





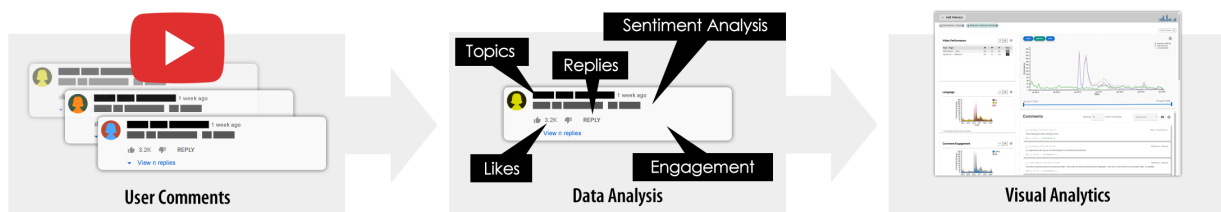


# SAMBAVis: Design Study of a Visual Analytics Tool for the Music Industry Powered by YouTube Comments

V. A. de Jesus Oliveira , C. Stoiber , J. Gröblbauer , C. Musik , A. Ringot  and A. Gebesmair 

St. Poelten University of Applied Sciences, Austria



**Figure 1:** SAMBAVis is powered by user comments from YouTube music videos and data analytics.

## Abstract

Data from comments on social media platforms offer valuable information about trends and market changes. Aiming at the music industry, we propose SAMBAVis: a visual analytics tool to handle user-generated content from comments left on YouTube music videos. SAMBAVis displays main key performance indexes, video lifecycle, and engagement with comments. It also performs sentiment analysis and extracts the main keywords from the comments, expanding YouTube capabilities. In this paper, we contribute with a design study, explaining the development of SAMBAVis and the rationale of our design. We present a usage scenario and reflect on our methods and results when creating a visualization tool for experts in the music business.

## CCS Concepts

• **Human-centered computing** → Visual analytics; Empirical studies in visualization;

## 1. Introduction

Companies have access to tools for continuous monitoring of sales and web traffic through social media monitoring (SMM). Yet, available monitoring tools differ in their depth of analysis [GH13]. Different types of SMM tools can be found on the market: a) Analysis tools of respective platforms, only for owned profiles [fac20b, you20b, spo20, goo20, the20a]; b) Artist-Account-Based Monitoring - across platforms [the20b, sou20a]; and c) Keyword-based monitoring - across platforms [tal20, bra20, lin20]. With YouTube Analytics [you20b] (a), for instance, content creators can monitor their channels and videos through different reports, such as the number of views, watch time, traffic sources, and demographic reports. However, it is fairly restricted to the artist content in the corresponding platform. On the other hand, *NextBig-Sound* [the20b] (b) monitors different social media platforms. It delivers data related to audience reach, metric trend, audience engagement, and artist social stage. Yet, it only shows relative values benchmarked against indexed artists and for a given date range. Finally, tools like *Talkwalker* [tal20], *Brandwatch* [bra20], and *Link-*

*fluence* [lin20] (c) can be very powerful in their paid-versions. Still, only a few of these SMM platforms focus on unstructured user-generated content based on artists or songs. Semi- and unstructured data from the Internet, such as blogs, websites, and comments on social media offer valuable information about trends and changes in markets [FMB13]. Therefore, it is important to make such data exploitable for strategic and operative management of labels, music publishers, and radio stations. Therefore, we propose SAMBAVis (smart data for music business administration visualization) as an extension of an analysis tool (such as YouTube Analytics), which is Artist- and/or Song-Based and focused on user comments.

In this paper, we present a design study [SMM12] on SAMBAVis, which is an open-source visual analytics (VA) tool focused on time-oriented and user-generated content based on comments from YouTube. Over the past three years (2017-2019), we have cooperated with experts in the fields of media economics and visualization; understanding the problem with the help of interviews and workshops with experts in the music industry; as well as designing and evaluating SAMBAVis in a workshop with experts in 2019.

In this design study, our main *contributions* are:

- a problem characterization based on semi-structured interviews with music experts (Sec. 2);
- the design of SAMBAVis, describing our visual analytics tool on the song and artist levels (Sec. 3);
- a usage scenario showing the application of SAMBAVis as an exploratory tool on the “Austrian Ibiza affaire” and the upturn of the Vengaboys’ “We’re Going to Ibiza” song (Sec. 4);
- the results and lessons learned from a workshop and usability test with experts, artists, and influencers (Sec. 6).

## 2. Problem Characterization and Abstraction

In the first phase of our design study, we conducted interviews to obtain a good understanding of our target group and requirements.

### 2.1. Interviews with Domain Experts

Semi-structured interviews [LFH09] were conducted with domain experts in 2017 to collect our first set of design requirements.

**Methods & Participants:** Five domain experts participated in the interviews: three Indie Labels (Artists and Repertoire - A&R); one Major Label A&R; and one Concert Agency (Head of Public Relations, Marketing, Promotion). All of the surveyed experts were male and had more than 10 years of experience. The interviews took approximately one hour, and they were performed in person and in German. We investigated how managers in the music business work, how the process is structured, which data and tools are used, and what potential problems accompany these tools. One important aspect was to find out the role social media plays in their daily work routines and which platforms are used in practice.

**Results from Interviews:** All interviewees have a strong interest in social media marketing to promote their own contracted artists and to monitor and evaluate the artists’ actions on social media platforms. The main questions they try to answer are *Who’s my fanbase?*, *Where are they?*, *What do they like (next to the music)?*, *How can I better approach them?*, *How can I monetize?*. In addition, music managers use social media and the Internet to gain information about external or new artists. Their first step is to run a *search* for artists and songs. They report using Google to search for what kind of online presence artists have and where, how do the artists communicate there, and how do fans engage with them.

Commercial SMM tools are used for different purposes by the interviewees, e.g. all labels use in-house platforms for analytics concerning contracted artists, ForTunes [for20] is used by Indie and Major labels, Social Blade [soc20] is used to estimate YouTube revenues by Indie labels, Hootsuite [hoo20], NextBigSound [the20b] and Soundchats [sou20a] are used by an Indie label; NextBigSound is used for ticket sales data including demographics by the Concert agency. In addition, Facebook [fac20a], Instagram [ins20] (especially for live concerts and as a networking tool), YouTube [you20a], Soundcloud [Sou20b], and Twitter [twi20] (especially to follow trends) are mentioned as the most important social media platforms. They mention that SMM is very time-consuming and emphasize that tools for social media monitoring need to be *easy* to use.

Interviewees mention that “*metrics are not the result, but the beginning of analysis and strategic management*”. They highlight the importance of metrics such as organic growth as the most suitable in the field contrarily to clicks and views. They comment that these metrics are dependent on artists’ stage, scene, genre, territory, aims, etc. A crucial task is to translate data into insights and then into action. They also comment that a smart interlinking of complex big data is important. Moreover, it is a process of learning how to interpret these data. All interviewees also emphasize the importance of recommendations from the personal network and that long-term mutual trust is crucial in the field. Besides, digital demos and concerts are an important source of A&R. A&R Major label mentions that “*In my work, the human and interpersonal level is a very, very big factor*”. Decisions are not only made by metrics and number, but the human and interpersonal level is also a factor in decision making. Thus, both the *input* from the public as well as long-term *trust* in the data used by the SMM tools are crucial.

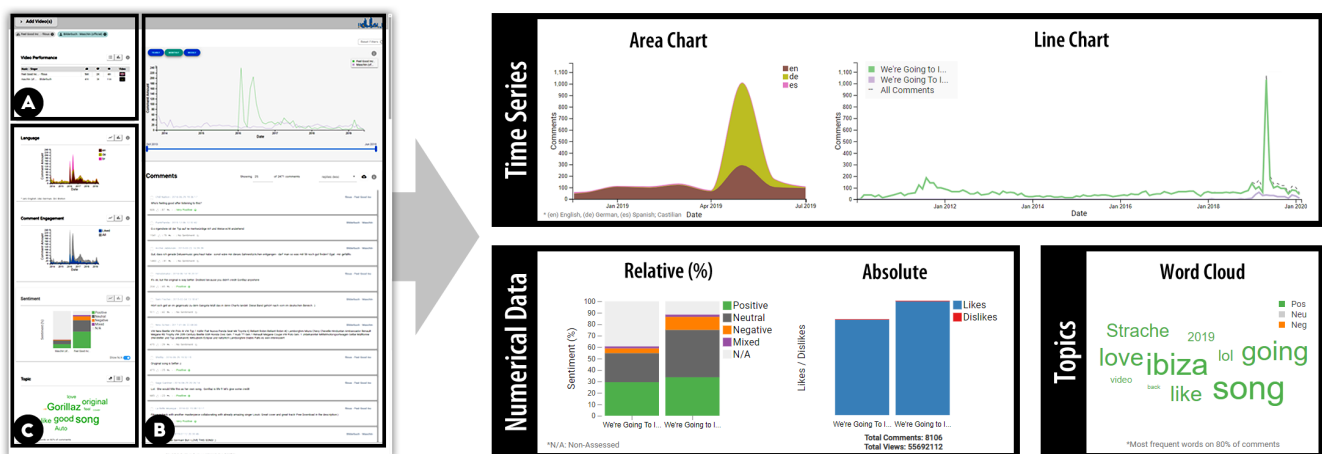
### 2.2. Design Requirements

Based on the results from the interviews, we identified three key functional and non-functional requirements (R):

- **R1 – To manage time-oriented and unstructured social media data:** The interviews reveal that a data analysis tool is still needed to analyze public comments (user-generated content) from a platform such as YouTube to convey the temporal development and correlations between metrics and these comments, which is not fully supported by other tools. Text mining techniques can extrapolate such unstructured data and provide potentially more significant analytics [GH13]. For instance, user-generated content provides the semantics of users’ behaviors [CLY17]; users manifest their sentiments by posting negative or positive messages [ZGWZ14]; the public’s opinions and attitudes can also be translated into topics or keywords [AGCH11, LYW\*16].
- **R2 – Search for artists and songs:** Due to the interviewees experience using search engines, a practical search interface needs to be provided for searching for artists, songs, and their related comments and metrics.
- **R3 – Ease of use:** The interviewees emphasized that tools for social media monitoring need to be easy to use. The interpretation of the provided metrics is not easy, so it needs to be clear and support learning and understanding it.
- **R4 – Data transparency:** The interviewees mentioned that long-term trust in the data used in SMM tools is crucial. Therefore, when using information coming from user comments, the raw data should still be accessible. Providing access to the original input data help users to check the validity of the analysis results and to gain trust in the system.

## 3. Conceptual Design and Implementation

The SAMBAVis main workflow is shown in Figure 1. First, a Python script is used to retrieve data via the Youtube API. Then, metadata about videos, users, comments, and replies, as well as text mining results, are stored in an ArangoDB database as JSON documents. Finally, an interface is implemented with AngularJS, which communicates with our database.



**Figure 2:** SAMBAVis interface. (left) The main areas of our dashboard. (right) The main visualizations used according to our data [SAM20].

The different interface elements are organized in cards (Figure 2, left). Each card contains its own set of controls and guidance (R3). The visualizations are connected by “brushing and linking” techniques [Kei02]. The cards are grouped into three main areas:

**A. Videos:** A button on the top left side is used to search and add videos to the dashboard. Users can search by song or artist (R2), and add videos or a group of videos from different channels corresponding to the same song. A card in this area displays information about the selected videos and their key performance indexes (KPIs), such as likes, dislikes, and views. Users can be redirected to the original video on YouTube from this card (R4).

**B. Comments:** To display the temporal development of the comments (R1) for each selected video (R2), we use a line chart [Har99], which is common to similar tools (R3). The number of comments can be aggregated by year, month, and day. Time filter can be applied using a slider or brush, which is reproduced on all other cards. Under this chart are the YouTube comments. The user can read the comment directly on the SAMBAVis or be redirected to its source on YouTube (R4). Comments can be sorted by replies, likes, and dates. They can also be filtered by time.

**C. Analytics:** Multiple cards display data directly from the comments. Each card includes an overview window, which aggregates data from all selected videos, and a detailed window, grouping data by video. Currently, cards in this area display information about *language*, *engagement*, *sentiment*, and *topics*. Both *language*, *engagement*, and *sentiment* cards display categorical data. To convey the proportion of each category over time, an overview is provided by an area chart [Har99]. To convey the proportion of each category per video, details are shown by stacked bar charts [Har99]. *Language* is classified with LangID [LB12] and uses a standard categorical color scheme. *Engagement* shows the number of comments (in grey) and how many of them were liked (in blue). *Sentiment* discriminate between positive (green), neutral (dark grey), and negative (orange) comments. We also identify “Mixed” comments (purple) detected as both positive and negative by different tools [Nie11, LB02, Lor18]. Non-English comments were not as-

essed (light grey). Finally, the *topics* card shows the most frequent words in the comments (i.e. keywords [LYW\*16]). An overview is provided by a word cloud [VWF09] and details by lists colored according to the sentiment of the comment they were extracted from.

#### 4. Observation of Real World Usage

As a first validation step, we present a usage scenario [SMM12]. To demonstrate the capabilities of SAMBAVis, we analyzed the video “We’re Going to Ibiza” by Vengaboys to observe the effect of the Ibiza Affair. The Ibiza affair (or Ibizagate [EHP20]), was a political scandal in Austria involving Heinz-Christian Strache, the former vice-chancellor of Austria and leader of the Freedom Party. On May 17, 2019, two German media outlets published a video of Strache at a resort in Ibiza, triggering the scandal [EHP20]. The same day, many users started associating the scandal with the song. In Figure 2 (right), two videos from “We’re Going to Ibiza” are selected in SAMBAVis. One is the official video released in the Vengaboys YouTube channel. The main line chart shows a noticeable peak for the official video in May 2019 with 1034 comments. In 2019, the video received a total of 2134 comments. Since its release ten years ago, it has received on average 601 comments per year (SD: 275.2). The area chart shows that most of these comments are in German (the yellow portion in Figure 2 (area chart)). Also, the word cloud shows that “Strache”, “2019”, and “video” are the most frequently occurring words. Yet, the majority of comments are not negative, as they are mostly provocative such as “*If ur here because of H.C.Strache, hit like. XD*” and “*The new Austrian National Hymn*”. Liked comments went from 9.5% before May 2019 to 41% showing particular interest in the discussion on the topic.

#### 5. Final Workshop and Usability Evaluation

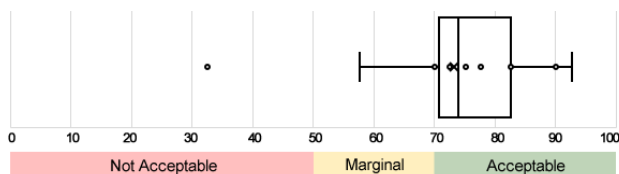
In our next validation step, we performed a second workshop with domain experts. The workshop took place at Music Austria (MICA) in June 2019. With three hours of duration, the goal of the workshop was to introduce the project, provide an overview of SMM tools, introduce SAMBAVis, and perform a usability evaluation.

**Table 1:** Performance of users with the SAMBAVis during final workshop

Question	Answer/Accuracy
Task 1: Search and select the song “Feel Good” from the artist Filous.	
1 How easy it was to find the songs?	6.8 (SD: 0.4) *
2 How easy it was to select/add the song?	5.9 (SD: 1.0) *
3 Which month (M) and year (Y) had the most comments?	M: 33.3%, Y: 58.3% **
4 What is the language of the majority of the comments?	100% **
5 What is the keyword that appears more frequently in the comments of this song?	100% **
6 How easy it was to find the official video?	6.8 (SD: 0.4) *
Task 2: Search the song “Maschin” from the artist Bilderbuch and add to the selection only the official video version.	
7 Which video has more likes?	100% **
8 Which video has more dislikes?	100% **
9 Which video received the highest proportion of negative comments?	100% **
10 Overall, how easy it was to compare the information between two songs/videos?	5.2 (SD: 1.5) *
<i>Mean and Standard Deviation for a 7-point scale [1-7: Very Difficult to Very Easy]</i> *	
<i>% of Correct answers</i> **	

**Methods & Participants:** The workshop had 14 participants (5 male, 9 female), from 23 to 55 years old. Among them, seven artists (musicians and a DJ), five managers (management of artists and booking, agency, label, events, and marketing), one social influencer, and one expert working into different areas of the Austrian music industry. They reported using Hootsuite (3x), ReverbNation (2x), Next Big Sound (1x), ForTunes (1x), Social Blade (1x), Swat.io (1x), sprinklr (1x), and individual corporate solutions (1x). For the workshop, our dataset was composed of 4081 YouTube videos from 1,236 songs of 17 Austrian artists. That means 132,953 comments and 42,645 replies. 49.5% of the comments were in English. For the usability test, participants were asked to perform three tasks: 1) search for a given music video, and answer questions about its data; 2) add a second video to the dashboard and compare data from both videos; and 3) explore the tool and search for any videos they like. After the tasks, they filled out a SUS Questionnaire [BKM08] and shared their impressions about the tool.

**Results:** Results are shown in Table 1 and Figure 3. Users could not properly assess time-oriented data (Table 1.3). Area charts were too small, while the line chart needed extra tooltips, contrast, and clearer labels. Participants, non-familiar with such interaction techniques, did not understand the use of brush as well. Their opinions and results were used to improve SAMBAVis: area charts were integrated into the main line chart; time filter is now applied by slider only; the dashboard was optimized to display content on-demand, and; topics are now better positioned in the dashboard as it was addressed by the participants as our main unique selling proposition.

**Figure 3:** SUS Scores for SAMBAVis ( $M: 73.1$ ,  $SD: 15.9$ ).

## 6. Reflection and Limitations

From our first interviews, we understood that music professionals consider collected data only as suggestions. That led us to develop a simple tool, aiming at providing a fair amount of data and leaving the interpretations to the users. The capabilities of SAMBAVis in providing insights are shown in our usage scenario (Sec. 4). It provided ways for understanding the reasons behind changes in the video lifecycle. Such knowledge should help professionals in marketing their content and engaging with their audience.

Our evaluation focused on the Austrian music industry as a use case because of its limited breadth and good access to experts. Yet, many of our data mining tools were optimized for English content. Besides, we did not explore more advanced types of visualization as the music professionals were not visualization experts. Overall, defining requirements for the domain experts for our tool was a challenge as our users can have different roles such as managers, artists, and social media influencers. Still, our final assessment shows that SAMBAVis is usable and effective for them.

## 7. Conclusion

In this design study, we presented the design, development, and evaluation of SAMBAVis over the past three years. We briefly reported on our problem characterization, the design of SAMBAVis, a usage scenario, and the results from a final workshop with experts. After this study, the interface of SAMBAVis was updated taking into consideration our results and feedback. Our next steps include further documentation and promotion of SAMBAVis as an open-source tool for music professionals.

## 8. Acknowledgments

This work was supported by the Austrian Research Promotion Agency (FFG - COIN program) via the "SAMBA - Smart Data for Music Business Administration" project (FFG 856363). Special thanks go to Dr. M. Zeppelzauer and M. Steinacker for their support and participation in this study.



## References

- [AGCH11] ARCHAMBAULT D. W., GREENE D., CUNNINGHAM P., HURLEY N. J.: Themecrowds: multiresolution summaries of twitter usage. In *SMUC '11* (2011). 2
- [BKM08] BANGOR A., KORTUM P. T., MILLER J. T.: An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction* 24, 6 (2008), 574–594. 4
- [bra20] Brandwatch. <https://www.brandwatch.com/de/>, 2020. Accessed: 2020-02-17. 1
- [CLY17] CHEN S., LIN L., YUAN X.: Social media visual analytics. *Comput. Graph. Forum* 36, 3 (June 2017), 563–587. URL: <https://doi.org/10.1111/cgf.13211>, doi:10.1111/cgf.13211. 2
- [EHP20] EBERL J.-M., HUBER L. M., PLESCIA C.: A tale of firsts: the 2019 austrian snap election. *West European Politics* (2020), 1–14. 3
- [fac20a] Facebook. <https://facebook.com/>, 2020. Accessed: 2020-02-17. 2
- [fac20b] Facebook Insights. <https://www.facebook.com/business/insights/tools/audience-insights>, 2020. Accessed: 2020-02-17. 1
- [FMB13] FRIEDRICHSEN M., MÜHL-BENNINGHAUS W.: Handbook of social media management : value chain and business models in changing media markets. 1
- [for20] ForTunes. <https://www.fortunes.io/>, 2020. Accessed: 2020-02-17. 2
- [GH13] GRÜBLBAUER J., HARIC P.: Social media monitoring tools as instruments of strategic issues management. In *Handbook of social media management*. Springer, 2013, pp. 671–687. 1, 2
- [goo20] Google analytics. <https://analytics.google.com/analytics/web/>, 2020. Accessed: 2020-02-12. 1
- [Har99] HARRIS R.: *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, 1999. 3
- [hoo20] Hootsuite. <https://hootsuite.com/de/>, 2020. Accessed: 2020-02-17. 2
- [ins20] Instagram. <https://www.instagram.com/>, 2020. Accessed: 2020-02-17. 2
- [Kei02] KEIM D. A.: Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (Jan 2002), 1–8. doi:10.1109/2945.981847. 3
- [LB02] LOPER E., BIRD S.: Nltk: the natural language toolkit. *arXiv preprint cs/0205028* (2002). 3
- [LB12] LUI M., BALDWIN T.: langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations* (2012), Association for Computational Linguistics, pp. 25–30. 3
- [LFH09] LAZAR J., FENG J. H., HOCHHEISER H.: *Research Methods in Human-Computer Interaction*. John Wiley & Sons, Chichester, UK, 2009. 2
- [lin20] Linkfluence. <https://www.linkfluence.com/de/>, 2020. Accessed: 2020-02-17. 1
- [Lor18] LORIA S.: textblob documentation. *Release 0.15.2* (2018). 3
- [LYW\*16] LIU S., YIN J., WANG X., CUI W., CAO K., PEI J.: Online visual analytics of text streams. *IEEE Transactions on Visualization and Computer Graphics* 22, 11 (Nov 2016), 2451–2466. doi:10.1109/TVCG.2015.2509990. 2, 3
- [Nie11] NIELSEN F. Å.: A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903* (2011). 3
- [SAM20] SAMBAVis Video Presentation. <https://phaidra.fhstp.ac.at/download/o:4003>, 2020. Accessed: 2020-04-07. 3
- [SMM12] SEDLMAIR M., MEYER M., MUNZNER T.: Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics* 18, 12 (2012), 2431–2440. 1, 3
- [soc20] Social Blade. <https://socialblade.com/>, 2020. Accessed: 2020-02-17. 2
- [sou20a] Soundcharts. <https://soundcharts.com/>, 2020. Accessed: 2020-02-17. 1, 2
- [Sou20b] Soundcloud. <https://soundcloud.com/>, 2020. Accessed: 2020-02-12. 2
- [spo20] Spotify Analytics. <https://analytics.spotify.com/>, 2020. Accessed: 2020-02-12. 1
- [tal20] Talkwalker. <https://www.talkwalker.com/de>, 2020. Accessed: 2020-02-17. 1
- [the20a] the Echo Nest. <http://the.echonest.com/>, 2020. Accessed: 2020-02-17. 1
- [the20b] Next Big Sound. <https://www.nextbigsound.com/>, 2020. Accessed: 2020-02-17. 1, 2
- [twi20] Twitter. <https://twitter.com/>, 2020. Accessed: 2020-02-17. 2
- [VWF09] VIEGAS F. B., WATTENBERG M., FEINBERG J.: Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (Nov 2009), 1137–1144. doi:10.1109/TVCG.2009.171. 3
- [you20a] Youtube. <https://www.youtube.com/>, 2020. Accessed: 2020-02-17. 2
- [you20b] Youtube analytics. <https://www.oberlo.com/blog/youtube-analytics>, 2020. Accessed: 2020-02-12. 1
- [ZGWZ14] ZHAO J., GOU L., WANG F., ZHOU M.: Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct 2014), pp. 203–212. doi:10.1109/VAST.2014.7042496. 2