Natural Language Processing in Arts Management

MARK CIELIEBAK[1], FERNANDO BENITES[2], LARA LEUSCHEN[3], MICHAELA HNIZDA[4], DIANA BETZLER[5]*

[1] Institute of Applied Information Technologies, Zurich University of Applied Sciences and SpinningBytes AG, Winterthur
[2] Institute of Applied Information Technologies, Zurich University of Applied Sciences
[3] Center for Arts Management, Zurich University of Applied Sciences
[4] SpinningBytes AG, Winterthur
[5] Diana Betzler, Center for Arts Management, Zurich University of Applied Sciences

Abstract

Natural Language Processing (NLP) opens up new possibilities for arts management in practice and research. This article introduces the typical research process of NLP and presents the most important methods and techniques like Sentiment Analysis, Author Profiling, Named Entity Recognition, Topic Modeling and Trend Detection. Using recent research results and new illustrative examples, we describe the possibilities and limitations of NLP for arts management.

Keywords

Arts Management Research, Natural Language Processing, Text analysis, Topic Modeling, Author Profiling, Named Entity Recognition

1. Introduction

The boost of digital archives and libraries in art, literature, and music; the shift of cultural marketing and cultural criticism to social media and online platforms; and the emergence of new digital art and cultural products lead to an enormous increase in digital data, creating challenges as well as opportunities for arts management practitioners and researchers.

Natural language processing (NLP) is used to analyze and interpret human language automatically. As a subfield of computer science and artificial intelligence, NLP is concerned with how to program computers to process and analyze large amounts of natural language data. The field of NLP has seen major breakthroughs in recent years. The main reasons

* The project is managed by Dr. Diana Betzler, who is the corresponding author. Email: bera@zhaw.ch.

doi 10.14361/zkmm-2019-010#

for this are (i) better methods in artificial intelligence (AI), (ii) increased availability of data in much larger quantities, and (iii) increased processing power at a lower cost. Thanks to these factors, the possibilities for using technology to analyze data, understand texts, and make reliable predictions have reached new and exciting levels.

NLP draws from many disciplines including computer science, statistics, and computational linguistics in its pursuit to fill the gap between human communication and computer-aided understanding. NLP methods are an essential component of research disciplines relevant to arts and cultural management, such as digital marketing, communication research, and the newly emerging discipline of digital humanities, which involves the implementation of computational methods to answer core as well as emerging questions presented by the humanities.

Despite the maturity of NLP as a research field and the increased effects of digitalization, there is still a high entry cost for introducing these methods to other disciplines. This is mainly due to the difficulty of specifying tasks to which NLP can be applied, and the cost of implementing new, large-scale automated processes. Human language has always played a central role when studying arts and culture, so the application of innovative methods such as NLP is essential: "The digitization of huge quantities of text has raised the stakes by enabling scholars to launch more ambitious projects, while requiring the development of new, more powerful analytic tools." (DIMAGGIO/NAG/BLEI 2013: 576).

We are witnessing a growing momentum in this area, but many opportunities to employ these technologies are squandered, owing to a lack of understanding of the immense potential of NLP tools. For example, simple statistical procedures such as finding the most common words used in association with an arts organization's name can offer valuable insight. NLP can aid qualitative researchers in discovering topics of interest and identifying corresponding texts. By following a mixed-method approach, it can help make qualitative research more transparent (CHAKRABARTI/FRYE 2017: 1357). Obviously, just detecting interesting information is not sufficient. In fact, insights have to be put into action, which depends on the settings and goals of your project or institution. Only then can the discovered information produce actual value to the project.

To those who have been reading about NLP, AI, or text analysis but have difficulty understanding how these might be of benefit for arts management research and practice, the following examples and applications may offer some clarification. They are based on the experience and

insight gained from our recent research project, which was conducted by an interdisciplinary team of NLP professionals and arts management experts. In Table 1, we compiled a summary of NLP applications with their different relevant aspects, as will be discussed in this article.

Method	Research Questions regarding	Data
Sentiment Analysis Section 5.1	• Polarity of texts • Determination of writers' attitude towards a specific topic, event, or product	• Tweets • Comments on event websites • Instagram comments
Author Profiling Section 5.2	• Uncovering details such as authors' age, gender, or native language	• Social media entries • Blogs
Named Entitiy Recognition Section 5.3	• Identifying named entities such as individuals, locations, or organisations	• Social media entries • News articles • Communities websites
Topic Modeling Section 5.4	• Identifying most dominant topics	• Document collections • social media entries
Trend Detection Section 5.5	• Identifying trending topics, long-term trends, and short-term tendencies, single peaks, outliers, etc.	• Google Search • Social media entries • Blog articles • News articles

Table 1: *Summary of exemplary NLP applications with relevant aspects.*

In the following sections, we explain what can be achieved by NLP methodologies and how NLP research processes can be designed. We then introduce five NLP methods in detail and demonstrate their potential for arts management, using intuitive examples and applications. The examples originate from our research and consulting practice or current research. Finally, we critically discuss the potential and limitations of NLP for arts management.

2. NLP – a Definition

As HAUSSER (2000) explains, computational linguistics is a highly interdisciplinary field, which researches language production and comprehension. When applying such methods to real-world problems, we speak of natural language processing (NLP), specifically when analyzing, manipulating, or generating written and spoken human language. At a fundamental level, NLP deals with the structural and syntactic analysis of text, which includes tasks such as tokenizing, parsing, and part-of-speech tagging. Based on this, NLP allows for semantic analyses such as searching, text classification, named entity recognition, topic detection, and argumentation mining, which are explained in greater detail below. Finally, NLP also includes advanced topics such as speech-to-text and text-to-speech, natural language generation, machine translation, and document summarization.

NLP is based on a long-standing research tradition in artificial intelligence (AI) and computational linguistics. It began in the 1950s during the first boom of AI with the goal of "understanding" human language, thereby facilitating a fluent, natural interaction between humans and machines. Its first international association was the Association for Computational Linguistics (ACL), which was founded in 1962. Today, NLP is applied to all forms of text: news articles, tweets, emails, Wikipedia articles, and reviews as well as to data based on spoken language such as audio interviews, YouTube videos, and news broadcasts.

Many solutions to NLP tasks were traditionally statistical or rule-based, mainly when dealing with large corpora and implementing applications such as machine translation or information retrieval. Recently, machine learning approaches have gained momentum and became state-of-the-art in many disciplines. For this reason, the following section briefly explains how machine learning works.

3. A (Very) Brief Introduction to Machine Learning

Machine learning works in two phases (Figure 1). In the learning phase, a system is trained to "learn" a particular task using selected examples. Then, in the application phase, the trained system is applied to a new object. For example, the task of sentiment analysis is to analyze whether a given text expresses a positive, negative, or neutral opinion. During the

learning phase, the system is given a set of texts, each of which has been labeled by humans as positive, negative, or neutral. This is the 'training data', which is typically about 10,000 texts for sentiment analysis. These texts are converted into a numerical representation (the so-called 'features') which represent the structure and content of each text. These features are then fed into a machine learning algorithm. There are hundreds of different algorithms, including Support Vector Machines, Decision Trees, or Neural Networks (note that Deep Learning, which has been successful for many tasks recently, is mainly a special kind of a neural network that basically uses more than one internal layer). The machine learning algorithm computes a mathematical model and depending on the algorithm, the task, and the amount of training data, which can take anything from a few minutes to several weeks.

In the application phase, this model can then be applied to new texts whose sentiment is unknown. First, the same features as in the training phase are computed for the new text. Then the model is applied to these features, and the resulting output suggests which sentiment the text might contain.

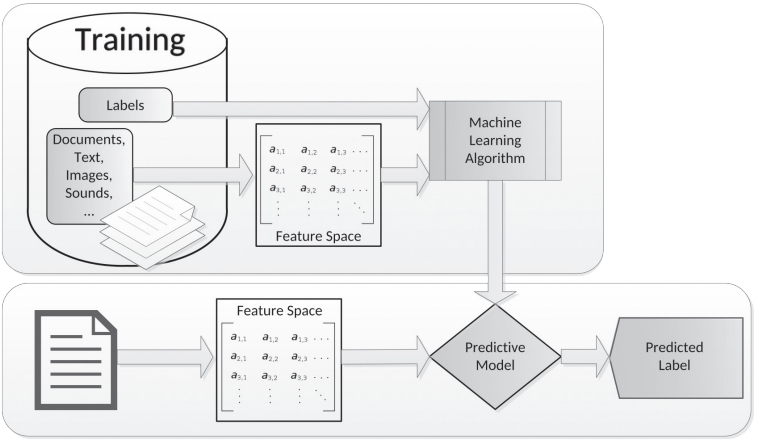

Figure 1: *Supervised machine learning workflow.*

The above process describes a supervised machine learning task, where the training data is labeled (e.g., 'positive', 'negative', or 'neutral'). There are also unsupervised methods, which are usually employed for data exploration and to find key features of the data, such as clustering or topic modeling.

One important thing to note is that (almost) every machine learning system is prone to error, which can vary widely depending on the task. For example, state-of-the-art solutions for sentiment analysis achieve an accuracy of only about 70% (ROSENTHAL/FARRA/NAKOV, 2017), while named entity recognition achieves a score of more than 90% (YADAV/BETHARD 2018). Also the quality can vary significantly with the domain, the algorithm chosen, and the amount and quality of training data, as demonstrated in a 2017 benchmark study (DERIU et al. 2017).

In principle, machine learning algorithms can be applied to any language; however, most current research focuses on "major" languages such as English, Spanish, Chinese, and Arabic, where plenty of data and tools are available. For other languages, appropriately labeled training data is sometimes hard to find, although a subfield of NLP for 'low-resource languages' does exist.

4. Research Process

A typical NLP project is split into the following phases:

1. *Target definition:* What should be analyzed? Which data is available and what information should be extracted? How should the resulting information be used? Ideally, this step defines a goal for the project and identifies which NLP technologies should be applied to reach this goal. The desired target is typically a report or a visualization of the findings.
2. *Data collection:* Is it a one-time analysis or an ongoing process? Can the data be gathered at all due to legal restrictions? How can the data be aggregated and stored? How much will it cost? Some data streams are freely available while others might have copyright restrictions and need to be purchased. If the data is coming in over time, it is essential to have a sustainable aggregation process.
3. *Data clean-up and preprocessing:* Most data is not immediately suitable for NLP analysis. Data from websites or blogs are typically in HTML format and include, besides the main text, navigation elements (menus, buttons, etc.), metadata, advertisements, and teaser texts from other pages which must all be removed. Another example is data verification such as duplicate detection, which is necessary for news articles which are often published in different media in an almost identical form. Other typical NLP preprocessing steps are tokenization (splitting a text into single words), stop word removal

(deleting unnecessary words), and lemmatization (the stem reduction of words).

4. *Data labeling:* A vast amount of labeled data exists which can be employed in NLP projects. However, there is often no suitable data available for the task at hand, either because it is a very specific task, or - more frequently - because no data exists in the target language. In this case, humans must label the training data by hand, which, depending on the task, can vary between several hundred and several thousand documents. Labeling can be done by domain experts but also via crowdsourcing platforms such as Amazon Mechanical Turk (www.mturk.com).
5. *Algorithm selection and optimization:* In this phase, an NLP expert selects the most promising NLP algorithms, trains them on the labeled data, optimizes their parameters, and evaluates their performance to select the best possible system.
6. *Application and interpretation:* Once everything is in place, the NLP system can be applied to real-life data. Depending on the setting, this can be a one-time application or an ongoing process. The application of NLP is usually followed by a methodologically careful interpretation of the data. The primary goal is to distill core information from the data and to generate actionable insights. This often includes visualization of the data and findings, thereby making the results more easily accessible.

5. Methods

In this section, we introduce five classical NLP methods and show, using illustrative examples, how they can be applied to cultural tasks. Each methodological description states the task, a tangible example, and applications from the literature.

Method 1: Sentiment Analysis

Task. Sentiment analysis is used to identify and categorize the polarity of a text, usually to distinguish whether it is positive, negative, or neutral. It is used to determine the writer's attitude towards a specific topic, event, or product. For example, the following sentences each demonstrate a different sentiment:

- I love the movie → positive
- I hate the movie → negative

- The movie starts at 8 pm → neutral
- I don't hate the movie → unknown

A straightforward solution for sentiment analysis would be just to look for specific positive or negative words (such as 'good', 'bad', 'happy', etc.) and determine the mood of a text based on this keyword lookup. However, this dictionary-based method does not always work since these word lists are never complete; word meanings can depend on context (e.g., 'hot') and negation has to be handled carefully. For this reason, most sentiment analysis solutions nowadays use machine learning, whereby the computer learns from examples of which text properties differentiate positive and negative texts. This yields much better and more reliable results and can be even adapted to the target domain, using appropriately labeled training data. CIELIEBAK (2018) provides a more elaborate introduction to sentiment analysis for the layperson, while ROSENTHAL/FARRA/NAKOV (2017) gives an overview of recent state-of-the-art technologies and performances.

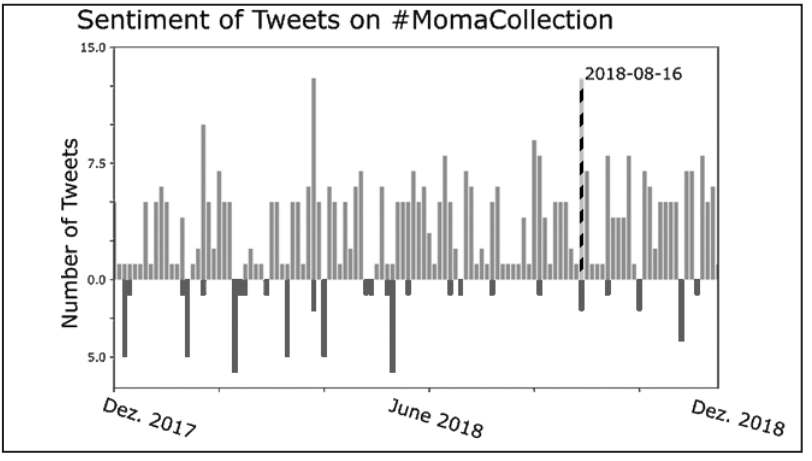

Figure 2: *Sentiment of tweets with hashtag #momacollection. Light bars depict positive tweets and dark grey bars depict negative tweets. Data collected between December 2017 and December 2018.*

Example. In Figure 2, we see the number of positives (light grey) and negatives (dark grey) tweets with the hashtag #momacollection. The high positive peak at 16 August 2018 (marked in green) was the day Aretha Franklin died. There is a not so obvious explanation for this: Andy Warhol created an album cover for Aretha in 1986. MoMA has displayed many exhibitions from Warhol and therefore circumstances linked to him are relevant for the Museum.

Applications. THET/NA/KHOO (2010) built a data set of movie review texts collected from the movie review site www.imdb.com to conduct detailed sentiment analysis. After preprocessing the data, they assigned previously agreed sentiment scores for each word, which they derived from a sentiment dictionary (SentiWordNet) and domain-specific lexicons. In addition, they also included contextual sentiment scores based on grammatical relationships and the dependencies of words in sentences. The authors concluded that this more refined sentiment score calculation method was especially effective for short documents such as message posts on discussion boards.

CORALLO et al. (2017) investigated the use of the official mobile application FolkTure developed for the Italian folk music festival La notte della Taranta in Salento (<www.lanottedellataranta.it>). By offering features such as augmented reality simulations, navigation to cultural points, and game dynamics, the application encouraged users to comment on their experiences and to share multimedia content. To evaluate the tourism and economic spillover effects on Salento, its ongoing potential, and the visitor opinion, CORALLO et al. (2017) combined correlation analysis and spatial analysis with text-analysis techniques. For example, they identified recurrent topics in web user posts published on FolkTure and applied sentiment analysis to this data. Based on this, geostatistics allowed them to estimate the 'sentiment score spatial variation', depicting geographical areas characterized by negative or positive average sentiment scores.

Remarks. It should be noted that sentiment analysis is a typical classification task, where text should be designated according to a set of predefined classes (here: positive, negative, or neutral). In principle, most sentiment analysis algorithms can be generalized to any classification task - one only needs enough training data to adapt to a new task. We have applied similar techniques already to age and gender detection (as explained below), hate speech detection, dialect identification, as well as news categorization, etc. In arts management, it could be applied, for example, to distinguish between reviews or comments on concerts or exhibitions. Combined with word frequency analysis, this could provide insights into which attributes are associated with which type of event.

Method 2: Author Profiling

Task. Author profiling is the method of analyzing one or many texts by the same author to uncover details such as his/her age, gender, or native language from characteristics in writing style and content. Author profiling has a number of applications: First, it is often used in the field of forensic linguistics to create profiles for certain types of messages or online texts, which in turn can help identify security threats. Second, from a marketing perspective, companies and organizations are interested in determining what groups of people use, like, or dislike their products or services, based on an analysis of blogs, online reviews, and comments. Also recently, election campaigns have become a prominent example of the use of author profiling (CONFESSORE 2018). Author profiling from texts exploits the fact that author age, gender, and other qualities are reflected in their writing style. For example, Table 2 shows grammatical structures that are typically used by female and male authors respectively. Similarly, topics in blog articles vary significantly depending on the age of the author, as can be seen in Table 3. The authorship resolution (profiling and identificaton) is such an important challenge that there are scientific competitions such as PAN (<https://pan.webis.de/tasks.html>). They range from author diarization to authorship attribution (KESTEMONT 2018) as well as gender prediction to celebrity profiling (tasks in 2019). One relatable example for (some) humanities scholars might be actual authorship research using stylometry, which analyzes whether a text was written by author A or B or others.

Female Authors	**Male Authors**
Pronouns	Determiners
For and with	Adjectives
Present tense	Of-modifiers (e.g., a pot of gold)

Table 2: *Grammatical features which are more likely to be found in texts by female and male authors respectively (KOPPEL 2002; cited in ROSSO 2016).*

Applications. VOLO (2010) evaluates tourist blogs as a research data source and the findings demonstrate cultural differences in the way tourists use blogs. They also demonstrate significant potential for revealing market-relevant aspects of a tourist experience, although they do not always convey the 'experience essence'. These applications could be interesting for all kinds of issues pertinent to the arts and culture sector.

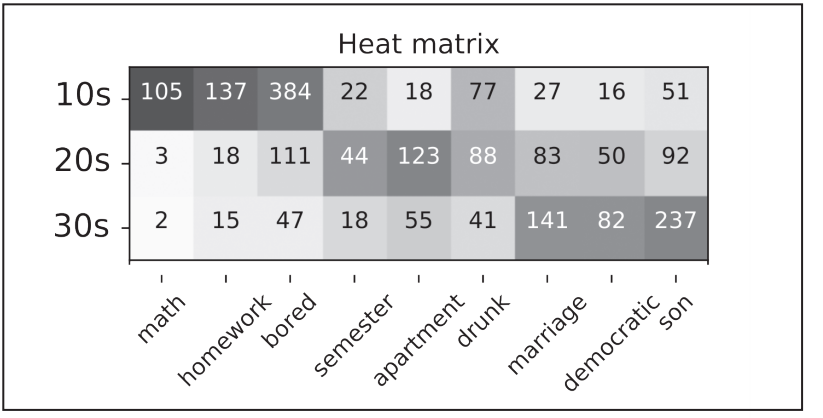

Table 3: *Frequency of topics depending on the age of the authors. The three rows show authors from 10-20 years, 20-30 years, and above 30 years of age, respectively. Values in the table show the number of occurrences of specific topics in blog articles of authors of the corresponding age. The darker the color of the cell, the more often the word is used by a particular age group (Data from ROSSO 2016).*

For example, with all online feedback for a new exhibition or a music event, one could ascertain the distribution of age, gender, and place of origin of its visitors, and – in combination with sentiment analysis – which groups most enjoyed that event. Author profiling could also be used to target marketing campaigns to specific groups via social media. For example, BARBARESI (2016) has shown that German-language tweets can be assigned to specific geographical regions based on their content, which would allow marketing efforts to targeted towards the cultural and national background of its recipients.

Method 3: Named Entity Recognition

Task. Named entity recognition (NER) is the task of identifying named entities such as individuals (e.g., Andy Warhol), locations (e.g., Zurich), or organizations (e.g., The Museum of Modern Art MoMA) from a text. It is commonly used as an intermediate step for further processing. For example:

> ***Bregenz** boasts one permanent cultural attraction in a strikingly modernist art gallery designed by the great Swiss architect **Peter Zumthor.***

Running the above sentence through an NER system would result in the retrieval of the entities ‘Bregenz’ and ‘Peter Zumthor’, which could then

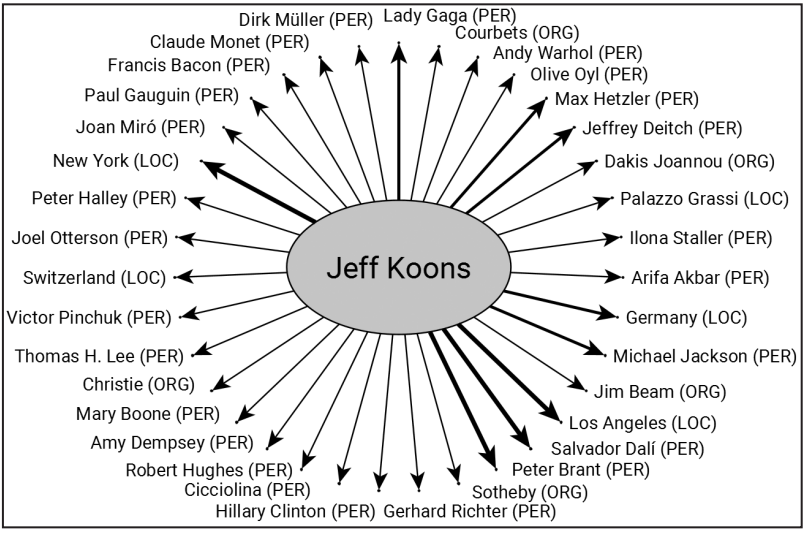

Figure 3: *Named entities extracted from the Wikipedia page of Jeff Koons (<https://en.wikipedia.org/wiki/Jeff_Koons>). Entity types: PER=person, LOC=location, ORG=organization. Arrow thickness indicates the frequency of the entity in the text. Shown is a random sample of 35 out of the 392 recognized entities.*

be used to assign tags to the article automatically. For an overview of recent NER approaches and technologies, see the survey article at YADAV/BETHARD (2018).

Example. NER can be applied to any text. In Figure B, there is a visualization of named entities that our algorithm detected in the Wikipedia page about Jeff Koons. Illustrations such as this are helpful to understand the "world" around a specified target (here, Jeff Koons). If NER is applied to many documents simultaneously, then the resulting data can be used to draw a social graph which shows the links and connections between any two targets.

Applications. HOOLAND et al. (2015) applied NER to the collection database of the Smithsonian Cooper-Hewitt National Design Museum in New York. Extracting unstructured descriptive metadata (e.g., abstracts, tables of content, or graphical representations), they explored the possibilities and limitations of NER and other techniques when mining for meaningful concepts that could be used to improve searching and browsing operations. In a quantitative analysis of precision and recall, HOOLAND et al. (2015) compared named entities retrieved by three different

third-party NER services (AlchemyAPI, DBpedia Spotlight, and Zemanta) to their own manually annotated gold-standard corpus (GSC). This comparative analysis demonstrated that despite some errors, even general-purpose NER services could provide relevant entities for cultural heritage collections at low cost.

ORAMAS et al. (2018) used NER to develop a knowledge base for flamenco music. For data acquisition, they gathered information about musical entities (e.g., artists, flamenco styles, recordings, etc.), including textual descriptions and available metadata (e.g., via Wikipedia, websites, recording databases, and other databases). They selected a pairwise classification approach based on string similarity between entity levels as an entity resolution method. Here, they asked: 'Do two entities from different sources refer to the same thing?' This procedure led to cohesion and ultimately to a reduction of duplicate entries in the database. The result of the procedure was a data knowledge base about flamenco music (FlaBase) with increased quality and cohesion of content.

Remarks. Classical NER targets only four entity types: persons, locations, organizations, and 'miscellaneous', the last of which contains all named entities which are not in the first three classes. However, there are hundreds of other named entity classes. We recently applied NER to more culture-specific entities such as events, movie titles, TV shows, and musicians (CIELIEBAK et al., 2017). Such information could be used, for example, to extract all references to artists in newspaper reviews and store them in a database. Combined with social network analysis, it would be possible to find out, for example, who frequently performs together or which artists constitute social clusters.

Method 4: Topic Modeling

Task. Topic modeling analyzes large collections of texts and produces two outputs: (i) what "topics" occur in the entire collection and (ii) what are the most dominant topics within every single text. It should be noted that topic modeling does not use predefined topics; on the contrary, it automatically detects frequently repeating topics in the data, where a topic is a collection of representative words. It is thus well suited for analyzing large document collections (usually starting with several hundred documents) to obtain an initial overview of the content. At a later stage, specific keywords and phrases can be used to analyze the data in greater detail. The related task with predefined topics is called "topic classification" and could be carried out with technologies similar to sentiment analysis. Topic modeling is helpful in determining which topics,

events, or products were most prominently discussed in a website's comment section; what aspects of a recent exhibition were valued by visitors (based on their Twitter comments); or which type of events were most often covered in the cultural sections of national newspapers. Besides, once you have a reliable topic model for the data, you can apply trend detection to track how it changes over time, which topics are booming ("trending topics"), and which are likely to disappear.

For a non-technical introduction to topic modeling, see, for example, BRETT (2013), while ALGHAMDI/ALFALQI (2015) provides a survey of implementation technologies.

Example. Figure 4 shows two topics that emerged from newspaper articles on culture and tourism. These topics were generated automatically by grouping and weighing the corresponding words based on their co-occurrences in the texts. This example uses the Reuters News dataset RCV1-v2 (LEWIS et al., 2004), which contains a total of 800,000 articles from 1996-1997. From this, we collected about 4,400 news articles related to tourism and culture by category, to which we applied latent dirichlet allocation (LDA), a standard algorithm for topic modeling.

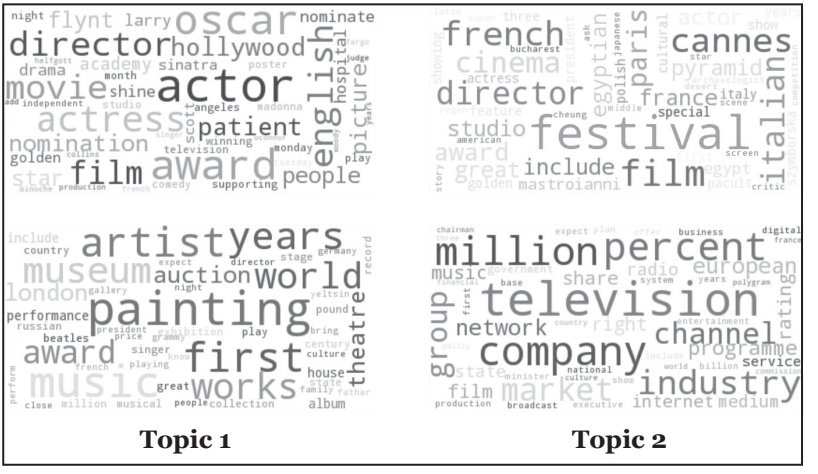

Figure 4: *Word clouds corresponding to two topics automatically extracted from the RCV1-v2 news corpus.*

Topic 1, depicted in Figure 4, is apparently related to the Oscar™ winning Film "The English Patient," whereas the second topic represents the Cannes Film Festival, which in 1997 included films by Mastroianni and Youssef Chahine (a French-Egyptian director), but also from many other countries such as Japan and Hong Kong (Cheung).

Topics for other areas such as sports, business, or weather could be similarly investigated. Armed with these topics, each new newspaper article can be analyzed and, by doing so, it would be possible to determine whether the number of cultural articles in a newspaper was increasing or decreasing over time. With more finely-tuned topics, it would even be possible to distinguish between articles about ballet, concerts, or exhibitions.

Applications. WEIJ/BERKERS (2017) focused on how people gave meaning to political music in informal, conversational settings by exploring the online reception to Pussy Riot (a political, punk-rock band from Moscow) on YouTube. They applied topic modeling to user comments to identify themes from the songs. Their findings show that the comments generally addressed (i) the geopolitical boundaries of activism, (ii) the legitimacy and commitment of the activists, (iii) the political content of the protests, and (iv) the relationship between the protests and religion.

JOCKERS/MIMNO (2013) applied statistical methods to identify and extract hundreds of topics (themes) from a corpus of 19th Century British, Irish, and American fiction. They used these topics as a measurable, data-driven proxy for literary themes and assessed how external factors such as author gender, author nationality, and date of publication might predict fluctuations in the use of themes and individual word choices within themes.

DIMAGGIO/NAG/BLEI (2013) analyzed how government sponsorship of artists and arts organizations was framed in almost 8,000 articles. These comprised all articles that referred to government support for the arts in the United States, published in five American newspapers between 1986 and 1997. Their analysis provided substantive insight into the response of the press to political attacks on the National Endowment for the Arts. They emphasize the usefulness of topic modeling in cultural sociology and legitimize it as an essential step in the overall research process.

Method 5: Trend Detection

Task. Trend detection analyzes the evolution of data over time. Originating in data science from time series analysis, it has many applications in other fields, among them also NLP, where it has immense potential in arts management. Trend detection can be applied to any of the methods mentioned above – in fact, to any time series of data. For example, it can be used to identify trending topics in culture or to detect significant

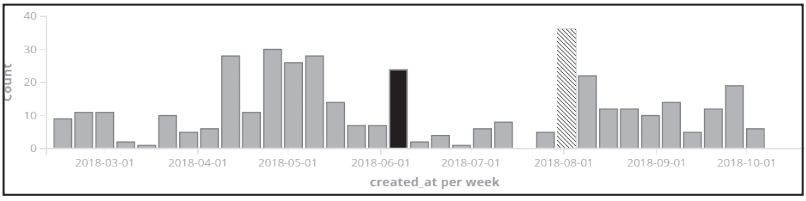

Figure 5: *Tweets about film festivals in Switzerland. Data were extracted from the Swiss Twitter Corpus (www.swisstwittercorpus.ch/) with filters for March - October 2018 and keyword "film festival" occurring in the tweet text. The black bar shows the Mobile Movies Festival (9-10 June 2018), while the stripped bar depicts the Locarno Film Festival (1-11 August 2018).*

changes in the sentiment of culture-related tweets. It can also identify long-term trends (upwards and downwards), short-term tendencies, single peaks, outliers, and similar observations in a large data series.

Example. Trend detection is especially useful when applied to social media data. For example, Figure 5 shows the number of tweets about "film festival" in the Swiss Twitter Corpus. Trend detection can find extreme peaks (such as the high-volume days in May 2018), but also isolated exceptional events, such as the higher black peak of June 2018. This could be used, for example, to detect new events or exhibitions for a particular area of arts.

ORAMAS et al. (2017) combined sentiment analysis of consumer reviews of music albums with album publication dates and the artist's biographical data to provide insights into the evolution of a musician's artistic quality and production cycles. ORAMAS et al. (2016) carried out a diachronic analysis of music criticism based on Amazon customer reviews of music albums. To detect any significant changes in the evolution of affective language in these reviews written between 2000 and 2014, they calculated an average sentiment score (i) for each year of review publication and (ii) for each review by year of album publication, including the average Amazon rating scores. In combining metadata and sentiment analysis results of these reviews, ORAMAS et al. (2016) established that there was a potential correlation between the language used in music reviews and major geopolitical events or economic fluctuations. Moreover, their analysis shows that this method can be a useful tool for understanding more about the evolution of music genres.

6. Challenges and Best Practices for NLP Applications

The wide range of NLP applications offers exciting opportunities for research as well as practical application. However, several potential pitfalls can impede a successful project. Below, we present and discuss the most significant challenges, and how to tackle or avoid them altogether. These best practices are condensed insights from more than 30 NLP projects that we conducted over the last four years in industry and academia.

Target Definition. First and foremost, every NLP project is guaranteed to fail if its expectations are unrealistic. NLP can process vast numbers of documents and achieve impressive results, but there are certain limitations which cannot be exceeded. Every NLP algorithm is prone to error (see 'Limited Capabilities' below), and any application must accept some level of inaccuracy. For this reason, it is crucial to shape the project carefully and align its aims with the technical capabilities of NLP. Here, computer scientists, NLP professionals, and domain experts need to work closely together to establish reasonable project goals. We cannot overemphasize the importance of this aspect, since in our experience, defining a manageable and realistic goal, i.e., one which can be achieved with the available technology, resources, and budget, is one of the most challenging yet essential tasks of the overall project.

Limited Capabilities. It is important to remember that NLP has limitations regarding what it can achieve. For example, state-of-the-art solutions for sentiment analysis achieve about 70% accuracy on Twitter posts - which means that 3 out of 10 tweets are being incorrectly classified. However, interpreting individual documents using sentiment analysis is not feasible. On the other hand, analyzing large numbers of documents can yield useful insights, e.g., by exploring historical changes or comparing sentiment for different topics.

Other NLP tasks have different levels of success: Named entity recognition achieves 90% accuracy but gender detection only 65%. This is sometimes due to the intrinsic difficulties of the task (e.g., the use of irony makes sentiment harder to determine), inconsistent labels (even humans can only agree on about 80% of sentiment annotations), or ambiguity in the data (e.g., 'Apple' could easily refer to a tech company, a fruit, or a music label). This last factor demonstrates a technology's inability to understand context. Current algorithms generally approach the text as stand-alone data and apply statistical analysis, mostly on

character, word, or phrase level. They ignore 'knowledge about the world', which is harder to incorporate. However, some progress has been made recently using semantic representations, contextual word embedding, and other technologies. Significant improvements in this area are anticipated in the near future.

Data Collection. Researchers should be aware that working with NLP methodologies requires a significant amount of data and, typically, much more than with qualitative, manual approaches. For supervised classification tasks such as age or dialect detection, there should be sufficient training data for every single class (e.g., 500-1,000 documents per dialect). Unsupervised methods often require a dataset that is much larger. For example, we processed word embedding for approx. 2 billion tweets, which also poses challenges to storage and computing power. Furthermore, the use of data is often restricted. Existing data protection rules and copyright laws must be considered, although many aspects of data access are unclear, both from a legal and an ethical perspective (CRAWFORD 2012). It should also be borne in mind that data collection is complicated and costly. JAMISON/GUREVYCH (2014) discusses many factors contributing to this challenge, such as training data collectors when samples are collected manually, or guiding the annotation process for these samples. This implicates that methods to assure data/labelling quality must be evaluated and implemented in order to obtain high quality predictions.

It is also critical that the 'right' data is available for the task at hand. There is a considerable choice of publicly accessible data and text corpora as well as ready-to-use solutions, although these are usually targeted towards a specific case or topic. A training sentiment analysis on millions of news documents will not help when analyzing tweets, and training NER on sports documents has little use in the cultural domain. For this reason, training data and models must be adapted to the specific domain and task.

Finally, the researcher must take into account that any NLP system can only learn what is available from the data. If the data provided is incomplete or skewed, then even the best algorithm cannot yield good results. The identification of Swiss-German dialects would not work correctly, for example, if samples from major geographical regions were missing.

Data Labeling. Manual data labeling can be a tedious task, but it is essential for generating good NLP solutions. Generally speaking, the more data available and the better the quality of the labels, the better the

outcome. In fact, the quality of any machine learning approach is limited by the quality of the training data. Data labeling requires a well-written and concise annotation guideline, describing in detail how documents should be annotated and when to assign a particular label, both with an abstract definition (e.g., 'did the author want to express a negative sentiment?') and practical examples. However, there will always be ambiguous samples which are not covered by the guideline. Consequently, after a while, the guidelines may need to be revised and some data re-labeled.

Label Quality. Even with good annotation guidelines in place, there is always a chance that the same document will be labeled differently by different annotators. This could be due to error, poor-quality work, or subjective opinion since annotators might interpret a text differently based on their subjective perception. One major issue in data labeling is, therefore, to ensure the homogenous and accurate quality of human annotation. To achieve this, a common approach is to have several annotators label the same documents (say, 3% of all documents) and then calculate the so-called inter-annotator agreement (also known as inter-coder agreement), which measures how well annotators concur. Also, the quality of a single annotator can be assessed using the self-annotator agreement (also intra-coder agreement), whereby one annotator labels the same documents more than once. Scores for these annotator agreements can vary significantly (e.g., between 0.25 and 0.80), depending on the human effort put into the labeling. It has been shown that the accuracy of, for example, sentiment analysis classifiers is typically about 0.1 points less than the annotator agreement (MOZETIČ/GRČAR/SMAILOVIĆ 2016). Thus, it is essential to have a dedicated process in place to monitor annotation quality during the labeling phase. However, even with precise labeling guidelines and proper monitoring of the annotation process, there will always be documents which obtain inconsistent labels. This is due to the intrinsic ambiguity of the data itself, where human annotators often cannot decide on the 'correct' label.

Ready-to-use Solutions. A considerable variety of ready-to-use software solutions for NLP already exist as commercial products or open source projects. They usually offer a simple interface which makes it easy to integrate them into an NLP processing pipeline. It is usually better to use an existing, established solution than implementing a new one from scratch. They are more likely to run better and perform well. However, it is necessary to ascertain that the software tool is appropriate for the task at hand and that it handles the data for analysis correctly. For example, if the tool was trained (and performs well) for sentiment analysis on

news articles, this does not imply that is will automatically be useful for processing other document types such as tweets or blog articles (DERIU et al., 2017). It might, therefore, be necessary to design and train a specific algorithm to handle the question of interest, even if this is a much larger task.

Scalability. A very practical issue is runtime. This applies to both the training and application phases. The time required to train a machine learning model is usually fast (several hours), but it can, in some cases, take up to several weeks, even on dedicated, large-scale hardware systems. This makes it hard to explore different settings, and careful experimental design is required in such cases. Conversely, the application of an NLP system to new data is typically much faster. Sentiment analysis takes a few milliseconds per document, but there are applications such as topic modeling which may need several minutes or even hours depending on the amount of data to be analyzed. While this may be acceptable in the case of a one-time analysis, it might become critical for interactive applications where researchers cannot wait that long. In this case, performance optimization, scalable hardware (in the cloud), distributed computing, or even trading accuracy for faster response times are potential solutions.

Interpretability. Another challenge in NLP projects is the question of why a particular result was generated. By using traditional qualitative research methods such as interpretative text analysis, text coding, or keyword research, explanations are given a priori by the selection of text arguments or keywords. The NLP approach is more cumulative and allows for a statistically-based identification of keywords and analysis of word and content relationships (see also DIMAGGIO 2015: 2). While these may produce similar results, it is often hard to deduce what triggered the decisions of the NLP algorithm.

Moreover, 'direct application of automated text analysis is highly successful at discovering patterns in the data, but simply reporting the output of models leaves the researcher vulnerable to oversimplification and naïve conclusions' (CHAKRABARTI/FRYE 2017: 1357). Therefore, carefully designed analytical models are essential for the generation of research-relevant insights. The output from an NLP algorithm usually indicates statistical correlations in the data, but the underlying algorithms cannot map these to causal relationships. To do this requires significant expertise on the part of the researcher. To exemplify that, some methods allow to grasp an idea of the data such as topic modelling with 3 topics (k=3), which are easy to visualize. However, it is much more difficult to evaluate the same approach with k=100. Some

approaches allow to use different resolutions or to use a hierarchical approach. Still, finding the right parameters to discover a meaningful explanation is a challenge which should be well considered.

7. Conclusion

Textual analysis has always been a central part of research and practice in arts and culture but has changed fundamentally with the advent of digitized and automated NLP methodologies.

For arts practitioners, NLP can be used to improve marketing and communication for target group analysis, event evaluation, (social) media analysis, pricing, social media optimization, advertisement targeting, or search engine optimization. In the field of archives, collections, and libraries, NLP can contribute to the improvement of indexing, consistency, and quality of databases as well as the development of suitable search algorithms. In the distribution of cultural products, online platforms can be improved and the markets analyzed.

However, applying NLP to a research or practice project in arts management often requires a high degree of skill and experience to avoid the typical pitfalls of such projects. The empowerment of cultural managers and cultural organizations in the use of NLP methods coupled with collaboration with computer science specialists from research and industry needs to be strengthened. In this paper, we have laid the foundation for such enterprises by describing fundamental technologies and approaches in NLP and giving illustrative examples from practice.

Applications in arts management research are as multifaceted as the discipline itself. They range from research into public opinion about arts funding and the evolution of art genres or artist biographies to the analysis of art criticism and research into the use of online apps at arts events. For arts management researchers, drawn primarily from the social sciences, engagement with computational text analysis requires a change in mindset:

> [It] entails more than adapting new methods to social science research questions. It also requires social scientists to relax some of our own disciplinary biases, such as our preoccupation with causality, our assumption that there is always a best-fitting solution, and our tendency to bring habits of thought based on causal modeling of population samples to interpretive modeling of complete populations. (DIMAGGIO 2015: 4)

NLP as an inductive method does not have to mean 'the end of the theory', but it will need more theoretical reflection on the construction of the big data research concepts. What kind of insights do we expect? How do we want to triangulate NLP with other methods? How should text corpora be prepared? What approaches confirm the fitness or robustness of results? And, finally, what old research questions can be dealt with in a new way using NLP methods? It is important that automated methods do not replace traditional social science and humanities analysis but exist as an additional methodology to reveal certain tendencies which can be very difficult to identify in purely manual analysis. Mixed-method research and interdisciplinary cooperation, such as in our project, are the key to applying NLP adequately to research into arts and culture, potentially eliciting new research questions.

Acknowledgments

The authors would like to thank Callum Williams for his help in generating the examples, and the reviewers for their helpful and encouraging comments.

This project has been funded by the *Internationale Bodensee-Hochschule* (IBH). It is part of an interdisciplinary project on *Strategies of Digital Cultural Communication in the Lake Constance Region* together with the following partners:

– *Center for Arts Management* at Zurich University of Applied Sciences (ZHAW), which is the project initiator and project leader; project leader is Dr. Diana Betzler (bera@zhaw.ch).

– *Institute of Applied Information Technology* at ZHAW, which selects and analyzes feasible NLP technologies; Prof. Dr. Marc Cieliebak (ciel@zhaw.ch).

– *Research Centre for Economic and Social Sciences* at Vorarlberg University of Applied Sciences (Austria), which undertake a quantitative survey; Prof. Dr. Frederic Fredersdorf.

– *Die Regionauten GbR* from Germany, which organize workshops with practitioners; Felix Pfäfflin (pfaefflin@die-regionauten.de).

– *SpinningBytes AG* from Switzerland, which provides implementations of NLP algorithms.

References

ALGHAMDI, Rubayyi/ALFALQI, Khalid (2015): A Survey of Topic Modeling in Text Mining. – In: *International Journal of Advanced Computer Science and Applications* 6/1, 147-153.

BARBARESI, Adrien (2016): Collection and Indexing of Tweets with a Geographical Focus. – In: *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC). Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia,* 24-27 <http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CMLC_Proceedings.pdf>.

BRETT, Megan R. (2013): Topic Modeling: A Basic Introduction. – In: *Journal of Digital Humanities* <http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/>.

CHAKRABARTI, Parijat/FRYE, Margaret (2017): A Mixed-Methods Framework for Analyzing Text Data: Integrating Computational Techniques with Qualitative Methods in Demography. – In: *Demographic Research* 37, 1351-1382.

CIELIEBAK, Mark (2018): *Sentiment Analysis: Distinguish Positive and Negative Documents,* blog post <https://www.spinningbytes.com/sentiment-analysis-distinguish-positive-and-negative-documents/>.

CIELIEBAK, Mark/DÄNIKEN, Pius von/FALKNER, Nicole/DOLCE, Stefano (2017): Swiss Chocolate at CAp 2017 NER Challenge: Partially Annotated Data and Transfer Learning, Zurich: ZHAW, <https://digitalcollection.zhaw.ch/handle/11475/1864>.

CONFESSORE, Nicholas (2018): Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. – In: *The New York Times* (14 November) <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>.

CORALLO, Angelo/FORTUNATO, Laura/RENNA, Clara/SARCINELLA, Marco Lucio/SPENNATO, Alessandra/DE BLASI, Christina (2017): Mobile App for Promoting Cultural Heritage: Geostatic and Textual Analysis. In: *IMEKO International Conference on Metrology for Archaeology and Cultural Heritage,* P. 9. Lecce, Italy, 194-201.

BOYD, Danah/CRAWFORD, Kate (2012): Critical Questions for Big Data. – In: *Information, Communication & Society,* 1575, 662-679.

DERIU, Jan Milan/WEILENMANN, Martin/GRUENIGEN, Dirk von/CIELIEBAK, Mark (2017): Potential and Limitations of Cross-Domain Sentiment Classification. – In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia/Spain: Association for Computational Linguistics, 17-24.

DIMAGGIO, Paul (2015): Adapting Computational Text Analysis to Social Science (and Vice Versa). – In: *Big Data & Society* 2/2, 1-5.

DIMAGGIO, Paul/NAG, Manish/BLEI, David (2013): Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding. – In: *Poetics* 41/6, 570-606.

HAUSSER, Roland (2000): *Grundlagen der Computerlinguistik*. Berlin, Heidelberg: Springer.

HOOLAND, Seth van/DE WILDE, Max/VERBORGH, Ruben/STEINER, Thomas/WALLE, Rik van de (2015): Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections. – In: *Literary and Linguistics Computing: The Journal of Digital Scholarship in the Humanities* 30/2, 262-279.

JAMISON, Emily/GUREVYCH, Iryna (2014): Needle in a Haystack: Reducing the Costs of Annotating Rare-Class Instances in Imbalanced Datasets. – In: *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation*. Phuket/Thailand: Department of Linguistics, Chulalongkorn University, 244-253.

JOCKERS, Matthew L./MIMNO, David (2013): Significant Themes in 19th-Century Literature. – In: *Poetics* 41/6, 750-769.

KESTEMONT, Mike/TSCHUGGNALL, Michael/STAMATOS, Efstathios/DAELEMANS, Walter/SPECHT, Günther/STEIN, Benno/POTTHAST, Martin (2018): Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. – In: *Working Notes Papers of the CLEF 2018 Evaluation Labs*. Avignon/France (September 10-14), 1-25.

KOPPEL, Moshe/ARGAMON, Shlomo/SHIMONI, Anat Rachel (2002): Automatically categorizing written texts by author gender. – In: *Literary and Linguistic Computing* 17/4, 401-412.

LEWIS, David D./YANG, Yiming/ROSE, Tony G./LI, Fan (2004): RCV1: A New Benchmark Collection for Text Categorization Research. – In: *The Journal of Machine Learning Research* 5, 361-397.

MOZETIČ, Igor/GRČAR, Miha/SMAILOVIĆ, Jasmina (2016): Multilingual Twitter Sentiment Classification: The Role of Human Annotators. – In: *PLOS ONE,* 11/5: e0155036 <http://dx.plos.org/10.1371/journal.pone.0155036>.

ORAMAS, Sergio/ESPINOSA-ANKE, Luis/GÓMEZ, Francisco/SERRA, Xavier (2018): Natural Language Processing for Music Knowledge Discovery. – In: *Journal of New Music Research,* 1-18 <www.tandfonline.com/doi/full/10.1080/09298215.2018.1488878>.

ORAMAS, Sergio/OSTUNI, Vito Claudio/EUGENIO, die Sciascio (2016): *Sound and Music Recommendation with Knowledge Graphs*. ACM TIST <http://10.1145/2926718>.

ORAMAS, Sergio/ESPINOSA-ANKE, Luis/LAWLOR, Aonghus/SERRA, Xavier/SAGGION, Horacio (2016): Exploring customer reviews for music genre classification and evolutionary studies. – In: *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016),* 150-156.

ROSENTHAL, Sara/FARRA, Noura/NAKOV, Preslav (2017): SemEval-2017 Task 4: Sentiment Analysis in Twitter. – In: *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, 3-4 August 2017,* 502-518 <http://www.aclweb.org/anthology/S17-2088>.

ROSSO, Paolo (2016): *Author Profiling in Social Media: The Impact of Emotions on Age and Gender*. Paper presented at the Swiss Text 2016, Zurich, June 8, <www.swisstext.org/docs/2016/talks/swisstext-2016-speaker-rosso.pdf>.

THET, Tun Thura/NA, Jin-Cheon/KHOO, Christopher S.G. (2010): Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards. – In: *Journal of Information Science* 36/6, 823-848.

VOLO. S Serena (2010): Bloggers' reported tourist experiences: Their utility as a tourism data source and their effect on prospective tourists. – In: *Journal of Vacation Marketing,* 16/4, 297–311.

WEIJ, Frank/BERKERS, Pauwke (2017): The Politics of Musical Activism: Western YouTube Reception of Pussy Riot's Punk Performances. – In: *Convergence: The International Journal of Research into New Media Technologies,* 1-20 <http://journals.sagepub.com/doi/10.1177/1354856517706493>.

YADAV, Vikas/BETHARD, Vikas (2018): A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. – In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2145-2158 <http://aclweb.org/anthology/C18-1182>.