Multimedia Systems (2024) 30:25
https://doi.org/10.1007/s00530-023-01195-7

REGULAR PAPER

AI and data-driven media analysis of TV content for optimised digital content marketing

Lyndon Nixon[1] · Konstantinos Apostolidis[2] · Evlampios Apostolidis[2] · Damianos Galanopoulos[2] · Vasileios Mezaris[2] · Basil Philipp[3] · Rasa Bocyte[4]

Received: 30 May 2023 / Accepted: 8 December 2023 / Published online: 19 January 2024

Abstract
To optimise digital content marketing for broadcasters, the Horizon 2020 funded ReTV project developed an end-to-end process termed "Trans-Vector Publishing" and made it accessible through a Web-based tool termed "Content Wizard". This paper presents this tool with a focus on each of the innovations in data and AI-driven media analysis to address each key step in the digital content marketing workflow: topic selection, content search and video summarisation. First, we use predictive analytics over online data to identify topics the target audience will give the most attention to at a future time. Second, we use neural networks and embeddings to find the video asset closest in content to the identified topic. Third, we use a GAN to create an optimally summarised form of that video for publication, e.g. on social networks. The result is a new and innovative digital content marketing workflow which meets the needs of media organisations in this age of interactive online media where content is transient, malleable and ubiquitous.

Keywords Digital content marketing · Topic prediction · Video retrieval · Video analysis · Video concept detection · Video summarisation

1 Introduction

The former gradual shift of media consumption from linear/broadcast to non-linear/online has become more of a flood as people have stayed home during the COVID-19 pandemic, spending more time in front of connected screens and wanted especially to consume both more news and more entertainment. Nielsen's Streaming Meter for 2020 found a 75% increase in cumulative weekly time spent streaming video compared to one year earlier. Streaming has been adopted particularly by those aged 55+, with Nielsen saying "now that they are here and entered the streaming realm, that behaviour is likely to stay".[1] The all important question for media organisations has to be: as consumers shift in their media consumption behaviour, would they be watching our content?

Communicated by B. Bao.

✉ Lyndon Nixon
lyndon.nixon@modul.ac.at

[1] MODUL Technology, Am Kahlenberg 1, 1190 Vienna, Austria

[2] CERTH-ITI, 6 km of Charilaou-Thermi road, 57001 Thermi-Thessaloniki, Greece

[3] Genistat AG, Seestrasse 30, 8806 Bäch, Switzerland

[4] Netherlands Institute for Sound and Vision, Media Parkboulevard 1, 1217 Hilversum, The Netherlands

[1] The Nielsen Total Audience Report: August 2020. https://www.nielsen.com/us/en/insights/report/2020/the-nielsen-total-audience-report-august-2020/.

Digital video content marketing refers to the digital marketing activities related to the promotion of the video assets of an organisation on digital channels. This can be either to directly acquire video views or click-throughs (from a summary or trailer) to the full video. Both broadcast and streaming media organisations as well as media archives promote their media assets on alternative digital channels (social media, Websites, apps) with the goal to attract viewers to their own channels (broadcast TV, Web streaming, Over-The-Top media services or collection portals). In doing so, they are forced to compete with a huge and growing scale of native video content on these channels (i.e. audiovisual assets made available to be directly consumed in the channel). Digital marketing on these channels, therefore, has to address the massive challenge to get the attention of an audience already overwhelmed by content choice and availability. Many organisations have already shifted to providing media content on these channels for direct consumption, typically shorter form content (possibly summarised from a longer original asset), e.g. AJ+, which provides short form videos around news stories from the Al Jazeera news channel, has become the second largest news video publisher on Facebook.[2]

Many organisations, therefore, now face the challenge of the cost and complexity of adapting their existing media asset collections and publication workflows to this new age of "interactive media". As media consumption shifts to online, social and mobile, it is obvious that media publishing and marketing must also make the shift. In this paradigm shift in media consumption, short form video is growing the fastest in popularity and social networks have become video discovery platforms. The 24-hour news cycle prompts media organisations to respond to constantly emerging stories. While this presents opportune moments to (re)publish relevant media assets, it also requires significant efforts in monitoring the wide range of digital platforms to understand what type of content would appeal to an audience at any particular moment. As these changes in consumer behaviour persist, media organisations—especially the traditional broadcasters and media archives—need more than ever new tools for digital video content marketing to ensure that their media assets can be found on digital channels in the right form at the right time to attract the audience's attention.

In this paper, we present "trans-vector publishing". This is a future digital content marketing workflow where the same content can be semi-automatically republished in different forms on different digital channels at the optimal publication time for each one. In the EU funded project ReTV,[3] we have developed the technical infrastructure for this, which we call the "Trans-Vector Platform" (TVP). Platform functionalities can be accessed through a user-friendly Web interface in a tool we call "Content Wizard" which enables digital content marketing teams to follow the end-to-end trans-vector publication workflow.

In Sect. 2, we present the state of the art in digital video content marketing and contemporary AI-based research contributing to this area. In Sect. 3, we introduce the Content Wizard tool, its architecture and the new and innovative Research and Development which enables its support for trans-vector publishing. The Research and Development covers (i) machine learning-based prediction of which topics the target audience will give most attention to at a future time of publication, (ii) an embeddings layer for matching video contents to text for accurate textual search over a video collection without sufficient textual metadata, and (iii) a method for the automatic summarisation of videos for a target digital channel. Section 4, presents the results of Content Wizard's evaluation with media professionals. Finally, we conclude with an assessment of our contribution to digital content marketing workflows through the AI and data-driven media analysis our new components enable and the outlook for the future from the perspective of media organisations.

2 State of the art

We refer to this new process for media organisations to "publish to all media vectors with the effort of one" as Trans-Vector Publishing. This can be compared to the state of the art in digital content marketing, typically manual content selection, preparation and publication workflows for each channel, often even with different teams working with different content to promote. These are time-consuming (and therefore costly) activities. In Fig. 1 the trans-vector publishing workflow is illustrated:

Here, millions of data points—Web pages, social media posts, blog entries etc.—are monitored, annotated and analysed to identify the topics of discourse among audiences on different digital channels. Predictive analytics is used with the collected data to identify trending topics in the future. Videos are retrieved from a media collection which are relevant to the topic through a text-to-video embeddings layer and re-purposed for the selected digital channels using a GAN. The social media post is prepared with the re-purposed video and scheduled for publication.

In the last years, the market for trans-vector publishing has expanded and at least a dozen tools have come to market which cover some parts of this workflow. Typically they allow users to prepare and schedule content (text and image) to post to social networks, but without the automatic trend prediction, topic-based video retrieval and digital

[2] https://variety.com/2015/digital/news/how-al-jazeeras-aj-became-one-of-the-biggest-video-publishers-on-facebook-1201553333/.

[3] https://www.retv-project.eu.

Fig. 1 Trans-Vector Publishing Workflow (WPn refers to the workpackages of the ReTV project which implemented each step)

channel-specific summarisation developed in ReTV. Later,[4] Facelift,[5] Tailwind,[6] Buffer,[7] hootsuite,[8] Sprout Social[9] and Falcon.io[10] are examples of well established tools in this area. They lack integration with Web and social media monitoring (to inform users of the topics their audiences are currently attentive to) as well as tools for video management (access to collections in archives or on a Multimedia Asset Management System; browsing or search via available or extracted metadata; rapid video re-purposing particularly to focus on pre-selected topics and to target social media channels). Content Wizard goes beyond the state of the art of the current tools, both by predicting the future topics of interest to the audience (as opposed to showing the topics from past content) and by automatically retrieving and re-purposing videos according to these topics and the target publication channel.

Predictive analytics for identifying future topics of interest for a media audience is a new research area. Separate social media and Web listening tools are usually used to track metrics of *past* discourse by media audiences (social media, Web fora), especially reaction to (e.g. sentiment) and opinion towards past media postings by the same organisation to learn audience preferences and optimise future publications. Such tools include Buzzcapture, Brandwatch, Meltwater, Nexis Newsdesk and Sysomos. They do not consider topics covered in the online sources their audience typically listens to, although a posting related to a currently popular topic among its target audience could be expected to get more attention. Digital marketers need to rely on trends from the recent past to decide the optimal topic for a posting in the near future; planning further into the future is based on known calendar events rather than predicting what would be interesting to the audience at that future time.

Forecasting has been common for predicting potential future TV audience size for a given program to be broadcast at a given time. Starting with time-based methods like ARIMA which use past patterns and the assumption of those same patterns occurring into the future [1, 2], researchers have also considered if other features correlate with changes in audience such as the demographic/behavioural segmentation of the audience [3] or social media metrics (frequency of mentions of the TV brand, or sentiment towards it) [4]. Past research has suggested correlations exist, e.g. "for 18–34-year-olds, an 8.5% increase in Twitter volume

[4] https://later.com/.
[5] https://facelift-bbt.com/.
[6] https://www.tailwindapp.com/.
[7] https://buffer.com/.
[8] https://www.hootsuite.com/.
[9] https://sproutsocial.com/.
[10] https://www.falcon.io/.

corresponds to a 1% increase in TV ratings for premiere episodes" [5]. The likes, shares and comments on TV show pages on Facebook or tweets and retweets on Twitter may also be indicators of the shows popularity and correlate to viewing figures of the next episode broadcast [6, 7].

Such work to date appears to suggest that social media metrics are a viable feature for a prediction model, but do not address the issue of the bulk of video content which is not subject to a critical mass of social media discussion or engagement. Furthermore, we consider the prediction of the audience on digital publication channels such as social media rather than the broadcast TV channel and based on the topics of the content rather than the content (e.g. TV series) itself, cf. Section 3.2.

Separate tools than those previously mentioned for social media publication are available for the task of video editing, especially the capture and re-purposing of scenes from (near-live) television. These include tellyo, grabyo and Wildmoka. Over the last few years, many AI-based technologies were introduced to help curate media content in general [8]. Such technologies include concept and event detection [9–11], video summarisation [12–17], and cross-modal modelling of video and text [18–23]; and, they have found their way in supporting/automating tasks in applications that offer optimised publication of audiovisual content across digital media vectors. Examples of such applications include domain-specific video analysis [24, 25] and online video summarisation tools [26, 27], content adaptation for archival media [28, 29], aspect ratio adaptation [30, 31] and TV content adaptation [32, 33]. For a comprehensive survey of television content adaptation methods (a.k.a personalisation in this context) the interested user is directed to [34] . In the Context Wizard, we employ selected state-of-the-art AI-based technologies, specifically the cross-modal retrieval of [35], which integrates a self-attention mechanism in both visual and textual modalities, and variations of this approach have been utilised in several cross-modal systems and in benchmarking activities [36–38]. We also employed the video summarisation method of [16], which is the first unsupervised method that embeds an Actor-Critic model in a Generative Adversarial Network, and formulates the selection of important video fragments (that will be used to form the summary) as a sequence generation task. Both selected methods were developed in the context of ReTV and in this paper, we present their integration into an interactive user interface, with such a unified interface being unique in the literature, at least to our knowledge. The cross-modal video retrieval and video summarisation methods are discussed in Sects. 3.3 and 3.4, respectively.

3 Content Wizard

3.1 Overview of the architecture

Content Wizard is an extension of a professional-grade tool for social media marketing, integrating the new components for trans-vector publishing developed in the ReTV project. The currently manual, labour-intensive task of selecting and adapting video content for publication to a growing set of digital channels (vectors) can be supported and, in many cases, automated by the newly developed online components which communicate with each other via secure Web services to cover all aspects needed:

- *Listening* - monitoring the audience's social media discourse and public online discussion. This helps to contextualise media content publication and transformation as a central KPI (Key Performance Indicator - success metric) for the media organisation.
- *Prediction* - for current world events and trending topics, predict the most relevant topics for content publication, at the best time (scheduling), and along the highest-impact vector.
- *Adaptation* - for an existing media asset, allow fully automatic content transformation for various audiences (personalisation) and different publication vectors (repurposing).

The aforementioned set of components forms a distributed platform following a service-oriented architecture, which we have called the Trans-Vector Platform (TVP).[11] The Content Wizard is the user facing application on top of this architecture, re-using some of its components as illustrated in Fig 2.

The *Prediction Service* component uses collected metrics about past topics of online discussion to predict the next trending topics (cf. Sect. 3.2). These predictions guide both the recommendation of which content to publish and the scheduling of when it should be published; for this, the *Recommendation and Scheduling* component makes use of both the Binary Repository containing the video assets and a Video Feature Storage where extracted metadata for each video asset is stored (containing structural metadata like scenes, shots and sub-shots as well as descriptive metadata like concept and object detection). Topics of audience interest are matched to topics in video assets using a text-to-video matching module (cf. Sect. 3.3). The *Video Adaptation and Re-purposing* component takes a video asset and produces a modified copy according to the topic(s) to focus on and the target publication channel while aiming to retain

[11] ReTV project deliverable D4.3 "Trans-Vector Platform, Technology Roadmap and Revised Prototype, Final Version", available at https://retv-project.eu/deliverables/.

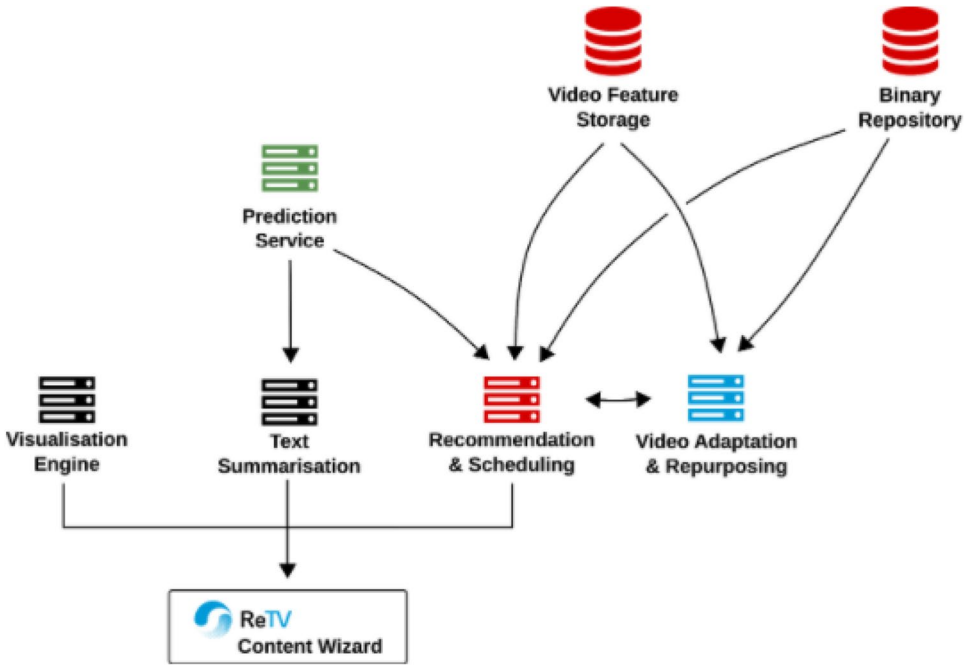

Fig. 2 Trans-vector platform components used by the content wizard application

as much as possible the original content of the video asset (cf. Sect. 3.4). The remaining components ensure that data visualisations are available to the Content Wizard user to guide their decision making in the workflow (Visualisation Engine) and that the final (digital video) posting can be published at the optimal time (cf. Sect. 3.5).

3.2 Trending topic detection

Trending topic detection is a part of predictive analytics, i.e. the use of algorithms, increasingly from the machine learning domain, to predict the future value of a chosen variable based on its past values and optionally additional features which exhibit a correlation to its values. To be able to offer a prediction, appropriate data has to be available. The online discourse on both Web and social media vectors has been collected by the webLyzard platform for several years for the news domain, while in the ReTV project we began both crawling TV/radio-specific Websites and social media monitoring for TV/radio related terms.[12] Our pipeline performs keyword and entity extraction from the text of the collected documents; we then calculate metrics related to communication success for each of those keywords or entities such as frequency (of mentions) and sentiment (based on the surrounding text). Since data collection takes place continually, these metrics can be expressed as time series data at daily or hourly granularity. We consider sets of keywords and entities as "topics" (from specific topics such as "Eurovision" which can include alternative labels - "Eurovision Song Contest" or "Grand Prix Eurovision" - as well as common hashtags or abbreviations, e.g. "ESC", to broad topics such as "cultural events" which might encompass matches with museum exhibitions, art gallery vernissage, concerts etc.). For any topic we can extract the time series data for its frequency of mentions for at least the past two and a half years.

Our heuristic is that the frequency of mentions of a topic on a channel positively correlates with the reach and engagement for content about that topic (posted on the same channel at the same time), just as tweets are considered more likely to get more reach and engagement when associated with a trending topic. We consider topics—sets of keywords—instead of single keywords in the prediction as we recognise that there are constantly new keywords emerging in the online discourse for which broader topics can still be effective (e.g. "non-fungible token" (NFT) may not be present in past online documents but "blockchain" would be). We use the frequency metric as the target variable for prediction of topic presence (and, by corollary, popularity) by channel and time, so that we could suggest the optimal time for publication of content about that topic on that channel. By comparing a group of topics on the same channel for a certain time, we can also suggest the optimal topic for which to choose or create content for publishing at that time.

Following linear regression and ARIMA based approaches reported in previous work [39, 40], we chose to look at LSTMs (Long Short Term Memories) which have been reported as being particularly effective for the forecasting task [41, 42]. We wanted to perform multi-step forecasting directly from the prediction model so we looked

[12] ReTV project deliverable D1.1 "Data Ingestion Analysis and Annotation", available at https://retv-project.eu/deliverables/.

Table 1 Table comparing accuracy of different prediction models for the topic "cycling"

Cycling	Val MAE	Val RMSE
SARIMAX	15.3	18.5
LSTM E-D AR(7, 1)	22.1	26.1
LSTM E-D TS AR(7, 1)	21.1	23.6
LSTM E-D (200, 30)	26.8	28.1
LSTM Seq2seq + attention with Swish (200, 30)	**9**	**10.4**

Our final implementation with best results in bold

at seq2seq models which were also being reported in the literature as the best performing LSTMs for multi-step forecasting [43]. Considering current research, we tested with the addition of attention mechanisms (Luong attention [44]), which further improved the accuracy. Finally, we also experimented with different activation functions and found a function proposed by Google engineers called Swish [45] to be the superior approach for our prediction task. Our final LSTM model, based on comparison with different configurations of accuracy with different evaluation datasets (based on keyword frequencies extracted from the webLyzard platform), is an Encoder-Decoder model with input–output sequences of length (200, 30), using Luong attention and Swish activation function.

Prediction accuracy for this LSTM model is illustrated below in Table 1 and Fig. 3 with one evaluation dataset for the topic "cycling" (frequency metric from global news media, 995 data points each where 1 data point represents 1 day). We provide MAE and RMSE metrics on the 'val'(idation) data which is the next 30 days of actual values (not in the input dataset for training/testing), where a comparison of predicted values is made to the 7-day moving average of the actual values (as explained in,[13] this better captures audience attention, as a recently present topic is not immediately forgotten). MAE and RMSE metrics are reported as the average from 5 runs. Different models are compared: E-D refers to Encoder-Decoder, AR to autoregression for single-step forecasts and TS is use of a TimeSeriesGenerator (from the Keras library for deep learning). The input–output sequences used are (7,1) and (200,30). Our SARIMAX model is also given as a comparative benchmark (baseline). The chart shows the similarity between the best performing model and the actual 7-day moving average. While this is just one evaluation, we tested across multiple topics. The main requirement for a good result was that the topic has not emerged recently and exhibits seasonality, i.e. past trends are indicative of future ones. Sports, for example, works well as the peaks and troughs of sports coverage—tied to events—tend to be similar each year (even when the average values have been lower for 2020 with the cancellations of sports events, there was still online discussion about the events around the time when they would have taken place).

Given that time series forecasting requires sufficient, relevant past data, we had to generate and pre-process almost three years of past data to predict the next 30 days for the

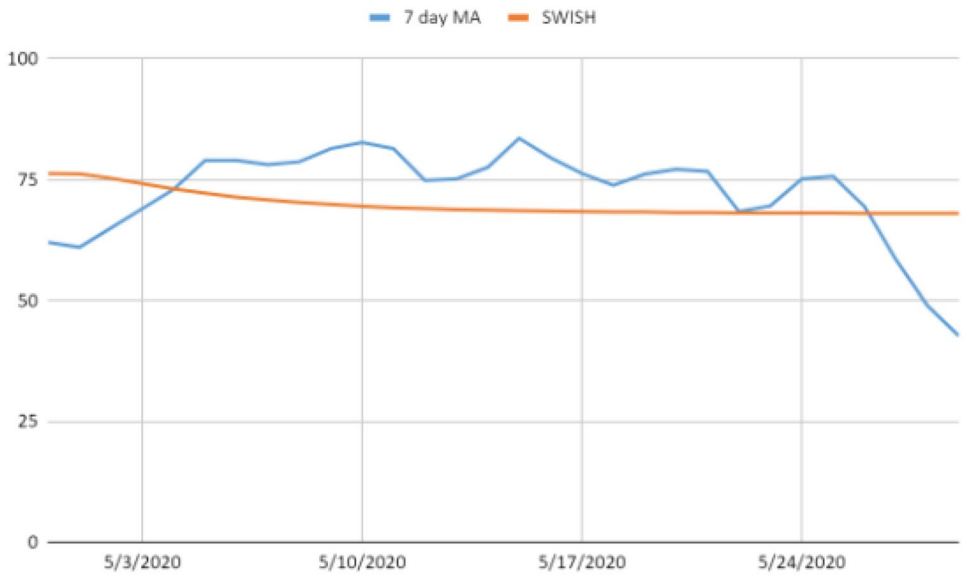

Fig. 3 Chart comparing best-performing prediction to an actual 7-day moving average for the topic "cycling" (x-axis: date (next thirty days), y-axis: frequency of mention (target variable for prediction))

[13] ReTV project deliverable D2.3 "Metrics-based Success Factors and Predictive Analytics", available at https://retv-project.eu/deliverables/.

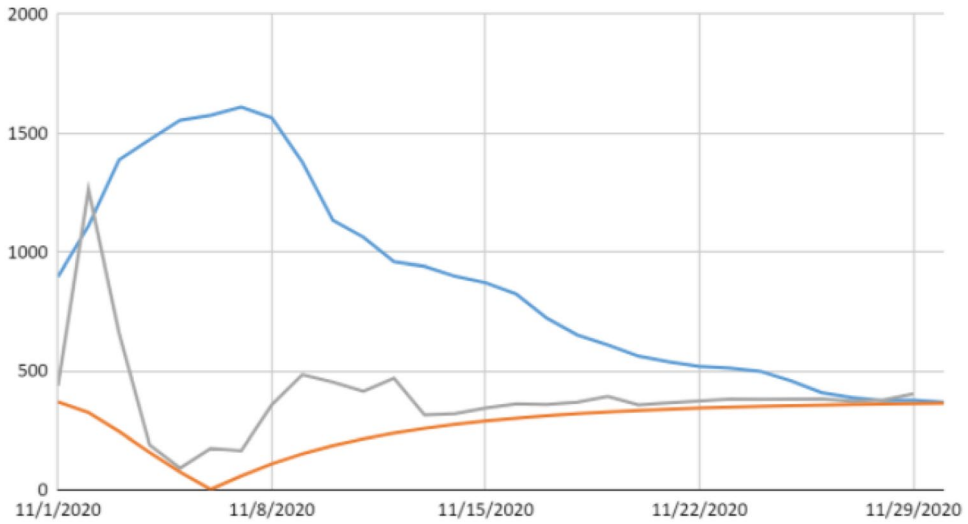

Fig. 4 Chart comparing prediction values, with (grey) and without (orange) temporal references, and actual values (blue) for 'elections' (x-axis: frequency, y-axis: date)

LSTM evaluations (making it impossible to apply this approach for predicting e.g. the topic coronavirus which would encounter almost zero data for the period before January 2020). Here, shorter time periods could be used for training with single-step auto-regressive models when topics are 'recently emerged'. However, this would still not be able to anticipate future appearances of the topic unrelated to past trends, e.g. due to a new event already known to be taking place in the future. We have, therefore, also processed all past documents to detect textual references to future dates, meaning we can also calculate a metric for the number of documents that associate a given topic with a given date. This metric, known as "temporal reference detection" is used in the prediction model as an additional feature to anticipate future spikes in mentions of a topic based not on past trends (which may not be present) but on a future event identified in the document corpus. The topic "elections" may be explored as an example of a future "out-of-trend" event. Consider a prediction task at the end of October 2020 for the next 30 days (being November 1–30, 2020) based on training with past data from 13 months (October 1, 2019 – October 31, 2020). Since an election is typically every four years, for example, the training data will not contain any frequency peaks for the topic "elections" (no election took place), and if there is an election in the prediction period, it can not be anticipated from this past data—as would be the case with the 2020 US election on November 3, 2020.

To evaluate, we use the frequency metrics from global English language news media for training, testing, and validation with the prediction time period of November 1–30, 2020. The training/testing dataset (past 13 months) has 397 data points. The validation dataset has 30 data points and uses the 7-day moving average of the frequency metric. As seen in Fig. 4, the predicted values from our best performing model (orange line), which are based on the "usual" news cycle without major election events, are much lower than the actual as the forecasting is unable to take into account the US election on 3 November. Looking at our "temporal reference detection" metric for the topic in the month of November, there is a significant peak on 3 November (as we would expect, with many documents referencing "election" and the 3rd November together) and generally a much larger association of the topic with dates in the first half of the month. The blue line is the real (7-day moving average) values for the topic "elections' in November. If we consider the addition of the prediction values with the temporal reference detection metrics (grey line), it can be seen that the predictions come closer to the actual values and interestingly the predicted value for 3 November is almost the same as the actual. The subsequent predicted values do drop off heavily from the actual values but it would probably have been impossible to anticipate the longer period of focus on the election several days after election day itself due to ongoing counts and disputes.

We conclude that the temporal reference detection metric can be valuable in combination with the LSTM-based forecasting model to incorporate out-of-trend peaks in topic popularity that would otherwise not be part of our predictions. We can then use the prediction model to identify which topic(s) can be expected to attract more attention from an online audience on some future date (and thus recommend a relevant video for publication). In an on-the-fly calculation situation, we can provide predictions almost immediately using only the temporal reference detection metric. Forecasting using LSTMs is more resource intensive and it is therefore not feasible to expect the same on-the-fly responsiveness yet such predictive capabilities can be provided using pre-calculations when the user knows the specific topics for which they want to forecast popularity over a future date range. The resulting prediction model enables organisations

Table 2 Experimental comparison of the employed cross-modal video retrieval method, ATT-ATV, with published SoA methods, on the AVS16, AVS17 and AVS18 datasets

Method	AVS16	AVS17	AVS18
VSE++ [47]	0.123	0.154	0.074
Video2vec [18]	0.087	0.150	–
W2VV++ [55]	0.151	0.220	0.121
Dual encoding [46]	0.165	0.228	0.117
Dual-task [56]	0.185	0.241	0.123
SEA [21]	0.164	0.228	0.125
Extended Dual Encoding [23]	0.159	0.244	0.126
ATT-ATV [35]	0.159	0.244	0.126
ATT-ATV re-training	**0.202**	**0.281**	**0.146**

MXinfAP is used as an evaluation measure (higher values are better). Bold indicates the best scores

to anticipate future moments of heightened interest in any preselected topic, with the intention to publish relevant content at that time such that it has optimal chances to achieve maximal reach and engagement metrics.

3.3 Cross-modal video retrieval

Cross-modal retrieval is used for performing a retrieval task across different modalities such as image-text, video-text, and audio-text. In Content Wizard, we employ such techniques to address the text-video cross-modal retrieval task. Given a set of unlabeled video shots and an unseen textual query, this task aims to retrieve video shots from a media collection ranked from the most relevant to the least relevant shot for an input textual query.

Inspired by the dual encoding network presented in [46], the ATT-ATV network that encodes video-caption pairs into a common feature subspace is created [35]. This network utilizes attention mechanisms for more efficient textual and visual representation and exploits the benefits of richer textual and visual embeddings. Let **V** be a media item (e.g., an entire video or a video shot) and **S** the corresponding caption of **V**. Both **V** and **S** are translated into a new common feature space $\Phi(\cdot)$, resulting in two new representations $\Phi(\mathbf{V})$ and $\Phi(\mathbf{S})$ that are directly comparable. For this, two similar modules, consisting of multiple levels of encoding, are utilised for the visual and textual content, respectively. As discussed in detail in [35], the levels of encoding that each module consists of are: (i) mean-pooling, (ii) attention-based bi-GRU sequential model, and (iii) CNN layers; and, the improved marginal ranking loss [47] is used to train the entire network. The produced representations are highly effective in generating a ranked list of relevant video shots, given a free-text query of what the user is looking for. The ATT-ATV network is trained using a combination of four large-scale video captioning datasets: MSR-VTTT [48], TGIF [49], ActivityNet [50] and Vatex [51]. In [35], for calculating the visual embeddings that are used as input to the visual module, a ResNext-101 network (trained on the ImageNet-13k dataset) and a ResNet-152 network (trained on the ImageNet-11k dataset) were used. Also, two word embeddings were used as input to the textual module: (i) a Word2Vec model trained on the English tags of 30K Flickr images, and (ii) a pre-trained language representation BERT, trained on Wikipedia content. Here, to improve the performance, we re-train the ATT-ATV architecture using additional visual and textual features, coming from the trained CLIP model (ViT-B/32) [52]; both visual and textual embeddings are calculated, by utilising the corresponding CLIP image or textual encoder. Table 2 presents the results of the utilised (re-trained) ATT-ATV network and comparisons with state-of-the-art literature methods, including the initial ATT-ATV network of [35], on three Trecvid AVS benchmark datasets (AVS16, AVS17, and AVS18) [53]. The presented results are those reported in the original paper of each method. We use the mean extended inferred average precision (MxinfAP) as an evaluation measure, as proposed in [54] since it is the typical evaluation measure when working

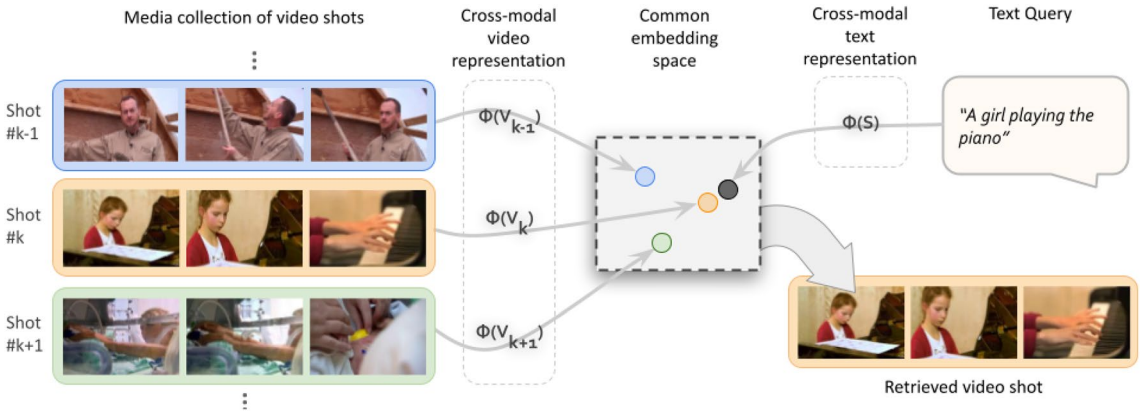

Fig. 5 An example of retrieving a relevant video shot to a text query utilising cross-modal representation

with these datasets. Our method outperforms all competitors on every dataset.

In the context of video content discovery in the Content Wizard, we focused on making it easier for editors to find videos in their collections that match the topics predicted to be more popular at the planned publication time. Specifically, given a text as search input, we want to return a set of relevant videos from the user's media collection (see Fig. 5). We pass the text (e.g. the set of labels associated with the topic chosen for the content publication) into the newly developed text-to-video search module. The text of the search query is mapped into a common embedding space with the video shots from the media collection, both represented via 2048-dimensional vectors. We compare the embedding vector of the text to all of the embedding vectors of the video shots that have been computed for each video at the time that it is ingested in the archive and return a ranked list of the closest matches.

We use Elasticsearch to store and query these embedding vectors, using the special, optimised field type DenseVector. Elasticsearch also efficiently supports the cosine similarity metric between the vectors to find the shots that are closest to our text embedding, where we only need to look for the n closest shots so the search does not have to be exhaustive. As we have an embedding vector for each shot of a video, we get a matching score for each shot yet some shots may come from the same video so we have to aggregate the results. We highlight how the scores of different shots are aggregated can change the search results. Let us assume we have two videos, each containing ten shots. The first video contains a scene that discusses a dinosaur movie and nine other scenes with no relation to the cinema or dinosaurs. The second video is a sequence of 10 shots of a dinosaur museum. For a search on dinosaurs and film, one could argue that the scene in the first video is the most relevant, but the second video would usually be ranked as more relevant on average due to many more shots of the video being close to our search. Therefore, we decided to aggregate the distance metrics for each video shot in such a way as to rank the first video higher based on our observation that social media managers prefer short content that is highly relevant to longer content with lower relevance. Our cross-modal video retrieval uniquely prefers the video with the single highly relevant shot whereas traditional systems would rank the other video with more relevant shots, due to our use case being the selection of suitable short form content for social media publication.

3.4 Video summarisation

Video summarisation aims to generate a concise synopsis of a full-length video that retains the most significant parts. Based on this, viewers can have a quick overview of the whole content without having to watch the entire video. In the context of Content Wizard, a new video summarisation method is utilised to create a media asset out of a longer form video into a topic-focused summary in the optimised short form for publication on social media.

Table 3 Comparison (in terms of F score (%)) of the top-5 unsupervised video summarisation approaches of the literature, on the SumMe and TVSum benchmarking datasets

	SumMe		TVSum		
Method	F score	Rank	F score	Rank	Avg. Rank
SUM-GAN-AAE [59]	48.9	5	58.3	5	5
SUM-GDA$_{unsup}$ [60]	50.0	4	59.6	2	3
CSNet+GL+RPE [15]	50.2	3	59.1	3	3
CSNet [61]	**51.3**	**1**	58.8	4	2
AC-SUM-GAN [16]	50.8	2	**60.6**	**1**	**1**

Bold indicates the best scores

In this Table, the "Rank" value indicates the relevant ordering of the methods, according to the obtained F score, from the best-performing one (Rank=1) to the worse (Rank=5) among these top-5 performers; and, the "Average Rank" indicates the ordering (from 1, for the best-performer, to 5) according to the average of the Rank values for the two datasets. Data source: [58]

In the core of our newly developed summarisation pipeline is an AC-SUM-GAN, an unsupervised deep-learning-based method that embeds an Actor-Critic model into a Generative Adversarial Network. The Actor and the Critic take part in a game that incrementally leads to the selection of the video key-fragments. Their choices at each step of the game result in a set of rewards from the Discriminator, according to the proximity of video-level representations of the original and the summary-based reconstructed version of the video in a learned latent space. The applied training workflow allows the Actor and Critic to discover a space of actions and states, and automatically learn a value function (Critic) and a policy for selecting the most important parts of the video (Actor). The pre-trained model of AC-SUM-GAN that is integrated into the Content Wizard, gets as input deep representations of the video frames that are obtained using the output of the pool5 layer of GoogleNet [57]. It is composed of the parts of the network architecture that are used at the inference stage, namely: (a) a linear layer for dimensionality reduction (from 1024 to 512), (b) the State Generator (formed using a 2-layer bi-directional LSTM with 512 hidden units) that formulates the state for the Actor's choices, and (c) the Actor (formed using four fully-connected layers that are followed by a softmax layer) that progressively picks 15% of the video fragments, based on the generated state and a categorical distribution of probabilities. Further details about the network architecture and the processing pipeline at the training and inference stages can be found in [16]. According to a recent survey on deep

learning methods for video summarisation [58], AC-SUM-GAN is ranked first (on average) among the top-5 unsupervised video summarisation approaches of the literature. As shown in Table 3, AC-SUM-GAN produces state-of-the-art results on two benchmarking datasets (SumMe and TVSum), reaching an F score that exceeds 50% on SumMe and 60% on the TVSum dataset (following the relevant established evaluation protocols).

To further adapt the results of this automatic summary generation to the video content retrieved by our previously described component, the analysis outcomes of AC-SUM-GAN are additionally combined with a set of rules and requirements about the content and characteristics of an optimal summary for social media, devised by considering requirements identified in the use cases of the ReTV project (e.g. "avoid talking-head shots") as well as the feedback provided by test users from ReTV partners in response to indicative summarisation results that were shown to them. As a result of this, the final video summarisation pipeline implements the following rules:

- discourages the selection of video segments with
  - blurry content
  - visual effects
  - content related to specific parts of a TV news program
  - extreme camera movement
- discourages the use of two or more visually similar segments
- supports the use of pre-defined filters (i.e. visual concepts) to adjust the content of the summary by filtering out segments depicting unnecessary or inappropriate content (e.g. avoid the parts that show the anchor person when summarising a news video)
- supports the use of user-defined parameters to regulate the length and the rhythm (i.e. the pace of segments changing) of the produced summary

To incorporate these additional rules into the video summarisation process, we proceeded as follows. First, the video is analysed by the initial video summarisation method which produces a set of frame-level scores indicating the importance of the visual content of each frame. Then, taking into account the extracted information about the shot segments of the video, we calculate shot-level importance scores by averaging the scores of the frames within each shot. Following this, we rank the shots based on the computed shot-level scores, thus producing a first ranking of the shots. Subsequently, we compute another ranking after filtering out less appropriate shots based on the set of predefined rules, making use of various hand-crafted features and concept detection results. The hand-crafted features include the Edge Change Ratio (ECR) measure and the variance of the Laplacian of each frame—which is used as a measure of the blurriness of an image—to avoid the selection of shots with extreme camera movement. Regarding concept detection, we employed publicly available pre-trained models for the ImageNet [62] and Places365 [63] datasets, and trained a new model of the EfficientNet B3 architecture [64] on the TRECVID SIN dataset [65], thus generating detection scores for more than 1.5K concept labels. We then selected and utilised specific concept labels to detect and exclude shots that include anchorpersons or related concepts (e.g. TV studio, TV talk show, speaking to camera, etc). Finally, the computed rankings are combined by averaging the rank of each video shot to form the final ranking.

To produce the summary, the shot with the highest rank (i.e. the most appropriate for being included in the video summary) is selected as a candidate and is removed from the ranked list. We then check if the next candidate shot is visually similar to the set of the already selected shots for inclusion in the summary, skipping it if it is and selecting it if not. To assess the visual similarity of shots, we compare their distance (computed based on their representations in intermediate layers of a CNN we use for processing the video for visual concept detection) against a pre-defined *min_visual_distance* threshold. Following this, two different pre-defined thresholds *min_shot_duration* and *max_shot_duration* are utilised to filter out very short and very long shots, in such a way to control the pace of change between video segments. The values of *min_shot_duration* and *max_shot_duration* were set as 1 and 5 seconds respectively. If the duration of a candidate shot is greater than *min_shot_duration* and lower than *max_shot_duration*, then the whole shot is included in the summary. However, if this duration is greater than *max_shot_duration*, then we select a sub-part of the shot that lasts *max_shot_duration* seconds, by maximising its importance based on the computed frame-level importance scores by the video summarisation method [16]. We continue until the desired duration of the video summary is reached.

The aforementioned shot selection process is different from the one implemented in [16], where the summary is constructed by solving the 0/1 Knapsack problem. The new shot selection process enables the use of a user-adjustable *rhythm* parameter to modify *min_shot_duration* and *max_shot_duration* thresholds. Specifically, setting a $rhythm < 1$, since we multiply *max_shot_duration* to *rhythm*, the value of *max_shot_duration* results being below 5 seconds, consequently leading to the selection of smaller segments of the original video for the summary. Thus a fast-paced summary is constructed, i.e. with quickly alternating shots. Respectively, setting $rhythm > 1$, results in a summary with a slower rhythm of alternating shots. Additionally, plainly solving the knapsack problem would often select very small segments to fulfil the time budget requirements, as observed during

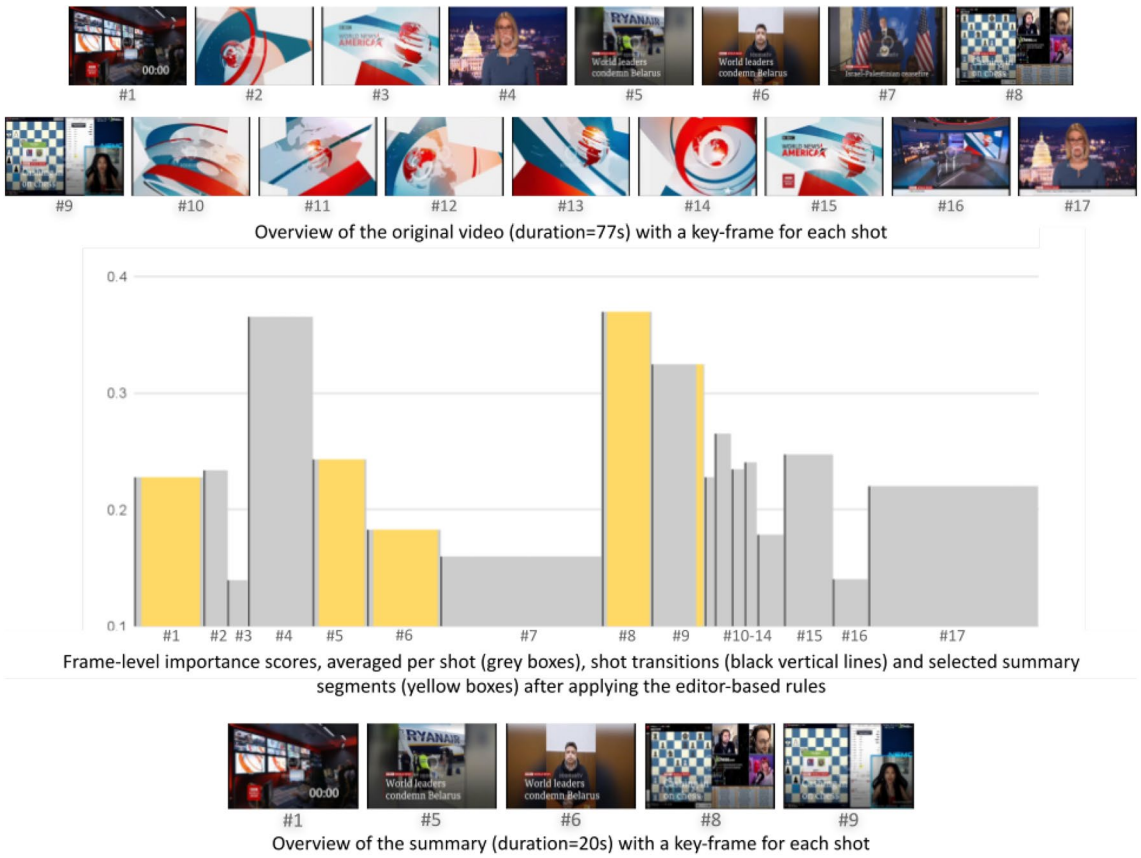

Fig. 6 A key-frame-based overview (using one key-frame per shot), and an example summary

preliminary experiments. This is avoided by employing the shot selection process discussed above.

An indicative example that demonstrates the impact of the additional rules developed in the ReTV project for video summarisation (suitable for social media publication) is presented in Fig. 6. The original video, which contains the start of a BBC news broadcast (the original video can be viewed at [14]), is 77 s long, for which we aim to generate a 20 s summary (e.g. for TikTok or an Instagram Reel). Constructing a summary, with the selection of segments being based solely on the AC-SUM-GAN method's frame importance scores, will select shots #4, #8, #9, #15 and #17, considering that a) the shots are the top-5 scoring shots (notice the height of the respective grey boxes at the middle of Fig. 6), and b) these shots' total duration suffices to construct a 20-second summary. However, employing the editor-based rules introduced in ReTV, shots #4 and #17 are excluded since they contain frames of an anchorperson, while shot #15 is excluded due to the included synthetic graphics (see the key-frames of the respective shots at the top of Fig. 6). Therefore, our algorithm selected segments from shots #1, #5, #6, #8, #9 to generate the summary (see the key-frames of the selected shots at the bottom of Fig. 6).

3.5 Content Wizard user interface

Content Wizard is implemented as a Web-based tool built on top of the social media management tool Levuro Engage.[15] Levuro Engage is a modern web application using React[16] for the front end. The backend is distributed among multiple servers. The processing of uploaded videos is done using AWS encoding servers and the processed videos are stored in S3 buckets. This distributed approach scales well with an increasing number of videos. The social media and Web monitoring backend deployed by webLyzard[17] which provides the documents used in the topic prediction uses Elastic indexes to scale up to support hundreds of millions of documents. The ReTV project did not only develop the innovative backend components for transvector publishing presented in the previous sub-sections which are called via

[14] https://youtu.be/49s6FybJDqQ.

[15] https://levuro.com/.

[16] https://reactjs.org/.

[17] https://weblyzard.com.

Fig. 7 Trending topics user interface in the Content Wizard. The predicted popularity of the topic ”James Bond” in the next four weeks is shown

Fig. 8 Text-2-video search user interface in the Content Wizard. Videos which closest match the search term are shown

REST interfaces by the Content Wizard but also Web-based user interfaces to interact with their functionality within the application.

The Content Wizard works with the video collections of the user. Since the process that determines the predicted popularity of topics is not instantaneous, several topics of interest may be predefined by the user and their predictions pre-computed for immediate access, with the user being able to request a recalculation using the latest Web and social media documents. The extraction of the text-to-video embeddings is also time-consuming, so we also pre-compute embeddings for the video collection and make them available in a database that allows for fast querying (also the user can request recalculations when new videos are added to the collection). To begin the transvector publication workflow, the user can look for future dates when a topic of interest to their audience is predicted to peak in interest, or simply see which of their topics are predicted to be rising in popularity on a specific future date. This view is available in the “trending stories” interface of the Tools tab of the Content Wizard (Fig. 7 shows the predicted popularity of the topic “James Bond” calculated on 25 March 2021 for the next four weeks.

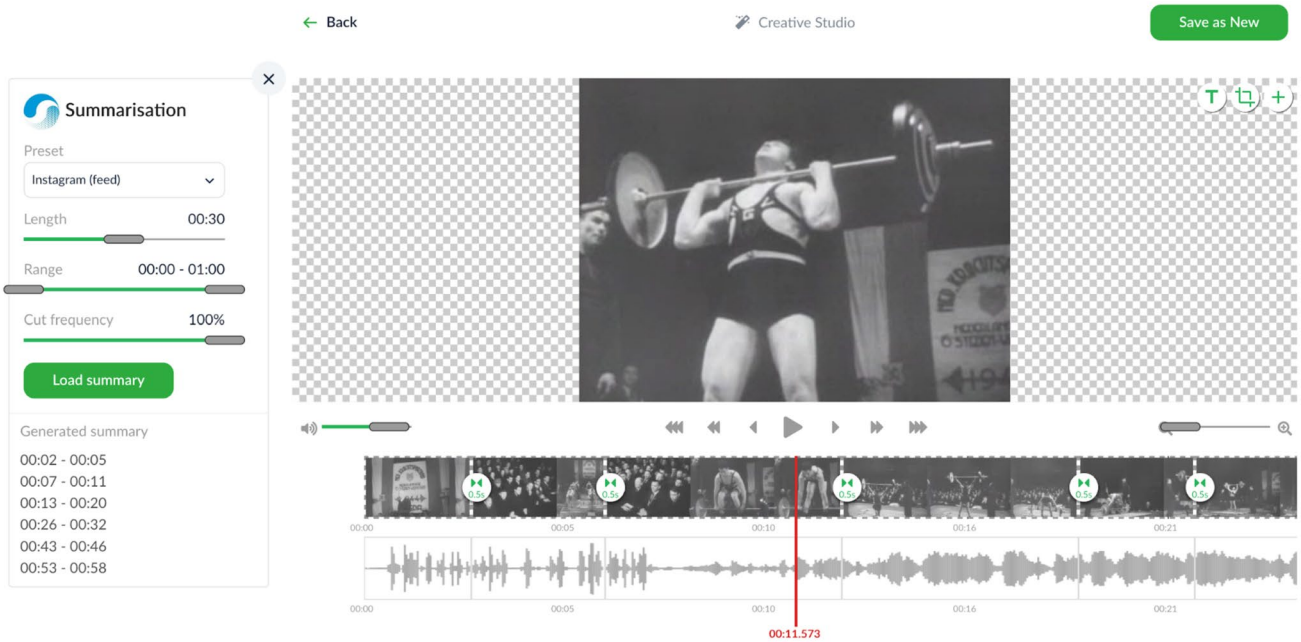

Fig. 9 Video summarisation user interface in the Content Wizard. Different presets are supported for common social media types

A peak in interest is predicted on 2 April 2021 - which was the originally planned release date of the next James Bond film "No Time To Die").

Since the process of identifying relevant topics and matching them to appropriate videos in the users media collection is often laborious, a user can select a topic in the "trending stories" interface and choose directly the presented option "Video Search". We immediately show the most related videos from their collection based on the keywords used by the topic as the textual input into the text-to-video search. The text-to-video search turns the list of keywords into an embedding vector and compares this embedding vector to all of the embedding vectors of video shots that are available in the collection and returns the best matches. For example, the videos matched to the topic "James Bond" all relate to film or cinema despite the search input not mentioning either term (see Fig. 8), an association which has been learnt by the embeddings layer. The user of the Content Wizard can then select the video, summarise it if desired using the ReTV video summarisation component, and post it on social media with an accompanying message.

When the user selects a video for editing in the Content Wizard, they are given the option to have it automatically summarised. The summarisation is then requested from the newly developed *Video Adaptation and Re-purposing* component. The video analysis process necessary for the video summarisation has been run in advance for all videos in the collection and the extracted features are stored in the Video Feature Storage, therefore the main summarisation process is almost instant (again, when new videos are introduced, the user can request their analysis so that their features are also available to the summarisation step). Figure 9 illustrates how the video editor cuts the video into the segments of the original video that the *Video Adaptation and Re-purposing* component proposed. The user can still make manual adjustments if so desired. Through this process, the video content is optimised for a particular publication vector.

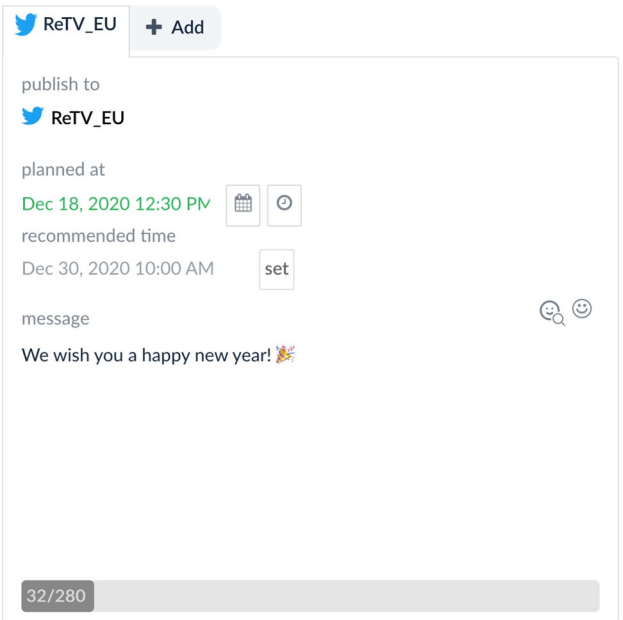

Fig. 10 Choosing a publication time for a social media post

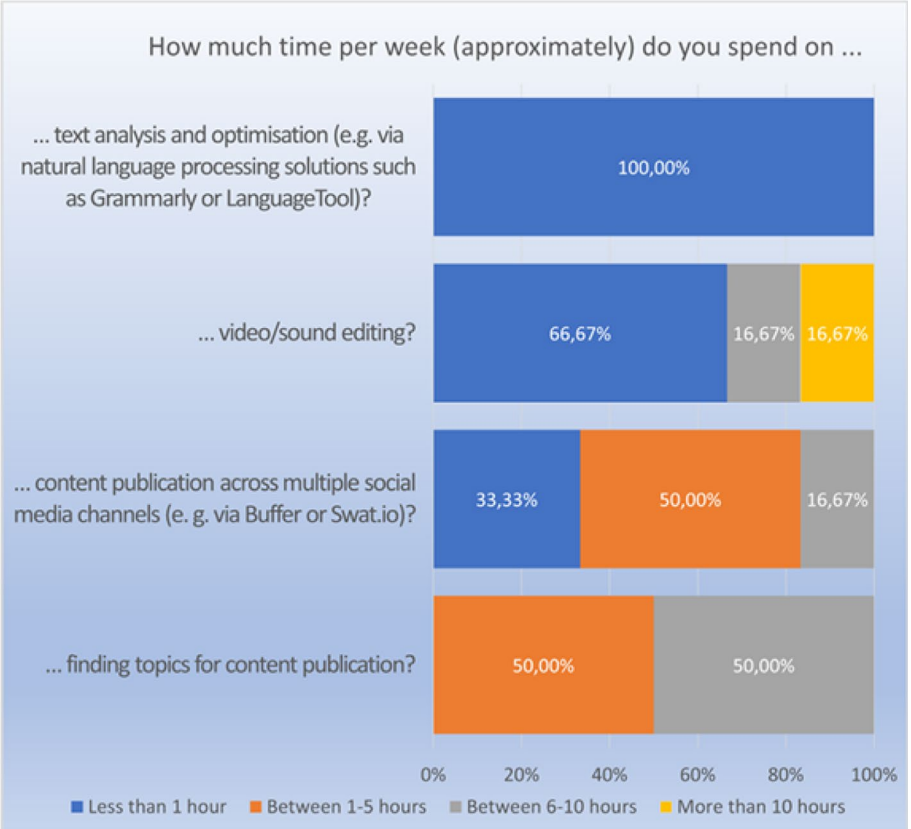

Fig. 11 The time currently spent per week by the participants on different tasks in content creation workflows

Finally, the user can publish to multiple social networks connected to the Levuro publication module, which currently supports Twitter, Instagram, YouTube, Facebook, and LinkedIn. Accompanying the video, a text can be drafted. When publishing to a social media channel, the Content Wizard allows the user to schedule the post (see Fig. 10). The Content Wizard now calls the *Recommendation and Scheduling* component to suggest the ideal date of publication. Specifically, the prediction model is run against the text of the posting (using keyword extraction to generate a list of keywords from the text, which are treated as a "topic" for publication) and the future date is selected where the predicted frequency for mentions of that topic is highest (within the given publication date range, e.g. from 1 to 14 days in the future).

This combination of steps aided by the user-friendly interface is intended to ensure that users of the Content Wizard can still retain control over the digital video content marketing decisions while supported by the automated functionality to find relevant topics, select and re-purpose appropriate video content in their collections, and publish it on digital channels at the optimal moment, enabled by the AI and data-driven media analysis implemented in the ReTV project.

4 Technical and user evaluation

Despite the Content Wizard making use of various Web services (accessed via REST APIs) which employ neural networks and embeddings to provide advanced functionalities (prediction, multi-modal video retrieval, video summarisation), the tool is very responsive and can be used in a production setting. This is largely thanks to the pre-calculation of predictive analytics and the video embeddings for each user, based on the topics they pre-identify as of interest for prediction as well as the provided media collections. The video summarisation does generate the summary 'on the fly', making use of a prior feature analysis of the video which is available via an Elastic index (the video's fragmentation and concept detection for each fragment).

As part of a user evaluation, the Content Wizard was evaluated using a longitudinal testing methodology with six media professionals from three organisations during March/April 2021. The goal was to assess the usability of the Content Wizard workflow and its individual features in operational contexts. Before the four-week testing period started, the tool was configured with data sources and media content to match the needs of the evaluation participants. In particular, relevant video collections were ingested and pre-analysed, a template was defined

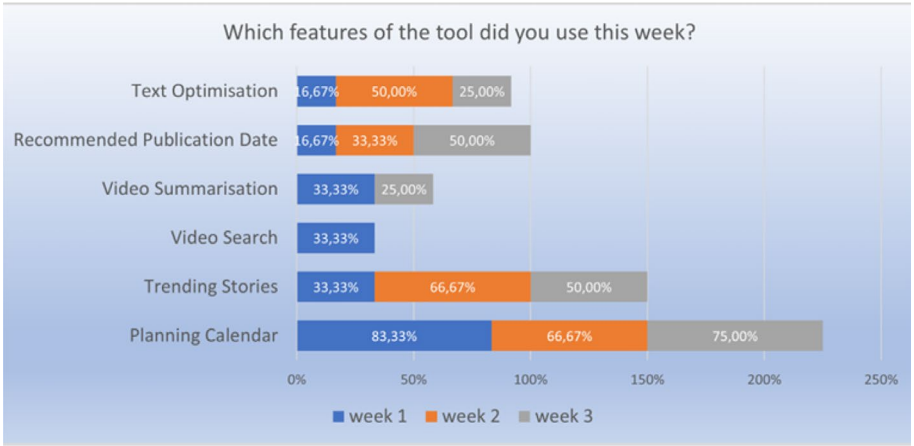

Fig. 12 The features the participants used during the first three weeks

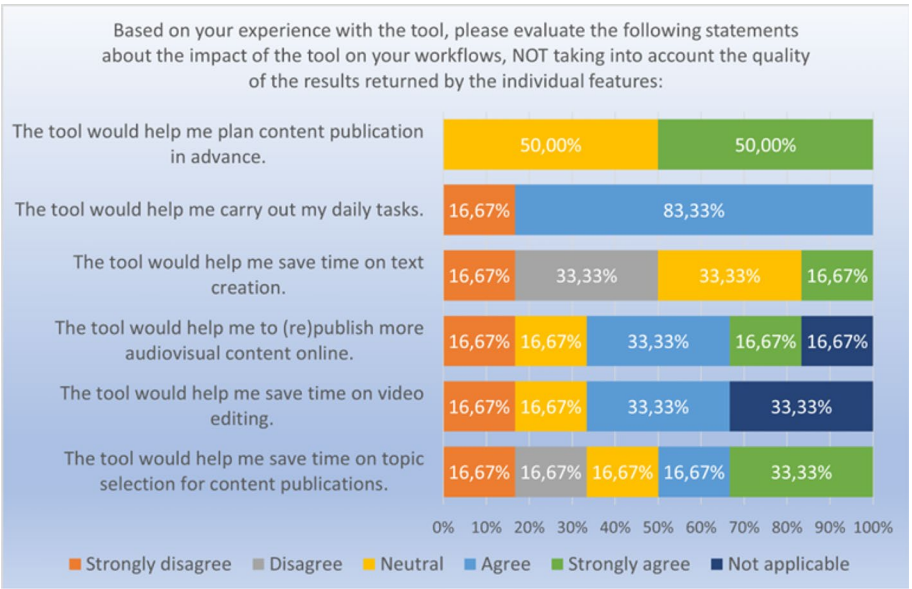

Fig. 13 Impact of the Content Wizard on the participants workflows

to show specific events in a calendar in the Planning tab of the Content Wizard UI (e.g. one user was interested in holidays and commemoration days across the European countries), and appropriate topics were defined and set up in consultation with the testers based on the topics or campaign that were interested in monitoring (e.g. one user was monitoring stories around Women's History Month and Black History Month topics). Testers were asked to use the tool at least twice a week and provided qualitative and quantitative feedback via surveys and direct communication with the project partners.

Before the testing started, all participants indicated that topic selection is the most time-consuming step in their workflows, with half of them spending between six to ten hours each week on it (Fig. 11). During the testing, the calendar in the Planning tab and the Trending Topics feature that showed the predictions for the future mentions of their topics were used the most and evaluated as the most satisfactory (Fig. 12). The calendar was rated positively for providing ideas at a glance. In particular, it helped participants to use popular hashtags on social media (such as #onthisday) to promote historic audiovisual content. Trending Topics was seen as the most novel feature. Although some participants were already using other trend monitoring tools, none of these tools offered comparable capabilities for the prediction of topic popularity in the future.

As for video summarisation, participants stated that they would not consider the automatically generated summary as a final product ready for publication; they would like to perform additional editing to it. The tool already supports some basic video editing features such as adding overlays and the ability to change the order of segments within a summary; the latter was rated very positively as users do not feel the need to stick with chronological ordering of

shots when they do manual video creation. A number of users noted that they would like to define the parameters for video summaries themselves. The current interface could be expanded to accommodate this and also users who need granular customisation could use the video summarisation as a standalone service. Users mentioned that the media consumption habits on social media are changing and that video with sound are becoming more important, hence combining the current video summarisation approach with audio analysis would be of high interest.

The vast majority (83%) of participants agreed that the tool would help them with their daily tasks and half of them strongly agreed that the tool can help them plan content releases in advance (Fig. 13). The opinion of participants was most divided on the potential of the text optimisation feature, with many already using other software in their workflows (e.g. Grammarly). We can note that this was in any case not the key focus of the Content Wizard development, where it is the finding and editing of optimal audiovisual material that is the more time consuming activity with current tools. Overall, we can remark that they confirmed that the tool would have a positive impact on the reuse and distribution of audiovisual content online. In particular, Content Wizard could provide significant support to organizations that do not have access to professional-grade video creation tools or do not have large teams with specialised expertise in media production and distribution. Equally, the feedback gathered from larger teams shows the potential for using individual features of the tool or possible integration with already existing solutions on the market.

Details of the complete questionnaire used in the survey to obtain the above results as well as the detailed responses to the individual questions are presented in the ReTV project deliverable D5.3 "Second Validation of Engagement Monitoring Prototype", available at.[18]

5 Outlook and conclusion

This paper presented the Content Wizard tool for trans-vector publishing, which is itself enabled by an innovative set of Web-based components developed in the ReTV project which we term the Trans-Vector Platform. As this platform is made up of loosely coupled components connected via REST-based Web services, applications can be provided that make use of different combinations of functionalities. We have focused here on the specific innovations developed for trans-vector publishing used by the Content Wizard: components for trending topic prediction, cross-modal video retrieval as well as social media-focused video summarisation. We believe this gives the Content Wizard tool a USP (Unique Selling Proposition) for media organizations which need to adapt their digital video content marketing to the new age of 'interactive media' where content is increasingly transient, malleable and ubiquitous.

User evaluations were positive about access to these innovative functionalities through the user interface of the Content Wizard and their potential impact on future video publication workflows. With minimal editing, users were able to prepare content for publication within minutes, significantly cutting down the effort of performing the same actions manually. Selecting an appropriate topic for publication is a regular challenge for social media managers, which is aided strongly by the trending topic predictions. Once a topic was selected, the text-to-video search eased the task of identifying relevant videos within their collections. The video summarisation functions helped professional users customise their content for publication across multiple digital media vectors while allowing them to maintain creative control. Overall, the feedback suggested that the data and AI-driven media analysis capabilities developed in ReTV can make their current workflows much more efficient.

We believe the media industry will recognise the importance and value of trans-vector publishing in a post-pandemic world where media consumption patterns, among many other aspects of life, have been permanently disrupted and media organisations will need to adapt to this new era. In the US, non-linear streaming content now represents the majority of media consumed, overtaking for the first time classical broadcast and cable TV.[19] As viewing time on traditional media distribution (linear broadcast, cable or satellite) drops, media owners will need to ensure non-linear access to their content (via Web platforms or social networks), which in turn requires targeted promotion to interested audiences on those same channels. To facilitate this, the "Content Wizard" functionalities as presented in this paper have been integrated as an extension to the social media management tool Levuro Engage[20] and are also available as standalone Web services and embeddable Web interfaces.

Author contributions LN provided the manuscript text for Chapters 1 and 2 as well as Chapter 3.2 and 5. Basil Philipp contributed the initial text for Chapter 3.1 and 3.5. The CERTH-ITI authors provided the text for Chapters 3.3 and 3.4. Rasa Bocyte contributed the text for Chapter 4. The entire paper was coordinated by LN, with all authors reviewing the completed manuscript and contributing suggestions and revisions.

Funding Open access funding provided by MODUL University Vienna GmbH. This work was supported by the EU Horizon 2020 research and innovation programme under grant agreement H2020–780656 ReTV.

[18] https://retv-project.eu/deliverables/.

[19] https://www.cnbc.com/2023/08/15/traditional-tv-usage-drops-below-50percent-for-first-time-ever.html.

[20] https://levuro.com/.

Data availability The video assets used by this work were provided by the Netherlands Institute of Sound and Vision from the open-access Open Images repository (https://www.openbeelden.nl/). The Web and social media metrics used for topic prediction are provided by the webLyzard platform (https://weblyzard.com), access on request for non-commercial research purposes. The extracted time series data used by the author for the predictive analytics are available on request. A public Web interface to automatically summarise videos is available at http://multimedia2.iti.gr/videosummarization/service/start.html.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Danaher, P.J., Dagger, T.S., Smith, M.S.: Forecasting television ratings. Int. J. Forecast. **27**(4), 1215–1240 (2011)
2. Weber, R.: Methods to forecast television viewing patterns for target audiences. Communication Research in Europe and Abroad Challenges of the First Decade. Berlin: DeGruyter (2002)
3. Meyer, D., Hyndman, R.J.: The accuracy of television network rating forecasts: the effects of data aggregation and alternative models. Model. Assist. Stat. Appl. **1**(3), 147–155 (2006)
4. Goodman, C., Donthu, N.: Using consumer-generated social media posts to improve forecasts of television premiere viewership: extending diffusion of innovation theory. Available at SSRN 4321891 (2023)
5. Wang, Y.: How do television networks use twitter? exploring the relationship between twitter use and television ratings. South Commun. J. **81**(3), 125–135 (2016)
6. Hsieh, W.-T., Chou, S.-C.T., Cheng, Y.-H., Wu, C.-M.: Predicting tv audience rating with social media. In: Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP), pp. 1–5 (2013)
7. Crisci, A., Grasso, V., Nesi, P., Pantaleo, G., Paoli, I., Zaza, I.: Predicting tv programme audience by using Twitter based metrics. Multimed. Tools Appl. **77**, 12203–12232 (2018)
8. Troncy, R., Laaksonen, J., Tavakoli, H.R., Nixon, L., Mezaris, V., Hosseini, M.: AI4TV 2020: 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4756–4757 (2020)
9. Markatopoulou, F., Mezaris, V., Patras, I.: Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation. IEEE Trans. Circuits Syst. Video Technol. **29**(6), 1631–1644 (2019)
10. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. (CSUR) **54**(10s), 1–41 (2022)
11. Gkalelis, N., Daskalakis, D., Mezaris, V.: ViGAT: bottom-up event recognition and explanation in video using factorized graph attention network. IEEE Access **10**, 108797–108816 (2022)
12. Zhao, B., Li, X., Lu, X.: Property-constrained dual learning for video summarization. IEEE Trans. Neural Netw. Learn. Syst. **31**(10), 3989–4000 (2019)
13. Chu, W.-T., Liu, Y.-H.: Spatiotemporal modeling and label distribution learning for video summarization. In: 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6 (2019). IEEE
14. Rochan, M., Wang, Y.: Video summarization by learning from unpaired data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7902–7911 (2019)
15. Jung, Y., Cho, D., Woo, S., Kweon, I.S.: Global-and-local relative position embedding for unsupervised video summarization. In: European Conference on Computer Vision, pp. 167–183 (2020). Springer
16. Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: AC-SUM-GAN: connecting actor-critic and generative adversarial networks for unsupervised video summarization. IEEE Trans. Circuits Syst. Video Technol. **31**(8), 3278–3292 (2021)
17. Li, H., Ke, Q., Gong, M., Drummond, T.: Progressive video summarization via multimodal self-supervised learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5584–5593 (2023)
18. Habibian, A., Mensink, T., Snoek, C.G.: Video2vec embeddings recognize events when examples are scarce. IEEE Trans. Pattern Anal. Mach. Intell. **39**(10), 2089–2103 (2017). https://doi.org/10.1109/TPAMI.2016.2627563
19. Francis, D., Anh Nguyen, P., Huet, B., Ngo, C.-W.: Fusion of multimodal embeddings for ad-hoc video search. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 1868–1872 (2019)
20. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. ICMR '18, pp. 19–27 (2018). ACM
21. Li, X., Zhou, F., Xu, C., Ji, J., Yang, G.: SEA: sentence encoder assembly for video retrieval by textual queries. IEEE Trans. Multimed. **23**, 4351–4362 (2021)
22. Yang, X., Wang, S., Dong, J., Dong, J., Wang, M., Chua, T.-S.: Video moment retrieval with cross-modal neural architecture search. IEEE Trans. Image Process. **31**, 1204–1216 (2022)
23. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. IEEE Trans. Pattern Anal. Mach. Intell. **44**(8), 4065–4080 (2022)
24. Zhang, X.: Research on design of news video retrieval system based on semantics. In: Proceedings of the 6th International Conference on Virtual and Augmented Reality Simulations, pp. 71–75 (2022)
25. Zwicklbauer, M., Lamm, W., Gordon, M., Apostolidis, K., Philipp, B., Mezaris, V.: Video Analysis for Interactive Story Creation: The sandmännchen showcase. In: Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery, at ACM Multimedia 2020, pp. 17–24 (2020)
26. Glasp: YouTube Summary YouTube with ChatGPT & Claude. https://glasp.co/youtube-summary. Accessed: 2023-08-22 (2023)
27. Collyda, C., Apostolidis, K., Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V.: A web service for video summarization. In: ACM International Conference on Interactive Media Experiences, pp. 148–153 (2020)

28. Cushing, A.L., Osti, G.: "So how do we balance all of these needs?'': how the concept of AI technology impacts digital archival expertise. J. Doc. **79**(7), 12–29 (2022)
29. Bocyte, R., Oomen, J.: Content adaptation, personalisation and fine-grained retrieval: applying AI to support engagement with and reuse of archival content at scale. In: ICAART (1), pp. 506–511 (2020)
30. Jin, J.-G., Bae, J., Baek, H.-g., Park, S.-h.: Object-ratio-preserving video retargeting framework based on segmentation and inpainting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 497–503 (2023)
31. Apostolidis, K., Mezaris, V.: A fast smart-cropping method and dataset for video retargeting. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 2618–2622 (2021). IEEE
32. Casado, M.À., Guimerà, J.À., Bonet, M., Llavador, J.P.: Adapt or die? how traditional spanish tv broadcasters deal with the youth target in the new audio-visual ecosystem. Critical Studies in Television, 17496020221076983 (2022)
33. Philipp, B., Ciesielski, K., Nixon, L.: Automatically adapting and publishing tv content for increased effectiveness and efficiency. In: Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery, pp. 51–52 (2019)
34. Nixon, L., Foss, J., Apostolidis, K., Mezaris, V.: Data-driven personalisation of television content: a survey. Multimed. Syst. **28**(6), 2193–2225 (2022)
35. Galanopoulos, D., Mezaris, V.: Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In: Proceedings of the 2020 ACM International Conference on Multimedia Retrieval, pp. 336–340 (2020)
36. Pantelidis, N., Andreadis, S., Pegia, M., Moumtzidou, A., Galanopoulos, D., Apostolidis, K., Touska, D., Gkountakos, K., Gialampoukidis, I., Vrochidis, S., Mezaris, V., Kompatsiaris, I.: VERGE in vbs 2023. In: Dang-Nguyen, D.-T., Gurrin, C., Larson, M., Smeaton, A.F., Rudinac, S., Dao, M.-S., Trattner, C., Chen, P. (eds.) MultiMedia Modeling, pp. 658–664. Springer, Cham (2023)
37. Galanopoulos, D., Mezaris, V.: Cross-modal networks and dual softmax operation for MediaEval NewsImages 2022. In: 2022 Multimedia Evaluation Workshop (MediaEval'22), Bergen, Norway (2022)
38. Gkountakos, K., Galanopoulos, D., Touska, D., Ioannidis, K., Vrochidis, S., Mezaris, V., Kompatsiaris, I.: ITI-CERTH participation in ActEV and AVS tracks of TRECVID 2022. In: TRECVID 2022 Workshop, Gaithersburg, MD, USA (2022)
39. Nixon, L.J.B.: Predicting your future audience: Experiments in picking the best topic for future content. In: ACM International Conference on Interactive Media Experiences. IMX '20, pp. 185–188. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3391614.3399398
40. Nixon, L.: Predicting your future audience's popular topics to optimize tv content marketing success. In: Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery. AI4TV '20, pp. 5–10. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3422839.3423062
41. Laptev, N., Yosinski, J., Li, L.E., Smyl, S.: Time-series extreme event forecasting with neural networks at uber. In: International Conference on Machine Learning, vol. 34, pp. 1–5 (2017)
42. Bykov, N., Skorohodov, A., Denisenko, E.: Predictive analytics in tv marketing for the mass segment. In: 2023 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), pp. 214–216 (2023). https://doi.org/10.1109/USBEREIT58508.2023.10158901
43. Du, S., Li, T., Yang, Y., Horng, S.-J.: Multivariate time series forecasting via attention-based encoder-decoder framework. Neurocomputing **388**, 269–279 (2020)
44. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421. Association for Computational Linguistics, Lisbon, Portugal (2015). https://doi.org/10.18653/v1/D15-1166
45. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions (2018). https://openreview.net/forum?id=SkBYYyZRZ
46. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9346–9355 (2019)
47. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
48. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288–5296 (2016)
49. Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., Luo, J.: Tgif: A new dataset and benchmark on animated gif description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4641–4650 (2016)
50. Caba Heilbron, F., *et al.*: ActivityNet: A large-scale video benchmark for human activity understanding. In: Proc. of IEEE CVPR 2015, pp. 961–970 (2015)
51. Wang, X., *et al.*: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proc. of IEEE/CVF ICCV 2019, pp. 4581–4591 (2019)
52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR
53. Awad, G., Butt, A., Fiscus, J., Joy, D., Delgado, A., *et al.*: TRECVID 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In: TRECVID 2017 Workshop. Gaithersburg, MD, USA (2017)
54. Awad, G., Butt, A.A., Curtis, K., Fiscus, J., Godil, A., Lee, Y., Delgado, A., Zhang, J., Godard, E., Chocot, B., Diduch, L., Liu, J., Graham, Y., Jones, G.J.F., , Quénot, G.: Evaluating multiple video understanding and retrieval tasks at trecvid 2021. In: Proceedings of TRECVID 2021 (2021). NIST, USA
55. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2vv++: Fully deep learning for ad-hoc video search. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1786–1794 (2019). ACM
56. Wu, J., Ngo, C.-W.: Interpretable embedding for ad-hoc video search. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3357–3366. ACM, New York, NY, USA (2020)
57. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
58. Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Video summarization using deep neural networks: a survey. Proc. IEEE **109**(11), 1838–1863 (2021). https://doi.org/10.1109/JPROC.2021.3117472
59. Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Unsupervised video summarization via attention-driven

adversarial learning. In: International Conference on Multimedia Modeling, pp. 492–504 (2020). Springer
60. Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., Shao, L.: Exploring global diverse attention via pairwise temporal relation for video summarization. Pattern Recogn. **111**, 107677 (2021)
61. Jung, Y., Cho, D., Kim, D., Woo, S., Kweon, I.S.: Discriminative feature learning for unsupervised video summarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8537–8544 (2019)
62. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee
63. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
64. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR
65. Over, P.: TRECVID 2013–an overview of the goals, tasks, data, evaluation mechanisms and metrics (2013)