Enhancing and Re-Purposing TV Content for Trans-Vector Engagement (ReTV)
H2020 Research and Innovation Action - Grant Agreement No. 780656

Enhancing and Re-Purposing TV Content for Trans-Vector Engagement

Deliverable 2.3 (M30)

Metrics-Based Success Factors and Predictive Analytics

Version 2.0

This document was produced in the context of the ReTV project supported by the European Commission under the H2020-ICT-2016-2017 Information & Communication Technologies Call Grant Agreement No 780656

DOCUMENT INFORMATION

Delivery Type	Report
Deliverable Number	2.3
Deliverable Title	Metrics-Based Success Factors & Predictive Analytics v2
Due Date	M30
Submission Date	June 30, 2020
Work Package	WP2
Partners	MODUL Technology, webLyzard technology, Genistat
Author(s)	Lyndon Nixon, Jakob Steixner, Adrian Brasoveanu (MODUL Technology), Arno Scharl, Max Goebel, Katinka Boehm (webLyzard), Basil Philipp, Krzysztof Ciesielski (Genistat)
Reviewer(s)	Martin Gordon (RBB)
Keywords	Event Extraction, Event Knowledge Base, Success Metrics, Audience Metrics, Predictive Analytics, Content Success Prediction
Dissemination Level	PU
Project Coordinator	MODUL Technology GmbH Am Kahlenberg 1, 1190 Vienna, Austria
Contact Details	Coordinator: Dr Lyndon Nixon (nixon@modultech.eu) R&D Manager: Prof Dr Arno Scharl (scharl@weblyzard.com) Innovation Manager: Bea Knecht (bea@zattoo.com)

Revisions

Version	Date	Author	Changes
0.1	30/4/20	L. Nixon	Created template and ToC
0.11	5/5/20	A. Scharl	Minor revisions and edits
0.2	19/5/20	L. Nixon	Filled in initial outlines
0.3	3/6/20	B. Philipp	Genistat content added
0.4	11/6/20	L. Nixon	MODUL content added
0.6	15/6/20	L. Nixon	Completion of prediction section
0.61	17/6/20	L. Nixon	Included tables and figures, conclusion
0.7	18/6/20	A. Brasoveanu	Revision of prediction section
0.8	18/6/20	A. Scharl	Content-based success metrics
0.9	24/6/20	L. Nixon, A. Scharl	Revision, success metrics and prediction
1.0	26/6/20	L. Nixon	Post-QA edits and final check

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

This deliverable reflects only the authors' views and the European Union is not liable for any use that might be made of information contained therein.

Contents

1 Introduction **8**

2 Event Extraction and Modelling **9**

2.1 Status . . . 9

2.2 Updates from Previous Report . . . 9

2.2.1 Event KB: Adding Awareness and World Days . . . 9

2.2.2 Event and Anniversaries API . . . 10

3 Success Metrics **12**

3.1 Audience-Based Metrics . . . 12

3.2 Content-Based Metrics . . . 12

3.2.1 Cross-Channel Impact Metrics . . . 12

3.2.2 Multidimensional Models of Affect . . . 12

3.2.3 WYSDOM Success Metric . . . 14

4 Predictive Analytics **16**

4.1 Audience Prediction . . . 16

4.2 Content Success Prediction . . . 16

4.2.1 Temporal References . . . 17

4.2.2 Time Series Data . . . 18

4.2.3 Hybrid Prediction Model . . . 19

4.3 Evaluation . . . 19

4.3.1 Temporal Reference-Based Prediction . . . 19

4.3.2 Time Series-Based Prediction . . . 21

4.4 Considerations for a Hybrid Solution . . . 25

5 Conclusion and Outlook **26**

5.1 Event Extraction and Modelling . . . 26

5.2 Success Metrics . . . 26

5.3 Predictive Analytics . . . 26

EXECUTIVE SUMMARY

This deliverable is an update of D2.2. It covers the topics of event collection and modelling, audience and success metrics, and predictive analytics for the success of future online content publication. Specifically, it updates on our collection and modelling of events, audience metrics and success metrics for the purpose of using them as inputs to predictive analytical models. It presents the results of using these inputs both individually and in combination to enable content-based success predictions for content owners. We conclude with a plan for a hybrid prediction model that can achieve optimal content-based success prediction for our scenarios.

ABBREVIATIONS LIST

Abbreviation	Description
API	Application Programming Interface: a set of functions and procedures that allow the creation of applications which access the features or data of an application or other service.
DCNN	Deep Convolutional Neural Network: a type of artificial neural network.
EPG	Electronic Program Guides: menu-based systems that provide users of television with continuously updated menus displaying broadcast programming or scheduling information for current and upcoming programming.
HTTP POST/GET	Types of method in the Hypertext Transfer Protocol (HTTP). The HTTP POST method is used to send data to a server to create/update a resource. The HTTP GET method is used to request data from a specified resource.
IPTV	Internet Protocol Television: is the delivery of television content over Internet Protocol (IP) networks.
JSON	JavaScript Object Notation: a data-interchange format.
LSTM	Long Short Term Memory networks: a type of recurrent neural network.
MTL	Multi-task learning: a field of machine learning in which multiple learning tasks are solved at the same time, exploiting commonalities and differences across tasks.
NEL	Named Entity Linking
NER	Named Entity Recognition
NLP	Natural Language Processing: subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human (natural) languages.
OTT	Over The Top: content providers that distribute streaming media as a standalone product directly to viewers over the Internet, bypassing telecommunications that traditionally act as a distributor of such content.
RDF	Resource Description Framework: a method for conceptual description or modeling of information that is implemented in web resources.
REST	Representational State Transfer: an architectural style that defines a set of constraints to be used for creating web services.
RNN	Recurrent Neural Network: a type of an artificial neural network.
SKB	Semantic Knowledge Base: a RDF-based triple store for a knowledge representation of keywords and entities during in annotation of documents
TVoD	Transactional Video on Demand: a distribution method by which customers pay for each individual piece of video on demand content.
URL	Uniform Resource Locator: a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it.

1. Introduction

This deliverable reports on the work done from month M21 to M30 in the ReTV project. In this reporting period the following tasks were active:

Firstly, the three individual tasks on the collection and modelling of events, audience metrics and success metrics as inputs for predictive analytical models officially finished with the past deliverable D2.2. However, as reflected in this deliverable, further improvements and extensions still occurred in this period as a result of supporting the evolving requirements of the ReTV use cases. Therefore we still present an overview of those developments with respect to event extraction and modelling (Section 2) as well as audience and content-based success metric collection (Section 3).

Secondly, the main focus on this period has been the preparation of predictive analytical approaches with the above data as input options. While the audience metrics forecasting can indicate patterns in viewership of TV shows and channels, the content-based success metrics can indicate trends in the popularity of topics among the audience on different digital channels. Since both of these metrics are time-based, forecasting is based on patterns in the past data and cannot alone take into account out-of-trend developments in the future. Therefore with both we bring in future event knowledge combined with training with past events as additional features, in order to include the variations caused by events in our future predictions. Section 4 presents the different prediction approaches which were implemented and evaluated by ReTV with the intention of combining them into a best-result hybrid model.

2. Event Extraction and Modelling

2.1. Status

Our Semantic Knowledge Base (SKB) which contains the entities of type Event alongside other entity types such as Person, Organisation or Location has been maintained throughout the period. It currently has 67 296 event entities, which includes 52 511 public holidays and awareness days (counting each recurrence up to the year 2099), 14 394 iCal events and 1 021 WikiData events.

We make a daily query against the WikiData API for all new events (matching the types we previously defined) which start in the next 180 day period. We also have a running process to check for updates in the descriptions of events we have already added to the SKB (using the Recent Changes API [1] in order to update the local metadata.

2.2. Updates from Previous Report

One change we made when extracting events for the SKB due to the observed lack of informative descriptions in WikiData events was to crossreference the entities with the associated Wikipedia entry and extract automatically the first paragraph of Wikipedia article text as a description. This is done when the description property in WikiData is empty.

2.2.1. Event KB: Adding Awareness and World Days

In particular, we ran an internal evaluation with our use case partners (RBB and NISV) regarding the events they would expect to see for a selection of dates and identified a significant gap in 'international days' (meaning official global days for recognized causes such as those defined by the United Nations [2]) and other 'awareness days' whether for Muffins, Birds, Coffee or anything else (The International Day of ... or the World Day of ...). These types of annual days were regularly used by the use case partners in their content publication planning so we prepared a list of relevant events (we used the UN page footnoted above as well as the National Day Calendar [3] for other awareness days). The international days were extracted from WikiData and compared with the online list to identify gaps; as a result we added manually eleven international days to WikiData as well [4]. We also checked and corrected inconsistencies in some of the WikiData metadata (e.g. we found sometimes that dates were missing or incorrect) and especially took care to add meaningful descriptions in many cases. Then we queried the WikiData API again following all of these edits and added the curated entities to the SKB.

In modelling the new events for addition to the SKB, we saw that the standard event classes could not capture topical differences in the subject of the events. These events are all typed (dct:type) as Awareness Days (as opposed to the other types already presented in D2.2: Public Holidays, wd:Q16466010 (association football matches added via iCals), or the respective class from WikiData). However, we felt that there were topical subjects of events that might cross event classes - e.g. sports, which can be a subject of differently typed WikiData events and equally a subject of an awareness day. So we added in our own list an additional column for event category, which is a freely chosen label although we focus on a set of known terms

[1] https://www.mediawiki.org/wiki/API:RecentChanges

[2] https://www.un.org/en/sections/observances/international-days/

[3] https://nationaldaycalendar.com/

[4] e.g. https://www.wikidata.org/wiki/Q85521697 International Day for Disaster Risk Reduction

(thus creating a form of controlled vocabulary) so that where possible an existing category is re-used rather than creating new terms unnecessarily. Based on the lists of awareness days (we had in the end 281 international days and 79 other awareness days) we labeled them with the following categories: politics, culture, news, entertainment, health, society, environment, weather, food, sport. Where possible, we added the categories also to the other events in the SKB, e.g. every event of type wd:Q16466010 (association football match) is categorized (with a new property skbprop:eventCategory) as 'sport'.

2.2.2. Event and Anniversaries API

We also continued development of the API for requesting events from the SKB and implemented a new, event-related API called Anniversaries API. This API takes a date (month-day pair) without a year and returns entities of type Person or Organization which have an anniversary on that date (for Persons, we use the given dates in birthdate and deathdate, for Organizations we use the given date in inception (founding)). As the SKB grew, we made the queries executed through the API calls more efficient e.g. we found that countries were represented differently between events as either strings or as URIs so we needed to handle both cases. We updated code to that we could make value matching possible across all property values, necessary for the scenario filter described next.

Both APIs are extended to support an additional parameter, which indicates for which scenario the API is being called. We could observe for a query on all matching events or anniversaries in the SKB could return a number of results in up to three digits (hundreds), thanks to the size of our event collection. However, for scenarios the user needs to see the events on that date that are most relevant to them, making some additional filter on event results necessary. To do this, we implement within the code for the SPARQL query creation a template which can be specified in different forms (i.e. on which property a filter should be applied, what value filter is used depending on the value type such as text, number or URI) for different scenarios.

Based on the internal evaluations with the use case partners, we identified the typical characteristics of most of the events they specified as relevant in their content publication planning processes apart from the aforementioned international and awareness days. The first two scenario templates that we implement therefore will be generically 'RBB' and 'NISV'; in the use case development, we will discuss with new scenario stakeholders their events of interest to provide additional templates on the events and anniversaries API as necessary. As an example, the filters for the first two scenarios are as follows:

RBB

- Person anniversaries: any German person.
- Person anniversaries: famous entertainers, actors, painters, etc.
- Organization anniversaries: inception of any with country = Germany
- Events: UN awareness days.
- Events: National Day calendar.
- Events: holidays in Germany.
- Events: Berlin-specific.
- Locations: in Berlin or Brandenburg with a date.

NISV

- Person anniversaries: any Dutch person.
- Person anniversaries: famous authors, painters, directors, etc.
- Organization anniversaries: inception of any with country = Netherlands.
- Events: UN awareness days.
- Events: holidays in Netherlands.
- Events: Amsterdam-specific.
- Locations: in Amsterdam or rest of the Netherlands with a date.

3. Success Metrics

3.1. Audience-Based Metrics

Details of the audience metrics provided via Genistat to the metadata repository were given in deliverables D2.1 and D2.2. The last extension we made was to link the Electronic Program Guide (EPG) data to the audience data to allow for easy aggregation of the audience numbers per show.

The push of data has shown to be reasonably robust in daily use, and the choice of using a functional programming paradigm has shown its advantages when faulty data had to be resent, as this could be done without side effects.

As of project M30 (June 2020), we expect to switch to a new source for real-time data from Zattoo. The real-time audience data used so far was calculated from the internal monitoring system that Zattoo uses. This system is being deprecated for technical reasons, which makes the change to a new data source necessary. We plan to use a read-only access to anonymized Google Analytics properties of Zattoo. This new access will completely replace the old access, with some advantages:

1. Since we access the data through an official API, instead of using an internal monitoring system, we expect the robustness of the system to increase.
2. The Google Analytics data allows us to split the data by additional dimensions, like the device used. We will decide at a later stage if those additional dimensions add value to the ReTV data analytics and visualisations.

3.2. Content-Based Metrics

The set of success metrics reported in D2.2, Section 3 were continually calculated throughout the period on the collected data sources and made available via the TVP Visual Dashboard: Frequency, Sentiment and WYSDOM. A new metric of Impact was added as a result of the work done on considering a normalization of content reach across digital channels. Sentiment has been extended to a more multidimensional model of Emotions.

3.2.1. Cross-Channel Impact Metrics

For each source, a normalized reach value is calculated in a [0,1] interval - based on the average site traffic for Web sites, for example, or a derived indicator based on the number of followers / following in the case of Twitter. Various dashboard components have been updated to reflect the new metric - including the source table, the scatterplot and the trend chart.

The table of sources accessible via the content area yields source-specific keywords that reflect what each source associates with the query term. The list can be sorted by name, frequency of mentions (count), reach, impact (frequency multiplied by reach), or average sentiment. The initial version only supported a local resorting of the Top 50 sources, the latest prototype supports global ranking by each of the attributes listed above.

3.2.2. Multidimensional Models of Affect

There are several affective models used for opinion mining that differ in terms of their goals and complexity. Sentiment analysis classifies content streams into positive and negative expressions, using sentiment lexicons in conjunction with artificial intelligence-based machine learning methods to determine the polarity of sentences and documents. ReTV has adopted more complex affective models, which tend to be based on the work of psychologists. They provide more fine-grained classifications into major emotional categories. Many of these models also define sub-categories for a more nuanced representation that considers the intensity of the expressed emotion. Figure 1 compares the different affective models that can guide the automated classification of human emotions, including Robert Plutchik's *Wheel of Emotions* and Erik Cambria's *Hourglass of Emotions*. We consider that these affective models are more appropriate for the representation of audience emotion towards TV series (or films, etc.) since e.g. fear is generally seen in bipolar sentiment analysis as negative, yet may be a desired emotion among an audience for a horror series.

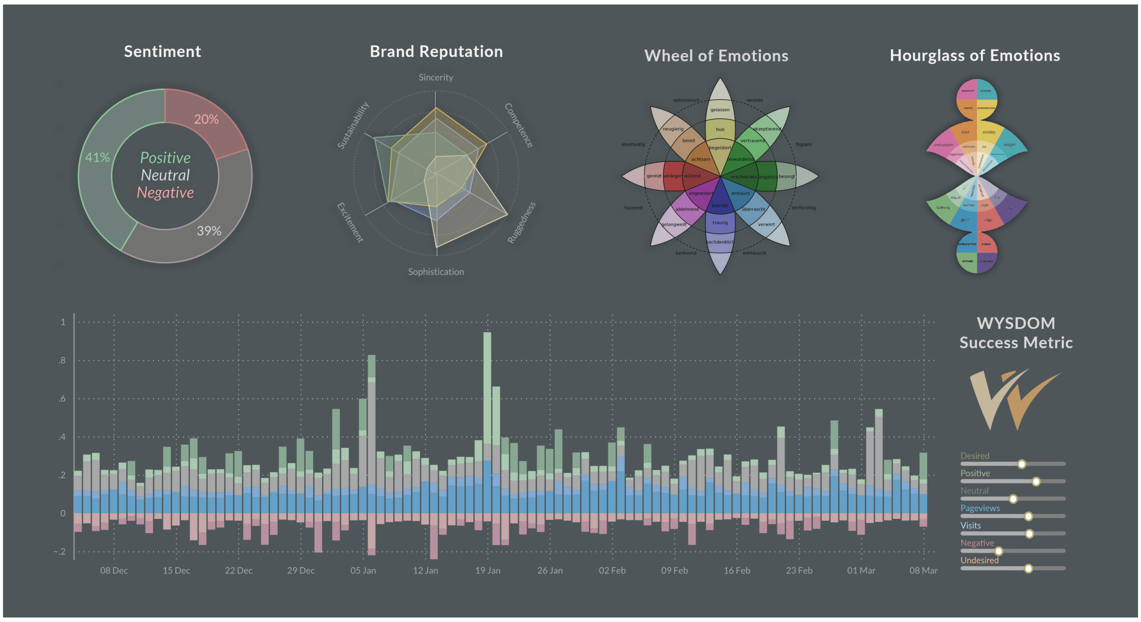
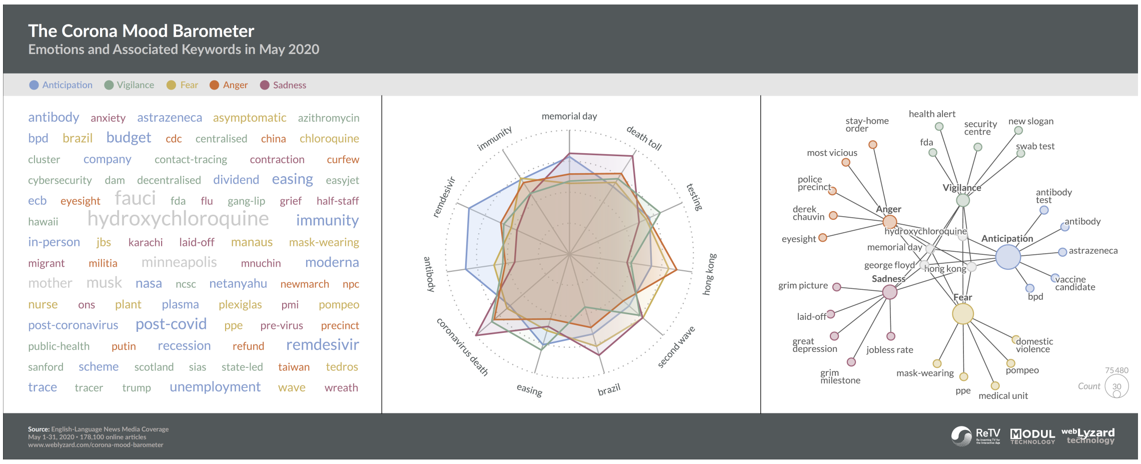


Figure 1: Overview of affective models including *Sentiment, Brand Personality, Wheel of Emotions, Hourglass of Emotions* and the *webLyzard Stakeholder Dialogue and Opinion Model* (WYSDOM)

Plutchik's popular model represents a good starting point, due to its combination of a workable structure (derived from the eight basic emotions: joy, trust, fear, surprise, sadness, disgust, anger and anticipation) with different degrees of expression along the covered affective categories. Similar to our sentiment detection work, the affective model was realised through creation of lexical resources which associate terms in the source languages (English, German and Dutch have been considered in ReTV) with each emotional dimension, each with three degrees of expression (weaker, normal, stronger). This is combined with the existing NLP pipeline as well as the detection of negation which was already implemented for sentiment.

Figure 2 shows a first application of this work as part of the *Corona Mood Barometer*.[5] The system uses a combination of story detection and emotion analysis techniques to better understand what drives the public coronavirus debate and how government responses to the COVID-19 pandemic are perceived across the various countries. The visualizations of Figure 2 explore the associations with five selected emotions (Anticipation, Vigilance, Fear, Anger and Sadness) to better understand the drivers of the public debate in May 2020.

[5] www.weblyzard.com/corona-mood-barometer



Source: English-Language News Media Coverage
May 1-31, 2020 • 178,100 online articles
www.weblyzard.com/corona-mood-barometer

Figure 2: Analysis of COVID-19 coverage based on Plutchik's *Wheel of Emotions*, using a tag cloud (left), a radar chart (middle) and keyword graph (right) with color coding to distinguish selected emotions incl. *Anticipation, Vigilance, Fear, Anger* and *Sadness*.

The tag cloud sorts associations alphabetically, color coding them by emotion. The radar chart projects the top keywords along multiple axes, revealing the relative strength of association with each emotional category. The keyword graph applies a hierarchical layout, with grey center nodes to represent keywords linked to multiple emotional categories. Further work will consider the use of deep learning NLP models such as BERT to train our affective model to associate arbitrary texts with the affective dimensions, going beyond our initial lexica.

3.2.3. WYSDOM Success Metric

Work related to the WYSDOM success metric, already introduced in Deliverable D2.2, included a new data model and mechanism to store WYSDOM settings together with the topic definition, since just one global setting would not suffice to cover all ReTV use cases.

The problem of real-time negation detection was tackled by means of automated prefix processing at runtime, using several standard prefix sets based on the part of speech of the search term. Users can now also specify the minimal number of regular expressions that a document must match to be included in the search results. This can improve the precision of defining desired and undesired concepts at the cost of lower recall, especially if the concept definitions contain terms that are ambiguous without additional context information.

To demonstrate WYSDOM via the Topics Compass to media organizations, we solicited short lists of desired and undesired attributes associated with selected news organizations. Domain experts provided a seed model, condensed into ten undesired and ten desired concepts:

- **Desired:** balanced, captivating, entertaining, informative, innovative, investigative, professional, reliable, transparent, trustworthy.
- **Undesired:** boring, censored, disrespectful, fake, irrelevant, offensive, partisan, self-referential, unbalanced, unprofessional.

Combining WYSDOM with the affective model presented in the previous sub-section considerably improves recall, as illustrated in Table 1. Sentences that would have been classified with a one-sided score (desirability) with the sentiment model contain significantly nuanced WYSDOM scores with the expanded model using the affective concepts.

Table 1: Positive (desired) and negative (undesired) WYSDOM results obtained with the expanded affective model.

title	desirability	affective concepts
"Just to make this perfectly clear, I was laughing at the joke and not at any group of people."	1.0	just: 1.0; make: 1.0; perfectly clear: 1.0; laughing: -0.05; group: -0.05
So while coverage for Democrats overall was a bit more positive than negative, that was almost all due to extremely favorable coverage for Obama.	1.0	coverage: 0.05; overall: 0.2; bit: 0.05; positive:1.0; negative: -1.0; extremely favorable: 1.0
The statement drew criticism to the network for being false.	-1.0	criticism: -0.55; false: -0.7
Jeet Heer, the national affairs correspondent at The Nation said "the big loser of the night was the network that hosted the event."	-0.715	affairs: -0.05; loser: -1; Nation said: -0.05

4. Predictive Analytics

4.1. Audience Prediction

In Section 5.2 of D2.2 we reported on our forecasting model for television audiences. We showed how event-based features improved the predictions, especially when applied to the channels that broadcast sports, since we found this domain in particular had a strong effect on audience numbers.

We have since extended our prediction approach and targeted it to specific use cases. We report on how we use feature embeddings in recommendation and scheduling content within the 4u2 Chatbot and the Content Wizard use-cases in ReTV deliverable D3.3.

4.2. Content Success Prediction

In Section 5.3 of D2.2 we reflected on how to forecast the success of a piece of content published on a digital channel at a future moment in time. We reasoned that predictions for the future value of a content success metric for a keyword or topic should be reflective of the content success (quantified by a metric like reach, engagement or views) when a piece of content related to that keyword or topic is published at the time of the prediction. This led to two cases for prediction that we needed to consider:

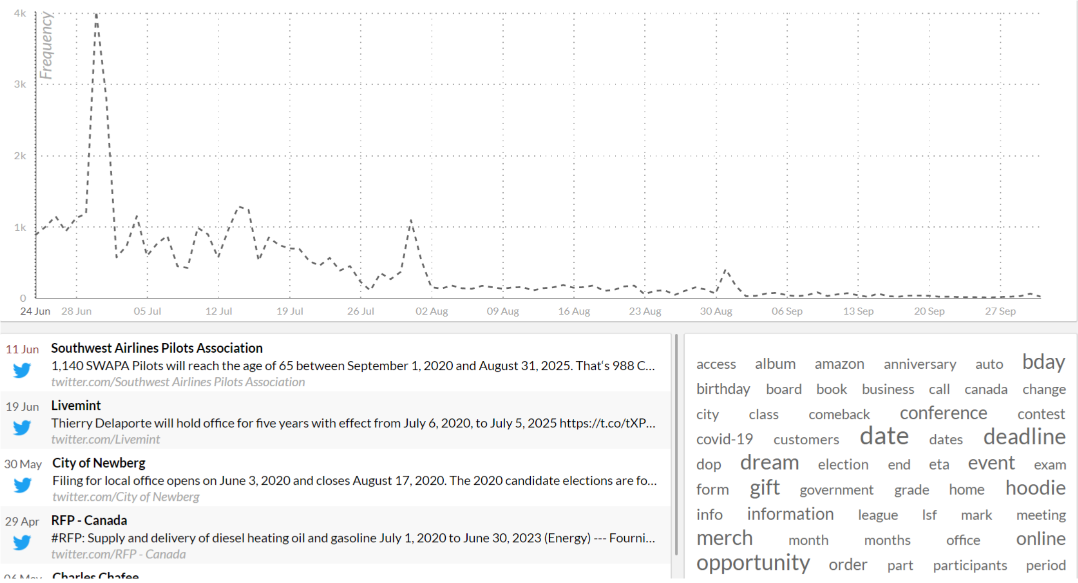
- For a chosen keyword or topic (of a piece of content to be published), determine the optimal time in a time range to publish the content
- For a chosen future time point, determine the optimal keyword or topic to use in reference to a piece of content

We also presented two approaches to prediction that had distinct advantages and disadvantages:

- A temporal reference extraction process from collected (news) documents so that, for any future date, a set of keywords could be aggregated from the documents which reference that date. This allowed a prediction to be made for dates more than a few days into the future but results could be sparse for some arbitrary future date.
- A linear autoregressive model with 4 day moving window and past/future events as one-hot encoded categorical variables. This model could generally [6] get close to future values for the next few days but accuracy dropped out severely when looking more than three or four days into the future due to the autoregression.

In this section, we present the improvements made to both approaches. Then we discuss the potential to combine both into a hybrid model where the time series-based prediction provides a stable baseline for prediction values and the temporal reference-based prediction the basis for a predicted variation in this baseline.

[6] As will be seen again in the evaluation section, each distinct keyword/topic can perform very differently in prediction as so many independent events can have occurred in the past to affect any natural trend in the keyword frequency, and as with any prediction task, we cannot anticipate every future independent event that will occur to affect the natural trend in future keyword frequency. The coronavirus/COVID-19 event alone affected every aspect of online communication this year.



11 Jun **Southwest Airlines Pilots Association**
1,140 SWAPA Pilots will reach the age of 65 between September 1, 2020 and August 31, 2025. That's 988 C...
twitter.com/Southwest Airlines Pilots Association

19 Jun **Livemint**
Thierry Delaporte will hold office for five years with effect from July 6, 2020, to July 5, 2025 https://t.co/tXP...
twitter.com/Livemint

30 May **City of Newberg**
Filing for local office opens on June 3, 2020 and closes August 17, 2020. The 2020 candidate elections are fo...
twitter.com/City of Newberg

29 Apr **RFP - Canada**
#RFP: Supply and delivery of diesel heating oil and gasoline July 1, 2020 to June 30, 2023 (Energy) --- Fourni...
twitter.com/RFP - Canada

06 May **Charles Chafee**

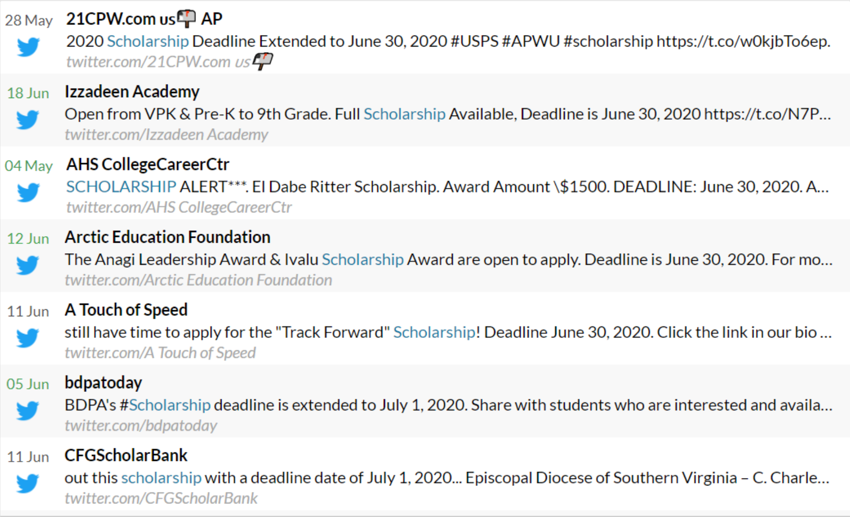
Figure 3: Distribution of frequency of future date references in Twitter

4.2.1. Temporal References

The results of the temporal reference-based prediction were made visible on the ReTV visual analytics dashboard. Through a separate Prediction Mode, future date ranges can be chosen and the number of matching documents displayed in the context of the current search. A match is considered when a document in the corpus processed for temporal reference detection (currently this is the past 12 months of English and German language news articles) contains a date in the given date range. The analytics for the documents selected in the prediction search are also visible, e.g. the sentiment of each document, the frequency of documents per day or according to a moving average, and most significantly for the prediction task the aggregated set of keywords and entities detected in the document corpus.

Following a first indexing of the news articles according to our implementation of temporal reference detection, an evaluation triggered corrections to the first implementation - e.g., we found problems with both absolute and relative references to date ranges. While specific date formats and relative references to specific days showed good accuracy, date ranges had multiple issues: unlike days of the week, month names can be ambiguous in text (e.g. 'May'/'may' in English) whereas counting a document which references a future month in a search for any date in that month introduced too much irrelevance to results. So we decided from our testing to restrict our acceptance of temporal references to those which are specific date references (considering too the differences in US and UK English orders - DD/MM/YYYY or MM/DD/YYYY) as well as a smaller subset of relative references to absolute dates (yesterday, tomorrow, next Monday etc. as opposed to last year, next week or the coming summer).

The news document corpus has been re-indexed according to this corrected implementation, to be used in prediction experiments. We have extended the data available in the prediction mode to include also a Twitter source. Here we monitor the Twitter Streaming API for references to any absolute date in 90 days into the future, collect those tweets, index them according to all temporal date references they contain and this is also now visible through the ReTV visual analytics dashboard. The screenshot (Fig. 3) shows the frequency distribution of date references in the Twitter source from June 24 through October 1, 2020. The peak indicates many



28 May **21CPW.com us📫 AP**
2020 Scholarship Deadline Extended to June 30, 2020 #USPS #APWU #scholarship https://t.co/w0kjbTo6ep.
twitter.com/21CPW.com us📫

18 Jun **Izzadeen Academy**
Open from VPK & Pre-K to 9th Grade. Full Scholarship Available, Deadline is June 30, 2020 https://t.co/N7P...
twitter.com/Izzadeen Academy

04 May **AHS CollegeCareerCtr**
SCHOLARSHIP ALERT***. El Dabe Ritter Scholarship. Award Amount \$1500. DEADLINE: June 30, 2020. A...
twitter.com/AHS CollegeCareerCtr

12 Jun **Arctic Education Foundation**
The Anagi Leadership Award & Ivalu Scholarship Award are open to apply. Deadline is June 30, 2020. For mo...
twitter.com/Arctic Education Foundation

11 Jun **A Touch of Speed**
still have time to apply for the "Track Forward" Scholarship! Deadline June 30, 2020. Click the link in our bio ...
twitter.com/A Touch of Speed

05 Jun **bdpatoday**
BDPA's #Scholarship deadline is extended to July 1, 2020. Share with students who are interested and availa...
twitter.com/bdpatoday

11 Jun **CFGScholarBank**
out this scholarship with a deadline date of July 1, 2020... Episcopal Diocese of Southern Virginia – C. Charle...
twitter.com/CFGScholarBank

Figure 4: Selection of tweets about scholarships from June 30 to July 1, 2020

different deadlines mentioned for June 30/July 1. We find associations to calls, submissions and scholarships around those dates, the screenshot (Fig. 4) shows tweets about scholarships showing the apparent importance of this date period for application deadlines.

4.2.2. Time Series Data

In the time series-based prediction, our main concern was to address the drop-off in accuracy seen in the autoregressive (AR) model. Firstly, we extended it to a full ARIMA model including integration of time-adjacent values (I) and the moving average (MA). This allows us to test different configurations and to include the event occurrences in training, testing and prediction through exogenous variables (X = the model is now known as ARIMAX). In consideration that at the same time in the annual news cycle each year it is likely that some keywords and topics will repeat a peak in frequency (e.g. Christmas, Eurovision Song Contest, Victory in Europe Day etc.), we further include an annual seasonal component to conclude with a Seasonal ARIMAX (SARIMAX) model for prediction.

For training and testing different configurations, we chose the sports domain to take keywords from, inspired by our scenario partner Europeana who planned a content publication strategy around a 'Summer of Sport'. An initial exploration of sports keywords showed that there was an annual trend in their frequency, since sports events tended to run according to the same schedule each year, suggesting the SARIMAX approach could be effective. We settled in our tests on three keywords which each showed a particular behaviour:

- Cycling: this sport starts to peak in frequency in early March and then has repeated peaks throughout the year to early November with the highest peaks at the start and end of July (with the Tour de France).
- Wimbledon: with a small peak in early January, this sport grows in frequency first in the second half of June and its peak is in early July (when the Wimbledon tournament takes place), after which it drops off.
- Formula One: this sport has a first peak in March (season start) and then repeated peaks throughout the year until end of November (season end).

For training and testing, we exported 28 months of past frequency metric data for each keyword and predicted the following month. So we had 841 data points in the export (daily figure from 1 Jan 2018 to 28 Apr 2020), and ran the prediction for the following 30 data points (29 Apr 2020 to 28 May 2020).

4.2.3. Hybrid Prediction Model

It has been noted that our initial autoregression-based method of time series forecasting only allowed us to make accurate predictions a few days into the future. As we will show below in the evaluation section, we have found now a configuration with SARIMAX which supports a more stable predicted value longer into the future. This value can be seen only as a 'baseline' for the prediction as the longer we look into the future, the less it can change according to the changes in the moving average of the most recent time period. Indeed, we find the predicted value tends to remain static apart from variations generated by the exogenous variable (an event occurrence).

However, the temporal reference-based prediction does return a constantly varying frequency value over future dates which is the absolute number of documents in the corpus referencing the date. While we cannot expect smaller variations to be significant, particularly when the overall distribution of document frequency proves to be sparse (zero or single digits), we can assume that significant peaks in mentions in documents with a future date may also be associated with predicted higher mentions on that date. Conversely, the periods in which documents are sparse would be indicative of periods where it is not expected to see the keyword/topic being particularly mentioned.

Since this allows for a prediction of lower or higher relative popularity of the topic compared to the static baseline value of the time series forecasting, we will conclude our evaluations of the two prediction methods with a consideration of how to get to an effective hybrid model, where the baseline of the time series forecasting would be modified by the relative frequency of mentions with a future date in the temporal reference-based prediction.

4.3. Evaluation

4.3.1. Temporal Reference-Based Prediction

We sampled documents at random to check if the temporal reference detection had correctly annotated them with the dates in their text. We found for the corrected implementation focused on absolute date references and relative references to absolute dates that we had effectively a 100% accuracy in detection (the only issue we found was with relative references like 'tomorrow' which would be resolved with respect to the publication date but could be reported in relation to some other date). To further evaluate the correctness of the temporal reference-based prediction, we considered its use in ReTV predictive analytics and defined manually a number of 'correctness' conditions:

- For a given future date or date range, a certain topic should be predicted as trending;
- For a given topic, a certain future date or date range should be predicted as trending.

We aimed for one example in English (global) events and German events each per month from July to December, i.e. testing with a total of 12 cases. Given the necessity to identify key events to test for in the domains of culture, sport or politics, the final list (below) has 7 global events, 4 German and 1 Austrian.

Table 2: Chosen future events for temporal prediction evaluation

date(s)	event
4 July (DE)	DFB Pokal Endspiel (German cup final)
30 July – 2 August 2020 (EN)	Lollapalooza (original dates)
23 August 2020 (EN)	UEFA Champions League final
29 August 2020 (EN)	Tour de France
11 September 2020 (EN)	Anniversary of WTC terrorist attack
9-18 October 2020 (EN)	Coachella festival
11 October 2020 (DE-AT)	Vienna city elections
18 October 2020 (DE)	Bonn marathon
31 October 2020 (DE)	Opening of new Berlin airport
3 November 2020 (EN)	US presidential election
8 November 2020 (DE)	Tegel airport closure
25 December 2020 (EN)	Christmas Day

The condition for successful prediction is that there are keywords related to the event in the list of associations with the given future date(s), using either the English or German language news as prediction source. We cannot avoid that other documents will also refer to the same date about other subjects, and in 'real world' usage we would expect the user to be searching for future dates already with some topic pre-chosen (such as culture or sports) so that results would be filtered to the related documents and the more relevant associations emphasized. We found eight of our chosen events in the prediction mode but it is also interesting to highlight the four events that did not occur in the prediction data:

Table 3: For each future event, found YES/NO and predicted keywords for that date

event	found yes with keywords / no
DFB Pokal Endspiel (German cup final)	YES - bundesliga, dfb cup competition
Lollapalooza (original dates)	NO
UEFA Champions League final	YES - champions league
Tour de France	NO
Anniversary of WTC terrorist attack	NO
Coachella festival	YES - coachella, coachella valley
Vienna city elections	YES - Michael Ludwig, postal vote, election date, polls
Bonn marathon	YES - Bonn
Opening of new Berlin airport	YES - FBB, airport, Schoenefeld, Tegel
US presidential election	YES - election day, ballots, voter registration
Tegel airport closure	YES - air traffic, FBB, Schoenefeld, Tegel airport
Christmas Day	NO

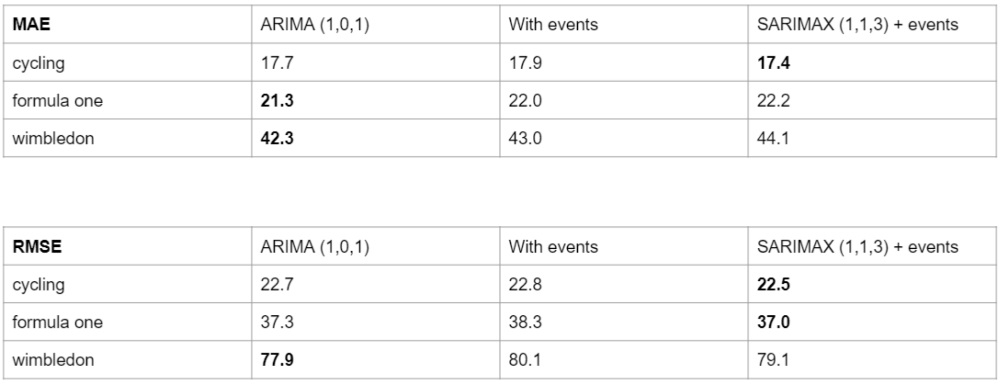
Among the matches, the associations in English new media for November 3, 2020 were the most interesting, with postal ballots and voter registration clearly being current issues discussed around the US election. Among the non-matches, Christmas was initially the most surprising but the news media is not yet talking much about the Christmas period and references often do not make explicit the date as it is well known (we may also note that Christmas can be identified as an event on December 24/25 through the events in the SKB). Similarly, the September 11 terrorist attacks are barely mentioned in the news so an anniversary would be best detected using the Anniversaries API. Both Lollapalooza and Tour de France are actually more surprising. For Lollapalooza, we were expecting an association with the original dates for the global music festival and if we search specifically on the keyword 'lollapalooza' we do find a peak in mentions with 1-2 August and again with 6 September (it turns out this was the original date of a Lollapalooza branded music event in Berlin). It appears that in the English media there were not enough associations but in the German media the keywords 'organizer' and 'festival' both are predicted for the dates 30 July - 2 August, and while Lollapalooza is mentioned 9 times in documents associated with this time period, the German media only references Lollapalooza's September dates in Berlin. The top associations are two other festivals 'Wacken' and 'Hurricane' which are also postponed from these original dates and Lollapalooza appears in association with them. For the keyword 'tour de france' the peak occurs on 31 August and it seems the news media currently mainly associates the start of the sporting event with the end of the month, so maybe this association is not yet strongly enough reflected in news coverage even though it seems 29 August is now confirmed as the new start date.

In brief, this approach complements our events and anniversaries entities stored in the SKB with further associations between events and future dates extracted from news media. This is clearly dependent on precisely what associations are made, as could be seen with Lollapalooza and Tour de France from the current prediction sources. In one case, Lollapalooza was more often associated with a different set of dates, in the other the association is still stronger with a different date in the current news corpus. As more news is collected and analyzed, we expect this becomes more accurate, especially as we come closer to such events and they are discussed comparatively more in the news media.

4.3.2. Time Series-Based Prediction

The results of the testing with the three chosen keywords as measured by standard metrics (MAE and RMSE) are shown in the table (Figure 5). We took a 90/10 training/testing split in the extracted data and used ARIMA(1,0,1) as a starting point as it is close to the original prediction model (which would have 1,0,0 i.e. only autoregression) incorporating also one moving average calculation. As can be seen, we saw different results for the three different keywords with 'cycling' prediction being the most accurate, 'wimbledon' the least accurate and 'formula one' performing inbetween. This is not so significant as we know that every keyword will act differently both in the activity of its past success metrics as well as in how the same success metric will behave in the future. However, we can postulate that, since the dates of the testing period were in early 2020 (Feb-Apr) cycling performed best because cycling events were still taking place up to early March and therefore the keyword was still close to the usual annual trend. As corona virus began to affect schedules, some cycling races still went ahead virtually. However, Wimbledon was cancelled on 1 April 2020 after a period of speculation, so the keyword would have acted quite differently to previous years. Then again Formula One also saw its usual season start in March delayed indefinitely too but the keyword performed quite well in the tests.

After establishing the performance of ARIMA(1,0,1) we added past and future events as an exogeneous variable (ARIMAX). This did not bring any improvements though we had seen in the D2.2 experiments already a potential benefit from capturing both when a related event occurred on a past date as well as when a similarly related event will occur on a future date (with Eurovision as the example). For cycling, we had multiple past events (using each official



MAE	ARIMA (1,0,1)	With events	SARIMAX (1,1,3) + events
cycling	17.7	17.9	**17.4**
formula one	**21.3**	22.0	22.2
wimbledon	**42.3**	43.0	44.1

RMSE	ARIMA (1,0,1)	With events	SARIMAX (1,1,3) + events
cycling	22.7	22.8	**22.5**
formula one	37.3	38.3	**37.0**
wimbledon	**77.9**	80.1	79.1

Figure 5: Table of results from testing forecasting models

cycling competition as an event occurrence) as well as included still the current future cycling calendar even if the competition may have been cancelled (with events taking place in March and April which were within the test data period). So it is worth noticing that the accuracy loss with the events was negligible here. Regarding the other keywords, neither 'formula one' nor 'wimbledon' had any events in the test data period (we excluded the formula one races that were originally scheduled in Mar/Apr as we knew they had been postponed) and we see a small loss in prediction accuracy as a result - it seems that when there are no events in the prediction period, it is better to not include any of the past events in the training for that prediction task.

We then tested different ARIMAX configurations and included annual seasonality into the model (SARIMAX with a seasonal order of 0,0,0,365), finding that our best results overall were achieved with an ARIMA(1,1,3) model. As hoped, this did have a small beneficial effect on prediction accuracy except for the case of wimbledon.

Since a month had passed since the testing, we decided to also look at predicted values outside of the extracted dates. As the extracted dates were up to April 28, 2020 we could now use the model to predict the next 30 days' values and compare them to the actual numbers. We also considered in the prediction evaluation what would be the most meaningful metric for prediction success, taking into account that our keywords will always be subject to strong external factors outside of our possible knowledge and therefore we can only reasonably expect to predict that a keyword will be relatively more or less popular on a future point of time. While MAE and RMSE are useful in testing the trained model as standard metrics within the field, as will be shown in the following figures and explanations, we found two other measures to use for the prediction accuracy: MPE (Mean Percentage Error) which addresses better just how far a prediction is relatively accurate with respect to the actual value (using percentages of error rather than absolute numbers) and an accuracy within 20% margin of error which reflects our consideration that we do not need exact matches between predicted and actual values in the ReTV case.

The three figures show the predictions for our three keywords. For the actual values, we show not only the absolute frequency of documents on each day but also the 7 day moving average (MA). As it can be seen that the 7 day MA tends to have a more similar shape to our prediction values, we also decided to measure the accuracy against the moving average rather than the absolute values. Since the moving average is more representative of the keywords

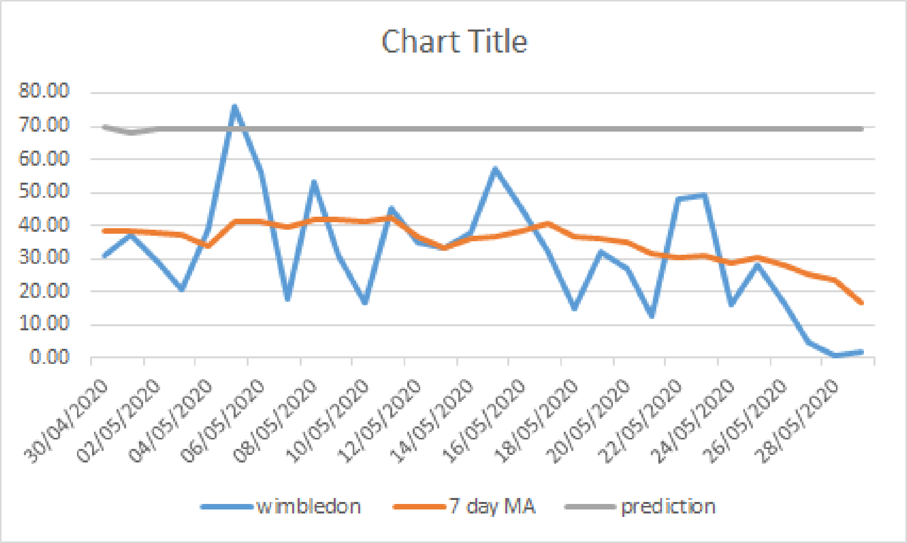


Figure 6: Predicted and actual values for keyword 'wimbledon'

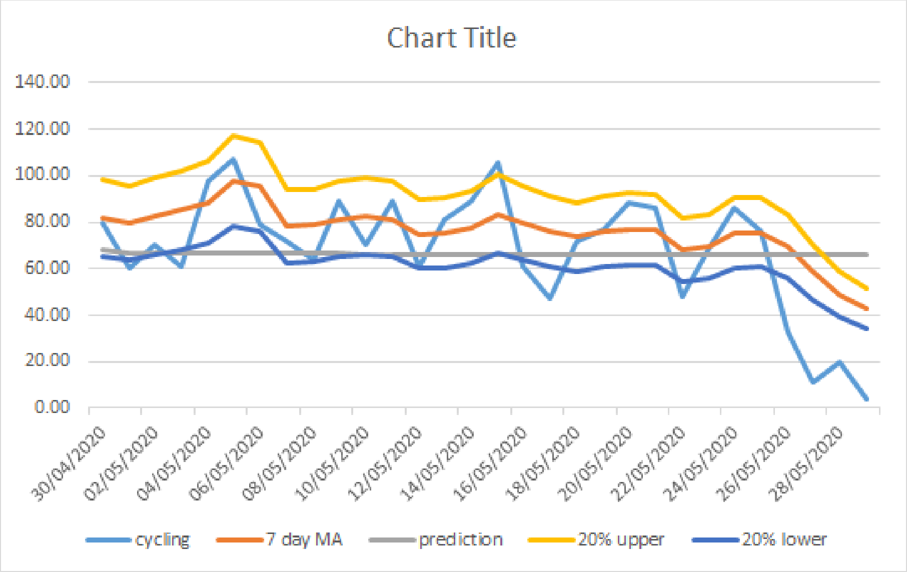


Figure 7: Predicted and actual values for keyword 'cycling'

relative visibility to the audience as a more enduring topic in the online discourse, this also seems a more reasonable assumption that using the absolute values, where there could be a large peak (or trough) on a single day which does not necessarily mean the keyword is visible to the audience any more or less the next day. For cycling and formula one, we include in the figures the 20% upper and lower bounds of the variation from the absolute figure in order to include the calculation of accuracy of the prediction within this 20% margin of error (this is not shown in the first figure for 'wimbledon' as the predicted values are clearly more than 20% higher than the actual values except for a single day in the prediction period).

Beginning with 'wimbledon' in Figure 6, the predicted baseline was consistently higher than the actual. Our finding was that wimbledon was normally (i.e. in previous years) mentioned more in this period but probably due to the cancellation, an exceptional event that no-one

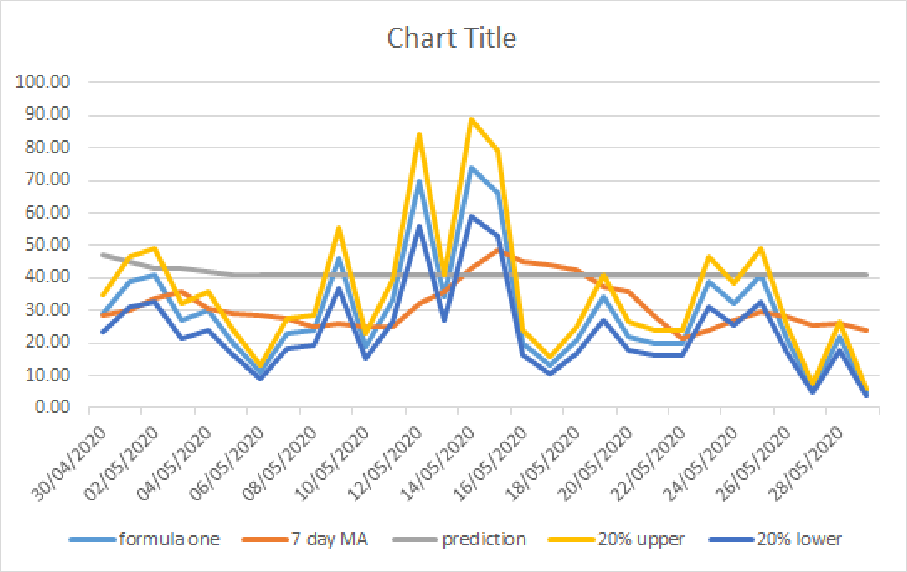


Figure 8: Predicted and actual values for keyword 'formula one'

could have predicted, the news coverage is much less this year. Therefore there is little that could be done to more accurately predict this unless we could consider some sort of indicator that an event is cancelled or postponed (e.g. a '-1' value in the exogeneous variable for event occurrence) and train our model with that; of course since one probably does not have already the case in the past data it is not possible to train the model how to deal with it in future prediction and creating an artificial behaviour (i.e. train the model that there are less mentions of a keyword when the related future event(s) is/are cancelled) falls into the danger that actual behaviour will differ (e.g. as we will see with 'formula one' the news mentions may actually be the same or more, as the news deals with reports on when and how the sport might start again). However, we could observe that while the reported RMSE on the moving average looked as if the accuracy was better (34.5 compared to 79.1 on the test dataset), the calculated MPE (105.4) was higher than any other keyword and thus better reflected the relative inaccuracy of the prediction in this case.

Turning to 'cycling' in Figure 7, which was already our best case in the test stage, it proves again to be the best performer also in the prediction. Here our baseline prediction is quite close to the 7 day MA throughout, and indeed half the time our predicted values are within the 20% margin of error from the actual values. We can also observe that we have a low MPE of 17.6 together with a similarly good RMSE of 14.0.

Finally, 'formula one' in Figure 8 which was in the middle of the test results and is also similarly in-between for prediction. Again our baseline values tend towards being higher than the usual actual values, but a number of peaks in May push up the moving average and we have 7 days in the 30 day period where our prediction is within that 20% margin of error. The MPE of 41 is similarly between the values of the other two keywords, while the RMSE of 12.5 showed a welcome closeness in prediction to the 7 day moving average.

The choice of keywords seem quite fortuitous as we could observe all worlds in this evaluation - a keyword which performs quite well in prediction, a keyword which didn't perform well at all and a keyword somewhere between the two.

4.4. Considerations for a Hybrid Solution

We have already stated several times that we cannot expect great accuracy in the form of prediction we are looking at in ReTV, since we consider various keywords and topics in the online discourse which are subject every day to previously unforeseeable events that change how and to what extent they may be a subject of discourse for the audience on the selected digital channel. In this year, we had a rather extreme example of this with the emergence of the corona virus pandemic and the resulting disruption to global life, work and of course events such as sports.

However, the work continued along these two paths of predictive modelling: time series-based forecasting and temporal reference-based detection. We found that forecasting could produce stable baseline prediction values for keywords/topics that, to a varying extent, would match the actual future values within a given margin of error (and could be closer to the actual 7 day moving average). Since this is using past keyword metrics with an annual seasonality, one could say that the closer a keyword is present in the online discourse in this year compared to past years, the more accurate our forecasting should be. While not all 'out-of-trend' future events can be knowable and therefore our prediction can always be knocked off by such external and uncontrollable factors, we still have two chances to modify the baseline according to knowable future events. One is the known events that are explicitly available through our own event extraction and modelling (as Event Entities in our Semantic Knowledge Base) and can be incorporated as both past and future events as exogenous variables in the training and use of the prediction models. The other is predicted peaks or troughs in keyword frequency compared to the usual level of mentions as provided by the temporal reference detection. Therefore, to find a model that optimally predicts keywords' relative future popularity for a content publication strategy, we want to conclude this prediction activity with consideration of how we could combine these approaches into a potentially better performing hybrid model.

The assumption is that the forecasting produces an acceptable baseline value but that the temporal reference detection can indicate future periods where a keyword can be expected to have a higher or lower than usual visibility among the audience. This can be used to produce a variation in the baseline that brings it closer to the actual values.

5. Conclusion and Outlook

5.1. Event Extraction and Modelling

Our event KB continues to grow and be updated; we have tested and shown the capability to also manually add further events directly into the SKB as required by any scenario. Events can also now be searched and browsed in the SKB through a Web-based interface called the SKBBrowser (see ReTV deliverable D1.3), as well as requested via API. This API will be configured for different scenario partners to return ranked lists of matching events most relevant to their scenario needs, starting with testing with our own use case partners RBB and NISV. The scenario-adapted API will be used to integrate events and anniversaries of interest into the Content Wizard tool being prepared in ReTV workpackage 5.

5.2. Success Metrics

The work on content-based success metric extraction has led to several options for the data analysis tasks. In the specific domain of television supported by the ReTV project, we needed to go beyond 'classical' metrics such as past audience when looking at how best to support media organizations to optimally publish about their media content. Going beyond this, there is still plenty of opportunity to explore the use of the other success metrics in the data analysis for media organisations as well, such as the use of WYSDOM as an alternative to sentiment for tracking TV content communication success or the tracking of emotions expressed towards TV content on digital channels. We will look at incorporating these aspects in the use case evaluations to be done in WP5 of ReTV.

5.3. Predictive Analytics

The predictive analytics work has reached a valuable milestone, with the temporal reference detection integrated into the Topics Compass scenario so that users may also now switch to a Prediction Mode, give a future date range and explore the relative frequency of topics and keywords associated in documents to that future period.

Furthermore, the time series forecasting has established a means to produce a stable predicted value for future frequency of a topic or keyword on a digital channel based on past frequency measurements. Finally, the work on event extraction and modelling ensures we have a knowledge base available for identification of events of relevance to a topic or keyword with both related past and future occurrences so that events may also be used in the training for the prediction model. Results in accuracy can vary greatly as the extent to which a keyword or topic has been affected by events independent of any trend or seasonality, and the extent to which a keyword or topic will be affected by such in the future, cannot be adequately incorporated into a forecasting model. The emergence of the topic of corona virus to dominate global news coverage is an extreme example of this.

However we have recognized that the temporal reference detection may be a further input to a hybrid prediction solution that bases its initial prediction on the time series-based forecasting and then modifies that prediction according to peaks and troughs in the number of documents associating the keyword or topic to each future date. This idea needs further investigation and we will perform sufficient data analyses combining the results of both prediction models together in comparison with the actual values in order to establish a hybrid implementation before the end of the project.

The prediction work of ReTV does not take place in a vacuum and indeed an important future step in this work will be the incorporation of prediction results into the interface of the Content Wizard, the primary end user tool in ReTV for the media organisation workflow (selecting, summarizing and scheduling the publication of media content). The topics predicted to be of most importance to the organization's audience may be pre-selected in the interface so that the relevant media assets are already displayed, for example. Visualisations from the Topics Compass for the relative associations of topics being tracked by the organisation with a future date range can also be included into the Content Wizard. A calendar view will include events and anniversaries filtered according to the respective scenario. All of these inputs will help guide the Content Wizard user to better choose which media content they should best post or promote (according to topic) on digital channels in future publications, the core objective of the ReTV project.