zhaw School of Management and Law

FH Vorarlberg University of Applied Sciences

Internationale Bodensee Hochschule

# Methods of NLP in Arts Management

Mark Cieliebak, Fernando Benites, Lara Leuschen, Michaela Hnizda, Diana Betzler

## NLP in Arts Management

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. 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.

### Method 1: Sentiment Analysis

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

In Figure 1, we see the number of positives (blue) and negatives (orange) 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.

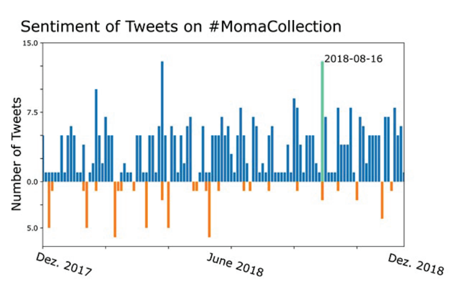

**Fig. 1:** Sentiment of tweets with"#momacollection." Blue bars depict positive tweets and orange bars depict negative tweets. Data collected between December 2017 and December 2018.

### Method 2: Topic Modeling

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 is helpful in determining which topics, events, or products were most prominently discussed in a website's commentsection; 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. Figure 2 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.

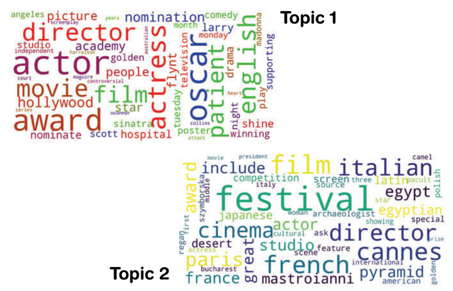

**Fig. 2:** Word clouds corresponding to two topics automatically extracted from the RCV1-v2 news corpus.

### Method 3: Author Profiling

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 from texts exploits the fact that author age, gender, and other qualities are reflected in their writing style. Author profiling could also be used to target marketing campaigns to specific groups via social media.

### Method 4: Named Entity Recognition

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. NER can be applied to any text. In Figure 3, 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.

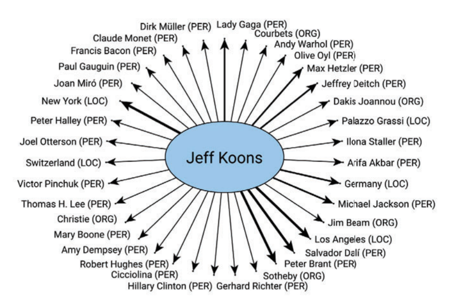

**Fig. 3:** Named entities extracted from the Wikipedia page of Jeff Koons. Entity types: PER=person, LOC=location, ORG=organization. Arrow thickness indicates the frequency of the entity in the text. Shown is a random sample of 35 out of the 392 recognized entities. https://en.wikipedia.org/wiki/Jeff_Koons

### Method 5: Trend Detection

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

### How to proceed

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 6 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.

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 Center for Arts Management at Zurich University of Applied Sciences (ZHAW), which is the project initiator and project leader; the Institute of Applied Information Technology at ZHAW, the Research Centre for Economic and Social Sciences at Vorarlberg University of Applied Sciences (Austria) as well as the practice partners Die Regionauten GbR (Germany) and SpinningBytes AG (Switzerland).