My-TRAC

My TRAvel Companion.

Deliverable D2.2

Model for analysing a user’s trip purpose (activities)

Project funded by the European Union’s Horizon 2020 Research and Innovation Programme (2014 – 2020)

Contract No. H2020 – 777640

Deliverable 2.2 Model for analysing a user's trip purpose (activities)

Due date of deliverable: 31/07/2018

Actual submission date: 12/08/2018

Start date of project: 01/09/2018 Duration: 36 months

Dissemination Level		
PU	**Public**	X
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

Document Control Sheet

Deliverable number:	**D2.2**
Deliverable responsible:	Model for analysing a user's trip purpose (activities)
Work package:	WP2
Main editor:	Evangelos Mitsakis

Editor name	Organisation
Eleni CHALKIA	Centre for Research and Technology Hellas/ Hellenic Institute of Transport
Charis CHALKIADAKIS	Centre for Research and Technology Hellas/ Hellenic Institute of Transport
Pablo CHAMOSO	Universidad de Salamanca
Anastasis DROSOU	Centre for Research and Technology Hellas/ Information Technologies Institute
Panagiotis IORDANOPOULOS	Centre for Research and Technology Hellas/ Hellenic Institute of Transport
Lucía Martín GÓMEZ	Universidad de Salamanca
Matina LOUKEA	Centre for Research and Technology Hellas/ Hellenic Institute of Transport
Evangelos MITSAKIS	Centre for Research and Technology Hellas/ Hellenic Institute of Transport
Alexandros ZAMICHOS	Centre for Research and Technology Hellas/ Information Technologies Institute

Modifications Introduced			
Version	Date	Reason	Editor
1.0	15/12/2017	First draft version including literature review	Evangelos Mitsakis, Charis Chalkiadakis, Panagiotis Iordanopoulos
1.1	21/02/2018	Additional input on Chapter 2	Evangelos Mitsakis, Eleni Chalkia
1.2	26/03/2017	Input on Chapter 3	Anastasis Drosou, Alexandros Zamichos
1.3	25/05/2018	SMNs crawler and mapping (USAL)	Pablo Chamoso
1.4	14/06/2018	Additional input on Chapter 2	Matina Loukea
1.5	18/06/2018	Additional input on Chapter 3	Anastasis Drosou, Alexandros Zamichos
1.6	21/06/2018	Additional input on Chapter 3	Anastasis Drosou, Alexandros Zamichos
1.7	22/06/2018	First edition of the Deliverable	Evangelos Mitsakis, Charis Chalkiadakis, Panagiotis Iordanopoulos
1.8	26/06/2018	1st Quality review on preferences analysis models	Ismini Stroumpou, Alexandros E. Papacharalampous
1.9	27/06/2018	Input on Chapter 2	Matina Loukea

2.0	28/06/2018	1st Quality review	Joan Guisado-Gamez
2.1	23/07/2018	Integration of partners' feedbacks Input on Chapter 2	Matina Loukea
2.2	24/07/2018	Integration of partners' feedbacks Input on Chapter 3	Pablo Chamoso, Anastasis Drosou, Alexandros Zamichos
2.3	25/07/2018	2nd Quality review on preferences analysis models	Eleni Vlachogianni, Ismini Stroumpou
2.4	26/07/2018	2nd Quality review on activity prediction mechanism	Joan Guisado-Gamez
2.5	27/07/2018	Changes on Chapter 2 according to the 2nd Quality review	Charis Chalkiadakis, Evangelos Mitsakis, Panagiotis Iordanopoulos
2.6	27/07/2018	Changes on Chapter 3 according to the 2nd Quality review	Anastasis Drosou, Alexandros Zamichos
2.7	31/07/2018	Final Quality review	Ismini Strompou, Joan Guisado
2.8	02/08/2018	Changes on Chapter 3 (Section 3.6) according to the final Quality review	Lucía Martín Gómez
2.9	09/08/2018	Final release	Evangelos Mitsakis

Legal Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The above referenced consortium members shall have no liability to third parties for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law. © 2018 by My-TRAC Consortium.

Executive Summary

This Deliverable aims at the design and development of i) a strategy for modelling user profiles based on their tweets and ii) an algorithm for predicting activities for the users, as the core of a activity recommending system. The design of the algorithm is based both on literature review of already existing algorithms for the collection of transport-related data and on the proposed attributes, which can affect the preferences of the users.

The report begins with the state of the art within this deliverable reviews the literature on transport-related data that can be derived from social media platforms, as well as the already existing and implemented transport-related data mining models. There is a variety of transport data that can be collected and analysed from social media platforms. The majority of the data mining models, which have been found during the literature review, share the following four main modules:

- Search for data and pre-process
- Elaboration
- Classification
- Geo-location

Another important aspect taken into account in the present Deliverable, is that of the factors/attributes, which may affect the users' activities preferences under certain circumstances. In this document, there is thorough discussion on travellers' characteristics, activities' preferences and travellers' preferences; factors which, to a greater or lesser extent, affect users' choices related to transportation systems.

Following the literature review, the design and development of an activity prediction mechanism is presented. The proposed mechanism collects transport-related data from social media platforms' APIs, analyses the data and finally provides information regarding the users' activities preferences. The developed mechanism is based on the main points of the literature review. The procedure for the design of the activity prediction mechanism is based on information that users themselves share in a public way. Further to the collection of data from the Twitter API, text mining techniques are applied to the collected tweets, in order to recognize the activities, through the development of two unsupervised methods. The fusion of the outcomes of the aforementioned methods leads to information about the category of the activity executed. After the identification of the activities of each user, the activities are ordered and activity sequences for each user are formed. Finally, the processing of user specific information (e.g. favourite tweets) leads to the user profiling and by using a Markov chain model to the prediction of the user's next activities.

Abbreviations and Acronyms

My-TRAC	My TRAvel Companion
PTra	Public Transport
AI	Artificial Intelligence
HMI	Human-Machine Interface
MaaS	Mobility-as-a-Service
O-D	Origin-Destination
PII	Privately Identifiable Information
SUM	Status Update Message
id	User's Identification
LoS	Level of Service
API	Application Programming Interface
RTDC	Real Time Data Collection
HDC	Historic Data Collection
CDR	Call Detail Records
URL	Uniform Resource Locators
SVM	Support Vector Machines
LR	Logistic Regression
RF	Random Forests
GIS	Geographical Information Systems
TAZ	Transportation Analysis Zone
ICF	International Classification on Functioning and Health

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

My TRAvel Companion (My-TRAC) is a project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 777640. The main goal of My-TRAC is to develop a novel transport services platform for the rail sector, designed for public and private transport users and operators in order to provide an improved passenger experience by developing and applying advanced behavioural and transport analytics and Artificial Intelligence (AI) algorithms to meaningful data gathered from diverse public transport and Open Data sources. My-TRAC aims to deliver an innovative application for seamless transport and an ecosystem of models and algorithms for Public Transport (PTra) user choice simulation, data analytics and affective computing. My-TRAC stands out from other technologies due to three main reasons. First, My-TRAC fosters unprecedented involvement of users during, before and after a trip through a smart Human-Machine interface (HMI) and numerous functionalities such as crowdsourcing, group recommendations, data exchange. Second, the application implements a vast array of technologies, such as affective computing, AI and user choice simulation that fuse expertise from multiple fields. Third, My-TRAC facilitates engagement of multiple stakeholders by seamlessly integrating services and creating connections between Rail operators and other PTra providers.

My-TRAC application applies behavioural analytics and AI techniques, in order to provide a seamless door-to-door experience that suggests solutions and available options when they are necessary during a journey. The existence of a user-based data provisioning module can provide data for application operators, as well as connecting with external services such as booking, ticketing and analytics modules. The trip tracking services can help in the guidance of the passengers/ users of the application by selecting the best means available in various scales and various provided information (e.g. smaller-scale information regarding the optimum exit from the train station, or larger-scale information concerning the availability of free car-parking lots in a parking area). The development of My-TRAC social market service enhances the interactions between the transport providers and the passengers during trip, offering additional products and services (e.g. discounted transport, leisure activities and Wi-Fi access). The Advanced Human Machine Interface, which will be developed in the framework of My-TRAC project will adapt the travel companion services with the most suitable interface to match users' profile, preferences and specific accessibility needs.

1.1 SCOPE AND STRUCTURE OF THIS DELIVERABLE

This Deliverable, Deliverable 2.2: Model for analysing a user's trip purpose (activities) aims to design and develop an algorithm for the collection of information as for My-TRAC application users' activities preferences. The design of the algorithm is based both on the literature review of the already existing algorithms for the collection of transport-related data and on the proposed attributes which can affect the preferences of the users.

This Deliverable is structured in 4 sections. Section 1 is an introduction to the My-TRAC project and Deliverable 2.2. Section 2 presents a detailed literature review on the already existing algorithms and methodologies for the collection and analysis of transport-related data from social media platforms. Also, in Section 2, based on the literature, the main attributes which may affect the users' activities preferences are provided. Section 3 deals with the design and development of the activity prediction mechanism. The proposed method is based on literature review and also integrates the attributes, which affect the activities preferences. Various algorithms, (which are described in detail in the relevant sections below), have been used for processing Status Update Messages data, in order to provide credible information. Finally, Section 4 highlights the most relevant conclusions of this document.

1.2 RELATION TO OTHER DELIVERABLES

This Deliverable is highly related to other Deliverables of My-TRAC project, as it provides a mechanism for collecting and analysing data from social media (Twitter), in order for the users activities preferences to be provided as the final outcome of the developed mechanism. Next follows a brief description of the relation of Deliverable 2.2 and of its outcome with the other Tasks of the My-TRAC project.

Deliverable 2.2 is complementary to the outcome of Task 2.1. Deliverable 2.2, in combination with Deliverable 2.1, provide a complete framework for the understanding of factors which affect users' choices and assists on identifying the dependencies between these factors. Deliverable 2.2 is also related with Deliverable 2.3. Tasks 2.1 and 2.2 provide a complete framework for the understanding of a user's trip by predicting activities and understanding which individual's factors affect user's choices in order to propose alternate routes. In Task 2.3 a demand and assignment model will be developed for simulating and predicting users' choices and behaviour concerning mode choice, station choice, departure time choice and route choice and incorporate the model presented in Deliverable 2.2. There is also a high interrelation between Deliverable 2.2 and the Deliverable of Task 2.4 (Deliverable 2.4). The factors identified both in Deliverables 2.1 and 2.2 can provide a key input for Task 2.4 in order for the identification of users and activities relevance, in the reputation algorithms of Task 2.4.

. The traveller behavioural characteristics, including social media data, as derived from WP2 will be used in Task 3.2 in order for an integrated data collection and structuring framework to be designed and developed. The findings of Task 2.2, and of WP2 in general, provide an input of importance for Task 3.3. Task 3.3 has as a main goal the user customised and relevant advice and information provision. The findings of WP2 can assist in the development of a framework for the provision of the desired information in the users of My-TRAC application.

The findings of WP2 will provide a significant input in the development of the HMI, in the framework of Task 4.1. Regarding the development of an overall HMI, the users' specific needs as defined in WP2 will assist in the development of tools and services which will provide the users with viable access, depending on their specific needs. The design of the HMI will receive as inputs: the functional description of tools and services, and the specific needs and habits of the addressed user groups from WPs 2 and 3. The design will target the specific features of each user group, as defined in WP2 that may affect the way participants will interact with My-TRAC platform. These will include access to consumer technologies, language skills, social habits and cultural factors. Based on the user needs and application scenarios identified in WP2 as well as the functional specifications from WP3, a thorough requirement engineering approach will be conducted. User interfaces and adequate information strategies for all user groups and beneficiaries of My-TRAC system will be specified and conceptualised, taking into account the urgency, the frequency and the safety relevance of the information/ warning.

Finally, the data collected in WP2 will contribute in the development of the My-TRAC application and the developed mechanisms will be integrated in the application. In Task 5.2 the analytics results of WP2 will contribute in the implementation of the functionalities for individuals and the implementation of the general design as it will be provided by Task 3.1. The outcome of Deliverable 2.2 will also contribute in the framework of Task 5.3 regarding the proper implementation of all the group functionalities allowing for the collaborative analytics and recommendation algorithms investigated both in WP2 and WP3. Finally, there is an indirect connection between WP2 and Task 5.5. This connection is sourced in the fact that Task 5.5 will coordinate with WP2 and WP3.

2 LITERATURE REVIEW

The following sub-sections aim at providing a literature review regarding the transportation data that can be derived from social media, as well as an extent review on the already developed and applied data mining models for social media. This literature review aims at providing the necessary information regarding the range of the data that can be retrieved from social media and the prospective that transportation-related data collection and further analysis may have.

2.1 TRANSPORTATION DATA THAT DERIVE FROM SOCIAL MEDIA CONTENT

During the last years several new applications and services have been developed that, according to M. Parameswaran (2007) [1], "facilitate collective action and social interaction online with rich exchange of multimedia information and evolution of aggregate knowledge". These applications tend to dominate the Web and they are referred to under a variety of terms such as Web 2.0[1], online communities and social computing [1].

Social media (also referred to as Social Networking Services or Social Networking Sites) are the product of Web 2.0 and their usage is growing due to the increased use of smart phones and tablets [2]. Facebook [, Twitter [] and Foursquare are among the most widely used applications/websites of this category.

A number of research activities based on data deriving from social media have been conducted in scientific disciplines such as social sciences, economics, politics, tourism etc. [6]. Especially for the transport sector, the most commonly exploited information by social media platforms, until now, is based on the use of the spatial information accompanying posts (geotagged information) and the language processing of posted content [6].

The most common uses and applications of data obtained through social media, as stated in E. Chaniotakis et al. (2016) [6], are the following:

1. Identification of spatial and temporal mobility patterns.
2. Investigation of the applicability of the social-media originated data for travel demand modelling.
3. Identification of user activities.
4. Definition of urban settings and related characteristics (such as points of interest, boundaries, and land uses).
5. Investigation of rider satisfaction.
6. Examination of the relationship between social networks and mobility.

In the field of transportation research, social-media originated data can be used to [6]:

1. Identify the mobility patterns of a population.
2. Identify zones and boundaries of cities.
3. Derive Origin-Destination (O-D) matrices.
4. Explore users' social networks and their effect on transportation-related behaviour.

The data derived from social media platforms, such as those mentioned before, are of high importance in the transportation field. A study conducted by T.H. Rashidi et al. (2017) [2] identifies the main advantages and disadvantages in the use of social media data as a supporting tool in the transport sector, which are described next.

[1] According to Wikipedia, Web 2.0, also called Participative (or Participatory) and Social Web, refers to World Wide Web websites that emphasize user-generated content, usability (ease of use, even by non-experts), and interoperability (this means that a website can work well with other products, systems, and devices) for end users

T.H. Rashidi et al. (2017) [2] indicate that there seems to be significant potential in using social media data for developing models related to estimation of travel demand, managing operation and long-term planning purposes. Taking into account that the number of social media users is growing, the sample could be considered as a close representation of the population. Another advantage that makes social media data appealing is the acquisition cost. Nonetheless, the total cost of having useful data for modelling purposes is significantly larger than the acquisition cost. Also, social media data encompasses information revealed by the users in realistic situations. This means that the data is unbiased. Last but not least, such data are (in most cases) associated with geolocation information, which is valuable for transportation planning and management purposes [2].

Despite the advantages mentioned above, there are some disadvantages in the use of social media data for transportation related purposes. The most challenging issue is associated with the techniques and the methodology required for extracting useful information from the content of the data. When such data is used for demand estimation purposes, it should be adjusted for over-representation of system users. Furthermore, such data can be over-representative for discretionary and leisure activities. The result of these drawbacks is that the activities should be inferred. Thus, the acquisition cost of social media data is free, but the processing cost of the derived information is still quite high. Another important concern regarding social media data is that of individual-specific information. Access to such information (Privately Identifiable Information - PII) is restricted and any analysis conducted on social media data requires careful attention to aggregate the geotagged information of people so that it is not identifiable [2].

2.1.1 Data Mining from Social Media Platforms

The most common use of the social media platforms is for the users to post messages/ Status Update Messages (SUMs). The posted SUMs may contain, except of text, meta-information such as user's identification, timestamps etc. The SUMs referring to a certain topic may provide, after proper analysis of their content, information about an event or a topic [7], [8].

The data from those applications is usually referred as "Social Signal Data" and it is characterized by big volume, wide spatial coverage, long observational period and real-time features. Nevertheless, these characteristics may vary, depending on the source of information. This information can then be analysed and finally travel patterns can be discovered by applying data mining techniques in these large data sets [7].

However, when considering the semi-automatic and automatic techniques to collect such data (and especially the automatic techniques) two features have to be taken into consideration. The first feature is that the available information is either structured or unstructured (80% of the data posted in social media is unstructured, i.e. free text SUMs). The second feature is that social media text is often ungrammatical (typographical errors, uses-specialized language etc.) [9].

According to the nature of social media and the examples of its use in a variety of domains, three characteristics of social media content may be identified [9]:

1. Content created by an individual usually refers to an event that the individual has experienced.
2. The event or action commented on occurs either shortly before or shortly after the time at which the content is created.
3. The issue raised is of importance to the individual.

These three characteristics form the basis for the goals in harvesting transport-service related information. This includes transport modes, physical facilities and services. The goal of harvesting is therefore to find three types of transport-related information [9].

The transport-related information [9] might be:

1. Information on travellers' journey needs.

2. Detection of an irregular event that has an impact on mobility.
3. Travellers' opinions on the quality of a transport service.

Such information can serve as the basis for at least three types of actions that can be taken by transport planners and operators, and it is therefore important, in terms of policy development and delivery [9].

The three types of actions [9] are:

1. Creating a new service or enhancing an existing one.
2. Undertaking an ad-hoc solution for a problem reported through social media.
3. Improving the Level of Service (LoS) of an existing service.

In order to develop a methodology for achieving the set goals, there are two hypotheses [9], [10] concerning the potential of social media for transport policy:

1. Social media contains valuable information for transport planning and management.
2. Such information can be harvested either automatically or semi-automatically.

The extracted information must be characterized by high volume, velocity and variety in order to require specific technologies and analytical methods for its transformation into value. This means that the information should be highly relevant to the whole concept and complete [9], [11].

There are three challenges, which are of importance in the context of transport and need to be taken into consideration. These challenges are summarised below [9], [11]:

1. Analysis of text: Social media text is often ungrammatical
2. Representative sampling: Social media content is spontaneous and may represent only a subgroup of the target population.
3. Integration of text and geographic information: Geographic information about a user assists in modelling and interpretation, but may be unspecified.

The following figure shows the features, analytical approaches and applications of social signal data [7].

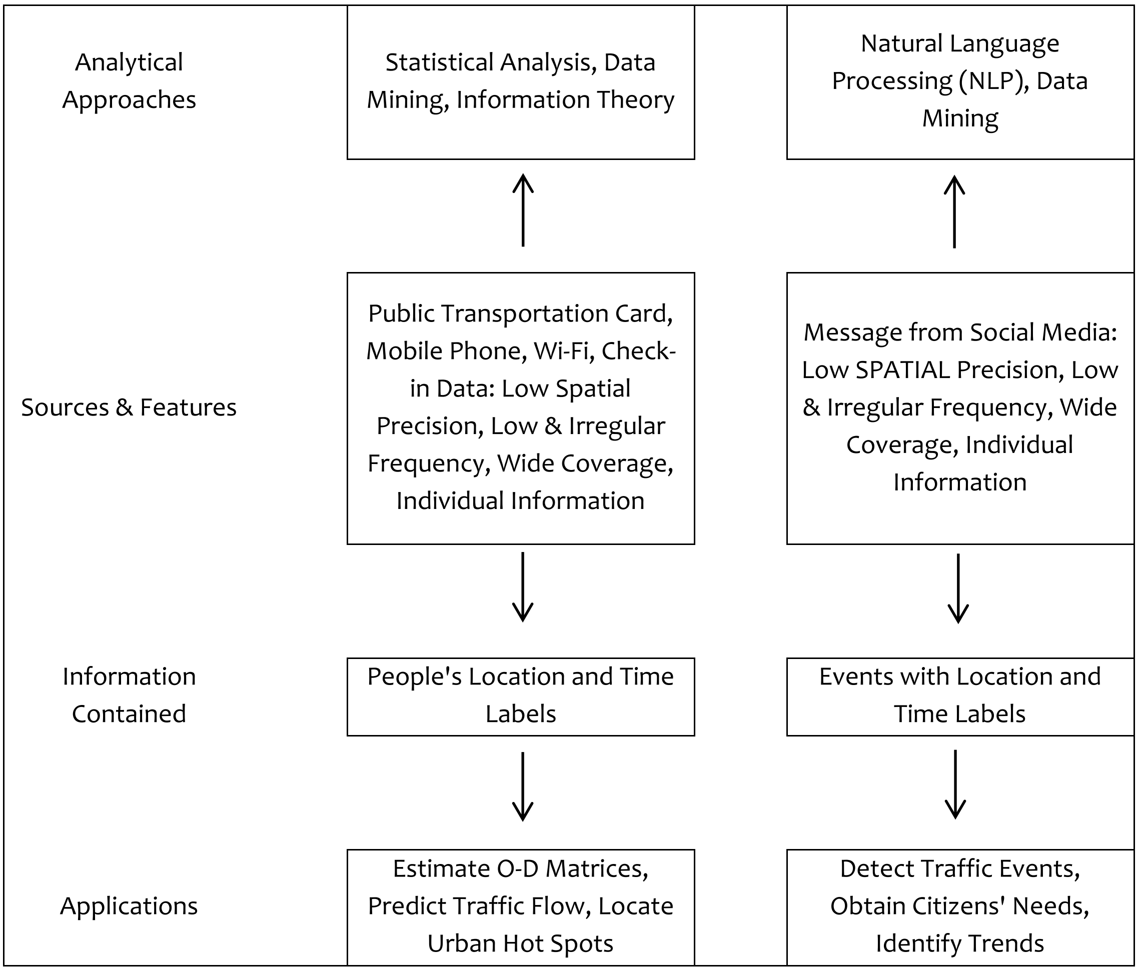

Analytical Approaches
Statistical Analysis, Data Mining, Information Theory
Natural Language Processing (NLP), Data Mining
Sources & Features
Public Transportation Card, Mobile Phone, Wi-Fi, Check-in Data: Low Spatial Precision, Low & Irregular Frequency, Wide Coverage, Individual Information
Message from Social Media: Low SPATIAL Precision, Low & Irregular Frequency, Wide Coverage, Individual Information
Information Contained
People's Location and Time Labels
Events with Location and Time Labels
Applications
Estimate O-D Matrices, Predict Traffic Flow, Locate Urban Hot Spots
Detect Traffic Events, Obtain Citizens' Needs, Identify Trends

Figure 1: Features, analytical approaches and applications of social signal data

The following figures (Figures 2, 3 and 4) [9], [10] illustrate the process through which the SUMs are analysed.

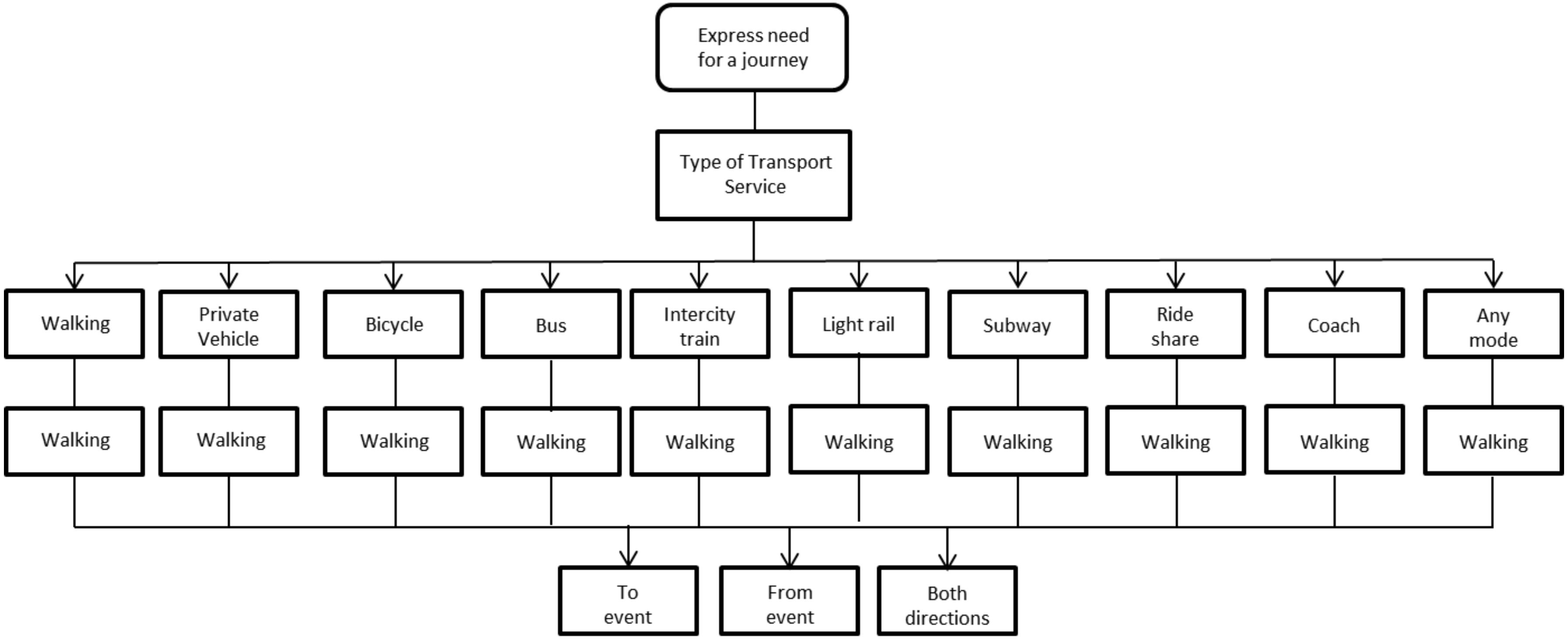


Figure 2: Process for the analysis of the Status Update Messages (SUMs) (a)

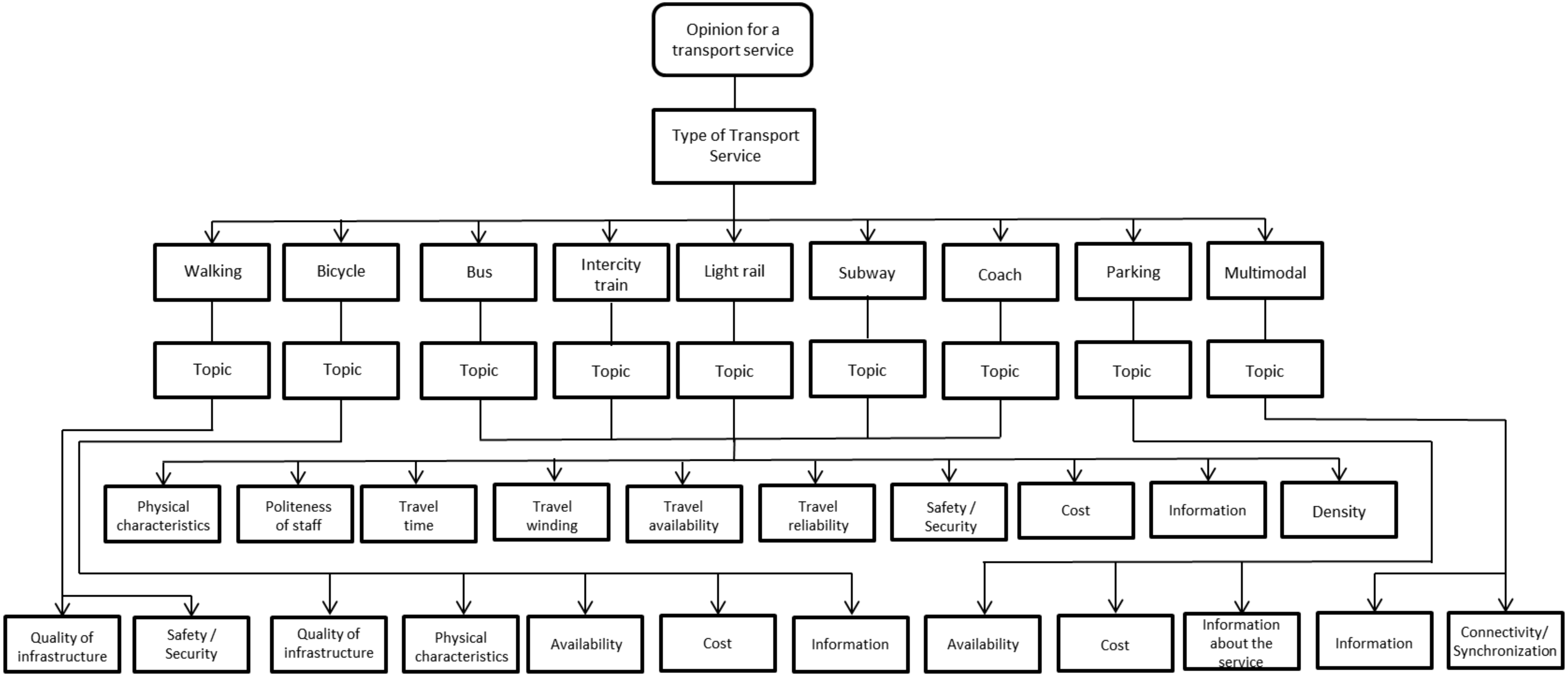


Figure 3: Process for the analysis of the Status Update Messages (SUMs) (b)

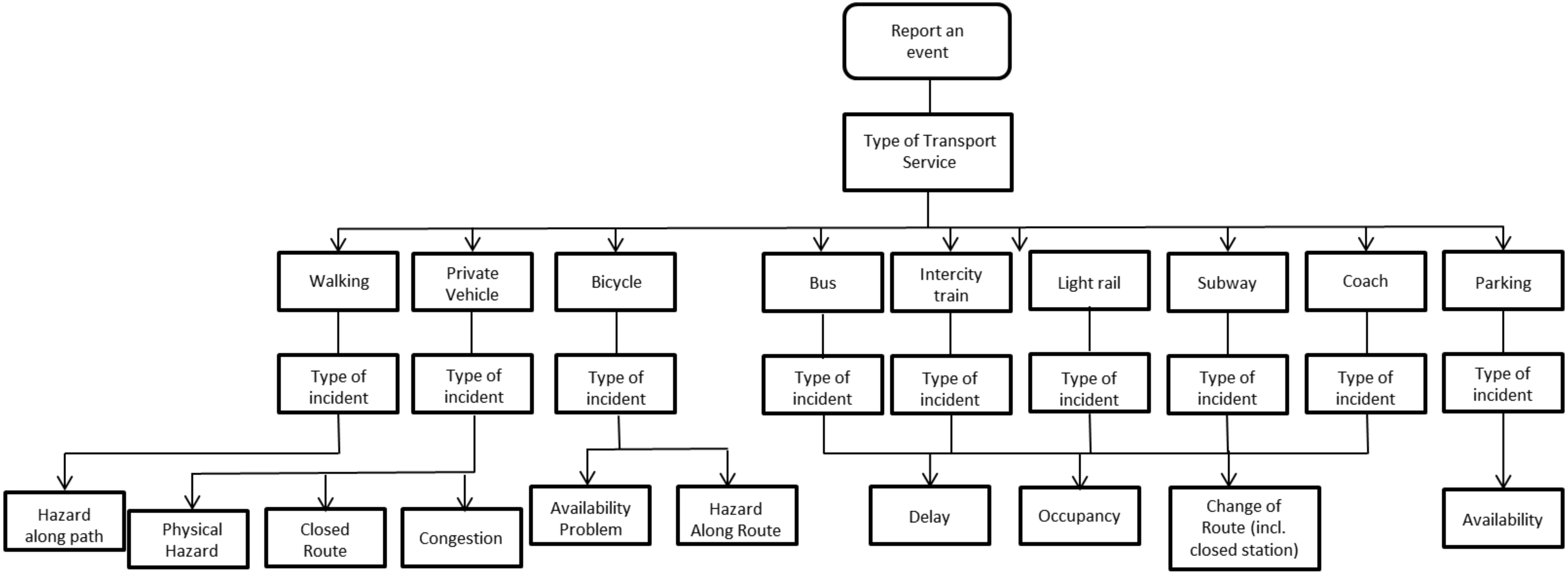


Figure 4: Process for the analysis of the Status Update Messages (SUMs) (c)

2.2 Data Mining Models for Social Media Data

Subsection 2.2 presents a review on the models used in studies for deriving information through social media. In general, the used models are service-oriented and event-driven. The models used for data mining purposes collect SUMs from social media and then they process the collected SUMs by applying text mining techniques in order to assign the appropriate class label to each SUM. Later, the models analyse the classified SUMs. Additionally, there is the capability to add geographic information in the SUMs so for the information to be geolocated.

The general pattern followed by the majority of the social media data mining models, as found in the literature review, consist of the following four main modules [8]–[10], [12]–[15]:

1. Search of SUMs and Pre-processing procedures
2. Elaboration of SUMs
3. Classification of SUMs
4. Geo-location

Below there is a brief description of the aforementioned modules [8]–[10], [12]–[15].

1. Search of SUMs and Pre-Processing Procedures

The first module extracts SUMs based on one or more search criteria (e.g., geographic coordinates, keywords appearing in the text of the tweet, etc.). The collection of data originates from social media applications and websites APIs. Each fetched raw data contains information such as the user id, the timestamp, the geographic coordinates and the text itself.

A two–phase data collection methodology could be applied, where first data is conventionally collected by performing API queries. This first data collection phase allows the creation of a user database. After the creation of the database, the SUMs are pre-processed. In order to extract only the text of each raw SUM and remove all meta-information, a Regular Expression filter is applied. Regular Expression is defined as the sequence of string characters that define a search pattern [16]. Finally, a case-folding operation is applied to the texts, in order to convert all characters to lower case. At the end of this procedure, each fetched SUM appears as a string sequence of characters.

2. Elaboration of SUMs

The second module aims to transform the set of pre-processed SUMs in a set of numeric vectors to be elaborated by the next module (Classification of SUMs). To this aim, text mining techniques are applied to the pre-processed SUMs.

3. Classification of SUMs

The output of this module is a collection of labelled SUMs with Road, Start point, and End point to classify traffic information. In order to label each SUM, there is a need to use a classification model under specific parameters. Link information is the traffic information that has all the attributes: Road, Start point and End point. Point information is the traffic information that has Road and Start or Stop point attribute or only Road or Start point attribute or Stop attribute.

4. Geo-location

The last step in that methodology is the geolocation of Start point and End point attributes in the text messages. Geolocation is a Data Enrichment technique, which helps the visualization of the traffic information on the map. This information can derive from applications and websites such as Foursquare.

The data mining models for social media data include two main components: data collected from the users through the Twitter's Application Programming Interface (API) [17] or other platforms' APIs (Real Time Data Collection - RTDC) component and historic data collected from those users to collect a number of Tweets per user in order to create historical matrices (Historic Data Collection - HDC) component [8]–[10], [12]–[15].

Figure 5 presents schematically the above methodology, as retrieved from [13], [15], [18]. The left flow presents the procedure for retrieving RTDC data and the right flow presents the procedure for retrieving HDC data.

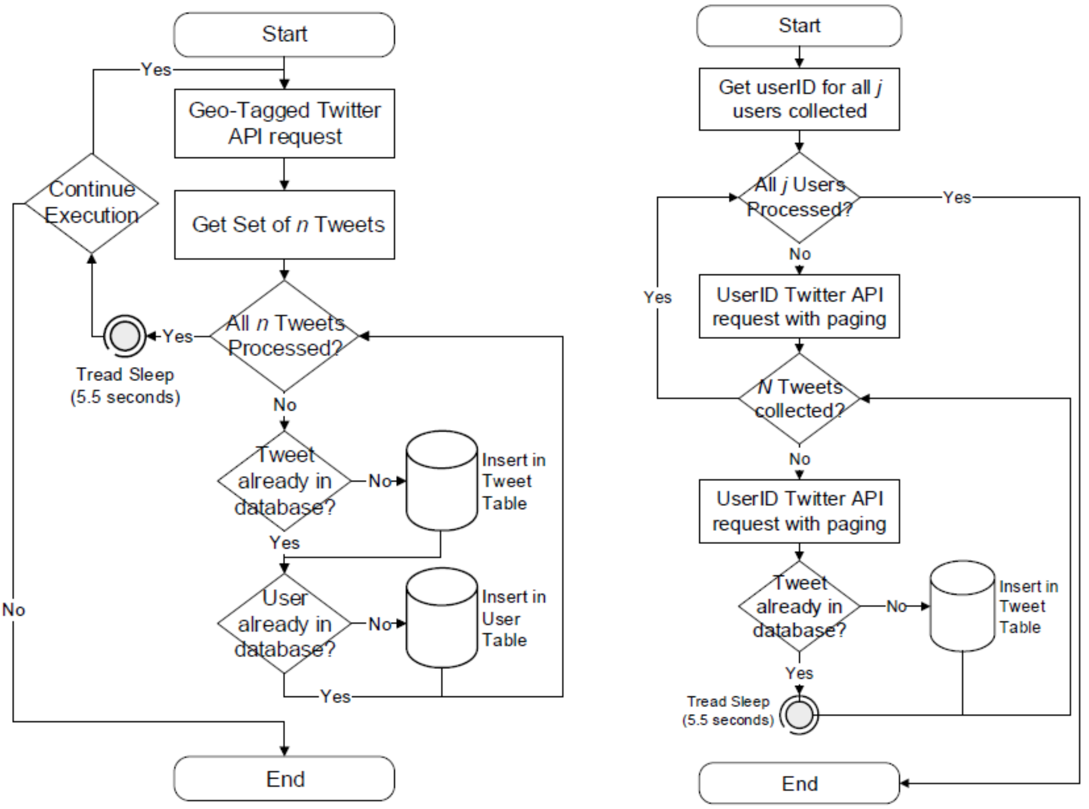

Start
Yes
Geo-Tagged Twitter API request
Get Set of n Tweets
Continue Execution
All n Tweets Processed?
Yes
Tread Sleep (5.5 seconds)
No
Tweet already in database?
No
Insert in Tweet Table
Yes
User already in database?
No
Insert in User Table
Yes
No
End
Start
Get userID for all j users collected
All j Users Processed?
Yes
No
Yes
UserID Twitter API request with paging
N Tweets collected?
No
UserID Twitter API request with paging
Tweet already in database?
No
Insert in Tweet Table
Yes
Tread Sleep (5.5 seconds)
End

Figure 5: Visualization of the methodology retrieved from E. Chaniotakis et al. (2017), E. Chaniotakis et al. (2015 a) and E. Chaniotakis et al. (2015 b)

Except these modules, M. Sinha et al. (2016) [14] include in their proposed system architecture the processes of "aggregation", "visualization" and "output", as presented in Figure 6.

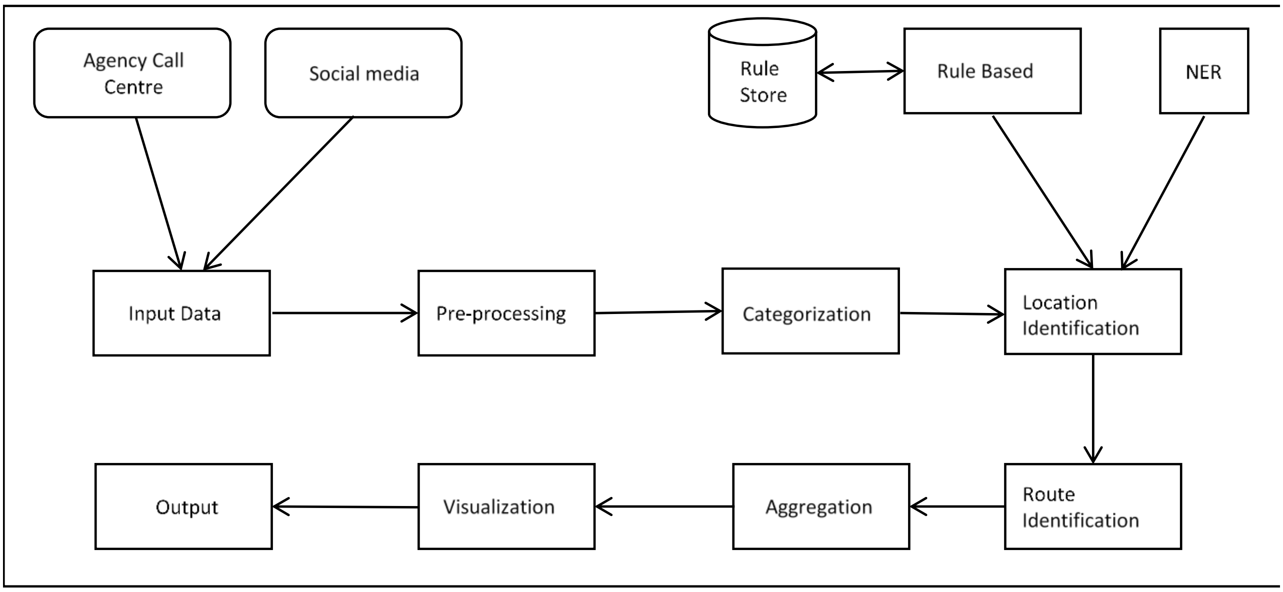


Figure 6: Proposed methodology from M. Sinha et al. (2016)

The proposed data mining model collects SUMs from social media and then it processes the collected SUMs by applying text mining techniques, in order to assign the appropriate class label to each SUM. The analysis of the classified SUMs provides the system with the ability to notify the presence of a traffic event. Finally, there is a capability to add geographic information in the SUMs.

In their study, Efthymiou & Antoniou, 2012 [19] mentioned that different twitter applications have been developed which process the collected data by using the statistical software R [20]. This provides the possibility for statistical analysis of the collected data. In the case of Twitter, the messages of the platform's users that are publicly available (mentioned above as SUMs) can be scrapped, parsed and analysed in R [20], by using the appropriate packages such as TwitterXML [19], [21], [22].

The first step of the proposed methodology from Efthymiou & Antoniou, 2012 [19] is the survey conduction and the statistical analysis of questionnaires. A script that can retrieve information about the number of the tweets containing words chosen by the user and the geographical location of Twitters' users (in case the application was enabled by the users) can be coded in R and applied in the collected data. The script reads the time format from the html page and it translates it in R and prints a graph using the ggplot2 package of R. The script also reads and stores the location of the users (in case that it is provided by them), which can then be plotted by the googleVis package [19].

Cottrill et al., (2017) [23] at their study designed and implemented a methodology, in order to share information among their customers related to planned and unplanned disruptions using *Twitter*. This methodology proposed a mixed-methods approach to data collection from social media platforms, combining both semi-structured interviews of persons and transport-related data collected from Twitter [23]. This methodology consisted of two main parts. The first part is about the semi-structured interviews and their qualitative analysis and the second part is the analysis of the Twitter's SUMs [23].

T. Ruiz et al., (2016) [11] mentioned at their study that the MINERVA Project, funded by the Ministry for Economy and Competitiveness of Spain, is planned to collect passive information (without the intervention of individuals) to inferred characteristics of social network interactions and activities and travels from Social Media and mobile phones, in order to enhance data collection methods for travel demand forecasting. The methodology of the MINERVA Project consists of four main steps [11], which are:

1. Data extraction from Social Media

2. Data collection from mobile phones
3. Estimation and calibration of data mining algorithms
4. Data fusion techniques

S. E. Middleton et al. (2014) [24] developed a real-time crisis-mapping platform which is "mapping real-time tweet flood reports for at-risk coastal areas near known geological fault lines that have the potential to cause a tsunami" [24]. The researchers developed an architecture, which is split into sets of offline and real-time services. The proposed methodology [24] uses both off-line and on-line information (historical data and real-time-data), in order to produce geo-tagged information through the use of Twitter API [17], OpenStreetMap [25] and GooglePlaces API [26].

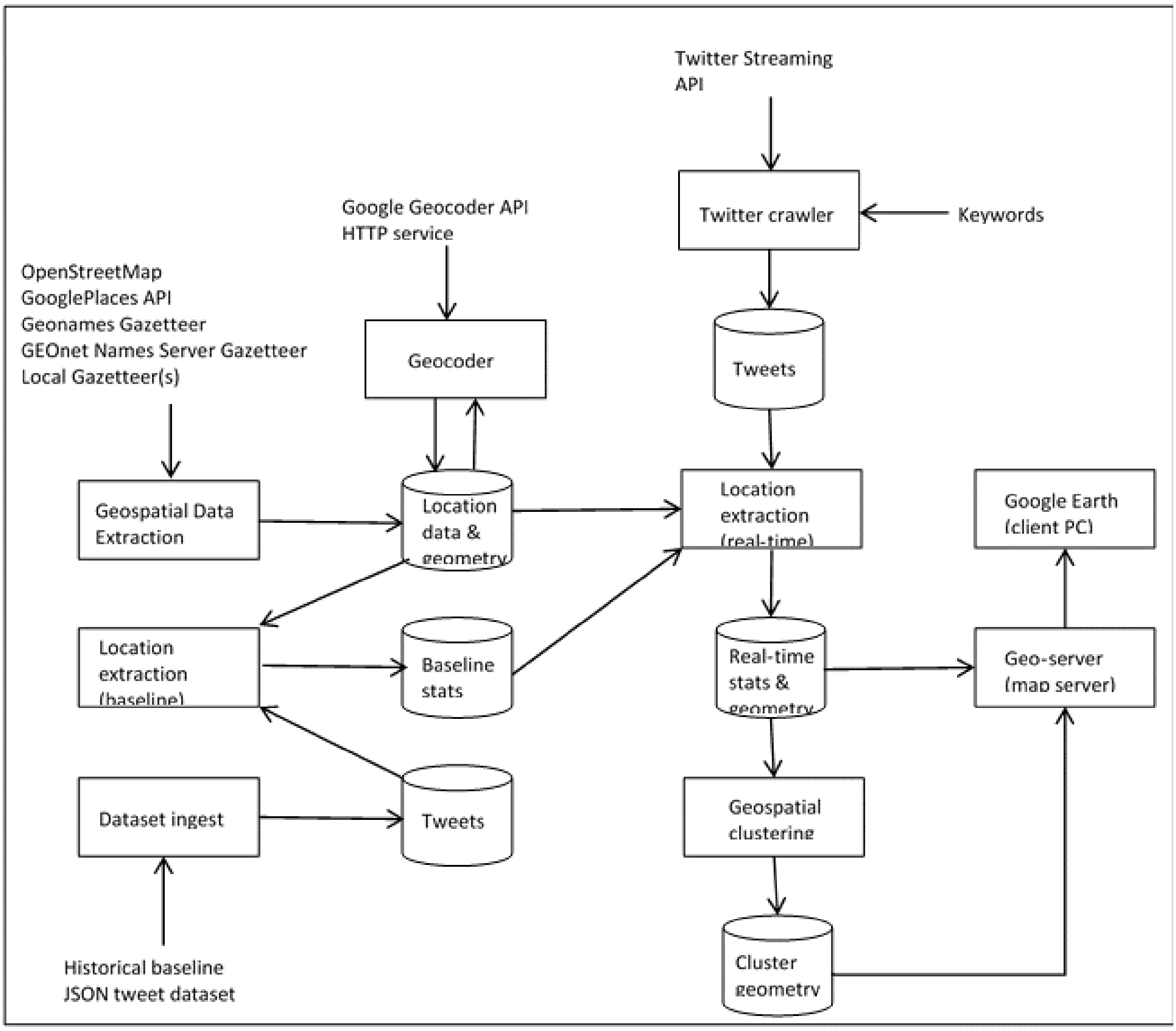


Figure 7: Proposed methodology by S. E. Middleton et al. (2014)

J. Pereira et al. (2017) [27] proposed a methodology in their study, which collected data using Twitter's Streaming API through a Python library (data derived through "Tweepy" [28]). The library was configured by activating the 'locations' filter. This process allows the retrieval of all tweets within a defined bounding-box.

The collected SUMs were submitted in three pre-processing operations [27]:

1. **Lowercasing:** Every message presented in a tweet was converted into lowercase.
2. **Transforming repeated characters:** Sequences of characters repeated more than three times were transformed.
3. **Cleaning:** Removing Uniform Resource Locators (URLs) and user mentions.

Then, the collected geolocated SUMs were classified through a classifier model developed in that study. The classifiers used were Support Vector Machines (SVM), Logistic Regression (LR) and Random Forests (RF) and they were trained by using different groups of features (Bag-of-words or Bag-of-embeddings or combination of both) [27].

Another methodology to derive data from social media, and especially from Facebook (check-ins), and to produce a tool with spatial information has been proposed by I. Toumpalidis (2017) [29] and I. Toumpalidis & N. Karanikolas

(2015a) [30], Toumpalidis & N. Karanikolas (2015b) [31] and partially used by J. M. Salanova Grau et al. (under review) [32] and J. M. Salanova Grau et al. (under review) [33].

In Figure 8 there is a representation of the flowchart used by [29]–[33].

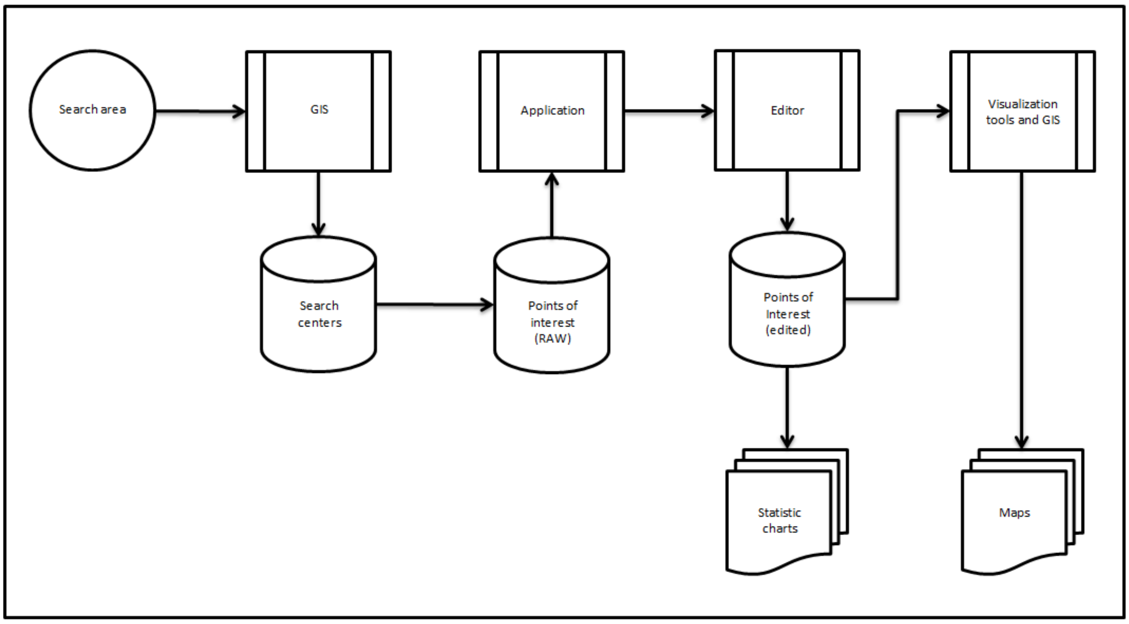


Figure 8: Methodology retrieved from I. Toumpalidis (2017), I. Toumpalidis & N. Karanikolas (2015a), Toumpalidis & N. Karanikolas (2015b), J. M. Salanova Grau et al. (under review) and J. M. Salanova Grau et al. (under review)

The methodology is comprised of four main modules:

1. Geographical Information System (GIS): Use of GIS to collect the coordinates from the study area, in order to be used for information retrieval.
2. Use of Facebook API: Connection with the Facebook API [34] and use of the geographical information mentioned above (coordinates), in order to collect information for a specific area.
3. Editor: Processing and formatting of the datasets.
4. Visualization Tools and GIS: Visualization and analysis of the final results.

It is evident from the above literature review that there is a variety of methods used in order for transport-related data to be collected from social media platforms and be further analysed in order for valuable information to be extracted.

2.3 Preference Analysis Models

The following subsections present and describe the factors which may affect the preferences of the users and which can lead to better estimation of the users' activities preferences.

2.3.1 Modelling the Traveller

Until 2030, an additional billion people will be in the world, and at least 20% of them will be travelling. The population is foreseen to reach 9.6bn by 2050, according to the United Nations [35], while according to more forecasts a 5% annual increase in passenger traffic will be reached from 2015 and the global passenger and freighter fleet will double [36].

During the past three decades, research in travel behaviour has changed and from aggregate level analysis (i.e. transportation analysis zones (TAZs) and neighbourhoods) is now focusing on more behaviourally explicit, and disaggregate policy inquiries, aiming to understand individual level attributes and their correspondence to travel choices [37]. Travel mode choice has been widely investigated based on the "random utility maximisation theory" and "choice behaviour theory", such as discrete choice model. The basic idea behind the use of "travel mode choice models" is to better understand the connection and correspondence between the choices of the traveller and other affecting factors, such as the socio-economic level and the level of transport services. [38].

In general, being able to understand what customers really want and expect is important to any successful business. But if this could also go a step even further to the factors that affect and shape travel experiences and to the preferences of travellers themselves, their expectations and needs would be more efficiently met [39]. It is clear nowadays that within the increasingly complex and interconnected world of travel that it opens up new markets and promotes wider travel, a more sophisticated approach is required, in order to be able to know more about how travellers currently behave but also will behave in the future.

Understanding why people choose specific travel modes and how their travel behaviour is affected by social media, as well as by other parameters (i.e. ethical concerns, desire for wellbeing, etc.) [39] is a prerequisite for improving the transport system, for developing useful and appropriate services and applications but also for encouraging behaviour alteration when required.

Taking this into account, within the framework of My-TRAC project, Table 1 has been developed. Table 1 describes in detail the travellers' characteristics that, to a greater or lesser extent, affect their choices related to the transportation system. These traveller characteristics are categorised in three main categories: dynamic, semi-dynamic and static:

A. The **Static** characteristics are the ones that are considered to be non-variable and that remain constant with respect to time (they don't change continuously while the system is activated). These are inserted once by the user. Of course, when necessary, they can be updated by the user or the system. The static characteristics include mainly the demographic features of the travellers, as well as their impairments (of any type). The classification of the characteristics related to impairments is based on the World Health Organisation's International Classification on Functioning and Health (ICF), which provides an authoritative and forward-looking framework for the understanding and, by extension, the management and delivery of services for people with disabilities [40].

B. The **Semi-dynamic** travellers' characteristics are the features that the traveller defines each time they start a trip, unless the relevant characteristics remain the same since they last had the same route [42, 43]. For example, mobility limitations that are not permanent (i.e. mobility impairments) apply in this category (such as use of stroller of carriage of luggage) or even some identity characteristics, such as parental status. Also, according to IM@GINE IT project [41] the users can be categorized based on the reason for travelling (context of use). The traveller has different needs if, for example, the reason for travelling is recreation or work, and if he is in his/her town or in a different town/ country. Thus, an effective system should be able to change accordingly when the user type changes, for the same person/user, as the needs for information vary significantly with the reason for transportation.

C. The **Dynamic** travellers' characteristics are the ones that are calculated automatically by the system and are updated according to the user's choices each time he/she uses the system. They are based on the assumption of changing behaviour during time. Such dynamic parameters are the route characteristics, as there are various possible priorities based on which the traveller can select the route that will follow in order to reach his/her destination [42]. Also, the emotional states are included here, which can be interchanged even during a trip (i.e. from excitement and anticipation to boredom or frustration, etc.), but also the psychological states, including –for example- a traveller's attitude that describes their predisposed state of mind regarding transport elements (i.e. regarding public transport means), which in turn influences their thoughts and actions [43].

Table 1: Traveller characteristics

Category	Sub-category		Variables
Static	Demographic		Age
			Attitudes
			Gender
			Car ownership
			Car usage
			Country
			Education
			Habit
			Home location
			Income
			Intention
			Norms
			Perceived control
			Perceived safety
			Activities execution (during travelling)
			Workplace location
	Impairments		Lower limb impairment
			Wheelchair users
			Upper limb impairment
			Upper body impairment
			Physiological impairment
			Psychological impairment
			Cognitive impairment
			Vision impairment
			Hearing impairment
			Communication producing and receiving difficulties
Semi-dynamic	Identity	General	Occupational status
			Marital status
			Parental status
			Physical activity related
		Social profile	Calm
			Certainty seeking
			City dweller
			Convenience seeking
			Countryside-lover
			Dedicated to family
			Fit or desire to be fit
			Environment conscious
			Efficient
			Exclusivity seeking
			Frustrated
			Patient
			Punctual

Category	Sub-category		Variables
			Socializer
			Spontaneous
			Status seeking
			Variety seeking
			Workaholic
		Context of use	Business traveller in a foreign country or new city
			Commuter (a person in his/her city that moves daily to/from his work)
			Leisure/recreational traveller in his/her city
			Tourist (a person in a foreign country or new city that travels for holidays).
			Emergency traveller (a person in his/her city that needs to travel urgently, e.g. to go or get someone at the hospital).
		Mobility profile	Cyclist
			Car driver
			Pedestrian
			Public transport user
		Travel behavioural profile	Aggressive
			Healthy
			Independent
			Risk taking
			Sporty
	Mobility limitations		Lower limb impairment
			Stroller user
			Carrying luggage-equipment
			Wheelchair users
			Cane user
Dynamic	Emotional states		Anxiety/ stress
			Boredom
			Calmness
			Excitement
			Fatigue
			Frustration
			Happiness
			Confidence
			Relaxation
	Psychological states		Intention
	Route characteristics (listed with a prioritisation order)		Accessible route (for travellers with mobility impairments)
			Cheapest route
			Most picturesque/interesting route
			Nearest route (the one with the less possible distance from the user’s starting point)
			Route with acceptable/ preferred (by the user) transportation mean types

Category	Sub-category		Variables
			Route with number of interchanges (change of transport mean) less than the maximum acceptable by the user
			Route with walking distance less than the maximum acceptable one by the user
			Shortest route

2.3.2 Modelling Traveller's Activities

Individuals generate different and complex travel-activity patterns, as they are performing their daily activities, in different times and in different locations. As many researchers have conceptualised these observed behaviour patterns as the outcome of choices made within constraints, attempts to understand and explain spatial choice behaviour have focused on the relationship of observed behaviour to those sets of variables believed to affect choices and constraints [44]. Activity-based models are part of the 3rd generation of travel demand models, being mainly used and applied during the last three decades and they treat travelling as a result of the demand for activity participation [45].

Each trip connects two distinct activities and each traveller moves across the physical plane to perform activities (for example, a traveller uses public transport to move from their home location to their workplace). As understanding, as well as predicting activities, is an important aspect of the My-TRAC project, in order to provide improved trip advice, the analysis of traveller's preferences needs to also take place. For this reason, Table 2 has been developed, including a categorisation of the travellers' preferences concerning their activities related to travelling. The activity preferences are directly linked and can be considered an extension of the semi-dynamic characteristics of Table 1 above.

Table 2: Activity preference

Category	Sub-category	Attribute	Sub-attribute
Activity preferences		Proximity from origin	Travel time between home and activity
			Distance from home
	Activity's (destination) characteristics	Opening hours of activity	
		Parking convenience of activity	
		Accessibility	Accessible by PT
			Accessible for vulnerable users
	Purpose of the trip	Commuting	Work
			Education
		Personal	Social activities (visits to friends/ families, etc.)
			Shopping for everyday needs (i.e. grocery shopping)
			Health related activities (i.e. doctors' appointments, etc.)
			Services-related activities (i.e. travel to post-office, bank, registry office, etc.)
			Religion related activities (i.e. travels related to practicing the worship of the religion of a traveller, etc.)
		Leisure	History/Culture
			Hobbies (i.e. sports and/ or other similar recreational activities)
			Eating & drinking
			Shopping for pleasure (i.e. visits to malls, touristic shopping, etc.)
		Business	Out of the office meetings
			Visits for work
	Day of trip	Weekday	
		Saturday	
		Sunday	
		Official holiday	
	Activity sequence and duration		
	Priorities for activities	Favourite activities	
	Telecommunications options	Indoors	Wi-Fi availability
			Internet availability (wired)
		Outdoors	3G/4G availability
	Access travel information	Traffic conditions	
		Route guidance	
		Parking availability	

Category	Sub-category	Attribute	Sub-attribute
		Public transportation schedules	

The analysis of a traveller's activities, as well as their connection to specific routes and specific travel habits and their correlation with additional parameters (i.e. selection of travel modes and personal limitations, etc.) can provide very important information for the personalisation of the journey and the provision of relevant suggestions and tips. Within the My-TRAC project (WP2), relevant data will be collected both from the Twitter API and from the users of My-TRAC application (through brief questions regarding their preferences during the initial uses of My-TRAC application).

2.3.3 Modelling Traveller's Preferences

New travel modes and services (i.e. car sharing, MaaS, etc.) can play an important role in improving the efficiency and sustainability of transportation systems and for this to be achieved the preferences of travellers should also be taken under consideration and be evaluated [46].

The preferences list included in Table 3 below, is linked mainly to the existing travel modes and the features that each one of them offers, as well as to the transit between different modes.

According to what has been also described in previous sections, each traveller's preferences are linked to his/her characteristics and his/her activities of each time. The following table is linked to the attribute "Route with acceptable/ preferred (by the user) transportation mean types" of the dynamic characteristics of Table 1. All these parameters combined together may provide a most valuable input for offering personalised information and services to My-TRAC application users.

Table 3: Travellers preferences

Category	Sub-category	Attribute	Sub-attribute
Travel mode	Bus	Walking time to bus stop	
		Waiting time	
		In-vehicle time	
		Fare	
		Comfort	Cleanness
			Crowdedness
			Air-conditioning
		Efficiency	Punctuality
			Disruptions (Cancelation, strike)
		Carbon footprint	
		Reliability	
		Flexibility	
		Protection against the weather	
		Activities execution (during travelling)	
		Safety	
		Control	
		Scenery	
		Mode accessibility	
		Type of tickets available	Mobile phone
			Personalised rechargeable ticket/card

			Anonymised rechargeable ticket (one way/return)
			Monthly card
			24hours travelling card
			Reduced price tickets (students, elderly, unemployed)
	Metro	Walking time to platform/walking time to station	
		Waiting time	
		In-vehicle time	
		Fare	
		Comfort	Cleanness
			Crowdedness
			Air conditioning
		Efficiency	Punctuality
			Disruptions (Cancelation, strike)
		Frequency	
		Carbon footprint	
		Reliability	
		Flexibility	
		Protection against the weather	
		Activities execution (during travelling)	
		Safety	
		Control	
		Mode accessibility	
		Type of tickets available	Mobile phone
			Personalised rechargeable ticket
			Anonymised rechargeable ticket (one way/return)
			Monthly card
			24hours travelling card
			Reduced price tickets (students, elderly, unemployed)
	Rail	Walking time to platform/walking time to station	
		Waiting time	
		In-vehicle time	
		Fare	
		Comfort	Cleanness
			Crowdedness
			Air conditioning
			Class
		Efficiency/reliability	Punctuality/delays
			Disruptions (Cancelation, strike)
		Frequency	
		Carbon footprint	
		Flexibility	

		Activities execution (during travelling)	
		Safety	
		Control	
		Scenery	
		Mode accessibility	
		Type of tickets available	Mobile phone
			Personalised rechargeable ticket
			Anonymised rechargeable ticket (one way/return)
			Monthly card
			24hours travelling card
			Reduced price tickets (students, elderly, unemployed)
	Tram	Walking time to platform/walking time to station	
		Waiting time	
		In-vehicle time	
		Fare	
		Comfort	Cleanness
			Crowdedness
			Air conditioning
			Class
		Efficiency/Reliability	Punctuality/ delays
			Disruptions (Cancelation, strike)
		Frequency	
		Carbon footprint	
		Flexibility	
		Activities execution (during travelling)	
		Safety	
		Control	
		Scenery	
		Mode accessibility	
		Type of tickets available	Mobile phone
			Personalised rechargeable ticket
			Anonymised rechargeable ticket (one way/return)
			Monthly card
			24hours travelling card
			Reduced price tickets (students, elderly, unemployed)
	Bicycle	Comfort	Bicycle lanes
		Scenery	
		Travel time	
		Flexibility	
		Protection against the weather	
		Safety	
		Control	

		Route accessibility	
	Bike sharing	Comfort	Bicycle lanes
		Scenery	
		Travel time	
		Flexibility	
		Protection against the weather	
		Safety	
		Control	
		Route accessibility	
		Type of tickets available	Electronic booking & paying platform
	Car	Comfort	Autonomous driving functionalities
			Route guidance
			Car size/class
		Travel time/ trip duration	
		Cost	
		Carbon footprint	
		Flexibility	
		Activities execution (during travelling)	
		Safety	ADAS
		Control	Autonomous driving functionalities
		Scenery	
		Mode accessibility	
	Car pooling	Cost	
		Travel time/ trip duration	
		Comfort	Cleanness
			Car size/class
		Carbon footprint	
		Flexibility	
		Protection against the weather	
		Activities execution (during travelling)	
		Safety	
		Control	
		Scenery	
		Mode accessibility	
		Type of tickets available	Electronic booking & paying platform
	Car sharing	Cost	
		Travel time/ trip duration	
		Comfort	Cleanness
			Car size
		Carbon footprint	
		Flexibility	
		Protection against the weather	
		Activities execution (during travelling)	
		Safety	
		Control	
		Scenery	

		Mode accessibility	
		Type of tickets available	Electronic booking & paying platform
	Moto	Comfort	Autonomous driving functionalities
			Route guidance
			Moto size/class
		Travel time/ trip duration	
		Cost	
		Carbon footprint	
		Flexibility	
		Safety	ARAS
		Control	Autonomous driving functionalities
		Scenery	
	Moto sharing	Comfort	Moto size/class
		Travel time/ trip duration	
		Cost	
		Carbon footprint	
		Flexibility	
		Safety	ARAS
		Control	Autonomous driving functionalities
		Scenery	
		Type of tickets available	Electronic booking & paying platform
	Walk	Travel time	
		Comfort	Walking paths
		Scenery	
		Protection against the weather	
		Safety	
		Route accessibility	

3 Activity Prediction Mechanism

3.1 Concept & Definition

Travellers move across the physical plane to participate in/to perform activities. Each traveller's trip serves in connecting two distinct activities that may take place in different locations. A list including a categorisation of the travellers' preferences concerning their activities related to travelling has been identified, through the relevant literature, and presented in Table 2. Based on this and on its contained field "Purpose of the trip", activities of a traveller are defined as the purposes of his/her trips. Provided that the update of the number of the "activity" categories (see **Table 2**) from 3 to 4, was suggested end of M10, only the original three categories are addressed within this version of the deliverable. In particular, the set of "activities" investigated in this Section is the following one:

1. Commuting
2. Personal
3. Leisure

The evaluation of the performance of the algorithms on the new set of categories will be provided later, in D2.3 and/or D2.4.

Each of the categories is further detailed in sub-categories as shown in Table 2. Besides understanding travellers' activities, in the framework of My-TRAC, it is equally important to provide predictions of users' anticipated activities. The latter will enable My-TRAC application to provide recommendations concerning "intermediate" activities for a traveller. To this end, an activity prediction mechanism is designed and developed, which, based on the input retrieved from traveller's social media, is able to predict their next activities.

3.2 Approach

In order to create an activity prediction mechanism based on user's preferences, the strategy depicted in Figure 9 is followed. The procedure is based on information that the users voluntarily share in a public way. Within this deliverable, Twitter is used as the source of information, due to the public character of users' comments. However, any other source/social media that provides such kind of data can be introduced to the pipeline of calculations depicted in Figure 9. Notice that Twitter data can be either about a specific Tweet (e.g. timestamp, user id, text, etc.) and also about the user behaviour (e.g. liked tweets, retweets, following users, etc.). Once data has been retrieved, the text of the tweet is processed using text mining techniques to infer his activities. For that purpose, an unsupervised method has been developed, which is detailed in the Section 3.4.1. After identifying the activities of each user, the activities are ordered by time forming, thus, an activity sequences for each user. Moreover, by processing information about users' tweets, users' favourite tweets and the account that they follow the user profiling for each user is succeeded. The identified sequences of activities and the results of the user profiling are presented as input to a Markov chain model for predicting the user's next activities.

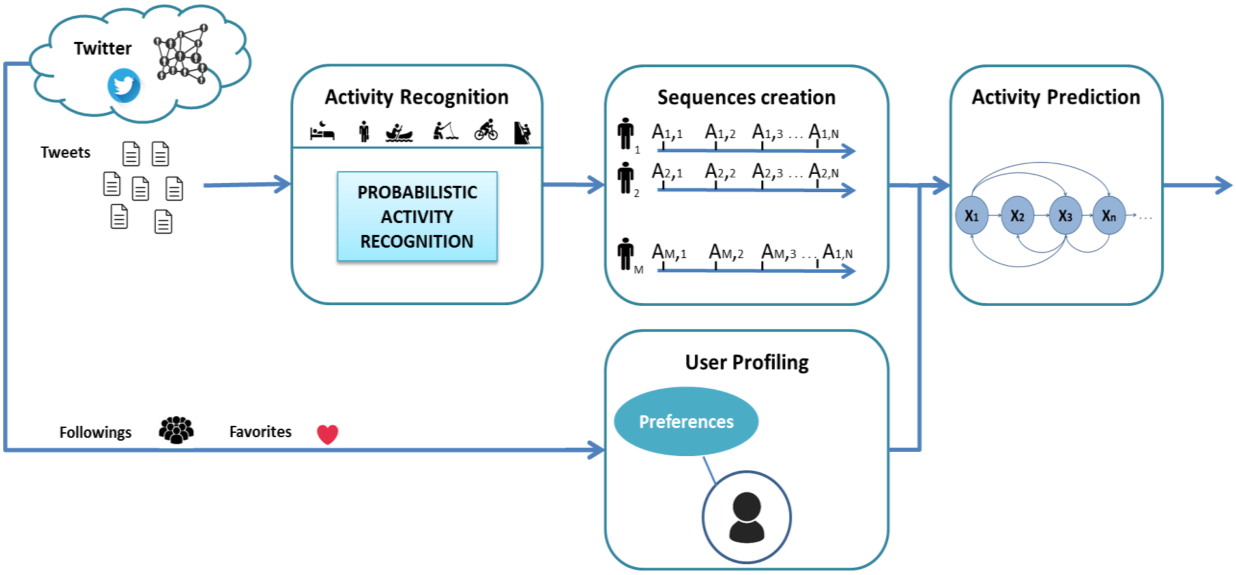

Twitter
Tweets
Activity Recognition
PROBABILISTIC ACTIVITY RECOGNITION
Sequences creation
$A_{1,1}$ $A_{1,2}$ $A_{1,3}$ … $A_{1,N}$
$A_{2,1}$ $A_{2,2}$ $A_{2,3}$ … $A_{2,N}$
$A_{M,1}$ $A_{M,2}$ $A_{M,3}$ … $A_{1,N}$
Activity Prediction
X_1 X_2 X_3 X_n
User Profiling
Preferences
Followings
Favorites

Figure 9: Activity Prediction Approach

3.3 Analysis of social network's data

In order to access user information from social media (in this case Twitter), the user needs to provide permission to analyse his information and username on the social network. This process is compatible with the use cases presented in D5.1 from the My-TRAC application.

The information will be retrieved with a crawler that makes use of the official API of the social network and will be analysed to map it into activities/preferences. It is assumed that if a user shares information on a topic, that topic is of interest to the specific user and, therefore, he may want to receive suggestions regarding this topic.

Analysing a Twitter account or a Twitter Activity implies crawling Twitter and analysing both historical tweets and live tweets. A Rest API for each action is publicly accessible.

Endpoints for historical tweets are very rich with over 80 fields providing data about the characteristics of the tweet (original tweet/a reply/a retweet), its text, its associated media (either pictures or videos), the geolocalization of the tweet, the publishing time and so on. Streaming endpoints are also very detailed with more than 100 fields even though we don't use that many fields anyway.

To analyse the profile of users from their tweets two inputs are needed. First the user information, after authentication we can retrieve what we call:

- CLIENT_TOKEN_KEY (a 25-character string of capital letters and figures),

- CLIENT_TOKEN_SECRET (a 32-character string of capital letters and figures).

Then we indicate on what topics we want to analyse data, keyword databases have been (or can be) created to identify the interests of the user.

A first way to analyse someone's interests is by analysing tweets he wrote or retweeted. Even though users can retweet tweets that they disagree with, it still demonstrates an interest about the subject. Until now we have used the free version of the API, which allows retrieving 20 tweets / pages, approximately up to 2000 tweets. Because of this page system, retrieved tweets cannot be the last 2000 tweets, but in the other hand we have access to old tweets which can still be interesting for us.

A second way to analyse someone's profile is to analyse the accounts he follows. As the description of an account is something very precise, is the only field we choose to analyse, while we ignore the tweets from each account, which would cost too many resources and would not be relevant. Using the free API we could retrieve up to 50 accounts, but a significant delay has been observed.

Last but not least, Twitter API enables us to retrieved tweets marked as favourite by the user. Approximately the same number of tweets as before can be retrieved.

In addition, the streaming API can be used. It allows to retrieve any live tweets within a given area and a set of keywords. This has been used to enrich the mechanism for identifying preferences/activities from the tweets. To be able to use it, five inputs are required in our application:

- A set of keywords to search;
- 2 set of GPS coordinate to create location box, representing the area;
- The frequency of analysis (i.e.: every 100 tweets)
- One or more language code to filter tweets by language;
- The keyword databases chosen or all of them.

3.4 Activity recognition from social media

For recognizing a traveller's activities from text data retrieved by their posts on social media, a probabilistic activity recognition algorithm was developed. The developed method use text-mining techniques and aims at recognizing with high accuracy the activities of a user. The method follows a frequency-based approach, which is detailed in the next section.

3.4.1 Probabilistic Activity Recognition Method (PAR)

The developed method that aims at recognizing the type of activities that a user performs, from text input data retrieved from social media is a combination of Natural Language Processing (NLP) techniques and of a probabilistic approach. As Figure 10 illustrates, the activity recognition procedure is divided into four phases:

1. Twitter crawling
2. Dictionary generation
3. Text pre-processing
4. Activity recognition

Each of the four phases is detailed in the next sections.

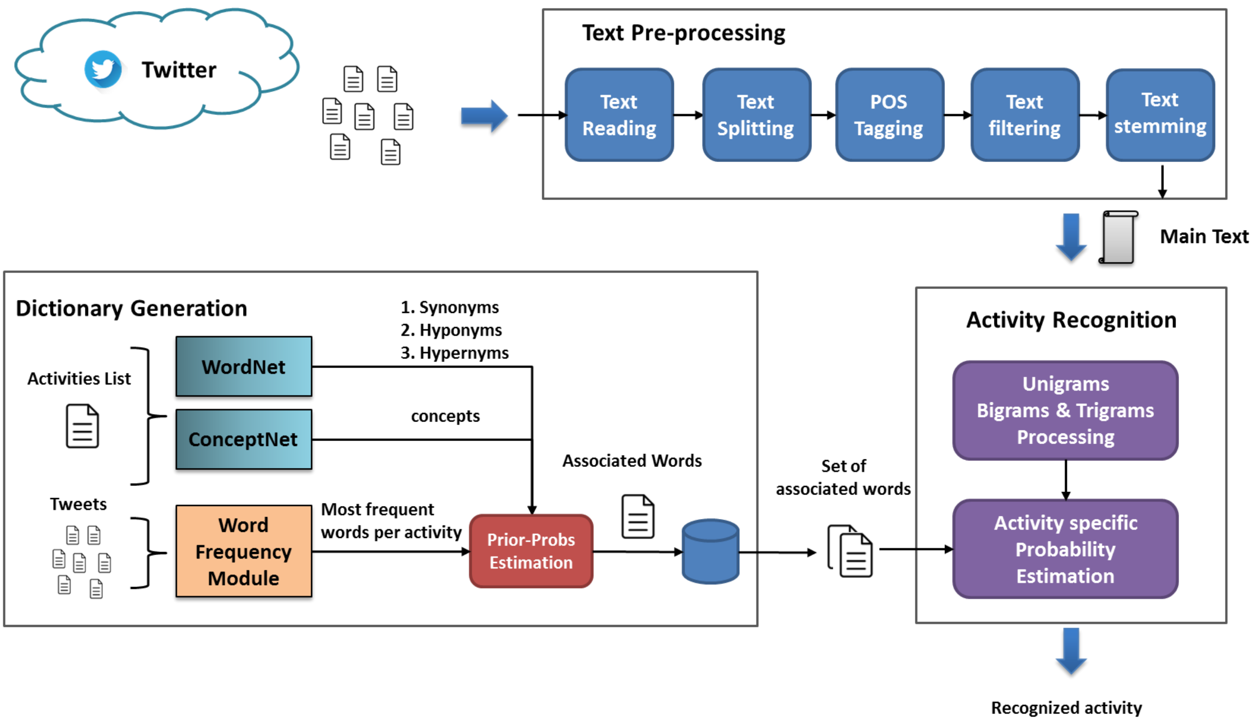


Figure 10: Probabilistic activity recognition method's architecture

3.4.1.1 Twitter Crawling

The first phase of the Probabilistic activity recognition method is the Twitter's crawling. As discussed in Section 3.3, data retrieved from the social network Twitter was decided to be utilized. More specifically, this algorithm takes as input information about a user's tweets that contain: a) the location, b) the timestamp and c) the raw text of the published tweet coupled with the unique user's ID.

3.4.1.2 Dictionary Generation

Apart from the input data retrieved from Twitter the activity recognition algorithm receives also as input sets of pre-associated words or phrases (a.k.a. dictionary) with the defined activities. In order to generate the dictionaries, three parsers were developed for retrieving data from

a) WordNet [47] which is a large lexical database of English,
b) fConceptNet [48] which is an open, multilingual knowledge graph and
c) Twitter Search API.

a) Wordnet Parser

Regarding WordNet, it groups nouns, verbs adjectives and adverbs into sets of cognitive synonyms that are called "synsets", each which expresses a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The main relation among words in WordNet is synonymy, as between the words car and automobile. Synonyms - words that denote the same concept and are interchangeable in many contexts - are grouped into sets (synsets). The most frequently encoded relation among synsets is the super-subordinate relation (also called hyperonymy, hyponymy or ISA relation).Taking as input a list of words (e.g. activity categories and sub-categories) the developed WordNet parser, utilizes the WordNet's API for retrieving synonyms, hyponyms and hypernyms of the given words, creating, thus, for each activity it creates 3 sets (i.e. synonyms-set, hyponyms-set, hypernyms-set)

ConceptNet Parser

ConceptNet is a knowledge graph that connects words and phrases of natural language (terms) with labelled, weighted edges (assertions). It uses a closed class of selected relations such as IsA, UsedFor, and CapableOf, intended to represent a relationship independently of the language or the source of the terms it connects. Taking as input the list of activities the developed parser utilizes the ConceptNet's API for enriching even more the generated dictionary with an extra set called concept-set.

Word Frequency Module

Apart from the WordNet and the ConceptNet parsers one more module is developed for extracting the most frequent words from tweets related to specific categories of activities. In order to do this,the related to the activities hashtags have been identified and by utilizing the Twitter Search API several tweets (~5000) were collected containing the provided hashtags. From the collected tweets, the developed module is able to extract the most frequent words for each category and create an extra set of words for each activity (a.k.a. frequentWords-set). Based on the above, 5 sets of pre-associated words for each activity were constructed (i.e. hyponyms-set, hypernyms-set, synonyms-set, conceptnet-set, frequentWords-set). The sets of an activity are constructed so as not to contain same words as well as to contain the stems of the collected words. The collected datasets may include single words or phrases that consist of more than one word. For this reason we call each associated to an activity word/phrase as an associated "n-gram". An *n*-gram is a contiguous sequence of *n* items in a given text. An n-gram of size 1 is referred to as a "unigram", size 2 is a "bigram", size 3 is a "trigram", etc. Based on the collected sets for all activities, 5 multisets (i.e. hyponyms-mset, hypernyms-mset, synonyms-mset, conceptnet-mset, frequentWords-mset) are created containing the sets of all activities. A multiset (e.g. synonyms-mset) of a certain category (e.g. synonyms) is the union of the generated synonyms-sets of all activities. Concerning the appearance of an n-gram in the pre-associated sets, each word is annotated with a value describing the correlation of the n-gram with the activity. The correlation of a n-gram is proportional the number of occurrences of a n-gram in the multisets. The 5 multisets are the final generated dictionary, which is the result of the dictionary generation procedure.

3.4.1.3 Text Pre-processing

Before applying the probabilistic approach to the collected tweets for recognizing a traveller's activities, a pre-processing procedure of 5 steps precedes which are the following:

1. **Text reading**: The reading of a user's tweet text takes place.
2. **Text splitting**: The text of the tweet is split on sentences.
3. **POS tagging**: Each tweet's sentence is part of speech (POS) tagged
4. **Text filtering:** The tweet's text is filtered out for removing certain parts of speech, hyperlinks, non-asci characters, etc.
5. **Text stemming:** The tweet's text is stemmed in order plurals, suffixes to be removed and only the stem of the words to remain.

After pre-processing the raw tweet's text, the text, in its final form, is analysed by the activity recognition core algorithm.

For instance, the following sample tweet: "Lunch/dinner is ready, shrimp and scallops. #Lunch #Dinner #Food #Eat #Eating #Shrimp #Scallops #Rice #Chef" will be finally converted to "lunch / dinner ready, shrimp scallops, lunch dinner food eat eating shrimp scallops rice chef", which be the text that the activity recognition algorithm will be applied to.

3.4.1.4 Activity Recognition Core Algorithm

Based on the pre-processed input text and generated dictionary with the sets of pre-associated words, described in section 3.4.1.2, the algorithm parses the tweet in n-grams and searches for all the occurrences of the pre-associated n-grams within the text. The number of occurrences of an n-gram that is pre-associated with an *activity* is considered to be proportional with the probability of the activity to be relevant to the provided text. Thus, we consider:

$$p(activity|tweet) \propto \; N^{tweet}_{n-gram_k} \qquad (1)$$

Where $p(activity|tweet)$ is the probability of the *activity* to be the related activity of a *tweet and* $N^{text}_{n-gram_k}$*, the* clean count of the appearance of a pre-associated n-gram within the analysed text.

Moreover, we consider that the $p(activity|tweet)$ is proportional to the number of activities that a pre-associated, identified within the tweet, n-gram is related to:

$$p(activity|tweet) \propto {}^{1}\!/_{N^{Set_k}_{n-gram_k}} \;, \qquad (2)$$

Where $N^{Set_k}_{n-gram_k}$: the number of occurrences of a n-gram in the multiset k, and $k \in \{\mathrm{hyponyms-mset, hypernyms-mset, synonyms-mset, conceptnet-mset, frequentWords-mset}\}$. The less the appearance of a word within the pre-associated sets of activities is, the greater is the probability the text to referred to the activity which the n-gram is related to.

Finally, we consider that the $p(activity|tweet)$ is proportional to the number of words of the identified word/phrase of the pre-associated multisets within the text, over the number of the phrase with the maximum number of words within the multiset where the identified $n - gram_k$is placed in. Large n-grams provide more evidence with respect the activity they belong to than smaller n-grams. Thus, the larger the n-gram is, the larger the probability of its pre-associated activity to be the related activity of the tweet.

$$p(activity|tweet) \propto \frac{CW_{n-gram_k}}{\max(CW_{type_{n-gram_k}})}, \qquad (3)$$

Where CW_{n-gram_k}: the count of words of the identified within the *tweet*, $n - gram_k$ and $\max(CW_{type_{n-gram_k}})$: the maximum count of words of an n-gram within the multiset that the $n - gram_k$is placed in.

In order for all the pre-associated n-grams identified within the tweet to be considered to the calculations of the $p(activity|Text)$ and based on the above relations, we can write:

$$p(activity|tweet) \propto \sum_{n-gram_k \in S_{activity}} {}^{1}\!/_{N^{Set_k}_{n-gram_k}} \; N^{tweet}_{n-gram_k} \frac{CW_{n-gram_k}}{\max(CW_{type_{n-gram_k}})} \qquad (4)$$

Where $S_{activity}$: the set of all the pre-associated n-grams of an activity.

The recognized as the most relevant to the text activity, can be considered as the maximum probability $p(activity|tweet)$ within the set of activities, thus:

$$recognized\ activity = \underset{activity}{\operatorname{argmax}}\, p(activity|tweet), \qquad (5)$$

Which by combining with the Equation (4) is written as:

$$recognized\ activity = \underset{activity}{\operatorname{argmax}} \left(\sum_{n-gram_k \in S_{activity}} p_j\, N^{tweet}_{n-gram_k} \frac{CW_{n-gram_k}}{\max(CW_{type_{n-gram_k}})} \right) \quad (6)$$

3.4.1.5 Evaluation

Description of the dataset

A dataset, based on Twitter data, was built for the step of the evaluation of the activities detection algorithm in this task. The dataset was constructed by crawling historical data from Twitter, through the designated Twitter API. It consists of 1704 tweets (rows) in a comma-delimited (CSV) format (Table 4). The tweets were collected randomly from specific location and on a specific date (Texas, 27/6/2018). Although, Twitter API allows for downloading a wide range of information, only 3 types - fields per Tweet were extracted for the process of evaluation: userID, tweetText, timestamp (Table 4). The language of the retrieved data was chosen to be English, as this is the one used throughout the development of My-Trac.

Table 4: Dataset format

<UserID_1>, <tweetText>, <Timestamp>
<UserID_2>, <tweetText>, <Timestamp>
<UserID_3>, <tweetText>, <Timestamp>
…

There were some pre-processing steps, before utilizing the dataset in the algorithm's performance evaluation to make it more suitable for the analysis steps of the algorithm. Also, to maintain anonymity of the users, the field of userID was altered in such a way to pseudo-anonymize the load of information. This change was required in order to be compliant with the data protection laws (GDPR). The dataset was manually annotated based on 4 categories presented in Section 3.2. The general rule, which the annotation was based on, was the distinction between tweets referring to an activity being performed related to one of the first 3 out of four categories (commuting, personal, leisure) and tweets that their subject is vaguely relevant with those categories (uncategorised). The definition of the categories of activities, as well as their distribution within the dataset are presented in the next table.

Table 5: Definition and distribution of categories of the dataset

Commuting	Defined as the travelling of some distance between one's home and place of work (or education) on a regular basis	783
Personal	Refers to transfers related to the subcategories of social activities, health-related activities and services-related activities	80
Leisure	Refers to transfers related to cultural, religious, food & drink, and other recreational activities	449
Uncategorized	Category for any other option	392

Two different kinds of evaluations have been performed, using the same sample. The first kind of evaluation was made per each category separately, treated as a binary classification. The second kind is a multiclass evaluation of the same sample. In the rest of this section we present the results for each kind of evaluation.

1. Binary classification

Based on the calculation of true positive, true negative, false positive and false negative as described in Table 6.

Table 6: Terminology

Metric	*Description*
True-Positive (TP)	correctly classified as positive
True-Negative (TN)	correctly classified as negative
False–Positive (FP)	wrongly classified as positive
False-Negative (FN)	wrongly classified as negative

The overall evaluation metrics defined for this kind of evaluation are the accuracy, precision, recall and the f1 score. Accuracy refers to the how well a binary classification test correctly identifies or excludes a condition. Precision refers to the proportion of positive identifications that was actually correct. Recall is the fraction of the relevant documents that are successfully retrieved. The F1 score is the weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The mathematical equations of each metric are presented below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}\,call = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 * (\mathrm{Pr}\,ecision * \mathrm{Re}\,call)}{(\mathrm{Pr}\,ecision + \mathrm{Re}\,call)}$$

The evaluation results per each category are:

Table 7: Evaluation results for the "commuting" category.

Commuting	
Precision	0.69
Recall	0.98
F1_score	0.81
Accuracy score	0.79

Table 8: Evaluation results for the "leisure" category.

Leisure	
Precision	0.60
Recall	0.84
F1_score	0.70
Accuracy score	0.81

Table 9: Evaluation results for the "personal" category.

Personal	
Precision	0.57
Recall	0.61
F1_score	0.59
Accuracy score	0.96

Table 10: Evaluation results for the "uncategorised" category.

Uncategorized	
Precision	0.99
Recall	0.57
F1_score	0.72
Accuracy score	0.90

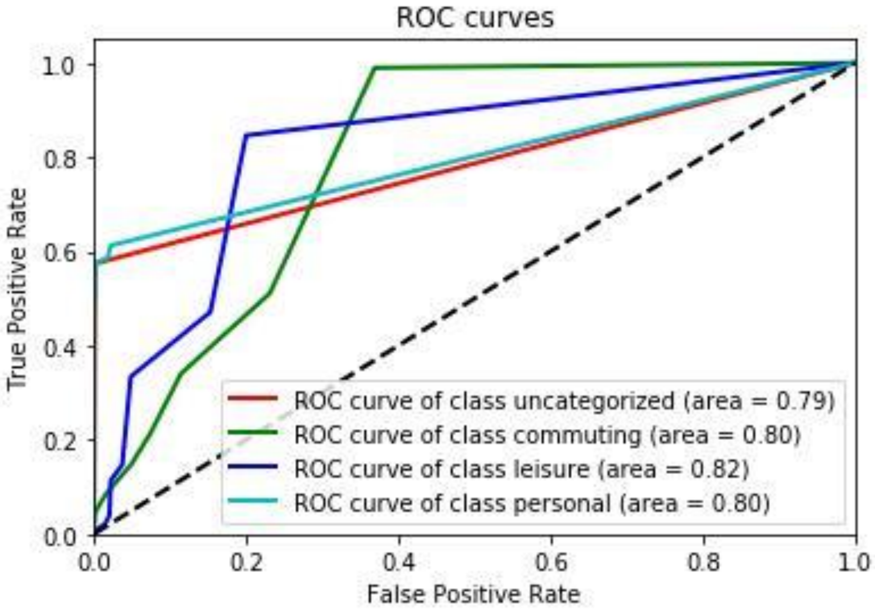


Figure 11: ROC curves diagram presenting the ROC curves of each category. The x-axis presents the false positive rates of the sample and the y-axis presents the true positive rate.

1. Multiclass classification

In this kind of evaluation we considered that there are 4 classes (i.e. commuting, leisure, personal, uncategorized). The dataset used for the evaluation is the same. The evaluation metrics calculated for this evaluation were the micro and macro precision, recall and F1 score. Micro calculates the metrics globally by counting the total true positives, false negatives and false positives scores. Though, macro calculates metrics for each label, and finds their unweighted mean. The precision, recall and F1 score in both micro and micro are presented in Table 11, where y defines the set of predicted pairs, $\hat{y}$ defines the set of true pairs, L is the set of labels and y_l is the subset of y with label l.

Table 11: Evaluation metrics on the multiclass evaluation

	Precision	Recall	F1 score
Micro	$P(y,\hat{y}) = \frac{\lvert y \cap \hat{y}\rvert}{\lvert y\rvert}$	$R(y,\hat{y}) = \frac{\lvert y \cap \hat{y}\rvert}{\lvert \hat{y}\rvert}$	$F_1(y,\hat{y}) = 2 * \frac{P(y,\hat{y}) \times R(y,\hat{y})}{P(y,\hat{y}) + R(y,\hat{y})}$
Macro	$\frac{1}{\lvert L\rvert}\sum_{l \in L} P(y_l,\hat{y}_l)$	$\frac{1}{\lvert L\rvert}\sum_{l \in L} R(y_l,\hat{y}_l)$	$\frac{1}{\lvert L\rvert}\sum_{l \in L} F_1(y_l,\hat{y}_l)$

The equal error rate (EER) was, also, calculated which indicates the common value of the false acceptance rate and the false rejection rate, when the rates are equal. The lower the equal error rate value, the higher the accuracy of the system

The results of the computed evaluation metrics are presented in the next table:

Results	
Precision macro	0.801
Precision micro	0.778
Recall macro	0.706
Recall micro	0.778
F1 score macro	0.735
F1 score micro	0.778
EER macro	0.271
EER micro	0.155

Moreover, the micro- and the micro- average ROC curve has been calculated as the Figure 12 displays, as well as the areas under the ROC curves.

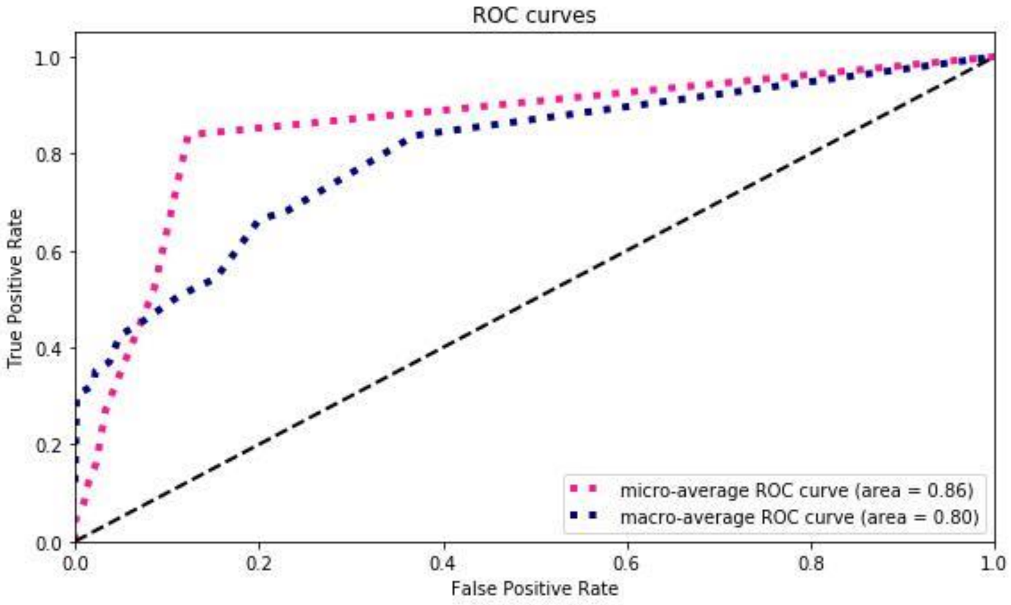


Figure 12: ROC curves diagram. The x-axis presents the false positive rate of the sample and the y-axis presents the true positive rate.

Alongside, the confusion matrix, displaying the performance of the algorithm, has been created. Each row of the confusion matrix represents the instances in a predicted class while each column represents the instances in an actual class (according to the ground truth).

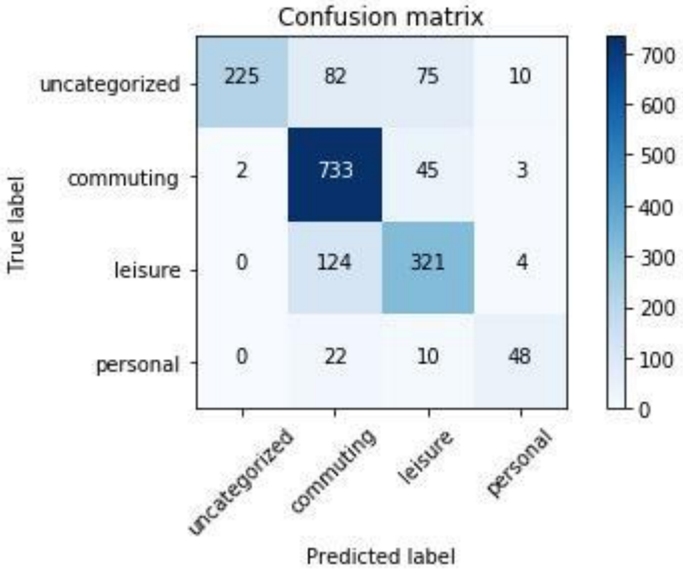


Figure 13: Confusion matrix for the recognized categories of activities.

Apart from the two different evaluation kinds referred, a confidence diagram has been created, presented in Figure 14, as the result of the activity detection algorithm and displays the confidence value on the x-axis and the value of the number of tweets on each category on the y-axis. Confidence reflects the reliability existing regarding the length of tweet. In order the confidence metric to be calculated, the length of each tweet and the maximum length of the whole sequence of tweets (number of words) is taken into account.

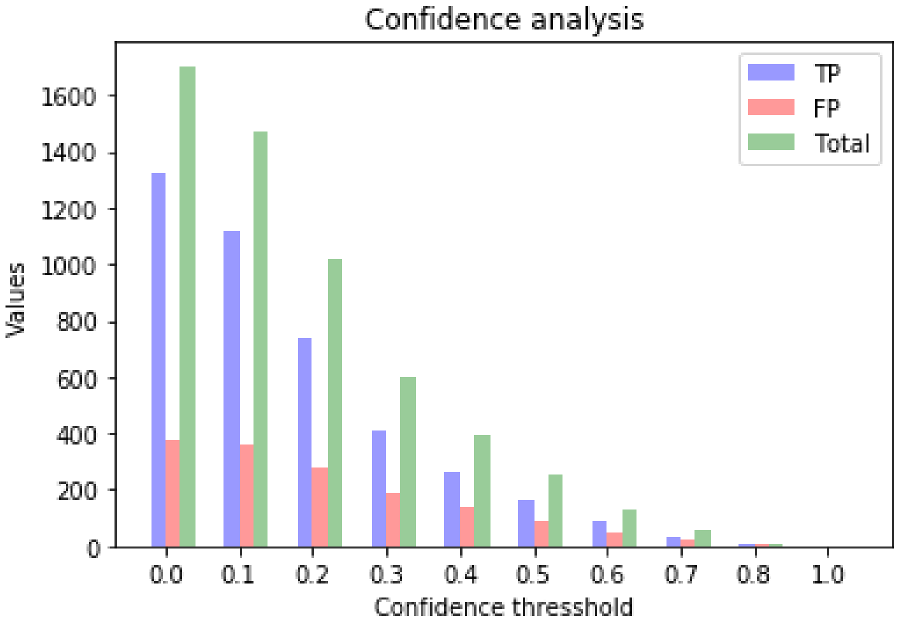


Figure 14: Confidence diagram resulting from the evaluation process.

3.5 Sequences Creation

Based on the activities recognised from the tweets analysis, sequences of activities are identified for each user. Thus, for each user a table describing the sequence of his activities is created. The Table 12 provides a sample sequence of activities for a user. The table contains information about the time of the user's activity and the category of the activity. Regarding the first it is a string value that represents the date and time that the tweet was published, while the activities are those described in Section 3.1.

Table 12: Sample of an identified user's activities sequence

Date-Time	Activity
2018-05-29 08:02:20	Leisure
2018-05-29 09:00:00	Commuting
2018-05-29 18:14:35	Personal
2018-05-29 22:14:35	Leisure
2018-05-30 09:20:35	Commuting
…	**…**
2018-06-10 10:14:35	Leisure

3.6 USERS PROFILING MECHANISM

The method designed by USAL aims to identify the type of activities that a user performs based on the information they share on their social networks, in this case we focus on the social network Twitter.

To be able to classify the information shared by the user in a series of activities, the analysis is based on the TF-IDF algorithm, mainly because of its speed, which allows classifying information quickly from dictionaries of relevant words already created.

Therefore, it is necessary to create a series of dictionaries for each language and for each category of activities to be analysed.

In this case, as the dictionaries can change according to the language and the type of activities can change over time, a dynamic mechanism has been developed that is able to build dictionaries automatically.

For each category, the system requires a set of keywords associated with the category. Then, with each of the keywords, the system performs a Google search and analyses the first n results returned by the search engine. It does not only analyse the snippet that Google displays, but also it accesses the site and extracts its content.

To perform the tests, we created dictionaries related to:

- Education
- Housework
- Religion
- Entertainment
- History/Culture
- Personal care
- Eating & drinking
- Shopping
- Work

For instance, with the dictionary "Education", the 15 first keywords are: education, English, social, subjects, primary, school, research, home, curriculum, students, policy, content, studies, information and system.

With "Housework" we obtained: home, equipment, household, cleaning, clothes, chores, vacuum, house, clean, contact, washing, floor, sweep, table and laundry.

This crawler doesn't use any API to prevent high pricing of Google.

The app simply uses the URL java method and queries Google Search itself with keywords chosen by the developer, then we retrieve results links in the HTML code and download each web page code. After removing all tags and unnecessary data, we do exactly the same TF-IDF analysis to build our database.

Before determining the most relevant terms with the TF-IDF algorithm (which is mainly based on the frequency of occurrence), a pre-processing is performed to remove punctuation and stop words (adjectives, pronouns, etc.) and stemming is done, so that only the part of the word that provides the meaning (without plurals, without suffixes, etc.) is used.

Users' activities are associated with activities with more common terms, always with a minimum of similarity.

Figure 15 summarizes the proposed platform schema.

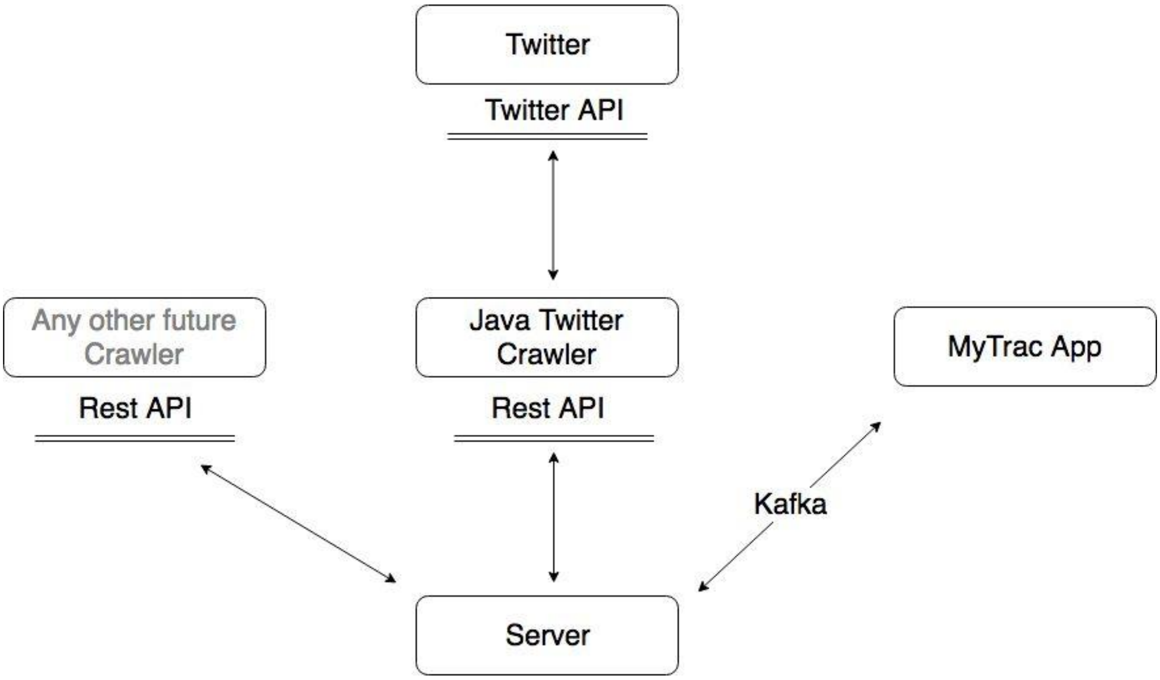


Figure 15. Platform schema

The Crawler in the middle communicates with the Twitter API; this Crawler implements a Rest API so we can trigger the analysis from another program. The server communicates with My-TRAC app (or a module used to obtain this information), this same server could communicate with any other future crawler working with social networks, in order to extend possibilities of the system.

In order to communicate with the server using Kafka, an authentication is required. Client has to send a request token with a randomly generated string; the server receiving this message will calculate the hash of this string, generates a token and send them both back to the client. Finally, the client calculates its hash based on the initial string and checks if it is equal to the hash result received. If so, the token is valid and can be used with every future communication. This workflow can be seen in Figure 16.

For each token sent by the client, the server will check the token attached and compare it with the currently authorized token.

As we have not implemented any persistent dataset of authorized users such as a database, the system is considered opened for now.

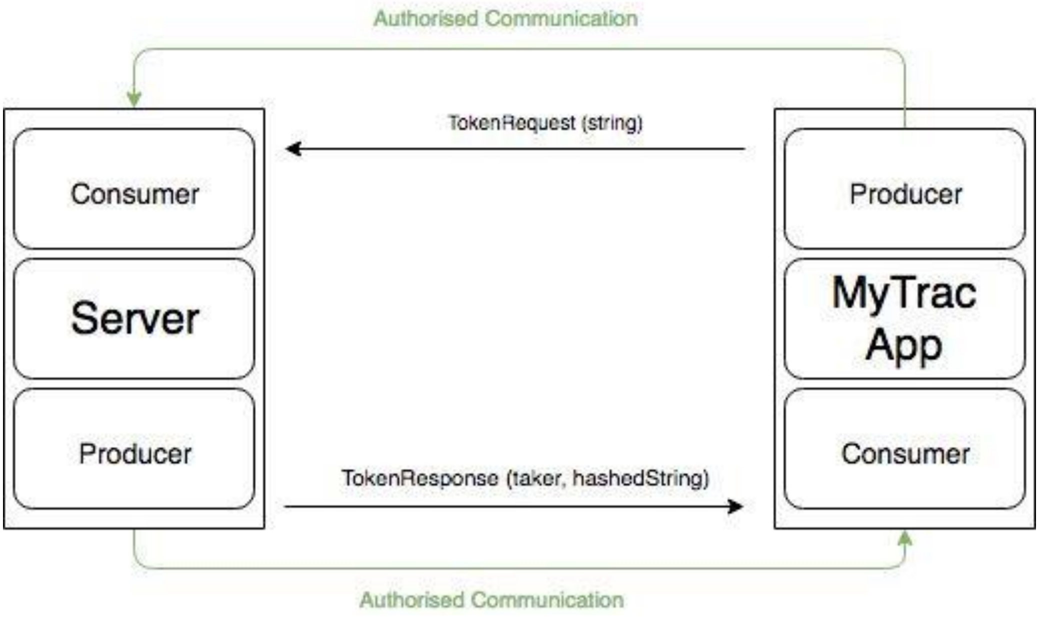


Figure 16. Communication schema.

Apache Kafka is used to communicate; it is an asymmetrical producer-consumer system, so both client and server will have to be producer and consumer in order to have symmetrical communication.

This system implies multithreading programming in both sides; however, we faced some significant but inconstant delays using this method. Messages exchanged can either be real-time or with more than a 20 second lag.

From the information previously extracted with the Google crawler, which allows you to create dictionaries for any type of activity dynamically and from the information extracted from the account of a Twitter user, who has provided permission and its user to analyse their preferences independently, you can create a profile of each user.

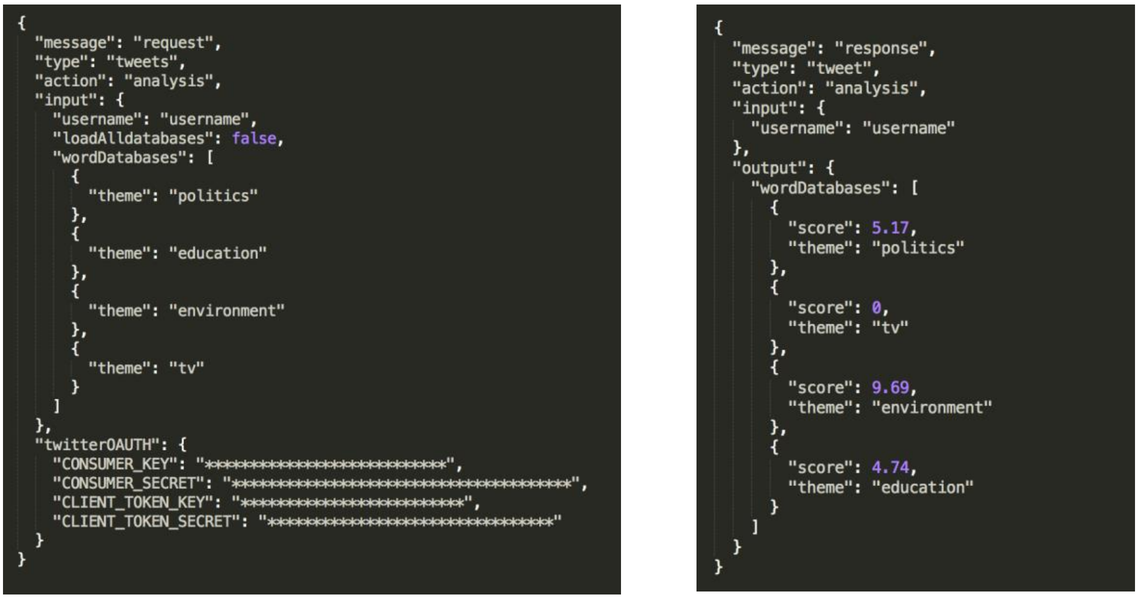
This profile links the information published with each category and provides a metric based on the frequency of publication on each topic.

In order to start the search on a profile, a request must be made by sending a JSON with the necessary information.

For every JSON file we will look at the attribute "type" to determine whether we want to analyse tweets, stream or create a word database.

Analysing tweets:

We can see with the examples shown in Figure 17 the endpoints of a tweets analysis. The attribute "action" is for the moment only "analysis", but this attribute can be used in the future to perform other kind of actions on historical tweets.



```
{
  "message": "request",
  "type": "tweets",
  "action": "analysis",
  "input": {
    "username": "username",
    "loadAlldatabases": false,
    "wordDatabases": [
      {
        "theme": "politics"
      },
      {
        "theme": "education"
      },
      {
        "theme": "environment"
      },
      {
        "theme": "tv"
      }
    ]
  },
  "twitterOAUTH": {
    "CONSUMER_KEY": "*************************",
    "CONSUMER_SECRET": "************************************",
    "CLIENT_TOKEN_KEY": "*************************",
    "CLIENT_TOKEN_SECRET": "********************************"
  }
}
```

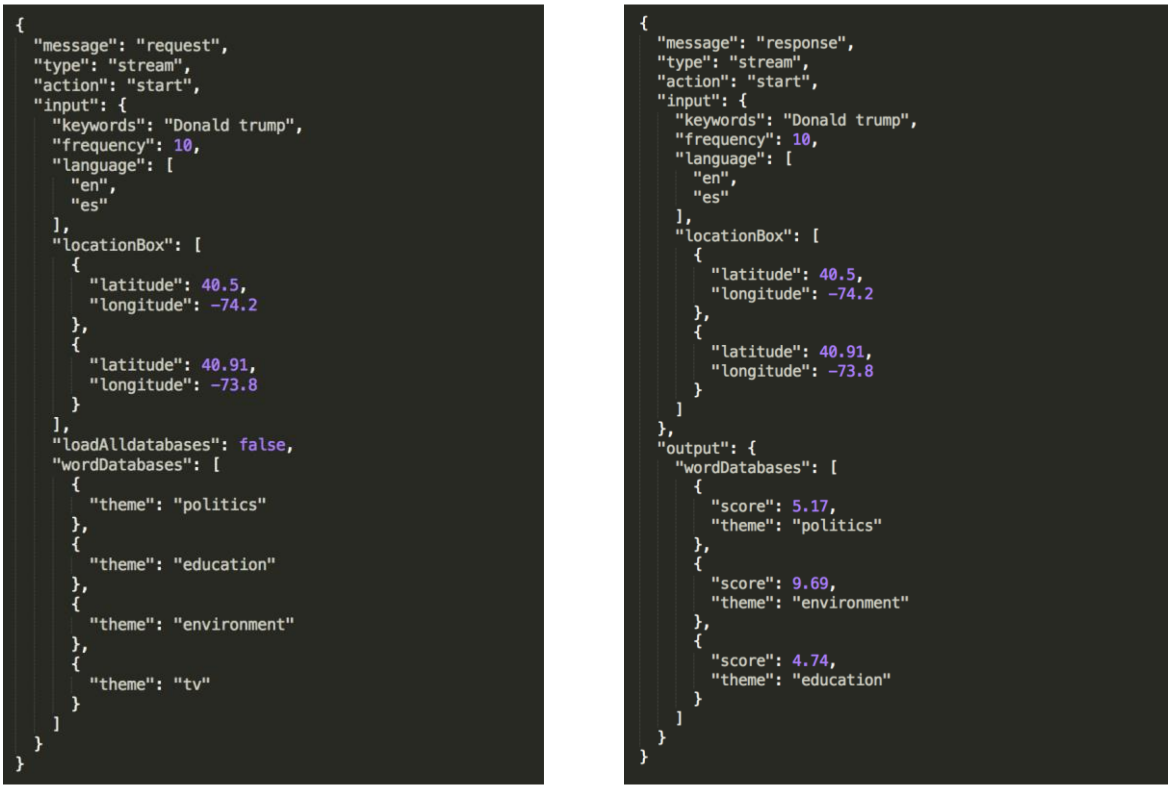
```
{
  "message": "response",
  "type": "tweet",
  "action": "analysis",
  "input": {
    "username": "username"
  },
  "output": {
    "wordDatabases": [
      {
        "score": 5.17,
        "theme": "politics"
      },
      {
        "score": 0,
        "theme": "tv"
      },
      {
        "score": 9.69,
        "theme": "environment"
      },
      {
        "score": 4.74,
        "theme": "education"
      }
    ]
  }
}
```

Figure 17. JSON examples for tweet profile analysis. Request (left) and response (right).

We can choose whether we want the analysis with all word databases or only some of them, finally we need the username and the various Twitter tokens.

Stream analysis:

As can be seen in Figure 18, the streaming endpoints look similar as the previous figure, as inputs you need the keywords, the frequency of analysis (here 10 means that we want an analysis every 10 tweets, but it can 100, 1000, 10000 etc. depending on the subject), which language do we want and the location box. Here we have an example with New York City coordinates.

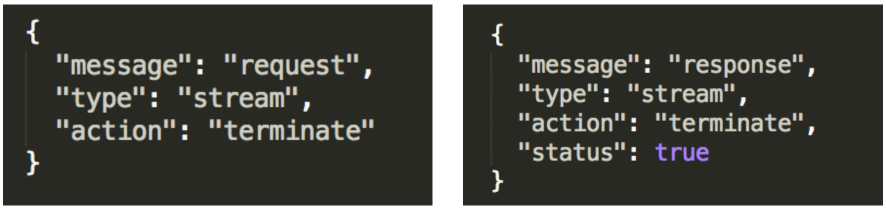


```
{
  "message": "request",
  "type": "stream",
  "action": "start",
  "input": {
    "keywords": "Donald trump",
    "frequency": 10,
    "language": [
      "en",
      "es"
    ],
    "locationBox": [
      {
        "latitude": 40.5,
        "longitude": -74.2
      },
      {
        "latitude": 40.91,
        "longitude": -73.8
      }
    ],
    "loadAlldatabases": false,
    "wordDatabases": [
      {
        "theme": "politics"
      },
      {
        "theme": "education"
      },
      {
        "theme": "environment"
      },
      {
        "theme": "tv"
      }
    ]
  }
}
```

```
{
  "message": "response",
  "type": "stream",
  "action": "start",
  "input": {
    "keywords": "Donald trump",
    "frequency": 10,
    "language": [
      "en",
      "es"
    ],
    "locationBox": [
      {
        "latitude": 40.5,
        "longitude": -74.2
      },
      {
        "latitude": 40.91,
        "longitude": -73.8
      }
    ]
  },
  "output": {
    "wordDatabases": [
      {
        "score": 5.17,
        "theme": "politics"
      },
      {
        "score": 9.69,
        "theme": "environment"
      },
      {
        "score": 4.74,
        "theme": "education"
      }
    ]
  }
}
```

Figure 18. JSON examples for tweet streaming analysis. Request (left) and response (right).

To terminate the stream the endpoints shown in Figure 19 can be used:



```
{
  "message": "request",
  "type": "stream",
  "action": "terminate"
}
```

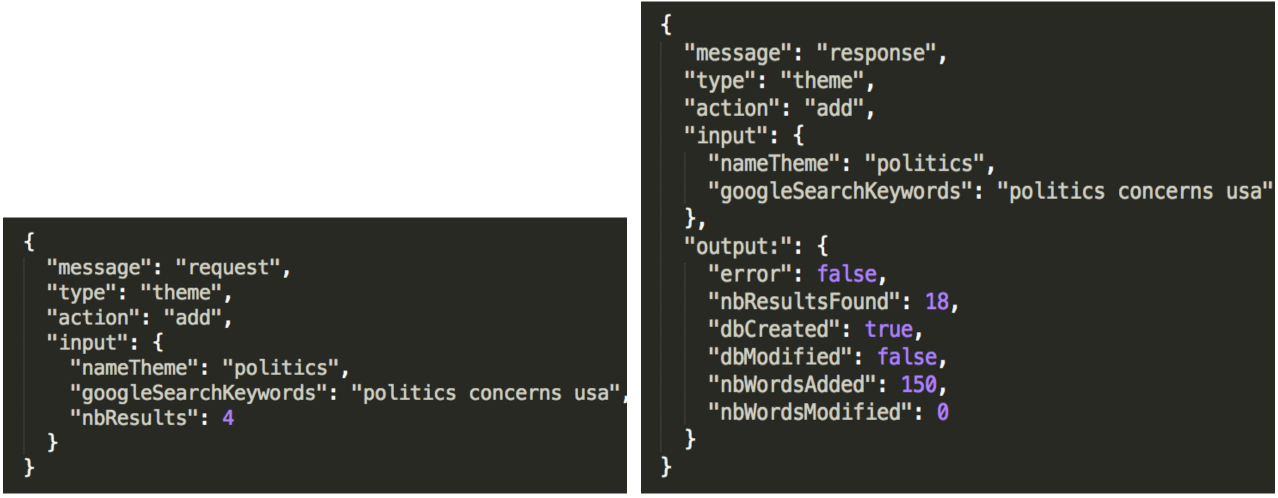
```
{
  "message": "response",
  "type": "stream",
  "action": "terminate",
  "status": true
}
```

Figure 19. JSON examples to stop the streaming analysis. Request (left) and response (right).

Creating databases:

Figure 20 shows how a database can be created giving the name, the Google Search keywords and the maximum number of web pages to crawl. The difference between the maximum and the actual number of pages crawled depends on how well the crawler parse the URL or read the HTML code, sometimes exceptions are triggered, and some pages cannot be read.

Once again, the "action" attribute is for now only "add" as can be seen in Figure 20. There is not the option to remove a database, however if the database already exists, it will be modified and not overwritten.



```
{
  "message": "request",
  "type": "theme",
  "action": "add",
  "input": {
    "nameTheme": "politics",
    "googleSearchKeywords": "politics concerns usa",
    "nbResults": 4
  }
}
```

```
{
  "message": "response",
  "type": "theme",
  "action": "add",
  "input": {
    "nameTheme": "politics",
    "googleSearchKeywords": "politics concerns usa"
  },
  "output:": {
    "error": false,
    "nbResultsFound": 18,
    "dbCreated": true,
    "dbModified": false,
    "nbWordsAdded": 150,
    "nbWordsModified": 0
  }
}
```

Figure 20. JSON examples to create the databases for every category. Request (left) and response, including a summary of the results (right).

3.6.1 Evaluation of the users' profiling mechanism

A prior data collection process was necessary to evaluate the quality of the proposed system. Specifically, a set of Twitter posts from several users with different interests is needed to check the accuracy of the classification algorithm. To conduct the validation process, several Twitter official profiles have been manually selected. These profiles have been chosen in such way that their main interests match one or several of the categories considered for this problem. Even though the proposed system supports several languages, as the project is developed in English, the evaluation process uses information from 20 Twitter profiles published in this language.

The selected Twitter accounts are presented hereunder:

- **Emma Watson (@EmmaWatson):** English actress, model, and activist.

- **Oprah Winfrey (@Oprah):** American actress, producer, philanthropist, businesswoman and television presenter.
- **Donald Trump (@realDonaldTrump):** current President of the United States.
- **Elon Musk (@elonmusk):** business magnate, investor and engineer.
- **Oxford University (@UniofOxford):** one of the leading universities in the world placed in the United Kingdom.
- **Pope Francis (@Pontifex):** current leader of the Catholic Church.
- **International Food Information Council (@FoodInsight):** foundation that communicates science-based information on health, nutrition and food safety for the public good.
- **Sephora (@Sephora):** French cosmetics chain.
- **AXE (@AXE):** men's personal care products company.
- **HISTORY (@HISTORY):** official account related to historical events.
- **Kristina Bazan (@KristinaBazanxx):** author, singer and one of the most famous digital influencers in the world.
- **Crissy Page (@Crissy):** writer and blogger interested in food, family and photography.
- **UCLA (@UCLA):** public research university in the district of Los Angeles that belongs to the University of California.
- **IKEA UK (@IKEAUK):** Multinational Swedish group that sells and distributes furniture and home accessories.
- **Bosch Home UK (@BoschHomeUK):** leading global supplier of technology and services.
- **Real Madrid EN (@realmadriden):** Spanish football team.
- **Gary Lineker (@GaryLineker):** English former professional footballer and current sports broadcaster.
- **iRobot (@iRobot):** the world's leading consumer robot company.
- **Save The Children (@SaveTheChildren):** an international non-governmental organization working to promote children's rights.
- **LEGO (@LEGO_Group):** Danish toy company.

The system aims to categorize these profiles according to their interests and considering their most recurrent publication themes. As the set of categories to be considered in My-TRAC has not been definitively established, a set of different and relevant topics is selected for the evaluation of this work. These categories are presented below:

- Education (E)
- Housework (HW)
- Religion (R)
- Entertainment (En)
- History (H)
- Personal care (PC)
- Eating & drinking (ED)
- Shopping (S)
- Work (W)

To this end, we establish a set of dictionaries in which each topic or category is associated to a set of related keywords. The first step in this process is to perform Google searches related to each category. In each case, between 200 and 320 pages of results are analysed by removing the stop words from the language (in this case English). As the dictionaries can be created in any language, the algorithm would work in an equivalent way in any other language. After a frequency analysis of the words obtained in the results of each search, the most relevant ones are selected. This frequency analysis is made by applying the well-known Term Frequency – Inverse Document Frequency (TF-IDF) algorithm. Specifically, this is a numerical measure that expresses how relevant a word is to a document in a collection. It is a combination of two measures: TF which counts the occurrences or frequency (f) of the term (t) in a document (d) and IDF which measures how much information the term t provides across all documents (D).

$$TF(t,d) = \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \quad (7)$$

$$IDF(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (8)$$

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D) \quad (9)$$

Subsequently, a manual revision of the keywords of each category is performed to filter them and eliminate the non-relevant ones. Additionally, keyword matches for distinct categories were analysed. Consequently, results were refined and keyword overlapping in several categories was avoided. This coincidence can be seen in Table 13 which shows the percentage of similarity between categories according to the overlap of their keywords. Specifically, both the rows and the columns refer to the categories considered in the problem. Each cell expresses the percentage of keywords that are repeated from the total of keywords of the two corresponding categories. The highest values are marked in red and orange.

Table 13 - Percentage of similarity between the categories considered (first version of the system)

	E	HW	R	En	H	ED	PC	S	W
E	-	0.00	0.00	0.00	1.94	0.00	1.23	0.83	4.24
HW	0.00	-	0.00	0.00	0.00	1.06	7.75	1.14	1.18
R	0.00	0.00	-	0.00	2.27	0.00	0.55	0.00	0.00
En	0.00	0.00	0.00	-	2.84	0.00	0.00	1.41	1.44
H	1.94	0.00	2.27	2.84	-	0.00	0.00	0.65	1.99
ED	0.00	1.06	0.00	0.00	0.00	-	4.60	0.91	0.92
PC	1.23	7.75	0.55	0.00	0.00	4.60	-	2.48	1.90
S	0.83	1.14	0.00	1.41	0.65	0.91	2.48	-	4.27
W	4.24	1.18	0.00	1.44	1.99	0.92	1.90	4.27	-

As can be seen, the percentage of similarity between categories is undesirably high in some cases. Examples of this are Housework - Personal Care (7.75%), Education – Work (4.24%), Shopping – Work (4.27%) and Eating & Drinking – Personal Care (4.60%). This can lead to problems in the prediction task: if a certain word appears as a keyword for two different categories, it can cause problems when classifying a profile.

To evaluate the quality of the results obtained by the system, it is necessary to have a list of previously classified profiles. The objective of the system is to obtain the percentage of similarity of each profile with each of the categories. For this reason, to evaluate the system, a set of tweets from each profile account is manually analysed selecting the four categories most related to its interests. For example, after the analysis of the tweets of IKEA's profile, the four categories chosen are: Housework, Shopping, Eating & drinking and Entertainment. In this way, a set of data consisting of the user names, a selection of tweets published in each account and four manually assigned categories is generated.

After obtaining the keywords for each category and carrying out the manual tagging process, the proposed system analyses, in each case, 3000 tweets. In this information, stop words are eliminated and the most relevant terms are extracted through a second frequency analysis. This analysis is a bit different as the previous one: we only use the IDF part of the TF-IDF method. Indeed, a tweet is only 140 characters long, and up to 280 for the latest ones. This means that the TF part equals number of occurrences in a single/number total of tweets, so basically 1/3000. We ended up having insignificant results considering the TF, so we choose to only measure the importance of each word in all documents with the IDF method.

Then, these words are compared with the keywords of the dictionaries defined for the distinct categories. The purpose is to obtain a percentage of membership for each profile in each category. In fact, the output of the algorithm is not a simple category, but a list of the most relevant categories for each profile and the estimated percentage of affinity for each of them.

The corresponding manual and automatic sorting processes lead to the results that are used to evaluate the system. Table 14 shows the manual classification for each of the selected Twitter profiles (the four most relevant in each case are considered), and Table 15 shows the classification obtained by the system. Additionally, the results of the automatic classification are shown in green if they match the manual classification and in red otherwise. In this way, the successes of the algorithm can be seen more easily. It is important to note that the order of classification is not considered when determining whether a hit or failure has occurred; the coincidence of manually and automatically estimated categories for a profile is considered, but not the order in which they appear.

Table 14 - Manual classification for the 20 Twitter profiles selected in the study

	CATEGORIES			
Emma Watson	Entertainment	Work	Education	Religion
Oprah Winfrey	Entertainment	Religion	Education	Work
Donald Trump	History	Work	Education	Entertainment
Oxford University	Education	Work	Entertainment	FoodDrink
Pope Francis	Religion	History	Education	FoodDrink
Food Information	FoodDrink	Education	PersonalCare	Shopping
Sephora	PersonalCare	Shopping	Entertainment	Work
History	History	Education	Entertainment	Religion
Kristina Bazan	Shopping	Entertainment	PersonalCare	FoodDrink
Crissy Page	FoodDrink	PersonalCare	Shopping	Entertainment
Bosch Home UK	Housework	Shopping	FoodDrink	Work
Real Madrid EN	Entertainment	FoodDrink	Shopping	PersonalCare
Gary Lineker	Entertainment	Work	PersonalCare	FoodDrink
AXE	PersonalCare	Shopping	Entertainment	Work
Elon Musk	Entertainment	Education	Work	Shopping
UCLA	Education	Work	Entertainment	FoodDrink
IKEA UK	Housework	Shopping	FoodDrink	Entertainment
iRobot	Housework	Entertainment	Shopping	Education
Islamic Relief	Religion	FoodDrink	History	Education
Total	Work	Shopping	Education	FoodDrink

Table 15 - Automatic classification for the 20 Twitter profiles selected in the study (first version of the system)

	CATEGORIES			
Emma Watson	Entertainment	History	PersonalCare	Religion
Oprah Winfrey	Entertainment	Religion	PersonalCare	History

	CATEGORIES			
Donald Trump	History	Work	Shopping	PersonalCare
Oxford University	Education	Work	Entertainment	History
Pope Francis	Religion	History	PersonalCare	Entertainment
Food Information	FoodDrink	Education	PersonalCare	History
Sephora	PersonalCare	Shopping	History	Work
History	History	Education	Entertainment	Work
Kristina Bazan	Shopping	Entertainment	PersonalCare	FoodDrink
Crissy Page	FoodDrink	PersonalCare	Shopping	History
Bosch Home UK	PersonalCare	Shopping	History	Work
Real Madrid EN	Entertainment	History	Shopping	Work
Gary Lineker	Entertainment	Work	PersonalCare	History
AXE	PersonalCare	Shopping	Entertainment	FoodDrink
Elon Musk	Entertainment	History	Work	Shopping
UCLA	Education	History	Entertainment	PersonalCare
IKEA UK	PersonalCare	Shopping	Work	History
iRobot	Housework	Entertainment	Work	PersonalCare
Islamic Relief	Religion	PersonalCare	History	Work
Total	Work	Shopping	Education	Entertainment

Bearing in mind that there are categories for which some keywords appear simultaneously in different dictionaries (there is a contextual overlap), some classification errors seem logical. On several occasions the system classifies as Personal Care what was manually established as Eating & Drinking (in the profiles of Pope Francis, Bosch, UCLA, IKEA and Islamic Relief). These categories had one of the highest percentages of similarity in Table 13. Thus, the misclassifications are strongly affected by the similarity of dictionaries.

Another interesting case is the classification of Crissy Page's profile. She is a blogger who daily deals with food, family and lifestyle. For this reason, the proposed manual classification for this account considered the categories of Eating & Drinking, Personal Care, Shopping and Entertainment. Three of the proposed categories coincide in the system's prediction; however, instead of relating this person's publications to the Entertainment category, the algorithm proposes the History one. The same misclassification occurs in the case of the Sephora account. According to Table 13, this error is understandable because the keywords of these dictionaries have some overlap.

On the other hand, the manual classification of accounts involves a subjective component by having to choose the four categories most related to each account. In some cases, three of the categories seem clear, but adding a fourth is complicated because there is no one that fits perfectly. In other cases, there are more than four categories that could be worthwhile, and it is necessary to decide which are the most appropriate.

The four most relevant categories obtained by the algorithm are extracted in each case and we compare them to the four manually selected categories. As there are 80 cases (20 users and 4 categories for each one), we manually calculate the success rate by taking into account the number of matches between the two classification processes.

Despite all the issues previously explained, the success rate of the algorithm is 62.50%. Considering that the study is a multi-label classification problem (it is necessary to match four categories for each profile to obtain 100% accuracy), the results obtained are satisfactory.

If we consider only the two most relevant categories instead of the four ones, the success rate reaches a slightly higher value. Specifically, 26 out of 40 categories are correct in this case, which implies 65% accuracy. This proves that the nature of multi-label problems is much more complex than the simple classification ones.

However, we propose a second version of the system in which the accuracy of the prediction is increased. In this second proposal, the system developed is refined through an analysis of the keywords obtained from the user profiles and their relationship with the category dictionaries. After this analysis, all the non-relevant words are removed from the dictionaries. Additionally, keywords obtained from user profiles that were not included in any category are added to the corresponding dictionaries. Finally, the prediction is made again, and the accuracy is analysed with the same procedure as in the previous case. The new overlap values between the keywords of the updated dictionaries can be seen in Table 11.

Table 16 - Percentage of similarity between the categories considered (second version of the system)

	E	HW	R	En	H	ED	PC	S	W
E	-	0,00	0,00	0,00	0,00	0,00	0,96	0,59	5,49
HW	0,00	-	0,00	0,00	0,00	0,87	5,73	0,85	0,88
R	0,00	0,00	-	0,00	0,00	0,00	0,00	0,00	0,00
En	0,00	0,00	0,00	-	0,56	0,00	0,00	1,00	0,51
H	0,00	0,00	0,00	0,56	-	0,00	0,00	0,00	0,00
ED	0,00	0,87	0,00	0,00	0,00	-	4,43	0,36	0,37
PC	0,96	5,73	0,00	0,00	0,00	4,43	-	0,99	1,01
S	0,59	0,85	0,00	1,00	0,00	0,36	0,99	-	3,14
W	5,49	0,88	0,00	0,51	0,00	0,37	1,01	3,14	-

The percentage of similarity between the categories is calculated from the number of repetitions of the keywords in the dictionaries. In general terms, as we can see in Table 16, the overlap problem has been reduced.

After updating the keywords of the dictionaries, a new version of the system is obtained. Table 17 shows the automatic classification made by the new system. If successful, the results are displayed in green, and they are written in red otherwise.

Table 17 - Automatic classification for the 20 Twitter profiles selected in the study (second version of the system)

	CATEGORIES			
Emma Watson	Entertainment	Work	Education	Shopping
Oprah Winfrey	Entertainment	Religion	Education	Work
Donald Trump	History	Work	Shopping	Entertainment
Oxford University	Education	Work	Entertainment	History
Pope Francis	Religion	History	Entertainment	Shopping
Food Information	FoodDrink	Education	Work	PersonalCare
Sephora	PersonalCare	Shopping	Education	Work
History	History	Education	Entertainment	Shopping
Kristina Bazan	Shopping	Entertainment	PersonalCare	FoodDrink
Crissy Page	FoodDrink	PersonalCare	Shopping	Entertainment
Bosch Home UK	Housework	Shopping	Education	Work
Real Madrid EN	Entertainment	Work	Shopping	Education
Gary Lineker	Entertainment	Work	Shopping	History
AXE	PersonalCare	Shopping	Entertainment	FoodDrink
Elon Musk	Entertainment	Education	Work	Shopping
UCLA	Education	Work	Entertainment	PersonalCare
IKEA UK	Work	Shopping	FoodDrink	PersonalCare
iRobot	housework	Work	Shopping	PersonalCare
Islamic Relief	Religion	FoodDrink	Work	Education
Total	Work	Shopping	Education	FoodDrink

After an in-depth analysis of keyword matching in tweets and dictionaries, it is concluded that most of the misclassifications are due to the dictionaries of Work and Shopping as can be seen in Table 17. According to Table 17 and Table 16, these dictionaries have high similarity values in all cases in general terms. Thus, the classification process involving these dictionaries is more complex.

Despite these issues, in this second version of the system 60 out of 80 categories are successfully predicted, what leads to an accuracy of 75%. As the validation of the system is performed as a multi-label problem, the results are quite satisfactory.

A system that considers a set of topics and builds a dictionary of keywords for each of them has been developed. Additionally, a classification of several Twitter official accounts based on their interests has been performed. A validation process has been carried out leading to satisfactory results.

3.7 COMPUTATIONAL PROCEDURE FOR PROPOSING PREFERRED ACTIVITIES

The results of the user profiling, which is the user's preferences as well as the recognized sequences of activities for each user, are used as input to the proposed model for predicting the user's next activities. The used model is a Markov chain.

A Markov chain is "a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event(s)". There are different types of Markov chains as displayed in Table 18.

Table 18: Variations of Markov Chains

	Countable state space	**Continuous or general state space**
Discrete-time	(discrete-time) Markov chain on a countable or finite state space	Harris chain (Markov chain on a general state space)
Continuous-time	Continuous-time Markov process or Markov jump process	Any continuous stochastic process with the Markov property, e.g., the Wiener process

In the proposed model we will use a discrete-time Markov chain on a finite state space. Each state is defined as a user's activity. For each user, the probability of moving from one state to another will be based on the sequence of activities of the mean user and the results of the user profiling, as well as on weather data and data considering emergent events.

Figure 21, shows an example of a 1st order Markov chain used for the prediction of the mean user's next activity. Each state stands for a user activity as defined in Section 3.1 (i.e. C: commuting, L: Leisure, P: personal).

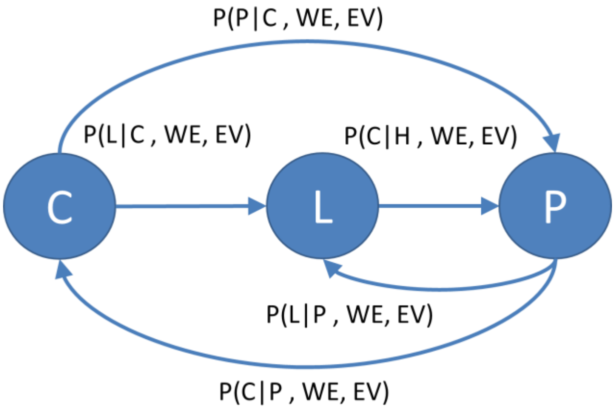


Figure 21: Example of 1st order Markov chain for the mean user

In order to predict the next activity of a user, the probabilities for moving from one state to another must be calculated. The calculation of which is based on the sequence of activities of all users. For each user, we define a sequence of activities as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_M], \boldsymbol{where}\ \boldsymbol{x}_t \in \boldsymbol{A}$, where $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots \boldsymbol{a}_N\}$the set of the activities performed by user, A = {C: "commuting", L: "Leisure", P: "Personal"}.

The transition probability for moving from a state x_t to state x is calculated as:

$$P(x|x_t, we, ev) = \frac{C^x_{x_t,we,ev}}{\sum_{x\in A} C^x_{x_t,we,ev}} \tag{7}$$

Where $C^x_{x_t,w,e}$are the counts of transitions from activities of type x_t to activities of type$x$ within the activity sequences of all users for the weather we and event ev.

The predicted next mean user's activity, x_{t+1} is computed as:

$$x_{t+1} = \underset{x\in A}{argmax}\, P(x|x_t, we, ev) \tag{8}$$

Based on the above a general model that describes the transitions of the mean user from one activity to another is generated. This model can be used in conjunction with the results of the user profiling in order to predict a user's next activity. Having information about a user's interests from the user profiling, the model of Figure 21 is updated to the model of Figure 22.

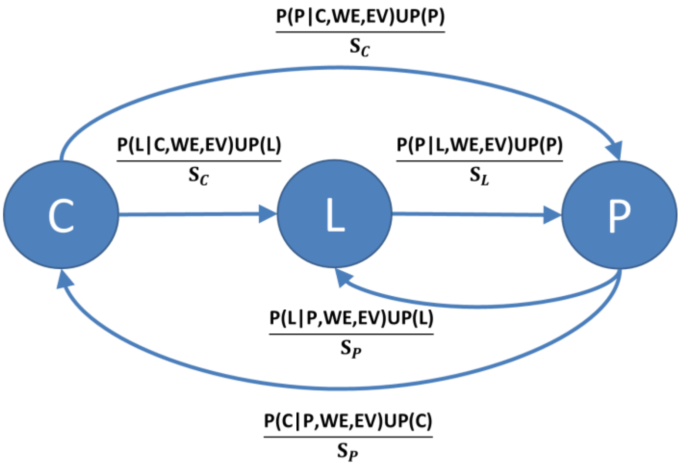

$\frac{P(P|C,WE,EV)UP(P)}{S_C}$
$\frac{P(L|C,WE,EV)UP(L)}{S_C}$
$\frac{P(P|L,WE,EV)UP(P)}{S_L}$
C
L
P
$\frac{P(L|P,WE,EV)UP(L)}{S_P}$
$\frac{P(C|P,WE,EV)UP(C)}{S_P}$

Figure 22: Example of 1st order Markov chain for the specific user

In order to predict the next activity x_{t+1} of a specific user based on model of Figure 22, the following equation is used:

$$x_{t+1} = \underset{x\in A}{argmax}\left(\frac{P(x|x_t, we, ev)UP(x)}{S_{x_t}}\right) \tag{9}$$

Where $P(x|x_t, we, ev)$: is the transition probability as defined in Equation (7), $UP(x)$: is the normalized resulted value of the user profiling, in the range from 0 to 1, and $S_{x_t} = P(x|x_t, we, ev)UP(x)$. The denominator is used in order to normalize the fraction so that it represents probability (i.e. to be in range of 0 to 1).

3.8 Evaluation

For the evaluation of the Markov model, a synthetic dataset was composed using the data received from Twitter. In addition, we enriched them with information from other sources. Specifically, we downloaded data regarding the weather conditions of each tweet from https://developer.worldweatheronline.com/ and event data from https://app.predicthq.com/events, based on each tweet's timestamp. For these kind of data we were only interested in the weather conditions around the location of the tweets and if there were any occurrences of events in the area.

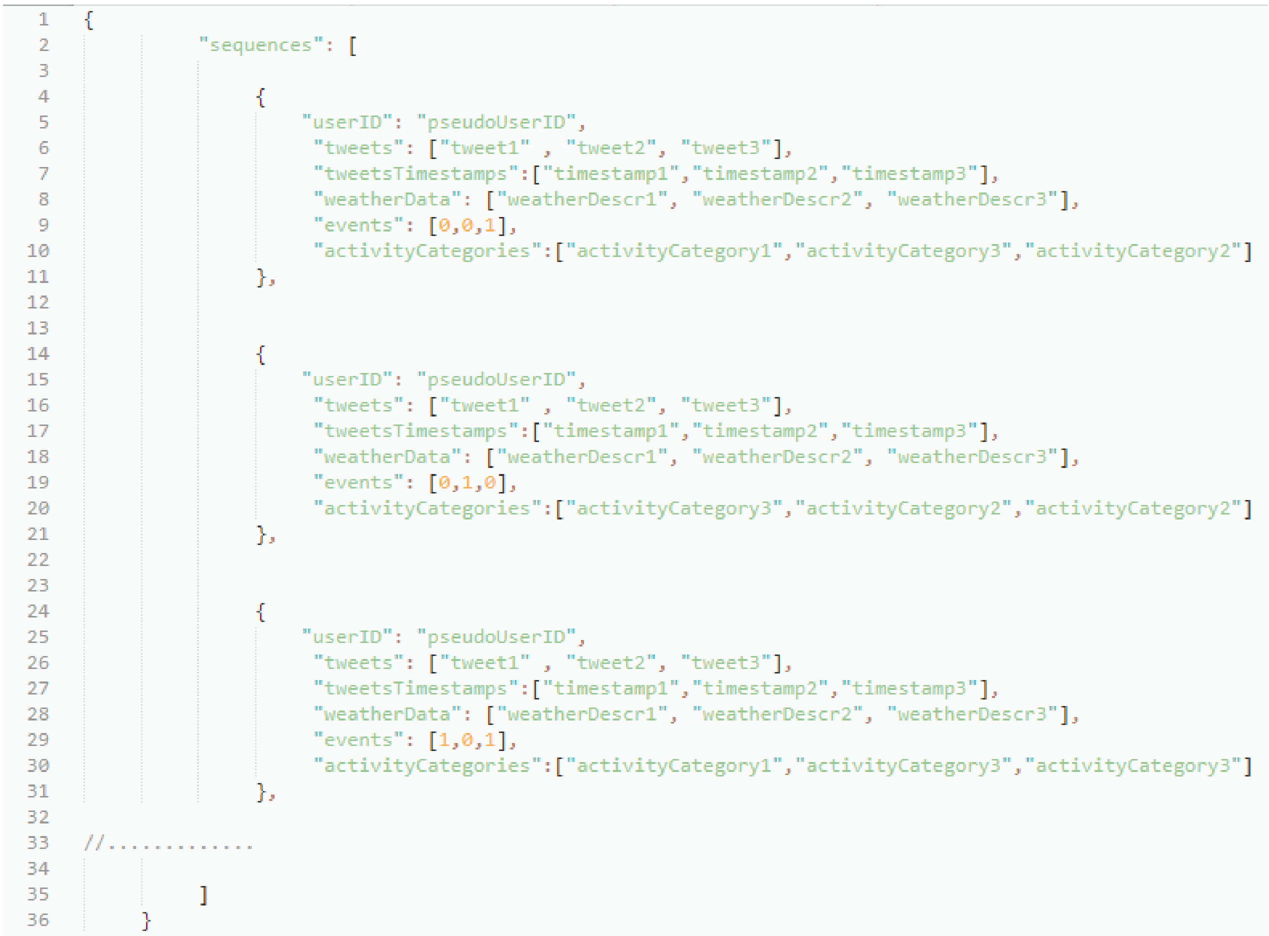
Both source sites' billing is subscription based, but they do provide the option of trial accounts, which is the way we obtained these data.

The goal was to have a set of users with a sequence of activities for each one of them, along with additional information for each sequence as described below. From the body of annotated tweets (1704) we managed to find and identify only three users with a sequence of activities (one user with 10 tweets, one user with 8 tweets and one user with 5 tweets). Due to the lack of more users with identified sequences, we created a synthetic dataset which includes 80 users (77 users with synthetic sequences and 3 users with identified sequence of activities), where each user has twitted 5 times throughout one day. The procedure of creating the synthetic sequences refers to the merging of the rest existing annotated data, in order to create a sequence of tweets per each user. The kind of data we used for this task are: IDs, Timestamp, Text, Activities, Weather, Event, as shown in Table 19.

Table 19: Description of synthetic data

Field	Type	Description
UserID	String	ID that represents the user
Timestamp	String	Date and time of the published text
Text	String	Text published by the UserID in their social accounts (Twitter)
Activities	String	UserID's ongoing activity as detected from the Text
Weather	String	Temperature, Weather conditions(e.g. sunny, cloudy, rainy, snow, windy, etc.), Humidity, based on the time and location of the published Text
Event	Boolean	Ongoing emergency events based on the time of the detected activity. Has a value of 0 or 1.

This synthetic is stored in a JSON files and contains information for all the users, whose activities sequence will be input, for the testing of the Markov model. Its structure and characteristics are shown in Figure 23.



```
{
        "sequences": [

                {
                  "userID": "pseudoUserID",
                   "tweets": ["tweet1" , "tweet2", "tweet3"],
                   "tweetsTimestamps":["timestamp1","timestamp2","timestamp3"],
                   "weatherData": ["weatherDescr1", "weatherDescr2", "weatherDescr3"],
                   "events": [0,0,1],
                   "activityCategories":["activityCategory1","activityCategory3","activityCategory2"]
                },


                {
                  "userID": "pseudoUserID",
                   "tweets": ["tweet1" , "tweet2", "tweet3"],
                   "tweetsTimestamps":["timestamp1","timestamp2","timestamp3"],
                   "weatherData": ["weatherDescr1", "weatherDescr2", "weatherDescr3"],
                   "events": [0,1,0],
                   "activityCategories":["activityCategory3","activityCategory2","activityCategory2"]
                },


                {
                  "userID": "pseudoUserID",
                   "tweets": ["tweet1" , "tweet2", "tweet3"],
                   "tweetsTimestamps":["timestamp1","timestamp2","timestamp3"],
                   "weatherData": ["weatherDescr1", "weatherDescr2", "weatherDescr3"],
                   "events": [1,0,1],
                   "activityCategories":["activityCategory1","activityCategory3","activityCategory3"]
                },

//............

        ]
    }
```

Figure 23: Synthetic dataset

The JSON dataset includes the user's sequences as JSON objects. Each object has nested values or arrays containing information about the aforementioned data fields. Each sequence is linked with a userID. There is an array filled with the recognized tweets that corresponds to that userID. There are also 4 more arrays of the same length:

- **weather data array** – contains a weather description for the location and time of each tweet,
- **timestamps array** – contains the timestamps for the tweets in tweets array
- **event data array** – has a values of 0s and 1s for each tweet depending if there was any occurrence of event during the time the tweet was posted
- **activity categories array** – holds for every tweet, the detected activity category.

In order to evaluate the Markov model the synthetic dataset was used, which was generated considering weather information and the existence of emergency events. The dataset was generated so as when there is no any emergent event the activities performed by the users to be random, while when a emergent event occurs the activities performed by the users to be mainly activities related to the personal category, such as visit hospitals, friends, etc.

The target of this was to prove that the model is able to predict certain patterns of movements from different categories of activities in normal and in emergencies considering also the weather conditions.

Thus, for each one of the possible combination of the last user's activity, the weather data and the existence of an emergent event different probabilities where calculated generate, thus, several Markov models for each one of the combination. Those models are able to predict how the travellers will react under specific conditions. Each state of the Markov model represents a category of activity *C: "Commuting", P: "Personal", L: "Leisure"*, which are highlighted (bold text) in the following Table, while the transition probabilities from state to state, considering also weather data and/or data related to emergent event, are presented in Table 20- Table 23. Due to the lack of real data and the use of synthetic data, a qualitative evaluation was preferred instead of a quantitative one.

Commuting	**Defined as the travelling of some distance between one's home and place of work (or education) on a regular basis**
Personal	**Refers to transfers related to the subcategories of social activities, health-related activities and services-related activities**
Leisure	**Refers to transfers related to cultural, religious, food & drink, and other recreational activities**
Uncategorized	Category for any other option

P(transition|normal)	Commuting	Leisure	Personal
Commuting	0.672	0.245	0.081
Leisure	0.435	0.435	0.129
Personal	0.192	0.307	0.5

Table 20: Transition probabilities for normal conditions

P(transition|event)	Commuting	Leisure	Personal
Commuting	0.571	0	0.428
Leisure	0	0	0
Personal	0.058	0	0.941

Table 21:Transition probabilities during an emergency

P(transition|weather)	Commuting	Leisure	Personal
Commuting	0.720	0.162	0.116
Leisure	0.529	0.411	0.058
Personal	0.090	0.272	0.636

Table 22: Transition probabilities for "Rainy" weather

P(transition|weather, event)	Commuting	Leisure	Personal
Commuting	0.333	0	0.666
Leisure	0	0	1
Personal	0.1	0	0.9

Table 23: Transition probabilities for "Rainy" weather and during emergent event

As a qualitative evaluation of the method we present two example models, Figure 24 presents the trained model for the case that the weather is *"Rainy"* and no unexpected event occurred.

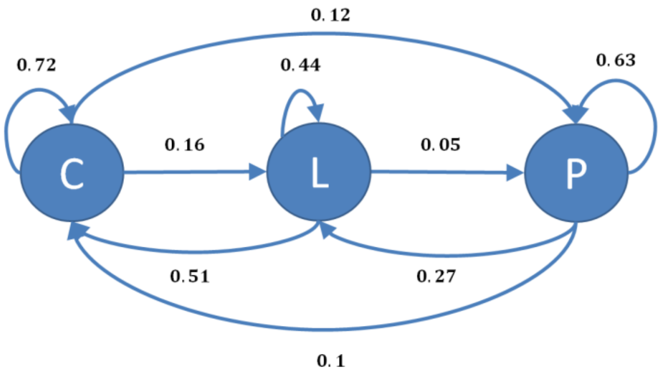


Figure 24: Sample Markov Model for "Rainy" day without emergent event

While for the case that the weather is "Rainy" and there is an unexpected event the model is depicted in Figure 25.

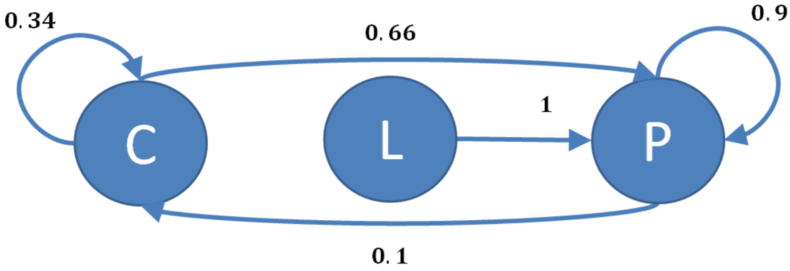


Figure 25: Sample Markov model for a "Rainy" with an emergent event

The calculated probabilities of the Markov models reflect the user's behaviour as described in the generated dataset. Consequently, while in the first example, the prediction of the next activity seems to be random, in the second example model, where the emergent event takes place, the users seem to move toward the personal ("P") activity with higher probability.

4 CONCLUSIONS

This Deliverable focused on two aspects. First, the presentation of existing literature regarding methodologies for collection and analysis of transport-related information from social media platforms and second the design and development of a model for the collection of transport-related information from the Twitter's API. The outcome of the developed model is the prediction of users' activities preferences, by applying algorithms for the (pre-) processing, the classification and the fusion of the collected data.

The literature review regarding transport-related data that can be derived from social media platforms, as well as a literature review for already existing and implemented transport-related data mining models has been presented. There is a variety of transport data that can be collected and analysed from social media platforms. Another important aspect, taken into account in the present Deliverable, is the factors/ attributes, which may affect the users' activities preferences under certain circumstances. Deliverable 2.2 provides information regarding travellers' characteristics, activities' preferences and travellers' preferences; factors which, to a greater or lesser extent, affect the users' choices related to the transportation system.

After the literature review, there is an extended presentation regarding the design and the development of an activity prediction mechanism, which collects transport-related data from social media platforms APIs, analyses the data and finally provides information regarding the users' activities preferences. The developed method of the Deliverable will be integrated in the assignment model for simulating and predicting users' choices and behaviour concerning mode choice, station choice, departure time choice and route choice of Deliverable 2.3.

5 REFERENCES

[1] M. Parameswaran, “Social computing: an overview,” vol. 19, pp. 762–780, 2007.

[2] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, “Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges,” Transp. Res. Part C Emerg. Technol., vol. 75, pp. 197–211, 2017.

[3] “Facebook.” [Online]. Available: www.facebook.com.

[4] “Twitter.” [Online]. Available: www.twitter.com.

[5] “Foursquare.” [Online]. Available: https://foursquare.com/.

[6] E. Chaniotakis, C. Antoniou, and F. Pereira, “Mapping Social media for transportation studies,” IEEE Intell. Syst., vol. 31, no. 6, pp. 64–70, 2016.

[7] X. Zheng et al., “Big Data for Social Transportation,” IEEE Trans. Intell. Transp. Syst., vol. 17, no. 3, pp. 620–630, 2015.

[8] E. D’Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, “Real-Time Detection of Traffic from Twitter Stream Analysis,” IEEE Trans. Intell. Transp. Syst., vol. 16, no. 4, pp. 2269–2283, 2015.

[9] A. Gal-Tzur, S. M. Grant-Muller, T. Kuflik, E. Minkov, S. Nocera, and I. Shoor, “The potential of social media in delivering transport policy goals,” Transp. Policy, vol. 32, pp. 115–123, 2014.

[10] T. Kuflik, E. Minkov, S. Nocera, S. Grant-Muller, A. Gal-Tzur, and I. Shoor, “Automating a framework to extract and analyse transport related social media content: The potential and the challenges,” Transp. Res. Part C Emerg. Technol., vol. 77, pp. 275–291, 2017.

[11] T. Ruiz, L. Mars, R. Arroyo, and A. Serna, “Social Networks, Big Data and Transport Planning,” Transp. Res. Procedia, vol. 18, no. June 2016, pp. 446–452, 2016.

[12] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit, “Social-based traffic information extraction and classification,” nternational Conf. ITS Telecommun., pp. 107–112, 2011.

[13] E. Chaniotakis, C. Antoniou, G. Aifadopoulou, and L. Dimitriou, “Inferring Activities From Social Media Data,” 96th Annu. Meet. Transp. Res. Board, pp. 1–8, 2016.

[14] M. Sinha et al., “Improving Urban Transportation through Social Media Analytics,” CODS ’16 Proc. 3rd IKDD Conf. Data Sci., pp. 1–2, 2016.

[15] E. Chaniotakis, C. Antoniou, and E. Mitsakis, “Data for Leisure Travel Demand from Social Networking Services,” hEART Conf., p. 6, 2015.

[16] R. Mitkov, The Oxford Handbook of Computational Linguistics. Oxford, New York, 2009.

[17] “Twitter API.” [Online]. Available: https://developer.twitter.com/.

[18] E. Chaniotakis and C. Antoniou, “Use of Geotagged Social Media in Urban Settings: Empirical Evidence on Its Potential from Twitter,” IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC, vol. 2015–Octob, no. February 2016, pp. 214–219, 2015.

[19] D. Efthymiou and C. Antoniou, “Use of Social Media for Transport Data Collection,” Procedia - Soc. Behav. Sci., vol. 48, pp. 775–785, 2012.

[20] “The R Project for Statistical Computing.” [Online]. Available: https://www.r-project.org/.

[21] “TwitterXML.” [Online]. Available: https://twitterxml.codeplex.com/.

[22] “TwitterXML BlogSpot.” [Online]. Available: http://twitterxml.blogspot.gr/.

[23] C. Cottrill, P. Gault, G. Yeboah, J. D. Nelson, J. Anable, and T. Budd, “Tweeting Transit: An examination of social media strategies for transport information management during a large event,” Transp. Res. Part C Emerg. Technol., vol. 77, pp. 421–432, 2017.

[24] S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time crisis mapping of natural disasters using social media," IEEE Intell. Syst., vol. 29, no. 2, pp. 9–17, 2014.

[25] "OpenStreetMap." [Online]. Available: https://www.openstreetmap.org/.

[26] "GooglePlaces API." [Online]. Available: https://developers.google.com/places/.

[27] J. Pereira, A. Pasquali, P. Saleiro, and R. Rossetti, "Transportation in social media: An automatic classifier for travel-related tweets," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10423 LNAI, pp. 355–366, 2017.

[28] "Tweepy." [Online]. Available: http://www.tweepy.org/.

[29] I. Toumpalidis, "Physical Spaces and Digital Flows: Navigating through the Informational Matrix," 2017.

[30] I. Toumpalidis and N. Karanikolas, "Spatial Data Mining from Social Media Services," Aristotle University of Thessaloniki, 2015.

[31] I. Toumpalidis and N. Karanikolas, "Spatial Data Analysis from Social Media Services."

[32] J. M. S. Grau, I. Toumpalidis, E. Chaniotakis, N. Karanikolas, and G. Aifadopoulou, "A spatio-temporal correlation between digital and physical world, case study in Thessaloniki."

[33] J. M. S. Grau, E. Chaniotakis, I. Toumpalidis, N. Karanikolas, and G. Aifadopoulou, "Big data for transportation analysis and trip generation."

[34] "Facebook API." [Online]. Available: https://developers.facebook.com/.

[35] United Nations, Department of Economic and Social Affairs, Population Division's World Population Prospects: The 2012 Revision.

[36] Boeing, Current Market Outlook, 2014-2033, p.2

[37] Muhs, C. D., "Understanding Travel Modes to Non-work Destinations: Analysis of an Establishment Survey in Portland, Oregon", Doctoral dissertation, Portland State University, 2013.

[38] Ding, L., & Zhang, N., "A travel mode choice model using individual grouping based on cluster analysis" Procedia engineering, 137, 786-795, 2016.

[39] TOMORROW'S, T. R. A. V. E. L. L. E. R. FUTURE TRAVELLER TRIBES 2030.

[40] World Health organisation, "Towards a Common Language for Functioning, Disability and Health ICF." Retrieved from: http://www.who.int/classifications/icf/icfbeginnersguide.pdf?ua=1) , 2002.

[41] Mizaras V., Manos T., Pachinis T., Batsis A., Beck P., Weisser J., Pretsch T., Petraki E., Spanoudakis N., IM@GINE IT Deliverable 1.1 'Use Cases and user/vehicle profile requirements', 2005

[42] Simões A., Gomes A., Bekiaris E., "ASK-IT (IST-511298) Deliverable 1.1.2 'Use cases'", 2006

[43] Allport, G., "Attitudes," in A Handbook of Social Psychology, ed. C. Murchison. Worcester, MA: Clark University Press, 789–844, 1935.

[44] Allsop, R. E., "Transport networks and their use: how real can modelling get?" Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 366(1872), 1879-1892, 2008.

[45] Chu, Z., Cheng, L., & Chen, H., "A Review of Activity-Based Travel Demand Modeling" In CICTP 2012: Multimodal Transportation Systems—Convenient, Safe, Cost-Effective, Efficient (pp. 48-59), 2012.

[46] Yang, L., "Modeling Preferences for Innovative Modes and Services: A Case Study in Lisbon" Master Thesis submitted to the Department of Civil and Environmental Engineering and Sloan School of Management, Massachusetts Institute of Technology, 2010.

[47] Princeton University "About WordNet." WordNet. Princeton University. 2010.

[48] Speer, R., Chin, J., & Havasi, C., "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." In *AAAI*(pp. 4444-4451), 2017.