Social Media and Google Trends in Support of Audience Analytics: Methodology and Architecture

Nikos Kalatzis, Ioanna Roussaki, Christos Matsoukas, Marios Paraskevopoulos, Symeon Papavassiliou
Institute of Communications and Computer Systems
Athens, Greece
e-mails:{nikosk@cn, ioanna.roussaki@cn, cmatsoukas@cn, mariosp@cn, papavass@mail}.ntua.gr

Simona Tonoli
European & Funded Projects Department
Mediaset
Milan, Italy
e-mail: Simona.Tonoli@mediaset.it

Abstract — **In recent years, there have been various research efforts aiming to investigate how social media are used to express or influence TV audiences and if possible to estimate TV ratings through the analysis of user interactions via social media. Given that these efforts are still in their infancy, there is a lack of an established methodology for designing such frameworks or services. This paper reviews the most dominant existing approaches and identifies the fundamental design principles aiming to generate best practices and guidelines for such systems. It then proposes a methodology and a reference architecture that can be employed by those that aim to exploit social media data to support audience analytics and extract TV ratings. Finally, this paper introduces and evaluates the utilisation of Google Trends service as an additional information source in support of audience analytics.**

Keywords-social media data; audience analytics; methodology; reference architecture; Twitter; Facebook; Google Trends.

I. INTRODUCTION

TV ratings have been crucial for the society due to the fact that they influence the popular culture, but also for the media industry as they are the basis for billions of dollars' worth of advertising transactions every year between marketers and media companies [1]. For more than 50 years, TV ratings are estimated by sampling the audience with specific installed hardware on TV devices. In the meantime, there is an increasing trend of people watching TV programs that on the same time are also interacting via social media services posting messages or other content mediating their opinions. In addition, new services are drastically changing the media consumption patterns as for example it is now possible to watch TV programs on YouTube regardless of location or time.

This proliferation of social media utilisation by large population portions along with recent advances in data collection, storage and management, makes available massive amounts of data to research organisations and data scientists. Exploiting the wealth of information originating from huge repositories generated for example by Twitter and Facebook has become of strategic importance for various industries. However, the availability of massive volumes of data doesn't automatically guarantee the extraction of useful results, while it becomes evident that robust research methodologies are more important than ever.

There are already various research efforts aiming to exploit data originating from social media sources in support of audience analytics. Until today, these efforts are mainly offered to complement the traditional TV ratings and do not aim to substitute them. However, as stated in [2], traditional approaches are demonstrating various limitations due to the fact that they are necessarily sample based with a relatively small number of installed metering devices due to the high cost of them. In addition, traditional approaches can hardly take into account new viewing behaviours such as the mobility of audience members and nonlinear viewing.

This paper aims to review the most dominant research efforts for social media analytics focusing on the extraction of additional insights about TV audiences. Based on this review and on authors' own evaluations, this paper proposes a five stage methodology and introduces a reference architecture that can be used as a starting point for any research work studying the usefulness of social media data for audience analytics purposes.

The rest of this paper is structured as follows. Section II reviews the main research initiatives that aim to extract audience analytics metrics by monitoring any related keyword-specific traffic across selected social media. Section III proposes a methodology for building a social-media based audience analytics framework and maps the most dominant related research initiatives to specific decisions made across the five steps of this methodology. Section IV introduces a reference architecture suitable for the implementation of such a framework. Section V presents experimental findings that support the extension of social media data sources with Google Trends data aiming to optimise the performance of the proposed audience analytics framework. Finally, conclusions are drawn and future plans are exposed.

II. RELATED WORK

In recent years, there is an increasing trend on analysing social media and Internet search engines utilisation for studying and examining behaviour of people with regards to various societal activities. The proper analysis of these services goes beyond the standard surveys or focus groups and has the potential to be a valuable information source leveraging internet users as the largest panel of users in the world. Researchers and analysts from a wide area of fields are able to reveal current and historic interests for

 ISBN: 978-1-61208-681-1

communities of people and to extract valuable information about future trends, behaviours and preferences. Some of the fields where social media analytics have been employed for such purposes include: *economy* (stock market analysis [3] and private consumption prediction [4]), *politics* (opinion polls [5] and predictions of political elections [6]), *public health* (estimate spread of influenza [7] and malaria [8]), *sports* (predict football game results [9]), *tourism* (places to be visited by observing the most frequently attended places in a given location [10]), *demographics* (identifying gender and age of selected user groups [11]) and *infotainment* elaborated upon hereafter.

There are numerous research initiatives that apply social media analytics to estimate potential popularity of multimedia content. For example, authors in [12] propose a mechanism for predicting online content's popularity by analysing the activity of self-organized groups of users in social networks. Authors in [13] attempt to predict IMDB movie ratings using Google search frequencies for movie related information. Similarly, authors in [14] are applying social media analytics for predicting potential box office revenues for Bollywood movies based on related content shared over social networks. In the work presented in [15], social media and search engines utilisation are analysed during the pre-production phase of documentaries in order to identify appealing topics and potential audiences.

Based on the findings of the aforementioned initiatives, social media data demonstrate a relevant and flexible predictive power. The underlying relations among social media data and predictive variables not known a priori. The extraction of these variables and the utilisation of the appropriate algorithms can lead to quantitative statistical predictive models of several social targets of interest. A research field that gains significant attention with huge economical potential is the application of social media analytics in support of audience estimation for TV shows. Some of the existing efforts are presented here after.

One of the first approaches towards this scope is presented in [16]. Authors introduced concepts such as "Textual relevance", "Spatial Relevance", and "Temporal Relevance" along with the respective formulas for measuring the relevance of a Tweet with a targeted TV show. These metrics were utilized along with the total volumes of tweets and users for calculating the popularity of TV shows. However, cross-validation of this approach with ground truth data is missing.

The research presented in [17] mainly focuses on TV drama series that airs once a week. Authors attempt to estimate future audience volumes of TV shows through a predictor which is based on three layer back-propagation neural network. The predictor is fed with input from Facebook (e.g., number of related posts comments, likes, and shares) along with the ratings of the first show of the season and the rating of the previous show. According to the authors, the prediction accuracy of this approach based on Mean Absolute Percentage Error (MAPE) was from 6% to 24%.

The work presented in [18] is one of the first attempts for creating a statistical model with the goal of predicting the audience of a TV show from Twitter activity. Authors collected a large number of Tweets containing at least one of the official hash-tags of the targeted political talk shows. After analysing the data, a significant correlation was discovered between Twitter contributors per minute during airtime and the audience of the show's episode. Based on these results, a multiple regression model was trained with part of the dataset and utilized as a predictor for the remaining observations to evaluate its performance.

In [19], interactions between television audiences and social networks have been studied, focusing mainly on Twitter data. Authors collected about 2.5 million tweets in relation with 14 TV series in a nine-week period aired in USA. Initially, tweets were categorised according to their sentiment (positive, negative, neutral) based on the use of a decision trees classifier. Further analysis included the clustering of tweets based on the average audience characteristics for each individual series, while a linear regression model indicated the existence of a strong link between actual audience size and volume of tweets.

Authors in [20] defined a set of metrics based on Twitter data analysis in order to predict the audience of scheduled television programmes. Authors mainly focus on Italian TV reality shows, such as X Factor and Pechino Express where audiences are actively engaged. According to the authors, the most appropriate metrics are related to the volume of tweets, the distribution of linguistic elements, the volume of distinct users involved in tweeting, and the sentiment analysis of tweets. Based on these metrics, audience population prediction algorithms were developed that have been validated based on real audience ratings.

Similarly, research efforts presented in [2] and [21] are utilising Twitter data in an attempt to improve TV ratings, but these approaches lack extensive validation.

III. METHODOLOGY FOR BUILDING A SOCIAL-MEDIA BASED AUDIENCE ANALYTICS FRAMEWORK

In the last years, the vast use of social media gave us the ability to accurately predict or discover various events exploiting data freely available online. The most suitable methodology that enables researchers to build an efficient audience analytics mechanism based on social-media data has been studied by several initiatives [22]-[24]. Building on these and based on the main big-data mining principles [25]-[27], as well as on the online social network data collection and analytics trends and approaches [28]-[30], we propose a five-stage methodology that is depicted in Figure 1 and is briefly described in this section.

The five main steps that compose the proposed methodology for social media data exploitation in support of audience analytics are the following:

1. SM Data Identification

- Identify the most appropriate social media sources to be used such as popular SM networks (e.g., Twitter, Facebook, Google+, YouTube, LinkedIn, Pinterest, Instagram, Tumblr, Reddit, Tumblr, Snapchat, etc.) or more focused sources (e.g., web fora, blogs, message boards, news sites, podcasts, Wikis, etc.).

- Identify the type of data (e.g., comments, tweets, posts, likes, shares, links, connections, user account details, etc.) to be collected.
- Specify the criteria to be used for data collection (e.g., keywords, hashtags, timeframe, geographic area, language, user account properties, etc.)

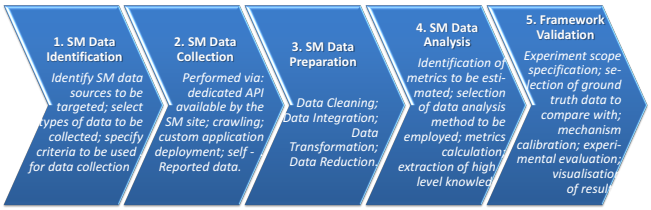

Figure 1. Social media data mining methodology in support of audience analytics.

2. SM Data Collection

- Most often, the SM sites provide an Application Programming Interface (API) that can be used by the developers to collect data based on the type and criteria specified in previous Phase. In this case, initially the necessary libraries need to be installed, the authorization needs to be obtained and finally a decision needs to be made regarding the platform/language to be used for writing the data collection software.
- In case no such API is made available by the SM site, crawling can alternatively be used with an automated script that explores the social media website and collects data using HTTP requests and responses.
- Another method that enables collection of SM data is the implementation and deployment of a custom application based on an SM site that monitors its usage.
- If the methods above cannot be employed, self-reported data can be used instead, which requires for directly asking the users about their interests, opinion, experience, etc., mainly via online questionnaires or in-situ surveys.

  Various shortcomings and limitations need to be carefully addressed during the "Data Collection" phase. For example, the standard version of Twitter API enforces "Rate Limits" allowing a certain number of calls on 15 minute intervals. In addition, the standard Twitter API allows the retrieval of Tweets for a period of the last 7 days. In a similar manner, Facebook enforces strict privacy protection policies which are limiting the amount of information that can be retrieved. These policies are subject to frequent updates which services consuming the APIs should follow.

3. SM Data Preparation

- Data cleaning (or cleansing) aims to fill in missing values, smooth out noisy data, identify and remove outliers, minimise duplication and computed biases, and correct inconsistencies within the data collected in the previous phase. Given that data are collected based on criteria such as the inclusion of a keyword or a hashtag it is highly possible that the actual Tweet or Facebook post to be irrelevant with the targeted show. In a similar manner, during this phase data generated by automated software scripts – known as bots – should also be identified and filtered out.
- Data integration combines data from multiple social media sources into a coherent store, while it aims to detect and resolve any data value conflicts or data redundancies that may arise in this process.
- Data transformation converts the raw social media data into the specified format, model or structure. Methods used for this purpose include: normalization (where numerical data is converted into the specified range, i.e., between 0 and one so that scaling of data can be performed), aggregation/ summarisation (to combine features into one), generalization or attribute/feature construction (where lower level attributes are converted to a higher standard or new attributes are constructed from the given ones in general) and discretization (where raw values of a numeric attribute are replaced by interval labels).
- Data reduction aims to obtain a reduced representation of the social media data set collected that is much smaller in volume but yet produces the same (or almost the same) analytical results. Methods employed for this purpose include: data compression, data sampling, and dimensionality reduction (e.g., removing unimportant attributes) and multiplicity reduction (to reduce data volume by choosing alternative, smaller/more compact forms of data representation).

4. SM Data Analysis

- At this stage the pre-processed SM data collected need to be analysed in order to extract results linked with the popularity of the broadcasted TV program. Specific metrics are defined indicating audience statistics aspects, such as: number of messages referring to the targeted show (e.g., tweets, posts), number of interactions associated with a post (e.g., liked, favorited, retweeted, shared, etc.), number of unique users participated in the overall interactions. In some cases, researches are setting their own objective functions and define their own scores for evaluating the generated Social Media buzz.
- Aiming to identify correlations between the Social media data and the actual audience's interest, the following statistical approaches are often utilised by the research community: Logistic Regression (LR), Density Based Algorithm (DBA), Hierarchical Clustering (HC), AdaBoost, Linear-Regression(Lin-R), Markov, Maximum Entropy (ME), Genetic Algorithms (GA), Fuzzy, Apriori, Wrapper, etc.
- Inference of higher level information based on the analysis of raw collected data aiming to extract additional characteristics of the audience. Processing raw data through sophisticated algorithms allows the extraction of information that are not provided by directly by SM APIs, such as the classification of a post/tweet with regards the sentiment (positive,

 ISBN: 978-1-61208-681-1

negative, neutral) and the profile of the SM user (e.g., gender, age, political views, hobbies, other preferences). Towards this scope, there are numerous data analytics techniques that have been applied by researchers [31], each suitable for specific problems and domains. These tasks are usually handled as a classification/ categorization problem.

5. Framework Validation

- Validation of the outcomes generated by the Data Analysis step allows drawing the respective conclusions whether and in which extent the overall approach achieved the desired outcomes. Within the scope of estimating audience volumes interested or watched a TV show through the analysis of SM data, often the Data Analysis results are cross-validated with published audience rates metrics.
- These metrics also contain audience profile information (e.g., age, gender, location, occupation) hence advanced analysis methods can also be cross validated. However, these metrics are not always publicly available or the published measurements are generic and not include details about the profile of audiences.
- So far, realisation of such techniques has revealed various shortcomings mainly related to the over or/and under representation of certain societal groups within SM services. Statistical approaches for validation include but are not limited to: Root Mean Square Error, Mean Absolute Error, Pearson correlation coefficient, Akaike Information Criterion, etc.

The proposed methodology described herewith has been used by the authors in several situations in the domain of social media data exploitation for audience analytics or prediction purposes. In all occasions, it has proven to be very effective, when proper elements and configuration are employed at each stage.

As already stated, the five-step methodology is the outcome of a thorough review of the most dominant state of the art approaches in the area of social-media data processing in support of TV show audience analytics. To this end, Table 1 presents a summary of this analysis, where the various steps employed by the most popular approaches presented in the state of the art review in Section II are mapped to the main elements of the methodology proposed herewith, while the specific mechanisms employed are identified.

TABLE I. MAPPING THE MOST DOMINANT STATE OF THE ART WORK TO THE PROPOSED METHODOLOGY STEPS

Refe-rence	Step#	Step Elements
[2]	**Step 1**	Japan geo-tagged tweets containing hashtags related with the show and the TV channel. Targeted genres of shows are: News, documentary, talk show, life style
	Step 2	Native Twitter API
	Step 3	Algorithm for calculating score reflecting the relevance of tweets with targeted show
	Step 4	Identification of overall popularity of the show based on the volume of "relevant" tweets and the absolute number of users
	Step 5	No cross-validation with ground truth data
[20]	**Step 1**	Tweets containing hashtag related to Reality shows in Italy
	Step 2	Collection of Tweets based on "Twitter Vigilance" multiuser tool developed for research purposes
	Step 3	Sentiment Analysis based on a score for positive, negative, and neutral mood
	Step 4	Predict audience of scheduled TV programmes based on volumes of (re)tweets, of unique users "tweeting", sentiment analysis scores for each textual element in the tweets. Statistical approaches utilised are: Principal Component Analysis, Multi-linear Regression & Ridge Models
	Step 5	RMSE-Root Mean Square Error, MAE - Mean Absolute Error
[19]	**Step 1**	Tweets containing the official hashtag of selected popular USA TV series, possible hashtag derivatives, official account of the television program.
	Step 2	Python script, interacting with Twitter Streaming API that allows for real-time downloading of tweets containing certain keywords.
	Step 3	Sentiment categorization based on Knime Decision Trees algorithm. Manual classification for an initial set of 14,000 tweets.
	Step 4	Audience prediction estimation based on the calculation of volumes of positive, neutral and negative tweets considering different time frame windows: (i) within 3 hours after episode start, (ii) within 24 hours after episode start Statistical approaches utilised are: Linear Regression Models
	Step 5	p-value and t-test significant test
[17]	**Step 1**	Data from Facebook "fan pages" regarding the TV shows: #page posts, #fans posts, #page posts comments, #fans posts likes, #fans posts shares, #fans posts comments, #page posts likes. Taiwan: TV drama series aired once a week.
	Step 2	Facebook API for collecting data from the official page of the show
	Step 3	Repeated respondents in the same article were filtered out to avoid large amount of increased responses due to special events (such as quizzes or Facebook Meeting Rooms, etc.)
	Step 4	Accumulation of the broadcasted TV programs' word-of mouth on Facebook and apply the Backpropagation Neural Network to predict the latest program audience rating.
	Step 5	Mean Absolute Error, Mean Absolute Percentage Errors
[21]	**Step 1**	Tweets containing only the most commonly used TV shows hashtags Netherlands: Country's top-25 TV shows with high number of tweets. #tweets #hashtags (tweets posted half an hour before broadcast, during broadcast and half an hour after the end on the shows)
	Step 2	Native Twitter API
	Step 3	None
	Step 4	Correlation of tweets volumes with audience measurements with the utilisation of Linear Regression Models.
	Step 5	Pearson correlation coefficient
[32]	**Step 1**	Official Facebook page of TV show. USA: Reality, Drama and Sports series. Average engagement per post, Fans on Facebook, Posts in Total, Links in Total, Photos in Total, Videos in Total, Included question post in Total, Unique engaged audiences on Facebook
	Step 2	Facebook API
	Step 3	Not specified
	Step 4	Data Visualization of Correlations, PCA for Dimension Reduction, Multiple Regression Analysis
	Step 5	Akaike Information Criterion, Bayesian Information Criterion, ANOVA, Word Cloud Analysis

IV. PROPOSED ARCHITECTURE FOR SOCIAL MEDIA-BASED AUDIENCE ANALYTICS

Based on the proposed methodology and building on the architectures proposed by the most popular state of the art related initiatives, this section proposes a reference architecture suitable for social media based audience analytics. Among the design principles of this architecture is the ability of the service to query various Social Media

Services (e.g., Twitter and Facebook) through pluggable connectors for each service. The respective high level functional view is presented in Figure 2 and its core modules are described hereafter.

RestAPI: It exposes the backend's functionality v a REST endpoint. The API specifies a set of SM data mining functions where the service consumer provides as input various criteria such as data sources, data types, keywords, topics, geographical regions, time periods, etc.

SM Data Query Management: This component orchestrates the overall execution of the queries and the processing of the replies. Given the input from the user (e.g., targeted keywords, targeted location, time frame) several properly formulated queries are generated that are forwarded to the respective connectors/wrappers to dispatch the requests to several existing Social Media services available online. The SM Data Query Management enforces querying policies tailored to each service in order to optimize the utilization of the services and to avoid potential bans.

Social Media Connectors: A set of software modules that support the connection and the execution of queries to external services through the provided available APIs or via tailored crawlers. Connectors are embedding all the necessary security related credentials to the calls and automate the initiation of a session with the external services. Thus, the connectors automate and ease the actual formulation and execution of the queries issued by the SM Data Query Management component. Some example APIs that are utilized by the connectors are: Twitter API, YouTube Data API v3, Facebook API.

SM Data Collection Engine: Given that each external service will reply in different time frames (e.g., a call to Google Trends discovery replies within a few seconds while Twitter stream analysis might take longer time periods) the overall process is performed in an asynchronous manner, coordinated by the Data Collection Engine,

SM Data Pre-processor: This module performs the necessary data cleaning aiming to improve the quality of the collected data and to prepare them for the actual data analysis through data transformation mechanisms that models raw data into a uniform model.

SM Data Analytics Engine: This module maintains a repository of statistical and data mining methods that can be applied on data sets previously prepared by the SM Data Pre-processor module.

Analysis Results Database Management: The outcome from various data analysis tasks are maintained to a local database. The Database Management module supports the creation, retrieval, update and deletion of data objects. Hence, it is feasible for the user to compare analysis tasks reports performed in the past with more recent ones and have an intuitive view of the evolution of trend reports in time.

Front End: The Front-End visualizes the results providing the following output: (i) various graphs presenting the absolute volume of retrieved messages (e.g., number of tweets, number of Facebook posts), (ii) calculated scores indicating audience interest for different shows, per location (country), time period, (iii) higher level information e.g., audience's sentiment or gender.

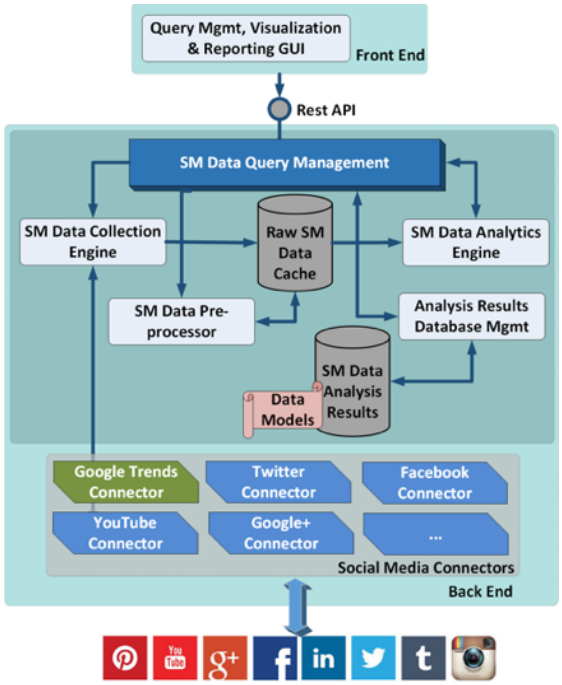

Figure 2. Reference Architecture for using social media data to support audience analytics.

A proof of concept of the described architecture has been implemented in Python 3 programming language. For the SM Data Analytics Engine the scikit-learn [33] python package has been integrated, while results are stored in a MySQL relational database. Currently connectors for Twitter, Facebook and Google Trends have been integrated.

V. WHAT ABOUT GOOGLE TRENDS?

Based on the state of the art review, the research work conducted so far by various initiatives on the domain of Social Media data usage to extract information related to audience analytics focuses on Twitter and Facebook. In certain occasions, the obtained results are not of adequate quality and reliability. In an attempt to treat this, the authors are currently experimenting with using additional information obtained via Google Trends [34]. Google Trends is a public web facility of Google Inc. that presents how often a specific search-term is entered on Google Search relative to the total search-volume across various regions of the world, and in various languages. The idea is to couple this information with data extracted by Twitter for example in the framework of a TV show or program in order to enhance the quality of the respective audience statistics extracted.

In an initial attempt to evaluate this approach, we focused on the Italian talent show "Amici di Maria de Filippi" that broadcasts for the last 17 years and lies among the most popular shows in Italy. The show airs annually from October until June, thus being appropriate for yearly examination of the data. In this study, data of the year 2017 have been used, split in two semesters as elaborated upon subsequently.

 ISBN: 978-1-61208-681-1

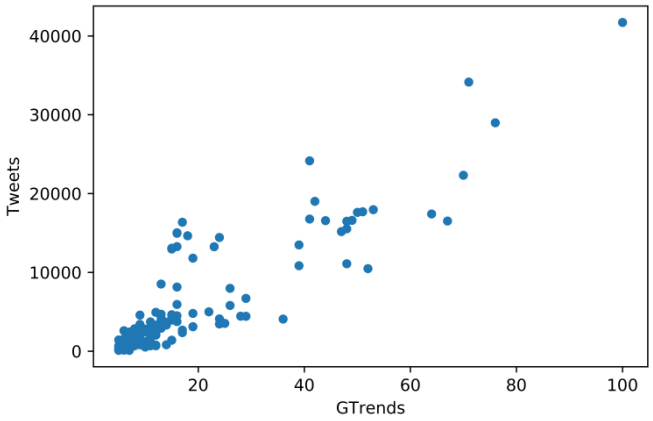

Figure 3. Correlation of Google Trends and Twitter data for the term '#amici16' targeting the first semester of 2017

Google Trends (GTrends) provides a variety of chronological and geographical metrics that show the search activity for the term in question. The one used in this study is a time series of the relative search figures -search volume for the term divided by the total volume of the day- normalized between 0 and 100. One limitation of the platform is that one can only get daily figures for a maximum range of 270 days, which of course is less than a year. To overcome this obstacle, the year has been split in two semesters, which also allows us not to mix results, since the TV season starts during the second semester of the year. Data from GTrends require no further processing, since they are provided in the format required for this experiment.

On the other hand, the data obtained via Twitter have been extracted on a monthly basis and have been grouped based on date in order to acquire the daily volume. Tweets retrieval was based on the hashtags '#amiciXX' where XX corresponds to the number of the consequent season that the show is aired. During the period January -June 2017 and based on the hashtag '#amici16' there where 882024 tweets collected while during the period July to December 2017 and based on the hashtag '#amici17' there where 135288 tweets collected.

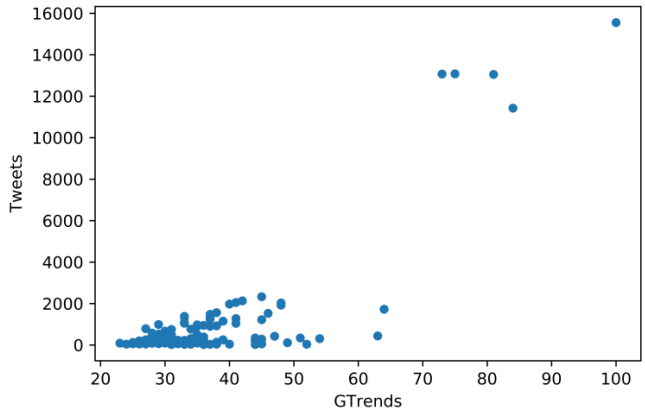

Figure 4. Correlation of Google Trends and Twitter data for the term '#amici17' targeting the second semester of 2017.

In order to verify the correlation between data originating from Google Trends and those originating from Twitter, the Pearson correlation coefficient was utilised. The obtained results for the first semester of 2017 are illustrated in Figure 3 and lead to coefficient of 0.893 and to significance of approximately 10-32. This indicates that the two datasets are strongly correlated, since we secured that the figures of each set are matched 1-1 and the low significance ensures that this result cannot be produced randomly. The respective outcomes for the second semester of 2017 are presented in Figure 4 and lead to correlation coefficient of 0.816 and to significance of about 10-30. The slightly lower correlation demonstrated can be fully justified by the fact that the show does not broadcast during the summer and thus there is lower activity both on Twitter, as well as on Google, resulting in lower correlation results. Nevertheless, the findings indicate a strong relation between Twitter and Google Trends data. The aforementioned results confirm what the authors originally expected: Data obtained from Google Trends and Twitter at the same period and subject are strongly (linearly) correlated and this of course can be further exploited in a variety of research purposes.

VI. CONCLUSIONS & FUTURE PLANS

This paper presents a review of existing research initiatives that exploit online user activity data across social media to extract information linked to audience statistics and TV ratings. Among the core findings of this analysis is that existing mechanisms can mainly act as a complement approach to the traditional TV ratings, but are not able yet to fully substitute them. However, there is an imperative need to further investigate such mechanisms, as traditional audience metering approaches are demonstrating various limitations, especially due to the change of viewing behaviours such as audience's mobility and nonlinear viewing.

Most of the investigated approaches employ common data analysis steps that have been studied herewith, along with the prevailing statistical and data mining methods utilised. Building on these and considering the online social network data collection and analytics trends and approaches, this paper proposed a five-stage methodology that enables any interested party to build an efficient audience analytics mechanism based on social-media data. Based on the defined methodology and building on the architectures proposed by the most effective state of the art related initiatives, a reference architecture in support of social-media based audience analytics extraction is introduced and elaborated upon. Finally, this paper identifies Google Trends as a valuable source of information that, to the best of the authors' knowledge, has so far not been investigated by any of the existing approaches on this topic. Future plans include further evaluation of the proposed methodology and architecture by extracting qualitative and quantitative audiences' characteristics and metrics in various settings, and cross validating these with ground truth data collected with the traditional audience rating measurement approaches. Moreover, the authors have already kicked off an extended evaluation of the usability of Google Trends in the

 ISBN: 978-1-61208-681-1

application domain of TV show audience analytics based on several shows and couple the respective data with other SM data to improve the quality and accuracy of the estimated and predicted TV ratings.

ACKNOWLEDGMENT

This work has been supported by the European Commission, Horizon 2020 Framework Program for research and innovation under grant agreement no. 65020601.

REFERENCES

[1] S. Sereday and J. Cui, "Using machine learning to predict future tv ratings", Data Science, Nielsen, Vol. 1, No. 3, pp. 3-12, Feb. 2017.

[2] S. Wakamiya, R. Lee, and K. Sumiya, "Towards better TV viewing rates: exploiting crowd's media life logs over twitter for TV rating", 5th ACM Int. Conf. on ubiquitous information management and communication, pp. 412-421, Feb. 2011.

[3] F. Ahmed, R. Asif, S. Hina, and M. Muzammil, "Financial Market Prediction using Google Trends", Int. Journal of Advanced Computer Science and Applications, Vol. 8, No.7, pp. 388-391, July 2017.

[4] N. Askitas and K.F. Zimmermann, "Google econometrics and unemployment forecasting", Applied Economics Quarterly, Vol. 55, No. 2, pp. 107-120, Apr. 2009.

[5] B. O'Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith, "From tweets to polls: linking text sentiment to public opinion time series", 4th AAAI Int. Conf. on Weblogs and Social Media (ICWSM 2010), pp. 122–129, May 2010.

[6] A. Tumasjan, T. Sprenger, P.G. Sandner, and I.M. Welpe, "Predicting elections with twitter: what 140 characters reveal about political sentiment", 4th AAAI Int. Conf. on Weblogs and Social Media (ICWSM 2010), pp. 178–185, May 2010.

[7] A. J. Ocampo, R. Chunara, and J. S. Brownstein, "Using search queries for malaria surveillance, Thailand", Malaria Journal, Vol. 12, pp. 390-396, Nov. 2013.

[8] S. Yang, et al., "Using electronic health records and Internet search information for accurate influenza forecasting", BMC Infectious Diseases (BMC series), Vol. 17, pp. 332-341, May 2017.

[9] S. Sinha, C. Dyer, K. Gimpel, and N.A. Smith, "Predicting the NFL Using Twitter", Machine Learning and Data Mining for Sports Analytics Workshop (ECML/PKDD 2013), pp. 137-147, Sep. 2013.

[10] A. Chauhan, K. Kummamuru, and D. Toshniwal, "Prediction of places of visit using tweets", Knowledge and Information Systems Journal, Vol. 50, No. 1, pp. 145-166 , Jan. 2017.

[11] O. Giannakopoulos, N. Kalatzis, I. Roussaki, and S. Papavassiliou, "Gender Recognition Based on Social Networks for Multimedia Production", 13th IEEE Image, Video, and Multidimensional Signal Processing Workshop (IVMSP 2018), IEEE Press, Jun. 2018, pp. 1-5, doi: 10.1109/IVMSPW.2018.8448788

[12] M.X. Hoang , X. Dang , X. Wu , Z. Yan, and A.K. Singh, "GPOP: Scalable Group-level Popularity Prediction for Online Content in Social Networks", 26th Int. Conf. on World Wide Web, pp. 725-733, Apr. 2017.

[13] A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke, "Predicting IMDB movie ratings using social media", 34th European Conf. on Advances in Information Retrieval (ECIR 2012), pp. 503-507, Apr. 2012.

[14] B. Bhattacharjee, A. Sridhar, and A. Dutta, "Identifying the causal relationship between social media content of a Bollywood movie and its box-office success-a text mining approach", Int. Journal of Business Information Systems, Vol. 24, No. 3, pp. 344-368, 2017.

[15] G. Mitsis, N. Kalatzis, I. Roussaki, E. Tsiropoulou, S. Papavassiliou, and S. Tonoli, "Social Media Analytics in Support of Documentary Production", 10th International Conference on Creative Content Technologies (CONTENT 2018) IARIA, Feb. 2018, pp. 7-13, ISSN: 2308-4162, ISBN: 978-1-61208-611-8

[16] S. Wakamiya, R. Lee, and K. Sumiya, "Crowd-Powered TV Viewing Rates: Measuring Relevancy between Tweets and TV Programs", Int. Conf. on Database Systems for Advanced Applications (DASFAA 2011), pp. 390-401, Apr. 2011.

[17] W. Hsieh, Y. Cheng, S.T. Chou, and C. Wu, "Predicting tv audience rating with social media", Workshop on Natural Language Processing for Social Media (SocialNLP) under 6th Int. Joint Conf. on Natural Language Processing (IJCNLP 2013), pp. 1–5, Oct. 2013.

[18] F. Giglietto, "Exploring correlations between TV viewership and twitter conversations in Italian political talk shows", Aug. 2013, doi: 10.2139/ssrn.2306512.

[19] L. Molteni and J. Ponce De Leon, "Forecasting with twitter data: an application to Usa Tv series audience", Int. Journal of Design & Nature and Ecodynamics, Vol. 11, No. 3, pp. 220–229, Jul. 2016.

[20] A. Crisci, et. al, "Predicting TV programme audience by using twitter based metrics", Multimedia Tools and Applications Journal, Vol. 77, No. 10, pp. 12203–12232, May 2018.

[21] B. Sommerdijk, E. Sanders, and A. van den Bosch, "Can Tweets Predict TV Ratings?", 10th Int. Conf. on Language Resources and Evaluation (LREC 2016), pp. 2965-1970, May 2016.

[22] C. Oh, S. Sasser, and S. Almahmoud, "Social Media Analytics Framework: The Case of Twitter and Super Bowl Ads", Journal of Information Technology Management, Journal of Information Technology Management, Vol. 26, No. 1, pp. 1-18, Jan. 2015.

[23] M.H. Cheng, Y.C. Wu, and M.C. Chen, "Television Meets Facebook: The Correlation between TV Ratings and Social Media", American Journal of Industrial and Business Management, Vol. 6, No. 3, pp. 282-290, Mar. 2016.

[24] W. Fan and M. Gordon, "The Power of Social Media Analytics", Communications of the ACM, Vol. 57, No.6, pp.74-81, June 2014.

[25] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", Int. Journal of Information Management, Vol. 35, No. 2, pp. 137-144, Apr. 2015.

[26] X. Wu, X. Zhu, G.Q. Wu, and W. Ding, "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, pp. 97-107, Jan. 2014.

[27] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics", Journal of Parallel and Distributed Computing, Vol. 74, No. 7, pp. 2561-2573, July 2014.

[28] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", Int. Journal of Information Management, Vol. 35, No. 2, pp. 137-144, Apr. 2015.

[29] X. Wu, X. Zhu, G.Q. Wu, and W. Ding, "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, pp. 97-107, Jan. 2014.

[30] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics", Journal of Parallel and Distributed Computing, Vol. 74, No. 7, pp. 2561-2573, July 2014.

[31] M.N. Injadat, F. Salo, and A.B. Nassif, "Data mining techniques in social media: A survey", Neurocomputing, Vol. 214, pp. 654–670, Nov. 2016.

[32] J. Min, Q. Zang, and Y. Liu, "The influence of social media engagement on TV program ratings", 2015 Systems & Information Engineering Design Symposium, pp. 283-288, Apr. 2015.

[33] http://scikit-learn.org/ [accessed July 2018]

[34] https://trends.google.com/ [accessed July 2018]