Extended Abstract

Advancement through technology? The analysis of journalistic online-content by using automated tools\textsuperscript{1}

Jörg Haßler, Marcus Maurer & Thomas Holbach

1. Introduction

Without any doubt, the Internet is continually gaining in significance for political communication research. At present, about 75 percent of the German population state that they use the Internet at least occasionally (van Eimeren & Frees, 2013, p. 363). All traditional mass media operate websites that provide real-time information. For citizens, these journalistic websites are the most important information sources online (Hasebrink & Schmidt, 2013, p. 8).

The growing importance of online media also gives rise to consequences for content analyses of journalistic online media coverage. Because such analyses consider wide-reaching media that are representative for the entire media system or media that are meant to be the basis for effect analyses, journalistic online media have to be included in many content analyses nowadays. On the one hand, such analyses appear very promising because online media use standardized programming languages and codes are available in digitized form. On the other hand, the quantity, dynamics, multimediality, hypertextuality or the personalization set boundaries for storage and analyses of websites. This article discusses frequently used strategies to address those challenges and presents five recently developed tools for automated storage, organization or coding of online content.

2. Challenges of the Content Analysis of Websites

The internet constantly changes. This holds true both for the available websites in their entirety and for individual articles within web services. For practical reasons and for reasons of intersubjective comprehensibility, websites have to be stored before analysis. As websites are standardized, the dynamics can be addressed by automatically storing their content. The overall quantity of online content and the dynamics of websites can thus be challenged by careful selection and storage of websites. To address the multimediality it is necessary to store and code not only the text of websites but also embedded pictures, videos and audiofiles. The same holds true for hyperlinks. A further challenge for online content analyses is personalization. Online content can be tailored individually using algorithms to

\textsuperscript{1} This publication was created in the context of the Research Unit “Political Communication in the Online World” (1381), subproject 4 which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).
address single users. This individualization can cause enormous problems for automated tools. For example, it is nearly impossible to analyze the placement of single articles as articles can be presented in different order to different users. But online content does not only challenge content analysis. It also provides opportunities like the wide availability of meta-data such as comments, likes and shares of articles.

3. Traditional procedures to store online content

Many studies in communication research use procedures like taking screenshots, webcrawling or storing RSS-Feeds to archive online content for content analyses. These procedures vary regarding the degree of automatization. Taking screenshots of websites is the easiest but also the most time-consuming procedure to store online content. While websites mostly appear in the same layout and style and there are formats that make hyperlinks usable, it is impossible to save videos and audiofiles with just a screenshot. Webcrawlers like screenshots save online content in the same layout and style as it appears online. They make hyperlinks available but they also demand manual storage of videos and audiofiles. Furthermore, both procedures do not bypass personalization algorithms. A procedure that can be used to store websites regardless of personalization is the access via RSS-Feeds. These feeds are often automatically created by the content-management-system and list published articles. As all articles appear in the same layout in reverse chronological order, they are not personalized. Unfortunately, the RSS-Feeds do not per se show the articles in the layout and style like they appear online. To store the articles in the outlook of the online versions it is necessary to use the RSS-Feeds as a register and store the articles online versions from there. This short overview shows, that conventional procedures do not address all challenges of websites for reliable content analyses. Therefore it is necessary, to combine these procedures and use tools that best fit the needs regarding the particular research questions.

4. The analysis of journalistic online-content by using automated tools

To address the challenges of online content analyses we want to compare five recently developed tools to store, organize and code online content. These tools are AmCAT, the NewsClassifier, the coding platforms ANGRIST/IN-TOUCH, the Facepager and ARTICLe.

AmCAT\(^2\) combines the organization and coding of online content (van Atteveldt, 2008). It allows to list great amounts of data in a SQL-Database. This data can be analyzed automatically or manually. AmCAT focusses on text formats like XML, RTF or CSV. Thus, AmCAT alone does not address the internet’s multimediality and hypertextuality. Furthermore it depends on the procedure of data storage whether AmCAT can be used for content analyses of videos, audiofiles with just a screenshot. Webcrawlers like screenshots save online content in the same layout and style as it appears online. They make hyperlinks available but they also demand manual storage of videos and audiofiles. Furthermore, both procedures do not bypass personalization algorithms. A procedure that can be used to store websites regardless of personalization is the access via RSS-Feeds. These feeds are often automatically created by the content-management-system and list published articles. As all articles appear in the same layout in reverse chronological order, they are not personalized. Unfortunately, the RSS-Feeds do not per se show the articles in the layout and style like they appear online. To store the articles in the outlook of the online versions it is necessary to use the RSS-Feeds as a register and store the articles online versions from there. This short overview shows, that conventional procedures do not address all challenges of websites for reliable content analyses. Therefore it is necessary, to combine these procedures and use tools that best fit the needs regarding the particular research questions.

---

2 A documentation of the AmCAT (Amsterdam Content Analysis Toolkit) is available at https://github.com/amcat/amcat. Furthermore, registration to use the tool is possible at http://amcat.vu.nl.
files or hyperlinks. This holds true for bypassing algorithms. The data has to be stored in a way that neutralizes personalization. Besides the organization of data, AmCAT allows various procedures of automated text analysis, like Natural Language Processing (NLP) or Part-of-Speech Tagging (POS). It combines the organization of the coding process, the allocation of the material and the generation of data as well as their export to a statistics software (van Atteveldt, 2008, p. 185). The opportunities to organize, code and generate data lead to a high complexity of the tool. Coding of videos and pictures has to be done manually, as the tool is specialized for text analyses. These disadvantages can be addressed by combining AmCAT with tools for data storage that are appropriate for research questions that focus on multimediality or hypertextuality and by combinations of automatic and manual coding.

The tool NewsClassifier3 was created to automate the whole process of content analysis from data storage to coding (Scharkow, 2012, p. 250). The tool allows to automatically store journalistic websites. It is possible to access the data via the RSS-Feeds of the websites. Thus, algorithms that personalize content can be bypassed. The data can be stored as HTML or text files. Like AmCAT NewsClassifier focusses on automated text analysis. To organize the coding procedure the tool is able to select a sample of data for automated or manual coding. Manual coding data can be used as training data for the automated coding. Furthermore, NewsClassifier calculates reliability tests and allows exporting data to a statistics software. The disadvantages of the NewsClassifier are similar to those of AmCAT. It is not possible to automatically code content information from pictures, videos or audiofiles. But the automatic storage of the NewsClassifier makes manual coding of such data possible, thus addressing the main challenges of online content analyses.

The tools ANGRIST and IN-TOUCH4 focus on the coding process itself (Wettstein, 2012; Wettstein, Reichel, Kühne, & Wirth, 2012). They allow computer-assisted half automatic coding. ANGRIST provides a step by step coding along categories within a programmed codebook. The texts for coding are displayed in the tool. Therefore, Unicode or ASCII formats are required. The user interface makes it unnecessary to code single numbers as it provides dropdown menus or checkboxes for coding. The tool IN-TOUCH complements ANGRIST as it is a tool for supervising the coding process. It provides reliability tests and controls the progress of the project. As both are tools for manual coding of text data, they do not per se account for multimediality and hypertextuality. Thus both tools can only complement storage tools if the research questions focus on videos, audiofiles or links.

A tool only for data storage is the Facepager5 (Keyling & Jünger, 2013). It was developed to collect information from social network sites. It accesses the appl-

3 The tool is available at https://github.com/mscharkow/newsclassifier.
4 A documentation of ANGRIST (Adjustable Non-commercial Gadget for Relational data Input in Sequential Tasks) is available at http://www.ipmz.uzh.ch/Abteilungen/Medienpsychologie/Recource/Angrist.html.
5 The Facepager can be downloaded at https://github.com/strohne/Facepager.
cation programming interface (API) of Facebook and Twitter. But it can also be used to save information from other JSON-based platforms, like YouTube. After adding the Facebook-Feeds or Twitter-Channels that should be collected, the Facepager saves information such as status updates, the number of page likes or the number of comments. The Facepager collects all data that is available from each platforms API. It might thus be insensitive with regards to personalization. To collect data it is necessary to have a user account at the social network sites of interest. The collected data is shown in the user interface and can as well be exported to a statistics software. Multimedia data that is shared in status updates can also be saved and is copied to the local hard disk. The text information is machine readable. The tool can thus be combined with one of the previously described tools for automated or half-automated coding. As the Facepager is developed for social network sites it cannot be used to store or analyze complete websites or articles from websites.

ARTICLe was developed for the automated storage of articles from journalistic websites by Thomas Holbach, Christoph Uschkrat and Jörg Haßler for the DFG funded project “Digital Knowledge Gaps”. In contrast to the previous tools it stores articles from websites fully automatic including all multimedia elements like pictures, videos and audiofiles. Furthermore, it stores meta-data such as the likes and shares of an article on social network sites. A third advantage of the tool is that it serves as a coding platform for manual coding. As ARTICLe saves articles via the RSS-Feeds of the websites it is able to bypass algorithms for personalization. Articles are stored as they appear online, as the focus of the database is to provide a platform for manual coding. Therefore screenshots in the formats HTML, PDF and JPG are saved in a relational database. As well as the texts and pictures, videos and audiofiles are collected automatically. The source codes of the articles are searched for keywords. If a keyword appears a regular expression (RegExp) extracts videos and audiofiles and a php-script allows to download these files. All stored articles are saved in a table together with all embedded multimedia files and the meta-data of the articles. This table serves as a coding platform were human coders can select, edit and comment all stored articles. Meta-data like Facebook-likes and –shares can be exported to statistics software. The main advantages of ARTICLe are the presentation of the articles like they appear online. It accounts for the multimediarity and hypertextuality of websites and it allows to bypass the personalization of websites. In combination with tools for automated coding ARTICLe might provide a fully automated content analysis of news websites.

5. Conclusion

The growing importance of the internet as a political communication channel as lead to a growing importance of online content analyses. To address the challenges of online content for scientific analyses, like the quantity, dynamics, multi-

---

6 The user interface of ARTICLe (The Automatic RSS-crawling Tool for Internet-based Content analysis) can be accessed at https://article.publizistik.uni-mainz.de/feeds/view.
mediality, hypertextuality and personalization of websites, it is necessary to use tools for data storage. Depending on the research questions there are a few recent tools that address these challenges and allow an automatization of several steps within the process of content analysis. Although there are technical obstacles like flash or javascript applications that are hardly storable, a careful planning of the content analysis and a mindful use of the presented tools allows to automate many working steps of the content analysis.

References


