Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 2499282, 12 pages
https://doi.org/10.1155/2022/2499282

Research Article

Sentiment Analysis Models with Bayesian Approach: A Bike Preference Application in Metropolitan Cities

Antonino Vitetta

Dipartimento di Ingegneria dell'Informazione, delle Infrastrutture e dell'Energia Sostenibile, Università degli Studi Mediterranea di Reggio Calabria, Feo di Vito, Reggio Calabria 89122, Italy

Correspondence should be addressed to Antonino Vitetta; vitetta@unirc.it

Received 25 November 2021; Revised 19 January 2022; Accepted 3 February 2022; Published 17 March 2022

Academic Editor: Luigi Dell'Olio

Social media data is an important source of information that can also be used for the study of the passenger mobility sector. In transport systems, user choice is studied through demand models that define how user behavior is affected by the performance of the supply system. Demand models are typically calibrated through data observed in the transport system. The observed data includes the choices actually made by users. This paper investigates how sentiment analysis of data available in social media can be adopted to specify, calibrate, and validate demand models in certain choice levels. In this work a model based on the Bayesian approach is specified, calibrated, and validated in the case of bike preference in some Italian metropolitan cities. The model takes into account the discrete choice approach. Specification, calibration, and validation made it possible to identify the relevant variables that influence sentiments and obtain the posterior distribution probability of the parameters. The prior and the posterior conditional probabilities are compared, and some indications are obtained on the elasticity and weight of the sentiments that influence the choice.

1. Introduction

This paper studies the problem of using sentiment analysis of data available on social media to evaluate the sentiments that influence user choice in the transport system.

Different types of models can be adopted for the modelling behavior of discrete users' choices in transport systems.

The most used ones are derived from the theory of random utility with the theoretical foundations reported in Domencich and McFadden [1], Mansky [2] and Williams [3]. In the context of random utility theory, discrete choice models are adopted to estimate the probability of choosing alternatives as a function of the user's utilities specification and of user's utility probability distributions. It is assumed that the expected value of the utilities depends on quantities that characterize the alternatives and the users (attributes) and on the weights of the attributes (parameters).

In the context of random utility models, different specifications are proposed for the depending on the assumptions made for the utility probability distributions. Among the proposed models are the Logit [4], the Nested Logit [5], the Cross Nested Logit [6], and the Probit [7]. The Logit assumes independent distribution of the utilities; Nested Logit assumes nested dependence for subset of utilities; Cross Nested Logit assumes cross dependences among the utilities; Probit assumes general dependences among utilities. The Logit family assumes Gumbel probability distribution for the utilities, while the Probit family assumes multivariate Gaussian probability distribution for the utilities. Other specifications are proposed for the probability distribution (i.e., multivariate Gamma).

On the other hand, the models can be considered in the broader Bayesian approach [8]. The Bayesian approach assumes a posterior probability distribution obtained given an a priori probability, a likelihood, and an evidence.

In the transport choice models, the Bayesian approach assumes that the parameters have an initial (a priori) probability distribution that is updated by a set of observations. From the observation, a likelihood function is

evaluated, and, together with the prior distribution probability, the posterior probability distribution is obtained. From the Bayes approach, there is the advantage that the parameters are not deterministic; the hypothesized a priori distribution is updated by observation, and the initial belief can be modified by the data.

The models need to be specified, and the utility parameters need to be calibrated and validated. The trial-and-error process of specification-calibration-validation allows you to use models in a real-world context for planning and evaluating policies.

Transport models are calibrated by adopting surveys that observe the choices of users, adopting attributes of the transport system and socioeconomic attributes. The user's choice can be observed with automatic measurement (e.g., GPS) or with explicit user declaration. Attributes can be defined by adopting observed values or modelled values.

Social media is a valuable source of information with continuously available and updated data also for mobility studies [9, 10]. It can be adopted for model specification and for calibration of transport choice model parameters. The language used in social media is very often informal, and a sentiment could be derived from the text. Many software and papers have studied sentiment analysis and the models and procedures to be adopted for the extraction of sentiment. Considering the objective of this paper, the main steps taken for the analysis of sentiments can be divided into natural language processing, text analysis, computational linguistics, and sentiment evaluation. The estimation of sentiment from social media is very important, but it is beyond the scope of this work, and this problem is not considered in this paper (a comparison between sentiment analysis methods is reported in [11]). The aim of this paper is to investigate how sentiment can be adopted for the study of the influence on the transport choice.

Sentiment analysis is applied in the transport system: the assessment of sustainability in the economic and environmental components is proposed in Serna et al. [12]; sentiments for the transport system (accident, level of service, services, flow, etc.) are proposed in Ali et al. [13, 14] and Candelieri and Archetti [15]. To the author's best knowledge, sentiment analysis in the transport area is applied for the statistical evaluation of supply and demand but is not extended for the estimation of demand parameters.

In this paper, the study of the influence of sentiment on the preference of transport alternatives is studied. This work considers the specification, calibration, and validation of models based on the Bayes approach and useful for estimating sentiments about the preference for bike. It can be extended to other mode preferences or to other levels of choice, even time-dependent ones, since sentiments are on the web and change over time based on the perception of users of the transport performances.

The main innovation of this work concerns (i) the proposal of a method for the calibration of the parameters of the transport choice model for the mode preference, based on the sentiment analysis starting from the data available on social media; (ii) the application of the proposed method in a real case for the preference of the bike. The proposed method can be applied to assess preference for other modalities or other levels of choice.

In sections 2, 3, and 4, (i) the models are reported: section 2 specifies the discrete choice models that can also be applied in the context of sentiment analysis; section 3 specifies the Bayes approach model; section 4 describes the adoption of the model for policy evaluation. In section 5, (ii) two case studies are considered: the first (subsection 5.1) considers two users, in order to explain the application of the models; the second (Subsection. 5.2) applies the method in a real case. Subsections 5.1 and 5.2 are divided by adopting the title of sections 2, 3, and 4. The conclusions are reported in section 6.

2. Discrete Choice Models

2.1. Definition and Notation. The following notations and definitions are adopted:

(i) n the user (or a homogenous set of users);

(ii) k the generic alternative available for the users;

(iii) $I = \{\ldots, k, \ldots.\}$ the set of user's alternatives, equal for all users;

(iv) $\mathbf{U}_n = [\ldots, U_{k,n}, \ldots]'$ the vector of perceived utility for the user n, with $U_{\mathrm{k,n}}$ the perceived utility for the alternative k and the user n;

(v) $v_{k,n} = E(U_{k,n}) = \varphi(\mathbf{b}, \mathbf{y}_n)$ the expected value of $U_{\mathrm{k,n}}$, function of unknown vector of weights $\mathbf{b}$ and vector of user's characteristics $\mathbf{y}_{\mathrm{n}}$;

(vi) $\mathbf{y}_n = [\ldots, y_{j,n}, \ldots]'$ the vector of user's characteristics;

(vii) $\mathbf{b} = [\ldots, b_j, \ldots]'$ the vector of unknown weights (to be estimated) of the user's characteristics for the evaluation of $v_{k,n}$;

(viii) $\mathrm{p}_n(k|\mathbf{b}) = \rho(\mathbf{U}_n|\mathbf{b})$ the conditional choice probability that the user n choice the alternative k, over I and $\mathbf{b}$, evaluated in relation to the utilities' specifications and parameters values; note that the set I is not indicated considering that is considered fixed and equal for all users in this paper;

(ix) $\mathbf{x} = [\ldots, x_i, \ldots, x_m]'$ the vector of observed sentiment over I;

(x) $P()$ a probability function.

In this paper, the symbol n is used in an equivalent way for a single user or for a set of users with characteristics and homogenous choice behavior in the case studied. Therefore, n is associated with a user or even with a set of homogenous users. Note that a homogenous group of users has a single probabilistic model of choice, but each one can make a different real choice, considering the probabilistic nature of the problem.

2.2. Sentiments. On the web and on social networks, a natural textual language is adopted. From the analysis of the text, we can derive the polarity of sentiment by applying many approaches developed in the literature [12]; Ali et al.

[12–15]. The methods adopted to extract the polarity of sentiment from a natural language are a relevant problem, and much research is developed in this area considering the complexity of the problem. The study of this method is beyond the scope of this work. This paper focuses on the application of behavioral models [4, 16] starting from the polarity of sentiment extracted from the web [12–15].

The sentiment can be classified with different levels of positive or negative polarity. In this paper, just for simplicity sake, 3 levels are adopted: positive (P), negative (N), and neutral (E).

For a user n it can be assumed that the set of possible alternatives is

$$I = \{P, N, E\}. \tag{1}$$

The problem can easily be extended by considering other levels of polarity. The considerations below are valid considering a different number of sentiments reducing or expanding the number of alternatives in the choice set I. It is sufficient to delete or to insert alternatives into set I. Modifying the size of set I does not change all model properties in terms of model specification and parameter calibration.

Over a defined period of time, each user n observes the number of positive (P), negative (N), and neutral (E) sentiments, respectively, defined as

$$\mathbf{x}_n = \left[x_{P,n}, x_{N,n}, x_{E,n}\right]'. \tag{2}$$

For all users, the vector $\mathbf{x}$ of sentiments is observed:

$$\begin{aligned}\mathbf{x} &= \left[\mathbf{x}_1', \ldots, \mathbf{x}_n', \ldots\right]' \\ &= \left[x_{P,1}, x_{N,1}, x_{E,1}, \ldots, x_{P,n}, x_{N,n}, x_{E,n}, \ldots\right]'.\end{aligned} \tag{3}$$

The vector $\mathbf{x}$ has $m = 3 \cdot n$ entries, and for general purpose, it can be reported as

$$\mathbf{x} = \left[\ldots, x_i, \ldots, x_m\right]'. \tag{4}$$

2.3. Choice Model. The sentiment observed by the analyst cannot be considered as a deterministic variable. The polarity of the sentiment is an estimate of true polarity.

In the method applied, various errors must be considered (for users and analysts), such as the following:

(i) (for users) incomplete information on the subject, human errors in writing, etc;

(ii) (for users and analysts) understanding of the text, ambiguity of the text, etc;

(iii) (for the analyst) reliability of the algorithm, model specification, etc.

For these, it is necessary to adopt a probabilistic model to simulate the user's choice in relation to the observed sentiments.

For a user n, choice is commonly modelled with the following decision levels [2]:

(i) Generation of perceived alternatives and set of choices (I): in this work, it is assumed that set I is fixed and equal for all users, and for this reason, I is not reported for simplicity in the probability function;

(ii) choice of the alternative given the set of perceived choices ($k|I$).

User n defines for each alternative $k \in I$ a utility $U_{\mathrm{k,n}}$ which cannot be assumed deterministic due to user and analyst errors. Sentiment is not observed directly from the analysis, and some variables are observed for sentiment estimation. For this reason, a probability can be assessed that the alternative will be chosen.

Considering that the utility $U_{\mathrm{k,n}}$, for each alternative k and user n, is defined with a random variable, the conditional choice probability that user n chooses the alternative k, on I and $\mathbf{b}$, is evaluated by the Random Utility Theory [4] with discrete choice model:

$$\begin{aligned}p_n(k|\mathbf{b}) &= \mathrm{prob}\left(U_{k,n} > U_{h,n}; \forall k, h \in I, k \neq h; I = \{P, N, E\}\right) \\ &= \rho\left(\mathbf{U}_n|\mathbf{b}\right)\forall k \in I,\end{aligned} \tag{5}$$

where prob stands for probability.

The probability that an alternative is chosen is equivalent to the probability that the utility associated with the alternative chosen is the highest compared to the utilities associated with the other alternatives.

Note that the correct specification of the probability is $p_n(k|\mathbf{b}, I)$, but I is not reported in all probability functions for simplicity considering that it is assumed to be equal for all users and fixed.

It is assumed that, for a user n, the expected value of the utility is a function of coefficients $\mathbf{b}$ and the characteristics of the users $\mathbf{y}_{\mathrm{n}}$:

$$v_{k,n} = \varphi\left(\mathbf{b}, \mathbf{y}_n\right)\forall k \in I. \tag{6}$$

Assuming an independent and identical probability distribution for each $U_{\mathrm{k,n}}$, with a Gumbel of expected value $v_{k,n}$ and parameter θ (the variance of the distribution for each alternative is equal to $\pi^2 \cdot \theta^2/6$), the choice probability is obtained with the Logit model:

$$p_n(k|\mathbf{b}) = \frac{\exp\left(v_{k,n}/\theta\right)}{\sum_{h \in I}\exp\left(v_{h,n}/\theta\right)}. \tag{7}$$

For the Logit model, very often, it is assumed that for a user n the ratio between the expected value of the utility $v_{k,n}$ and the parameter θ is a linear combination of the coefficients $\mathbf{b}$ and the characteristics of the users $\mathbf{y}_{\mathrm{n}}$:

$$\frac{v_{k,n}}{\theta} = \frac{\varphi\left(\mathbf{b}, \mathbf{y}_n\right)}{\theta} = b' \cdot \mathbf{y}_{k,n} = \sum_{j \in k} b_j \cdot y_{j,n}\forall k \in I. \tag{8}$$

Other assumptions can be made for the probability distribution of utilities. Examples of choice models adopted in the literature are Nested Logit, with nested Gumbel distributions; Probit, with multivariate Gaussian

distribution; Gammit, with multivariate Gamma distribution; and so on.

Probability can also be evaluated with a mixed Logit by considering the probability distribution for the parameters.

3. Bayes Approach

With the Bayes approach, the prior (initial) probability distribution of **b** is improved from a set of sentiments **x**.

Bayes' approach considers that

$$P(\mathbf{b}|\mathbf{x}) = \frac{P(\mathbf{b}) \cdot P(\mathbf{x}|\mathbf{b})}{P(\mathbf{x})}, \tag{9}$$

with

(i) $P(\mathbf{b}|\mathbf{x})$ the *posterior* conditional joint probability **b** over **x** and I;

(ii) $P(\mathbf{b})$ the *prior* conditional joint probability **b** over I;

(iii) $P(\mathbf{x}|\mathbf{b})$ the conditional joint probability **x** over **b** and I called also *likelihood* as explicated in Subsection 3.2;

(iv) $P(\mathbf{x})$ the *evidence* conditional joint probability **x** over I.

3.1. Prior. The prior probability $P(\mathbf{b})$ could be assumed from a known distribution defined in similar context or in the same context in earlier time. Assume that the conditional independence between the estimated parameters b_i, $P(\mathbf{b})$ is

$$P(\mathbf{b}) = \Pi_j P\left(b_j\right), \tag{10}$$

with $P(b_j)$ the prior probability b_j over I for the parameter j.

3.2. Likelihood. The conditional joint probability $P(\mathbf{x}|\mathbf{b})$ is evaluated in relation to the type of observed variables **x**. This conditional probability is also called likelihood as it evaluates the probability (how probably given a parameter or how likely) of observing **x**, given a predefined value of the **b** parameter of the distribution. Very often in the evaluation of the transport analysis, a sample of independent users is considered, and for each user n the chosen alternative $j(n)$ is observed. In this context, $P(\mathbf{x}|\mathbf{b})$ is the joint probability of observing the alternative actually chosen by independent users ($P(\mathbf{x}|\mathbf{b}) = \Pi_n p_n(j(n)|\mathbf{b})$).

In the sentiment approach adopted in this paper, for each independent user n, the sentiments of frequency $\mathbf{x}_n$ (2) are observed (independence between users and between observations, naïve condition). For all users, the vector **x** (3) is observed. Assuming that the observed events are independent, $P(\mathbf{x}_n|\mathbf{b})$ is the joint probability of observing $\mathbf{x}_n$ and it is given by a multinomial distribution:

$$P(\mathbf{x}_n|\mathbf{b}) = \left(\frac{\left(x_{P,n} + x_{N,n} + x_{E,n}\right)!}{\left(x_{P,n} \cdot x_{N,n} \cdot x_{E,n}\right)!}\right) \cdot \left(p_n(P|\mathbf{b})^{x_{P,n}}\right) \cdot \left(p_n(N|\mathbf{b})^{x_{N,n}}\right) \cdot \left(p_n(E|\mathbf{b})^{x_{E,n}}\right). \tag{11}$$

Considering the problem of **b** estimation, the term $A_n = (x_{P,n} + x_{N,n} + x_{E,n})!/(x_{P,n} \cdot x_{N,n} \cdot x_{E,n})!$ is a constant, because it depends on the user n but it does not depend on **b**. The likelihood for the user n can be written as

$$P(\mathbf{x}_n|\mathbf{b}) = A_n \cdot p_n(P|\mathbf{b})^{x_{P,n}} \cdot p_n(N|\mathbf{b})^{x_{N,n}} \cdot p_n(E|\mathbf{b})^{x_{E,n}}. \tag{12}$$

Considering independent users, the likelihood is

$$P(\mathbf{x}|\mathbf{b}) = \Pi_n A_n \cdot p_n(P|\mathbf{b})^{x_{P,n}} \cdot p_n(N|\mathbf{b})^{x_{N,n}} \cdot p_n(E|\mathbf{b})^{x_{E,n}} = A \cdot \Pi_{i=1..m} p_n(k(i)|\mathbf{b})^{x_i}, \tag{13}$$

with

(i) $A = \Pi_n A_n$

(ii) $k(i)$ the alternative (sentiment) belonging to I associated with the observed variable x_i.

3.3. Evidence. Considering the problem of **b** estimation, the evidence $P(\mathbf{x}|I)$ is a constant B because it does not depend on **b**:

$$P(\mathbf{x}) = B. \tag{14}$$

From the probability axiom $\int_{\boldsymbol{\beta}} (P(\boldsymbol{\beta}) \cdot P(\mathbf{x}|\boldsymbol{\beta})/P(\mathbf{x}))d\boldsymbol{\beta} = 1$ it follows that

$$B = P(\mathbf{x}) = \int_{\boldsymbol{\beta}} P(\boldsymbol{\beta}) \cdot P(\mathbf{x}|\boldsymbol{\beta}) \cdot d\boldsymbol{\beta}. \tag{15}$$

3.4. Posterior. Assuming the conditional independence between the estimated parameters, the independence between the users and between the observations, the posterior conditional joint probability **b** over **x** and I can be obtained as

$$P(\mathbf{b}|\mathbf{x}) = C \cdot \Pi_j P\left(b_j\right) \cdot \Pi_{i=1..m} p_n(k(i)|\mathbf{b})^{x_i}, \tag{16}$$

with $C = A/B$

4. Policy Evaluation

The application of the discrete choice model allows the estimation of the probability of the different sentiments (specification reported in section 2). The application of the discrete choice model requires the vector of the parameters to be estimated and/or updated from the observation.

The parameters are estimated and/or updated in terms of the posterior conditional joint probability **b** over **x** and I (specification reported in section 3), obtained with the Bayesian approach. For the definition of the vector of the estimated parameters **b**, the posterior conditional joint probability is considered.

Given a sample of user, starting from the input variables;

(i) $P(\mathbf{b})$, the prior conditional joint probability **b** over I,

(ii) $\mathbf{x}_n$, the observed sentiments for each user n,

(iii) $\mathbf{y}_n$, the attributes for each user n.

with the Bayesian approach, the posterior conditional joint probability **b** over **x** and I, $P(\mathbf{b}|\mathbf{x})$, is obtained. The whole method is reported in Figure 1.

The figure is divided into three rows and three columns:

(i) the first row shows the input data (prior conditional joint probability **b** over **I**; observed sentiment for each user n; observed users' characteristics for each user n);

(ii) the second row shows the models in terms of application of the Bayes approach (second column) and application of the model for the sentiment prediction (third column);

(iii) the last row shows the intermediate (second column, posterior conditional joint probability **b** over **x** and **I**) and the final (third column, probability evaluation for each user n and for each alternative k) output data.

Sentiments are predicted with the discrete choice model reported in Subsection 2.2 considered as input characteristics of the users, $\mathbf{y}_n$, and the probability distribution obtained from the joint posterior conditional probability **b** over **x** and I, $P(\mathbf{b}|\mathbf{x})$. Note that the model shown in the (5) can be applied by adopting the parameters in term of values, i.e., the expected value or the mode, (i.e., Logit model) or in terms of probability distribution (i.e., mixed Logit model).

Different indicators (i.e., the parameters' elasticity) can be also evaluated as reported in the experimental section.

Sentiments are predicted in relation to the users' characteristics, and the influence of each characteristic on sentiments can be estimated. It allows evaluating different aspects (i.e., probability estimate; direct or elasticity of probability with respect to characteristics; direct or relative weight of the user characteristics).

The proposed model evaluates the impact of variables on user sentiments; it does not estimate the quantitative variations in transport demand values; it estimates the average fraction of sentiment for each alternative. The estimate of the weight of each variable provides information on user perceptions and therefore indicates the way forward to improve the utilities perceived. The estimate of the choice probability of a transport alternative cannot take place with the use of the models proposed in this paper.

The proposed model can be useful for the support of the decision maker regarding intervention policies in transport systems. The estimated weight for the parameters and their elasticity can be useful to evaluate the positive or negative effects of each variable and its percentage variation. The comparison between parameters provides indications on the relative weight with respect to the sentiments of the users. These indications can support the decision maker in identifying the variables to be modified considering the objectives to be achieved in order to improve the positive sentiments related to the alternatives that require an increase in the percentage of choice of users and vice versa.

This approach has some weak points to consider, reported in the following paragraphs.

The model reported is based on individual choices and user's characteristics that are generally not available or are available with an unknown level of reliability. In the application (section 5), the variables considered are of an aggregate nature, and the average value (or the average fraction, or the percentage) of users who choose each alternative is evaluated.

The web sentiments detected do not refer to a sample of users extracted from the population using the sampling rules. Sentiments refer to particular categories of users who use social media and the web; from the web, a sample of users with the same percentage distribution of a specific characteristic (i.e., age) of the population could be extracted. This aspect remains influenced by the reliability of the information made available by users on the web.

It is necessary to highlight the considerable potentials that derive from the use of data continuously present on the social media and web and that can allow an update of the models with the use of few resources. The model could estimate also the evolution of sentiments over time, also considering the actions implemented on the transport system, adopting a probabilistic process.

5. Experimentation

The main objective of the experimentation concerns the validation of the proposed method in order to verify the applicability in real contexts. For this reason, in Subsection 5.1, a numerical example is reported in order to apply step by step the proposed method; in Subsection 5.2, the method is applied in a real context.

In the two cases considered in this section, three subsections are reported: the subsections report, respectively, the contents of sections 2 (Discrete choice models), 3 (Bayes approach), and 4 (Policy evaluation).

In the numerical example, two users are considered, and for each user, it is assumed that three alternatives concerning sentiments are available. The numerical test concerns the use of the models specified in the paper and the possible application in a simple case. This application has the purpose of applying the models step by step and therefore supporting and explaining the application of the procedure by having all

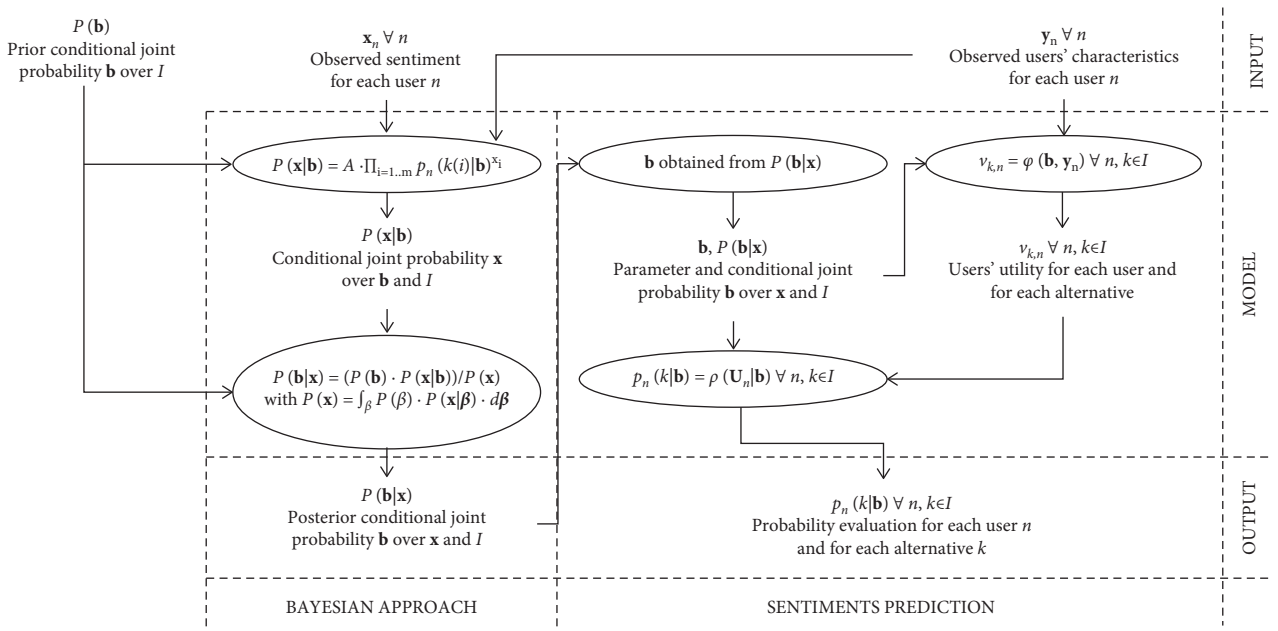

Figure 1: Bayesian approach and new sentiment prediction.

the necessary data. The numerical test is reported to provide the first indications, also in terms of the structure of figures and tables, which will also be adopted for the development of the test case.

In the test case, five Italian metropolitan cities are considered, and for each city, two alternatives concerning sentiments are considered. In the test case, the models proposed in the paper are applied, with the primary aim of obtaining the attributes relevant to users and the posterior probabilities of the corresponding parameters.

5.1. Numerical Example. In order to fix the attention, a small numerical example is reported. The scope is to test the models exposed in sections 2, 3, and 4.

It is assumed that 2 users ($n = 2$) named Z and W are observed. It is also assumed that each user considers three sentiments in the set of alternatives: positive (P), negative (N), and neutral (E). In this section, the characteristics do not take on a particular meaning, considering that the objective of the test concerns the application of the proposed method in a simplified general case that could not concern the transport sector. User characteristics are defined in section 5.2, which considers a real case.

5.1.1. Discrete Choice Models. It is assumed that

(i) the positive, negative, and neutral sentiments observed for the two users are, respectively, $x_{P,Z} = 8, x_{N,Z} = 8, x_{E,Z} = 6$ and $x_{P,W} = 6$, $x_{N,W} = 4, x_{E,W} = 3$; in this context, the vector $\mathbf{x}$ has $6\,(m = 6)$ entries:

$$\mathbf{x} = [8; 8; 6; 6; 4; 3]'. \tag{17}$$

(ii) the alternative characteristics are

$$\begin{aligned} \mathbf{y}_{P,Z} &= [2; 2; 3; 21]', \\ \mathbf{y}_{P,W} &= [3; 6; 1; 27]'. \end{aligned} \tag{18}$$

(iii) the choice model is Logit with parameter θ (7)

(iv) the initial values for the vector of unknown weights is

$$\mathbf{b} = [-0,5; -0,5; -1,0; 0,1]'. \tag{19}$$

(v) the specifications of the ratio between the expected value of the utility and the parameter of the distribution are (8)

$$\begin{aligned} \frac{v_{P,Z}}{\theta} &= b_1 \cdot y_{1,Z} + b_4 \cdot y_{4,Z} = -0,5 \cdot 2 + 0,1 \cdot 21 = 1,1, \\ \frac{v_{N,Z}}{\theta} &= b_2 \cdot y_{2,Z} = -0,5 \cdot 2 = -1,0, \\ \frac{v_{E,Z}}{\theta} &= b_3 \cdot y_{3,Z} + b_4 \cdot y_{4,Z} = -1,0 \cdot 3 + 0,1 \cdot 21 = -0,9, \\ \frac{v_{P,W}}{\theta} &= b_1 \cdot y_{1,W} + b_4 \cdot y_{4,W} = 1,2, \\ \frac{v_{N,W}}{\theta} &= b_2 \cdot y_{2,W} = -3,0, \\ \frac{v_{E,W}}{\theta} &= b_3 \cdot y_{3,W} + b_4 \cdot y_{4,W} = 1,7. \end{aligned} \tag{20}$$

(vi) With this assumption, the probabilities are

$$p_Z(P|b) = \frac{\exp(1,1)}{(\exp(1,1) + \exp(-1,0) + \exp(-0,9))} = 0,795 = 79,5\%,$$

$$p_Z(N|\mathbf{b}) = 9,7\%,$$

$$p_Z(E|\mathbf{b}) = 10,8\%,$$

$$p_W(P|\mathbf{b}) = 37,5\%,$$

$$p_W(N|\mathbf{b}) = 0,6\%,$$

$$p_W(E|\mathbf{b}) = 61,9\%. \tag{21}$$

5.1.2. Bayes Approach. The same specification adopted in Subsection 5.1.1 for the utility is assumed for Bayes approach; it has four parameters ($\mathbf{b} = [b_1, b_2, b_3, b_4]'$). For the prior probability $P(\mathbf{b})$, it is assumed that each parameter j has a priori Gaussian independent distribution with expected value (and mode) $[-0,5; -0,5; -1,0; 0,1]'$ and variance $[0.5, 0.5, 0.5, 0.5]'$. The prior distribution is represented in the Figure 2(a). For each parameter, the expected value, the mode, and the variance of the probability distribution are reported under each figure.

The likelihood function has to be evaluated for each value of the vector **b**. One of these values, without considering the constant A, in the point relative to the prior expected value **b** is

$$\Pi_{i=1..n} p_n(k(i)|\mathbf{b})^{x_i} = 0.795^8 \cdot 0,097^8 \cdot 0,108^6 \cdot 0,375^6 \cdot 0,006^4 \cdot 0,619^3 = 1,3E-27. \tag{22}$$

The marginal posterior probabilities $P(\mathbf{b}|\mathbf{x})$ relative to each parameters are represented in Figure 2(b). They are obtained applying the Bayes model reported in the (9). The constant C is evaluated assuming that the area under the posterior probability is equal to one.

Starting from a priori Gaussian distributions, posterior not Gaussian distributions (Figure 2(b)) are obtained (the expected value and the mode in each distribution are different). The posterior distributions are modified in the form respect the prior distribution, and they have a variance reduction.

It can be concluded that the observed sentiment values allow the calibration of the posterior distributions.

5.1.3. Policy Evaluation. Adopting the prior probabilities distribution for the parameters, the probability values are reported in Subsection 5.1.1.

The probabilities for the sentiment prediction could be evaluated. The comparison can be made on a user-by-user basis (Table 1). The probabilities evaluated with a priori and posterior parameters give different results, in some cases with a different level of magnitude. For an aggregate evaluation, for simplicity sake, in the last line of Table 1, the comparison is reported considering an aggregate indicator obtained as the average of the probabilities relating to the same sentiment. The probabilities in the two cases are different with a high difference in terms of values.

Neutral sentiment is not reported considering that, for each user, it can be obtained by difference. The initial value of **b** is chosen with arbitrary values; as expected, the posterior probability is closer to the prior probability (evaluating it with or without the posterior probability distribution of **b**).

Elasticity is a measure of the percent change in probability in response to a percent change in a user's characteristic. The elasticity is close to −1 for user W and sentiment N in response to the variation of characteristic number 2 (Table 2). It can also be observed that characteristic number 4 does not influence sentiment. The negative value (with difference in level of magnitude) provides the information that the increase in the characteristic generates a decrease in the corresponding sentiment and vice versa.

5.2. Test Case. The model is tested in a real case considering the sentiments acquired on the website for the bike mode in some Italian metropolitan cities (Florence, Rome, Bari, Naples, and Reggio Calabria).

The aim of the experimentation is to test the real applicability of the proposed model.

The procedure for obtaining the sentiment is developed in Ferrara [17] and from this work, the results on the number of positive and negative sentiments for the metropolitan cities considered are obtained. The sentiments observed relate to the period December 2020-January 2021. The positive and negative sentiments observed are reported in Table 3. The set I reported in (1) is in this specific case: $I = \{P, N\}$. The same table shows some characteristics useful for the estimation of the model.

The data used in this paper refer to 5 cities, and two sentiments are observed for each city. To obtain statistically more significant results, it is necessary to observe a larger number of cities.

In the test case, the observed variables are the percentages of positive and negative sentiments. A limit that should be considered concerns the use of sentiments extracted from the web, which could refer only to certain groups of users who are more likely to post comments on the social media (for example younger users). The probability model intends to estimate the average fraction (or percentage) of users with positive or negative sentiments as a function of the aggregate characteristics of the cities. The model can be used to estimate how the fraction

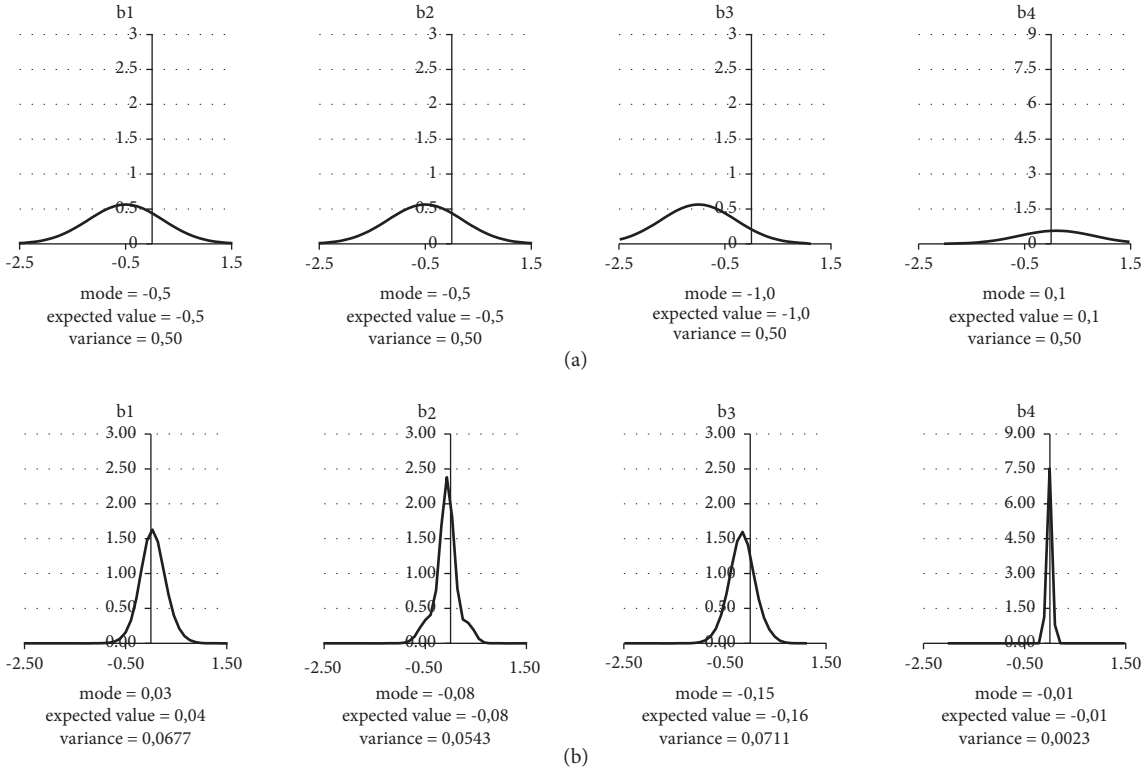

Figure 2: Prior and posterior probability of the parameters **b** in the numerical example. (a) Prior probability of the parameters **b** over I adopted in the numerical example. (b) Posterior probability of the parameters **b** over I and x obtained in the numerical example.

Table 1: Sentiment probabilities and observed frequencies evaluation in the numerical example.

Sentiment	Observed frequency (%)		Priori probability (%)		Posterior probability (%)			
					With probability distribution of **b**		Whit expected value of **b**	
	P	N	*P*	N	*P*	N	*P*	N
User Z	36,4	36,4	79,5	9,7	39,3	37,6	39,7	37,8
User W	46,2	30,8	37,5	0,6	40,1	29,0	41,0	28,4
Average	41,3	33,6	58,5	5,1	39,7	33,3	40,4	33,1

Table 2: Elasticity evaluation adopting the posterior parameters in the numerical example.

User	*Sentiment*	*User characteristics y*			
		1	*2*	*3*	*4*
Z	*P*	0,05	—	—	−0,12
	N	—	−0,10	—	—
	E	—	—	−0,37	−0,16
W	*P*	0,05	—	—	−0,16
	N	—	−0,36	—	—
	E	—	—	−0,11	−0,18

of the population with positive or negative sentiments is affected by the characteristics of the considered attributes. In this section, the probability model and the results obtained are to be considered as an estimate of the average fraction (or percentage) of users with positive or negative sentiments.

Table 3: Number of positive and negative sentiment for the Italian metropolitan cities considered.

Metropolitan city	x Number of sentiment *		y Population density **	Incident with bike ***
	Positive (P)	Negative (N)	Resident/km^2	Number (year 2019)
Firenze	34	13	288,7	478
Roma	67	42	811,8	459
Napoli	123	33	2653,4	122
Bari	37	54	326,2	185
Reggio Calabria	50	11	172,5	34
Average	62	31	846,9	255,6

Elaboration starting from: *Ferrara [17]; **Finocchiaro and Iaccarino [18]; ***Dati Statistici Generali, Automobile Club d'Italia, Localizzazione degli incidenti stradali, 2019.

5.2.1. Discrete Choice Models. It is assumed that

(i) five cities reported in Table 3 are observed (column 1);

(ii) positive and negative sentiments are observed with number of sentiments observed in the roll horizon, reported in Table 3 (columns 2 and 3);

(iii) two attributes are considered as alternative variables, the residential population density (y_1) and the number of incident with bike in the year 2019 (y_2), reported in Table 3 (columns 4 and 5);

(iv) the choice model is Logit with parameter θ (7)

(v) the initial values for the vector of unknown weights are obtained with a minimum square method (minimum of the sum of the square of the difference between the modelled average percentage and the observed frequency for the sentiments): $\mathbf{b} = [0,218; -0,542; 0,673]'$ with residential population (y_1) expressed as 'number of resident/(10^3 km^2)' and the number of incident with bike in the year 2019 (y_2) 'number of incident/(10^3 year)'; the ratio b_2/b_1 is about −2,5 considering the adopted unit of measurements;

(vi) the specifications of the ratio between the expected value of the utility and the parameter of the distribution for the generic metropolitan city M and positive (P) or negative (N) sentiment are (8):

$$\frac{\left(v_{P,M} - v_{N,M}\right)}{\theta} = b_1 \cdot y_{1,M} + b_2 \cdot y_{2,M} + b_3. \tag{23}$$

Whit these assumptions, the prior probabilities evaluated for the positive (P) sentiment and for the five cities reported in Table 3 are, respectively, ($p_C(P|\mathbf{b})$): 61,70%; 64,59%; 76,52%; 65,56%; 66,65%. The negative sentiment can be evaluated because for each city the sum of negative and positive sentiments is 100%.

5.2.2. Bayes Approach. For the prior probability $P(\mathbf{b})$, it is assumed that each parameter j has an independent a priori (normal) Gaussian distribution with expected value (and mode) $[0,218; -0,542; 0,673]'$ and each coefficient of variation equal to 1. The prior probability distributions are represented in Figure 3(a).

With the application of the Bayesian approach reported in (9), the posterior probability distributions of parameters **b**, over I and **x** obtained in the experiment, are represented in Figure 3(b).

As expected, the posterior distribution is modified from the prior distribution. The variances are greatly reduced, and the coefficient of variation is reduced from 1 to values in the range 0.3–0.7.

5.2.3. Policy Evaluation. A comparison between the observed frequency, the prior estimated average percentage, and the posterior estimated average percentage for positive sentiment in the analyzed cities is reported in Table 4. The estimated average percentage for negative sentiments can be evaluated considering that the sum with the estimated average percentage for positive sentiment in the same city is equal to 100%.

In the experimentation in real context, the initial values of the parameters are obtained with a minimum quadratic optimization between the observed frequency and the prior estimated average percentage. For this reason, the Bayes approach cannot give the posterior average percentage that is closer to the observed frequency than the prior average percentage. The main objective of this experimentation is to test the effect of the Bayes approach on the expected value, on the mode and on the variance, starting from optimized values.

Table 5 shows the elasticities of the posterior parameters. Parameters 1 and 3 (residential and constant density) have positive elasticity; parameter 2 (number of incidents) has negative elasticity. Parameters 1 and 2 have similar absolute value of elasticity, and in relation to the metropolitan city, the first or the second is the largest. Parameter 3 (constant) has the greatest elasticity, and this gives the information that the positive and negative sentiments on the bike have a high background in users not dependent on external variables. The elasticity of the model with respect to the two considered attributes (1) and (2) is very low.

The ratio between the expected value of the posterior probability of the parameters 1 and 2, $E(b_2)/E(b_1)$, is about −1,9. It is lower that of the a priori ratio. A similar value is obtained adopting the modes of the posterior probabilities. The attribute y_2 is relative to the number of incident with bike in the year 2019 (unit of measurement

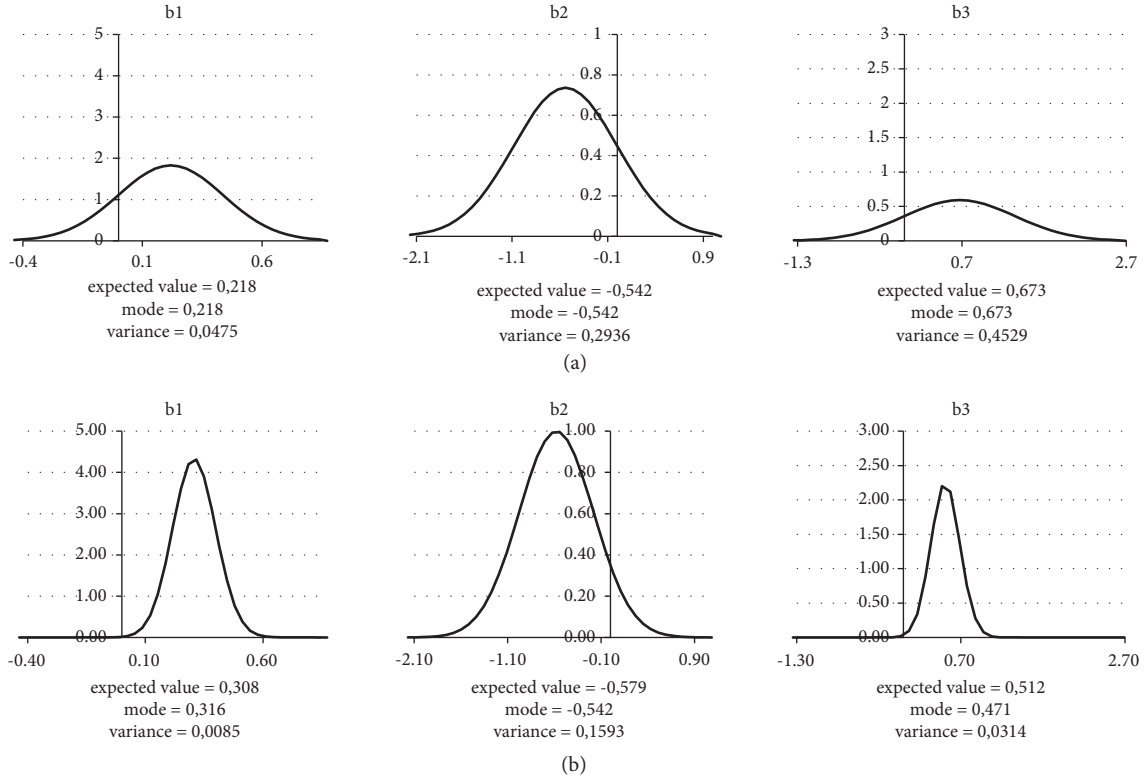

Figure 3: Prior and posterior probability of the parameters **b** in the experimentation case. (a) Prior probability of the parameters **b** over I adopted in the experimentation case. (b) Posterior probability of the parameters **b** over I and x obtained in the experimentation case.

Table 4: Sentiment average percentage and observed frequencies evaluation in the experimental case.

Sentiment	*Observed frequency (%)*	*Priori percentage (%)*	*Posterior percentage (%)*	
			With probability distribution of **b**	*Whit expected value of* **b**
	P	*P*	*P*	*P*
Firenze	72,3	61,7	58,0	58,0
Roma	61,5	64,6	62,1	62,2
Napoli	78,8	76,5	77,6	77,8
Bari	40,7	65,6	62,3	62,4
Reggio Calabria	82,0	66,7	63,2	63,3
Average	*67,1*	*67,0*	64,7	64,7

Table 5: Elasticity evaluation adopting the posterior parameters for the Italian metropolitan cities considered.

City	*Sentiment*	*User characteristics y*		
		1	*2*	*3*
Firenze	*P*	0,04	−0,12	0,21
Roma	*P*	0,09	−0,10	0,19
Napoli	*P*	0,18	−0,02	0,11
Bari	*P*	0,04	−0,04	0,19
Reggio Calabria	*P*	0,02	−0,01	0,19
Average	*P*	0,07	−0,06	0,18

'number of incident/(10^3 year)'), and the attribute y_1 is relative to the residential population (unit of measurement 'number of resident/(10^3 km^2)'). It means that, in terms of sentiments for the users, the reduction of 1 incident/year with bike is equivalent to the increase of 1,9 resident/km^2, considering fixed the other attributes.

Considering that the elasticity of the constant is high with respect to the elasticity of attributes 1 and 2, it can be assumed that the model indicates a preference for the bike mode but does not have a good level of explanation with respect to the attributes considered. Others specifications for the utility function are tested without obtaining signs or

statistical results satisfactory. This problem is probably a consequence of the low number of cities and sentiments observed compared to the number of calibrated parameters.

6. Conclusion

The most widely adopted demand models in the transport system to model the user's discrete choices are based on the theory of random utility. These models evaluate the behavior of the users' choices according to the trend of the supply system and the socioeconomic and territorial characteristics.

Choice models are typically calibrated using data observed directly in the transport system. In this paper, the possibility of calibrating choice models is evaluated starting from the text continuously available on social media and from the sentiments estimated. The specification of choice models based on a Bayesian approach is therefore adopted; it allows updating the prior average percentage distribution from the sentiment observed in social media.

After specifying the model, two numerical applications are reported. A first numerical example considers two users; the purpose is to verify the applicability of the method. A second application considers a real case; the sentiments of people living in five Italian metropolitan cities with respect to bike mode are analyzed, and a preference model for the mode is specified, calibrated, and validated.

The results obtained are encouraging: the sentiments present in the text available on social media allow us to define a demand model without using data observed directly in the transport system. Therefore, the objectives set out in the introduction of this manuscript have been achieved.

It opens up a new area of research. The results are preliminarily considering the low number of data involved and the study of a single level of choice. Future developments may involve extending the field and the levels of choice and integrating sentiments and traditional data. A possible extension could concern the use of the Bayes approach for the study of the evolution of sentiments over time as a function of the actions implemented in the transport system. The model should be tested considering a larger number of cities also considering the segmentation of the model according to the category of users who use the web and social networks.

Data Availability

The data that can be shared are included in the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was partially supported by the Dipartimento di ingegneria dell'Informazione, delle Infrastrutture e dell'Energia Sostenibile, Università Mediterranea di Reggio Calabria, and by the project "La Mobilità per the authors passeggeri come Servizio–MyPasS", Fondi PON R&The authors 2014–2020 e FSC "Avviso per la presentazione di Progetti di Ricerca Industriale e Sviluppo Sperimentale nelle 12 aree di Specializzazione individuate dal PNR 2015-2020", codice identificativo ARS01_01100. The author thanks Giulio Erberto Cantarella and Francesco Russo for the constructive comments and for the interesting suggestions received during the development of this research.

References

[1] T. A. Domencich and D. McFadden, *Urban Travel Demand: A Behavioural Analysis*, American Elsevier, New York, 1975.

[2] C. F. Manski, "The structure of random utility models," *Theory and Decision*, vol. 8, no. 3, pp. 229–254, 1977.

[3] H. C. W. L. Williams, "On the formation of travel demand models and economic evaluation measures of user benefit," *Environment & Planning A: Economy and Space*, vol. 9, no. 3, pp. 285–344, 1977.

[4] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, Mass, 1985.

[5] C. F. Daganzo and M. Kusnic, *Another Look at the Nested Logit Model. Technical Report UCB-ITS-RTR 92-2*, Institute of Transportation Studies, University of California, Berkeley, 1992.

[6] D. McFadden, "A benchmark comparison of state-of-the-practice sentiment analysis methods," in *Frontiers of Econometrics*, P. Zarembka, Ed., pp. 105–142, Academic Press, New York, 1978.

[7] C. F. Daganzo, *Multinomial Probit: The Theory and its Application to Demand Forecasting*, Academic Press, New York, 1979.

[8] S. Washington, P. Congdon, F. L. Mannering, and F. L. Mannering, "Bayesian multinomial Logit," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2136, no. 1, pp. 28–36, 2009.

[9] T. Arentze and H. Timmermans, "Social networks, social interactions, and activity-travel behavior: a framework for microsimulation," *Environment and Planning B: Planning and Design*, vol. 35, no. 6, pp. 1012–1027, 2008.

[10] B. Wessels, S. Kesselring, and P. O. Plaut, *How to Define Social Network in the Context of Mobilities?. Digital Social Networks and Travel Behaviour in Urban Environments*, Routledge Taylor and Francis Group, London and New York, pp. 27–42, 2019.

[11] P. Gonçalves, M. Araújo, F. Ribeiro, F. Benevenuto, and M. Gonçalves, *A Benchmark Comparison of State-Of-The-Practice Sentiment Analysis Methods*, EPJ Data Science, 2015.

[12] A. Serna, J. K. Gerrikagoitia, U. Bernabé, and T. Ruiz, "Sustainability analysis on urban mobility based on social media content," *Transportation Research Procedia*, vol. 24, pp. 1–8, 2017.

[13] F. Ali, D. Kwak, P. Khan, S. M. R. Islam, K. H. Kim, and K. S. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 33–48, 2017.

[14] F. Ali, D. Kwak, P. Khan et al., "Transportation sentiment analysis using word embedding and ontology-based topic modeling," *Knowledge-Based Systems*, vol. 174, pp. 27–42, 2019.

[15] A. Candelieri and F. Archetti, *Detecting Events and Sentiment on Twitter for Improving Urban Mobility*, ESSEM@ AAMAS, 2015.

[16] E. Cascetta, *Transportation Systems Engineering: Theory and Methods*, Springer, New York, 2009.
[17] A. Ferrara, *Modelli di domanda di mobilità: sperimentazione della sentiment analysis nelle città metropolitane per la scelta del modo. Tesi di Luarea in Ingegneria Civile*, Università degli Studi Mediterranea di Reggio Calabria, 2021.
[18] G. Finocchiaro and S. Iaccarino, "La funzione turistica dei territori delle città metropolitane," *Qualità dell'ambiente urbano – XIII Rapporto, ISPRA Stato dell'Ambiente*, vol. 74/17, pp. 557–563, 2017.