentiment analysis on product reviews in English, Dutch, German, French, Spanish, and Italian.

Experiments that aimed at optimizing the model's hyperparameters yielded the following settings: a batch size of four, a dropout value of 0.3 to prevent overfitting, an Adam optimizer with a learning rate of $\alpha = 3 \times 10^{-5}$, a binary cross-entropy (BCE) loss function, and a training limit of 35 epochs with early stopping if the validation loss does not improve.

Both models were fine-tuned on the domain-specificity classification task using the gold standard created by our domain experts. In our experiments, the German BERT Base model achieved an F1 score of 92.5% with both a precision and a recall of 92.5%. The multilingual BERT model even topped this performance with an F1 score of 95.8% (recall: 95.6%, precision: 95.8%). This considerable boost in performance came at the cost of a lower explainability of the provided classification results.

7. Knowledge Extraction

The selection of the knowledge extraction techniques was guided by the research framework discussed in Section 4, which requires (i) the analysis and tracking of major stakeholders, (ii) the identification of dominant issues, and (iii) the classification of media criticism as either optimistic–constructive or negative. Weichselbraun et al. [37] discussed the use of domain-specific affective models, which support capturing emotions that go beyond standard sentiment and emotion models. An assessment that compared the optimistic–constructive and negative dimensions from communication science literature with standard sentiment polarity (dimensions: positive and negative) concluded that, for the purpose of the joint research project, the use of sentiment polarity is an efficient (availability of high-quality sentiment lexicons) and effective (sufficiently high correlation between both metrics) strategy.

Consequently, we deployed named entity linking for identifying stakeholders (Section 7.1), the phrase extraction method described by Weichselbraun et al. [38] for tracking associations and dominant issues, and sentiment analysis (Section 7.2) to automatically determine whether the feedback was positive or negative.

7.1. Named Entity Linking

Named entity linking identifies mentions of named entities, such as persons or organizations that are important stakeholders in the public discourse on media criticism, as well as locations, and links them to structured knowledge sources, such as the DBpedia, GeoNames, and custom linked open data repositories. Therefore, it paves the way for analytics that assign sentiments to entities, identify trends, and reveal relations between these entities. Transforming the domain knowledge on media criticism assembled by the domain experts to a linked data format (Section 5.2) enables us to leverage these data for named entity linking.

The webLyzard platform used in this project draws upon Recognyze [39], a named entity linking component that queries linked open data sources to obtain textual, contextual, and structural information on entities, which is then used to identify entity mentions in text documents. Figure 3 outlines this process. Recognyze uses analyzers, i.e., graph mining components, that retrieve the relevant information from the available linked data sources.

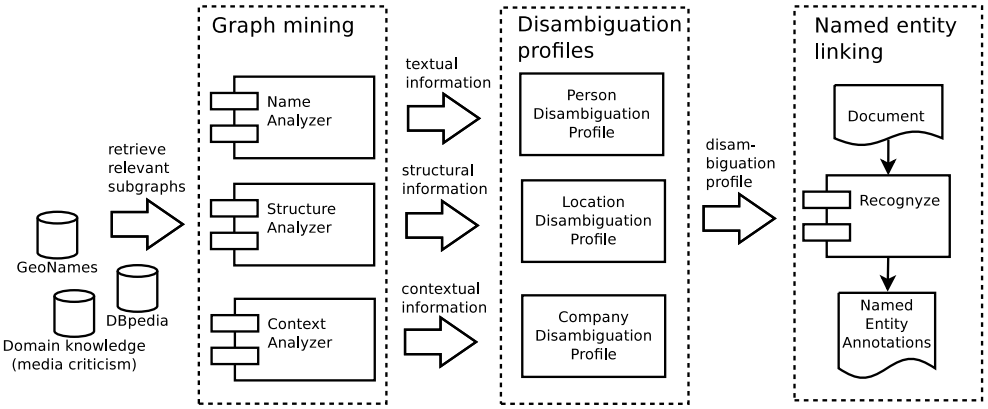

Figure 3. Named entity linking with Recognyze.

Name analyzers yield textual information that comprises different named variants and abbreviations used to refer to an entity (e.g., BBC, British Broadcasting Corporation, etc.). Context analyzers mine contextual information, such as keywords obtained from the organization's description, products and services offered by an organization, its web address, and mail and email addresses. Structural information is obtained by structure analyzers, which reveal relations between the extracted entities—for instance, that John Reith was the founder of the BBC or that Future Media is an operational division of the BBC.

The obtained information was then used to create disambiguation profiles that identified potential named entity mentions in textual content, thus helping to disambiguate these mentions and ground them to the correct entity in the linked open data repository.

7.2. Sentiment Analysis

Sentiment analysis computes the polarity (positive versus negative) of targets, such as documents, sentences, topics, and named entities. Therefore, it is useful to determine how targets are perceived by the public or to identify conflicting targets, i.e., targets with a high standard deviation of the sentiment value, which indicates that the coverage of these targets is framed differently in the analyzed outlets.

The Swiss Media Criticism portal uses a context-aware sentiment analysis approach to determine the text sentiment, which considers the text's context prior to evaluating its sentiment. Therefore, it combines a static sentiment lexicon that contains sentiment terms t_i that either indicate a positive or negative sentiment (e.g., good and excellent versus bad and horrible) with a contextualized sentiment lexicon, which determines the sentiment value for ambiguous sentiment terms ($t_i \in T_{ambig}$) based on the text's context C_i, expressed as a set of non-sentiment terms within the text. The term expensive, for instance, is considered negative in conjunction with context terms such as overpriced, while the context terms high value and quality might indicate a positive usage of this term.

Aggregating the sentiment value of all terms t_i within a sentence with context C_i yields the total text's sentiment (Equation (3)).

$$s_{\text{sentence}} = \sum_{i=1}^{n} \mathcal{N}(t_i) \cdot s''(t_i, C_i) \tag{3}$$

The sentiment lexicon $s''(t_i, C_i)$ yields the contextualized sentiment value for ambiguous sentiment terms t_i and falls back to the static sentiment lexicon $s(t_i)$ for unambiguous sentiment terms or if no context terms are available.

$$s''(t_i, C_i) = \begin{cases} s'(t_i, C_i) & \text{if } t_i \in T_{\text{ambig}} \text{ and } C_i \neq \emptyset \\ s(t_i) & \text{otherwise} \end{cases} \tag{4}$$

A function $\mathcal{N}(t_i)$ considers negations by inverting the term sentiment if t_i has been negated.

$$\mathcal{N}(t_i) = \begin{cases} -1 & \text{if } t_i \text{ is negated} \\ +1 & \text{otherwise} \end{cases} \quad (5)$$

The system also propagates the sentence sentiment to topics, sources, and entities, i.e., it allows assessment of whether a certain entity occurs frequently in a positive or negative context or how a certain new media site frames a particular topic.

8. Visualizations and Analytics—The Swiss Media Criticism Portal

The expert system that was created, the "Swiss Media Criticism portal and analytics dashboard", was developed within the Radar Media Criticism Switzerland project, funded by the Swiss National Science Foundation, and was built upon the webLyzard platform (https://www.weblyzard.com accessed on 21 February 2021), which provides components for scalable knowledge acquisition and extraction [40], advanced analytics [16], and visualizations [6]. The expert system has been actively used by communication scientists from the project consortium, and efforts towards extending this system to further countries, domains, and research groups are planned.

Figure 4 illustrates the Swiss Media Criticism portal and analytics dashboard, which enables users to search, refine, analyze, and interpret media-critical coverage from Swiss online sources.

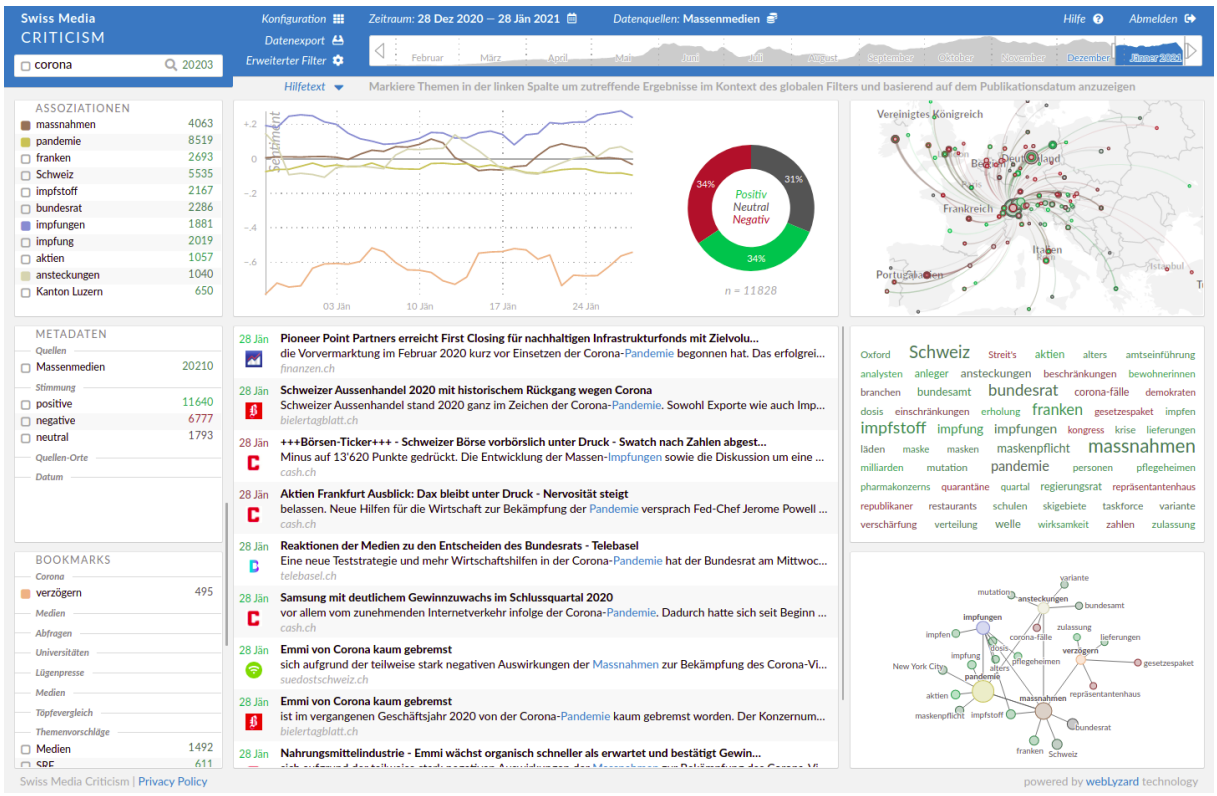

Figure 4. Screenshot of the Swiss Media Criticism portal featuring a query for the search term "corona".

The system aggregates media-critical content and yields keyphrase statistics (left bottom), relevant documents (center), real-time analytics, such as the geographic distribution of the media coverage (right top), a tag cloud highlighting terms associated with the current query (right middle), and a keyword graph outlining how the associated terms are connected to each other (right bottom). The case study in the following section outlines how the computed associations and the corresponding visualizations provide clues con-

cerning the framing behavior of the agents and support the development of hypotheses that describe the effects that the coverage has among the target audience.

The created expert system provides powerful analytics for addressing the major issues and questions raised within the research framework:

- The platform provides real-time coverage of current (e.g., Figure 4) and past (e.g., Section 8.1) issues, thus enabling experts to analyze the dynamics of both current and past issues and to participate in the discourse with data-driven insights.
- Stakeholder groups (i.e., mass media, professional public, and social media) are organized in different samples, allowing domain experts to analyze and contrast the output and impact of these groups.
- Named entity linking enables analyses that focus on the stakeholder groups that have been defined by the domain experts (Section 5).
- Phrase extraction automatically identifies terms and concepts that are associated with stakeholders, locations, and queries, supporting domain experts in uncovering important topics, dominant issues, and their framing.
- Sentiment analysis provides insights into the perception of stakeholders and issues.
- Drill-down analyses ensure that aggregated results and trends presented in the portal are valid and help in understanding the underlying reasons for the observed effects.

The following case study demonstrates how these analytics and visualizations are instrumental in understanding the media-critical discourse and its dynamics by supporting analyses that (i) aid in answering fundamental questions on the temporal course and lifespan of issues, (ii) highlight important agents participating in the discourse, and (iii) identify the sentiment (positive versus negative) of its coverage.

8.1. Case Study: New Year's Eve Sexual Assaults in Cologne

During the night of New Year's Eve in 2015/2016, numerous women were robbed and sexually molested at the Cologne main station. Generally, the suspects were described as African- or Arabic-looking. A public debate emerged over immigrants, sexism, and cultural values. In this context, the role of mass media coverage of delicate topics was critically discussed, and the term *Lügenpresse* (lying press) was suddenly socially acceptable again.

Searching for media-critical coverage of these events using the keywords *Silvesternacht* (night of New Year's Eve), *Köln* (Cologne), and *Medien* (media) during the period from 31 December 2015 to 31 March 2016 yielded 72 documents in the mass media and professional public categories, as well as 227 Tweets from media-critical stakeholders.

Figure 5 compares the media-critical coverage in mass media and in Tweets authored by media-critical stakeholders.

For mass media, the most relevant concepts were *Köln* (Cologne, 36 mentions), *Täter* (offenders, 36 mentions), and *Übergriffe* (assaults, 16 mentions), indicating that the discourse was focused on these particular events.

The discussion of media-critical stakeholders, in contrast, focuses on the implications of the events for Switzerland, as indicated by the keywords *Frauen* (women, 17 mentions), *Schweiz* (Switzerland, 12 mentions), and *Durchsetzungsinitiative* (an upcoming people's vote regarding the deportation of criminal foreign citizens, nine mentions). It is also interesting to note that the term "refugeeswelcome", which was present only in the Twitter tag cloud, became a hashtag for some of these discussions.

The geographic distribution of the locations mentioned in the articles indicates that the coverage focused on Cologne, Stuttgart, and Sweden, where similar events were reported, as well as in Poland due to an article that covered the drastic means by which Poland's government tried to strengthen its influence on the media, mentioning the Polish media coverage of the events in Cologne.

Figure 6 reveals the topic's life cycle, which seems to be typical for mass media coverage, with a heavy increase in coverage when the issue first became apparent three days after New Year's Day. The mostly episodic mass media coverage, which has been criticized

by media scholars, seemed to continue as far as media-critical coverage is concerned. Media critics seem to operate similarly to their colleagues.

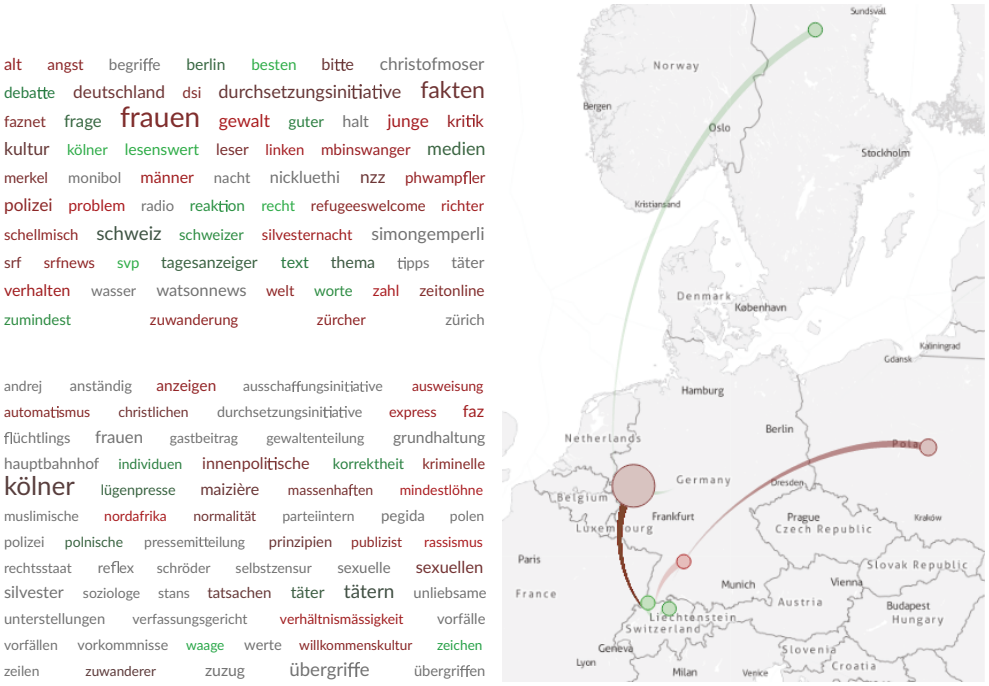

Figure 5. Tag clouds of the media-critical coverage of the New Year's Eve sexual assaults in Cologne on Twitter (**top**), in news media (**bottom**), and in locations mentioned in the news media coverage (**right**).

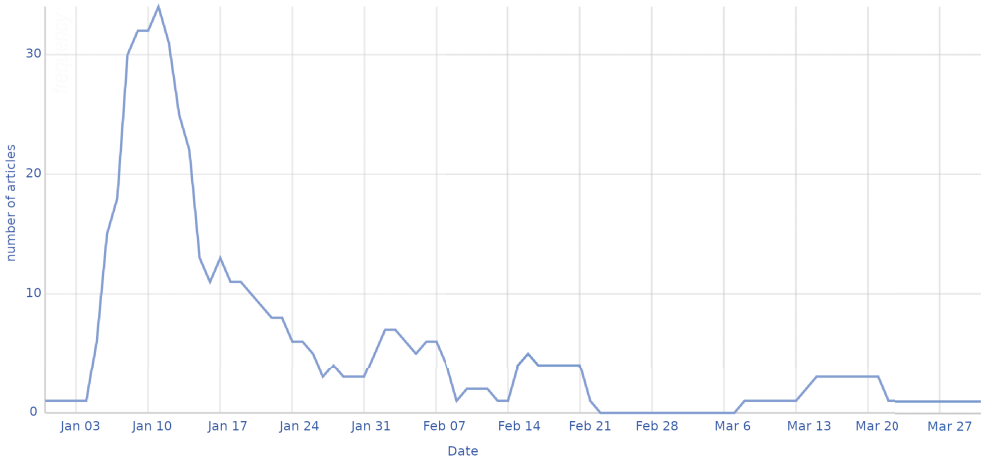

Figure 6. Frequency graph illustrating the topic's life cycle in mass media in the first quarter of 2016.

After a fortnightly high during which the issue was dominant, the media lost interest, as hardly any new information became available or there was a lack of similar follow-up events. The steep decline was followed by a steady loss of importance, which ended in disappearance. When considering that almost three hundred monitored agents only yielded 72 hits over three months, it became obvious that this event did not attract a lot of media-critical coverage in Switzerland. Hence, the hypothesis of media-critical coverage not being widespread is supported. Swiss journalists held themselves back from criticizing their professional colleagues in Germany, but also omitted a discussion of what should be done differently by Swiss media if a comparable event would happen in Switzerland.

At the same time, the system revealed which sources participated in a public debate and how strongly each source had been involved (Figure 7). From a media studies perspective, this feature is relevant concerning the distinction of mass media with institutionalized

forms of media criticism and editorial teams without similar structures or for recognizing differences between tabloid press and quality media. Institutionalization means that there is at least a personal specialization or an organizational beat in the topic field of media journalism. According to structuration theory, institutionalization of media criticism should lead to more regular and more qualitative coverage [41]. Moreover, the coverage's geographic distribution becomes visible; therefore, it is also more discernible if one regional press cooperation reports more often than another.

The deepening analysis in Table 3 points out that only 22 of the 185 mass media titles covered the event from a media-critical point of view. This corresponds to an average of 0.31 contributed articles per title in the category of the mass media. The reaction of the professional media in this case was even lower. Only 8 of the 100 professional public media titles covered the issue. This corresponds to an average of 0.13 contributed articles per title in the category of the professional public. This result shows that mass media criticism cannot be fully substituted with criticism by professional public agents. Moreover, the analysis allows a comparison between tabloid papers and quality press. In the sample, we found 24 articles coming from five quality press titles, but only five articles in five tabloid press titles. We could argue that substantial media criticism is not something that tabloid press is predestined for due to its meta-level nature and its complex context with ethics or law. Therefore, media criticism is a topic that tabloids seem to avoid, fearing they could perform poorly.

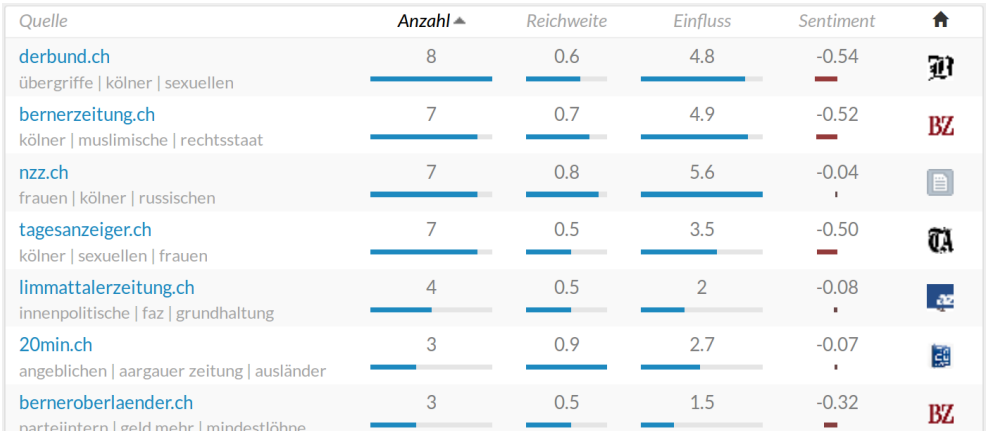

Figure 7. Cropped screenshot of the Swiss Media Criticism portal demonstrating the most frequent sources of the selected articles. Included are the source name (*Quelle*), number of documents (*Anzahl*), reach (*Reichweite*), influence (*Einfluss*), and sentiment.

Table 3. Average number of articles per medium.

Category	Articles (Agents)	Average per Medium
Top Categories		
Mass media	59 (185)	0.31
Professional public	13 (100)	0.13
Sub-Categories		
Quality media	24 (5)	4.80
Tabloid press	5 (5)	1.00
Institutionalized media criticism	17 (5)	3.40

The data show that forms of institutionalized media criticism structures within the newsroom do indeed lead to more coverage. In the category of the media with institutionalized media criticism, we found an average of 3.4 contributed articles per title, whereas editorial teams without such specialization published far fewer articles.

Although the analysis shows certain tendencies, it is clear that one case study is not enough to determine whether institutionalization is meaningful or not. Just by skimming the articles, it becomes evident that the expression "lying press" is hardly existent in Swiss mass media coverage. The Swiss journalists do not seem to be willing to support the polemical use of this inadequate term. In addition, mass media only perform "embedded media criticism", which means that critical content is not the main topic of an article, but is peripherally mentioned in a few sentences.

Figure 8 shows the sentiment for the press coverage of the New Year's Eve sexual assaults in Cologne. The figure indicates the framing behavior of the involved agents—more precisely, whether an issue is mainly framed with positive, neutral, or negative terms. This allows, for example, the creation of hypotheses regarding the effects of this coverage on the audience.

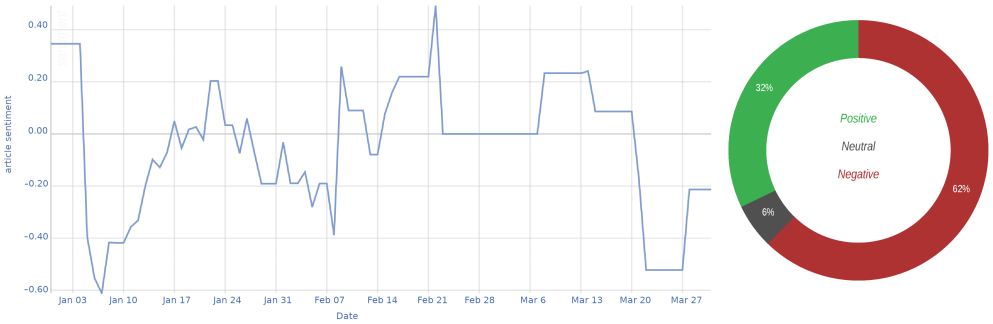

Figure 8. Sentiment analysis of the topic's perception in the first quarter of 2016.

In the case of Cologne, once more, we recognized a typical pattern known from general media coverage: There is often a first outrage wave among journalists when such an issue emerges. A highly emotionalized coverage arises, leading to a very negative issue sentiment. Indeed, most of the media-critical Cologne coverage was clearly negatively framed. After a few weeks, the sentiment becomes more neutral, as journalists gain distance from an event and contextualize an incident. Later, the sentiment even becomes positive, as voices often arise that demand that society learns and improves and that things have to be handled differently the next time in order to avoid a similar event. After this conclusion phase, the sentiment stays neutral as media coverage vanishes. If there is a new aspect to be released to the public or a new event occurs that can be associated with the previous one, the neutral phase ends and is replaced by more emotional coverage again.

8.2. Validation of the Results

Validation of automated methods plays a key role in ensuring a high data quality and in obtaining reliable results and insights. Within the Swiss Media Criticism project, we implemented multiple validation routines that have been tailored to meet the specific needs of the domain experts' use cases, some being fully automated tests, and others requiring heavy human interaction.

For instance, the recall of the data acquisition pipeline (i.e., whether all available media-critical coverage can be found in the expert system) was of particular importance. Therefore, we complemented standard automatized tests that computed the precision, recall, and F1 measure of the domain-specificity components with manual checks, in which domain experts scanned newspapers for relevant articles and verified that the articles were also actually available in the portal.

To evaluate information extraction tasks, such as named entity linking and sentiment analysis, we introduced group exercises into lectures and tasked students with evaluating data quality in seminars. These efforts were complemented by evaluations performed by computer scientists using tools such as Orbis [42] that aid validations of computer-generated annotations by visually comparing them with gold-standard annotations.

The feedback received from the validation steps was collected in an issue management system, which allowed all stakeholders to track the improvements in data quality and reliability, and has been proven to strengthen participation and involvement.

8.3. Discussion of the Case Study

For newly evolving discourses, the Swiss Media Criticism portal provides an automated quantitative overview for crucial parameters and answers to basic research questions in almost real time. What is the temporal course and lifespan of an issue? Which agents participate to which extent in an ongoing discourse? How is an issue framed and which sentiment does it encounter during coverage? In addition, the portal instantly delivers a sample of articles that can easily be used and modified for more in-depth, manual content analysis, as search results are exportable to various data formats.

Several benefits can be highlighted from the application of data-driven research methods to the domain of media criticism: (i) availability of a complete inventory, such as extensive document repositories, (ii) volatile sources of information (such as Twitter and comments in forums) can be captured, and (iii) the efficiency of the content analysis is much higher than with conventional methods.

In addition, the system provides powerful analytics and visualization to gain a better understanding of the target domain, of spatial and temporal effects, and of the framing of topics due to the computation of associations and the sentiment. The Swiss Media Criticism portal also supports exporting documents and visualizations, enabling the application of further linguistic and statistical methods to the document corpus.

9. Conclusions

Recent literature has raised serious concerns about the effectiveness of computer scientists that apply methods from their field to other disciplines without a proper understanding of the necessary research background within the target domain [1,4]. This article addresses these concerns by suggesting a strong collaborative adaptation process that has been developed within a number of research projects and that aids interdisciplinary research groups in building a common understanding of (i) the research framework within the target discipline and (ii) the benefits that data-driven research methods could yield. The process guides researchers in:

- Aligning their research design and hypotheses with the target discipline's research framework and background to ensure that the envisioned research has a real impact on that discipline;
- Leveraging theories and concepts from the target domain in the design of indicators and metrics;
- Considering the target domain's research framework in the development of the entire data acquisition, processing, and analytics process and integrating domain expert input (e.g., on formalized domain knowledge and data sources) into their development;
- Organizing the collaboration between the groups by defining interfaces and feedback loops.

We then applied this approach to the field of communication science and discussed its impact on the design of the whole data-driven research process. Our experiences with the domain-specificity component, which was improved based on multiple feedback loops, demonstrate that iterative feedback and refinements—for example, of system components—are crucial for ensuring success. The presented process also guides the evolution of the system in a controlled and structured way, enabling computer scientists and domain experts to systematically monitor the impact of each feedback loop and to determine when a sufficient quality has been reached.

A case study that drew upon the developed system demonstrated how a close alignment between the target discipline and computer scientists helps in creating research designs that yield insights of high relevance to the target domain. The Swiss Media Criticism portal, for example, provided sophisticated analytics that helped communication

scientists in identifying, tracking, and understanding public media criticism debates. An analysis that would take weeks with conventional means can be performed in close to real time, enabling the research team to provide continuous reports on media-critical events and to investigate temporal effects. The alignment of data-driven research methods with the research framework from communication science also ensures that the obtained results adhere to the target domain's scientific standards and yield relevant contributions to this field. Computer scientists also benefit from this process, since the interdisciplinary approach provides them with insights into the target domain's research frameworks and with innovative views on the strengths and weaknesses of their technology, since each scientific discipline poses its own typical questions for its objects of investigation, and technology cannot answer all questions ad hoc.

Author Contributions: Conceptualization, A.W., P.K., M.S., U.D. and V.W.; Data curation, P.K., V.F. and M.S.; Funding acquisition, A.W., U.D. and V.W.; Methodology, A.W., U.D. and V.W.; Project administration, U.D. and V.W.; Resources, V.F. and M.S.; Software, A.W. and P.K.; Supervision, A.W., U.D. and V.W.; Visualization, P.K. and V.F.; Writing—original draft, A.W., P.K., V.F., M.S., U.D. and V.W.; Writing—review & editing, A.W., P.K., V.F., M.S., U.D. and V.W. All authors have read and agreed to the published version of the manuscript.

Funding: The work presented in this paper was conducted as part of the "Radar Media Criticism Switzerland" project funded by the Swiss National Science Foundation under the project number: 150327.

Data Availability Statement: Publicly available datasets have been created and utilized within this study. These data can be found at https://github.com/media-criticism/swiss-dataset. The gold standard dataset used for the training and evaluation of the domain-specificity component contains documents that are copyrighted by third parties and, therefore, cannot be legally redistributed in Switzerland.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Connolly, R. Why computing belongs within the social sciences. *Commun. ACM* **2020**, *63*, 54–59, doi:10.1145/3383444.
2. Susanto, Y.; Livingstone, A.; Ng, B.C.; Cambria, E. The Hourglass model revisited. *IEEE Intell. Syst.* **2020**, *35*, 96–102.
3. Ekman, P. An Argument for Basic Emotions. *Cogn. Emot.* **1992**, *6*, 169–200, doi:10.1080/02699939208411068.
4. Bartlett, A.; Lewis, J.; Reyes-Galindo, L.; Stephens, N. The locus of legitimate interpretation in Big Data sciences: Lessons for computational social science from -omic biology and high-energy physics. *Big Data Soc.* **2018**, *5*, 2053951718768831. doi:10.1177/2053951718768831.
5. Ranganath, S.; Hu, X.; Tang, J.; Liu, H. Understanding and Identifying Advocates for Political Campaigns on Social Media. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16), San Francisco, CA, USA, 22–25 February 2016.; ACM: New York, NY, USA, 2016; pp. 43–52, doi:10.1145/2835776.2835807.
6. Scharl, A.; Herring, D.; Rafelsberger, W.; Hubmann-Haidvogel, A.; Kamolov, R.; Fischl, D.; Föls, M.; Weichselbraun, A. Semantic Systems and Visual Tools to Support Environmental Communication. *IEEE Syst. J.* **2017**, *11*, 762–771, doi:10.1109/JSYST.2015.2466439.
7. Khatua, A.; Cambria, E.; Ho, S.S.; Na, J.C. Deciphering Public Opinion of Nuclear Energy on Twitter. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8, ISSN: 2161-4407, doi:10.1109/IJCNN48605.2020.9206903.
8. Cavalcante, R.C.; Brasileiro, R.C.; Souza, V.L.F.; Nobrega, J.P.; Oliveira, A.L.I. Computational Intelligence and Financial Markets: A Survey and Future Directions. *Expert Syst. Appl.* **2016**, *55*, 194–211, doi:10.1016/j.eswa.2016.02.006.
9. Xing, F.Z.; Cambria, E.; Welsch, R.E. Natural language based financial forecasting: A survey. *Artif. Intell. Rev.* **2018**, *50*, 49–73, doi:10.1007/s10462-017-9588-9.
10. Yang, D.; Huang, C.; Wang, M. A social recommender system by combining social network and sentiment similarity: A case study of healthcare. *J. Inf. Sci.* **2017**, *43*, 635–648, doi:10.1177/0165551516657712.
11. Xiao, S.; Wei, C.P.; Dong, M. Crowd intelligence: Analyzing online product reviews for preference measurement. *Inf. Manag.* **2016**, *53*, 169–182, doi:10.1016/j.im.2015.09.010.
12. Chung, W.; Zeng, D. Social-media-based public policy informatics: Sentiment and network analyses of U.S. Immigration and border security. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 1588–1606, doi:10.1002/asi.23449.

13. Scharl, A.; Herring, D.D. Extracting Knowledge from the Web and Social Media for Progress Monitoring in Public Outreach and Science Communication. In Proceedings of the 19th Brazilian Symposium on Multimedia and the Web (WebMedia'13), Salvador, Brazil, 5–8 November 2013; ACM: New York, NY, USA, 2013; pp. 121–124, doi:10.1145/2526188.2526219.
14. Napoli, P.M. Social TV Engagement Metrics: An Exploratory Comparative Analysis of Competing (Aspiring) Market Information Regimes. *SSRN Electron. J.* **2013**, doi:10.2139/ssrn.2307484.
15. Wakamiya, S.; Lee, R.; Sumiya, K. Towards Better TV Viewing Rates: Exploiting Crowd's Media Life Logs over Twitter for TV Rating. In Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication (ICUIMC'11), Seoul, Korea, 21–23 February 2011 ; ACM: New York, NY, USA, 2011; pp. 39:1–39:10, doi:10.1145/1968613.1968661.
16. Scharl, A.; Hubmann-Haidvogel, A.; Jones, A.; Fischl, D.; Kamolov, R.; Weichselbraun, A.; Rafelsberger, W. Analyzing the Public Discourse on Works of Fiction—Automatic Emotion Detection in Online Media Coverage about HBO's Game of Thrones. *Inf. Process. Manag.* **2016**, *52*, 129–138, doi:10.1016/j.ipm.2015.02.003.
17. Li, Q.; Wang, T.; Li, P.; Liu, L.; Gong, Q.; Chen, Y. The effect of news and public mood on stock movements. *Inf. Sci.* **2014**, *278*, 826–840.
18. Xing, F.; Malandri, L.; Zhang, Y.; Cambria, E. Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; International Committee on Computational Linguistics: Barcelona, Spain (Online), 2020; pp. 978–987.
19. Kim, E.H.J.; Jeong, Y.K.; Kim, Y.; Kang, K.Y.; Song, M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *J. Inf. Sci.* **2016**, *42*, 763–781, doi:10.1177/0165551515608733.
20. Hubmann-Haidvogel, A.; Scharl, A.; Weichselbraun, A. Multiple Coordinated Views for Searching and Navigating Web Content Repositories. *Inf. Sci.* **2009**, *179*, 1813–1821, doi:10.1016/j.ins.2009.01.030.
21. Malik, M. *Journalismusjournalismus. Funktion, Strukturen und Strategien der journalistischen Berichterstattung*; Springer: Berlin/Heidelberg, Germany, 2004.
22. Wyss, V.; Keel, G. Media Governance and Media Quality Management: Theoretical Concepts and an Empirical Example from Switzerland. In *Press Freedom and Pluralism in Europe: Concepts and Conditions*; Intellect: Bristol, TN, USA; Chicago, IL, USA, 2009; pp. 115–128.
23. Sutter, T. *Medienanalyse und Medienkritik. Forschungsfelder einer Konstruktivistischen Soziologie der Medien*; VS Verlag: Wiesbaden, Germany, 2010.
24. Schmidt, S.J. Zur Grundlegung einer Medienkritik. In *Neue Kritik der Medienkritik. Werkanalyse, Nutzerservice, Sales Promotion oder Kulturkritik*; Herbert von Halem Verlag: Köln, Germany, 2005; pp. 21–40.
25. Scodari, C.; Thorpe, J. *Media Criticism. Journeys in Interpretation*; Kendall Hunt Publishing: Dubuque, IA, USA, 1993.
26. Meier, C.; Weichert, S. *Basiswissen für die Medienpraxis. Journalismus Bibliothek 8*; 2012. Available online: https://www.halem-verlag.de/wp-content/uploads/2012/09/9783869620237_inhalt.pdf (accessed on 21 February 2021).
27. Wyss, V.; Schanne, M.; Stoffel, A. Medienkritik in der Schweiz—Eine Bestandsaufnahme. In *Qualität der Medien. Schweiz-Suisse -Svizzera. Jahrbuch 2012*; Schwabe: Basel, Switzerland, 2012; pp. 361–376.
28. Puppis, M.; Schönhagen, P.; Fürst, S.; Hofstetter, B.; Meissner, M. Arbeitsbedingungen und Berichterstattungsfreiheit in Journalistischen Organisationen. Available online: https://www.bakom.admin.ch/dam/bakom/de/dokumente/2014/12/journalistenbefragungimpressum.pdf.download.pdf/journalistenbefragungimpressum.pdf (accessed on 21 February 2021).
29. Eberwein, T. *Raus aus der Selbstbeobachtungsfalle! Zum medienkritischen Potenzial der Blogosphäre*; Springer: Berlin/Heidelberg, Germany, 2008.
30. Eberwein, T. *Typen und Funktionen von Medienblogs*; Springer: Berlin/Heidelberg, Germany, 2008.
31. Eberwein, T. Von "Holzhausen" nach "Blogville"—Und zurück. Medienbeobachtung in Tagespresse und Weblogs. In *Journalismus und Öffentlichkeit. Eine Profession und ihr gesellschaftlicher Auftrag*; Festschrift für Horst Pöttker; VS Verlag: Wiesbaden, Germany, 2010; pp. 143–165.
32. Kleiner, M.S. *Einleitung; Grundlagentexte zur sozialwissenschaftlichen Medienkritik*; VS Verlag: Wiesbaden, Germany, 2010; pp. 13–85.
33. Russ-Mohl, S.; Fengler, S. *Medien auf der Bühne der Medien. Zur Zukunft von Medienjournalismus und Medien-PR*; Dahlem University Press: Berlin, Germany, 2000.
34. Maali, F.; Cyganiak, R.; Peristeras, V. A Publishing Pipeline for Linked Government Data. In Proceedings of the 9th Extended Semantic Web Conference, Heraklion, Greece, 27–31 May 2012.
35. Lang, H.P.; Wohlgenannt, G.; Weichselbraun, A. TextSweeper—A System for Content Extraction and Overview Page Detection. In Proceedings of the International Conference on Information Resources Management (Conf-IRM), Vienna, Austria, 21–23 May 2012.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
37. Weichselbraun, A.; Steixner, J.; Brasoveanu, A.M.P.; Scharl, A.; Göbel, M.; Nixon, L.J.B. Automatic Expansion of Domain-Specific Affective Models for Web Intelligence Applications. *Cogn. Comput.* **2021**, doi:10.1007/s12559-021-09839-4.
38. Weichselbraun, A.; Scharl, A.; Gindl, S. Extracting Opini Targets from Environmental Web Coverage and Social Media Streams. In Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS-49), Koloa, HI, USA, 5–8 January 2016; IEEE Computer Society Press: Los Alamitos, CA, USA, 2016.

39. Weichselbraun, A.; Streiff, D.; Scharl, A. Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. *Int. J. Artif. Intell. Tools* **2015**, *24*, 1540008, doi:10.1142/S0218213015400084.
40. Scharl, A.; Weichselbraun, A.; Göbel, M.; Rafelsberger, W.; Kamolov, R. Scalable Knowledge Extraction and Visualization for Web Intelligence. In Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS-49), Koloa, HI, USA, 5–8 January 2016; IEEE Computer Society Press: Los Alamitos, CA, USA, 2016.
41. Wyss, V. Journalismus als duale Struktur. Grundlagen einer strukturationstheoretischen Journalismustheorie. In *Theorien des Journalismus*; Ein diskursives Handbuch; VS Verlag: Wiesbaden, Germany, 2004; pp. 305–320.
42. Odoni, F.; Kuntschik, P.; Brasoveanu, A.M.; Rizzo, G.; Weichselbraun, A. On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance. In Proceedings of the 14th International Conference on Semantic Systems (SEMANTICS 2018), Vienna, Austria, 10–13 September 2018; Elsevier: Vienna, Austria, 2018.