PISSN : 2571-9904 - EISSN : 2602-7763

# *Automated Content Analysis for the Elecronic-Newspapers Studies*

**Rahmouni Loubna [*،1] , Adel Sid [2]**

**Media Studies and Digital Media Lab**

**[1] University - Oum El Bouaghi – (Algeria), Rahmouni.loubna@univ-oeb**
**[2] University - Oum El Bouaghi – (Algeria), Adelinffo@hotmail.com**

**Received**: 27 / 10 / 2021 **Accepted**: 09 / 02 / 2022 **Published**: 31 / 03 / 2022

**Abstract**:

The digital age has brought about fundamental changes in the field of media, so studies have shown that many researchers from different scientific fields have already begun to analyze the contents of digital media and social networks, including researchers in the field of computers, who created research in the field of electronic newspapers analysis. The availability and volume of data in digital media, in parallel, has transformed its contents into a rich research subject in psychology, media, sociology, political science, and others. However, when analyzing the content of digital media, one faces Researchers have a number of fundamental differences compared to traditional content.

The vast amount of data and unique features of digital content require the application of new technologies of value associated with the transformations taking place in the digital environment. But despite the advantages that new methods of digital content analysis enjoy at the expense of traditional content analysis methods, they are not common in digital media research, despite the increasing rates of spread, in an environment dominated by the features of continuous information flow, and big data (Big Data). ) available to users with ease and ease, but the question that arises in this field is whether it is possible to analyze the contents of the Internet and its various platforms with the same ease of accessing the content therein? Or is there a clear difference, which requires researchers in the field of electronic media to pay attention to it? Will the huge data resulting from the digital press.

**Keywords:** *news paper, electronic news paper,digital transformations, content analysis, automated content analysis*

* **Rahmouni Loubna**

## INTRODUCTION

The current digital age has brought about fundamental changes in the field of journalism, and studies have shown that many researchers from different scientific fields have already begun to analyze the contents of digital journalism, including researchers in the field of computers, for example, as they created research in the field of analyzing electronic newspapers and websites. Or social media or blogs..., and the availability and volume of data in digital newspapers, in parallel, has transformed its contents into a rich research subject in psychology, media, sociology, political science and others (Trilling, 2021, p. 8_ 23). However, when analyzing the content of digital newspapers, researchers face a number of problems.

Fundamental differences compared to traditional journalistic content, the huge amount of data and unique features of digital content call for the application of new techniques of value associated with the transformations taking place in the field of electronic journalism, and there are various other scientific fields that already apply computational methods to study digital journalism data, and in many cases, interests are related Research here is closely linked to the interests of the researchers themselves.

However, despite the advantages that new methods of digital content analysis enjoy at the expense of traditional content analysis methods, they are not common in digital journalism studies, despite their increasing prevalence rates, in an environment dominated by the features of continuous information flow and big data (Big Data). ) available to users with ease, but the question that arises in this field is whether it is possible to analyze the contents of these newspapers with the same ease of accessing the news in them? Or is there a clear difference, which calls for researchers in the field of electronic journalism to pay attention to it? Will the huge data resulting from the digital press lead to more flexibility in defining media agendas, with the same features of the news flow inherited from traditional newspapers, and in the same way in which the contents are conceived and produced in terms of size, diversity and homogeneity? And the need to move to research methods, especially in related fields such as computers, for example?

### 1. Content analysis, from the bigenning

Content analysis has been used for several decades as a microscopic tool through which to focus on the contents of different media, but its uses with different names and levels of scientific abstraction preceded the emergence of the media. In the seventeenth century, when a careful examination of newspapers was conducted by the church as a result of its keenness to prevent the spread of non-religious matters, as well as the existence of a documented case for a quantitative analysis of the content in Sweden in the nineteenth century, which included the conflict between the church and scholars (J, spring 2000, pp. 88- 98).

Many researchers point out that there is no exact date for the beginnings of content analysis in the twentieth century, but most agree that its beginnings go back to **Lasswill** and his colleagues in the year 1930 AD when they were at the School of Journalism in Columbia, America, and then followed by the study conducted by Speed to compare the change In the nature of the reduction of New York newspapers after the New York Times newspaper attempted to increase their circulation by reducing the price and increasing the volume and its tendency to excitement in editing press articles.

Studies that apply content analysis have become distinguished studies, and among these studies is Willey's study of regional newspapers, in which he used the same categories and the same standards to study the development of the weekly regional newspapers, which he relied on alone during the American War of Independence.

In 1940, there was an organized use of the method in journalism research, after the studies presented by **Laswell** and **Letts**, through the scientific efforts of the study of propaganda at the University of Chicago. And a reference in content analysis, after categorizing them into categories for the purposes of analysis, conferences and seminars are held in the same context, including: **the American National Conference** held in 1967 for content analysis, the first

conference devoted to this topic, during which many researches related to content analysis systems were discussed (المدخلي, 2021).

Content analysis has become more widespread and used, especially with the studies conducted by **Berelson** and Lazar Sfield in the 1940s and the **Berelson** studies in the 1950s, and it has also become widely recognized as a research tool used by many researchers from different research disciplines. The definitions that researchers develop for content analysis, the most famous of all, is the definition developed by **Bernard Berelson** (1952) as "a research method that is applied in order to reach a purposeful and organized quantitative description of the content of the communication method". It is clear from **Berelson's** definition that there are several elements of content properties (عباس، 2007، صفحة 76):

- Content analysis, by classifying and tabulating the data, seeks to describe the apparent and explicit content of the material under analysis, and it is not limited to the substantive aspects, but also the formality.
- Depends on the repetitions received or the appearance of sentences, words, terms, symbols or forms of meanings included in the analysis material based on what the researcher objectively determines the categories and units of analysis.
- It must be characterized by objectivity and subject to methodological requirements (such as honesty and consistency) so that its results can be taken into account, as being generalizable. - The analysis should be systematic, and it should mainly adopt the quantitative method in the analysis processes with the aim of doing the qualitative analysis later, and on objective bases.
- The results of the content analysis must be identical in the case of re-examination of the same tool and the material (under analysis), to ensure the stability of the results
- consistency over time - or through their application and closeness of their results by other analysts (external arbitration).
- The results of the content analysis are linked with the descriptive, analytical and theoretical results in a general and comprehensive framework, according to which the phenomenon or problem is explained, that is, in this case it is considered a complement to other methodological procedures that precede it, or follow it within the framework of the comprehensive study.

2. **The spread of the use of the Internet and its impact on the methodology of analyzing the content of electronic newspapers**:

The increasing spread of the Internet with the emergence of Web 2.0 technology and the unlimited opportunities it provided in the manufacture, publishing and receiving of news led to the transformation of paper newspapers into electronic copies in Algeria and other countries of the world,how to publish it, and on the social level, many studies have shown the increasing recourse of Algerians - and they are no different in this matter from other peoples - to the Internet as a main source of news compared to other media, we affirm today that there is no newspaper without a more interactive and readable electronic version, And widespread among users, in order to respond to the lifestyle and changing reading habits of the masses, and to ensure the long-term survival of these institutions.

In the field of research, researchers shifted their focus to online content. This content includes the contents of electronic newspapers, which, unlike traditional newspapers, have a different format, frequencies and speed, besides, they have unique features such as audio, image, visual elements, hypertext, hyperlinks, and interactive systems that are Part of primarily digital media text. In light of these distinctive features, the application of content analysis as a technique to the contents of electronic newspapers is considered a major challenge for researchers. In the past, content analysis was mainly used to study texts from traditional media such as newspapers, television and radio programmes, official interview transcripts and various documents. Independent that was created to serve a specific journalistic purpose.

It is worth noting that news on the Internet has different features, and therefore the application of the technique of content analysis as it is classically accepted becomes an issue that requires a complete review, or a systematic adaptation of it to suit Internet texts that are characterized by being flowing and with limited time validity.

The news in the electronic newspaper - for example - is Liquid News and can be defined as irregular, continuous, and is published according to a participatory, multi-media and interconnected process, and is produced according to journalistic principles, and also includes liquid news stories that are published in different drafts, created in cooperation With users, it is published in multiple ways, and it also has hyperlinks to different sources and documents (photos, videos, texts, ...), and it is constantly changing, which theoretically means the impossibility of its existence in specific forms, including its continuous change and the difficulty of filling it in the case of applying content analysis In its traditional form (Buyong, 2017, pp. 164- 174).

In this regard, **Carlson** points out that the fundamental problem is that liquid news is constantly changing, which poses a challenge to traditional content analysis, which deals with content on the basis that it is clear content. , Which was confirmed by a number of researchers, but they went to the hypothesis of the impossibility of applying content analysis, which was specifically designed to study the contents of traditional media on the contents of digital media because of their fundamental differences (Buyong, 2017, p. 168). Content analysis, as we know it, is directed primarily towards certain types of traditional media, which have their own characteristics and characteristics, and are differentiated according to these characteristics and features. Film or advertising, for example, requires special methods to identify categories and units.

3. **Automated Content Analysis ACA: Concept, Mechanisms and Roles**:

Given the expertise of computer scientists in handling and evaluating digital data, their techniques can be very rewarding for researchers in the field of digital journalism. In press releases that traditional content analysis fails to identify, automated content analysis is defined as a method of coding messages with the help of computer algorithms, and depends on the method in which the actual coding decision is made (i.e. allocating symbols to documents or textual or audio-visual elements, typifying the content, and defining its indicators) Automatically, it does not require judgment or human intervention, and therefore it is executed automatically, since it relies on the computing capabilities of computers rather than human coders (it has to do with artificial intelligence), and this type of coding can be applied to very large data (Big Data). Moreover, automatic content coding is very reliable as any analysis of the same material and content can be reproduced. Its errors are few and what is found is related to the machine (error in specifying search fields, for example) and not to human biases. The history and development of ACA digital content analysis can be understood through three basic developments (SCHARKOW, s.d.):

(1) The concept of content analysis and its relationship to the computer.

(2) developing software for automated data processing and analysis,

(3) Availability of a large amount of digital content and documents or that can be read and analyzed automatically.

The first stage was characterized by several attempts to develop computer-aided analysis (in an automated way) in the late fifties, mainly through computer experiments. Sociologists and media researchers first had to learn computer programming, and in many cases the support of science departments was required, while the first few experiments were roughly aimed at providing textual statistics, by the number of words being repeated. Conceptually, the development of automated content analysis in its infancy was slow compared to the

development of traditional content analysis, which relies on coding and human intervention, due to the deep problems that this field was experiencing due to the slow development of technology and computerized data processing systems, offset in the field of content, the lack of data The digital media that can be analyzed, due to the delay in the integration of traditional media with technical media linked to the Internet.

The emergence and development of search package programs such as General Inquier represented the first milestones in the development of automated content analysis, which necessitated researchers to go further in the field of computer analysis, and the Germans were among the pioneers in this field, and they initially relied on archival materials only , due to the unavailability of machine-readable digital documents, **DeWeese** (1977) is considered the first to automatically collect everyday media content using electronic typesetting machines increasingly common in electronic newspapers. In 1979, service provider **LexisNexis** began creating digital editions of newspapers from America. With the development of the personal computer, and the additional availability of digital media content, the 1980s saw a period of renewed interest in ACA, especially in media, communication and political sciences. . Fan (1988) demonstrated the potential of this type of analysis in studies of media agenda and long-term media framing. A research team led by **Schrute** developed software to automatically extract international events from the Reuters wireline messaging service.

With the progress in artificial intelligence research, the techniques of automatic analysis of the contents of the press and digital media have developed, and interest in them has increased, despite the many reservations made by researchers in the field of scientific research methods, given the impossibility of a machine replacing human beings, and despite the debate in this field, no Computer scientists are still using programs that are able to analyze the contents and huge data in electronic newspapers, like other digital media. Repetition of units and categories and in the actual analysis of the results, there are computer programs to assist in these tasks, such as **The General Inquirer and Oxford Concordance Program**, where the **General Inquirer** program is used to identify and compare vocabulary repetitions, and the computer can identify important concepts in Text automatically, calculate repetitions, display results in graphs, arrange them into categories, perform some basic statistical tests, Using software such as KWIC (keywords in context), **KWOK** (keywords out of context), **INTEXT**, **MAX**, and **ATLAS**.ti (سارانتاكوس، 2017، صفحة 528).

Among the most important current programs in analyzing the contents of electronic newspapers, we find:

- General Inquirer:

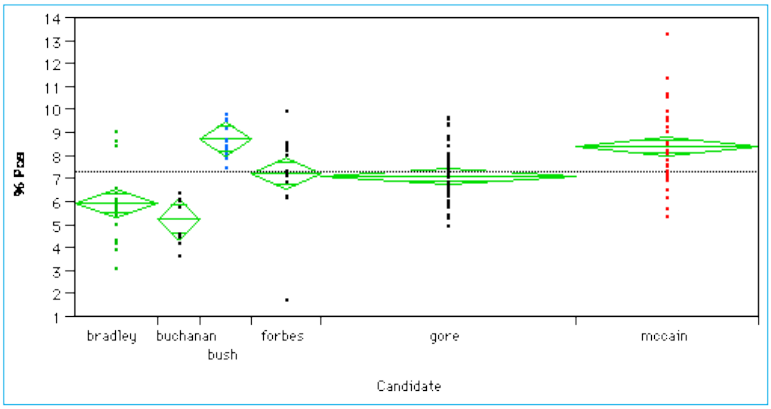
The Inquirer is one of the most advanced programs in quantitative data analysis. Text analysis tools span a wide range, and there is growing talk about whether they provide only a container for text processing, or whether they come with content categories and language-specific divisions. . The tools for automated content analysis are primarily text processing vessels and the definition of words and units, and although the Inquirer provides more than 182 categories, it differs from some automated content analysis programs in providing criteria for frequency and frequency of the category, but it imposes on the user to determine what are Domains to be compared statistically.

The contents analyzed in it may include a sample of editorials from different online newspapers, speeches by political candidates, reports of shareholders from different companies...etc. The 182 **General Inquirer** categories have been developed for application to content analysis research in the social and human sciences, and not for text archiving, automatic text routing, automatic text classification, or other NLP goals, although they may be related to each other. These classes are created by the researchers here, not the computer, although some of the class developers here have relied on analyzes and algorithms produced by

computers. Many categories were also initially created to represent different social sciences and humanities concepts, and for many of the major theories that were prominent at the time the Inquirer was first developed, including those put forward by Harold Laswell, Talcott Parsons and Charles Osgood.

It also included categories related to "middle-range" theories, such as **David McClelland's** theories of needs for achievement, power, and belonging. With further revisions to the categories over the years.

**Figure (1): Results Shown by GI**




- **The TEXTPACK**:

a program used to analyze digital content, as it provides procedures for text processing, encoding and reconnaissance, and is equipped with an automatic means to encode text with the help of a dictionary chosen by the researcher and containing symbols, and this program can calculate the frequency of units and other aspects, where these procedures are consistent with other statistical programs Such as SPSS, and SAS. TEXTPACK offers a host of other features: - Repetitions of the word in the entire text or in certain parts of it, as it is possible to specify the occurrences of the word in the text by different options. It can be obtained in the latter in different forms, such as the alphabetical order of it, or according to the size of its repetitions ... etc. Keywords can be displayed in context, out of context, or multiple word combinations in context. Cross-references and word matching - Word comparison between two different numeric texts. - coding. Selection of units of analysis. Hardware and operating system: The program can be run on all computers that run on the Windows operating system.

- **AQUAD**

Aqua dis a program with multiple functions, as in the previous program, and the content in this program is analyzed starting from summarizing the text to rebuilding it, then comparing it with others.

in the second and third stages, the level of analysis rises to a higher and more complex level, where the data is presented in tables or in arrays of text and symbols. **Aquad** provides classic tools for analyzing content as text search: search for segments in texts, markup, mark parts of text, audio, images, videos, write notes (annotation), insert comments related to fragments, whole texts, audios, or Images or video clips and word analysis such as counting word occurrences according to certain criteria, in addition to retrieval by file name, code, keyword or parts of the texts of a digital document, it also enables to retrieve segments, analyze tables: (create tables that collect criteria and arrange data in rows and columns , and building links hypotheses: looking for relationships between symbols and comparing states/files: coding

variation between files...). (525 صفحة ،2017 ،سارانتاكوس)

The benefits of automated content analysis for electronic newspapers are numerous, including:

• Reduces analysis cost and time - automated analysis software like the aforementioned has been found capable of processing up to 300MB of text data per analysis (about 8500 articles assuming 5000 words per newspaper article)

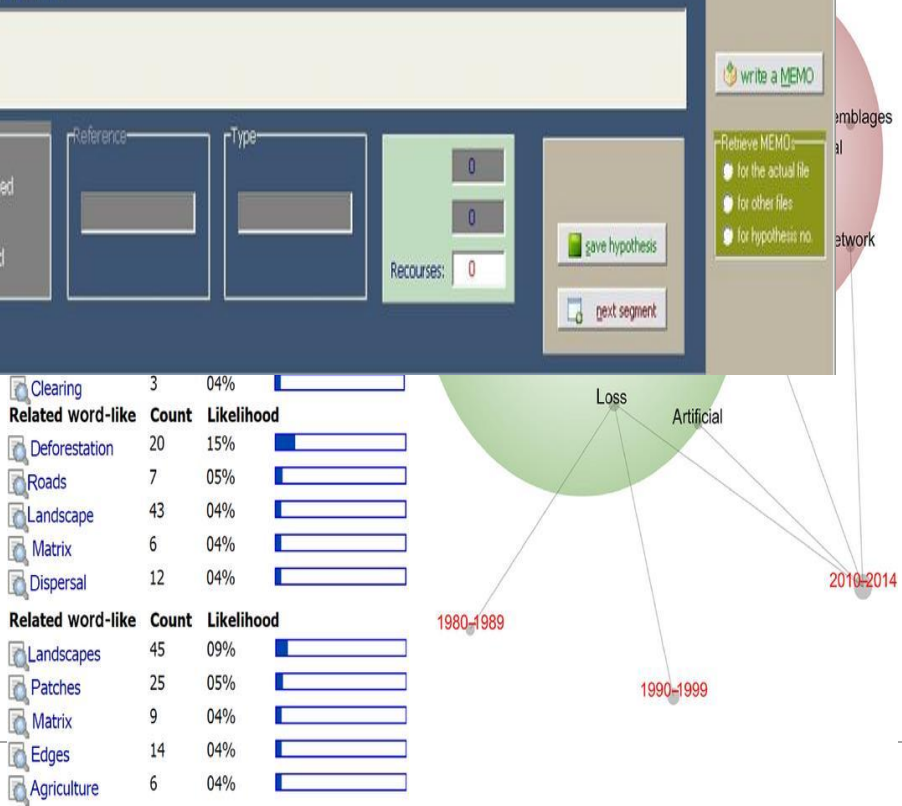
• It provides researchers with various results, both quantitative and qualitative. • Reduces subjectivity, reduces researcher influence, bias, and subjectivity

• Uses concepts instead of single words – counting the frequencies of semantic and linguistic complexities – such as synonyms, co-occurrence frequencies of users, syntax ....

• Creates graphic summaries of the results (as in the following image)

**Figure (2): Interface Of the GI**


**Figure (3):**



**Automated CI Results**

**4. Automated content analysis steps:**

The process of automated content analysis ACA - in principle - is similar to the process of traditional content analysis, but differs from it only in those areas in which human intervention is replaced by rules determined by the computer, Automated content analysis begins with providing and compiling the digital content of electronic newspapers, and the digitization of the content of electronic newspapers has changed, and turned to more effective methods, then sampling, storage and management across very large groups of texts and other topics, and then removing content that is not relevant to the processed content, especially in the case of data Unstructured as web pages or documents in pdf format.

In the case of analyzing the content of newspapers in a traditional way, the analysts can filter the content by ignoring the contents that are not related to the subject being treated by ignoring them, but the automation of content analysis is only through the tribal filtering, - automatically - or by the human for the unnecessary contents So that the computer does not have to process it, and thus affect the results obtained.

In order to ensure that the material from the electronic journal is preserved for as long as possible, for analysis or re-analysis, messages are often converted into standardized formats such as Unicode text or XML.

(Unicode assigns a unique number to each character, regardless of platform, regardless of software, and regardless of language. The Unicode standard has been embraced by leaders in the computing industry, such as Apple, HP, and IBM (IBM), JustSystem, Microsoft, Oracle, Sun, Unisys, etc. Also, modern standards such as XML ), Java, ECMAScript (JavaScript), LDAP, CORBA 3.0, WML etc. Requires the use of Unicode, which is the official way to implement the ISO/IEC 10646 standard. Unicode is also supported in Many operating systems, all modern browsers and many other products.The emergence of the Unicode standard and the availability of systems that support it is one of the most important recent global trends in software technology.)

In general, it can be said that there are three basic steps in analyzing the content of electronic newspapers in an automated way, which are:

**First**: Defining the concept: At this stage, the concepts through which the content of the electronic newspaper will be categorized are determined, through the use of keywords that constitute the most important concepts in the researcher's study and work to guide him to the content directly.

**Second**: Defining the concept: Here, the thesaurus can be built for the basic concepts in the study and the topic being addressed, through different methods and depending on the design and capability of the ACA system used. At the end of this stage, the program adopted in the analysis is able to create a set of prevalent concepts and knowledge in the vocabulary.

**Third**: Classifying the text, that is, the content of electronic newspapers according to the specific concepts identified in the previous two stages. The compositions are generally classified in high fidelity (eg by sentence or by font fragments, hereinafter referred to as “sections of text”). (GabrielaC, 2016)

**Conclusion**

All methodological methods, both traditional and new, have their advantages and disadvantages, and it is clear that they depend on the research questions being asked, but what we hope today and with urgency is that the doctoral students in Algeria have clear visions and perceptions about content analysis of digital data, especially Massive ones, and those related to electronic newspapers in particular. Despite the many challenges facing this type of analysis, especially when it comes to validation or inspection procedures, the development of digital media and the spread of electronic newspapers should draw the attention of academic researchers to the idea of the necessity of adapting content analysis studies according to the nature of the medium that The analysis process is applied to it, as no traditional content analysis has become able to analyze digital contents because of the vast gap in the characteristics of the traditional and digital mediators, nor have we mastered the analysis according to the nature of the digital mediator, of which automated analysis is only a part, and it is multiple and varies according to the programs that are applied, and has Its efficacy has been proven many times.

**References :**


- Jelle W. Boumans & Damian Trilling, *Taking Stock of the Toolkit, Digital Journalism*, Digital Journalism, Vol 4, No 1, p-p 8–23.
- Maier Daniel, Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology, Journalism and Mass Communication Quarterly, V 77, N 1, spring 2000, p-p 80- 98
- محمد المدخلي، منهج تحليل المحتوى، متاح عل الرابط: www.cau.edu.sa/30732/ تاريخ الزيارة 21/ 03/ 2021، 19 م
- محمد خليل عباس واخرون، مدخل الى مناهج البحث في التربية وعلم النفس، دار المسيرة ط 1، عمان، 2007 ، ص76.
- Zeti Azreen Ahmad & Mazni Buyong, Content Analysis Of Online News Portal, Issues and Challenges, Journal of Social Sciences and Humanities, Special issues 2 (2017), University of Kebaangsaan, Malaysia, 2017, P-P 164,174/
- Karlsson, M. Charting the liquidity of online news: Moving towards a method for content analysis of online news. International Communication Gazette, Vol.74 (4),2012, pp. 385-402.
- MICHAEL SCHARKOW, Content Analysis Automatic, he International Encyclopedia of Communication Research Methods, DOI: 10.1002/9781118901731.iecrm0043
- سوتيريوس سارانتاكوس، البحث الاجتماعي، تر شحدة فارع، المركز العربي للأبحاث ودراسة السياسات، ط1، بيروت، 2017، ص 528.
- For Rurther information see Inquirer -- Use and Comparisons (harvard.edu)
- Free Qualitative Data Analysis Software | by Manuel van Hoben | Medium
- see What is Unicode? In www. http://unicode.org- Gabriela C. et al, Automated content analysis: addressing the big literature challenge in ecology and evolution, Methods in Ecology and Evolution, N 7, 2016, p-p 1262,1272.