



# Using Social Media Data for Online Television Adaptation Services at RTÉ Ireland

**Final Project Report**

*ADDRESSED TO:*
Raidió Teilifís Éireann

*AUTHORS:*
Andrea Barraza-Urbina
Hugo Hromic
Benjamin Heitmann
Himasagar Tamatam
Andrea Yañez
Conor Hayes

**INSIGHT CENTRE FOR DATA ANALYTICS**
National University of Ireland, Galway, Ireland
Unit for Information Mining and Retrieval (UIMR)
GALWAY, IRELAND
2016

# FOREWORD

RTÉ (Raidió Teilifís Éireann) is the national provider of Television (TV) and radio in Ireland. RTÉ broadcasts its content online through the RTÉ Player and provides services to interact with its users using social media, such as Twitter and Facebook. However, RTÉ wishes to exploit the full power of knowledge that can be obtained from social media, and with that knowledge enhance their online services to further engage users. Towards this goal we present the RTÉ XPLORER prototype. This prototype offers services based on both *Social Analytics* and *Information Adaptation*. The overall goal is to offer services to RTÉ end users to support them in exploring the RTÉ product catalogue and understanding what is happening in social media related to RTÉ programming. In this manner, users could find interesting content faster and be encouraged to participate in social media communities discussing RTÉ content.

The goal of this document is to present the project outcomes, the process to achieve these outcomes and conclusions from our social analytic studies. In the document, we first present the concepts that laid foundation to our research and related work. Next, we review the data that characterizes the RTÉ case study, including data sources and data collection strategies. We proceed towards presenting the services we propose for both the end-users of the RTÉ Player service and directives at RTÉ. Finally, we present the implementation of a functional prototype meant to represent how proposed services could integrate with the current RTÉ Player service.

# TABLE OF CONTENTS

# TABLE OF FIGURES

# TABLE OF TABLES

# 1 INTRODUCTION

RTÉ (Raidió Teilifís Éireann) is the national provider of Television (TV) and radio in Ireland. RTÉ broadcasts its content online through the RTÉ Player and provides services to interact with its users using social media, such as Twitter and Facebook. However, RTÉ wishes to exploit the full power of knowledge that can be obtained from social media, and with that knowledge enhance their online services to further engage users. For this goal, RTÉ joined forces with The Insight Centre for Data Analytics. This document outlines the project outcomes of this collaboration.

In this section, we first outline the main deliverables of the project. Next, we highlight the restrictions and key challenges faced. Finally, we present the document outline.

## 1.1 Main Deliverables

The three main deliverables are:

- ***RTÉ XPLORER Prototype:*** The demo offers services based on both Social Analytics and Information Adaptation approaches. The RTÉ XPLORER prototype is meant to be a tangible representation of how services proposed in this document could be integrated in to the RTÉ Player service.
- ***Publication at International Conference:*** Barraza-Urbina, A., Hromic, H., Hulpus, I., Heitmann, B., Hayes, C., Cantle, N. Using Social Media Data for Online Television Recommendation Services at RTÉ Ireland. 2nd Workshop on Recommendation Systems for Television and Online Video, 9th ACM Conference on Recommender Systems, 20/09/2015.
- ***Project Outcomes Document:*** Barraza-Urbina, A., Hromic, H., Heitmann, B., Tamtam, H., Hayes, C., Using Social Media Data for Online Television Adaptation Services at RTÉ Ireland.

## 1.2 Restrictions and Key Challenges

The main research barrier was on the amount and quality of the data that we had access to, concerning RTÉ.

On the one hand, in terms of user preference data related to programmes, it is understandable that RTÉ cannot freely share user profile data. In addition, given that RTÉ users are not obligated to sign-in to use the RTÉ Player service, RTÉ has explained that most of their users do not have user accounts. This means that if they could share data freely, nevertheless the amount of data might not be enough to offer quality Information Adaptation services.

On the other hand, in terms of programme-related data, RTÉ's programme catalogue is highly dynamic, as they add programmes and remove them on a daily fashion.

As a consequence, in order to offer services tailored to the user's unique characteristics, we face the following challenges:

- Lack of personal preference data such as ratings.
- Relatively little historical user session information.
- Dynamic inventory and limited life span of recommendable items.
- No integration of social media analytics.
- Users would be considered anonymous.

We address these challenges by identifying a key opportunity in using data immersed within social media as a valuable resource that can be exploited to better understand user show preferences. Consequently, Insight has addressed the endeavour of providing a set of solutions based on social media that resulted in analytics tools for decision makers and Information Adaptation services to enhance the RTÉ Player service. The goal of this document is to offer a detailed presentation of the proposed solutions.

## 1.3 Document Outline

This section offers an outline of the document structure. The goal of this document is to present the project outcomes, the process to achieve these outcomes and conclusions from our social analytics studies.

We have focused on providing RTÉ with a solution that offers two types of services, Social Analytics and Information Adaption services. These services are intended for two types of end-users: the directives/employees at RTÉ and end-users of the RTÉ Player service. The services and target users can be viewed in *Figure 1*.

For the first case, we offer a customization of the SociaLens platform [47] for RTÉ, which is intended for directives/employees of RTÉ to aid decision making processes. SociaLens [47] is a business insight platform for enterprise social media.

For the second case, we propose the RTÉ XPLORER prototype service, which is directed towards the end-users of the RTÉ Player service. The overall goal is to offer RTÉ end users, services to support them in exploring the RTÉ product catalogue and understanding what is happening in social media related to RTÉ programming. In this manner, users can find interesting content faster and be encouraged to participate in social media communities discussing RTÉ content. In order to materialize the ideas proposed in this document, we developed a functional prototype, called the RTÉ XPLORER.

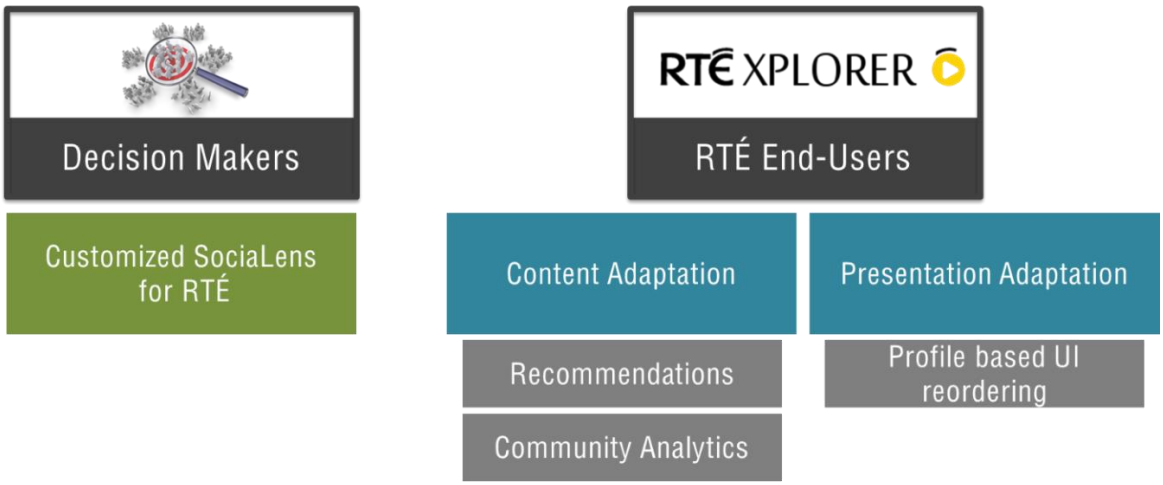


**Figure 1. Adaptation and Analytics Services for RTÉ**

In this manner, the document is structured as follows:

- ***Chapter 2. Background and Related Work***: This chapter highlights what Information Adaptation is and how social media and Information Adaptation have been important factors in online television services today.
- ***Chapter 3. Review of RTÉ Data***: This chapter outlines the data available within the RTÉ use case, including data sources and data collection strategies. The data identified in this section offers support to all the proposed services in this project.
- ***Chapter 4. Social Analytics on Twitter:*** This chapter presents an analysis of the data captured from Twitter. We offer special focus to the study of

communities, and how the information captured from social media can be used for Information Adaptation purposes.

- ***Chapter 5. SociaLens for RTÉ***: This chapter describes the process of adapting SociaLens to the RTÉ use case. In addition, we show the potential use of the data visualizations provided by SociaLens for decision makers.
- ***Chapter 6. RTÉ XPLORER Prototype Services:*** This chapter presents services that will be provided by the RTÉ XPLORER prototype.
- ***Chapter 7. RTÉ XPLORER Prototype Implementation****:* Describes the technical aspects of the development of the RTÉ XPLORER prototype.
- ***Chapter 8. Conclusion:*** This section presents final conclusions of the project and a review of potential future work that can be further explored.

# 2 Background and Related Work

In this section, we define the concepts that lay foundation to our current work. First, we explore the connections between TV, Social Media and Twitter. Then, we argue that TV Recommender Systems are needed tools to cope with Information Overload.

## 2.1 TV, Social Curiosity and Social Media

Watching television can lead to a number of social experiences. People can gather around the television set with family and/or friends to view and discuss their favourite programmes. Furthermore, television facilitates social interaction, giving common ground for even strangers to establish a conversation and feel in some way connected [40][21].

However, the way we consume media today has changed, it is no longer necessary to watch a show live given non-live alternatives such as Video on Demand (VoD), Internet Television (IPTV) and streaming services like Netflix [3]. This removes restrictions over where, how and when TV is watched. In this setting, given the wide variety of TV shows available and the freedom of selecting when to view a programme, it is increasingly less likely for two viewers to share a common experience and thus socially interact. Nevertheless, "*all indications point towards a lack of ability to communicate, not a lack of desire"* [40]. As a solution, users have re-purposed social networks to share their watching experience with friends and even total strangers [23][53]. However, understanding the new ways people communicates about TV presents challenges and opportunities to create solutions that support the characteristics of modern TV viewers.

Social Television is about offering support to social interactions when watching TV, and getting as close as possible to a physically shared experience [40][37]. For example, the project CollaboraTV [40] focuses on providing users "*a sense of social presence*" and supporting asynchronous viewing in a unified interface. In [15], sociability heuristics for social interactive television are defined, from these we

highlight: allow both synchronous and asynchronous use, encourage shared activities and offer different channels and levels for communicating freely.

In spite of Social TV efforts, these are very much focused on encouraging interaction during the TV watching experience. We argue that viewers are also interested in what is happening in social media before and after they watch a programme. People want to catch up on what has been happening or are curious on what others have to say about the programme they have recently watched. In [14], *social curiosity* is defined as the "*desire to acquire new information about other people and the resulting exploration of the social environment*". We observe from related work that the development of tools to help satisfy social curiosity has been under explored for Online Television solutions. Thus, the RTÉ XPLORER prototype has focused on solutions that deliver social awareness, by offering a structured view of what is going on in social media related to RTÉ content. Furthermore, our tools provide quick mechanisms for users to engage with the communities in social media they could be the most connected to.

## 2.2 Twitter: The Grapevine for TV Fans

Twitter is a social network microblogging tool where users post about their current activities, opinions and status in short messages called tweets [28]. Twitter has become largely popular since its launch in late 2006 [28], and more than a communication tool, it is a "*social awareness stream*", given the: public nature of communication, brevity of messages and existence of social relations [39]. In this fashion, conversations in Twitter tend to have a broader audience compared to other social media platforms [8], promoting instead a more open ambient for socializing.

Kaplan and Haenlein [30] define ambient awareness as "being updated about even the most trivial matters in other peoples' lives", and explain how Twitter is a type of awareness system: "*different tweets sent out over time can paint a very accurate picture of a person's activities, just like the distinct dots in a pointillist painting can collectively create the beautiful images of a Vincent van Gogh, John Roy, or Chuck Close*" [30]. Given that Twitter can provide users a window on other people's lives, it can also provide an idea on what is happening within groups of individuals or communities.

In [32], Twitter is characterized as a valuable "*information spreading medium*", principally because of: (*a*) *Non-symmetric following:* Users can follow other users to gain access to their posting activity. This connection does not need to be mutual or approved. As a result, Twitter has a low level of reciprocity [32] and this could indicate that the act of following is more likely to represent the desire to receive content from users than the intention to contribute back. (*b*) *Retweets:* In [32], authors found that: "*any retweeted tweet is to reach an average of 1,000 users no matter what the number of followers is of the original tweet*". Also, they found that once retweeted, a tweet gets retweeted almost instantly on the following three hops [32]. Hence, the retweet is a powerful means for the fast spread of information.

Equally as important, users can offer context to their posts by means of hashtag annotations, thus linking messages to topics or events. This mechanism makes it easy to later read/share about focused streams of tweets by following hashtags. Specifically, Television has found an important marketing ally in Twitter, incentivizing users to discuss TV content by: highlighting during transmission official programme hashtags, creating contests and opinion polls, and even offering exclusive promotional trailers for their social media audience. Twitter has become such an important source of information for Television, that Nielsen has found a business opportunity in predicting TV ratings from Twitter [41]. In this project, we study Twitter and its value as a source of information to characterize RTÉ TV content, mostly to define how related two programmes are according to relevant user activity (tweets, re-tweets and mentions).

## 2.3 Information Adaptation and Recommender Systems for TV Services

A television service offers hundreds, even thousands, of products to their customers. However, users spend more time "channel surfing", and ultimately end up watching from the top ten most popular channels [1]. Information Overload is a very clear problem for Television services, and solutions are needed to help users find, better yet discover, faster and more efficiently, the appropriate content to suit their unique tastes. This, in turn, will keep them engaged and as loyal customers.

In face of Information Overload, Information Adaptation tailors services based on certain knowledge about the context in which these services are provided. Context

can be defined as anything that influences the interaction between the user and the system, *e.g.,* time, location, among others. Personalization is a type of Information Adaptation where services are customized to user characteristics. Another special type of service that provides adapted/personalized information is the one provided by a Recommender System. By offering products/services that are compatible with the particular characteristics of a user, her/his preferences and her/his context; Recommendation Systems offer adapted/personalized information through pull (initiated by the user) or push (initiated by the system) services to support a user in decision-making processes that involve great amounts of information and a wide space of alternatives. Recommender Systems (RS) have emerged as tools that by means of relevant proactive suggestions help users discover interesting products. Ideally, if users are automatically presented with content they could like, they can spend more time watching TV than searching for something interesting. From a business point of view, RS help expose under explored *section*s of the product catalogue, products that users hardly find on their own or might not even had known are available. As a result, RS help reduce churn and generate higher revenues by capitalizing on long-tail TV shows [1].

There are various works on RS for TV. TV Genius [1] proposes the construction of a Bayes network defining the likelihood that a viewer could enjoy a TV programme. In [43], the use of an item-based RS approach is described, able to suggest less popular items under a high traffic load. In [6][26][2], the integration of both collaborative and content-based techniques is proposed. However, these RS approaches heavily rely on the existence of a user profile; feed by both implicit (from past user behaviours) and explicit user ratings. Alternatively, current Social RS approaches, such as those described in [20][31], are strongly based on user relations and trust-networks. Given the unique requirements of our case study, the RTÉ XPLORER prototype is presented as a non-user-based solution, which generates programme recommendations using social media as a source of collective knowledge.

In this section, we have discussed the fundamental concepts that guide our research and analysed related work. In the following chapter, we will describe available data for the RTÉ case study.

# 3 REVIEW OF RTÉ DATA

In this section, we first identify all data that defines the RTÉ use case. Then, we detect and evaluate potential data sources to assess the current availability of data. Next, we define data collection strategies for the different data sources and finally define a data integration strategy.

## 3.1 Use Case Data Overview

In this section, we will bring to light data related to the RTÉ use case and discuss ideas on their potential usefulness. This analysis was carried out independently of the current availability of the data, but on the awareness that if the data were not available, it could be captured.

Data was identified for two purposes: (*a*) to describe the RTÉ Player service domain and, (*b*) to identify data that could be used as input for Information Adaptation Services. Therefore, we structure data in a *Domain Model* and an *Adaptation Model*.

### 3.1.1 Domain Model

The RTÉ Player service is a video-on-demand service that makes TV content available online. TV content includes: single-production programmes (*e.g.,* movies), multiple-production programmes (*i.e.,* television series) and advertisement. As well, RTÉ publishes through other channels (*e.g.,* YouTube), videos to advertise their TV content (*e.g.* trailers). Nevertheless, in this section we will focus on characterizing *RTÉ Player Items* in profiles, defining items as only the TV content that can be found within the RTÉ Player service.

For our Domain Model, we define in *Figure 2* a general *Programme Profile* which describes all programmes. If a programme is a TV series it can be augmented with information found in the *TV Series Profile*. Finally, we also define an *Advertisement Profile*.

# Programme Profile

## Basic Data

- Programme Name
- Description
- Type of Programme (*e.g.,* Movie, Talk Show, Soap Opera…)
- Category (*e.g.,* Romance, Drama, Sports … )
- Contains Mature Content
- Target demographic
- RTÉ Channel (RTÉ One, RTÉ Two, RTÉ News Now, and RTÉ Jr)
- Is Available Online
- Keywords
- Release Date
- Length of video content

## Television Statistics

- Nielsen Information
  - Live viewing, Time shifted viewing, TV Rating (% population watching), Share of all watching TV at a time, Rating

## Online Statistics

- **RTÉ Player Statistics** + ComScore data
  - Number of Streams, User Actions (Didn't finish watching, how much of a video is watched – user retention)
  - Number fo Favorites
- **YouTube Statistics**
  - Likes / Dislikes, Comments, View count, Favorites
  - Number of Channels
    - Subscribers, Viewing Statistics, Video count
- **Facebook Statistics**
  - Mentions, Shares, Likes
- **Twitter Statistics**
  - Number of Tweets, Trending

## Relation to other online resources

- Linked data information.
- **Associated Webpages**: Programme Website, Wikipedia page, IMDb page
- **Associated Social Media resources:** Facebook page, Youtube User accounts and Channels, Official twitter user accounts and hashtags
- Resources (*e.g.,* social media) asssociated to related content (such as: actors, locations, music, sponsors, books, among others).
- **Sentiment** sorrounding the programme and associated to related topics. For example, does the audience like the actors?
- **News and Events** - Gossip, Announcements.

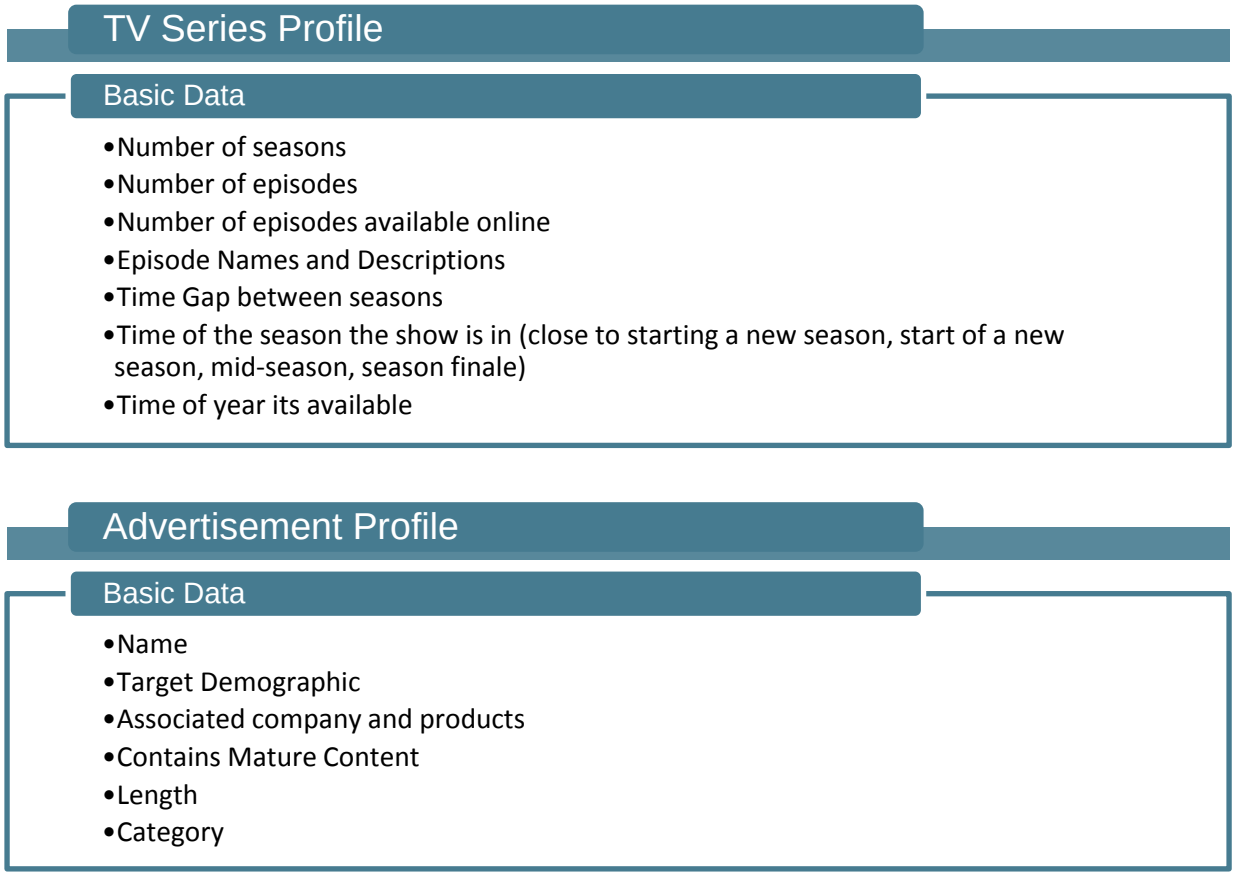


**Figure 2. Domain Profiles**

In our profiles we consider data from social media and external online sources. To further engage with their audience, RTÉ publishes content using social media channels such as Twitter, Facebook and YouTube. On the one hand, in Twitter and Facebook, the company makes available, for example, announcements on news and events surrounding shows. On the other hand, in YouTube the company publishes videos such as trailers, season premieres and promotional videos. In addition, programmes can have their own website where they publish pictures and information about the plot and cast.

We identify a risk in having RTÉ content scattered across the web, which is that users might not interact directly with the company's website to access RTÉ-related content. It would be interesting for RTÉ to create services that would encourage users to mostly use the RTÉ Player, and further engage users to continue interacting with the site. Such services would help users find everything they would want associated to a programme in an *integrated view*. Having detailed Programme

Profiles related to external online resources would certainly be key towards the creation of such a view.

For example, in the case of sports, a dedicated page with scores and statistics on teams, tools to engage discussion around a game, and previous games between the teams; are examples of resources surrounding a match that would encourage users to interact with the RTÉ service not only during the match, but also before and after it. Concerning episode-based programmes, for example, if we know that a new season is going to begin, the service should be able to offer users resources they might be searching for on external sites, such as: show recaps, previous interviews, episode summaries, and others. Furthermore, users could be presented with different views of what is happening in social media concerning the programme, with the aim of further encouraging users to participate in the online communities formed around TV content. User interaction with social media could offer valuable feedback on programme acceptance and viewer engagement. We will further discuss this topic in *section 3.1.2.2* and *section 4*.

### 3.1.2 Adaptation Model

We characterize data that can be used for Information Adaptation purposes in an Adaptation Model. This model includes data that can be used to portray users and their context. We use the following definition for context: *"Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application…"* [12]. For this, we design a User Profile to describe RTÉ users and a Contextual Profile to portray their situation. As an initial definition for contextual characteristics, we explore the user's physical and the user's social media contexts.

#### *3.1.2.1 User Profile*

To define the User Profile, we asked ourselves *what would we like to know about the user that could potentially be used to offer tailored services?.* In general, the goal should be to capture and interpret all user interactions (both implicit and explicit) to get to know the user well enough to anticipate user needs: *"Netflix even tracks how many people start tuning out when the credits start to roll"* [24].

After an analysis of the use case, the data identified is structured in the User Profile shown in *Figure 3.*

Some of the data can be captured directly from the user, given for example a questionnaire. Other data could be captured from the user's implicit interaction with not only the RTÉ Player service, but also external services (*e.g.,* social media). For example, it would be interesting to know if the user has a pet in order to see if a related programme could be suggested. We could attempt to infer this information from the way the user interacts with social media (*e.g.,* the user published a tweet about their pet). Furthermore, sources such as Facebook, already hold basic data on users (*e.g.,* gender, date of birth, and others). If we can create mechanisms to connect RTÉ user accounts with their social media profiles, we could analyse the way user's interact with social media (what they publish, like/dislike, their friends, their content, etc.) to infer information related to their hobbies, habits, interests, among others. Ideally, social media would open alternative doors to get to know the RTÉ audience, and in this way, have the means to tailor services to user preferences without the need of expensive user studies.

It is easy to see how the general information about users could be used for Information Adaptation. For example, an employed user might only be able to access the RTÉ Player service when not in business hours. Alternatively, an unemployed user could have more time to explore the catalogue at different times of the day. However, to be competitive RTÉ should get to know their audience in a more in depth level. For example, study which are their hobbies, which are their interests; and in this way offer tailored programme suggestions so users are aware that RTÉ does have programming that fits their tastes.

## User Profile

### General Data

- Age (Date of Birth)
- Gender (Female, Male, Undetermined)
- Civil Status (Single, Married)
- Place of Residence (Country, City)
- Has a pet? What pet? (Dog, Cat…) (Type, Size)
- Health Status? Disabilities (Could be sensible information.)
- Diet (Is vegetarian?)
- Family Structure (Has kids? How many? Ages…)
- Socio-economic Class (Stratification)
- Culture
  - Ethnic Group, Race (Could be sensible information.)
  - Nationality (Country)
  - Religion (Could be sensible information.)
- Political Point-of-View (Could be sensible information.)
- Occupation (Work, Study, Unemployed)

### Hobbies

- Likes to travel, Likes to cook...

### Habits

- Schedule
- Watching times
- Watching patterns
- Average time spent per Session
- What the user is watching and how long?

### Access Device

- Most used Access Device (Television, Computer, Mobile)
- Characteristics of Access Device

### Interests

- Actors, Shows, Books, Movies

### Preferences

- What programmes does the user prefer over others?
- Preference in Access Device
- Preferred watch times
- Preferred type of programme in different context

### Likes/Dislikes

- Programmes the person likes, actors, topics...

**Figure 3. User Profile**

#### 3.1.2.2 Contextual Profile

As we defined before, context is anything that influences the interaction between the user and the system. There are unlimited possible contextual features that could be considered, however not all are relevant to the use case. For an initial contextual profile, we consider two types of context: the user's *physical context* and the user's *social media context.* We know that users participate in social media, and that the user's social media context might contain useful information to understand the user's decisions and preferences. Thus, we offer special focus to data related to the user's social media context in this project.

In *Figure 4,* we define possible features for each of the profile categories: Physical Context and Social Media Context.

On the one hand, the *physical context* could be used to generate interesting statistics. For example, do users access the RTÉ service from their homes or from work?, what seasons generate the most user viewing?, do user viewing habits change when there are school vacations?, what types of programmes do users prefer to watch during the weekend?, among others. Even the weather conditions when the user is watching a programme could offer interesting insights on user habits: how does the rain affect viewing?, what type of programmes do user's prefer to watch when it's raining?. All this information can be used for decision making and in addition to generate services tailored to the context the user is in.

On the other hand, the *social media context* is meant to characterize the user's environment within social media: with what entities/users does the user interact with, why, where and how. We believe that when a user shares in social media, they engage in different communities emerging from the user interactions. In [16], it is explained that "*Friendship and interaction among users lead naturally to the formation of communities either by explicit connections or by connections that can be inferred through the similarity of the users or the online trails of their interactions (implicit communities)"*. We are interested in observing both explicit communities formed given friendship connections and implicit communities emerging from user interactions. Understanding information about the opinions discussed in communities, the trending topics, the influencing users, among others; can offer

valuable insight on RTÉ users. Furthermore, this information can serve as a foundation to design and deploy services meant to further engage users.

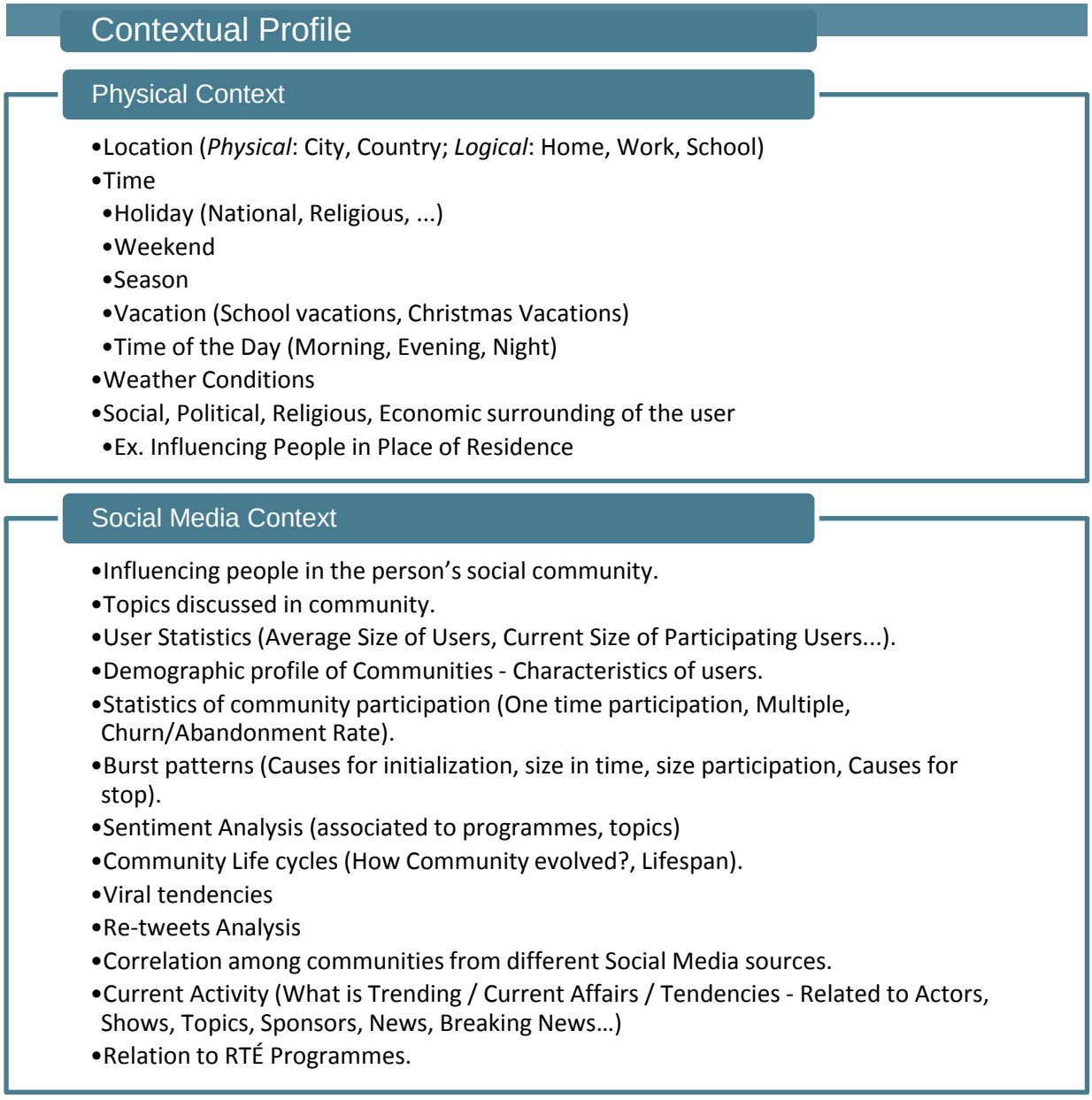


**Figure 4. Contextual Profile**

More specifically, it would be interesting to investigate: how to detect communities discussing RTÉ content?; what events influence the community?; which factors lead to community growth and encourage user participation?; can influencing social media contribute to the success rate of a programme?; in addition to detecting current trending topics, can future trends be predicted?; should social media opinion influence a television series story line?; among many others.

In this section, we have discussed potential interesting data that could be used to generate Information Adaptation services and characterize RTÉ programming. In the following *section*, we will analyse the data sources we currently have available and the data that we can currently access.

## 3.2 Data Sources

In this section, we identify potential data sources and the information that each data source holds. Mainly the data sources identified are:

- *RTÉ Player Service*
- *Social and Media Platforms*: Twitter, Facebook and YouTube.
- *Online Databases*: Wikipedia/DBpedia, IMDb, The Movie DB and Rotten Tomatoes.

In the following *sections* we will describe in detail each data source.

### 3.2.1 RTÉ Player Service

The RTÉ player site is the principal data source on the product catalogue offered by RTÉ. From the site we aim to capture information on programmes and episodes, which are available, when videos will expire and so on. Information describing programmes includes, but is not limited to, the following: name, description, season number, episode number, category and broadcast date. However, we must highlight that programmes to the moment are not described with tags or keywords. Content features would be very helpful in order to compare programmes and make inferences such as: if the user liked programme A then the user will probably also like a similar programme B. TV content is currently classified by RTÉ in broad categories, but in order to carry out insightful inferences on user tastes, it would be best to enhance programme descriptions with content features. We will discuss the possibility of extracting content features from online databases in *section 3.2.3*.

To retrieve information from the RTÉ Player service, we have developed a web crawler which we describe in *section 3.4.1.* From the RTÉ Player Service, we could also, in the future, capture information referring to click-through and user interaction with the service.

In terms of user information, RTÉ allows users to login, initially captures basic user information and registers user activities such as favourite programmes. However, in this stage of the project we do not have access to this information on user profiles.

### 3.2.2 Social Media Platforms

RTÉ has a strong presence on social media, specifically in the platforms of YouTube, Twitter and Facebook. On *YouTube*, promotional videos and trailers can be found concerning the programming broadcasted on the player service. Given that *Twitter* is a microblogging service, RTÉ uses this platform to communicate interesting events surrounding programmes with small messages or tweets. Alternatively, in *Facebook*, RTÉ can publish longer posts. In general, the platforms are used to highlight events using not only text, but also related media (*e.g.,* pictures, videos) and external links (*e.g.,* articles/news from the main RTÉ site). Most importantly, social media platforms allow for RTÉ to interactively communicate with their audience. Viewers can directly respond to messages from RTÉ, ask questions and offer explicit feedback on their preferences using features such as the “Like” button in Facebook and “Favourite” in Twitter.

### 3.2.3 Online Databases

In order to extract content features that could describe TV content, we searched for online databases that would offer information on movies and television series. These databases usually describe content with data on plot, cast, trivia, among others. Specifically, we looked at the Internet Movie Database (IMDb), The Movie DB and Rotten Tomatoes. However, after analysing the possibility of using these databases, we observed that the use of their data is most likely not public for commercial purposes. Furthermore, the data on these databases can be retrieved by other means that wouldn’t have restrictions on the use of the data, such as manual curation or extraction from open sources.

Wikipedia/DBpedia is another open source of information. However, information found on Wikipedia is mostly unstructured. This means information extraction techniques need to be employed to detect structured content features. In addition, manual curation to match TV content with their Wikipedia/DBpedia resources is needed. Nevertheless, we strongly believe that labelling programmes with content

features will open the possibilities for innovative services. However, this goal is outside the scope of the current project.

In this section, we have identified the potential sources that hold data described in *section 3.1.* In *section 3.3,* we analyse the access restrictions for these data sources and establish which we will use for the project.

## 3.3 Data Access Restrictions

For this project we do not have direct access to user profiles maintained by the RTÉ Player service. In addition, we do not have a mechanism to connect user profiles to their social media accounts. It is important to highlight that the sensibility of the information requires the study of possible privacy policy issues and this can be carried out in future work.

From social media, we decided to listen to Twitter, as will be explained in the following *section*. It is important to highlight that Twitter has restrictions over the amount of data that can be captured without payment. These restrictions can be found in [5]. In the future, we could study the benefits of paying Twitter in order to be able to have complete access to data.

We decided to use Twitter as our only source from Social Media given the facilities of the open API. Facebook, on the other hand, has stronger restrictions over user data. In other words, accessing Facebook data is not as straightforward and a number of permissions must be acquired.

Finally, as we mentioned in *section 3.2.3,* the use of data from online databases is generally restricted to non-commercial use. Given the data that can be potentially extracted from these sources, we believe that a better choice is to find alternative sources to extract content features.

In this section, we have identified the data sources that will be used for our project. In the following *section*, we describe how we collected data from the selected data sources.

## 3.4 Data Collection Strategies

In this section, we describe the strategies employed to capture data from the identified data sources. We explain the Web Page Crawler created to capture data from the RTÉ Player service and our methods of capturing relevant tweet information from Twitter.

### 3.4.1 Web Page Crawler

The main goal of the *Web Page Crawler* component is to crawl the RTÉ Player site (http://www.rte.ie/player/ie/) to extract data related to programmes and episodes. This crawl will run periodically in a set time interval and save retrieved information into a local database. The RTÉ Crawler is able to:

**1. Crawl data from the RTÉ Player site.**

In a nutshell, the crawler captures information about programmes and their associated episodes. Examples of captured information include: episode broadcast date, available episodes per programme, episode description, among others.

**2. Save crawled data to a local MongoDB database.**

In order to facilitate the integration of the crawled data to other components of the project, it has been decided that data should be saved into a MongoDB database. More information on the prototype's Data Architecture can be found in *section 7.*

**3. Report errors encountered when crawling data.**

The designed crawler is dependent on the HTML sources of RTÉ web pages, and it is important for the crawler to detect errors when extracting data. This is due to the fact that the RTÉ Player is constantly updated, and code is changed in an un-deterministic fashion.

Our proposed strategy is to implement a robust crawler that is able to detect repetitive errors when retrieving a piece of data and reports these errors by email. The person responsible must then verify and update the crawler's code in order to rectify the data capturing method.

**4. Report new programmes found in new crawl iteration.**

The crawler should report the new programmes it has detected compared to those that have been previously crawled. This information is reported by email and is used to update other project related components.

**5. Run in a periodic fashion given a specific time interval.**

Given the dynamicity of products in the RTÉ Player site, the crawler must run in a periodic fashion to be continuously detecting changes and to keep updated the database information.

The *Web Page Crawler* component was developed in Java, as a Maven Project.

In order to communicate with a database, the Java Persistence API (JPA) was used. Given the mapping of data objects to Entities, the crawler can be used to interact with different databases with small effort to changes of configurations. The *persistence.xml* file is currently configured to communicate with a MongoDB database. However, the crawler has also been tested to be used with a more traditional SQL database (specifically PostgreSQL). To communicate with MongoDB, Hibernate OGM was used (http://hibernate.org/ogm/).

### 3.4.2 Twitter Data Collection

Twitter provides two APIs for accessing its data: (*a*) the REST API, intended for crawling data such as lists of followers, friends and tweets matching certain search criteria, and (*b*) a Streaming API, intended for live real-time capturing of tweets using provided filter terms such as hashtags, keywords and users to follow. In both of those APIs, Twitter imposes request restrictions with the purpose of mitigating potential abuse of their platform [5]. Unfortunately, these restrictions also create key challenges for retrieving data from Twitter. For example, the REST API typically allows for a maximum of fifteen requests each fifteen minutes and the Streaming API provides a fraction of the total tweets being published based on the current rate of tweets the system is transporting. In addition, the REST API (through the *search* endpoint) is stated by Twitter to be focused on relevance instead of completeness, in contrast to the Streaming API [50]. This distinction makes the REST API especially sensitive to the query terms used to retrieve tweets and the time frame of the results.

For retrieving tweets potentially related to RTÉ programming we decided to experiment with both APIs.

On the one hand, for listening to the Streaming API we used the approach proposed in [25], where an adaptive method is employed. This approach consists on using individual sets of *seed* terms (hashtags, users to follow and keywords) for initially listening and then *extending* those sets dynamically according to frequently co-occurring terms based on pre-defined thresholds. This is done periodically using tweets aggregated in sliding time windows, whose size and shift are given as input parameters and their values depend on the dynamicity of the received tweeting activity.

The above terms extension is sensitive to spam and noisy tweets that can lead to including terms that are not actually related to our listening goals. For this reason, we have turned the adaptive listener in to a semi-automatic process instead, where an expert's intervention is needed to monitor if the new listening terms should be added to the extended sets. In the future, we expect to fully automate this task using an enhanced adaptive approach.

On the other hand, for the more static REST API, we decided to capture tweets using the *search* endpoint using a curated list of terms in the form of Twitter Search Queries [50]. In our first experiments, this list was manually created for a chosen list of key RTÉ programmes, as described in *section 3.4.2.1*.

Regardless of the source API of the retrieved tweets, one important task identified is to define how to link each incoming post according to which RTÉ programme it may belong to. Because we are already listening and searching using a curated list of programme-specific hashtags, users and keywords; we can directly annotate the target tweets with their associated programmes.

#### *3.4.2.1 Data Selection: What to listen to from Twitter?*

Now that a mechanism has been proposed to listen to Twitter data, the key question remains on which users, hashtags and keywords to listen to that relate to RTÉ. In other words, the first main challenge for retrieving RTÉ-related content from Twitter would be to determine: who to listen to in Twitter, when to start and when to stop listening.

We experimented with different approaches. We performed an analysis of the different objects we could listen to from Twitter. In first place, RTÉ has provided us with a list of 61 official Twitter user accounts used to post information, and a list of official hashtags to monitor in Twitter for a number of their programmes. However, these official accounts tell us what RTÉ is posting about RTÉ. We need to listen to what RTÉ users are saying about RTÉ and about RTÉ programmes. Nevertheless, we don't have a list of which are the user accounts of the RTÉ Player service users. To counteract this challenge, we decided to listen to what all users are tweeting about associated to RTÉ programmes. For this, we created a curated list of hashtags, user accounts and keywords where each can be mapped to an RTÉ programme.

We do observe that listening to everything that is tweeted about specific programmes has limitations, such as:

- Given that international programmes are watched by a broader audience not exclusive to RTÉ or Ireland, we have to be careful of what is listened from Twitter about those programmes in order to prevent introducing excessive noise to our system.
- More tweets will be received about a certain set of more popular programmes. This could bias our dataset.
- The curated list must be constantly updated to stay current with the RTÉ catalogue. This requires constant manual work.
- Tweets captured will solely be about a programme, which means it will be difficult to characterize RTÉ users with this information. The information captured will help describe the specific programmes.

In order to retrieve information on the actual RTÉ users, a better approach could be to listen to all the tweets from all RTÉ Player users. This brings up the challenge of identifying who are these RTÉ Player users in Twitter, and on the side, handling the limits the Twitter API imposes if we listen to all these users. However, if we listen to all tweets from RTÉ player users, maybe we can receive tweets that are discussing specific programmes, even though the user does not use specific RTÉ hashtags or user mentions. A proposed potential set of rules to identify a list of RTÉ Player users within Twitter is:

- Users following official RTÉ user accounts (these accounts where provided to us by RTÉ itself).
- Users that use one of the hashtags RTÉ has provided as their official hashtags.
- Users that mention any RTÉ user account.
- Users that use hashtags of those programmes exclusive only to RTÉ, *i.e.,* national programming.
- Users that follow those programmes exclusive only to RTÉ.

The resulting list of users should be updated frequently. In a day users can start or stop following programmes and RTÉ user accounts. For example, a user that posted a tweet using a known RTÉ programme hashtag once and since a very long time, should be removed from this list of users to listen because this user has actually not displayed a real interest on the programming.

Even though listening to user accounts of specific RTÉ users would be interesting, further work would need to be carried out to be able to identify these users. For the scope of this project, a good approach was to listen to tweets directly generated about RTÉ programmes. As was mentioned, this was carried out with the creation of a manually created list of users, keywords and hashtags that can each be mapped to an RTÉ programme. In this way, we can annotate each Tweet with the RTÉ programme it is related to.

In this section, we have proposed approaches for capturing RTÉ related tweets from Twitter. In the following *section*, we will offer a brief description of evaluations carried out to test the quality of the data retrieved by our listener and crawler.

#### *3.4.2.2 Data Cleaning*

The Twitter listener must perform data cleaning to ensure relevant tweets are fed to the later parts of the pipeline. We perform two tasks for this: (*1*) for each tweet we calculate a spam score based on [25] to decide if a tweet does not provide relevant content and (*2*) the manually curated list is ensured to not have too generic hashtags and keywords (*e.g.*, For the programme “Love/Hate” the keywords “love” or “hate” alone can introduce a great amount of noise) to ensure we do not capture purely non related tweets. For some cases, we used more generic terms (*i.e.,* "#castle") only

after a manual inspection confirmed that the terms yield more relevant Tweets than non-relevant. For those programmes with too generic names we preferred the usage of official accounts mentioning.

## 3.5 Data Integration Strategy

We integrate all information in a data store. Programmes captured from the RTÉ web crawler are given a unique identifier. If the programme is new, then a manually created list of related hashtags, users and keywords are assigned to the programme. This list is added to the Twitter listener. Consequently tweets captured can be directly mapped to programmes identified by the RTÉ web crawler.

## 3.6 Summary

In this chapter, we have analysed RTÉ related data from different perspectives. In first place, we considered all possible data that could be captured and what it would be useful for. Then we identified possible data sources. We choose to capture information from the RTÉ Player service and from Twitter. Next, we defined strategies to listen to each of the data sources and explained how data is integrated.

In the following chapter, we will offer an in depth analysis of the data captured from Twitter.

# 4 Social Analytics on Twitter

In this chapter, we propose and analyse approaches that use Social Media and Community Detection to understand how RTÉ users (*i.e.*, its online audience) use Twitter to organise and discuss RTÉ content, in particular, programming broadcasted and consumed using the RTÉ Player service. Moreover, we also aim to understand how different groups of people engage with RTÉ through this social platform to provide adaptation of the content according to the current social ambience.

The chapter is mainly divided according to two perspectives: (*1*) we first provide an exploratory analysis of the raw Tweets data that was directly obtained from Twitter in *sections 4.1 and 4.2*, and (*2*) we study higher-level implicit user community data that was built from the simpler-level Tweets interactions in *sections 4.3 and 4.4*. In addition, we provide an analysis of how users talk about multiple RTÉ programme and how we can use this behaviour for adaptation services in *section 4.5*. Finally we provide a summary of findings and conclusions in *section 4.6*.

## 4.1 Basic Data Analysis Methodology

In this section we describe our methodology for analysing raw received Tweets in the context of RTÉ broadcasts. The general view for integrating Twitter data with RTÉ Programming data for our Social Analytics is shown in *Figure 5*.

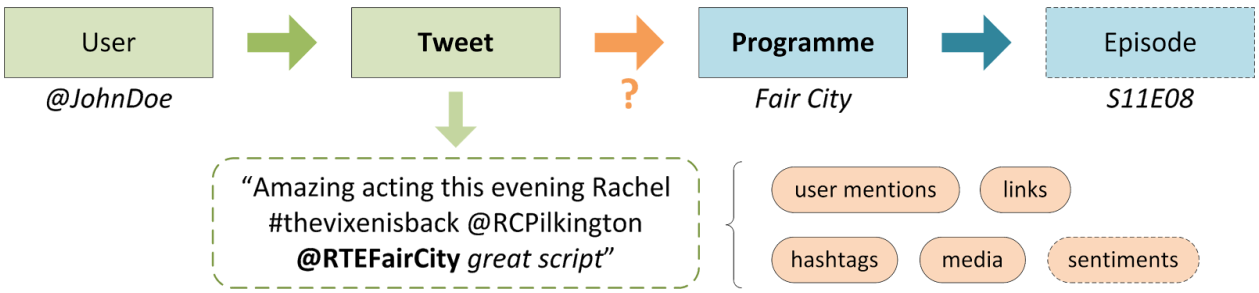


**Figure 5. Overview of Twitter data and RTÉ Programming data integration.**

In Twitter, users post Tweets related to their experience while watching TV. These Tweets can have *explicit* entities such as user mentions (*e.g.,* @RTEFairCity), hashtags (*e.g.,* #thevixenisback), links, media (*e.g.,* photos or videos) and *implicit*

information such as sentiments or opinions, *e.g.* the phrase “great script”. On the other hand we have a collection of current programmes that are being broadcasted by RTÉ (*e.g.,* Fair City, Season 11, Episode 8), and that the Social Media users could be potentially talking about before, during and after the broadcast time. One challenge is how to map Tweets like those to specific RTÉ programmes or even to specific episodes. We address this problem using a manually curated list of Twitter terms for each available programme as explained in *section 3.4.2.1* to annotate the received Tweets according to matching hashtags, keywords or users. With these annotated posts, we can then analyse their statistical and conversational properties, *e.g.,* user retweeting and mentioning behaviours. Moreover, we can build higher-level structures such as implicit user community’s formation and evolution (*section 4.3*) and study how those discussions are further interrelated (*section 4.4*).

In the next section we examine the characteristics of the captured Twitter data, *i.e.* provide descriptive statistics about the captured Tweets. Then we investigate how people tweeted about different RTÉ programmes and finally we study how the users engaged in potential reciprocal conversations.

## 4.2 Discussion on User Behaviours related to RTÉ shows

In this section we analyse the user behaviours captured in Twitter interactions related to the different RTÉ programmes configured for listening. For understanding how users in Twitter, either potential RTÉ audience or RTÉ itself, employ this social media service for exchanging information about the programmes they might be watching, either live or as re-runs, the following questions are stated for discussion during this section:

**Q1.** Do users tweet enough about programmes being broadcasted by RTÉ?
**Q2.** Are the users tweeting exclusively from Ireland?
**Q3.** What kind of programmes users seem to be more interested on in Twitter?
**Q4.** Do users tweet about more than one programme in the same Tweets?
**Q5.** What kind of programmes RTÉ itself promotes more in Twitter?
**Q6.** Do users participate together in conversations at the Tweet level?

To address the above questions, first in *section 4.2.1* we present a basic set of descriptive statistics that characterises the captured Twitter data, in *section 4.2.2* we discuss this Twitter dataset in respect to the Tweets activity towards the TV shows used for listening and finally in *section 4.2.3* we propose reciprocity and popularity of users as measures for studying conversational behaviours.

### 4.2.1 Descriptive Statistics

The following descriptive statistics are reported for Tweets captured between **July 17, 2015**, and **February 17, 2016**, comprising **7 months** of Twitter data for analysis. The Twitter listener was configured to capture tweets from **151** programmes that were available in the RTÉ Player service at the time.

During this period, a total of 11,376,849 Tweets written by 2,288,163 users were received. The spam detection approach described in *section 3.4.2.2* filtered out 780,452 Tweets (6.86%) from the above pool, leaving a total of **10,596,397** Tweets and **2,131,195** users tweeting about **138** (91%) annotated RTÉ programmes (**77** exclusive and **61** non-exclusive to RTÉ). A programme is considered "exclusive" if it is or was produced to some extent by the RTÉ Corporation (*e.g.,* The Late Late Show).

A breakdown of the tweeting activity captured from those 2.1M users can be seen in *Figure 6*. The chart displays the weekly number of Tweets received during the seven months of Twitter listening as well as the cumulative growth in the number of posts.

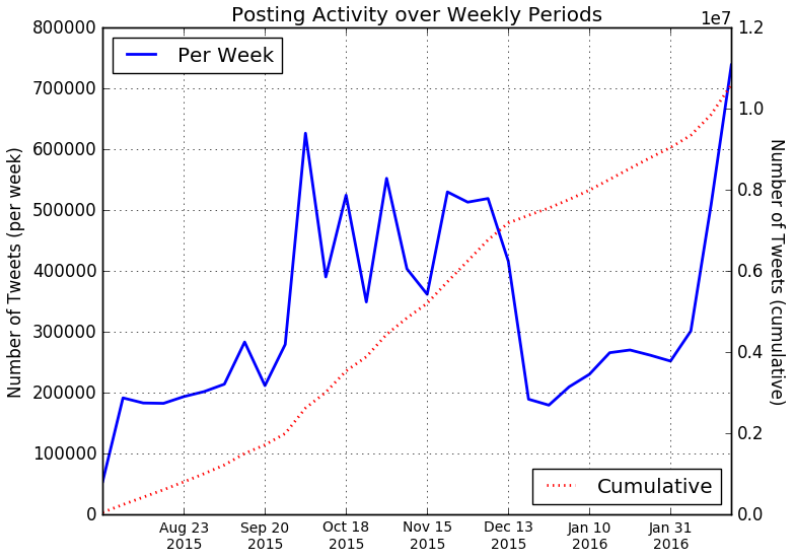


**Figure 6. Overview of the Twitter posting activity in the captured data.**

From the plot it can be observed that the activity behaviour is very dynamic over time, ranging from around 200,000 to 600,000 Tweets on a single week. For example, users posted significantly more Tweets during a concrete period between the weeks of September 20, 2015 and December 13, 2015. This increase in activity is strongly correlated with the Season 5 premiere of the highly popular American TV drama "The Walking Dead". Another increase in activity can be observed after the week of January 31, 2016, where another similar premiere event occurred. The above volume of posts indicates a very strong activity in Twitter related to the TV programming of interest, suggesting a positive answer for Q1.

Furthermore, a steadily growing number of Tweets was received as shown by the cumulative dotted line. A linear regression analysis over the accumulated growth reports a very high goodness of fit ($R^2$ value of 0.989) during the entire capture period. This suggests that despite different events occurring on different programmes, we still are able to receive a stable input stream of Tweets for processing.

We identified **218,620** Tweets (2.06%) having Place data (provided by Twitter), and **1,287,322** (60.4%) Twitter user profiles having Location data (provided by users themselves). From the Places data, users tweeted from a total of **190** different countries, with the top-10 most frequent ones shown in *Table 1*.

| Country | Number of Tweets |
|---|---|
| **United States (USA)** | 128,167 |
| **United Kingdom (UK)** | 39,299 |
| **Republic of Ireland** | 10,163 |
| **Canada** | 9,219 |
| **Brazil** | 3,361 |
| **New Zealand** | 2,679 |
| **France** | 2,395 |
| **Australia** | 2,352 |
| **Philippines** | 2,308 |
| **Italy** | 1,949 |

**Table 1. Top 10 countries identified in Places data and their number of Tweets**

There is a clear bias towards the USA, heavily influenced by the fact that the most popular programme captured, The Walking Dead (Drama), is an American

production. The United Kingdom follows with the second biggest share, driven by the fact that other very popular programmes are of British production, *e.g.,* EastEnders (Drama), Holby City (Drama) or Top Gear (Entertainment), all of them bringing a significant British audience with them. Next is Ireland with 4.6% of the Tweets that had Location data, which is still higher than the rest of the list. This data strongly supports that not all users are expected to tweet from Ireland (Q2), however many of the exclusive RTÉ programmes do contain a majority of Tweets created from within Ireland.

Places data set by Twitter is considered more trustful than Location data because users are freely allowed to input any string into the latter in their profiles, *e.g.,* "*land of sand, sun, and surf*", "*Somwhere between and Heaven*", "*Edinburgh, United Kingdom*", etc. Because of this and despite it being significantly less, we only consider Places data and leave user profile Location information for future work. Nevertheless, as a preliminary exploration we identified the following approximate user profiles frequencies from Location data: United States (**60,850**), United Kingdom (**39,813**) and Ireland (**25,106**). These results may suggest that actually there is a significant Irish-based user presence.

Summarising the descriptive statistics results, we found that a significant amount of 2.1 million users tweet at least once about a generous 91% of the available programmes of the RTÉ Player catalogue during the period of time of listening. Additionally, based on the Places data of the captured Twitter users, we found that if we ignore the bias of the USA-based users tweeting about popular American programmes, the next big majority of users are located within Ireland and the United Kingdom, with the latter being much more predominant. This suggests that the Irish audience might be not fully aware or engaged with all the spectrum of the RTÉ programming offer.

### 4.2.2 How Do Users Tweet about RTÉ Programmes?

In this section, we provide a more in-depth analysis of the RTÉ Programmes that the captured Tweets related to. As expected, the users did not create posts distributed equally among programmes. In fact, this distribution had a very strong long-tail shape as shown in *Figure 7*, which includes all of the 151 programmes configured for

listening. To the left of the Figure is the distribution of Tweets considering all users and to the right the distribution of Tweets written only by the official RTÉ accounts.

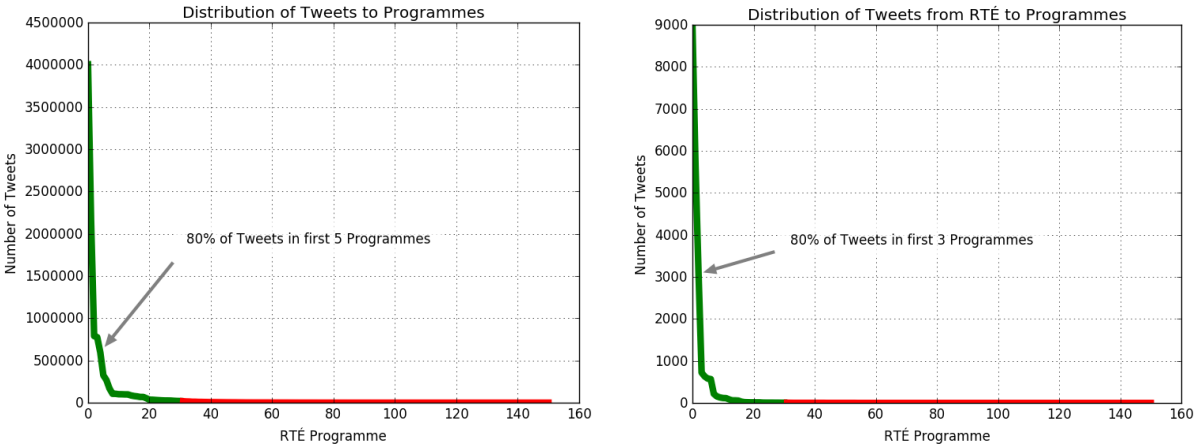


**Figure 7. Distribution of received Tweets to Programmes for all users (left) and only posted by the official RTÉ accounts (right) configured for listening.**

For the case of all users, 80% of the received Tweets are for the top 5 programmes and the top-30 programmes comprise almost all of the Tweets volume. In the case of RTÉ-generated content, the long-tail sharpens even more, suggesting a strong focus from RTÉ to certain programmes for advertising and engagement. Details of the top-10 programmes most mentioned by all users (the first eight of which fall in the Drama category and the last in Entertainment) can be seen in *Table 2*.

| Programme | Number of Tweets |
|---|---|
| **The Walking Dead** | 4,005,844 |
| **Grey's Anatomy** | 2,156,435 |
| **Gotham** | 788,451 |
| **Castle** | 775,556 |
| **EastEnders** | 596,573 |
| **Criminal Minds** | 324,133 |
| **Person of Interest** | 268,298 |
| **Hawaii Five-0** | 168,314 |
| **The Office** | 105,527 |
| **Top Gear** | 104,898 |

**Table 2. Top 10 programmes and their number of mentioning Tweets.**

As revealed in the previous *section*, the table is dominated by American shows that are well-known to be popular in a world-wide scale. However, EastEnders (British production with a big Irish audience) managed to appear within the first five shows.

The 138 programmes that users posted about are spread in the following categories: Drama (**34**), Children (**24**), Factual (**23**), Entertainment (**22**), Lifestyle (**16**), News and Sports (**15**), Comedy (**2**), and Religious and Irish Language (**2**).

In addition, we also studied the distribution of users tweeting about Irish-exclusive programming in contrast to global shows. As mentioned before, the most referenced programmes are mostly American-based and all of them extremely popular global TV shows that are also broadcasted by RTÉ. The Walking Dead dominates by far the amount of Tweets generated, proving to be highly trendy TV show. Those programmes are mostly of Drama and Crime categories. Because of the huge popularity of those shows, it is highly probable that the majority of those users are not only from Ireland. On the other hand, the top-5 most referenced Irish-exclusive programmes at the time of data capture were Six One News, The Saturday Night Show, Prime Time, The Sunday Game and The Late Late Show, all of them nearly having 10 times less Tweets than their global counterparts. However, the categories for those programmers are more diverse, ranging from Factual, News and Entertainment. This data suggests that there is a clear difference in programme categories between global and local shows. This is surprising considering that the vast amount of Tweets refer to global programmes. However, we observe that in the case of Irish-produced content, the Tweets are more spread (*i.e.* cover more available programmes) than the global content where it is heavily focused on popular shows only.

It is also interesting to study if users tweeted about more than one programme in the same Tweets and if more than one show is mentioned by the same users (Q4). For this we computed the number of simultaneous programmes mentioned in the same Tweet and by the same users, both shown as histograms in *Figure* 8 (left for tweets-based and right for users-based). Because there are huge frequency steps across bins, we display those histograms using a logarithmic scale.

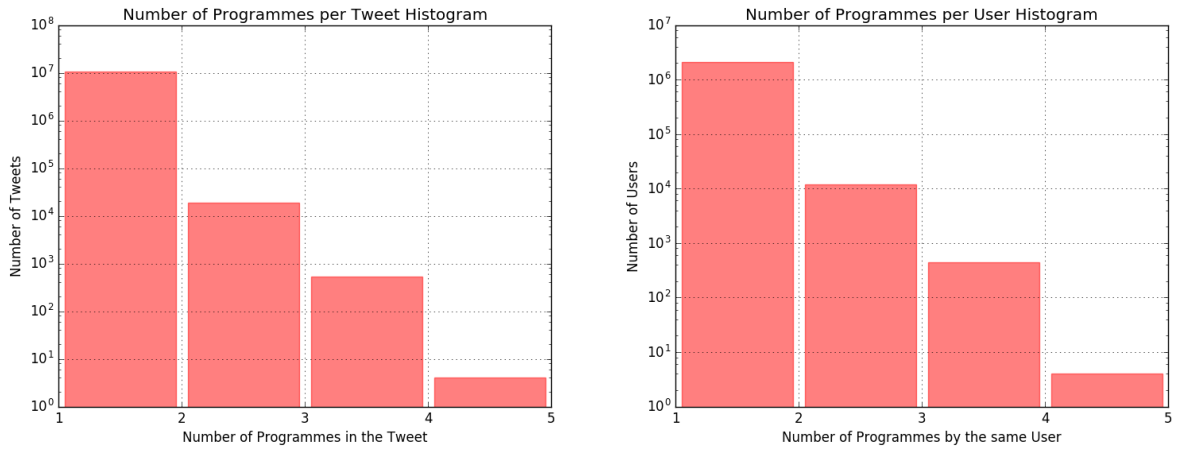


**Figure 8. Log-scale Histograms of the Number of Programmes in the same Tweet (left) and by the same user (right).**

The vast majority of Tweets (10,577,208) and users (2,118,774) only mention one programme at a time. This strongly suggests that users are very specific when tweeting about the programmes they are interested on. Nevertheless there are cases where users post about more than one programme (*e.g.,* 18,652 Tweets and 11,973 users mention two programmes together). Those cases with more than one simultaneous reference are interesting because they allow for programme co-occurrence that enables potential content adaptation based on the joint interests from the social media users (see more details in *section 4.5*).

As described in *section 3.4.2*, we listened to **61** RTÉ official Twitter accounts. In total, we received **21,226** Tweets generated from those official accounts during the same period as for the rest of the data. Those accounts posted Tweets about **54** different programmes (**41** exclusive and **13** non-exclusive to RTÉ), with the following number of programmes per categories: Drama (**12**), News and Sports (**10**), Factual (**9**), Entertainment (**8**) and Children (**6**). This indicates that RTÉ is strongly focused on advertising its own exclusive programming in very particular categories.

In conclusion, users seem to be mostly interested on tweeting about Drama programmes in the case of global shows, but lean towards News, Sports and Entertainment in the case of Irish-exclusive shows (Q3). It is surprising that the Comedy category did not receive much attention from the Twitter users. This could be explained, for example, because Dramas and Factual shows are much more interesting for engaging in longer discussions than Comedy programmes. In this same line, Children programming seems to also spark activity among the younger

audience. On the other hand, RTÉ accounts seem to promote Drama as well as Children and Entertainment programmes the most, which is aligned with the previous finding (Q5).

### 4.2.3 Do Users Engage in Conversations in Twitter?

The final analysis of simple user behaviour in Twitter is to assess if users engage in conversations about the programmes or if they simply post content without entering into active discussions with others. In contrast to more traditional online social systems such as the popular *Boards.ie forums*, Twitter was not designed explicitly for organized discussions in threads. Instead, users can tweet content more openly without the need to always receive replies or retweets. However forms of discussions are still possible in Twitter and they can further lead to implicit communities (see *section 4.3*).

We define the concept of "discussion" in Twitter that we are interested in as follows: *a discussion occurs when two or more users engage in a reciprocal form of communication, i.e. Tweets they generate are replied or retweeted*.

Under this assumption, we borrow the Reciprocity and Popularity measures described in [17], which are part of a collection of user features designed for role-decomposition in online forums:

- **Reciprocity:** *the average percentage (%) of bi-directional users that have replied or retweeted to each other at least once*. Range: [0.0, 1.0].
- **Popularity**: *the average percentage (%) of users' messages that are retweeted at least once*. Range: [0.0, 1.0]

Reciprocity can be subdivided in Reply and Retweet Reciprocities according to which relationship we are considering for computing the reciprocal proportion. For Popularity, and because of lack of tracking metadata, we can only consider Retweet relationships.

We computed those two measures over all the data captured and found that **1,680,015** users (78.8%) replied or retweeted at least once and **961,190** users (45.1%) posted Tweets that were not only retweets. As expected due to the open nature of Twitter (*i.e.* users do not require to be replied or retweeted or to form strong

relationships), a high proportion of users had very low reciprocities and popularities (*i.e.* less than 0.1). In particular, 1,667,542 users (99.3%) had low Reply Reciprocity, 1,669,855 users (99.4%) had low Retweet Reciprocity and 801.312 users (83.4%) had low Popularity. We can directly observe two behaviours from this result: (1) users significantly tend to form more reciprocal relationships using replies than retweets and (2) the majority of users retweet content but they are not retweeted back in the same proportion, functioning more as disconnected information spreaders than active participants in discussions.

Despite the vast amount of users not reciprocating very much or not popular enough, we studied the rest of the computed Reciprocity and Popularity values (*i.e.,* greater than or equal to 0.1) and they can be seen in the histograms shown in *Figure 9* (for both Reciprocities) and *Figure 10* (for Popularity).

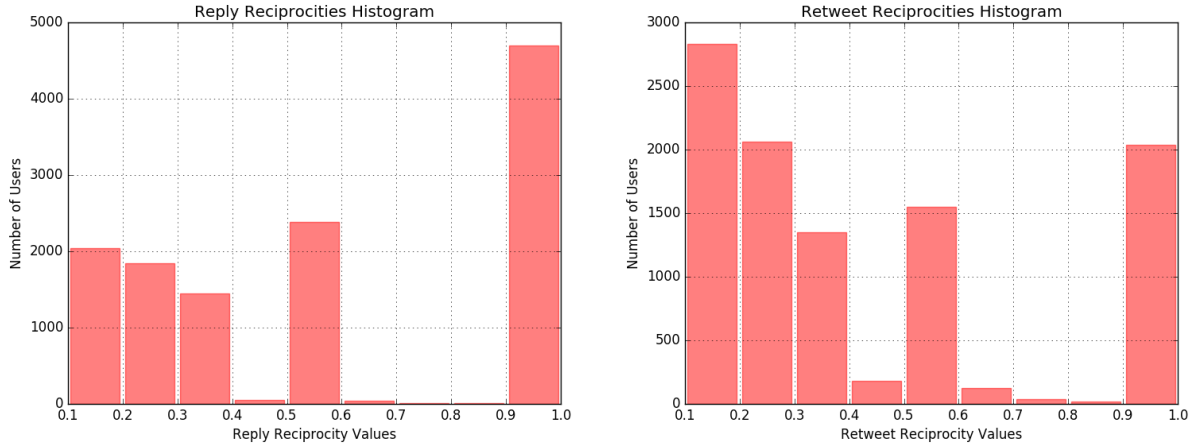


**Figure 9. Histogram of User Reciprocities based on Replies (left) and Retweets (right).**

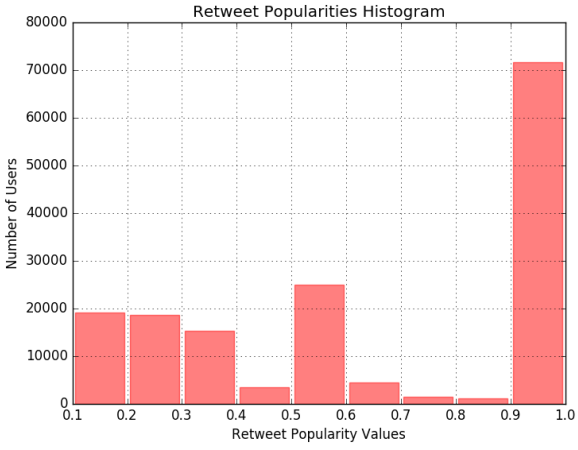


**Figure 10. Histogram of the User Popularities based on retweets.**

In all the histogram charts, three corner cases stand out clearly: (1) the (almost) **no reciprocity/popularity** range [*0.0, 0.1*] –not shown in the figures due to the very high frequency of users–, (2) the **half reciprocity/popularity** range [*0.5-0.6*], and (3) the **full reciprocity/popularity** range [*0.9, 1.0*].

Interestingly, the second most frequent range is of those users with an almost complete reciprocal/popular behaviour (3) for Reply Reciprocity and Popularity, but not for Retweet Reciprocity. This further highlights the notion that Retweets relationships are not good indicators of reciprocal communication but more a driver for diffusion.

The second largest group of users are those that lie in the half reciprocity/popularity value range (2), again for Reply Reciprocity and Popularity but not for Retweet Reciprocity (however this range also has a clear stand-out in this type of Reciprocity). This observation suggests that there is a group of users that mildly participates in reciprocal behaviours with an interesting potential to become more engaged.

Finally, we investigated the user overlaps, *i.e.* how many users are the same between the two Reciprocities and the Popularity value ranges. A new set of histograms describing the overlaps found for each range bin is shown in *Figure 11* (Popularity vs. both Reciprocities) and *Figure 12* (Reply vs. Retweet Reciprocities).

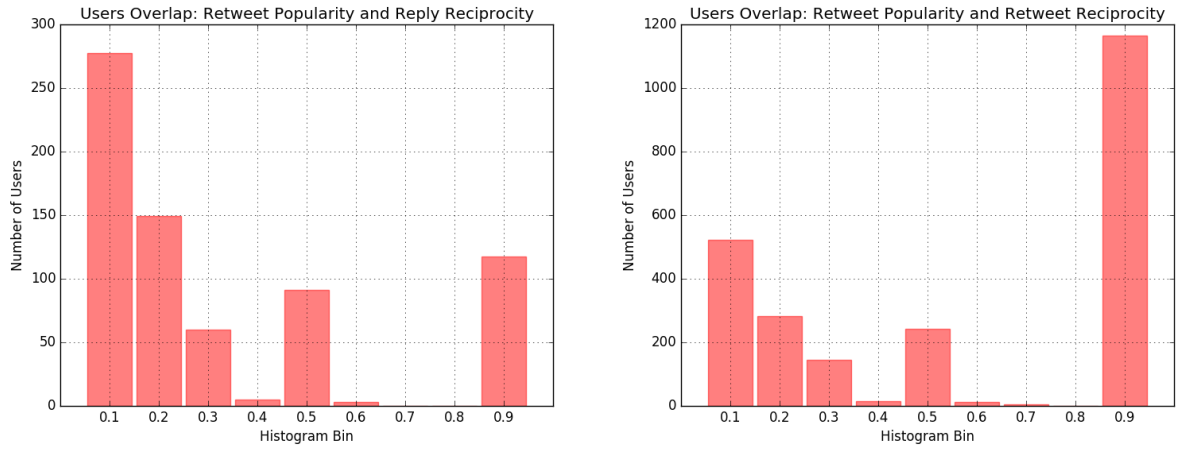


**Figure 11. Histogram of Users Overlap for Retweet Popularity against Reply (left) and Retweet (right) Reciprocities.**

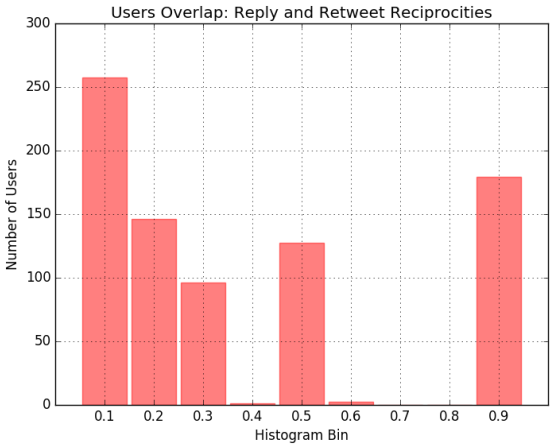


**Figure 12. Histogram of Users Overlap for Reply against Retweet Reciprocities**

The most important conclusion from this user overlap analysis is that (1) there is very little overlap between the cases in terms of number of users in bins, and (2) for the case of Popularity vs Reply Reciprocity there are more users in common not reciprocating at all than in the case of Popularity vs. Retweet Reciprocity, where more of the same users are both popular and reciprocal in terms of content being retweeted and users retweeting them back. Between both Reciprocity types, the overlap is again very low.

In summary, the observations from all the above analyses suggest that despite being a very big group of disconnected users (Reciprocity and Popularity values < 0.1), there are still groups of users that already are participating in discussion-like patterns (values >= 0.1) with varying degrees (Q6). Furthermore, the big group of isolated users potentially could be proactively motivated to join discussions about RTÉ programming.

## 4.3 Community Data Analysis Methodology

In this section, we describe our methodology for building and analysing higher-level implicit user communities extracted from the lower-level raw Tweets interactions.

The basis of a community is that individuals tend to interact more often with members within their same group than those outside. This same effect behaviour can be observed in social media as well [17]. Defining the concept of a community in a social network is challenging, however the following definition is often found in the literature: *"a group of nodes more densely connected to each other than to nodes*

*outside the group"* [49]. In Twitter, the same definition of community is often adopted [28][19][36][45][46]. Users in one community have certain common interests and they share their personal feelings and daily experiences as well. Individuals might have different intentions or reasons for joining communities in Twitter [28], or even same users can serve multiple intentions or serve multiple roles. Our sense of community is then broader given the unique features of the Twitter platform [32], *e.g.* weak user relationships, non-committed reciprocity, restricted Tweet length, distribution of subscribers, fast-pace, among other aspects. Given the lack of close friendships and low reciprocity, trends and topics that persist over time become more prominent, and thus the idea of community must be defined in a different way for our purposes.

Engeström discusses in [13] the notion of "object-centred sociality", arguing that social networks require an intermediate object that connects people together and become social. Without them, people have no interest to get along. Social objects can be of any nature, *e.g.* events, jobs, people, images and music. In the context of Social TV the concept of community usually refers to a close group of friends that share interests about TV shows [22], *i.e.* the social objects. Therefore, we refine the definition of a **Twitter Community** to: *"a group of nodes more densely connected to each other in terms of TV programme social objects, than to nodes outside the group"*.

In Twitter, users can interact with each other using three well-defined mechanisms:

1. **Mentions:** using the syntax "*@username*" in a Tweet. A mention is normally used to gather attention towards a certain Twitter user.
2. **Replies:** a special case of a user mention where the mention is in the beginning of a Tweet. A reply functions as a Tweet directed to a particular user.
3. **Retweets:** referencing someone else's Tweet for it to appear in one's own timeline of Tweets. Retweets are the main mechanism for information spread.

We use the above described mechanisms to build a network of interactions at regular intervals (described in *section 4.3.1*) for later discovery of implicit user communities that could be forming from these interactions (described in *section*

*4.3.2*). In addition, we also propose an approach for tracking the evolution of those found communities over time in *section 4.3.3*.

### 4.3.1 Data Processing Windows

Twitter data is captured continuously by the listener component, creating an *endless* input stream of annotated Tweets related to different RTÉ programmes that can be processed for community detection. However, processing this stream of Tweets is not practical as old data becomes obsolete and does not contribute anymore to any current set of discovered communities. Because of this, we need to aggregate Tweets using a data windowing scheme for detecting up-to-date implicit user communities at regular intervals without exhausting the available computational resources.

The proposed windowing scheme considers windows of fixed *temporal* size, *i.e.* accumulate input Tweets during a fixed period of time (we use a **one hour period**) and then process this aggregated data using the community detection component. Once a window is built and processed, a new window is started with new incoming data. This approach allows having a near real-time view of the user communities formed around the different programming offered by RTÉ.

### 4.3.2 Community Detection

Now we describe our proposed community detection approach. As we stated in *section 4.2*, users post Tweets about TV programmes they are watching and can engage in discussions with other online users about what they like or dislike of those shows, alongside sharing related links or extra content. However, in Twitter there is no notion of explicit user communities other than users having followers, which is more a publisher/subscriber approach than a true community organisation. Despite this, we can detect implicit communities by analysing the posting behaviour of those users. For example when users mention each other, reply, or quote Tweets, *i.e.* retweets.

After capturing, annotating and aggregating relevant Tweets generated by users about RTÉ Programmes, we represent them into a graph of interactions that model how users (nodes) relate to each other using a number of mentions, replies and

retweets (weighted edges) interactions. This User-User graphical model is shown in *Figure 13*.

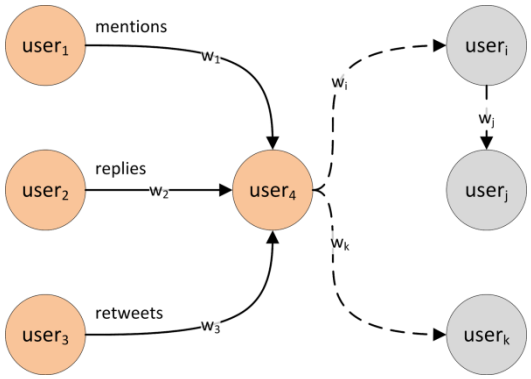


**Figure 13. Twitter User-User directed weighted network.**

This network of interacting users quickly becomes complex over time as the interactions between users increases significantly when more Tweets are received. This graph of interactions is built every time a processing window is completed and is the input for any chosen graph-based Community Detection algorithm. The notion is that users are put together in communities according to how tightly connected they interact.

For our proposed system, community discovery from the above User-User graph is performed using the state-of-the-art OSLOM (Order Statistics Local Optimization Method) algorithm designed for user networks [34]. OSLOM is capable of finding community structures from user graphs like ours based on optimizing the modularity and statistical significance of clusters with respect to random groups. This algorithm takes in account edge directions, weights, community hierarchies (*i.e.* sub-communities) and overlapping (*i.e.* users belonging to more than one community at the same time).

The OSLOM algorithm can induce user communities but it only provides groups of users as output, not their Tweets directly. We assume that those communities are mined from relatively short processing windows (*e.g.*, one hour periods) and Twitter users do not tend to post too much or diverse content within this time frame. With this assumption we then assign to each found user in each found community their latest published Tweet inside the current data window. After the Tweets are assigned to users and therefore to found communities, a variety of content-based sub-

analyses can be done such as hashtag and entity extraction, sentiment and opinion mining and RTÉ Programmes annotation distributions statistics.

As an example, see in *Figure 14* (left) a set of communities (blue hubs) mined from Tweets posted during an hour of activity in July 2015, and based on the interactions between users modelled in a User-User graph. The Tweets found in those communities jointly reference different RTÉ programmes (yellow circles) and hashtags (red circles). A close-up of a sample community can be seen in *Figure 14* (right).

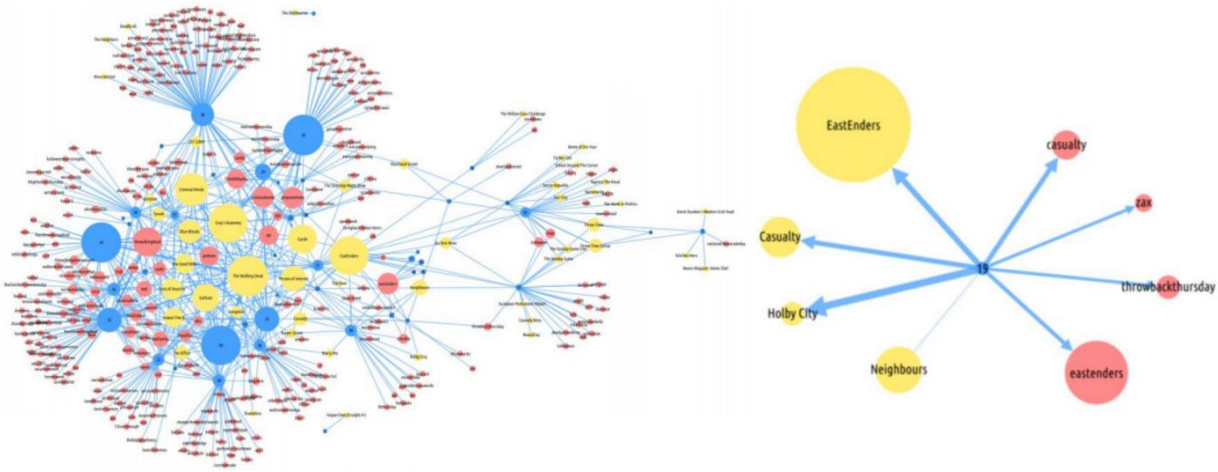


**Figure 14. (Left) sample communities (*blue*) identified around RTÉ programmes (*yellow*) and hashtags (*red*) during an hour of Tweets capturing in July, 2015. A total of 36,648 users and 23,232 Tweets are used. (Right) close-up view of a sample community involving the *EastEnders*, *Casualty*, *Holby City* and *Neighbours* shows.**

### 4.3.3 Community Tracking

In this section, we propose a potential approach for user community tracking over time. Sets of user communities are generated at regular intervals by the community detection approach described in *section 4.3.2*. However, the drawback is that each set of communities is independently mined from every processing window with no notion of the previous ones. We are interested on understanding not only communities in their current states but also their life-cycles over time.

For example, in *Figure 15* a set of communities discussing various RTE Programmes (in different colours) during various time periods can be seen. As depicted, communities appear and re-appear at different moments ($C_1$), discuss different programmes within the same community ($C_4$), and moreover some merge ($C_4$ with

$C_3$) and split ($C_2$ and $C_3$). It is then possible to follow the behaviour of those dynamic communities using community tracking algorithms, *i.e.* the work from Greene *et al.* [18].

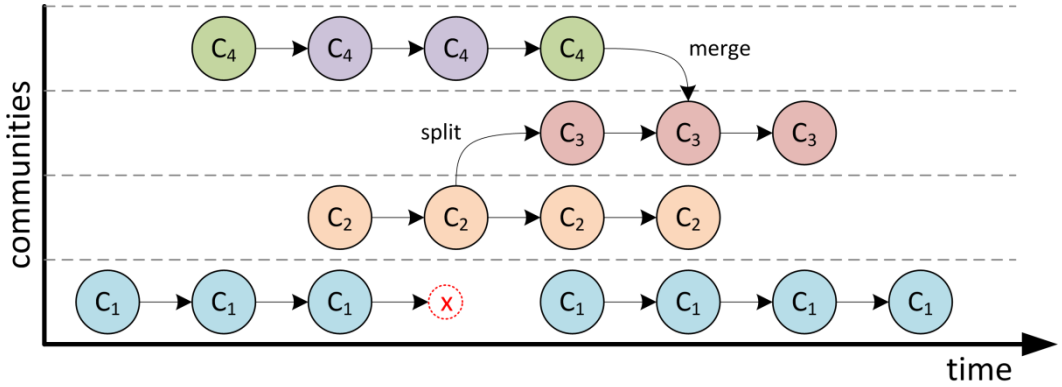


**Figure 15. Community Evolution over Time for Discussions about RTÉ Programmes.**

With this approach, a generic set of user communities can be studied across advancing time periods using well-defined behaviours such as birth, split, merge, intermittence and death. Furthermore, because user Tweets are assigned to the found communities, and annotations about RTÉ Programmes are present in those Tweets, we can study the evolution of user discussions over time not only about a single show, but also a set.

## 4.4 Discussion of RTÉ Programmes in Communities

In this section we analyse the behaviours of users from the perspective of implicit communities identified from their basic tweeting interactions (as detailed in *section 4.3*). The following questions are stated for discussion in this section:

- **Q1.** Do users form enough implicit communities about RTÉ programming?
- **Q2.** What kind of programmes group more users together in communities?
- **Q3.** Do user communities discuss more than one programme simultaneously?
- **Q4.** How grouping users in communities compares to a simple Tweets analysis?

To address the above questions, first in *section 4.4.1* we present a set of statistics that characterises the discovered implicit user communities and then in *section 4.4.2* we discuss the formed communities towards the RTÉ TV shows used for listening.

### 4.4.1 Community Statistics

In this section, we detail descriptive statistics for the implicit Twitter user communities found using the approach described in *section 4.3.2*. In total, **5,165** one hour processing windows (as detailed in *section 4.3.1*) were created and used for discovering communities. Overall inside those windows, we identified a total of **17,299** implicit user communities, with an average of **3.35** communities per window, *i.e.* per hour. However they are not necessarily unique across different windows. In the community tracking *section 4.3.3*, we discussed a potential approach for analysing persistent communities that might span multiple windows.

In *Figure 16*, the left chart shows the distribution of the total number of one hour windows that have 0, 1, 2, etc. number of communities found inside them. It can be seen that the majority of windows are concentrated in the range of equal or less than five implicit communities. In addition, almost 700 windows (13.5%) have no communities found because of very late night times where users do not expose enough activity. On the other hand, there are also windows with a big number of communities, *i.e.* more than 15, however those are very few.

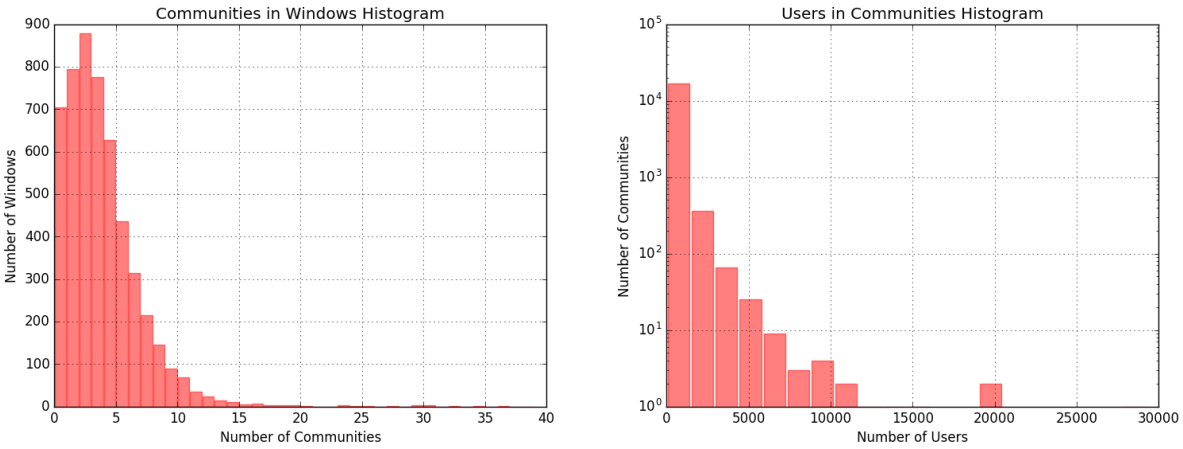


**Figure 16. Histograms for number of Communities in Windows (left) and log-scaled number of Users in Communities (right).**

In the same *Figure 16*, to the right, the distribution of the number of users that were assigned inside communities is shown using a logarithmic scale. The vast majority of communities (16,822; 97.2%) have between 3 and 1,400 member users, followed by 362 communities (2.1%) that contain between 1,400 and 3,000 members. Beyond this point, very few communities have big amounts of users. In particular, two

communities have around 20,000 users and only one community reached 30,000 members. These rare big communities are formed after big shows episode premieres such as The Walking Dead. The community statistics shown so far suggest that a very rich amount of communities formed around RTÉ programming are potentially useful for our approach (Q1). In the next *section* we examine more closely how those communities form around the RTÉ Programmes.

### 4.4.2 How Do Users Gather in Communities About RTÉ Programmes?

In this section, we discuss the characteristics of the found implicit communities in relation to the RTÉ Programmes they reference. Previously in *section 4.2.2*, we discussed the distribution of Tweets posted by users towards the set of TV shows that was configured for listening to Twitter. For that analysis, we found a long-tail shaped distribution as shown in *Figure 7*. Now we perform the same analysis but anchoring by Communities instead of Tweets. The resulting distribution can be seen in *Figure 17*.

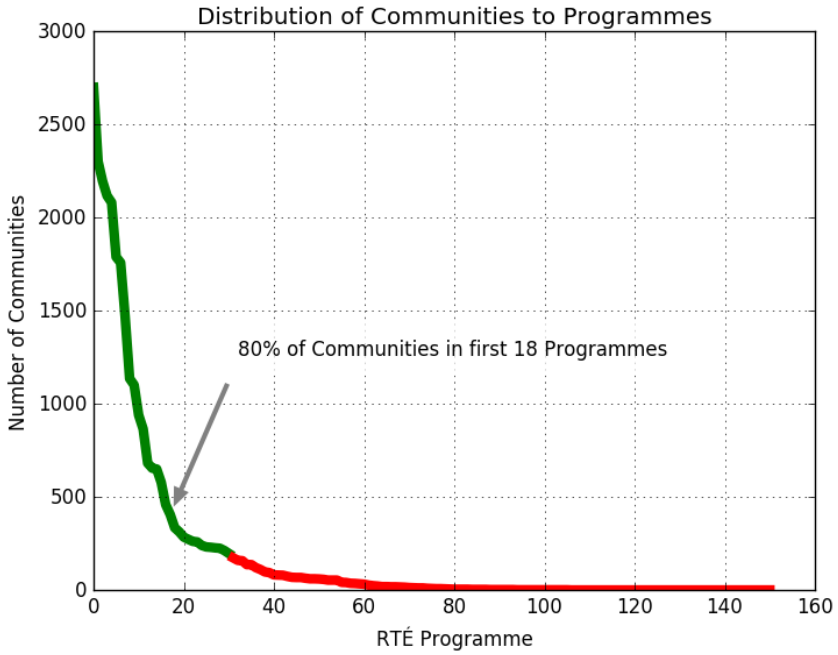


**Figure 17. Distribution of the number of Communities found per RTÉ Programme**

Similar to the Tweets-based distribution, the Communities-based is also shaped as a long-tail curve, however, this curve is much less pronounced because communities are spread more evenly among different programmes. In other words, grouping users in communities allow for less popular programmes to become more visible as

the majority of standalone Tweets from the more popular shows are not necessarily conversational. Using communities instead of individual Tweets allows covering of a wider amount of 18 programmes, in contrast to the five programmes that are covered by 80% of the standalone Tweets that not necessarily are participating in discussions (Q4).

Because the underlying Tweets that are used to mine the user communities are heavily biased towards popular global programmes, the top-5 TV shows in the distribution remains the same as for the Tweet-level. However, for the case of the Irish-exclusive programmes, the top-5 most mentioned shows in the user communities are: “The Sunday Game”, “Neven Maguire: Home Chef”, “The Million Euro Challenge”, “The Saturday Night Show” and “Six One News”. This list differs from the top-5 list of the Tweet-level distribution and also consist of more variate categories (Q2), suggesting that when considering users grouped in implicit communities they may be discussing programmes that cannot be identified naively when only considering simple isolated Tweets (Q4). In respect to global programmes mentioned at least once vs. Irish-produced TV shows, again they remain almost balanced (**52** global and **54** local).

The programme categories used by user communities are distributed as follows: Drama (**33**), Factual (**17**), Entertainment (**16**), Lifestyle (**15**), Children (**14**), News and Sports (**9**) and Comedy (**2**). The major distributional difference compared to only using Tweets lies in the fact that Children-type programmes are less frequent in communities, suggesting that actually most of the standalone Tweets about Children programming does not actually make users to engage in significant discussions (Q2).

Finally we studied the mentioning of multiple programmes in the same implicit user communities (Q3). In *Figure* 18, a histogram of the number of programmes mentioned per community can be seen. The red bars depict programme mentioning by singleton Tweets and the blue bars represent mentioning by Retweet-type posts. It can be observed that there was not much difference between those two types of mentioning.

Nearly half of the total communities (9302; 53.7%) only refer to a single programme. The second half however is spread among communities that simultaneously refer to more than one programme (*e.g.* 2458; 14.2% refer to two, 1433; 8.28% refer to three, 974; 5.6% refer to four). Interestingly, a number of communities (1486; 8.6%) did not refer to any known programme (first bar of the histogram). This is because there are received Tweets that could not be annotated using any programme in the curated list, however they could still be grouped into user communities of *unknown shows* (Q2).

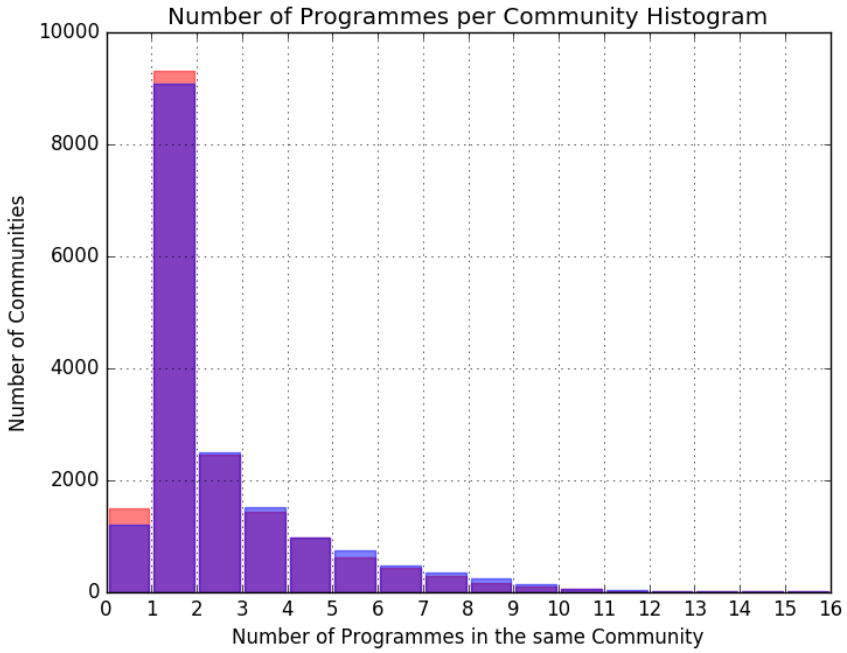


**Figure 18. Tweet (red) vs Retweet (blue) annotations**

To summarise, we have shown that users provide enough Twitter activity signal to extract implicit communities behaviour from their interactions (Q1). We also have shown that users gather in more diverse categories for Irish-exclusive programming when discussing in communities than when posting standalone Tweets (Q2). Users also exhibit that when discussing in communities, they provide conversation about more than one programme, allowing for joint relationships among shows (Q3). Community discovery provide implicit grouping of users tweeting about unknown programmes, which is not straightforward for simple annotation of Tweets (Q4). In the next section, we discuss different approaches for exploiting the underlying knowledge of programmes mentioned together by users in single Tweets and in Communities.

## 4.5 RTÉ Programmes Co-occurrence in Twitter

In this section we explore how RTÉ programmes are mentioned simultaneously (co-occur) at different levels of generalisation. In particular, we can study co-occurrence from two perspectives: (1) at the Tweet-level and (2) at the Community-level. In the following two *section*s we describe each approach.

### 4.5.1 Tweet-based Co-occurrence

In this section, we study the information that can be obtained on the relatedness of RTÉ programmes from **inter-tweet** co-occurrence –programmes mentioned in the same tweet– and **intra-tweet** co-occurrence –programmes mentioned by the same user in separate Tweets–.

For those two types of Tweet-based co-occurrence we build a square NxN matrix containing the N known programmes in both, the columns and the rows. Each cell of this matrix represents the number of times a particular pair of programmes ($P_i$, $P_j$), i,j ≤ N is mentioned in the same tweet, for the inter-tweet case, or by the same user in different tweets, for the intra-tweet case. We can then define the *sparsity* of this matrix as the proportion of zero-valued cells with respect to the total number of cells:

$$sparsity = \frac{\sum_{i,j \leq N} empty(i,j)}{N^2}, empty(i,j) = \begin{cases} 1, & (i,j) = 0 \\ 0, & otherwise \end{cases}$$

Conversely, we can define the *coverage* as the complementary measure of sparsity:

$$coverage = 1 - sparsity$$

We computed the matrix sparsity for both of the Tweet-based co-occurrence approaches on a weekly basis over the seven months of data, accumulating the Tweets data from one week to the next. The results are shown in *Figure 19*. It can be observed that, as expected, the sparsity decreases steadily over time –reaching a minimum of 0.74 for intra-tweet co-occurrence at the end of our captured data– and that the inter-tweet scheme has a much higher sparsity than the intra-tweet. This is because users very hardly post tweets about two or more programmes at the same time given the restrictive 140 characters limit. This restriction prompts users to write much focused and concise tweets. Moreover, the majority of tweets that mentions

more than one programme are used for announcements such as contests or promotions. Nonetheless, it is worth to mention that there still are tweets that genuinely give opinions about two programmes.

We found a strong separation between Irish and global programmes in both types of matrices, *i.e.* there is a low overlap between those two classes of programmes. However there are still weak connections bridging the two islands, allowing for recommendations from one side to eventually be able to reach the other side. From another perspective, there is a high cohesion among categories of programmes. People tweet often about two or more programmes of the same categories, *i.e.* they create connections between programmes that are stronger within the same categories.

Beyond the explicit categories associations, interestingly users also formed connections according to the topics of the programmes. For example, we observed that cooking programmes are tied together by Twitter users and clearly separated from other "Lifestyle"-categorised shows. The same effect could be noted for Children programmes. Another interesting example is the close connection between three programmes, two of them about news broadcasted using sign language and the third about weather news. This association is purely and automatically created as an effect of the social collective in Twitter, without requiring any explicit information in the dataset. For the case of dramas, the strongest category in the dataset, those programmes created a big cohesive cluster and did not show any particular sub-categorisation. This suggests that users tend to discuss about all those programmes at the same time, indicating a strong fan-base for popular drama shows.

In conclusion, our analysis suggests that our captured Twitter users post content over an interesting variety of programmes despite not covering all of the broadcasted shows, and that they keep increasing this coverage slowly over time. More important than having a complete coverage of the programmes is to gain coverage of those programmes that are most interesting and engaging for the users, showcased by the way the socially created connections among those shows and their categories.

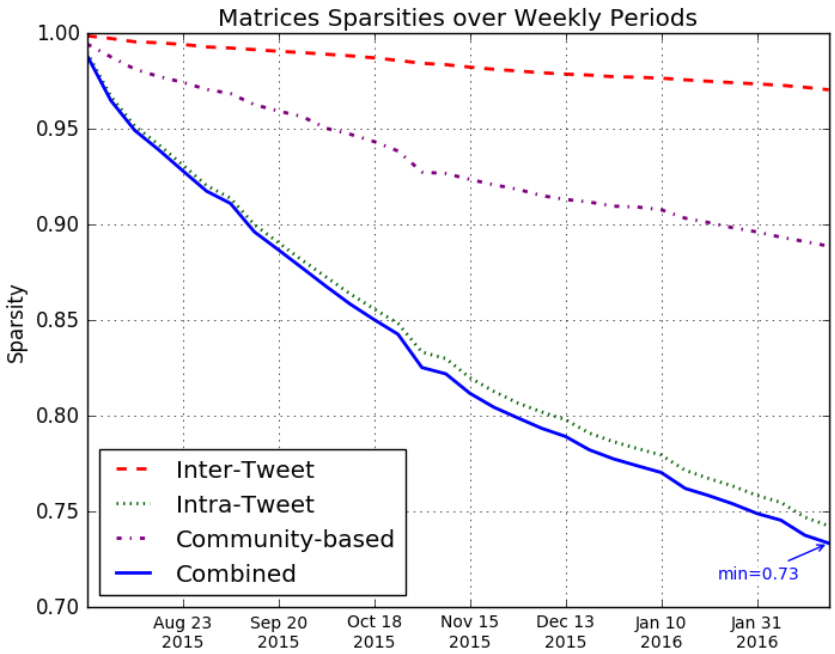


**Figure 19. Matrices Sparsities over Weekly Periods for the different perspectives of co-occurrence of Programmes (Tweets-based and Community-based).**

### 4.5.2 Community-based Co-occurrence

In this section, we study the information that can be obtained on the relatedness of RTÉ programmes from community co-occurrence, *i.e.* programmes mentioned in the same communities, in contrast to tweet-based co-occurrence.

Similarly to the Tweet-based co-occurrences described in *section 4.5.1*, we also take communities detected at regular intervals of one hour as the pivot for a third type of programme co-occurrence. We again build a square NxN matrix for the N known programmes containing the number of times a pair of those programmes is mentioned together within the same user community. The same sparsity measure can be computed and is shown alongside the sparsity values for the Tweet-based co-occurrences in the same *Figure* 19. The sparsity of the community-based co-occurrence lays in-between the inter-tweet and intra-tweets curves, making for an improvement from the inter-tweet but not as good as the intra-tweet approach. However, this result does not imply that community-based co-occurrence is not appropriate. In fact, there is a positive effect of using communities as a relatedness connector in the form that it reduced the noisy connections between programmes. The resulting matrix, despite being sparser than the intra-tweet case, was able to capture most of the same connections with only 24% less coverage, mostly of

programmes with very low tweeting activity. The same types of associations could be also observed in this matrix as discussed in *section 4.5.1*. In other words, the decrease in coverage is not critical for inducing programme relatedness.

Finally, we combined the three matrices obtained using our two co-occurrence approaches (Tweet-based and Community-based) by superimposing them and adding their respective frequencies. This combined matrix sparsity is also shown in *Figure 19*. The intuition for this merge is that the three matrices provide different perspectives for social relatedness of the programmes and by combining them we can blend the social connections from each into a single matrix with enhanced coverage (as shown in the Figure) that can be used for programme recommendations.

## 4.6 Summary

In this chapter, we proposed and analysed approaches that use Social Media and Community Detection to understand how the online RTÉ audience use Twitter to organise and discuss RTÉ programming with the potential of providing not only social awareness but also means of content adaptation.

The observations from all basic Tweets-based analyses suggest that despite our Twitter data having a very big group of disconnected users (Reciprocity and Popularity less than 10%), there are still groups of users that already are participating in discussion-like patterns and in varying degrees. Furthermore, this big group of isolated users potentially could be proactively motivated to join discussions about RTÉ programming.

We have also shown that users provide enough Twitter activity signal to extract implicit communities behaviour from their basic Tweets interactions. In communities, users gather in more diverse categories for Irish-exclusive programming than when posting standalone Tweets. In addition, when discussing in communities, users provide conversations about more than one programme hence allowing for joint social relationships among shows. Furthermore, community detection is able provide implicit grouping of users tweeting about unknown programmes, which is not straightforward for simple annotation of Tweets.

Our co-occurrence analysis suggests that our captured Twitter users post content over an interesting variety of programmes despite not covering all of the broadcasted shows, and that they keep increasing this coverage slowly over time. We conclude that more important than having a complete coverage of the programmes is to gain coverage of those programmes that are most interesting and engaging for the users, showcased by the way the socially created connections among those shows and their categories.

For the future work, a number of further analyses could be done for a better understanding of the social behaviour of Twitter users towards RTÉ programming. For example, Twitter users are known to have different communicator roles [52]: idea starters, amplifiers, curators, commentators and viewers. These roles could be exploited by RTÉ to better engage users towards desired programming, *e.g.* irish-exclusive productions. In this same line, information spread analyses such as Retweet cascading can be done to track how RTÉ-published advertisement or announcements travel across the online Twitter audience. For example, a Tweet could be spread very quickly and also promptly disappear from the audience's interest. We need to better understand how to create Tweets to circumvent this kind of behaviour if RTÉ wishes to optimise engagement with their online media user base. Longer standing Tweets could be exploited as seed content for generating more reciprocity among users and thus potentially generating more community discussions.

Lastly, another interesting analysis that can be carried is a finer-grained user locations study per programme as demographics insights. We already stated that location-based information is present on the captured Tweets and that heuristics can be proposed to estimate from where users are tweeting regarding the shows being discussed. With this information, we could be able to profile programmes according to their popularity at different locations within the country or even on the global scale.

# 5 SOCIALENS FOR RTÉ

SociaLens [47] is a business insight platform for enterprise social media. SociaLens serves to "*automatically analyse and diagnose the health of complex social or collaborative networks*" [47]. We adapted SociaLens to the RTÉ use case in order to take advantage of SociaLens technology and gain understanding of Twitter data related to RTÉ. These services are meant for RTÉ directives/employees.

SociaLens was originally designed to analyse forum-type data modelled as forums, threads and posts written by users. Forums and threads are often employed by communities of users to organize their discussions. Twitter data does not relate directly into this model, hence we developed a data mapping strategy for Tweets generated by RTÉ programmes. Our approach then is able to migrate data captured from Twitter into the SociaLens data model. Specifically, we mapped a forum to all the tweets generated by a particular RTÉ programme in Twitter. As a result, we can now access SociaLens tools, from which we highlight the following:

- **Community Roles:** Displays in a pie chart distribution of users according to the type of role they partake in the community. The possible roles are *Ignored*, *Elitists*, *Supporters* and *Grunts*. In *Figure 5,* an example of the widget can be observed with the definition of each type of role.
- **Community Sentiments:** Displays the number of comments related to positive sentiments versus those related to negative sentiments over time. In *Figure 21,* an example of the widget can be viewed.
- **Community Key Terms:** Key terms are tokens of text that best describe the content posted by users. This widget displays the most frequent found key terms in the community.
- **Community Social Graph:** The Social Graph shows the connections in terms of interactions between users of the community. Users are classified as Supporters, Ignored, Grunts or Elitists. In addition, the size of each node depicts the authority of the user in the community.

- **Community Named Entities:** A named entity is chunk of text such as a name, expression or acronym, and that can be of a Person, Organization or Location type. This widget displays the most frequent found named entities in the community, coloured by their type. In *Figure 21,* an example of the widget can be viewed.
- **Most Influential Users:** Ranks users according to the influence their comments have over the community. User names and the number of posts are displayed.
- **Related Communities:** Users of a community can also participate on other communities implicitly to discuss similar topics. This widget shows all the neighbouring communities whose users have shown discussions related to the community being inspected.
- **Posting Activity:** Statistics on posting activity, specifically shows total posts, total threads, posts per day, threads per day and posting average per day.

In general, SociaLens presents a dashboard composed of diverse widgets that offer different views/perspectives/representations of social data. These views would provide a visual aid to support decision making processes. Specifically, it would allow RTÉ decision makers to better understand the behaviour of social media users and their interaction with RTÉ content. For example, in *Figure 20*, we observe the widget of *Community Roles*. From this information, knowing that a big portion of the community roles is *Grunts*, the social manager could attempt to design strategies to intervene in this community and encourage users to interact.

In *Figure 21*, two more sample widgets are shown: community named entities and community sentiments. It should be noted how the component of time is strongly considered by SociaLens. All widgets can be viewed in different time intervals. In addition, widgets such as the community sentiments can be viewed in an even more detailed granularity.

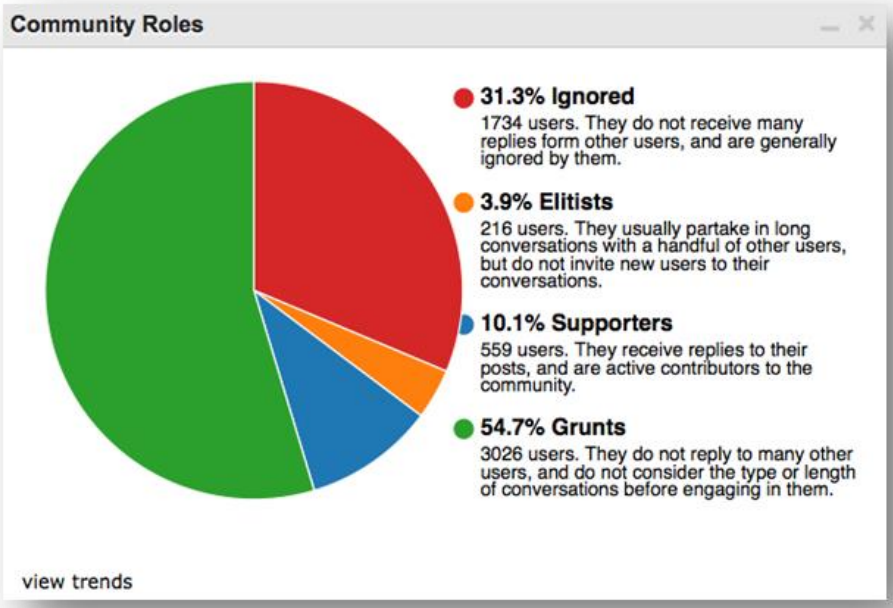


Figure 20. Community Roles Widget in SociaLens

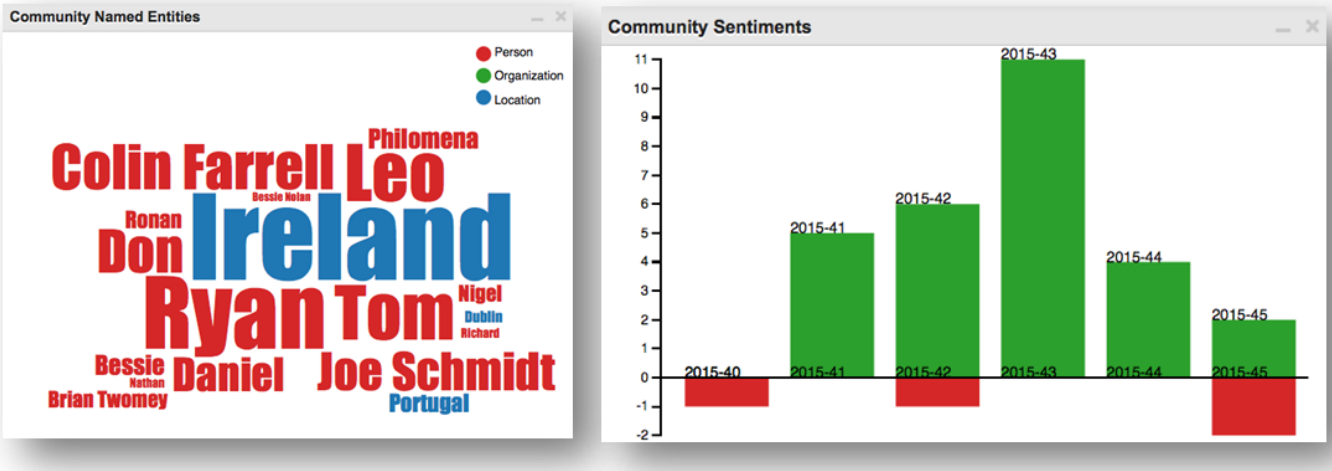


Figure 21. Example Widgets in SociaLens

In this section, we have presented the integration of RTÉ data to the SociaLens framework. SociaLens provides valuable tools that can be used by decision makers at RTÉ. In the following *section*, we will present the RTÉ XPLORER prototype, which integrates both social media-based and Information Adaption services intended for end-users of the RTÉ Player service.

# 6 RTÉ XPLORER PROTOTYPE SERVICES

The RTÉ XPLORER prototype is a set of services that aim to enhance the user's experience when using the RTÉ Player service. The RTÉ XPLORER prototype offers services based on both social analytics and Information Adaptation. The overall goal is to offer RTÉ end users, services to support them in exploring the RTÉ product catalogue and understanding what is happening in social media related to RTÉ programming. In this manner, users could find faster interesting content and be encouraged to participate in social media communities discussing RTÉ content.

First, we will present the core components that structure the RTÉ XPLORER prototype architecture, which include three layers: Data, Adaptation and Presentation. Guided by the architecture, in the following *sections* we further focus on the diverse set of services that are provided by the RTÉ XPLORER prototype. Regarding the Data layer, we have already defined data collection approaches in *section 3*. As for the Presentation layer, our user interface will be described in *section 7*. Therefore, this section will focus on describing in detail the Adaptation layer components.

## 6.1 RTÉ XPLORER Prototype Logical Architecture

In this section, we aim to present the core components that make up the RTÉ XPLORER prototype architecture. By means of the logical architecture, we explain the diverse set of services that are provided by the RTÉ XPLORER prototype.

In the architecture depicted in *Figure 22*, the following layers can be found (read in a bottom-up fashion):

- **Data Layer**: offers the data foundation to enable services provided by superior layers. The *Data* layer performs ETL (Extract Transform and Load) operations to feed a Data Store.
- **Adaptation Layer**: carries out analytical processing over data provided by the *Data* layer to customize services according to contextual features.

Specifically, analytics is carried out for the purposes of content and presentation adaptation.

- **Presentation Layer**: uses information from the Adaptation layer to provide services to the end users. For our system, there are two types of end-users: the directives/employees at RTÉ and end-users of the RTÉ Player service.

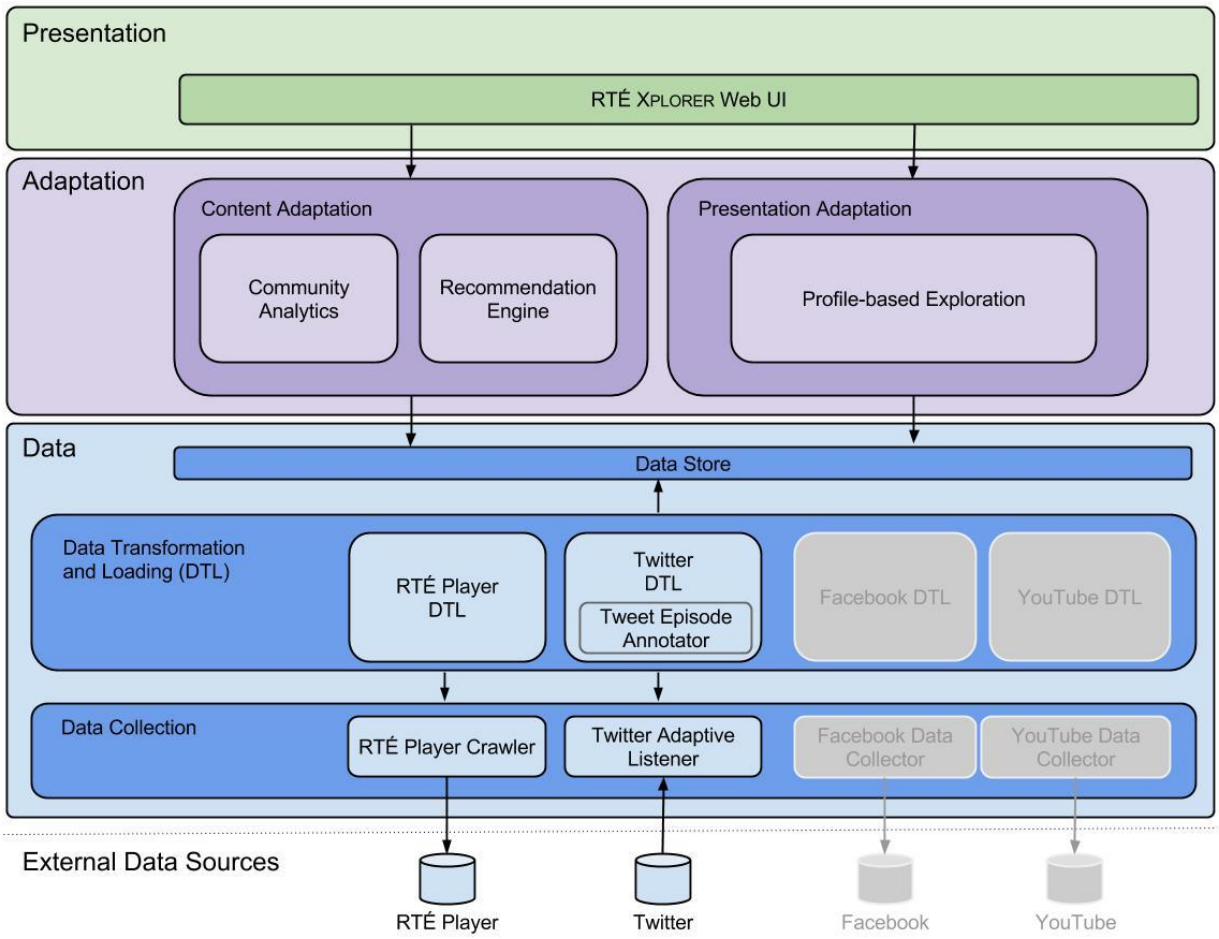


**Figure 22. RTÉ XPLORER Prototype Logical Architecture**

Next, we will focus on each of the components that belong to the presented layers.

### Data Layer

- *Data Collection*: collects useful data from external data sources (*i.e.,* RTÉ player, Twitter, Facebook, YouTube). For the initial phase the data sources that will be used would only be RTÉ player and Twitter. Data collection strategies are explained in *section 3.4.*

- *Data Transformation and Loading*: takes as input data collected from the external data sources, and carries out the necessary operations needed to update the Data Store. In order to do this, data must be transformed and processed to be later loaded to the Data Store.
- *Data Store*: manages all stored data and provides an interface that would enable access. A detailed description of the data captured in this component can be found in *section 7.1.2.*

**Adaptation Layer**

- *Content Adaptation*: holds components that carry out analysis over adaptation data (explained in *section 3.1.2*) in order to offer users a customized version of services adjusted to their context. More specifically, we propose the services related to Community Analytics and Recommendations for the RTÉ player.
- *Presentation Adaptation*: holds components that adapt/customize the way content is presented to the user. Specifically, we propose to tailor the navigational structure of content to the user characteristics.

**Presentation Layer**

- *RTÉ XPLORER Prototype Web UI*: Provides services through a web user interface along/within the RTÉ player website. These services are meant for the end-users of the RTÉ Player.

In this section, we have described the logical architecture of the RTÉ XPLORER prototype. Details on the Data Layer can be seen in both *section 3.4* and *section 7.1.2.* Details on the presentation layer can be seen in *section 7.2.* For this reason in this chapter we will focus on presenting the Adaptation Layer. Thus, in *section 6.2* we will offer special focus to the Presentation Adaptation component and in *section 6.3* to the Content Adaptation component.

## 6.2 Presentation Adaptation

Presentation Adaption refers to customizing the way content is presented to the user according to their contextual characteristics. Specifically, we propose to tailor the navigational structure of content tailored to user and social media data.

Our overall goal is to offer different views of RTÉ content that can support *Product Catalogue Exploration* which would allow users to better browse items from the product catalogue. Exploration views serve to showcase the most interesting parts of the RTÉ catalogue. However, the way views are presented could help users find, in less time, potentially interesting items.

In *Table 3,* we highlight different views that support product catalogue exploration. The Presentation Adaptation component organizes these views according to the user's browsing history and according to what is happening in social media. More specifically, we first identify programmes the user has interacted in his browsing history, and use their content features as input to re-rank the display of the proposed views. Also, the popularity of programmes in social media would influence the ordering of views. In this way, items the user would most probably like would be placed first considering the user's tastes and overall item popularity.

| Type of View | Description |
|---|---|
| ***Because You Watched*** | This view offers programmes given the user's past browsing actions. If the user has watched a given programme, then other related programmes are suggested in this view.<br>For this, the view uses the Recommendation Engine component explained in *section 6.3.2.* |
| ***Most Discussed*** | This view offers programmes that have been the most discussed in social media (primarily on twitter, *i.e.,* most trending programmes). The word popularity is used to refer to items that are the most talked about, or have generated the most buzz/hype. The programmes would be ordered according to popularity.<br>Information on how to determine a programme's popularity in social media can be obtained from statistics over the recollected tweets. We can define popularity of a programme in many ways. In this case, we determine that most popular/trending programme are those that in the last days have had the most action (tweets and retweets) from the most people. |
| ***Most Discussed Category*** | This service will showcase the programmes from the most trending category. For example, if the category Drama is the most popular |

| | |
|---|---|
| | (*i.e.,* Most Discussed Category – Drama), we would like to give spotlight to Drama programmes ordered according to popularity.<br>We define the most popular/trending category as those that in the last days have had the most action (tweets and retweets) from the most people. Action in this case comes from aggregating twitter actions over all programmes that belong to the category. |
| ***Most Discussed New Programmes*** | This view highlights new programmes that are starting to become popular in social media. It could be interesting to separate programmes that have had long term popularity from newer programmes that are becoming popular. This because new programmes that are recent risers might still not generate enough tweets to compete with the number of tweets generated by the older popular programmes, and thus will not make it to the Most Discussed list. However, users are more likely to know the old popular programmes and it would be more beneficial to introduce users to programming that will become popular in the future.<br>This view can be generated observing the trends of tweets for new programmes in different time windows. |
| ***Irish Exclusive*** | This view highlights exclusively Irish programming. It is important to offer a special *section* for Irish shows to help the RTÉ audience identify faster the programmes that are nationally produced.<br>We believe that Irish have an inherent sense of national pride and are very interested on what is happening in their country. For this reason, offering a special space for Irish programming could incentivize audiences to explore programmes that they might not have even known are exclusively made in Ireland. This view would also highlight the differential items the RTÉ Player service offers compared to other similar online television services. |
| ***One time runners*** | This view would highlight programmes that are running once. It can happen that users might not find out about these type of programmes during the time frame the programme is available on the RTÉ Player service. For this reason, it is important to suggest these programmes to users and make them aware of the |

| | |
|---|---|
| | programming they can access.<br>The difference between One time runners and recurrent shows, is that a user is already use to coming back to the RTÉ Player service to watch their series. In comparison, users might not become aware of One time runners and could miss a programme that if they had known about would have watched.<br>The order of the list could be according to expiry date, giving priority to one time runners that expire sooner.<br>One time runners can be obtained from data captured from the RTÉ Crawler. If the programme is new, but has only one episode, no next broadcast date and no season information, then we assume it will run once. |
| ***Last time to see*** | This view would highlight programmes that are expiring soon. It can happen that users are not aware of expiry dates for programmes that they would like to watch and then miss the programme. We propose that the *Last time to see* view should highlight items that are going to be removed from the RTÉ Player service in the following days (*e.g.,* in the following three days). The order of display could be according to the expiry date of the item, giving priority to those that expire sooner.<br>*Last time to see* are programmes that can be obtained from information captured by the RTÉ Crawler and the expiry date. |
| ***Most Recent*** | This view would showcase newest or recently added programmes. This is different than showcasing the newest episodes to be added for programmes that have other episodes available in the player service. Another way to describe the service, would be to add programmes for which there have been no episodes available in the recent past. In this way, series that have been on a break and a new season has started, could be showcased in the New Programming service. The needed information to provide this view is captured by the RTÉ Crawler. |
| ***Coming Up*** | Because the RTÉ product catalogue is constantly changing, promoting upcoming programming is advisable so users are aware |

| | |
|---|---|
| | of what items they can access in the future. This service would showcase programmes that will be added in the future to the RTÉ Player service.<br>This information should be provided by RTÉ. |

**Table 3. RTÉ XPLORER Prototype Exploration Views**

Among other presentation adaptation ideas to explore in future works we highlight:

- *Customized Emails:* It would be interesting to let users know about the programmes mentioned in *Table 3* through other communication channels (*e.g.,* email or social media), that do not depend on the user visiting the RTÉ Player site. It is possible that if a user doesn't need to return to the RTÉ Player service (*e.g.,* the user's television series is over), then they will not find out about other interesting parts of the product catalogue. Customized emails can be sent to users: reminding users if their favourites shows are being broadcasted soon, recommending shows to watch, letting users know that their favourite shows are about to be expired and they haven't watched them, among others.
- *Exploration based on semantic knowledge:* If products could be semantically annotated, users would be able to explore the product catalogue in different ways. For example, users could search for a product that is: a movie, which belongs to the category drama, where a particular actor participates and that is less than two hours long. With semantic content features programmes can be related to other semantically annotated external resources (*e.g., RTÉ Archive*). In general, more content features about programmes could open the doors for new search tools for users and also new views of the product catalogue.
- *Customized programme sites:* Customized sites could also be created for important programmes offered by the RTÉ Player. The goal would be to offer users a quick sense of background and environment surrounding the programme. A clear example can be offered for sports. For sport programming it is not only important to highlight information on current games, news and scores. It is also important to offer summary on the history of previous games played by the involved teams, give users the sense of why a

particular game is important, best players and plays, engage users with questions and polls to generate discussion, offer game preview and post-game information. In this way, users could feel closer to the teams and sports they are passionate about. Correspondingly, this could be equally important for television series. Information such as an overview of the main characters, plot twists and discussions on social media may encourage the user to become interested on a programme. The overall goal is to offer each user the information they would be most interested on about a programme, and encourage the user to engage himself with what is currently happening in social media around the programme. The purpose is to generate interest towards the events surrounding the programme, a deeper connection and encourage participation. Moreover, users are interested in the things they are observing in the video: people, clothes, products, scenarios, among others. It is would be interesting to offer users answers to questions such as: what sport gear their favourite player is wearing?, what jacket is a certain actress wearing?, where was that scene shot?, what restaurant was that scene shot?, among others. Projects such as the LinkedTV project [35] take a big step towards contextualizing television content and connecting it to other web resources.

In this section, we have presented different views that would support product catalogue exploration for the RTÉ Player service. The presentation of these views is customized to the characteristics of users and their social context. In the following *section* we will present services related to Content Adaptation.

## 6.3 Content Adaptation

Content Adaptation refers to providing users a customized version of services adjusted to their context. In this section, we propose services related to Community Analytics and Recommendations for the RTÉ player.

### 6.3.1 Community Analytics

In order to incentivize user engagement with social media, we propose to offer widgets customized to the social media characteristics of programmes. Specifically, we have explained that programmes can belong to communities and how to identify

these communities in *section 4.3.* These widgets would display the most interesting characteristics of communities. We propose the widgets shown in *Table 4* to be shown related to a specific programme.

| Widgets | Description |
|---|---|
| ***Interesting Conversations*** | Identify for users potentially interesting conversations/discussions that are happening in Twitter, associated to a programme. In this manner, the user can better understand what other viewer's opinions are and also more efficiently engage in conversation with other viewers. Incentivizing conversation could lead to the creation of more viewer communities formed around programmes in Twitter. Towards this purpose, we can identify the most recent tweets in a community about the programme, the most retweeted tweets in the community and related tweets (*i.e.,* tweets that were found in the community but that are not necessarily related to the current programme). |
| ***Relevant Hashtags*** | Suggest the most relevant hashtags for users to use when tweeting about a given programme. In this way users will be encouraged to use the official hashtags for programmes and in addition use hashtags we have found that tend to co-occur with the programme and that are also relevant. Encouraging the use of hashtags that are previously known will help us collect better information from twitter. |
| ***Top Users*** | Highlight the most influential users from Twitter that are commenting on a given programme. These users could be interesting users to follow.<br>A user is influential if his/her activity contributes to the overall spread of information in the network, *e.g.*, tweets this user posts reach potentially multiple other communities/regions of the network. Top users are identified according to their PageRank centrality within the community sub-network. |
| ***Public Sentiment*** | This service summarizes the public's sentiment towards a programme. We would like users to be able to compare among |

| | |
|---|---|
| | three types of sentiment: Positive, Controversial and Neutral.<br>It must be noted that sentiment analysis over tweets is still an ongoing topic in research. Also, it might not always be beneficial to showcase public sentiment if it is negative. For these reasons, this feature is proposed for future work. |
| ***Live Tweets*** | To better allow users to interact with Twitter while watching programmes, this service displays the live tweets about a programme. In addition, users could tweet from within the RTÉ Player service and in this way interact with social media and receive responses in real-time. This feature would go very much in line with promoting the second screen phenomena. |
| ***Relevant Media*** | This service would show to the user the most popular media components that are being shared in Twitter. Media components could include images and links to other web resources. |

**Table 4. RTÉ XPLORER Prototype Community Analytics Widgets**

For future work, it would be interesting to map tweets to the specific episodes they are referring to, and not only to the programme in general. In this way, a timeline of what people have been discussing in twitter can be offered to new audiences so they can catch up on the social context surrounding the programme.

Also, it would be interesting to run topic analysis over tweets and show in a timeline which have been the “bursty” topics.

In this section, we have explained how we can use the communities identified in *section 4.3* to offer users services that would allow them to get a view on the social media context surrounding a programme. In addition, users could more efficiently interact with the communities identified through the proposed widgets. In the following *section*, we will discuss the topic of offering users programme recommendations.

### 6.3.2 Recommendation Engine

Recommendation Systems (RS) help users find in less time potentially interesting items that the user might not have found easily on their own. These services help users find faster niche maybe not so popular but interesting products.

In [7] we explored the challenges of using conventional Recommender System techniques given the unique requirements of the RTÉ case study. In addition, in [7] diverse Recommender solutions are proposed based on concepts such as Collaborative Filtering, Content-based Recommendation systems and Linked open Data.

In this section, we will layout the method chosen as the strategy for the Recommendation Engine and which was implemented in the prototype (*section 7*).

We propose a RS approach based on the data captured from Twitter. First, we establish that Twitter has information on *how RTÉ programmes are related to each other*. We adopt the notion that relatedness is a broader concept than similarity [9]. Thus, two items can be related without necessarily being similar (*e.g.*, car and wheel) [9]. In our case, we would like to determine if two programmes are related by virtue of their social connections without necessarily having to be similar in their content features.

In the context of Twitter, as a heuristic, we define that two programmes are related proportionally to the amount of times they are found together in different Twitter settings. Specifically, we define that two programmes are related if they satisfy one of the following types of co-occurrence: (a) inter tweet: both are mentioned in the same tweet, (b) intra tweet: both are mentioned by the same user in separate tweets, and (c) community based: both are mentioned in the same community –communities are defined in *section* 4.3–. An analysis of these co-occurrence settings has already been carried out in *section 4*.

The ***Recommendation Engine*** uses knowledge acquired from social media about programme relatedness, to offer a ranked list of *N* programmes, ordered in terms of relevance in relation to a given input programme. A programme is relevant if it could be interesting for the user considering his/her current context. Context is restricted to what the user is currently viewing and what is happening in social media only, as in our system users are anonymous and we do not have past browsing history.

The Recommendation Engine component uses an Item-Based Recommendation approach, founded on the concept of Conditional Probability-Based Similarity as formulated in [11].

We define $P(j|i)$ as the conditional probability of programme $j$ being relevant to the current context, given that programme $i$ has already been determined as relevant. Because we do not have user explicit input, we must rely on the assumption that if the user is currently watching a programme then it is relevant to the user's current context. As a result, given the current programme the user is viewing, the Recommendation Engine delivers the top-N programmes that have the highest probability of being relevant to the current programme being watched.

The key relies now in how to define $P(j|i)$. The formulation of $P(j|i)$ as defined by [11] is in *Equation 1.*

$$P(j|i) = \frac{frequency(i,j)}{frequency\ (i)}$$

**Equation 1. Conditional Probability**

We have argued that Twitter can offer information on the relatedness of *RTÉ* programmes based on three co-occurrence situations: inter-tweet, intra-tweet and community-based. If programmes co-occur in these situations, we resolve they are to an extent related or relevant to each other. Thus, from each situation we can derive a different conditional probability distribution to define $P(j|i)$. *Table 5* defines $frequency(i,j)$ and $frequency(i)$ for each type of co-occurrence situation.

| Co-occurrence Type | | $frequency(i,j)$ | $frequency(i)$ |
|---|---|---|---|
| *Inter-tweet* | $P_1(j\|i)$ | Number of tweets that mention both programmes $i$ and $j$. | Number of tweets associated to programme $i$. |
| *Intra-tweet* | $P_2(j\|i)$ | Number of times the same user, in separate tweets, mentions both programmes $i$ and $j$. | Number of tweets associated to programme $i$. |
| *Community-based* | $P_3(j\|i)$ | Number of communities associated to both programmes $i$ and $j$. | Number of communities associated to programme $i$. |

**Table 5. Types of programme relevance**

Furthermore, we can deal with popularity bias by separately scaling results with a value that depends on $P(j)$ [11]. We multiply each $P_k(j|i)$ by $-\log_2(P_k(j))$, inspired from the inverse-document frequency scaling.

Finally, to determine a final value for $P(j|i)$, we linearly combine the different sources of evidence with different weights as in *Equation 2*. In *Equation 2,* $K$ is the number of sources of evidence. In this case, $K = 3$ given each co-occurrence type.

$$P(j|i) = \sum_{k=1}^{K} (w_k \cdot P_k(j|i) \cdot -\log_2(P_k(j)))$$

**Equation 2. Combined Conditional Probability**

Currently weights are defined by business rules and empirical intuitions. For example, if it is determined that evidence from inter-tweet co-occurrence ($w_1$) should be the most influential over the final probability, then weight values have to comply with the following restrictions: $w_1 > w_2$ and $w_1 > w_3$ .

In the future, extensions to the formulation of $P(j|i)$ can be further achieved given different heuristics, for example based on the shared content or semantic features of programmes, or even on user explicit feedback (when it can be obtained). As a consequence, diverse information sources could be considered, such as linked open data and other social media platforms. Proposals towards these approaches where suggested in [7], but still leave ground for future work.

For future work as well, different types of products that can be recommended when knowing a user is watching a particular programme episode can be considered. Related content to the programme episode could include: news stories, gossip, items to buy, publicity, events, YouTube videos, clothes, among others. This service would help users find other content to look at about the programme. Usually users are interested in knowing more about the programme and the real-world environment that surrounds it.

In this chapter, we have presented the logical architecture of the RTÉ XPLORER prototype and offered special attention to services provided by the Adaptation Layer. In the following *section*, we will present the RTÉ XPLORER functional prototype.

# 7 RTÉ XPLORER PROTOTYPE IMPLEMENTATION

This section describes the technical aspects of the RTÉ XPLORER functional prototype. By means of the prototype, services proposed in this document are materialized in a proof-of-concept product.

The RTÉ XPLORER prototype is meant to be a tangible representation of how services proposed in this document could be integrated into the RTÉ Player service. It offers a preliminary view of the real-world application of services and a test platform that could be used in a controlled environment.

First we will define the design considerations that were taken in to account when developing the product. Subsequently, the technical design of the prototype is presented, including the logical architecture and data model. Finally, we will present the final product and functional tests that were carried out.

## 7.1 Prototype Implementation and Deployment

In this section, we first present the enabling technologies used to develop the RTÉ XPLORER prototype, within the context of the logical architecture. Next, we offer special focus to the data model that lays foundation to the prototype. Finally, we describe the deployment conditions for the prototype to run.

### 7.1.1 Frameworks and Technologies

In this section, we outline the technologies used to implement the prototype per each component of the logical architecture.

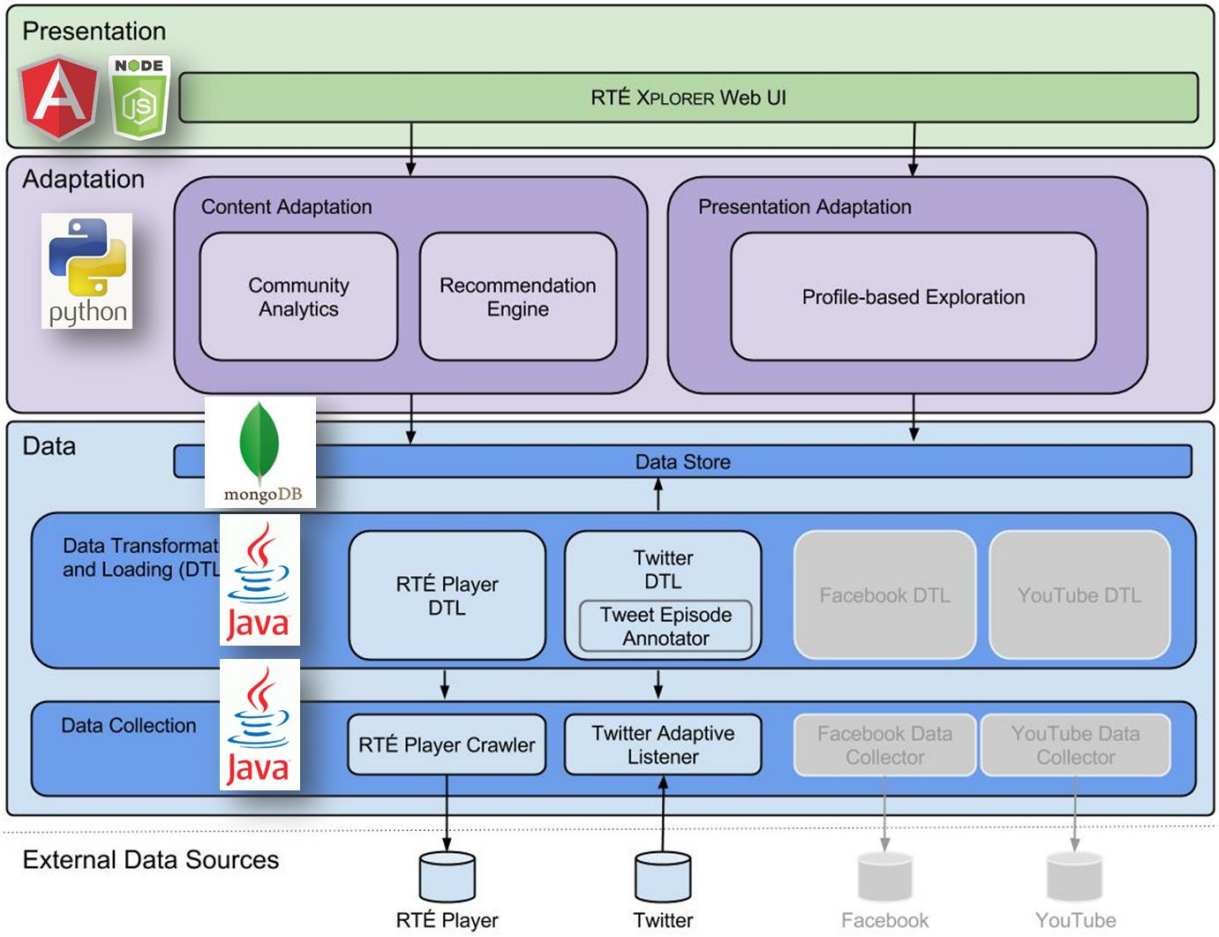


**Figure 23. Implementation Frameworks and Technologies**

The relation between technologies and the logical architecture can be seen in *Figure 23*. These technologies are:

- **Data Layer**: In the Data Layer, Java [29] was used to develop the RTÉ Player Crawler, RTÉ Player DTL, Twitter Adaptive Listener, and Twitter DTL components. The data store was developed on MongoDB [38]. In *section 7.1.2*, we explain why MongoDB was used and describe the data model implemented.
- **Adaptation Layer**: The components for the Adaptation Layer where developed in Python [44]. Python integrates a set of useful tools for analytics which were broadly used in the different components, such as *NetworkX* (library for graph-based algorithms), *NLTK* (natural language processing toolkit) and *Numpy* (high performance numerical library). In addition, Python easily integrates with the technologies used in the Presentation Layer.

- **Presentation Layer**: For the front-end of the RTÉ XPLORER prototype service, AngularJS [4] and Node.js [42] where used as the enabling technologies for the visual interface. These technologies make responsive MVC-style applications very easy to develop and deploy. For example, Node.js has an event-driven engine that makes our API design run efficiently. Moreover, communication to the Data Store component is rapidly achieved as all the involved technologies (AngularJS, Node.js and MongoDB) natively use the JSON data model.

In this section, we have outlined the technologies used to develop the RTÉ XPLORER prototype. In the following *section*, we will describe in further detail the data model of the Data Store component.

### 7.1.2 Data Model Implementation

We have selected MongoDB [38] as the enabling technology to deploy the Data Store component (view *section 6.1*).

MongoDB is a document-oriented storage engine designed for scalability and flexibility thanks to its schema-less approach. Documents in MongoDB are represented using an extended version of the JSON data model, widely used within the JavaScript language and as a data interchange format between remote services (*e.g.* REST and Streaming APIs). JSON allows for complex objects to be easily modelled and quickly parsed and generated. MongoDB can handle millions of documents with ease and are organized into collections that can be distributed and configured independently across a cluster of machines to maximize storage and availability. The engine offers fast indexing and adaptability when used with unstructured, rapidly changing data. All of the above features allow our system to conveniently integrate all of its components.

A description of the document collections that can be found in the prototype's database can be viewed in *Table 6*.

| Collection Name | Description |
|---|---|
| **analytics.twitter.communities** | Output community analytics. |
| **crawler.categories** | Categories used by RTÉ to classify programmes. |
| **crawler.episodes** | Information on television episodes published in the RTÉ Player service. |
| **crawler.programmes** | Information on television programmes published in the RTÉ Player service. |
| **curation.programmes** | Hashtags, keywords and users to follow associated to each RTÉ programme. |
| **listener.twitter.extended_hashtags** | Hashtags that frequently co-occur with seed terms, which are found by the *Adaptive Listener*. These are calculated each hour. |
| **listener.twitter.extended_keywords** | Keywords that frequently co-occur with seed terms, which are found by the *Adaptive Listener*. These are calculated each hour. |
| **listener.twitter.extended_users** | User mentions that frequently co-occur with seed terms, found by the *Adaptive Listener*. These are calculated each hour. |
| **listener.twitter.follow_suggestions** | Users to follow suggested by an external component which is not the *Adaptive Listener.* |
| **listener.twitter.grey_hashtags** | Hashtags that frequently co-occur with seed terms, which are found by the Adaptive Listener, but which are also trending topics. |
| **listener.twitter.seed_hastags** | Set of hashtags always used for listening to Twitter. |
| **listener.twitter.seed_keywords** | Set of keywords always used for listening to Twitter. |
| **listener.twitter.seed_users** | Set of user IDs always used for listening to Twitter. |
| **listener.twitter.statuses** | All collected Tweets by the listener (a "status" is a Tweet in Twitter's terminology.). |
| **listener.twitter.trending_topics** | Periodic trending topics in Ireland from Twitter. |

**Table 6. MongoDB Collections**

In this section, we have described the implementation of the data store component. In the following *section* we will outline the required conditions for the execution environment of the prototype.

### 7.1.3 Deployment Requirements

In this section we define the required technological infrastructure to run the RTÉ XPLORER prototype.

The prototype's external data sources include Twitter and the RTÉ Player service. These platforms need to be accessible via the Internet at all times for the normal operation of the system. In particular, a Twitter Application set of access tokens[1] is required to be obtained and configured to be able to connect to their Streaming API. The RTÉ Player service is accessed using a Web Crawler and access tokens are not required.

We assume that the RTÉ Player service is not constantly updated. However, when it is, our web crawler sends notifications so it can also be in turn updated to adjust to the new changes of the RTÉ Player service.

All the components of RTÉ XPLORER prototype are designed to be run under a Unix-compatible environment that provides the following application runtimes: (1) Oracle Java JRE version 8 or higher, (2) CPython interpreter version 2.7.9 or higher, and (3) NodeJS version 4 or higher.

In addition to the above runtimes, a working deployment of MongoDB version 3.0 or higher is required for the data store of the system. Collections can be deployed using a sharded (clustered) setup for better scalability and performance but it is not strictly required by RTÉ XPLORER prototype to work normally.

The front-end web service is run using NodeJS and includes its own integrated HTTP server for serving both, static and dynamic API requests. However, NodeJS is a single-threaded runtime and to achieve optimal performance the PM2 process management software[2] is highly recommended. A pre-configured multi-thread cluster mode settings file for PM2 is included with the server component for this purpose.

The all social analytics is designed to run at regular intervals with minimal memory requirements. However, the actual amount of memory and CPU necessary depends on the amount of Twitter data captured in each interval. Typically, a commodity-sized

[1] Can be requested using https://apps.twitter.com/
[2] More information: http://pm2.keymetrics.io

machine (single processor and 8G of RAM) is sufficient. For the community finding sub-system, the OSLOM algorithm version 2.3 or higher is required to be installed on the processing machine.

In this section, we have explained the implementation and deployment characteristics of the RTÉ XPLORER prototype. In the following *section*, we will describe the graphical interface of the demo.

## 7.2 Prototype User Interface

The RTÉ XPLORER prototype has two views: (*1*) ***Exploration View:*** the main landing page where the user can explore different programmes organized in tailored *section*s, *i.e.*, according to the user's viewing history and to social media, and (*2*) ***Video View:*** a page that offers all information related to a particular video content (*e.g.,* an episode of a particular programme). All video content is related to a programme (be it a single or multiple production programme). Information would include current communities around the related programme, the latest Tweets in Twitter and a set of related programmes.

Overall, the design of the prototype is inspired on the original RTÉ Player service. In this section, we will explain in detail both the Exploration View and the Video View.

### 7.2.1 Exploration View

The *Exploration View* offers different types of programme carrousels to allow a user to explore the product catalogue in a more efficient way. Given the carrousel design, users are offered diversity of content both horizontally and vertically. This augments the probability users can find faster interesting content and efficiently navigate a large product catalogue such as the one offered by RTÉ.

The Exploration view or main landing page can be seen in *Figure 24*. To the moment it has three core *section*s but it can be enhanced in the future to have more. These *section*s are "*Just For You*" (view *Figure 25*), "*Today In Social Media*" (view *Figure 26*) and "*More RTÉ Content*" (view *Figure* 27). The remainder of this section will offer a description for each landing page *section*.

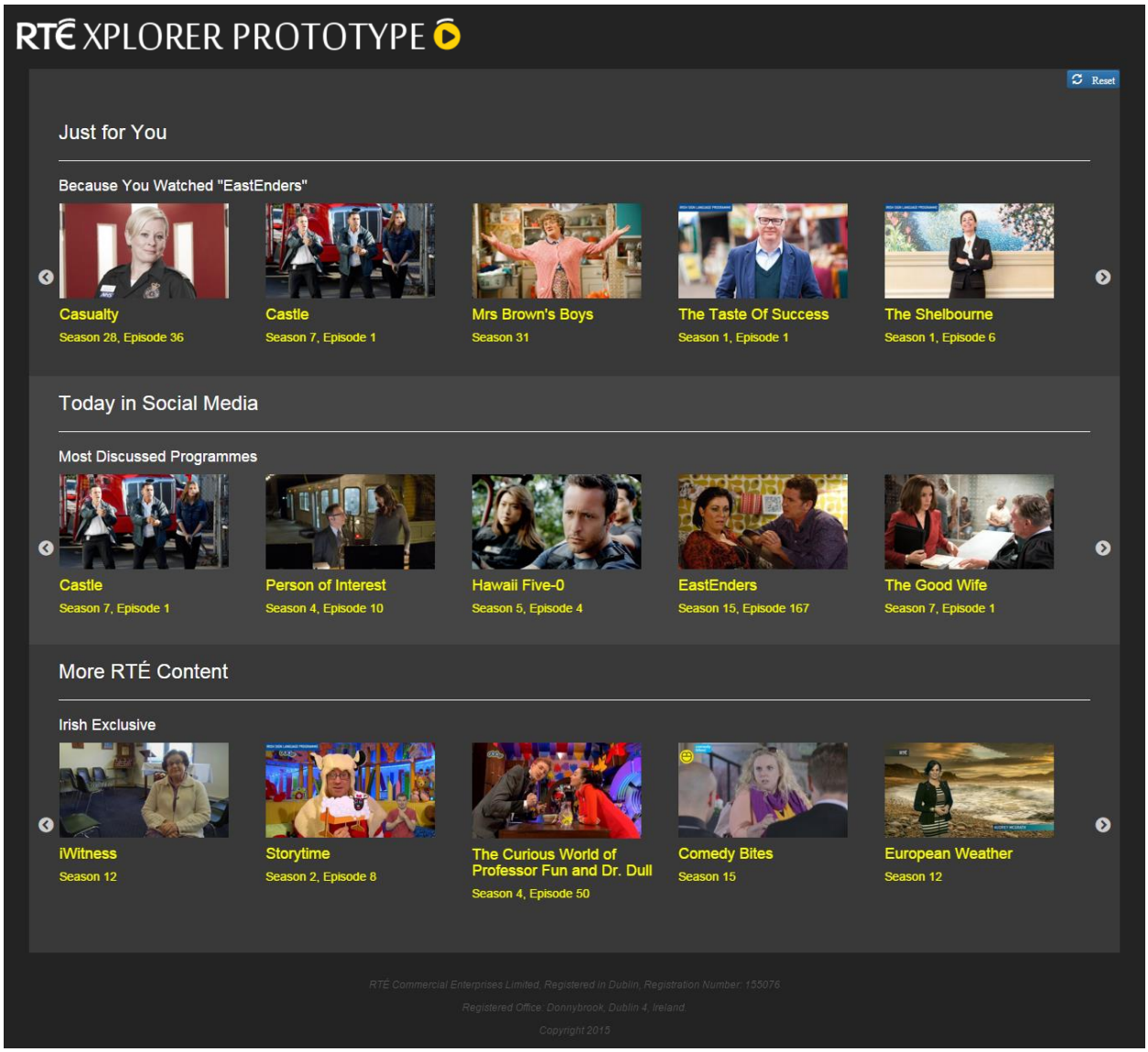


**Figure 24. Exploration View of RTÉ XPLORER Prototype**

The first part of the main landing page is the "*Just For You*" *section*, as shown in *Figure* 25*.* This section dynamically adjusts its contents depending on the current user's browsing activity within the RTÉ XPLORER prototype, *i.e.*, the different programmes the user visits. After user browsing-related carousels, a series of related programme carousels are generated for the user using the Recommendation Engine service, *i.e.,* "Because You Watched". For example the carousel in *Figure 25* (1) is for the programme "EastEnders" and the carousel in *Figure 25* (2) is for the programme "Person of Interest". In each carousel of all *section*s of the landing page, the user is able to click on any specific show picture or title as the one in *Figure 25* (3), to navigate to its detailed episode page or Video view. To improve the range of offered programmes to the current user, this section also generates carousels for the

most frequent categories of the programmes that he/she has visited, for example "More From Drama" viewed in *Figure 25* (4) because "EastEnders" is in that same category.

After some time of interacting with the system, the user will have a built profile of programmes that she visits and the most frequent categories from those. The user has the choice to restart this collection of profiling data using a reset button *Figure 25* (5). This button is added for the purpose of the prototype. In a real world application, this information could be managed by the user directly in his user profile.

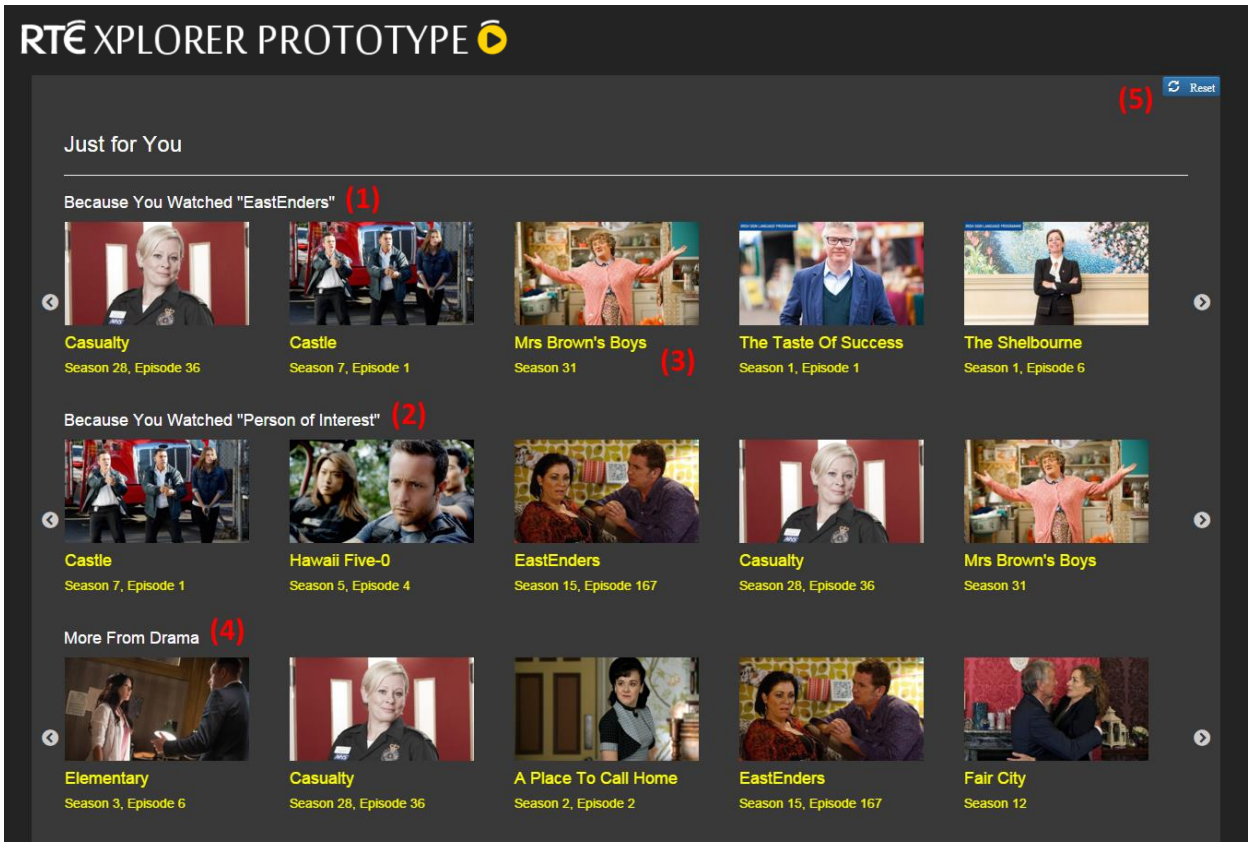


**Figure 25. "*Just For You*" *section* of the main landing page**

Right after the user tailored content, the Social Media based *section* shown in *Figure 26* as "Today in Social Media" is presented to the user. In this section, a carousel with the latest most discussed programmes (view *Figure 26* (6)) is generated together with one or more carousels containing programmes from the most discussed categories (view *Figure* 26 (7)). This panel offers the user dynamic spot-on insight of what programmes other users are discussing currently, regardless of the user's own viewing behaviour. Also, this functionality increases the chances for a

user to explore programmes that he/she hasn't considered, and are currently popular in social media.

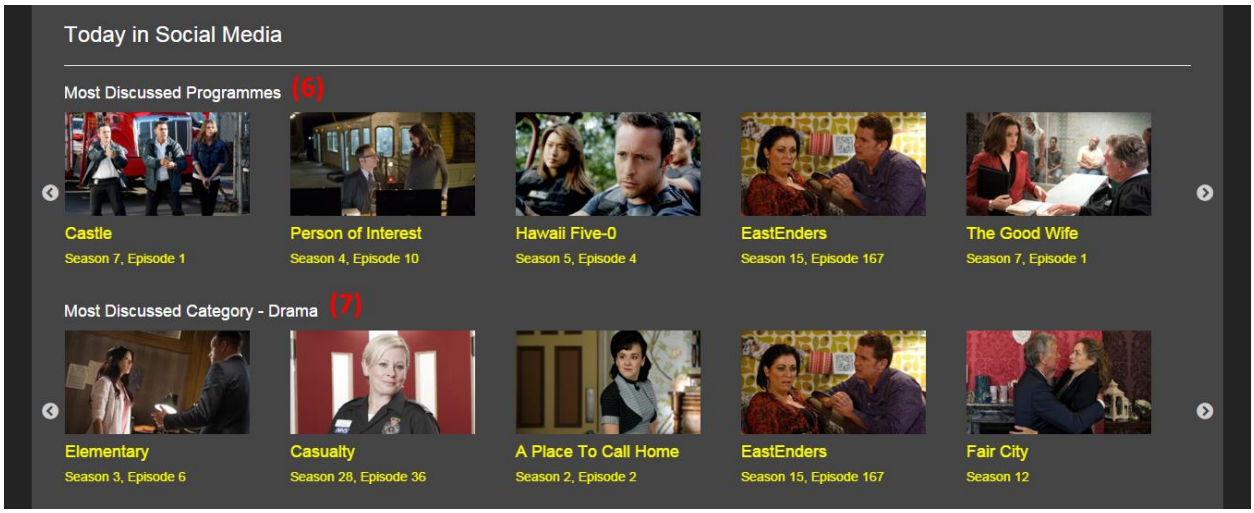


**Figure 26. "*Today in Social Media*" *section* of the main landing page**

Finally, the RTÉ XPLORER prototype presents the user a semi-static, *i.e.,* updated about once a day, "More RTÉ content" *section* purely generated using data from the RTÉ online catalogue. This section can be viewed in *Figure 27*. The purpose of this panel is to allow the user to discover content specifically produced by RTÉ such as *Figure 27*(8), most recently broadcasted as in *Figure 27*(9) and content about to expire soon from the online catalogue as in *Figure 27*(10). To further add variety to this section, the system provides programme suggestions from randomly chosen categories available from RTÉ, for example "Religious and Irish Language" as in *Figure 27* (11) and "Comedy" as it can be viewed in *Figure* 27 (12).

In this section, we have presented the Exploration View and offered an in-detail description of the different *section* it is composed of. In the following *section*, we will present the Video View.


**Figure 27. "*More RTÉ Content*" *section* of the main landing page**

### 7.2.2 Video View

In the Exploration View, the current user can click on any programme picture or title displayed inside any carousel of any *section*. This action takes the application to a dedicated *Video View,* which includes an embedded video player and displays other components such as metadata, social analytics details and related programmes, as shown in *Figure 28*. A clickable picture of the programme is presented in *Figure*

*28*(1) that allows the user to directly go to the Episode in the official RTÉ Player service.

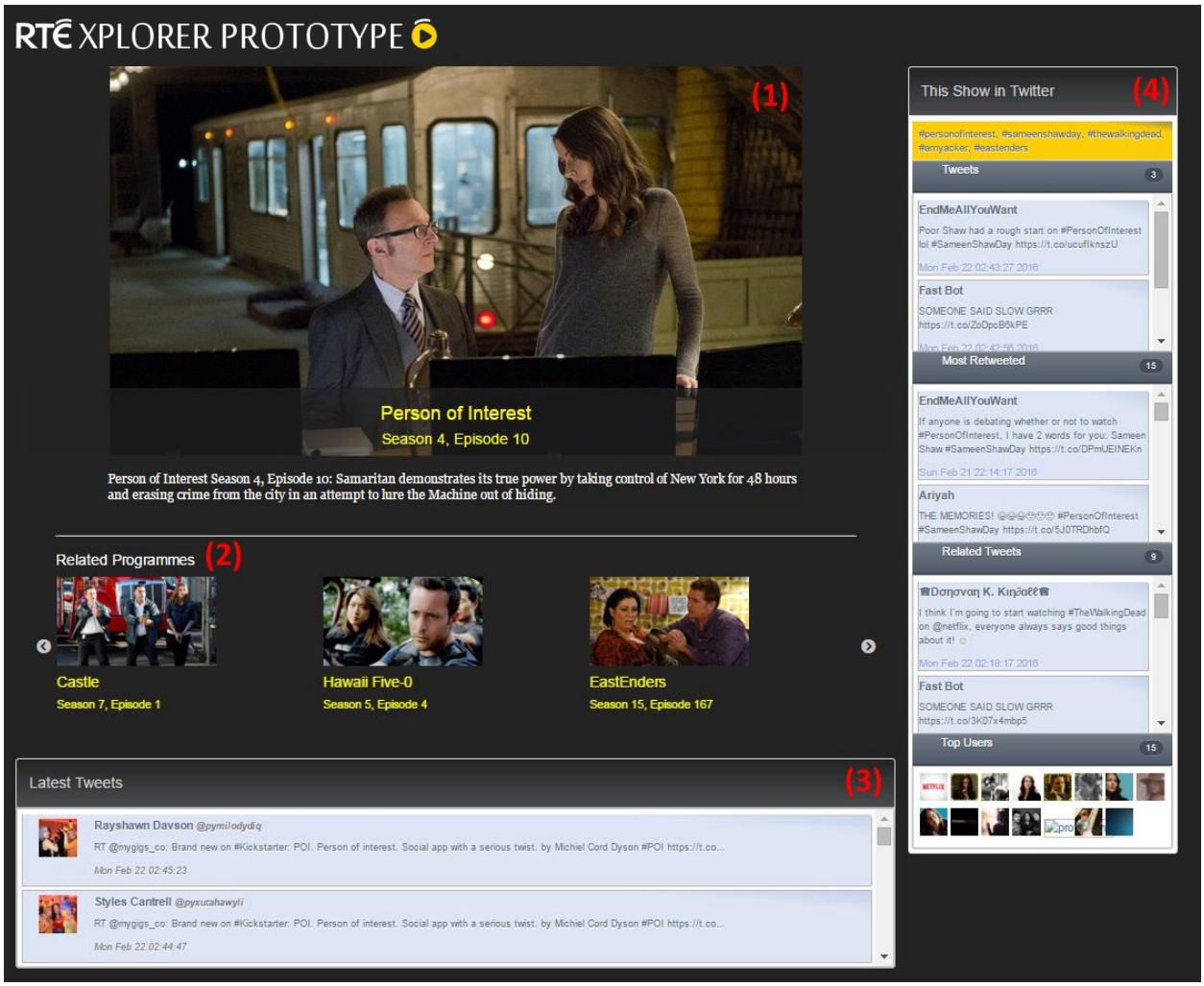


**Figure 28. Video View of RTÉ XPLORER Prototype**

Similarly as done in the "Just for you" *section* of the main landing page, the Video View provides a carousel of programmes related to the current video as in *Figure 28*(2). The items are also clickable. In case that there is not enough data to suggest related programmes, the system defaults to displaying more programmes of the same category.

At the bottom of the Video View, the user can see a feed of latest Tweets that are related to the programme being visited as seen in *Figure 28*(3). These Tweets come directly from the real-time Tweets capture component and are selected using their automatically assigned annotation. To the right hand side, the user is offered a panel

containing all the implicit communities discovered related to the show *Figure 28*(4). This panel includes different elements for each live user community discovered:

- **Hashtags:** displays the top most frequent hashtags used in the Tweets posted by the users of the community.
- **Tweets:** displays a scrollable list, in descending chronological order, of recent Tweets published by the users of the community that are exclusively related to the current programme. Retweets are not included in this list.
- **Most Retweeted:** displays a scrollable list also in descending order in time of the most retweeted Tweets generated inside the community.
- **Related Tweets:** displays a list of Tweets that were found in the community but that not necessarily are related to the current programme, *i.e.,* Tweets about related programmes being discussed at the same time by the community.
- **Top Users:** displays clickable profile pictures of the most prominent users of the found community. Top users are identified according to their PageRank centrality within the community sub-network. If a picture is clicked, the browser goes to the official Twitter profile page of the chosen community user, where it can be followed or further explored using Twitter's own user interface.

In this section, we have offered a complete presentation of the graphical interface of the RTÉ XPLORER prototype and of the services offered. In the following *section* a summary of the prototype implementation is offered.

## 7.3 Summary

In this section, we have described the technical aspects of the RTÉ XPLORER prototype. This demo, offers services based on both Social Analytics and Information Adaptation concepts. The overall goal is to offer RTÉ end users, services to support them in exploring the RTÉ product catalogue and understanding what is happening in social media related to RTÉ programming. In this manner, users can find faster interesting content and be encouraged to participate in social media communities discussing RTÉ content.

The prototype was developed to offer a view of how the integration of services proposed in this document would work in a real-world application. The developed

demo follows a logical architecture which has three layers: Data Layer, Adaptation Layer, and Presentation Layer. In this section, we have explained the technologies used for development and offered a detailed presentation of the RTÉ XPLORER prototype user interface.

# 8 CONCLUSION

This document has presented the project outcomes for the RTÉ project. In this section, we will present conclusion and future work.

## 8.1 Conclusion

The way we consume television today has changed. Similarly, the way we share our TV viewing experience has shifted towards the use of social media platforms. In this work, we aim to exploit the potential use of social media as a source of collective TV knowledge in order to enhance online television services and support the characteristics of modern TV viewers. Specifically, we focus on the case study put forth by Raidió Teilifís Éireann (RTÉ), which is the national public provider of television and radio in Ireland. The company broadcasts its content online through the RTÉ Player service. To support end users in exploring the RTÉ catalogue and understanding what is happening in social media related to RTÉ shows, we propose the RTÉ XPLORER prototype. Given our solution, users will be able to find interesting content faster, as well as participate in an enhanced and rich social experience around TV programmes. Furthermore, we presented the implemented RTÉ XPLORER prototype to RTÉ for consideration.

The **main contributions** presented include:

(1) *An in depth analyses of Twitter data in the context of RTÉ*

We defined a methodology for capturing and analysing live Twitter data based on the online catalogue of programmes available on the RTÉ Player service. In particular, we proposed a Web Crawler component that periodically retrieves the current programming being broadcasted on this catalogue and, after a systematic manual curation of Twitter terms for newly found shows, continuously keep a Twitter listener updated.

We captured and studied seven months of live RTÉ data from Twitter for 138 programmes available on the RTÉ Player service (77 exclusive and 61

non-exclusive to RTÉ), posted from over 100 countries. Our analyses were carried under two main perspectives: (1) only considering simple Tweets and (2) user behaviour based on discovered implicit communities. For the second perspective, we proposed a method for identifying user communities in Twitter discussing RTÉ content based on more basic Tweets interactions such as mentioning, retweeting and replying. For this we employ the state-of-the-art OSLOM algorithm for user networks.

We observed many different characteristics of the Tweets and communities such as sparsity, coverage, long-tail distributions, conversational reciprocity and popularity, programme co-occurrence in two levels (tweets and communities), geographical locations distributions, among others. With these metrics, we were able to identify interesting phenomena. For example, users seem mostly interested on tweeting about Drama programmes in the case of global shows, but lean towards News, Sports and Entertainment in the case of Irish-exclusive programmes. On the other hand, RTÉ accounts seem to promote Drama as well as Children and Entertainment programmes the most, which is aligned with the previous finding. Also the observations from our conversational analyses suggest that, despite a very big group of disconnected users, there are still groups of people that already are participating in discussion-like patterns with varying degrees. Those groups have the potential to be proactively motivated to join discussions about RTÉ programming and further enhance engagement.

Finally, our exploration of Twitter in respect to multiple programmes being discussed simultaneously revealed that users post content over an interesting variety of shows despite not covering all of the broadcasted content. More important than having a complete coverage of the programmes is to gain coverage of those shows that are most interesting and engaging for the users, based on the way the socially created connections among those programmes and their categories are related.

*(2) A tailored version of SociaLens for RTÉ*

SociaLens presents a dashboard composed of diverse widgets that offer different views/perspectives/representations of social data. These views would provide a visual aid to support decision making processes. Specifically, it would allow RTÉ

decision makers to better understand the behavior of social media users and their interaction with RTÉ content.

*(3) The portrayal of RTÉ XPLORER Prototype solutions*

The RTÉ XPLORER prototype is a system that provides a well-defined set of solutions which aim to enrich the RTÉ Player experience. These solutions were designed considering the conditions and restrictions depicted by the RTÉ case study. The most significant limitations are the lack of explicit user feedback (such as ratings) and user session data. This keeps us from maintaining a User Profile and also from applying traditional recommendation approaches. As a solution, we view a key opportunity in using Twitter as a source of collective knowledge to determine if RTÉ programmes are related, and in turn, use this knowledge to provide users with customized services tailored to the social media context.

The RTÉ XPLORER prototype system includes services to support content navigation, recommendations and community awareness. By offering an adaptive Exploration view of the product catalogue and by means of Recommendations, the RTÉ XPLORER prototype aims to help users find faster potentially interesting content. Furthermore, the RTÉ XPLORER prototype offers users a structured view of what is happening on social media associated to a programme. For this, Twitter communities are identified, and interesting discussions, media and users.

*(4) The development of a functional prototype*

The RTÉ XPLORER prototype was developed to offer a tangible representation of how the proposed RTÉ XPLORER services can be integrated into the RTÉ Player service.

There are numerous benefits for RTÉ resulting from this project. From these, we highlight the following:

- *Better understanding of RTÉ online users*: Using our social analytic approaches, RTÉ can further understand their online audience. RTÉ can observe the way users participate in social media communities, what and how they discuss about RTÉ related content and how their content is related in terms of the way users discuss about programmes.

- *Increase of user engagement with the RTÉ Player site*: We propose Information Adaptation services that can be added to the RTÉ Player service based on the programme preferences extracted from social media. An example of such services would be tailored recommendations, which can be offered to provide users with programme suggestions to incentive them to further explore less-known *section*s of the RTÉ product catalogue.
- *Increase of user engagement in social media about RTÉ*: Using outputs from social analytics, we propose services that RTÉ can add to their RTÉ Player service so users are encouraged to participate in social media. Such services include a view of current discussions about RTÉ programming, most influencing users and comments, among others. If users generate more information on social media, then more information can be captured to further enhance Information Adaptation services.

We believe that the RTÉ project presents a representative case study to explore the potential use of microblogging data to enhance Information Adaptation services. In this document, we have proposed a set of novel approaches that can be used for any online TV Player. As well, our method to study social media can be used with other annotated Twitter datasets.

## 8.2 Future Work

The richness of this case study allows for plenty of room for future work. Along the chapters of this document a few have already been mentioned. From possible directions we highlight the following areas and proposals:

- ***Evaluation*:** A more formal evaluation of our solution can be done using standard user testing. These tests can be carried out in a controlled environment using the developed RTÉ XPLORER prototype. In addition, A/B testing strategies could be used to temporarily test our recommendation approach on a live setting. In particular, for the Recommendation Engine, we could evaluate if placing a recommendation service incentivizes users to access less popular long tail products. This can be done in measuring the popularity of niche products before and after the recommendation service is placed in to action.

- ***Community Finding*:** Our approach for mining implicit communities can be improved further. For example, OSLOM's sensitivity can be tuned using different parameters. For those we used values obtained empirically, but they could also be automatically learned by more complex machine learning methods. On the other hand, there are alternative algorithms such as RankClus [48] that can also discover community structures. However, in contrast to OSLOM, this algorithm can consider Tweet data directly when clustering. This can be achieved because RankClus uses a heterogeneous network clustering approach, as opposed to OSLOM which only works with homogenous graphs.
- ***Community Tracking*:** Our current community detection approach mines Twitter data at regular intervals in independent processing windows, however same communities are known to be persist longer in time. Using community tracking methods such as the one from Greene et al. [18], a generic set of user communities can be studied across advancing time periods using well-defined behaviours such as birth, split, merge, intermittence and death. Furthermore, this would allow for studying the evolution of user discussions over time not only about a single show, but also a set of them simultaneously.
- ***Information Spread in Twitter*:** Users share information in Twitter mainly using the Retweets mechanism. However, Twitter does not directly allow for observing how the Retweets multiply and spread in the network. One way to circumvent this limitation is to superimpose the network of followers of the involved retweeting users to estimate who saw the original tweet or retweets from whom. With this estimation, we can build Cascade Trees of the retweets spread across users and compute a variety of graph-based metrics to identify influential, i.e. central, users and the content that goes farthest inside the social network.
- ***Entity Extraction*:** Our current analytics solution only takes Twitter #hashtags in consideration for labelling communities and as a part for annotating programmes. Despite hashtags being explicit entities provided from users in Tweets, they also provide other potential implicit entities such as actors, locations and organizations names, along with other higher-level entities such as dates. These implicit entities could be extracted using methods such as the

fine-tuned extractors from Derczynski et al. [33] and further used in our solutions to improve the annotation of programmes and their underlying connections.

- ***Sentiment Analysis*:** Thelwall et al. [51] have an extensive experience and research regarding the extraction of sentiments from Tweets. Sentiments are identified based on different textual features such as adjectives, negations, nouns, emoticons, among others and rated using a [-1, 1] range in any individual Tweet. With this information, it is possible to create more in-depth analysis about how users are reacting to different RTÉ programmes, and even profile them accordingly. On another example, it could be possible to automatically provide prompt live feedback to showrunners according to users liking or disliking parts of a show while it is being broadcasted.
- ***Recommendation Engine*:** In reference to our recommendation approach, we could extend data sources that can offer evidence on programme relatedness (*e.g.,* using linked open data sources such as Wikipedia). For this, multiple relatedness metrics have been proposed in the literature (*e.g.* Hulpus et al. [27]) to relate different objects according to their semantic similarity and distance. On another note, in previous works it has been shown that explanations help users understand why they are being recommended products and increase user's trust on recommendations. Future work can also focus on different techniques for offering explanations.
- ***Personalization*:** A new range of research topics opens if access to user profiles is granted, such as user historical transactions, viewing habits, and basic demographic data. User profiles would serve to propose personalized services customized to the individual user, and not just adaptive services tailored to a group of users, such as our current approach. Given personalized services, users could feel more identified with the RTÉ Player by receiving services that adjust to the unique user preferences, characteristics and context.

# 9 REFERENCES


[1] "An Integrated Approach to TV & VOD Recommendations." TV Genius. (2011).

[2] Ali, Kamal, and Wijnand Van Stam. "TiVo: making show recommendations using a distributed collaborative filtering architecture." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[3] Amatriain, Xavier, and Justin Basilico. "Recommender Systems in Industry: A Netflix Case Study." Recommender Systems Handbook. Springer US, 2015. 385-419.

[4] AngularJS: https://angularjs.org/ Last Access: 24th March, 2016

[5] API Rate Limits: https://dev.twitter.com/rest/public/rate-limiting. Last Access: 27th October, 2015.

[6] Bambini, R., *et al.* (2011). A recommender system for an IPTV service provider: a real large-scale production environment. In Recommender systems handbook

[7] Barraza-Urbina A., *et al.* "Using social media data for online television recommendation services at RTÉ Ireland." (2015).

[8] Boyd, Danah, Scott Golder, and Gilad Lotan. "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter." System Sciences (HICSS), 2010 43rd Hawaii International Conference on. IEEE, 2010.

[9] Budanitsky, Alexander, and Graeme Hirst. "Evaluating wordnet-based measures of lexical semantic relatedness." Computational Linguistics 32.1 (2006): 13-47.

[10] Chan, Jeffrey, Conor Hayes, and Elizabeth M. Daly. "Decomposing Discussion Forums and Boards Using User Roles." ICWSM 10 (2010): 215-218.

[11] Deshpande, Mukund, and George Karypis. "Item-based top-n recommendation algorithms." ACM Transactions on Information Systems

(TOIS) 22.1 (2004): 143-177

[12] Dey, A., Abowd, G. Towards a Better Understanding of Context and Context-Awareness. In: Gellersen, H. (eds): Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing (HUC 1999) (Karlsruhe, Germany, September 27-29, 1999), vol. 1707. Lecture Notes in Computer Science, Springer-Verlag, Berlin-Heidelberg (1999), pp. 304-307.

[13] Engeström, Jyri. "Why some social network services work and others don't—Or: the case for object-centered sociality." Blog post]. In: Zengestrom. < http://www.zengestrom.com/blog/2005/04/why_some_socialhtml (2005).

[14] Freda-Marie Hartung. Social Curiosity and Its Functions. PhD Dissertation. University of Konstanz. 2010.

[15] Geerts, D. (2010). The Sociability of Mobile TV. In A. Marcus, A. C. Roibás, & R. Sala (Eds.), Mobile TV: Customizing Content and Experience (pp. 25–28). Springer London

[16] Georgios Paliouras, Symeon Papadopoulos, Dimitrios Vogiatzis and Yiannis Kompatsiaris. User Community Discovery. Springer, 2015.

[17] Granovetter, Mark S. "The strength of weak ties." American journal of sociology (1973): 1360-1380.

[18] Greene, Derek, Donal Doyle, and Padraig Cunningham. "Tracking the evolution of communities in dynamic social networks." Advances in social networks analysis and mining (ASONAM), 2010 international conference on. IEEE, 2010.

[19] Gupta, Aditi, Anupam Joshi, and Ponnurangam Kumaraguru. "Identifying and characterizing user communities on twitter during crisis events." Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media. ACM, 2012.

[20] Guy, Ido. "Social recommender systems." Recommender Systems Handbook. Springer US, 2015. 511-543.

[21] Harboe, G. (2010). Introduction to Social TV. In A. Marcus, A. C. Roibás, & R. Sala (Eds.), Mobile TV: Customizing Content and Experience (pp. 21–24). Springer London.

[22] Harboe, Gunnar, et al. "Ambient social tv: drawing people into a shared experience." Proceedings of the SIGCHI Conference on Human Factors in

Computing Systems. ACM, 2008.

[23] Holanda, Pedro, *et al.* "TV Goes Social: Characterizing User Interaction in an Online Social Network for TV Fans." Engineering the Web in the Big Data Era. Springer International Publishing, 2015. 182-199.

[24] How Netflix is turning viewers into puppets. Available at: http://www.salon.com/2013/02/01/how_netflix_is_turning_viewers_into_puppets/ Last seen: 4th February 2016

[25] Hromic, H., Karnstedt, M., Wang, M., Hogan, A., Belák, V., & Hayes, C. (2012). Event panning in a stream of big data. In LWA Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML).

[26] Hsu, Shang H., et al. "AIMED-A personalized TV recommendation system." Interactive TV: a Shared Experience. Springer Berlin Heidelberg, 2007. 166-174.

[27] Hulpuş, Ioana, Narumol Prangnawarat, and Conor Hayes. "Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation." The Semantic Web-ISWC 2015. Springer International Publishing, 2015. 442-457.

[28] Java, Akshay, *et al.* "Why we twitter: understanding microblogging usage and communities." Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007

[29] Java: https://java.com Last Access: 24th March, 2016

[30] Kaplan, Andreas M., and Michael Haenlein. "The early bird catches the news: Nine things you should know about micro-blogging." Business Horizons 54.2 (2011): 105-113.

[31] Kim, Mucheol, et al. "IPTV Contents Recommender System Based on a Social Network." Embedded and Multimedia Computing Technology and Service. Springer Netherlands, 2012. 123-127.

[32] Kwak, Haewoon, *et al.* "What is Twitter, a social network or a news media?."Proceedings of the 19th international conference on World wide web. ACM, 2010

[33] L. Derczynski, A. Ritter, S. Clarke, and K. Bontcheva. 2013. "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data". In Proceedings of the International Conference on Recent Advances in Natural Language

Processing, ACL.

[34] Lancichinetti, A., *et al.* "Finding Statistically Significant Communities in Networks." PLoS ONE 6.4 (2011): e18961.

[35] Linked TV Project: http://www.linkedtv.eu/
Last Access: 24th March, 2016

[36] Lu, Xin, and Christa Brelsford. "Network structure and community evolution on twitter: human behavior change in response to the 2011 Japanese earthquake and tsunami." Scientific reports 4 (2014).

[37] Mitchell, Keith, *et al.* "Social TV: The impact of social awareness on content navigation within IPTV systems." Computers in Entertainment (CIE) 9.3 (2011): 19.

[38] MongoDB: https://www.mongodb.org/ Last Access: 24th March, 2016

[39] Naaman, Mor, Jeffrey Boase, and Chih-Hui Lai. "Is it really about me?: message content in social awareness streams." Proceedings of the 2010 ACM conference on Computer supported cooperative work. ACM, 2010.

[40] Nathan, Mukesh, *et al.* "CollaboraTV: making television viewing social again." Proceedings of the 1st international conference on Designing interactive user experiences for TV and video. ACM, 2008

[41] Nielsen Twitter TV Ratings: http://www.nielsensocial.com/product/nielsen-twitter-tv-ratings/ Last Access: 24th March, 2016

[42] Node.js: https://nodejs.org/en/ Last Access: 24th March, 2016

[43] Pripuzic, Kresimir, et al. "Building an IPTV VoD recommender system: An experience report." Telecommunications (ConTEL), 2013 12th International Conference on. IEEE, 2013.

[44] Python: https://www.python.org/ Last Access: 24th March, 2016

[45] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[46] Shamma, David A., Lyndon Kennedy, and Elizabeth F. Churchill. "Tweet the debates: understanding community annotation of uncollected sources."Proceedings of the first SIGMM workshop on Social media. ACM, 2009

[47] SociaLens: http://socialens.insight-centre.org/ Last Access: 24th March, 2016

[48] Sun, Yizhou, *et al.* “Rankclus: integrating clustering with ranking for heterogeneous information network analysis.” Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. ACM, 2009

[49] Tang, Lei, and Huan Liu. "Community detection and mining in social media." Synthesis Lectures on Data Mining and Knowledge Discovery 2.1 (2010): 1-137.

[50] The Search API: https://dev.twitter.com/rest/public/search.
Last Access: 27th October, 2015.

[51] Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

[52] Tinati, Ramine, et al. "Identifying communicator roles in twitter." Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012.

[53] Wohn, D. Yvette, and Eun-Kyung Na. "Tweeting about TV: Sharing television viewing experiences via social media message streams." First Monday 16.3 (2011).