Soft Comput (2018) 22:2731–2752
https://doi.org/10.1007/s00500-017-2530-x

METHODOLOGIES AND APPLICATION

Adaptive contents for interactive TV guided by machine learning based on predictive sentiment analysis of data

Victor M. Mondragon[1,2] · Vicente García-Díaz[2] · Carlos Porcel[3] · Rubén González Crespo[4]

Published online: 7 March 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract This paper describes a new proposal for interactive television which is an answer to a continuous change in the traditional television as consequence of the inclusion and evolution of the digital social networks, the Internet and the different elements of the digital age. The digital evolution has encourage the interaction of the viewers with the content and also increases the need to evolved the content, the methods, formats, tools and architectures to adapt the content to the sentiment expressed by the viewer while watching a show. The present paper contains the following objectives: The first objective is to create guidelines that can be used to construct adaptive contents for television, which can be modified in real time by the production team or the director of the show. The second objective is to develop applications that allows to obtain, collect and analyze the sentiment inside of the expressions, data or opinions of the viewers, who interact with the show through social networks or communication channels as: Facebook, Twitter, Instagram and WhatsApp. The third objective is to develop a machine learning to predict the preferences of the viewers, generating options and changes in the sequence of the scenes of the TV show that will be broadcasted in real time. All the objectives explained above are applied to two TV shows which are different in the content but share the live condition. During the broadcasting of the show, the guidelines are applied, the results are obtained, analyzed and the final result is more participation of the viewers and a better perception of the content. As a result of the research and the application in real life of the proposal, this paper contributes with an alternative solution for interactive TV where a viewer can interact with the show and the production team can modify the content according to what the viewers express and expect to watch based on an analysis of sentiment of data using a machine learning.

Keywords Sentiment analysis · Adaptive content · Television interactive · Machine learning · Modeling predictive · Real time · Plebiscite

Communicated by V. Loia.

✉ Rubén González Crespo
rubenagc@gmail.com

1 National Historical Memory Center, Bogotá, Colombia

2 Department of Computer Science, University of Oviedo, Oviedo, Spain

3 Department of Computer Science, University of Jaén, Jaén, Spain

4 School of Engineering, Universidad Internacional de La Rioja (UNIR), Madrid, Spain

1 Introduction

With the massive usage of mass media, the inclusion of social networks and the transformation of the traditional television to become digital and an appropriate environment is set to create new scenarios that will transform the traditional one way television toward a massive interactive and bidirectional television (Collazos Ordóñez and Mondragón 2008), which means, the new television will allow the active participation of the viewer. This transformation is achieved by integrating the communication channels (feedback channel) to the content of the TV show where the viewers interact through expressions, symbols and opinions, reflecting the emotion and sentiment felt by the viewer. The description given before is part of the definition of interactive television iTV (Collazos Ordóñez and Mondragón 2008), which in its digital evolution has allowed new models, architectures and original formats of audiovisuals contents, characterized by the inclusion of interactive functions.

The way people watch television has changed, the evolution of the digital era is bringing new ways and channels for the viewers to access the content. The mass use of Tablets and smartphones together with the widespread of Internet have caused a change where the TV set is not the main channel for entertainment anymore. This digital change has also made the content evolve toward an adaptive format where it adapts to the viewer's preferences and so changing the viewer's customs (Neira 2016). The digital evolution also promotes the generation of different platforms where the content is consumed by several ways, and the audience has personalized characteristics and take decisions according to what to watch. The metamorphosis in digital television also promotes the expansion of the advertising market to other spaces like: Facebook, Twitter, Instagram, WhatsApp, Netflix and so on. Since the Television is evolving, the TV set as main means of interaction needs also to evolve and about a decade ago the Smart TVs started to appear allowing the interaction between the content and the digital channels supported by the Internet. However, the expectations of the evolution of the Smart TVs was overshadow by the mobile devices, since the new generations rather to be informed and interact in the Social Networks by other means different to the TV sets (Neira 2016). According to the description above, now there is a need to develop new contents, methods, strategies and tools that allow a half-way point between the broadcasting of the content in real time (broadcasting live) and the interaction of the audience through the digital channels, what leads to the research objective of this paper, which can be sum up in the next question: ¿How to develop content for live TV that allows the participation of the viewers over the content broadcasted?

Supported by the analysis describe above, there are activities of research that have been done, as in the project: "Live Staging of Media Events (LIVE) (Jiang et al. 2011)", which is financed by the program "Marco Europeo". In the research, activities the viewers interact with a live TV show and allow the production team to change the broadcasting of the show according to the criteria of the production team, improving the quality of the content transmitted.

The results obtained by Jiang et al. (2011), where in the transmission of the 2008 Olympic Games by ORF, where the changes in the content cannot be done in the competitions themselves but over the comments, commercials and other elements that surround the competition and the athlete, meaning the scene does not suffer any modification by the perception of the viewer and the interaction of the viewers with the show is not reflected in the actions/script of the hosts of the show.

The work described in this paper contributes with new criteria to apply to the adaptation of the content of a conventional live TV show; allowing the viewers to interact in real time and deliver results of perception and sentiment to the production team so that the sequence of emission of the content changes according to the results of interaction.

One of the innovations shown in this paper is the inclusion of social networks as Facebook, Instagram, Twitter or WhatsApp (González et al. 2016), all of them used to obtain in real time the interactive information of the viewers. The work done by Li et al. (2014) takes the information just using Twitter.

Another contribution of this work is the inclusion of the MLA with learning intelligence that allows to threat the messages in a regional context, as well as to measure the sentiment inside the messages of the viewer to predict and propose the sequence of the scenes in the content.

During the development of the present proposal the concepts, techniques and methodologies used in the implementation of the different components that integrate the model of interactive television supported by adaptive contents. A relevant contribution of the present proposal is the realization of the experimental tests over two (2) TV shows of different topics, but both shows are transmitted life. The results of the tests with the two (2) shows are taken to support and prove the hypothesis consider in this research.

The rest of this paper is structured as follows: First, it presents an analysis of the state of art of similar works, making an emphasis in the concepts and techniques used by other authors, analyzing also advantages and disadvantages. Next, describes the architecture of interactive television propose as result of the definition of adaptive content, as well as the techniques used in this research to obtained the messages of the viewers through social networks (Facebook, Twitter, Instagram and WhatsApp). Afterward, it analyzes the way to treat the messages obtained to apply the techniques of measure of sentiment of the data, to later describe the design and training of the MLA. At the end, two test cases with results are shown as examples of the application of the concepts and techniques of the research.

2 Background

The work done by Jiang et al. (2011), introduces a new concept of LIVE TV in broadcasting that includes the interaction between the viewer and the production team in real time, at the same time that modifies the traditional concept of "TV content director" to "TV conductor", whose function is to regulate the content transmitted based on the feedback of the viewers which is analyzed by a behavior analyzer. One of the weaknesses of the work done by Jiang et al. (2011) is the limitation of the channel used to interact with the viewers, because it is an application "Real-time notification and feedback channel" (Jiang et al. 2011), implemented in the Set Top Box (STB) that transmits the viewers' actions to the TV production team. The application in the STB is a limitation because it needs to be installed and configured in the device (Collazos Ordóñez and Mondragón 2008). Also, the expansion of the Digital Terrestrial Television (DTT) with the inclusion of the applications of the digital TV limits the participation of the viewers, since a TV receptor supporting the manufacturer framework is needed (Shih-Hsuan et al. 2016). In the other hand, the work done by Jiang et al. (2011) limits the participation of the viewers that interact in social networks or massive communication channels, as described by Cesar and Geerts (2011); in this case, social networks are used as encounter point to hold debates and discussions, but no analysis of sentiment is done, narrowing the criteria of the conductor of content and production team when making decisions over the content. What is different inside this paper and the research done by its authors are the activities done get the opinion of the viewers through Facebook, Instagram, Twitter and the communication channel WhatsApp. So, the messages inside the text, icons of emotion or emotive characters are consolidated and normalized to be analyzed in the context of the sentiment of the viewers. Another disadvantage in the work done by Jiang et al. (2011) is that the content broadcasted, and the collection of the data of the viewers is oriented to a particular content (sports events), what limits the analysis of other types of TV shows or different formats of shows.

One contribution of Jiang et al. (2011) is the statistical method applied to the collected data; although no sentiments analysis is done, it gets an statistical approximation of the preferences of the group of viewers which allows to characterize the preferences over the broadcasted content, to later deliver the preferences of the viewers, as commercials, comments and other elements that go around the television content transmitted.

Another work analyzed is one developed by Cesar and Geerts (2011), who described the characteristics of the past, present and future of Social TV, analyzing the technological transformation of conventional TV toward digital TV and the development of interactive TV. Pablo Cesar and David Geerts make a description of services and platforms in development at that time, which allowed interactive activities in Social EPG's. Another contribution of this work is the inclusion of social networks as Facebook and Twitter as integrated applications to platforms and frameworks of GLOBAL TV and SMART TV. However, it is also a technological limitation linked to the STB, since as in Jiang et al. (2011), it also limits the interaction with users that do have a conventional receptor or do not support the installation of applications. Another important inclusion is the concept of "Social TV program formats", where Cesar and Geerts (2011) define the concept of interactivity using a social network; however, the lack of definition of formats do not allow the measure of the sentiment of the data, in consequence the interaction does not influence nor change the dynamic of the emission of the show, which means the content is not modified in real time as result of the expectations of the viewers.

As consequence of the above, there is a need to obtain the data from the interaction of the viewers; for this, the work done by Del Val et al. (2016) is taken as reference, where they describe a multi-agent system that automates the process to collect the data of the activity of the users in social networks (developed for Twitter) and makes a deep analysis over the social behavior in different levels of granularity in online events based on network theory metrics. The methods and structure used in the development of multi-agents works as reference in the construction of applications to obtained events coming from the interaction of the users, to later select and analyze messages and symbolic characters, which is one of the objectives of this research.

In the other hand, the growing use of social networks as interaction and feedback channel of the broadcasted content has become a significant source of information, not just to get information of the topic of the content but also to obtain findings about groups or profiles of users in social networks as Kianian et al. (2016) and analyze the groups that participate in social networks by applying Fuzzy semantic techniques to find superpositions of different communities in social networks. This process will allow not just to explore the analysis of sentiment of the users in a context of community but also to explore the characterization of a community in a social network domain.

Different authors have research about text analysis in the context of big data, one of the authors is Moreno A....who consider the analysis of text as the next process in the analysis of Big Data, and it is defined as a group of subprocesses that includes: extraction of information, entity or characters recognition, representation of the domain of analysis, among others. Moreno and Teófilo (2016) also highlight the use of techniques with supervised machine learning, which is a referent for the present research, since the nature of the information and the scalability of the machine learning are similar. In the other side, the work executed by Bernabé-Moreno et al. (2015) has developed a system for acquisition and loyalty of customers supported by information coming from Social Networks (CARESOME), where an analysis that quantifies the impact of the different interactions of every member in a period of time is done through different metrics and variables (intrinsic and extrinsic). The work just described is a reference in the adaptation of a model in the measure of impact of a set of viewers that interact with a TV show through the social networks.

As described previously, the present work intends to unify and normalized the messages of the viewers, so that they can be analyzed in the context of the sentiment of the data. To do the unification, the authors follow the work done by Miller (2005, 2014) and take the techniques to process and analyzed the text. Nevertheless, some weaknesses are found inside the lexical analyzer dictionary, since it was developed for English uses, according to Hu and Liu (2004). To be used in Spanish, the Lexical analyzer dictionary is implemented, including key words (Stop Words) classified as idioms in the regional context of the viewers and which help understand the context of the messages. Another work taken as reference is the one describe by Baldominos Gómez et al. (2015), where there is an evolutionary algorithm to classify and analyze the sentiment of a text which is made more efficient by using the technique "Stemming Words" to analyze the flow of words. In the other hand, the evolutionary algorithm works as reference for the required scalability needed to identify the groups of words that belong to a region (idioms) and be classified as positive or negative words.

The work done by Serrano et al. (2015) analyzes and defines terms related to the analysis of sentiment and compares the techniques in the implementation of the Automatic machine learning. This work is taken as reference for the present research in the implementation of semi-supervised machine learning. Another concept of reference are the supervised learning techniques based on lexical analysis, which used terms of the analysis of sentiment defined in advance to incorporate them in the machine learning as more tests are run. Both concepts are complemented by the Corpus-based technique that uses dictionaries in a specific domain. *These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either statistical or semantic techniques* (Serrano et al. 2015).

For the design of the machine learning and the implementation of the learning algorithm, an analysis of different algorithms of learning is done, so that the most appropriate is chosen. The work done by Settouti et al. (2016) take ten algorithms used in Data Mining, compare them and classify them according to statistical tests. Taking the research done by Settouti et al. (2016), it is concluded that the classification of an algorithm with respect to another requires an statistical analysis that demonstrate the difference between the analyzed models. The conclusion is taken as reference to do the tests of the different algorithms and techniques of measurement of sentiment exposed by Miller (2014), where are included: difference, regression difference, Word/item analysis, logistic regression, support vector machines and random forests. When comparing the results of Settouti et al. (2016), Miller (2014), the result demonstrates that the Learning Algorithm: Support Vector Machines (SVM) is an effective technique for the text classification in the machine learning context. The technique Support Vector Machines is improved and complemented with the Random Forest technique by using thousands of classifiers over estructured trees that finish in just one prediction.

Additionally to the references consulted previously, in the development of the machine learning (ML), it is necessary to explore different techniques that evaluate, filter and generate predictions or recommendations to the production team, based on the obtained information from the interaction of the audience. Works like the ones done by Tejeda-Lorente et al. (2014) use the concept fuzzy linguistic and product quality to generate recommendations that inside this research can be applied to the quality of the content the viewers perceives.

Another need is to evaluate the level of trust in the results of the machine learning (recommendations), and to do so, several techniques are available; however, the work done by Majd and Balakrishnan (2017) is chosen as reference for the development of the model and tests of the machine learning of the present paper because it allows to evaluate the level of trust in the following variables: reliability, similarity, satisfaction and transitive trust.

3 Proposed interactive TV architecture

The digital interactive television (iTV) is born because of the convergence of the traditional television and the different channels of communication available for the final users to interact with the TV shows. One of these communication channels are social networks, widely accessed through mobile devices (Wages et al. 2006) and which are influenced by conventional television (broadcasting, iptv); influence clearly seen through comments and posts that show and give feedback about what has been broadcasted (Cesar and Geerts 2011), as in the definition of ITV (Collazos Ordóñez and Mondragón 2008).

Applying the concept of iTV and taking advantage of the amount of data (Big Data) obtained in the interaction with the viewers through the social networks, these work includes new components to the architectures previously defined in Collazos Ordóñez and Mondragón (2008), Jiang et al. (2011) and Cesar and Geerts (2011). The new components have the following functions: component to obtain the data from the different social networks and communication channels. Component to treat and adequate the data. Component to learn and analyze the sentiment of the data, Fig. 1 describes the new architecture result of the fusion of the architecture of iTV in Collazos Ordóñez and Mondragón (2008), Jiang et al. (2011), Cesar and Geerts (2011) and these new components. The flow goes as follow:

1. The content is broadcasted.
2. The viewer watches the show and posts on social networks.
3. The opinions and expressions of the users are obtained by means of the applications APIs (Ravindran and Garg 2015), then;
4. These data are normalized in text files which will be later

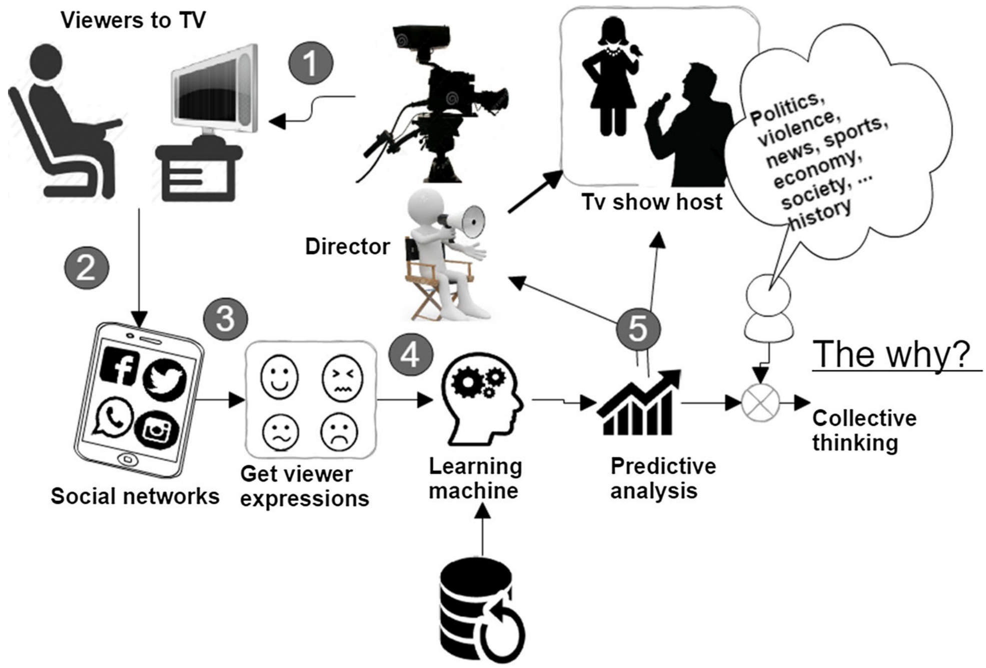


Fig. 1 ITV architecture

5. Consolidated and analyzed by a MLA which measures the sentiments inside the data given by the viewer. The MLA, named ML-ITV (Machine Learning-Interactive Television), also allows predicting the preferences of the viewers over future broadcasted content applying the concept adaptive content (Mondragon Maca et al. 2016).

The contribution of the work described in this article is to correlate the result of the analysis of sentiments done by the ML-ITV with variables of influence in the historic or present daily thoughts of the viewers to obtain information about trends of collective thinking that can be applied for future broadcasting (Russell 2013).

3.1 Adaptive content

As defined and described by Mondragon Maca et al. (2016), in the new format of adaptive content applied to a live TV show ("Login"), the director moves to a new role "TV content conductor" (Jiang et al. 2011). This TV content conductor acts in function of constant real-time interaction with the viewers, who allow the director to change the content in real time by guiding the actors or hosts according to the feedback received, meaning at the end that the public watch what expect.

The guidelines to construct an adaptive content are generated as result of experimental tests over different live TV shows, for example: News, talk shows, musical video and technological learning programs. All the TV shows experimented interaction with the viewers through social networks (Facebook, Twitter, Instagram and WhatsApp). Figure 2 shows the proposed new general format of design of adaptive content for different Live TV show. As result of the application of experimental guidelines in the construction of adaptive content, a new format is born, that allows to add, eliminate and modify the scenes in the script. This novel format contributes to the production team with the possibility to respond in real time to the changes identified in the analysis of sentiment of the data. Although currently there are a big number of shows that used Twitter as interaction channel with the viewer, the changes in the emission of the content are run by the TV host(s) based on improvising formats previously defined; for this new format, a previous stage in the production of the program is needed to design alternative content, scenes and scripts to be broadcasted according to the changes of adaptation of the show. It is important to mention that this new format requests more resources and production costs to have the alternative content ready. The new format was developed as result of many experimental TV shows broadcasted, where the production team gathered scripts and changes done in the first emissions of the TV shows, finding several similarities and common actions in the sequence of the scenes. However, it is concluded that the scenes of the initial script and the alternative scripts will depend on the topic and other characteristics of the show to broadcast.

With the proposed format, the production, postproduction and direction team could predict and design several options

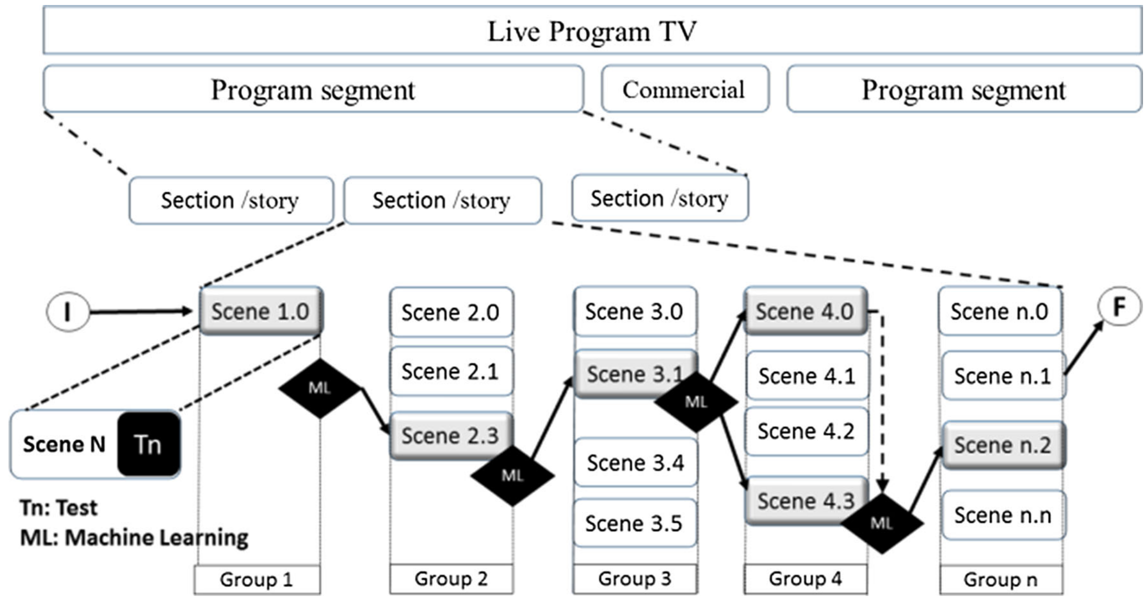


Fig. 2 Adaptive content structure

of contents that adapt to the viewers trends, following the viewers' sentiments.

Inside this article, besides the proposed format in Fig. 2, there is a new introduced concept by incorporating the analysis of a MLA to the sequence of adaptive scenes. Another innovative change from the proposed model (Mondragon Maca et al. 2016) is the possibility to select and combined different scenes from the same group (Group n)

The incorporation of the adaptive content described in Fig. 2 starts with the segmentation of the content of a TV show into smaller segments (Lian 2012). The TV show is divided in sections, the sections are divided then in scenes, permitting to create an adaptive script and sequences of scenes that take different routes during the TV show time line.

In a traditional script, the route of sequences that are part of a section would be: $R\{I : F\} = \{\text{Scene } 1 \rightarrow \text{ Scene } 2 \rightarrow \ldots \rightarrow \text{ Scene n}\}$, where R represents the route of sequences of the TV content.

Inside of the adaptive system, a section is conformed by a sequence in time of groups of scenes $\{G1 \rightarrow G2 \rightarrow G3 \ldots \rightarrow Gn\}$, and each group contains different scenes {Gi} = {Scene i.1, Scene i.2 ... Scene i.n}. The group of scenes follows the same pattern of topic, and the route of sequence R {I: F} is conformed by one or more scenes {Scene i,n} of each group Gi. To choose an scene {Scene i,n}, the TV content conductor and the production team consult the results of the analysis done by the machine learning "ML".

To deliver the analysis, the ML takes the viewers' answer, previously induced by the Test during the emission of the show {Scene X:T} (see Fig. 2). The Test {Tn} is elaborated taking into account the context of the scene and becomes the guideline to construct the adaptive content. The objective of the test is to induce the viewer to react and interact with the show through posting on social networks using special symbols and expressions.

Besides the analysis, the ML also works as a learner which will recognize and incorporate expressions coming from the viewers to measure and analyze the sentiments inside the data. The ML also must deliver prioritized options to the director and production team to allow them to take decisions in less time and construct the route of scenes R {I: F} based on reliable information coming from the audience. To make the ML obtain a value of prediction based on the events (grade, click, etc) previously analyzed, it is necessary to implement a correlation between the broadcasted TV video and the information of the video through a function of diffused neuronal network and learn the states of the sentiment (positive, negative). A similar work was developed by Duong et al. (2016), which uses a diffused neuronal network to filter the recommendation of video from a set of videos. In this case, the diffused neuronal network is used to learn the scoring of the users regarding the users' behavior.

4 Data extraction from social networks

Taking as reference the proposal presented in the work of Moreno and Teófilo (2016), it is required the extraction of the information to identify words, phrases and idioms and relations inside of the text to obtained. The functionality in the extraction of the information is based on the coincidence of patterns named regular expressions (Moreno and Teófilo 2016), which is taken as start point for the extraction of the information of the viewers in the present research.

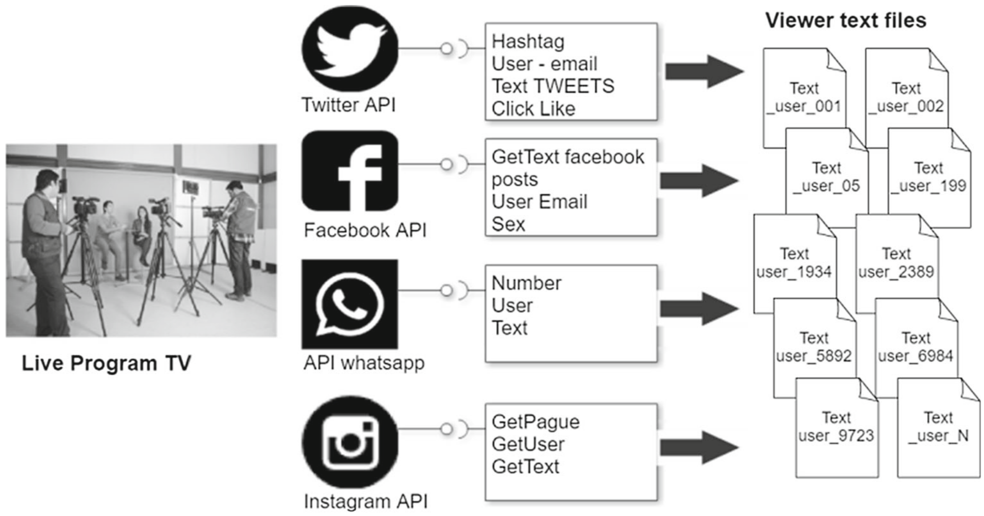


Fig. 3 Get messages from viewers on social networks

To obtain the messages, symbols and expression of each viewer that interact with the show through the social networks, an API (Application Programming Interface) is used. Every social network (Facebook, Twitter and Instagram) or communication channel (WhatsApp) has a different API. The functions and methods used for each social network or communication channel depend on the environment of development and the technology of the API.

Figure 3 shows the high level components used to obtain and consolidate the messages coming from the viewers. For each social network, there is an active account which is used to connect and get the information. During the broadcasting of the show, the hosts prompt the viewers to post on social networks using different expressions, symbols and messages. The applications developed by the authors and described inside this paper convert the viewers' posts into text saved in text files ((*.txt). For WhatsApp, the application takes the messages received in a given phone number, normalized them and store them in a file.

The API used to get the messages from social networks was developed to obtain other information of the user as: location, age, user name, name, e-mail among others. It is important to mention the collected information follows the politics and restrictions of privacy of each social network as configured in the account by the viewer; actually, this represented the main difficulty during the experimental tests, since some variables were not available.

To mitigate the restrictions in the variables and information, the social network could not deliver, during the tests Test {Tn}(see Fig. 2); the hosts in the shows followed some scripts and induced the viewers to send/post this information (age, location, preferences, name, profession, etc.) in exchange of prizes, tickets, gifts from sponsors, etc.

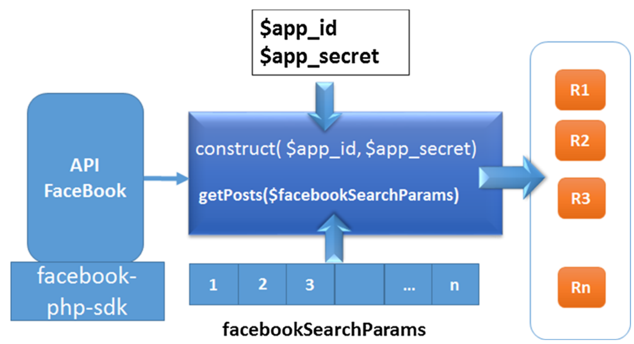


Fig. 4 Get messages using Facebook API

The development of APIs and data treatment for each social network is described next.

4.1 Get viewer messages on Facebook

For the views expressed by text messages posted by viewers on "fanpage" Facebook, initially you access data Facebook user account (user television content, such as Facebook "login"). Graph API is used for published data, combined with the Facebook Query Language (FQL), Graph API to be a graphical data structure that represents the node-based social interactions (Russell 2013), giving the tree interactions from viewers. The client application obtains the data and developed in Python and PHP, Fig. 4 shows the main components:

Access through "OAuth token" using the safety parameters and *app_secret*, *app_id* (in this work belong to the fanpage "login" program). The consultation implemented through the "Node ID" on which calls the search method Open Graph API "getPost()" returning the list of post that match the criteria.

A problem in building the application to get messages from Facebook is the limitation of TOKEN by the API, i.e., when you make use of the Facebook API, you automatically get a TOKEN that expires in 1–2 hours. However, to solve this error, Facebook will test us an option so that TOKEN does not expire. It was simply solved by obtaining the long-lasting TOKEN. The next call refers to the long-running TOKEN.

4.2 Get viewer messages on Twitter

To obtain the opinion of the viewers that post in Twitter, the work done by González et al. (2016) is taken as reference. González et al. (2016) describe a new standpoint for the analysis of sentiment inside of the ontology domain, making a semantic analysis of the obtained tweets using the processing of natural language to classify the sentiment of positive or negative type in each tweet. The concept described above (González et al. 2016) is applied to search and extract text and characters as described next:

For the opinions of viewers that are published via Twitter, an application was developed considering: Text messages on Twitter are short (maximum 140 characters), the views contain grammar and specific lexical a region and environment characterized by emotions, slang, hash tags, etc. Figure 5 shows the functional components of an application that described for the opinions or messages from viewers which shows: Twitter REST API version 1.1 that through the OAuth (Ravindran and Garg 2015) authentication allows for private consultations on followers and “hatch #” as the filter criterion in matters of opinion, getting more specific data viewer.

The function for messages posted on Twitter, with filter “hatch #” and keywords (1, 2 .., n). The result of this application implemented in Python; Rn is the vector dynamic (R1, R2... Rn) containing the name Twitter and message viewer field (opinion piece) field. To obtain regular consultations and events and complement the functionality the Streaming API that provides low latency times in the queries are used to access the data using a Push method which is recovered through the “Stream ()” (Tran et al. 2016) function and allows asynchronous data.

The streaming process gets Tweets flow in real time, which allows for short time in the data stream viewers on”hatch #” in particular, later to be analyzed in conjunction with search and filter criteria (words key). Next, the description of the developments and additional implementations over the Twitter API Rest version 1.1 is found, which through the OAuth (Ravindran and Garg 2015) authentication allows to obtain particularly consults about followers and over the “hatch #” as the criteria for the filter in the opinions. To find the viewer’s age on Twitter, the Login hosts asked the viewers to share their age by retweeting (Twitter 2016) their age to the Twitter account “@amigosdelogin” using the hatchtag : “miedades:

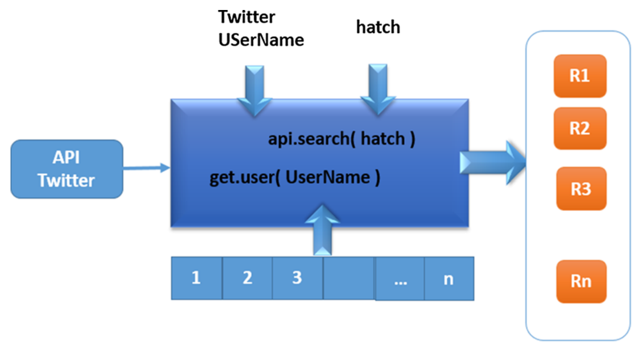


Fig. 5 Get Twitter messages using API

?[0-9]+”, which will deliver a number of one (1) or two (2) digits representing the viewers’ age.

One of the difficulties in implementing the Twitter API is the request limit. Twitter provides several API’s, including REST API, Tweepy API (an API of Twitter for Python). In the case of Tweepy API, it restricts the requests after a period of time. As solved, a dead-time process was implemented in the requests so that the script does not end session and can continue to make query requests.

4.3 Get viewer messages on WhatsApp

WhatsApp instant messaging application for mobile that is shifted the concept of traditional SMS text messaging (Karen and de Oliveira 2013). A considered percentage of viewers of “Login” program interact with the content via the WhatsApp application, which has replaced in the context of the program the old system of telephone call from viewers, as the WhatsApp an identifiable channel by its number of origin of the message (Caller ID), has become a channel of opinions.

The presenters receive text messages and emit some messages more relevant, however given the number of text messages received from different viewers, for the production of the program is not possible to analyze the time line. With the above, this paper presents a proposal as shown in Fig. 6 where the reception of messages of WhatsApp is implemented and stored in a temporary arrangement with the following storage fields:*ID - identifier* of the mobile where the message came, *SMS_WHT* - text message viewer opinion, *MHour*- Date and time of the message. The above fields are used to validate the search criteria and thematic classification of the message, according to the timeline of the program being broadcast.

4.4 Get viewer messages on Instagram

To get the feedback of the user on Instagram, the API from Instagram was called by a Python development. The block diagram and the data flow are shown in Fig. 7.

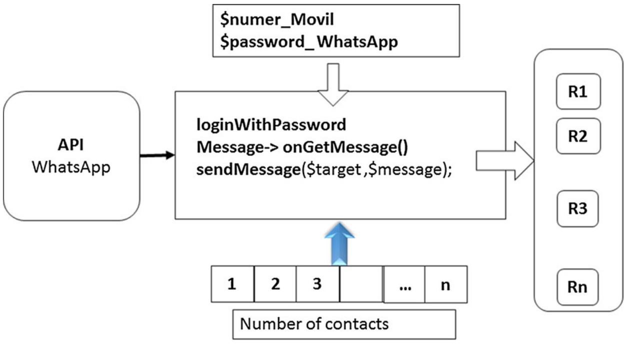


Fig. 6 Get and respond to messages from WhatsApp

Through the function "urlopen" is possible to connect to Instagram and look for the caption (hashtag) required (for instance "Plebiscito"), then using "href" all the links and the expressions are found (for example to find /p/BMMbJkRAQfX the next expression is used: (r′([″]\ w*[: ., / + -]*\w*[: ., / + -]*\ w*[: ., / + -]*\w*[: ., / + -]*\ w*[″]*\w*)′)).

To get the ID of each picture on Instagram the IDs for the regular expression required is done by using the function re.compile ((r′(\w*[B]\w*\S*\w)′)", get ID(i.e ID= BMM-bJkRAQfX) Once you have the ID list of all the pictures, to filter by caption and obtained the user's opinion (text) a Java Script JASON is executed.

A script was executed in PHP, which will load all the end_cursors, that is, all the pages and at the same time, obtaining the variable 'code' and storing it in a file. Figure 8. Script in PHP where you get all the "code" of all the pages. *Ini_set ('max_execution_time', 0)* is not to get a time limit when trying to get all the 'code'. Then, you create an infinite *while()*; at the same time you are writing and you get each of the codes of the images ('code') to the last page of the hashtag.

```
<?php
ini_set('max_execution_time', 0); //300 seconds = 5 minutes
$baseUrl = 'https://www.instagram.com/explore/tags/plebiscito/?__a=1';
$url = $baseUrl;
$array = array();
while(1) {
        $file = fopen("links.txt", "a");
        $json = json_decode(file_get_contents($url));
        $filter_array = array_filter($json->tag->media->nodes);
    // working
        foreach($filter_array as $object_return){
           echo $object_return->code.'</br>'; // will output code string
           fwrite($file, $object_return->code);
           fwrite($file,"\n");
        }
        fclose($file);
        if(!$json->tag->media->page_info->has_next_page) break;
 $url=$baseUrl.'&max_id='.$json->tag->media->page_info->end_cursor;
      }
?>
```

Fig. 8 Script get data Instagram

Running it on a local server, we can see that it prints and saves the respective codes.

5 Measurement of the sentiments in the data

The analysis of sentiments also known as analysis of opinion is maybe the most popular analysis applied to text, and the key aspect of this process is to understand the opinion of the viewer expressed inside of a message. The analysis of sentiments is the use of techniques as PNL (Mondragon Maca et al. 2016), lexical resources, linguistics and automatic learning to extract subjective information (opinions, emotions, attitude, humor) to be used later to calculate the polarity inside a text. Polarity refers to find out whether a text expresses a positive, negative or neutral sentiment. A more advance

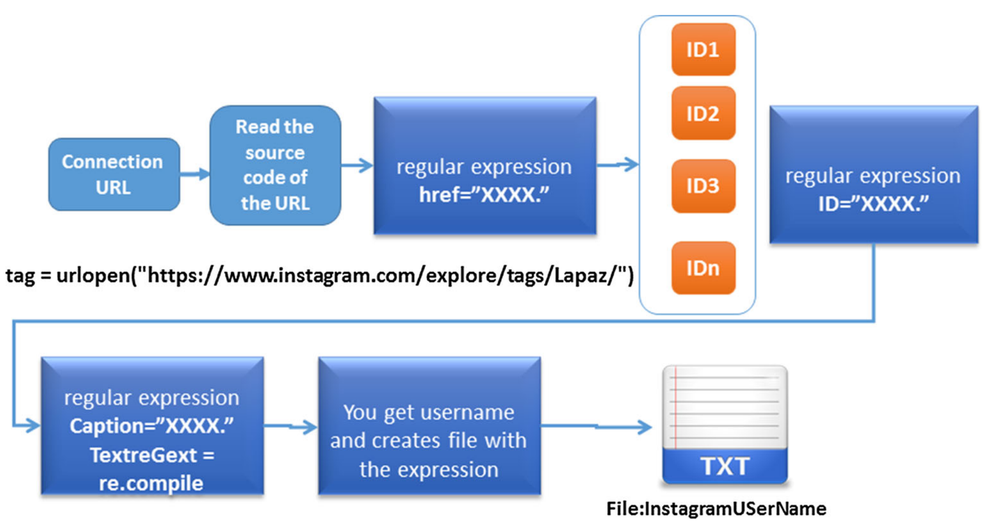


Fig. 7 User's expression on Instagram

analysis includes the detection of more complex sentiments as sadness, happiness, anger and sarcasm.

To measure polarity, it is necessary to assign values or scales defined as positive or negative where the analyzed text can be set. The applied technique in the analysis of the sentiment in this paper uses a list of words classified as POSITIVE and NEGATIVE (Miller 2014; Mertz 2013) in Spanish language, complemented with specific expressions from the region or idioms.

In general, social networks and the feedback received through them, express the beliefs, sensations, judgment, emotions and sentiments of human beings (Julian 2016), which means there is subjective context. In the other side, it is possible to find an objective context, meaning the text includes sentences without any emotion, feeling nor mood.

The analysis of sentiments works best in a subjective context and to analyze the messages, expressions and symbols saved inside the (*.*txt)* files the MLA described inside this paper takes into account the natural language of the user and puts the words into a context, for example for the experimental tests, the subjective context is the topic of the show.

The techniques used in this paper for the analysis of sentiments are based on the work of Dipanjan (Sarkar 2016): unsupervised lexicon-based and supervised machine learning, both techniques combined to implement a MLA in interactive TV {ML-ITV}. The technique unsupervised lexicon-based allows to recognize, validate and interpret regular expressions, idioms defined in each language (Hu and Liu 2004), as well as to construct a table of symbols, a lexical, syntactic and semantic structure (Mondragon Maca et al. 2016). For the tests of this paper, Spanish is used with a list of 2,006 positive and 4,783 negative native expressions.

The MLA depends on the focus of the show and the broadcasted content, for example: The show “A FONDO” (Telepacifico 2016) is a journalism and opinion program, with political in a regional context, while the show “LOGIN APP” (Login de Telepacifico 2016) (for example) is a youthful musical and technological show with idioms and youthful lexical.

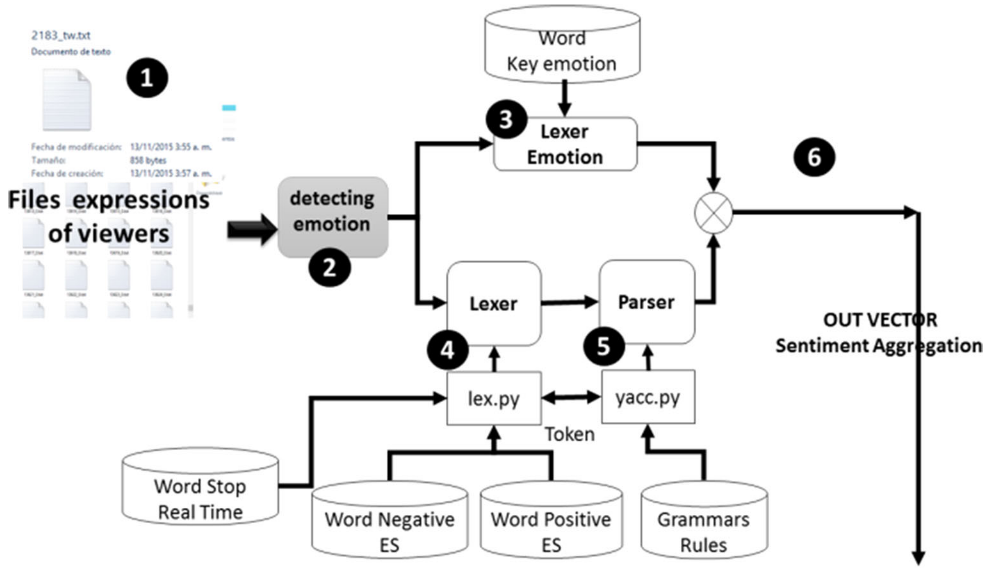
The implementation of analysis of sentiments in the data, the following definition of sentiment is used: *Sentiment = {data source, source, target, sentiment, polarity}* (Nathan Danneman 2014). Where “data source” represents the source of data that could be Twitter, Facebook, Instagram or WhatsApp. “source” represents the sentiment expression in an opinion. “target” represents the entity who the sentiment is related to. “sentiment” represents the categorization of the sentiment, which are related to frequently used expressions in the show of TV and for the analysis can be: I like it, I hate it, I love it, I am not interested in or I adore it. “polarity” defines the assessment of the sentiment as POSITIVE or NEGATIVE. In a different form, the emoticons (emoji) (Anasse et al. 2014), defined as a nonverbal expressions, are being used with a higher frequency in social networks, and its significance in the comprehension of the attitude, intension and emotion of the viewer. At the moment, Twitter supports more than 1100 emojis, Facebook supports 420 emojis and WhatsApp 236 emojis. This work incorporates the correlation of the sentiment in the text message with the emotion in the emoji-based expression to obtain a measure of the sentiment of the viewer.

It was found in the analysis of emojis that users do not grade the text by using methods like giving a number of stars, instead, there is a better way to assets the objects explicitly with a Positive or Negative Method, like explained by Nuñez, Cueva, Sanjuán, Montenegro-Marin e Infant (Nuñez-Valdez et al. 2011), so that the “Like” or “Unlike” on Facebook are taken like an appropriate way to assets the comments of users for the tests done in the present research.

Once the data, expressions and emotions are obtained and consolidated, they must be analyzed, so a system with specialized components is built. The system adequates, identifies, classifies and assets the data following the architecture shown in Fig. 9. The architecture shown in Fig. 9 joins several works (Cesar and Geerts 2011; Miller 2005; Russell 2013; Settouti et al. 2016) related to the analysis and data treatment; however, it incorporates new components that are used and needed to identify and classify none conventional expressions or idioms. Another contribution in the proposed architecture is the flexibility to add new words, phrase or idioms with the supervision of the production team of the TV show, this as part of the development of the semi-supervised machine learning. Other models such as the one suggested by Miller (2014) are implemented based on functions and libraries of the R language, the adaptation and analysis of the expressions; nevertheless, they do not incorporate the recognition of new lexical and syntactic expressions nor the emojis. The disadvantages in the work of Miller are solved in the work done in this research and shown in Fig. 9.

The Fig. 9 also contains the elements design to treat, adequate and normalize the data for the analysis of sentiment, where:

1. The messages, expressions and symbols posted by the viewers are stored in a file
2. The files are stored in two types of vectors, the first one contains text messages and grammatical expressions and the second vector stores symbols or characters that represent emotions.
3. Lexer emotion analyzes the different characters or special symbols used on Facebook, Twitter, Instagram and WhatsApp that represent a pattern of sentiment in the found emotions.
4. Lexer is a lexical analyzer developed in Python PYL (lex.py) (Sarkar 2016) which identifies the positive, negative, neutral and stop words (Anasse et al. 2014). The



	USER	API	SENTIMENT
TWEETS 3795	363_tw.txt	0	NEUTRAL
	364_tw.txt	1	POSITIVE
	365_tw.txt	-1	NEGATIVE
	366_tw.txt	-1	NEGATIVE
	371_tw.txt	0	NEUTRAL
Facebook 2571	1067_fb.txt	0	NEUTRAL
	1071_fb.txt	0	NEUTRAL
	1072_fb.txt	1	POSITIVE
	1074_fb.txt	-1	NEGATIVE
	1078_fb.txt	1	POSITIVE
Instagram 3876	2891_in.txt	-1	NEGATIVE
	2892_in.txt	-1	NEGATIVE
	3189_in.txt	0	NEUTRAL
	3190_in.txt	1	POSITIVE
	3191_in.txt	1	POSITIVE
Whatsap 172 Number Phone	4023_wh.txt	1	POSITIVE
	4024_wh.txt	-1	NEGATIVE
	4029_wh.txt	-1	NEGATIVE
	4030_wh.txt	1	POSITIVE

Fig. 9 Architecture of adaptation of data and measurement of sentiment

stop words are added in real time according to the topic of the broadcasted show, are controlled by the production team and used to describe names, songs, videos, politicians, current politics and general key words relevant to the topic of the show.

The key words and idioms found in the experiments are: "mi pana," "parce," "re-chimba," "papi," "amistad," "bororó," "háblame," "plebicito," "Expresidente," "elenos" "paramilitares," "paracos," "farc," "epl," "ELN" among other, that are described later in this paper during the explanation of the practical tests. The lexical analyzer in the context of the TV show analyzes the words above (Moreno and Teófilo 2016).

5. Lexer-Parser is a syntactic analyzer (yacc.py) in charge of the recognition of the syntactic expressions posted/sent by the viewer, which exchanges "tokens" with the lexical analyzer to recognize semantic expressions and idioms; for example in the expression "…Hoy nos acompañó la banda Cirkus Funk con toda su música es una puta chimba," the idiom words "puta" and "chimba" are Negative expressions; however, when they are combined with the STOP word "Cirkus Funk", the expression becomes POSITIVE. To achieve a better quality in the semantic level of the messages, since this kind of texts happen not to be in an structure text, it is necessary to use structure techniques of extraction, using a local grammar analysis (Ghoulam et al. 2015), that initially will work as a training schema for the machine learning, to later make an structured recognition and analysis autonomously.
6. When the result of the analysis of the emotions (lexer emotion) is correlated with the analysis of the sentiment of the expression (lexer-perser) a new vector "sentiment aggregation" is obtained with the correspondence between the two vectors.

5.1 Tests the measure of viewers sentiments

Sentiment measurement tests were carried out in different "Adaptive Test" (see III) along different emissions of television content in real time, with a total of 358 measurements. In

Table 1 Example of analysis and measurement of feeling

No	Viewer	Filetv	Channel.tv	Total words	POSITIVE	NEGATIVE	EMOTION (%)
12331	carlos345.ariza	12332_8.txt	Facebook	63	10.587	1.587	9.00
12332	j.romero23	12333_9.txt	Facebook	188	23.319	5.851	17.47
12333	@lucia_campino	12334_7.txt	Twitter	47	4.255	0	4.26
12334	@becpress	12335_9.txt	Instagram	57	1.754	3.509	−1.76
12335	geovana.hern370	12336_9.txt	Facebook	76	7.895	1.316	6.58
12336	@valennccia	12337_7.txt	Twitter	110	0.909	3.636	−2.73
12337	adriana.hpotes.7	12338_10.txt	Facebook	55	7.273	1.818	5.46
12338	@yamiles2502	12339_10.txt	Twitter	53	22.642	3.774	18.87
12339	jmera.cuesta	12340_8.txt	Facebook	105	10.476	3	7.48
12340	+57 3207190223	12341_7.txt	Whatsapp	82	6.098	1.22	4.88
12341	@mattitefaQ1	12342_8.txt	Twitter	184	3.804	2.174	1.63
12342	moureSantiago	12343_7.txt	Instagram	68	10.294	2.941	7.35
12343	@lomaselite	12344_8.txt	Twitter	61	26.557	6.557	20.00
12344	laddieAndrade	12345_10.txt	Facebook	73	12.329	1.37	10.96
12345	@foodservice1971	12346_10.txt	Twitter	56	17.857	21	−3.14
12346	Chiaracavero	12347_9.txt	Instagram	71	15.493	4.225	11.27
12347	@eltimo16	12348_9.txt	Twitter	56	7.143	5.357	1.79
12348	muequito.eelnene	12349_10.txt	Facebook	35	20	2.857	17.14
12349	@judaresey	12350_9.txt	Twitter	119	5.042	3.361	1.68
12350	Diuddy	12351_8.txt	Instagram	209	4.785	1.914	2.87
12351	@Julin39589895	12352_8.txt	Twitter	39	12.821	2.564	10.26

Table 2 Summary of descriptive variables

Summary	83 Televidentes		1783 Televidentes		16294 Televidentes	
	Comp.1	Comp.2	Comp.1	Comp.2	Comp.1	Comp.2
Standard deviation	1.0962283	0.8934672	1.0642983	0.9312728	1.0642983	0.9312728
Proportion of variance	0.6008582	0.3991418	0.5663654	0.4336346	0.5663654	0.4336346
Cumulative proportion	0.6008582	1.0000000	0.5663654	1.0000000	0.5663654	1.0000000

order to illustrate, analyses procedures and results obtained in the present study, three measurements of sentiment have been taken with the participation of 83, 1783 and 16294 viewers, who expressed their opinions on each “adaptive test” shown in Table 1. Each expression of the viewer that has been previously stored in a text file “filetv” (for example 104_tw.txt), applying the process of measuring feelings and emotions described above, gets a consolidated for the following variables: total of words, positive words, negative words, other words, positive emoji, emoji negative. By uniting each of the results of each “fileTV,” viewer gets the array of measurement of the sentiment that is stored in the object: “text.measures.data.frame”, as shown in example of Table 1.

The array of feelings obtained allows calculating the POSITIVE variable, which correspond to the ratio between the total numbers of positive words about the total number of words found in each expression (filetv), in a similar way the NEGATIVE variable. Applying the matrix of feelings basic descriptive measures of cumulative variables, obtaining: Standard deviation, Proportion of Variance and Cumulative Proportion, as it can be seen in Table 2.

The discrete variables obtained POSITIVE (Table 2 Comp.1), negative (Comp.2) undergo an analysis to find out if the negative and positive dimensions tend to be on a common scale. The table above shows a positive relationship moderate, numerically the relationship between variables NEGATIVE and POSITIVE can be quantified using the covariance and correlation, as shown in Table 2, where each measurement produces a positive covariance and the “*Cumulative Proportion*” ratio is still +1, which leads to the conclusion that the POSTIVE and NEGATIVE variables are proportional.

If the above results are plotted on a map of dispersion as shown in Fig. 10 (for measurement of 1783 viewers), you

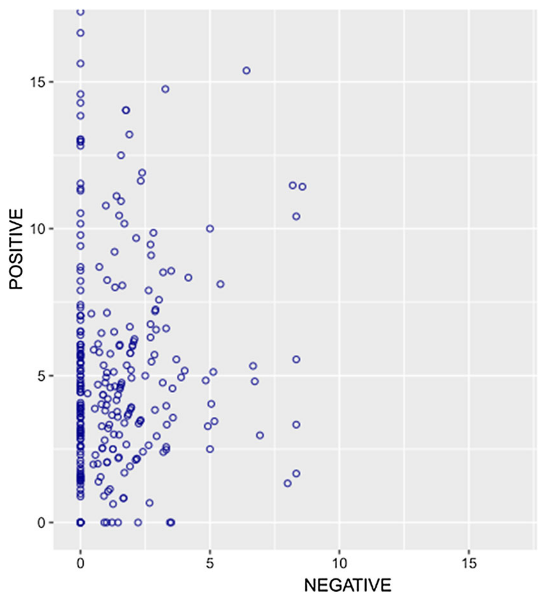


Fig. 10 Example of scatter to the extent of feeling

will see that the trend of dispersion diagram, which shows that a high number of viewers have expressed through positive words the view about the topic to evaluate the program through the "Adaptive Test" opposite a discrete number of negative words, for example: The feeling of viewers far distances has a tendency POSITIVE

Biplot (BBVA 2010) is used to illustrate a map representing together the rows and columns of the matrix of measuring sentiment in such a way that products scalar between the vectors and row column approximate as much as possible to the corresponding values in the array. Figure 11 shows two diagrams Bitplot 83 and 16294 viewers measurement.

Figure 11 shows the measurement of two-dimensional set of data sensation plane related by the variables: POSITIVE and NEGATIVE, uniform distribution in accordance with the underlying vectors pointing in opposite directions and orthogonal to each other. We conclude that the tendency of viewers is positive on the content delivered, however, evidence a relationship with a degree of inconformity, which leads to a subsequent predictive analysis on viewer preference.

6 ML–ITV predictive model

To obtain a matrix of reference, this research work takes as reference an automatic learning algorithm, as are the machine learning methods, which usually perform better than the traditional ones (linear or logistic regression methods). "*The underlying algorithms can yield thousands of formulas or nodal splits fit to the training data*" (Miller 2014). "*When some observations in the training set have coded responses and others do not, we employ a semi-supervised learning approach. The set of coded observations for the supervised component can be small relative to the set of uncoded observations for the unsupervised component*" (Liu 2011).

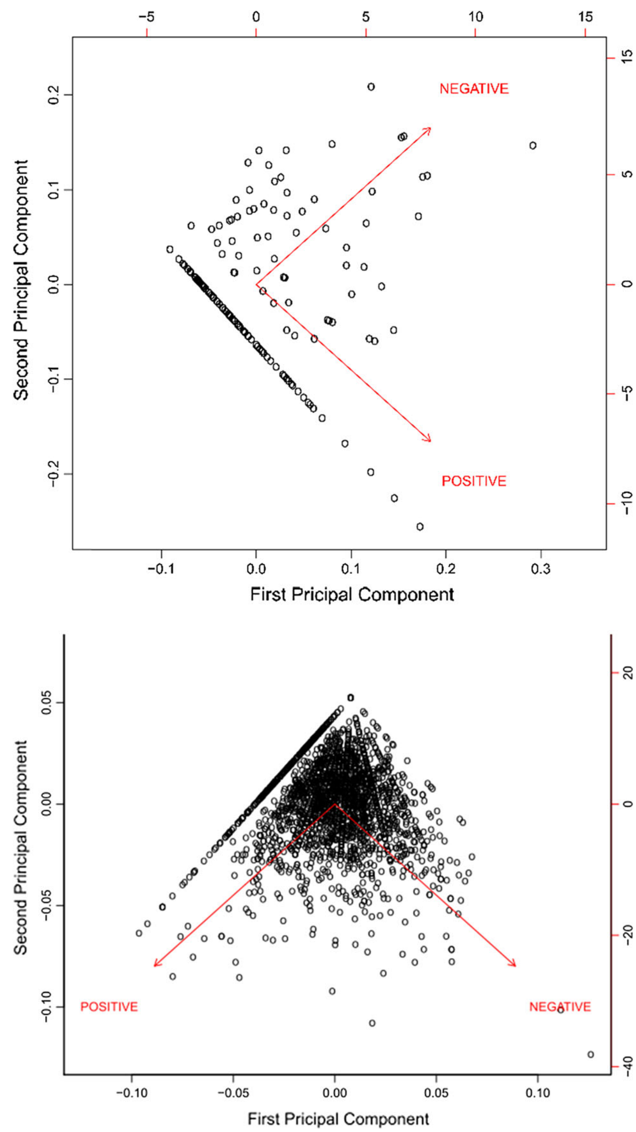


Fig. 11 Diagrams Bitplot measurement de 83 y 16294 viewers

According to the description above, a semi-supervised algorithm is applied. This algorithm consists in learning from each video broadcasted in the show, with the obtained values of feelings and emotions (V item), then these values are introduced into the matrix of the predictive learning model, where the level of feeling is stored for each viewer into the variable EMOTION.

Table 3 describes an example of the results of a second day of broadcasting of ten musical videos. In the matrix of learn-

Table 3 Learning matrix calculation of the variable: EMOTION

ID_viewer	Channel	M1 (%)	M2 (%)	…	M10 (%)	…	M20 (%)
carlos345.ariza	Facebook	2.68	−1.50		6.06		
lucia_campino	Instagram	3.43	2.13		1.59		
@becpress	Twitter	−1.33	7.08		−2.00		
jmera.cuesta	Facebook	−3.70	0.00		13.21		
+57 3207190223	WhatsApp	5.41	4.60		1.70		
@lomaselite	Twitter	5.15	1.47		5.33		
laddieAndrade	Facebook	−2.31	13.51		3.03		
@foodservice1991	Twitter	0.00	0.00		12.77		
eltimo16	Instagram	16.00	0.00		8.98		
muequito.eelnene	Facebook	6.25	2.08		7.14		

N	Music video	Music video list
1	M1	Pitbull Messin Around
2	M2	Ariana Grande IntoYou
3	M3	Martin Garrix Lions In The Wild
…	…	…
20	M10	Enrique Iglesias Duele El Corazón

ing is stored all the information for each viewer, including the participation channel.

6.1 ML–ITV experimental test for a week

Next is found an analysis of the implementation of the experimental predictive model over the answers gotten from 16339 viewers, while the video "Martin Garrix ft. Third Party - Lions In The Wild" was broadcasted (Garrix 2016). A first analysis shows a direct proportional relationship between the viewer's age and the amount of negative words, Fig. 12, and of course an inverse proportional relationship with the variable "positive.word". It was found that to an older age the NEGATIVE feelings and opinions are more common; also it is observed that the number of viewers between ages of 21 and 24 reduces.

The other variables were analyzed through an experimental prototype model, Fig. 13, which was built using the Alteryx tool (Alteryx 2016) by the modules: "Directory", "R tool", "Linear Regression", "Association Analysis" and "Nested Test" (Alteryx 2016).

The module "R Tool" is used to analyze the expressions and the result data of the processed expressions of the viewers in the social media (Facebook, Twitter, Instagram and WhatsApp), supported by the module Directory (49), which takes the files and allow the data to be adequated, extracted and processed, as described in the item III.

Later, some additional variables from Table 4 are added, so that the "R Tool" out is the data matrix of the viewers with a feeling analysis.

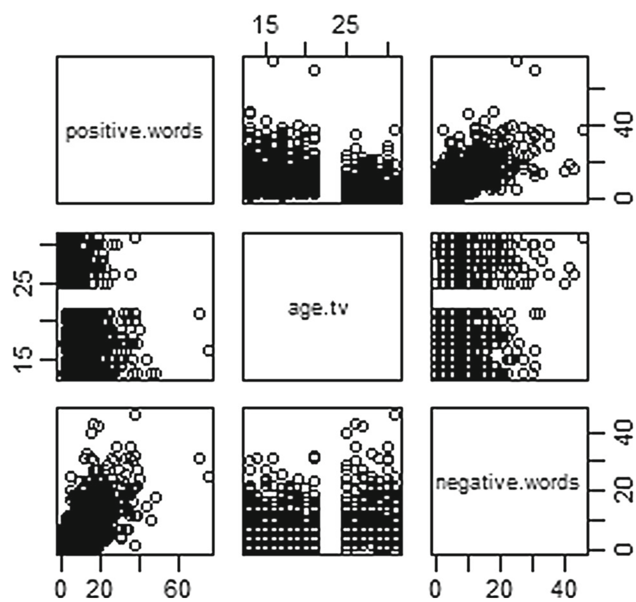


Fig. 12 Age scatterplot and negative words (in R)

Thanks to the module "Association Analysis (33)" [43] is possible to determine which fields in a database have a bivariate association with one another, in this case the variables *age.tv, negative.words y degree.tv* are related. Over the variables, age, negative.words and degree, the Pearson correlation is calculated through the object "Browse 38" (Alteryx 2016) as shown in Table 4. The previous analysis concludes that there is a strong correlation between the variable positive.words and age.tv. There is a low correlation between *positive.words* variable and the level of education of the

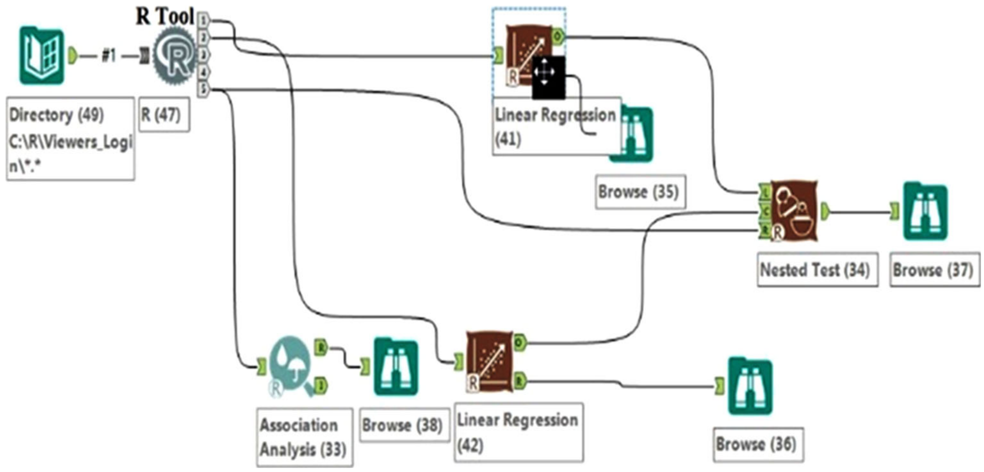


Fig. 13 Prototype MLA predictive model implemented in Alteryx

Table 4 Pearson correlation analysis

	positive.words	negative.words	age.tv	degree.tv
Full correlation matrix				
positive.words	100.000.000	0.54992449	0.71275	−0.00583528
negative.words	0.54992449	100.000.000	0.75399	0.00052843
age.tv	0.81275	0.02114252	100.000.000	0.59334439
degree.tv	−0.00583528	0.00052843	0.59334439	100.000.000
Matrix of corresponding p values				
positive.words		0.00000	0.0000e+00	0.66580
negative.words	0.00000		0.11756	0.96880
age.tv	0.0000e+00	0.11756		0.00000
degree.tv	0.66580	0.96880	0.00000	

viewer (degree.tv). For this reason, it is not used for the predictive model *degree.tv* variable.

An analysis of linear regression is obtained with the component “Browse 35” (Alteryx 2016) that correlates the variables “gender.tv + age.tv + degree.tv + county.tv + city.tv”, and shows the results in Table 5.

Table 5. Concludes that there is low correlation of the variables: *contry.tv, state.tv, city.tv* with the positive variable *positive.words* based on what the viewer types.

Continuing with the flow of data analysis on the linear regression model (see Fig. 8) with the object Linear Regression Tool and Browse 36 (Alteryx 2016), for variables *Positive.word* and *age.word, county.tv*, the results of correlation and regression are shown in Table 6 and Fig. 14.

When analyzing the output of component: “Nested Test” examine two models, one of which contains a subset of the variables contained in the other and evaluates if they are statistically equivalent in terms of their predictive capability. The object Nested Test is used to analyze models: TV_ALL (Table 5) y TV_AGE (Table 6); where the model TV_AGE contains a subset of the variables contained in the model TV_ALL that are statistically equivalent in terms of their predictive, the result (see Table 8) indicates that the predictive model is reduced to model TV_AGE, which confirms the results found in Fig. 14.

The results shown in Table 7 shows a low level of correlation between viewers and country of residence. The representative countries found in the experimental samples are: Colombia, Spain and the USA, the above follows a Colombian migratory behavior as shown in Fig. 15, i.e., the country where the viewers reside does not affect their opinion.

The results above require analysis of other external factors (variables) that influence the viewer preference (*positive.words*). With the experience of computer programming hypothesis on means of marketing that influence the preference of each country it is generated. Analyzing data proposed in the architecture for the measurement of feeling (Fig. 4), they added keywords or word stop with names of radio stations in different countries. As a result, this experiment found phrases like *“la X”, “X música”, “40 principles”, “los 40”, “emisora”, “La Kalle”, “Qué Onda”*. The above words occupy 7.8% of all words found in the comments and opinions from viewers.

Table 5 Linear model TV_ALL: Pearson correlation analysis: "gender.tv + age.tv + degree.tv + county.tv + city.tv"

Basic summary call: lm(formula = positive.words ∼ gender.tv + age.tv + degree.tv + county.tv + city.tv, data = the.data) Residuals:

Min	1Q	Median	3Q	Max
−11.65	−4.94	−1.83	2.66	72.62

Coefficients: (2 not defined because of singularities)

	Estimate	SE	t value	Pr(> |t|)
(Intercept)	10.727.395	124.117	864.295	< 2.2e-16
gender.tvM	−0.113917	0.20991	−0.54270	0.58736
age.tv	0.007206	0.02892	0.24916	0.80325
degree.tv	−0.009192	0.02654	−0.34638	0.72907
county.tvEspaña	−0.156844	155.913	−0.10060	0.91987
county.tvEstados Unidos	−0.418888	130.372	−0.32130	0.74799
city.tvBarcelona	−1.509.510	125.562	−120.220	0.22934
city.tvBogotá	1.920.632	155.918	123.182	0.21807
city.tvBuenaventura	−1.107.613	120.211	−0.92139	0.35689
city.tvCali	−0.870774	117.191	−0.74304	0.45749
city.tvGijon	0.844339	133.952	0.63033	0.5285
city.tvGinebra	−0.777516	158.047	−0.49195	0.62277
city.tvJamundí	−1.775.081	138.260	−128.387	0.19924
city.tvLa Dorada	−0.693252	160.441	−0.43209	0.66569
city.tvLa Unión	−0.800144	153.949	−0.51975	0.60326
city.tvMadrid	−0.145339	120.140	−0.12097	0.90372
city.tvMedellin	−0.092226	131.109	−0.07034	0.94392
city.tvNueva Jersey	−0.522906	164.072	−0.31870	0.74996
city.tvOviedo	−1.566.312	131.723	−118.910	0.23445
city.tvPalmira	−0.823779	121.232	−0.67950	0.49685
city.tvPereira	0.518144	161.321	0.32119	0.74808

Residual standard error: 7.7344 on 5430 degrees of freedom multiple, R^2: 0.0105, adjusted R^2: 0.001385, F statistic: 1.152 on 50 and 5430 DF, p value: 0.2159

Table 6 Linear model TV_AGE: Pearson correlation analysis: age.tv lm(formula = positive.words ∼ age.tv, data = the.data)

Residuals					
Min	1Q	Median	3Q	Max	
−208.3	−67.2	−16.1	35.0	985.9	
Coefficients					
	Estimate	SE	t value	Pr(> |t|)	
(Intercept)	10.26	154.826	0.6626	0.50772	
age.tv	6.19	0.7545	82.041	6.86e−16	***

Residual standard error: 109.68 on 1032 degrees of freedom multiple R^2: 0.06123, adjusted R^2: 0.06032 F statistic: 67.31 on 1 and 1032 DF, p value: 6.661e−16

New variables created *colombia.tv, spain.tv* and*eeuu.tv* measure the level of popularity of the music video in the three countries with the largest viewing audience. The video is measured on the scale of one (1) to three (3) for each of the above variables according to the most popular radio station in the genre of each country, with the following scale: one (1) does not appear in the list of the Top 10 in the country, two (2) appear in the list of the Top 10 in the country with more than four weeks and three (3) appears on the list of Top days country and takes less than two weeks in the list.

The predictive model incorporates the multivariable analysis for correlations on the score list music video of the moment as shown in example of Table 8.

According to the previous results obtained for the development of the machine learning, which gives recommendations as the list shown in Table 8, arises the need to incorporate methods that allow the characterization of the viewer's pro-

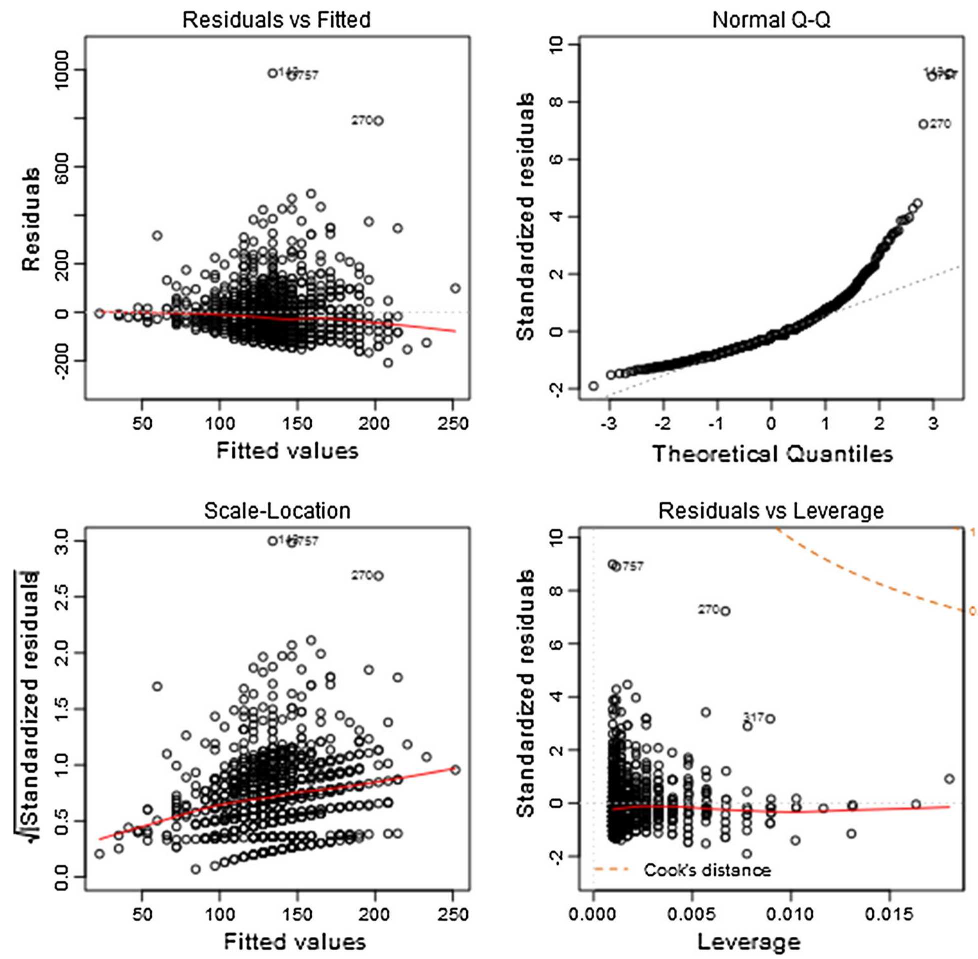
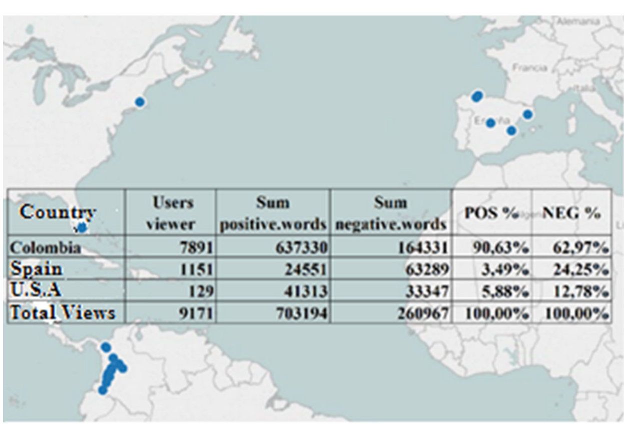


Fig. 14 Diagnostic linear regression positive.words ∼ age.tv (Alteryx)

Table 7 Effect of removing the variables degree.tv, county.tv, state.tv and city.tv from TV_ALL

DF	Sum of squares	*F*	Pr(>F)
46	7379.07	0.9417	0.58458



Country	Users viewer	Sum positive.words	Sum negative.words	POS %	NEG %
Colombia	7891	637330	164331	90,63%	62,97%
Spain	1151	24551	63289	3,49%	24,25%
U.S.A	129	41313	33347	5,88%	12,78%
Total Views	9171	703194	260967	100,00%	100,00%

Fig. 15 Distribution of viewers participation by country

file. Applying the ontology concept (Martinez-Cruz et al. 2015) is possible to characterize a group of users or viewers to improve the recommendations or the predictions made by the machine learning (ML).

7 Practical case: plebiscito of peace in Colombia

The construction of the model and the proposal of adaptive contents in the environment of iTV presented in this paper required to apply different stages of research; however, the design was refined applying experimental hypothesis over a pilot TV show. After applying the processes described before, it is concluded that the Deming cycle was follow in its phases: Plan, Do, Check, Act (Moen and Norman 2015).

The TV show where the Deming cycle was applied is named “A FONDO” (Telepacifico 2016), a life program of opinion with debates and analysis of current politic, socioeconomic topics and the role of the citizen in the region. To do the pilot, the most relevant characteristics were iden-

Table 8 List of music videos for prediction

N	Music video	Music video list	colombia.tv	españa.tv	eeuu.tv
1	M1	Pitbull Messin Around	3	1	2
2	M2	Ariana Grande IntoYou	2	2	1
3	M3	Martin Garrix Lions In The Wild	2	1	3
4	M4	Ariana Grande Let Me Love U	2	2	2
5	M5	Jesse & Joy No Soy Una De Esas	3	1	3
6	M6	Blasterjaxx The Silmarillia	2	1	2
7	M7	Adele Send My Love	2	1	1
8	M8	Kygo Im In Love	3	3	2
..	..	..	2	3	2
20	M10	Enrique Iglesias Duele El Corazon	3	2	1

Consulted on: http://www.wopvideos.com/ (February 2016)

tified at first, finding the following: 1. a strong trend to present politic and social topics. 2. Well known people who master the topic and has different opinions is invited to talk and debate. Shows of this kind stimulate the audience to participate and give their opinion through social networks.

Next is the explanation of the plan to design the adaptive interactive content by the production and direction team of the show:

7.1 Definition of the topic of the life TV show

Once the characteristics and resources of interactivity to be used in the show "A FONDO" are identified, the production and postproduction teams define the topic of debate; in this paper, the emission of " El plebiscito por la paz en Colombia" (Presidencia 2016) is explained. The debate "El plebiscito por la paz en Colombia" is presented in three (3) transmissions of one (1) hour each. The design of the script of the content in the show starts with the definition of Plebiscito:"…*El plebiscito es un mecanismo de participación ciudadana establecido por el Congreso para la refrendación popular del #AcuerdoDePaz alcanzado por el Gobierno Nacional y las Farc en la Mesa de Conversaciones de La Habana, Cuba…*" (Presidencia 2016). Given the national significance of the plebiscito for Colombia and the international impact it can have, the direction team of the show with a group of experts in the topic identifies the following question to be answered by the teleaudience: ¿Usted votaría por el SI o por el NO al Plebiscito? - ¿Would you vote YES (SI) or NO (NO) to the Plebiscite?. The question is presented in the different transmissions of the show "A Fondo" segmented in an estructure of questions that can change according to the dynamic and interaction between the tele audience and the guests.

Table 9 Script of the show "A FONDO"

SE1	In the peace agreement, can there be impunity for FARC-related crimes?
S1	Does the system of transitional justice lead to impunity?
T1	Test for trend S2
S2	Crime committed by the FARC
S2.1	Does the Transitional Justice make equals the ARMY and the FARC?
S2.2	Will the guerrillas go to jail?
S2.3	What will happen with the crime against humanity?
S2.4	What will happen with the crimes committed by the ARMY and the Peace Agreement?
T2	Test for trend S3
SE2	*Compensation to the victims*
S3	Which is the opinion of the victims regarding the Transitional Justice?
S3.1	Who are the victims of the conflict?
S3.2	What will happen with the crime result of the assault to the towns?
S3.3	How will the compensation to the victims be?
S4	How is mechanism to compensate the victims?
T3	Test for trend S6
SE3	*Benefits for the FARC*
S5	What are the benefits for the FARC?
S6.1	What kind of allowance will the guerrilla receive?
S6.2	What is the participation of the leaders of the guerrilla inside of the government?
S6.3	Will the FARC have a political party?
S7	*Which is the trend of the voting to Plebiscite?*

7.2 Scenes and script design

Applying the definition of the model of adaptive contents presented in this paper in Sect.3, the direction and production team together with the guests design the adaptive content described in Table 9.

The script describes the different topics and options of subtopics to be follow during the life transmission. It is important to mention the sequence of scenes and topics of debate {S2n, S3n y S6n} will depend in the interaction with the viewers {TEST: T1, T2, T3} through the social networks and the analysis done by the MLA.

7.3 Selection of the channels of interaction

The production team of the show “A Fondo” configures the accounts in the social networks (Facebook, Twitter, Instagram and WhatsApp) needed to receive the opinions, messages and symbols of the viewers through them during the transmission.

7.4 Transmission of the pilot

After the design and staging of the script defined previously, the host of the show starts the debate by introducing the guests who are experts in the topic and assigns the order and time of intervention. The first test “Test T1” is started over the debate presented {S1}; the results are obtained by LM-ITV to detect the sentiment in the data expressed in {T1} as in Fig. 16. The trends of the sentiment of the viewers is negative about the exposed topic {S1}, getting also a trend of the expressions related with the words “jail”, “prison”, be in jail”, “prisoner” - “cárcel, cana, pagar, presos”.

Based on the previous analysis, the director or TV conductor of the show indicates to the production team, host and guests that should continue the debate by using {S2} oriented

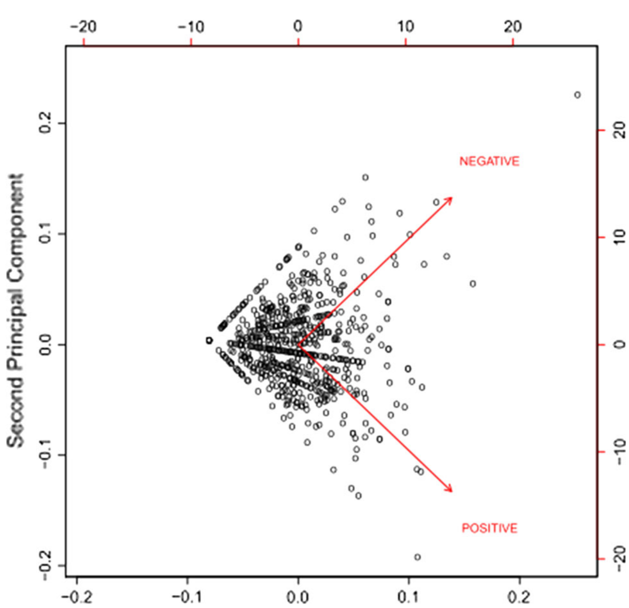


Fig. 16 Results of LM-ITV for test {T1}

to {S2.2} which will be more interesting for the tele audience according to the results in {T1}. With the methodology exposed previously, the sequence of transmission of the life show continues during the transmission of “A Fondo”. To illustrate the dynamic of the adaptive content, Fig. 11 shows the results of sentiment, trends and expressions analyzed by t LM-ITV for the sequence of scenes broadcasted during the show {S1, S2.2, S3.3, S4, S5, S6.3, S7}.

As shown in Fig. 17, the results of the analysis and the expressions delivered by the LM-ITV suggested the directions the show should follow according to the trends in the posts of the viewers, generating at the same time an environment of controversy and debate. Additionally to the results, the ML-ITV is fed by historical information about the conflict and violence phenomenon in Colombia. An interesting finding in LM-ITV is the presence of the expression “paramilitary” in the different messages of the viewers, where the expressions “paramilitares–paracos–para” take relevance in the different discussions of the debate, even when the topic “paramilitarismo” is not included.

As a result of the description above, it is recommended to do an analysis of text mining to quantify the expressions focus in the fenomenun paramilitar, doing the match between the location of the viewers and the negative expressions in the different messages of the viewers, where the expressions “paramilitares–paracos–para” appear (see Table 10).

The obtained results show common trends and characteristics by region, and it can be seen that the words related to the groups of violence name “paramilitares” has a big representation in the region named “Antioquia” and “Santanderes”. The previous results can be compared with the studies done about the phenomenon of violence “paramilitary” in Colombia (Centro Nacional de Memoria Histórica 2015) which gives as final result that the prediction delivered by the machine learning is successful. However, it shows the need to improve the characterization of the viewer’s profile, and to do so, the work done by Martinez-Cruz et al. (2015) can be taken into account, since it takes the fuzzy linguistic modeling to do the characterization based on the trust of the viewer.

8 Conclusion and future work

Television in its transformation influence by the digital evolution has promote the appearance of multiple platforms in the access to the content causing also the need to modify the traditional standards in the generation of audiovisual content. This paper defines a new proposal in the creation of adaptive audiovisual contents, with options to change the sequence of scenes and scripts in real time.

The adaptive content presented in this paper presents the advantage that allows to be adapted to any live show format.

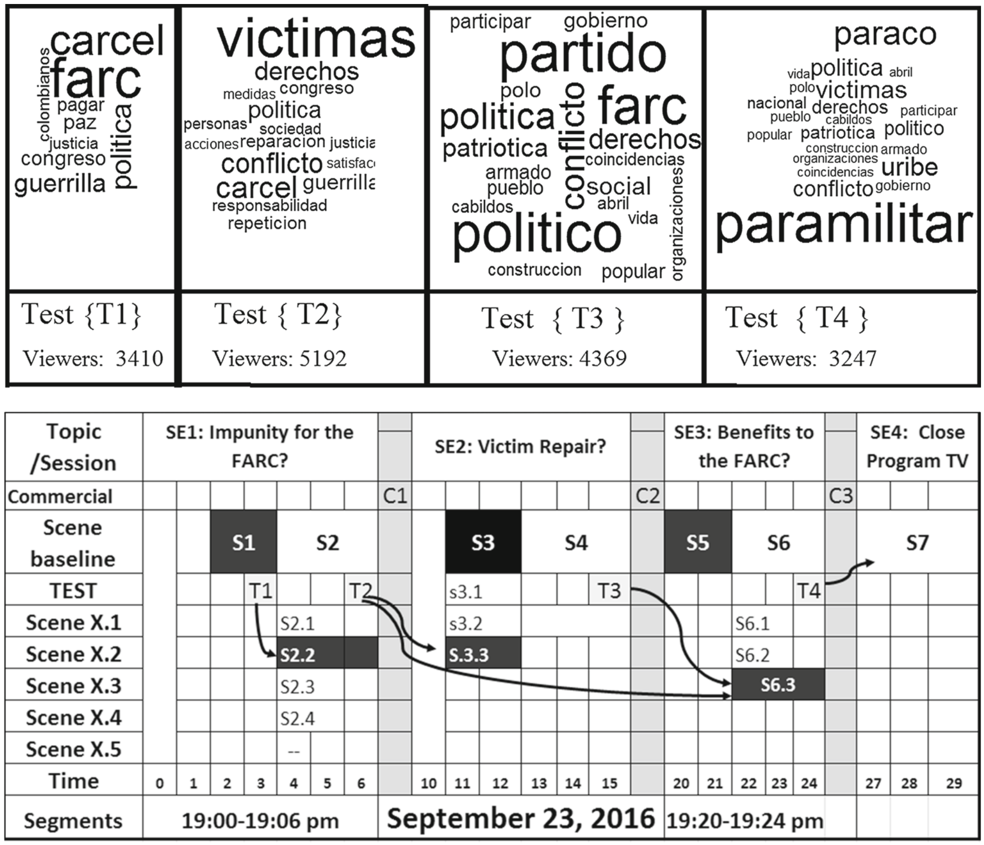


Fig. 17 Example case development of the adaptive content in the program “A Fondo”

Also, it is concluded thanks to the experimental stage that the adaptive content presented has disadvantages related to the production costs when it is compare to the traditional audiovisual contents, since the more flexible options per scene there are the more resources are required to produce them.

Another advantage for the present work is the feedback channels used, since it was found in the experimental stage that social networks are highly used by viewers in comparison with other communications channels, like phone calls or chats.

The quality in the measurement of the sentiment of data depends directly in the historical information collected and analyzed, which is represented by new words, symbols, idioms and current topics that influence the expressions and opinions of the viewers.

The method in the analysis of sentiment of data allows to do an efficient analysis of a group of viewers in real time; however, there is a disadvantage related to the assisted training needed, what leads to an increase in the resources or human interactions. The description above could be addressed by a future work, where methods with evolutionary algorithms are applied to make a deep analysis about the interaction of the viewers between them.

The efficiency in the predictions or recommendations of the machine learning developed depends in the quality of training assisted by the production team of the TV show. In consequence, the experience of the production team and process of decision making in the application of the adaptive contents requires the intervention of humans and cannot be replaced by the machine learning.

The experimental tests done in the present work over TV pilots were broadcasted live allowed to apply the concept of adaptive content, improving the participation of the viewers. In the same way, the expressions and messages posted by the viewers added new topics related with the main topic of the show, which become a new source of information that can be analyzed in future works.

As part of the results obtained in the pilot/experimental TV show, it is proposed for future work to do the analysis of information in the context of collective knowledge with the goal of detecting the behavior and answer of a group of participants in an influenced environment of discussion. This experiment will help evaluate the concepts of: social influence, rank of reduction and auto-confidence of the viewers.

Table 10 Thermal map of measurement of "paramilitary" expressions against the intention to vote for "NO to the plebiscite"

Province	Number of words found by department
Amazonas	2
Antioquia	4619
Arauca	26
Atlántico	582
Bogotá	964
Bolívar	539
Boyacá	373
Caldas	191
Caquetá	65
Casanare	207
Cauca	185
Cesar	1574
Chocó	221
Córdoba	1700
Cundinamarca	293
Guainía	1
Guaviare	6
Huila	86
La Guajira	123
Magdalena	984
Meta	448
Nariño	110
Norte Santander	436
Putumayo	25
Quindío	84
Risaralda	319
San Andrés	1
Santander	968
Sucre	354
Tolima	182
Valle del cauca	400
Vichada	5
Total expressions	16, 073

Acknowledgements This study has no funding.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Alteryx (2016) Alteryx: The Leader in Self-Service Data Analytics, ALteryx, 2016. [En línea]. http://help.alteryx.com/9.5. [Último acceso: 23 03 2016]

Anasse B, Mohamed C, Jung T (2014) Predictive analytics for dummies, p 360

Baldominos Gómez A, Mingueza N, García del Pozo M (2015) OpinAIS: an artificial immune system-based framework for opinion mining. Int J Artif Intell Interact Multimed 3(3):25–34

BBVA (2010) Biplots in practice, BBVA foundation manuals

Bernabé-Moreno J, Tejeda-Lorente A, Porcel C, Fujita HH, Herrera-Viedma E (2015) CARESOME: a system to enrich marketing customers acquisition and retention campaigns using social media information. Knowl Based Syst 80:163–179

Centro Nacional de Memoria Histórica, Rearmados, R, de Desmovilización R (2015) Panorama posacuerdos con las AUC, 1 edn, vol 1. C. N. d. M. Histórica., Ed. Panorama posacuerdos con las AUC, Bogota DC, 2015, pp 513–520

Cesar P, Geerts D (2011) Past, present, and future of social TV, consumer communications and networking conference (CCNC), pp 347– 351

Cesar P, Geerts D (2011) Past, present, and future of social TV. Consumer Commun Netw Conf 49(1):347–351

Collazos Ordóñez CV, Mondragón Pañeda (2008) Lineamientos de usabilidad para el diseño y evaluación de la televisión digital interactiva. Avances en Sistemas e Informática, vol 5, n°. ISSN:19090056, pp 214–218

Del Val E, Martínez C, Botti V (2016) Analyzing users' activity in online social networks over time through a multi-agent framework. Soft Comput 11(20):4331–4345

Duong HT, Nguyen DA, Van Huan N, Nguyen VD (2016) Behavior-based video recommendation using adaptive neuro-fuzzy system on social TV. J Intell Fuzzy Syst 1:12

Garrix M (2016) Compositor, lions in the wild. [Grabación de sonido]

Ghoulam A, Barigou F, Belalem G, Meziane F (2015) Using local grammar for entity extraction from clinical reports. Int J Interact Multimed Artif Intell III(3):16–24

González CB, García-Nieto J, Delgado IN, Montes JFA (2016) A fine grain sentiment analysis with semantics in Tweets. Int J Interact Multimed Artif Intell 3(6):22–28

Hu M, Liu yB (2004) Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery. http://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf Original source

Jiang J, Kohler J, Williams C, Zaletelj J, Guntner G, Horstmann H, Weng Y (2011) LIVE: an integrated production and feedback system for intelligent and interactive TV broadcasting. IEEE Trans Broadcast 57(3):646–661

Julian D (2016) Designing machine learning systems with Python. In: Design efficient machine learning systems that give you more accurate results, 2 edn. Packt Publishing, p 232

Karen C, de Oliveira R (2013) What's up with whatsapp? Comparing mobile instant messaging behaviors with traditional SMS. ACM, New York

Kianian S, Khayyambashi M, Movahhedinia N (2016) FuSeO: Fuzzy semantic overlapping community detection. J Intell Fuzzy Syst 1:12

Li WL, Zhang C, Qiu X (2014) Computational intelligence and security (CIS), 2014 Tenth International Conference on identifying relevant messages for social TV vol 53(3), pp 288–292

Lian S (2012) TV content analysis: techniques and applications. Auerbach, 19 March 2012, pp 222–223

Liu B (2011) Web data mining. Springer, Berlin

Login de Telepacifico, Telepacifico, [En línea]. http://www.telepacifico.com/login

Majd E, Balakrishnan V (2017) A trust model for recommender agent systems. Soft Comput 221(2):417–433

Martinez-Cruz C, Porcela C, Bernabé-Morenob J, Herrera-Viedm yE (2015) A model to represent users trust in recommender systems

using ontologies and fuzzy linguistic modeling. Inf Sci 311:102–118
Mertz D (2013) Text processing in Python, 2 edn. Editorial. Addison-Wesley, pp 55–156. ISBN:0-321-11254-7
Miller TW (2005) The data and text mining: a business applications approach. 9780131400856, Pearson Prentice Hall, Upper Saddle River, p 259
Miller TW (2014) Modeling techniques in predictive analytics: business problems and solutions with R. http://www.ftpress.com/miller, Pearson Education, pp 107–120
Miller TW (2014) Modeling techniques in predictive analytics: business problems and solutions with R. http://www.ftpress.com/miller, Pearson Education, pp 107–120
Moen R, Norman C (2015) Evolution of the PDCA cycle, [En línea]. http://pkpinc.com/files/NA01MoenNormanFullpaper.pdf. [Último acceso: 10 05 2016]
Mondragon Maca VM, Garcia Diaz V, Pascual Espada J, Bhaskar Semwal V (2016) Measurement of viewer sentiment to improve the quality of television and interactive content using adaptive content. International conference on electrical, electronics, and optimization techniques (ICEEOT), vol3(1), pp 143–154
Moreno A, Teófilo R (2016) Text Analytics: the convergence of big data and artificial intelligence. Int J Interact Multimed Artif Intell 3(6):7
Nathan Danneman RH (2014) Social media mining with R. Birmingham, Position
Neira E (2016) Redes Sociales y Televisión: un támden que funciona. de La otra pantalla: redes sociales, móviles y la nueva televisión, EBOOK, Ed., Barcelona, UOC, pp 45–55
Nuñez-Valdez ER, Cueva-Lovelle JM, Sanjuan O, Montenegro-Marin CE, Infante Hernandez G (2011) Social voting techniques: a comparison of the methods used for explicit feedback in recommendation systems. Spec Issue Comput Sci Softw Eng 1(4):61–66
Presidencia de la Republica de Colombia, Acuerdo de Paz, Presidencia de la Republica de Colombia, 01 08 2016. [En línea]. http://www.acuerdodepaz.gov.co/plebiscito. [Último acceso: 04 09 2016]
Ravindran SK, Garg V (2015) Mastering social media mining with R. Mastering social media mining with R: extract valuable data from social media sites and make better business decisions using R. Packt, Birmingham, p 248
Russell MA (2013) Mining the social web,data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More, 2 edn, vol 2. O'Reilly Media, p 448
Sarkar D (2016) Text analytics with Python: a practical real-world approach to gaining actionable. Apress, Bangalore, p 397
Serrano J, Olivas JA, Romero F, Herrera-Viedma E (2015) Sentiment analysis: a review and comparative analysis of web. Inf Sci 3(1):18–38
Settouti N, El Amine Bechar M, Amine Chikh M (2016) Statistical comparisons of the top 10 algorithms in data mining for classification task. Int J Interact Multimed Artif Intell 4(1):4
Shih-Hsuan Y, Xiu-Wen L, Ying-Chen L (2016) A design framework for smart TV: case study of the TaipeiTech smart TV system. 2016 IEEE international conference on consumer electronics—Taiwan (ICCE-TW), vol 5(3), pp 241–248
Tejeda-Lorente A, Porcel C, Peis E, Sanz R, Herrera-Viedma E (2014) A quality based recommender system to disseminate information in a university digital library. Inf Sci 261:52–69
Telepacifico AF (2016) Canal Regional Telepacifico, Telepacifico, [En línea]. http://www.telepacifico.com/afondo/. [Último acceso: 10 8 2016]
Tran VC, Hoang DT, Nguyen NT, Hwang D (2016) A named entity recognition approach for tweet streams using active learning. J Intell Fuzzy Syst 11 (Preprint)
Twitter, Preguntas Frecuentes sobre Retweets (RT). Twitter, 17/3/2016. [En línea]. https://support.twitter.com/articles/230754
Wages R, Grunvogel SM, Zaletelj J, Mac Williams C, Trogemann G (2006) Future live iTV production: challenges and opportunities. Conference automated production of cross media content for multi-channel distribution (AXMEDIS), pp 325–328