




Article

My-Trac: System for Recommendation of Points of Interest on the Basis of Twitter Profiles

Alberto Rivas ^{1,2,*} , Alfonso González-Briones ^{1,2,3} , Juan J. Cea-Morán ¹, Arnau Prat-Pérez ⁴ and Juan M. Corchado ^{1,2} 

¹ BISITE Research Group, University of Salamanca, Edificio I+D+i, Calle Espejo 2, 37007 Salamanca, Spain; alfonso@usal.es (A.G.-B.); juanju_97@usal.es (J.J.C.-M.); corchado@usal.es (J.M.C.)

² Air Institute, IoT Digital Innovation Hub, Carbajosa de la Sagrada, 37188 Salamanca, Spain

³ Research Group on Agent-Based, Social and Interdisciplinary Applications (GRASIA), Complutense University of Madrid, 28040 Madrid, Spain

⁴ Sparsity-Technologies, 08034 Barcelona, Spain; arnau@sparsity-technologies.com

* Correspondence: rivis@usal.es

Abstract: New mapping and location applications focus on offering improved usability and services based on multi-modal door to door passenger experiences. This helps citizens develop greater confidence in and adherence to multi-modal transport services. These applications adapt to the needs of the user during their journey through the data, statistics and trends extracted from their previous uses of the application. The My-Trac application is dedicated to the research and development of these user-centered services to improve the multi-modal experience using various techniques. Among these techniques are preference extraction systems, which extract user information from social networks, such as Twitter. In this article, we present a system that allows to develop a profile of the preferences of each user, on the basis of the tweets published on their Twitter account. The system extracts the tweets from the profile and analyzes them using the proposed algorithms and returns the result in a document containing the categories and the degree of affinity that the user has with each category. In this way, the My-Trac application includes a recommender system where the user receives preference-based suggestions about activities or services on the route to be taken.

Keywords: users' profiling; data extraction; natural language processing; recommender system; mapping application



Citation: Rivas, A.; González-Briones, A.; Cea-Morán, J.J.; Prat-Pérez, A.; Corchado, J.M. My-Trac: System for Recommendation of Points of Interest on the Basis of Twitter Profiles.

Electronics **2021**, *10*, 1263. <https://doi.org/10.3390/electronics10111263>

Academic Editors: Dimitris Apostolou and Osvaldo Gervasi

Received: 13 April 2021

Accepted: 20 May 2021

Published: 25 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Humans are social beings; we always seek to be in contact with other people and to have as much information as possible about the world around us. The philosopher Aristotle (384–322 B.C.) in his phrase “Man is a social being by nature” states that human beings are born with the social characteristic and develop it throughout their lives, as they need others in order to survive. Socialization is a learning process; the ability to socialize means we are capable of relating with other members of the society with autonomy, self-realization and self-regulation. For example, the incorporation of rules associated with behavior, language, and culture improves our communication skills and the ability to establish relationships within a community.

In the search for improvement, communication, and relationships, human beings seek to get in contact with other people and to obtain as much information as possible about the environment in order to achieve the above objectives. The emergence of the Internet has made it possible to define new forms of communication between people. It has also made it possible to make a large amount of information on any subject available to the average user at any time. This is materialized in the development of social networks. The concept of social networking emerged in the 2000s as a place that allows for interconnection between people, and, very soon, the first social networking platforms appeared on the Internet that

served to bring people together. Among the first platforms that emerged were Fotolog, MySpace, Hi5, Buzz, and SecondLife; however, most of them have declined in popularity or disappeared. Today, Facebook, Twitter, and Instagram stand out; they are used by millions of people all over the world. Thanks to these technologies, people from different parts of the world can engage in conversation, post photos of their latest trip, or keep their followers updated by sharing their opinions or experiences.

With regard to writing opinions or experiences, Twitter is the social network par excellence. Twitter is based on the concept of microblogging, i.e., users can post messages about their opinions, preferences, experiences, etc., with a maximum of 280 characters. Twitter allows its users to follow other accounts that interest them, or to comment on events in real time using hashtags. All this translates into one word: information. The information that users provide on social networks can be used in a variety of ways, many of them negative. Exposure on the Internet means that anyone can access the users' data and use it for financial gain. However, it can also be used to make life easier for users who choose to do so, always bearing in mind that there must be express consent on their part. This is precisely the case of the work presented here.

Since the emergence of the first social networks much progress has been made towards the current state of maturity of the social network life cycle. As presented above, they are of vital importance to society, as they fulfill the innate communication function of human beings. In addition to their use as a means of communication, they have begun to be exploited for business purposes in order to profit from the enormous amount of information that is generated on a daily basis. This information, which is generated by the society, is of great value once analyzed and processed correctly.

The data generated by users on social networks allows for the development of commercial and advertising actions that are much more effective than with traditional formats. Advertisement platforms have developed the ability to segment advertising on the basis of the behavior of each user, to show products and services to those who are really interested in those products and services. This increases the effectiveness of advertisements.

Social network users publish practically everything that happens in their daily lives, their opinions, where they are, what they eat, what they would like to buy, where they are going on holiday, and a long list of behaviors that are transformed into valuable information for analysis. The information obtained through the analysis is very interesting as it allows for the elaboration of demographic, socio-economic, and consumer trend profiles. The companies that own these social networks sell high-value information to other companies to enable them to carry out much more powerful and effective marketing and advertising strategies. Another remarkable aspect of data analytics on social networks is the ability to perform real-time analysis of the information to offer products and services according to their characteristics.

Information is a very precious commodity, and, as presented above, Twitter is a great source of data when analyzing human behavior and interactions or when learning about the opinion of certain users on certain topics. This information can be used to improve the multi-modal experience of users when they use the My-Trac application. Therefore, an adaptation of these systems for adoption in mapping applications is proposed.

The European My-TRAC project focuses on providing user-centered services to improve the multi-modal experience of passengers from door-to-door. This helps citizens develop greater confidence in and adherence to multi-modal transport services. In addition, My-TRAC improves customization to users' needs through data, statistics, and trends provided by passengers' experiences when using the proposed platform. Part of the tailoring of services and recommendations to users is determined by the knowledge obtained from their Twitter posts through the use of NLP techniques to classify and understand users.

There are other services that offer similar functionalities to My-Trac, such as ROSE [1], CTRR, and CTRR+ based systems for city-based tourism [2], or to participate in solidarity projects in rural environments [3]. From among the above, ROSE (ROuting Service) stands out, which is a mobile phone application that suggests events and places to the user and

guides them via public transport. There are many different systems that incorporate both recommendation and navigation. However, there is no system that combines event recommendation and pedestrian navigation with (real-time) public transport. However, it does not employ multi-modal navigation between different public transport modes (bus, train, carpooling, plane, etc.) in different countries and that would use information from the user's social network profile. Instead, current systems utilize a set of information initially entered into the application which is not updated afterwards. Finally, Tables 1 and 2 present a review of similar works.

Table 1. Review of similar works: Part I.

Title / Publication	Functionality	Advantages	Shortcomings
ROSE (ROuting SErvice) [1]	Mobile phone application that suggests events and places to the user and guides them via public transport.	The current systems utilize a set of information initially entered into the application which is not updated afterwards.	There is no system that combines event recommendation and pedestrian navigation with (real-time) public transport. It does not employ multi-modal navigation between different public transport modes (bus, train, carpooling, plane, etc.) in different countries and that would use information from the user's social network profile.
Systems for city-based tourism [2]	A personalized travel route recommendation based on the road networks and users' travel preferences.	The experimental results show that the proposed methods achieve better results for travel route recommendations compared with the shortest distance path method.	It does not use information from public transport services in route recommendations.
Tourism routes as a tool for the economic development of rural areas—vibrant hope or impossible dream? [3]	This paper argues that the clustering of activities and attractions, and the development of rural tourism routes, stimulates co-operation and partnerships between local areas. The paper further discusses the development of rural tourism routes in South Africa and highlights the factors critical to its success.	The article analyzes the realization of routes that include activities and attractions in a way that encourages and enhances rural development in Africa.	Preliminary project that requires public cooperation (institutions, transport, services) for a comprehensive improvement of the proposal.

This article improves on the previous system for the extraction of information regarding Twitter users [4]. The system is capable of obtaining information about a particular user and of elaborating a profile with the user's preferences in a series of pre-established categories. A review of existing reputation systems is presented in Section 2. Section 3 describes the proposal. Section 4 presents the assessment made with synthetic data. Section 5 shows how the system is integrated in My-Trac app. Finally, Section 6 presents the conclusions.

Table 2. Review of similar works: Part II.

Title / Publication	Functionality	Advantages	Shortcomings
Social Recommendations for Events [5]	Outlife recommender assists in finding the ideal event by providing recommendations based on the user's personal preferences.	In addition to the user's preferences, the recommender uses information from the user's group of friends to make event recommendations more satisfactory.	Although it uses information from the user's groups of friends, no use is made of information from the user's social networks to complement the analysis and recommendation.
Smart Discovery of Cultural and Natural Tourist Routes [6]	This paper presents a system designed to utilize innovative spatial interconnection technologies for sites and events of environmental, cultural and tourist interests. The system discover and consolidate semantic information from multiple sources, providing the end-user the ability to organize and implement integrated and enhanced tours.	The system adapts the services offered to meet the needs of specific individuals, or groups of users who share similar characteristics, such as visual, acoustic, or motor disabilities. Personalization is done in a dynamic way that takes place at the time and place of the service.	The very comprehensive system that uses external services, scraping, crawling, geo-positioning but does not include information from social networks to complement the analysis and recommendation of events.
Enhancing cultural recommendations through social and linked open data [7]	Hybrid recommender system (RS) in the artistic and cultural heritage area, which takes into account the activities on social media performed by the target user and her friends	The system integrates collaborative filtering and community-based algorithms with semantic technologies to exploit linked open data sources in the recommendation process. Furthermore, the proposed recommender provides the active user with personalized and context-aware itineraries among cultural points of interest.	The main drawback is the absence of extensive control over the semantics that are not taken into account. It generates difficulties in justifying, explaining, and hence analyzing the resulting scores.
Personalized Tourist Route Generation [8]	Intelligent routing system able to generate and customize personalized tourist routes in real-time and taking into account public transportation.	We have modeled the tourist planning problem, integrating public transportation, as the Time Dependent Team Orienteering Problem with Time Windows (TDTOPTW). We have designed a heuristic able to solve it in real time, precalculating the average travel times between each pair of POIs in a preprocessing step.	Future works consists on extending the system to more cities with a different public transport network topology. The next one consists on integrating an advanced recommendation system in a wholly functional PET. The systema don't use social network capabilities, that allows to store, share and add travel experiences to better help tourists on the destination.

2. Natural Language Processing Techniques Applied to Twitter Profiles

In this section, we review the main techniques applied in the analysis that make it possible to get to know the users preferences through their tweets. This allows for recommendations to be made according to the user profile.

2.1. Word Embedding Techniques

NLP techniques allow computers to analyze human language, interpret it, and derive its meaning so that it can be used in practical ways. These techniques allow for tasks, such as automatic text summarization, language translation, relation extraction, sentiment

analysis, speech recognition, and item classification, to be carried out. Currently, NLP is considered to be one of the great challenges of artificial intelligence as it is one of the fields with the highest development activity since it presents tasks of great complexity: how to really understand the meaning of a text, how to intuit neologisms, ironies, jokes, or poetry? It is a challenge to apply the techniques and algorithms that allow us to obtain the expected results.

One of the most commonly used NLP techniques is Topic Modeling. This technique is a type of statistical modeling that is used to discover the abstract “topics” that appear in a series of input texts. Topic modeling is a very useful text mining tool for discovering hidden semantic structures in texts. Generally, the text of a document deals with a particular topic, and the words related to that topic are likely to appear more frequently in the document than those that are unrelated to the text. Topic Modeling collects the set of more frequent words in a mathematical framework, which allows one to examine a set of text documents and discover, on the basis of the statistics of the words in each one, what the topics may be and what the balance is between the topics in each document.

The input of topic modeling is a document-term matrix. The order of words does not matter. In a document-term matrix, each row is a question (or document), each column is a term (or word), we label “0” if that document does not contain that term, “1” if that document contains that term once, “2” if that document contains that term twice, and so on.

Algorithms, such as Bag-of-words or TF-IDF, among others, make it possible to represent the words used by the models and create the matrix defined above, representing a token in each column and counting the number of times that token appears in each sentence (represented in each row).

- **Bag-of-words.** This model allows to extract the characteristics of texts (also images, audios, etc.). It is, therefore, a feature extraction model. The model consists of two parts: a representation of all the words in the text and a vector representing the number of occurrences of each word throughout the text. That is why it is called Bag-of-words. This model completely ignores the structure of the text, it simply counts the number of times words appear in it. It has been implemented through the Genism library [9].
- **Term Frequency - Inverse Document Frequency (TF-IDF).** This is the product of two measures that indicate, numerically, the degree of relevance that a word has in a document within a collection of documents [10]. It is broken down into two parts:
 - *Term frequency:* Measures the frequency with which certain terms appear in a document. There are several measurement options, the simplest being the gross frequency, i.e., the number of times a term t appears in a document d . However, in order to avoid a predisposition towards long documents, the normalized frequency is used:

$$tf(t, d) = \frac{f(t, d)}{\max\{f(t, d) : t \in d\}}. \quad (1)$$

As shown in Equation (1), the frequency of the term is divided by the maximum frequency of the terms in the document.

- *Inverse document frequency:* If a term appears very frequently in all of the analyzed documents, its weight is reduced. If it appears infrequently, it is increased.

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}. \quad (2)$$

As shown in Equation (2), the total number of documents is divided by the number of documents containing the term. Term frequency—Inverse document frequency: The entire formula is as shown in Equation (3).

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D). \quad (3)$$

Word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning [11]. Word embeddings can be obtained using a set of language modeling and feature learning techniques where words or phrases from the vocabulary are mapped to vectors of real numbers [12].

- **Word2vec.** This technique uses huge amounts of text as input and is able to identify which words appear to be similar in various contexts [13–15]. Once trained on a sufficiently big dataset, 300-dimensional vectors are generated for each word, forming a new vocabulary where "similar" words are placed close to each other. Pre-trained vectors are used, achieving a wealth of information from which to understand the semantic meaning of the texts.
- **Doc2vec.** This technique is an extension of Word2Vec and is applied to a document as a whole instead of individual words, it uses an unsupervised learning approach to better understand documents as a whole [16]. Doc2Vec model, as opposed to Word2Vec model [17], is used to create a vectorized representation of a group of words taken collectively as a single unit. It does not only give the simple average of the words in the sentence.

2.2. Topic Modeling

As already presented in the previous section, topic modeling is a tool that takes an individual text (or corpus) as input and looks for patterns in word usage; it is an attempt to find semantic meaning in the vocabulary of that text (or corpus).

This set of tools enables the extraction of topics from texts; a topic is a list of words that is presented in a way that is statistically significant. Topic modeling programs do not know anything about the meaning of the words in a text. Instead, they assume that each text fragment is composed (by an author) through the selection of words from possible word baskets, where each basket corresponds to a topic. If that is true, then it is possible to mathematically decompose a text into the baskets from which the words that compose it are most likely to come. The tool repeats the process over and over again until the most probable distribution of words within the baskets, the so-called topics, is established.

The techniques executed by the proposed system are used to discover word usage patterns of each user on Twitter, and they make it possible to group users into different categories. To this end, a thorough review of the main tools for topic modeling has been carried out. Most of the algorithms are based on the paradigm of unsupervised learning. These algorithms return a set of topics, as many as indicated in the training. Each topic represents a cluster of terms that must be related to one of those categories. Precisely for this reason, a large number of tweets have been retrieved as training data. Keywords have been searched for for each category. As part of this research, a total of three algorithms have been evaluated: LDA, LSI, and NMF. In the NMF experiment, the best results were obtained, although the techniques applied in other works have been reviewed in order to contrast their results with this method.

Apart from the comparison itself, there are numerous studies that have made similar comparisons between these techniques so that the decision is supported by similar studies. In the work of Tunazzina Islam, in 2019 a similar experiment was carried out to the one proposed in this paper [18]. In this paper, Apache Kafka is employed to handle the big streaming data from Twitter. Tweets on yoga and veganism are extracted and processed in parallel with data mining by integrating Apache Kafka and Spark Streaming. Topic modeling is then used to obtain the semantic structure of the unstructured data (i.e. Tweets). They then perform a comparison of the three different algorithms LSA, NMF, and LDA, with NMF being the best performing model.

Another noteworthy work is that carried out by Chen et al. [19], in which an experiment is carried out to detect topics in small text fragments.

This is similar to the proposal made in this paper, since tweets can be considered small texts. In this work a comparison is made between the LDA and NMF methods, the latter being the one that provided the best results.

- **Latent Dirichlet allocation (LDA).** Is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar [20–22]. For example, if observations are collections of words in documents, each document is a mixture of a small number of topics and each word's presence is attributable to one of the document's topics. LDA is an example of a topic model and belongs to the machine learning toolbox and in wider sense to the artificial intelligence toolbox.
- **Nonnegative Matrix Factorization (NMF).** Is an unsupervised learning algorithm belonging to the field of linear algebra. NMF reduces the dimensionality of an input matrix by factoring it in two and approximating it to another of a smaller range. The formula is $V \approx WH$. Let us suppose, observing Equation (4), a vectorization of P documents with an associated dictionary of N terms (weight). That is, each document is represented as a vector of N dimensions. All documents, therefore, correspond to a V matrix

$$V \in \mathbb{R}^{N \times P} = \begin{pmatrix} \square & \cdots & \cdots & \square \\ \square & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \square & \cdots & \square & \square \end{pmatrix}, \quad (4)$$

where N is the number of rows in the matrix, and each of them represents a term, while P is the number of columns in the matrix and each of them represents a document. Equations (5) and (6) shows matrices W and H . The value r marks the number of topics to be extracted from the texts.

Matrix W contains the characteristic vectors that make up these topics. The number of characteristics (dimensionality) of these vectors is identical to that of the data in the input matrix V . Since only a few topic vectors are used to represent many data vectors, it is ensured that these topic vectors discover latent structures in the text.

The H -matrix indicates how to reconstruct an approximation of the V -matrix by means of a linear combination with the W -columns.

$$W \in \mathbb{R}^{N \times r} = \begin{pmatrix} \square & \cdots & \cdots & \square \\ \square & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \square & \cdots & \square & \square \end{pmatrix}, \quad (5)$$

where N is the number of rows in matrix W , and each of them represents a term (weight), and r is the number of columns in matrix W , where r is the number of characteristics to be extracted.

$$H \in \mathbb{R}^{r \times P} = \begin{pmatrix} \square & \cdots & \cdots & \square \\ \square & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \square & \cdots & \square & \square \end{pmatrix}, \quad (6)$$

where r is the number of rows in matrix H , r is the number of characteristics to be extracted, and P is the number of columns, with one column for each document. The result of the matrix product between W and H is, therefore, a matrix of dimensions $N \times P$ corresponding to a compressed version of V .

The use of Machine Learning techniques for the analysis of information extracted from Twitter is a very common case study today. It is convenient to study what kind of

research is being carried out on this subject. One of the main applications is the use of Twitter and Natural Language Processing techniques in order to extract a user's opinion about what is being tweeted at a given time. The article "A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle", written by Hao Wang et al. [23], presents a system for real-time polarity analysis of tweets related to candidates for the 2012 U.S. elections.

The system collects tweets in real time, tokens and cleans them, identifies which user is being talked about in the tweet, and analyzes the polarity. For training, it applies Naïve Bayes, a statistical classifier. It uses hand-categorized tweets as input. Another study similar to this one is the one proposed by J.M.Cotelo et al. from the University of Seville: "Tweet Categorization by combining content and structural knowledge" [24]. It proposes a method to extract the users' opinion about the two main Spanish parties in the 2013 elections. It uses two processing pipelines, one based on the structural analysis of the tweets, and the other based on the analysis of their content.

Another possible line of research is based on categorizing Twitter content. This is the case of the article "Twitter Trending Topic Classification" written by Kathy Lee et al. [25]. It studies the way to classify trending topics (hashtags highlighted) in 18 different categories. To this end, Topic Modeling techniques were used. The key point lies in providing a solution based on the analysis of the network underlying the hashtags and not only the text: "our main contribution lies in the use of the social network structure instead of using only textual information, which can often be noisy considering the social network context".

As it can be seen, there are many studies currently oriented to the analysis of Twitter using Machine Learning tools. The challenge to be faced in this work is to find the optimal way of classifying users according to their tweets. The sections that follow describe the objectives of the project and detail the research and testing that led to the construction of a stable system fit for the purpose for which it has been designed.

3. Proposal

This section proposes a system for the extraction of information about Twitter users. The system is capable of obtaining information about a particular user and of elaborating a profile with the user's preferences in a series of pre-established categories. From an abstract point of view, the proposal could be seen as a processing pipeline, as shown in Figure 1. The different phases of this pipeline contribute to the achievement of the main objective: user classification.

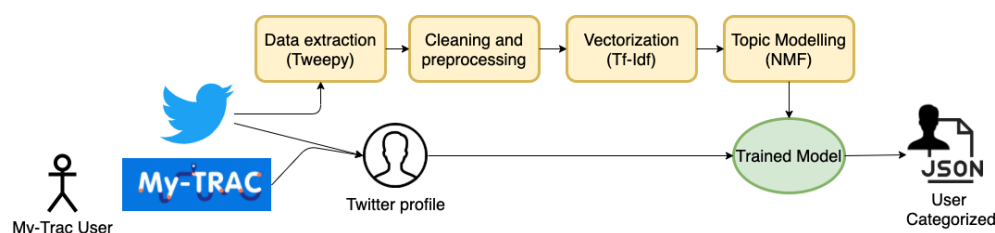


Figure 1. Pipeline representing the system processing steps.

3.1. Category Definition

Matching a given profile to a specific category or topic is one of the objectives of NLP algorithms. As a starting point, it is necessary to prepare the training dataset that is used when investigating the algorithmic model. The strategy followed is based on the model of the Interactive Advertising Bureau (IAB) association [26]. Today, IAB is a benchmark standard for the classification of digital content. In particular, the IAB Tech Lab has developed and released a content taxonomy on which the present categorization is based. This taxonomy proposes a total of 23 categories with their corresponding subcategories covering the main topics of interest. In this way, 8000 tweets from each of these categories have been ingested. As a result, 23 datasets with examples of tweets related to each category

were obtained, these datasets have been used to train the system at a later stage. Specifically, the list of topics is shown in Table 3.

Table 3. Categories taxonomy.

Topics
Arts & Entertainment
Automotive
Business
Careers
Education
Family & Parenting
Health & Fitness
Food & Drink
Hobbies & Interests
Home & Garden
Law, Gov't & Politics
News
Personal Finance
Society
Science
Pets
Sports
Style & Fashion
Technology & Computing
Travel
Real Estate
Shopping
Religion & Spirituality

3.2. Twitter Data Extraction

The Twitter data extraction mechanism is a fundamental element of the system. The goal of this mechanism is to recover two types of data.

On the one hand, the system extracts a set of anonymous tweets related to each of the defined preference categories; these tweets are used to train the data classification algorithms.

On the other hand, the mechanism extracts information about the given user for the analysis of their preferences.

Twitter's API enables developers to perform all kinds of operations on the social network. It is, therefore, necessary for our system to use this powerful API. This API could be used by elaborating a module that would make HTTP requests to the API so that the endpoints of interest are executed. However, this involves a remarkably high development cost.

Another option would be to make use of one of the multiple Python libraries that encapsulate this logic and offer a simple interface to developers. The latter option has been chosen for the development of this system, more specifically, library Tweepy [27].

3.3. Preprocessing of Tweets

Once the data has been extracted, it must be prepared for the classification algorithms. Cleaning and preprocessing techniques must be applied, so that the text is prepared for topic modeling algorithms. Libraries, such as NLTK and Spacy, have been used, as can be observed in Listing 1.

The first step involves cleaning tweets, by removing content that does not provide information for language processing. More specifically, this task consists in eliminating URLs, hashtags, mentions, punctuation marks, etc.

Another of the techniques applied to obtain more information from tweets is the transformation of the emojis contained in the text into a format from which it is possible to extract information. To do this, a dictionary of emojis is used as a starting point for the

conversion of the data. This dictionary contains a series of values that interpret each of the existing emojis when applying the corresponding analysis. In this way, it has been possible to identify and give a certain value to each emoji for its treatment.

The key activity performed during the preprocessing consist of eliminating stopwords and tokenization. Whether it is a paragraph, an entire document or a simple tweet, every text contains a set of empty words or stopwords. This set of words is characterized by its continuous repetition in the document and its low value within the analysis. These words are mainly articles, determiners, synonyms, conjunctions, and others.

Listing 1. Preprocessing step pseudocode.

```
from nltk.tokenize import word_tokenize
import~spacy

sp = spacy.load('en_core_web_sm')
stopwords_dict = sp.Defaults.stop_words

def tweet_preprocessing(tweet):
    tweet = hashtag_removal(tweet)
    tweet = mentions_removal(tweet)
    tweet = url_removal(tweet)
    tweet = html_removal(tweet)
    tweet = punctuation_removal(tweet)
    tweet = emojis_removal(tweet)
    tweet = word_tokenize(tweet)
    tweet = [word for word in tweet if not word in stopwords_dict]
    return tweet
```

Table 4 shows the results obtained after the tweets have gone through the preprocessing and preparation process which had been carried out using the tools listed above.

3.4. Vectorization

Vectorization is the application of models that convert texts into numerical vectors so that the algorithms can work with the data. Two algorithms have been considered for the performance of this task, “Bag-Of-Words” and “Tf-Idf”. Both are widely used in the field of NLP, but, in general, creation of tf-idf weights from text works properly and is not very expensive computationally. Moreover, NMF expects as input a Term-Document matrix, typically a “Tf-Idf” normalized.

The vectorizer have been tuned manually with some parameters according to the dataset, as can be observed in Listing 2. *Min_df* was set to 100 to ignore words that appear in less than 100 tweets. In the same way, *max_df* was set to 0.85 to ignore words that appear in more than 85% of the tweets. Thanks to that feature, it is possible to remove words that introduce noise in the model. Finally, the algorithm only takes into account single words, so, in order to include bigrams, the parameter *ngram_range* was set to (1, 2)

Table 4. Preprocessing results using NLTK tokenization.

	Text	Nltk_tokenized
0	Read This Before Taking a Road Trip with a Pet	[read, taking, road, trip, pet]
1	@kenwardskorner @Senators @Canucks In addition, does ...	[also, name, imply, take, acid, road, trips, I...]
2	Our Art is our Passion \n#apnatruckart #truck	[art, passion, apnatruckart, truck, art, uniqu...]
3	Lelang drop acc budget 40–50 k dong?	[lelang, drop, acc, budget, dong]
4	We agree...and want everyone to know that ou	[agree, want, everyone, know, tours, relaxed, ...]
5	Choosing a hotel for a break away with the fam...	[choosing, hotel, break, away, family, special...]
6	@_JassyJass Are you,camping?	[camping]
7	How to Pack Your Electronics for Air Travel ht	[pack, electronics, air, travel]
8	Dasar low budget! https://t.co/2YUmUrGjj5	[dasar, low, budget]

Listing 2. Vectorization step pseudocode.

```

from sklearn.feature_extraction.text import TfidfVectorizer

def tfidf_vectorization(tweets):
    vectorizer = TfidfVectorizer(
        min_df=100,
        max_df=0.85,
        ngram_range=(1, 2),
        preprocessor='_'.join,
        use_idf=True
    )
    vectorized_tweets = vectorizer.fit_transform(tweets)
    return vectorized_tweets

```

3.5. Topic Modeling

Topic Modeling is a typical NLP task that aims to discover abstract topics in texts. It is widely used to discover hidden semantic structures. In the present work, this technique has been used to discover the main topics of interest of the My-Trac application users based on their Twitter profiles, which should correspond to some of the previously defined categories.

Regarding the features of the model, in Section 3.4, the training tweets were vectorized to create a Term-Document matrix which has been the input of the NMF model. In addition, NMF needs one important parameter, the number of topics to be discovered “*n_components*”. In this case, *n_components* was set manually to 23, which is the number of topics that were defined initially in the categories taxonomy. Following this approach,

the algorithm is trained with 184,000 tweets (8000 per category) with the aim of obtaining as many topics as categories were defined in the taxonomy. Once the model has been trained, it has been possible to determine in which topics a user's profile fits on the basis of their tweets. The implementation of the topic modeling algorithm has been carried out on the basis of NMF using SKLearn library, as is deailed in Listing 3.

Listing 3. NMF Sklearn implementation.

```
from sklearn.decomposition import NMF

def train_model(vectorized_tweets):
    nmf_model = NMF(n_components=23, alpha=.1,
                    l1_ratio=.5, init='nndsvda')
    nmf_model.fit(vectorized_tweets)
    return nmf_model
```

Finally, it is worth mentioning the use of some extra parameters which were set in the implementation of the model. The method used to initialize the procedure was set to "NNDsva" which works better with the tweet dataset since this kind of data it is not sparse. *Alpha* and *l1_ratio* both are parameters which helps to define regularization.

4. Evaluation and Results

In order to evaluate the results of the algorithm, the most relevant terms have been identified for each resulting topic. Then, by reviewing the main terms for each topic, it is possible to determine if that words really represent the content of the topic. An example is shown in Figure 2, where the most relevant terms have been identified for 4 different topics, proving how well the algorithm identifies the terms associated with each one. As it can be seen, all of them are unambiguously related to their defined categories. Topic 1: Travel. Topic 2: Arts & Entertainment. Topic 9: Personal Finance. Topic 10: Pets.

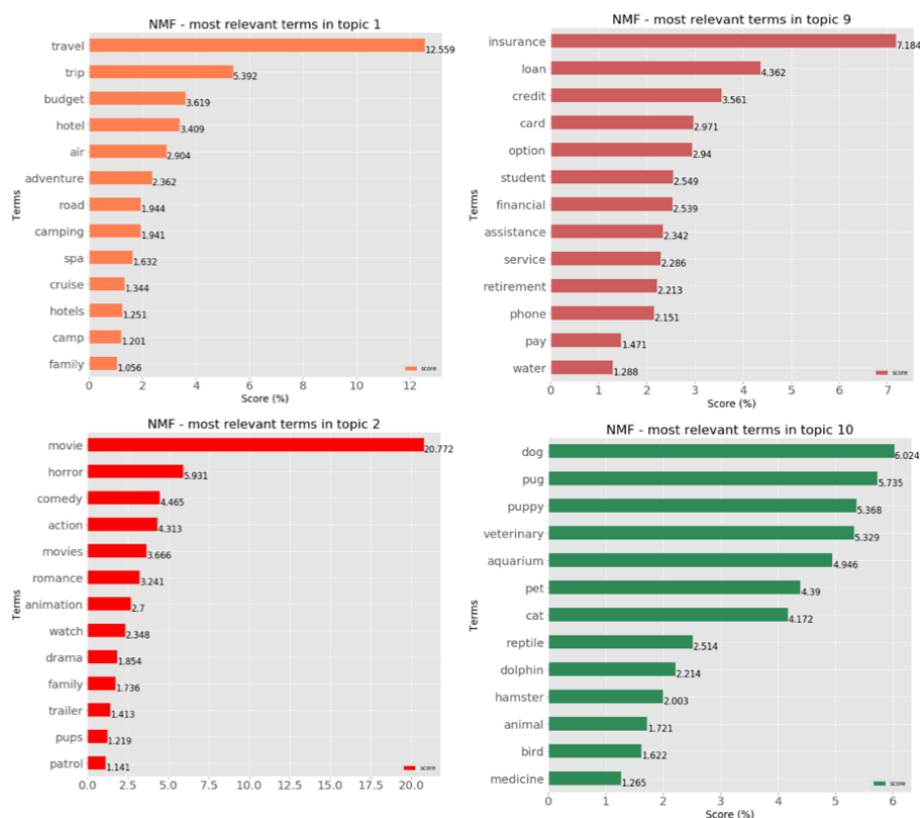


Figure 2. Example topics generated by NMF.

The full list of topics and their top 10 related keywords identified by the algorithm can be seen in Table 5. It should be noted that some of the previously defined categories in Table 3 have been removed during the evaluation of this model. This fact is due to the lack of tweets that would fit into those categories, as well as some topics were quite overlapped amongst them. The initially defined categories that have been removed during training process and evaluation are: “Home & Garden”, “Real State”, “Society”, and “News”. In the same way, the algorithm has been able to discover new categories related to the original ones, such as: “Movies”, “Videogames”, “Music”, “Events”, and “Medicine & Health”, leaving a total of 23 categories in the system.

Table 5. Topics obtained by the algorithm.

	Topic	Top 10 Words
1	Travel	travel, trip, budget, air, hotel, adventure, road, camp, family, day
2	Movies	movie, horror, action, comedy, movies, romance, watch, animation, family, drama
3	Videogames	game, xbox, pc, mmo, nintendo, videogame, esports, rpg, play, console
4	Careers	apprenticeship, internship, job, search, career, interview, vocational, training, remote, advice
5	Events	amusement, concert, cinema, restaurant, birthday, match, holiday, football, funeral, park
6	Health & Fitness	health, nutrition, therapy, physical, fitness, workout, exercise, wellness, medicine, weight
7	Religion & Spirituality	islam, christianity, hinduism, judaism, buddhism, spirituality, religion, astrology, atheism, sikhism
8	Shopping	grocery, lotto, shopping, gift, discount, sale, card, coupon, sales, code
9	Personal Finance	insurance, loan, credit, option, card, student, financial, service, assistance, phone
10	Pets	veterinary, dog, pug, puppy, aquarium, pet, cat, reptile, dolphin, hamster
11	Automotive	truck, car, auto, motorcycle, tesla, van, scooter, pickup, luxury, minivan
12	Science	chemistry, geography, geology, biology, physics, genetic, astronomy, environment, math, science
13	Law, Gov't & Politics	election, political, law, issue, vote, news, state, people, country, trump
14	Education	preschool, college, university, exam, electoral, education, homework, student, school, language
15	Food & Drink	coffee, vegetarian, beer, eat, vegan, cook, drink, wine, tea, dining
16	Family & Parenting	marriage, parent, single, baby, daycare, teen, date, life, toddler, adopt
17	Style & Fashion	wear, beauty, clothing, perfume, deodorant, wallet, casual, fashion, shave, trainer
18	Technology & Computing	software, app, developer, mongodb, database, email, android, internet, computer, ai
19	Hobbies & Interests	meme, draw, puzzle, collect, comic_strip, antique, guitar, art, woodwork, painting
20	Sports	martial, rugby, golf, sport, climb, pool, racing, cricket, skating, basketball
21	Business	industry, agriculture, construction, startup, recall, economy, business, automotive, butterfly, turkey
22	Medicine & Health	vaccine, menopause, pregnancy, health, mental, surgery, injury, disease, psychology, substance
23	Music	music, radio, rock, funk, pop, soul, classic, songwriter, listen, classical

Once the resulting model has been evaluated and verified, the next step is to check the effectiveness of the model with real Twitter profiles. The tests have been performed extracting 1200 tweets from different users and predicting for each user the most related topics based on their tweets. The final test results are shown in Table 6, where it can be observed how each profile name match with related topics according to the profile.

As an example, the main topics for the profile “Tesla” are “Automotive”, “Technology and computing”, and “Travel”.

Finally, in order to suggest the main topics of a specific user in the My-Trac app, for each user, the model returns the associated categories, along with the percentage of weight that each category has on the user. The lower the percentage, the less relation the user has with the category. The results of the final classification using some known Twitter accounts are given in Table 7. It should be noted that only the three main categories are shown in the table (together with their associated percentage), as they are the most accurate for categorizing the user.

Table 6. NMF evaluation with data from real Twitter profiles.

	Cat 1	Cat 2	Cat 2
Pontifex	Religion & Spirituality	Family & Parenting	Tech & Computing
Tesla	Automotive	Tech & Computing	Travel
BBCNews	Law, Gov’t & Politics	Sports	Family & Parenting
NintendoAmerica	Videogames	Hobbies & Interests	Sports
Theresa_may	Law, Gov’t & Politics	Business	Personal Finance
Oprah	Events	Family & Parenting	Sports
SkyFootball	Sports	Events	Hobbies & Interests
ScuderiaFerrari	Sports	Automotive	Law, Gov’t & Politics
IMDb	Events	Movies	Hobbies & Interests
ScienceMagazine	Science	Medicine & Health	Tech & Computing
Spotify	Music	Hobbies & Interests	Family & Parenting
Airbnb	Travel	Events	Careers

Table 7. Final results with different accounts.

	First Category	Second Category	Third Category
Tesla	Automotive (70.92%)	Tech & Computing (10.07%)	Travel (3.86%)
RealDonaldTrump	Law, Gov’t & Politics (28.02%)	Business (11.36%)	Sports (8.02%)
ScienceMagazine	Science (24.41%)	Medicine & Health (16.73%)	Tech & Computing (12.03%)
NintendoAmerica	Videogames (41.65%)	Hobbies & Interests (12.19%)	Sports (9.74%)

5. Final System Integration in My-Trac Application

Having passed the entire research and evaluation process, a trained algorithm has been obtained capable of classifying different Twitter accounts according to defined and discovered categories. In addition, a reliable data extraction method has been developed. Therefore, the next step consists of applying the algorithm to the My-Trac app to create a system that allows recommendations to My-Trac users based on their Twitter profiles, which is the objective of the present work.

The final system for My-Trac app consists of a mobile app where the user logs in, as it can be seen in Figure 3, and is asked to grant access to their Twitter data. Once the user signs in to the application, My-Trac seeks for the optimal means of transport to reach a specific destination given by the user and suggests the best conveyance for the trip, as Figures 4 and 5 show.

Finally, when the user chooses the route and mean of transport that best fits his trip, based on the present work, My-Trac app recommends some activities and points of interest for the user during the way based on its Twitter information, as can be observed in Figures 6 and 7.

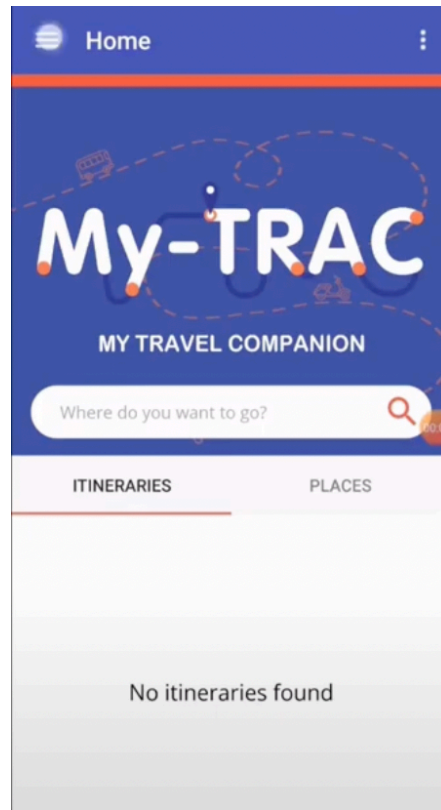


Figure 3. My-Trac application.

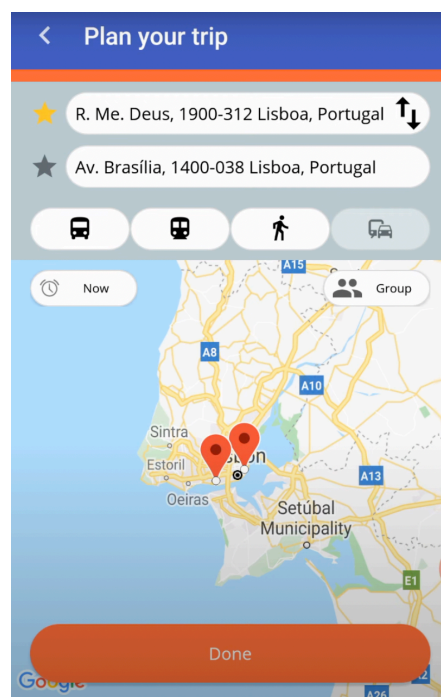


Figure 4. Trip planification using My-Trac app.

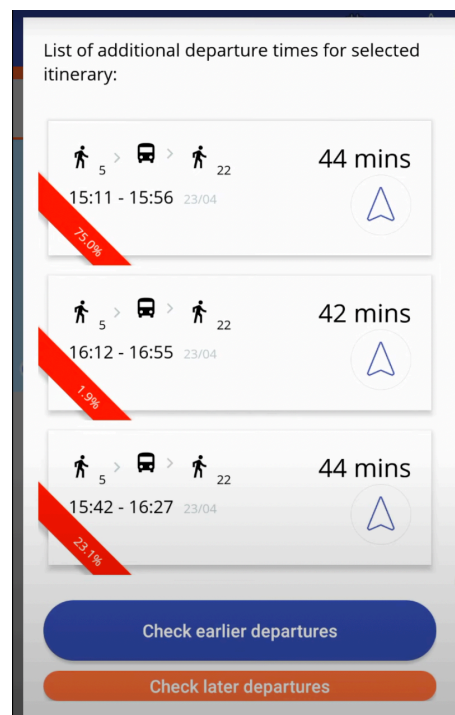


Figure 5. My-Trac suggests optimal means of transport for the destination.

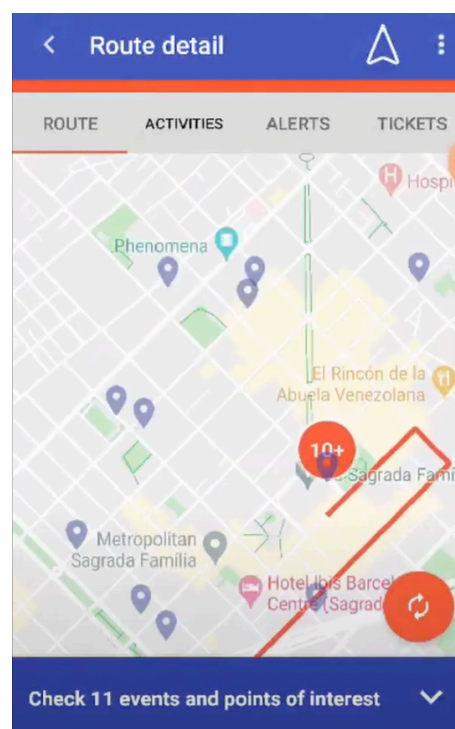


Figure 6. My-Trac recommends activities and point of interest for the user using its Twitter information.

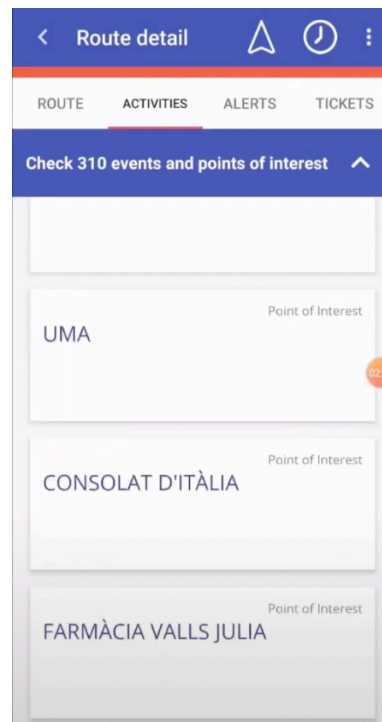


Figure 7. My-Trac suggestions.

Moreover, it is possible to get some detailed information for each activity recommended, as Figure 8 shows. In this way, thanks to My-Trac app, the user can improve his experience not only by receiving suggestions for the best conveyance for the trip but also receiving customized activity recommendations and points of interest.

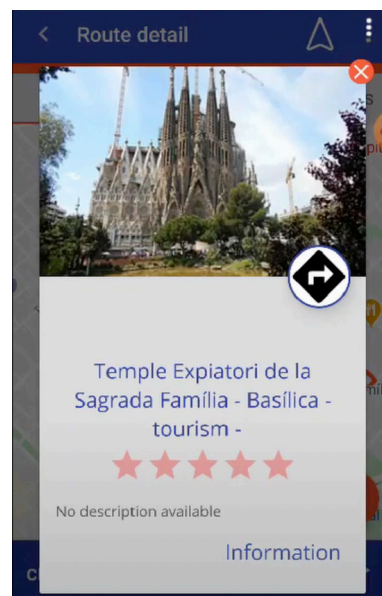


Figure 8. Detailed information about a suggested point of interest.

6. Conclusions and Future Work

This article presents a novel approach to extracting preferences from a Twitter profile by analyzing the tweets published by the user for use in mapping applications. This approach has successfully defined a consistent and representative list of categories, and the mechanisms needed for information extraction have been developed, both for model training and end-user analysis. It is a unique system, with which it has been possible to

develop an important feature in the My-Trac app, whereby it is possible to recommend relevant point of interest to the end users.

Regarding future work on this system, many areas of improvement and development have been identified. Tweets are not the only source of information that allows to discern the interests of a profile. It may be the case that a user only writes about football but is the follows many news-related and political accounts. The current system would only be able to extract the sports category. Therefore, one of the improvements would be the implementation of a model that would analyze followed users. This has been started, by extracting the followers and creating wordclouds with the most relevant ones. Similarly, hashtags also provide additional information suitable for analysis. Another line of research is the training of a model that allows to analyze the tweets individually. This would open the doors to performing a polarity analysis that would allow us to know if a user who writes about a certain category does it in a positive, negative, or neutral way.

As for the limitations of the system, it is possible that, in some regions, there may be restrictive regulations on the use of information published on social networks for this type of analysis. Therefore, the user should carry out a study of data protection and the legal framework adapted to each region where the service is to be provided. Furthermore, in terms of performance, it is possible that specific context-dependent systems training an algorithm for each individual user may perform slightly better than the proposed solution.

Author Contributions: Conceptualization, A.R. and A.G.-B.; methodology, A.R. and A.G.-B.; software, A.R. and J.J.C.-M.; validation, A.R., A.G.-B., J.J.C.-M. and A.P.-P.; formal analysis, A.R. and J.J.C.-M.; investigation, A.R., A.G.-B. and A.P.-P.; resources, A.R., A.G.-B., J.J.C.-M.; data curation, J.J.C.-M.; writing—original draft preparation, A.R., A.G.-B., J.J.C.-M.; writing—review and editing, A.R., A.G.-B., J.J.C.-M., A.P.-P. and J.M.C.; visualization, A.R. and J.J.C.-M.; supervision, A.G.-B., A.P.-P. and J.M.C.; project administration, A.G.-B., A.P.-P. and J.M.C.; funding acquisition, J.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Spanish Ministerio de Ciencia e Innovación under grant number TIN2017-89314-P.

Acknowledgments: This research has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 777,640.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ludwig, B.; Zenker, B.; Schrader, J. Recommendation of Personalized Routes with Public Transport Connections. In Proceedings of the International Conference on Intelligent Interactive Assistance and Mobile Multimedia Computing, Rostock-Warnemünde, Germany, 9–11 November 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 97–107.
2. Cui, G.; Luo, J.; Wang, X. Personalized travel route recommendation using collaborative filtering based on GPS trajectories. *Int. J. Digit. Earth* **2018**, *11*, 284–307. [CrossRef]
3. Briedenhann, J.; Wickens, E. Tourism routes as a tool for the economic development of rural areas—Vibrant hope or impossible dream? *Tour. Manag.* **2004**, *25*, 71–79. [CrossRef]
4. Cea-Morán, J.J.; González-Briones, A.; De La Prieta, F.; Prat-Pérez, A.; Prieto, J. Extraction of Travellers' Preferences Using Their Tweets. In Proceedings of the International Symposium on Ambient Intelligence, L Aquila, Italy, 17–19 June 2020; pp. 224–235.
5. De Pessemier, T.; Minnaert, J.; Vanhecke, K.; Dooms, S.; Martens, L. Social recommendations for events. In Proceedings of the CEUR workshop Proceedings, Miami, FL, USA, 1 October 2013; Volume 1066.
6. Stathopoulos, E.A.; Paliokas, I.; Meditskos, G.; Diplaris, S.; Tsafaras, S.; Valkouma, E.; Pehlivanides, G.; Riggas, C.; Vrochidis, S.; Votis, K.; et al. Smart discovery of cultural and natural tourist routes. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence-Companion, Thessaloniki, Greece, 14–17 October 2019; pp. 208–214.
7. Sansonetti, G.; Gasparetti, F.; Micarelli, A.; Cena, F.; Gena, C. Enhancing cultural recommendations through social and linked open data. *User Model. User-Adapt. Interact.* **2019**, *29*, 121–159. [CrossRef]
8. Garcia, A.; Arbelaitz, O.; Linaza, M.T.; Vansteenwegen, P.; Souffriau, W. Personalized tourist route generation. In Proceedings of the International Conference on Web Engineering, Vienna, Austria, 5–9 July 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 486–497.
9. University, Y. About Yale: Yale Facts. 2017. Available online: <https://www.yale.edu/about-yale/yale-facts> (accessed on 24 May 2021).

10. Demestichas, K.; Kosmides, P. An offline, statistical method for cost efficient design of experiments and field trials involving electric vehicles. In Proceedings of the 11th ITS European Congress, Glasgow, Scotland, 6–9 June 2016.
11. Ferrari, A.; Donati, B.; Gnesi, S. Detecting domain-specific ambiguities: An NLP approach based on Wikipedia crawling and word embeddings. In Proceedings of the 017 IEEE 25th International Requirements Engineering Conference Workshops (REW), Lisbon, Portugal, 4–8 September 2017; pp. 393–399.
12. Wallace, E.; Wang, Y.; Li, S.; Singh, S.; Gardner, M. Do nlp models know numbers? Probing numeracy in embeddings. *arXiv* **2019**, arXiv:1909.07940.
13. Church, K.W. Word2Vec. *Nat. Lang. Eng.* **2017**, *23*, 155–162. [[CrossRef](#)]
14. Rong, X. word2vec parameter learning explained. *arXiv* **2014**, arXiv:1411.2738.
15. Lilleberg, J.; Zhu, Y.; Zhang, Y. Support vector machines and word2vec for text classification with semantic features. In Proceedings of the 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), Beijing, China, 6–8 July 2015; pp. 136–140.
16. Bilgin, M.; Şentürk, İ.F. Sentiment analysis on Twitter data with semi-supervised Doc2Vec. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), London, UK, 5–7 July 2017; pp. 661–666.
17. Chen, Q.; Sokolova, M. Word2Vec and Doc2Vec in unsupervised sentiment analysis of clinical discharge summaries. *arXiv* **2018**, arXiv:1805.00352.
18. Islam, T. Yoga-veganism: Correlation mining of twitter health data. *arXiv* **2019**, arXiv:1906.07668.
19. Chen, Y.; Zhang, H.; Liu, R.; Ye, Z.; Lin, J. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowl.-Based Syst.* **2019**, *163*, 1–13. [[CrossRef](#)]
20. Resnik, P.; Armstrong, W.; Claudino, L.; Nguyen, T.; Nguyen, V.A.; Boyd-Graber, J. Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, CO, USA, 5 June 2015; pp. 99–107.
21. Mehrotra, R.; Sanner, S.; Buntine, W.; Xie, L. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland, 28 July–1 August 2013; pp. 889–892.
22. Tajbakhsh, M.S.; Bagherzadeh, J. Semantic knowledge LDA with topic vector for recommending hashtags: Twitter use case. *Intell. Data Anal.* **2019**, *23*, 609–622. [[CrossRef](#)]
23. Wang, H.; Can, D.; Kazemzadeh, A.; Bar, F.; Narayanan, S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, Jeju Island, Korea, 8–14 July 2012; pp. 115–120.
24. Coteló, J.M.; Cruz, F.L.; Enríquez, F.; Troyano, J. Tweet categorization by combining content and structural knowledge. *Inf. Fusion* **2016**, *31*, 54–64. [[CrossRef](#)]
25. Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M.M.A.; Agrawal, A.; Choudhary, A. Twitter trending topic classification. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 11 December 2011; pp. 251–258.
26. IAB Categories | MoPub Publisher UI | MoPub Developers. Available online: <https://developers.mopub.com/publishers/ui/iab-category-blocking/> (accessed on 4 May 2021).
27. Tweepy. Available online: <https://www.tweepy.org/> (accessed on 4 May 2021).