Information Processing and Management xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Marketing analysis of wineries using social collective behavior from users' temporal activity on Twitter

Gema Bello-Orgaz[*,a], Rus M. Mesas[b], Carmen Zarco[c], Victor Rodriguez[d], Oscar Cordón[e], David Camacho[a]

[a] Departamento de Sistemas Informáticos, ETSI de Sistemas Informáticos, Universidad Politécnica de Madrid, Calle de Alan Turing, s/n, Madrid 28031, Spain
[b] Telefonica de España, Ronda de la Comunicacion, 2, Madrid, 28050, Spain
[c] Department of Market Research, Universidad Internacional de La Rioja, Av. de la Paz, 137, Logroño, La Rioja 26006, Spain
[d] Computer Science Department, Universidad Autónoma de Madrid, Francisco Tomás y Valiente, 11, Madrid, 28049, Spain
[e] Instituto Andaluz Interuniversitario de Ciencia de Datos e Inteligencia Computacional (DaSCI), University of Granada, C/Daniel Saucedo Aranda, s/n, Granada 18071, Spain

ARTICLE INFO

Keywords:
Social networks
Marketing analysis
Temporal Twitter Activity
Social collective behavior
Temporal clustering
Hidden Markov models
Wineries

ABSTRACT

Marketing professionals face challenges of increasing complexity to adapt classic marketing strategies to the phenomenon of social networks. Companies are currently trying to take advantage of the useful collective knowledge available on social networks to support different types of marketing decisions. The appropriate analysis of this information can offer marketing professionals with important competitive advantages. This work proposes a new methodology to extract the social collective behavior of Twitter users concerning a group of brands based on the users' temporal activity. Time series of mentions made by individual users to each company's Twitter account are aggregated to obtain collective activity data for the companies, which is a consequence of both the company's and other users' actions. These data are processed using classical unsupervised machine learning techniques, such as temporal clustering and hidden Markov models, to extract collective temporal behavior patterns and models of the dynamics of customers over time for a single brand and groups of brands. The derived knowledge can be used for different tasks, such as identifying the impact of a marketing campaign on Twitter and comparatively assessing the social behaviors of different brands and groups of brands to assist in making marketing decisions. Our methodology is validated in a case study from the wine market. Twitter data were gathered from four regions of different countries around the world with important wineries (Italy: Veneto, Portugal: Porto and Douro Valley, Spain: La Rioja, and United States: Napa Valley), and comparative behavior analysis was carried out from the perspective of the use of Twitter as a communication channel for marketing campaigns.

1. Introduction

Social networks (SNs) have become one of the most important data sources for analyzing and extracting trends and collective opinions on a large number of topics (Bello-Orgaz, Jung, & Camacho, 2016; Cambria, Grassi, Hussain, & Havasi, 2012). Interactions

* Corresponding author.
E-mail addresses: gema.borgaz@upm.es (G. Bello-Orgaz), rusmaria.mesasjavega@telefonica.com (R.M. Mesas), carmen.zarco@unir.net (C. Zarco), victor.rodriguezf@uam.es (V. Rodriguez), ocordon@decsai.ugr.es (O. Cordón), david.camacho@upm.es (D. Camacho).

https://doi.org/10.1016/j.ipm.2020.102220
Received 30 July 2019; Received in revised form 4 February 2020; Accepted 5 February 2020

among users and the opinions they express within SNs provide public information about their preferences. Analysis of this information can offer marketing professionals some competitive advantages. Currently, this research area is facing increasingly complex challenges to adapting classic marketing strategies to the SN phenomenon. Companies try to take advantage of the useful collective knowledge provided by SNs to analyze the preferences and interests of their potential customers to match products to them, maximize the impact of their marketing campaigns on customers (increasing the company's profits as well), improve their reputation in the market, track how customers respond to their products, uncover groups of similar consumers, etc. (Alamaki, Pesonen, & Dirin, 2019; Bruckhaus, 2010; Jansen, Zhang, Sobel, & Chowdury, 2009; Maurer & Wiegmann, 2011). Upcoming changes stemming from the implementation of 5G, the fifth generation of broadband digital cellular network technology, will strongly impact digital and social media marketing (Singh, Saxena, Roy, & Kim, 2017). This new technology will not only achieve the quickest mobile connections ever but also completely change the way the Internet is used by consumers. As stated in Fang, *"The impact of this is simple: it will enable consumers to experience richer content offerings much more quickly than they have done so before.* Concerning marketing strategies on SNs, some changing factors are expected, such as the increasing importance of personalization (e.g., through programmatic advertising using ad networks based on big data and machine learning (ML) as well as through optimized location-based marketing strategies) and the real-time nature of big data analytics (owing to latency reduction and the exponential increase of the Internet of Things (IoT)). Moreover, because of the boom of these technologies, the concept of social IoT (Araniti, Orsino, Militano, Wang, & Iera, 2016; Atzori, Iera, Morabito, & Nitti, 2012) is experiencing increasing interest by marketing professionals because the combination and integration of ML and IoT methods allow us to generate new types of intelligent (and customized) services to end users. These methods also allows us to make massive use of augmented reality and video advertising (van Esch, 2020). However, the use of 5G communication technology also raises new challenges in the area of data transmission in terms of finding efficient methods of mass data transmission, such as the use of opportunistic SNs (Guan & Wu, 2019; Wu, Chen, & Zhao, 2019a; 2019b).

One of the most popular SNs is Twitter, which offers new challenges in different research fields, such as marketing campaigns (Bello-Orgaz, Menéndez, Okazaki, & Camacho, 2014; Chen, Wang, & Wang, 2010), financial prediction (Asur, Huberman et al., 2010), cybersecurity (Javed, Burnap, & Rana, 2019), and public healthcare (Bello-Orgaz, Hernandez-Castro, & Camacho, 2016; Collier, 2012). A very extended application of social network analysis (SNA) in marketing is related to sentiment analysis (Cambria, Das, Bandyopadhyay, & Feraco, 2017; Cambria et al., 2012) as sentiment mining techniques can be used for different marketing issues. Companies are progressively more interested in collecting and predicting the attitudes of users toward their products and brands. Users are more enthusiastic about sharing their opinions on social media every day, and it is assumed that these opinions will define future products and services Owyang. For example, as described in Cambria (2016), affective computing and sentiment analysis can enhance the ability of customer relationship management to reveal which features customers enjoy.

An increasing amount of graphical data is being extracted from SNs, and an important question that arises is how to represent this information for SNAs (Cavallari, Zheng, Cai, Chang, & Cambria, 2017; Fang et al., 2016). Traditionally, researchers propose multiple recursive or iterative algorithms to process such structured knowledge for analytical purposes. However, in the last few years, the concept of graph embedding has been introduced to address more complicated analytics tasks, requiring sophisticated algorithms that do not scale well to the size of modern problem settings. The graph embedding concept projects a topological graph structure into a Euclidean space, i.e., a low-dimensional space, for further application (Liu et al., 2017). Community structure is one of the main network properties for extracting the collective behavior of SNs, and it too has been considered in graph embedding (Cavallari et al., 2017).

In this contribution, we tackle a different task to analyze Twitter data for marketing applications. We propose a new methodology for extracting the social collective behavior of Twitter users concerning a group of brands based on the users' temporal activity within the SN. The timing of users' activity in social media and web communities has become an interesting aspect of study for social communication in the last few years (Rybski, Buldyrev, Havlin, Liljeros, & Makse, 2009). The global activity of a user group in an SN, represented as an aggregation of the individual activities of its members, can lead to correlations emerging from collective behavior in communication patterns at the level of the entire community (Rybski, Buldyrev, Havlin, Liljeros, & Makse, 2012). This explanation could be the social effects resulting from the SN dynamics that induce persistent fluctuations as information cascades; i.e., activity patterns could be a consequence of the superimposition of many cascades (Rybski et al., 2012). Taking all the latter as a base, we consider different levels of dynamic social collective behavior. First, the temporal user's behavioral pattern in terms of **mentions to each company's Twitter account** is considered. This time series reflects the collective behavior of consumers in response to both the brand's actions and other users' actions about the brand on Twitter. Second, the time series of the different companies from a group (e.g., a market sector or a geographical region) are aggregated to represent a higher level of collective behavior. This allows us to extract the **prototypical behavior of the group of companies** (a cluster's dynamic centroid) and obtain a **time-dependent collective behavioral model of the group**. The analysis of these centroids and their respective temporal models provides additional insights into the social behavior of users for the different groups of brands. Such analysis can also be used to study the cohesion of a group according to the individual behavior of each brand and identify similarities and differences among the different groups. Thus, it can be applied as a technological watch tool for different marketing tasks, such as identification of the impact of a marketing campaign on Twitter (brand attribution models), targeting and personalization of companies' customers generating a potential competitive advantage, and comparative assessment of social behaviors from different brands and groups of brands to assist in making marketing decisions.

Our methodology is based on the analysis of Twitter data using well-known artificial intelligence techniques. These techniques, which belong to the ML and data mining areas (Larose & Larose, 2014), are particularly useful for extracting both knowledge and patterns from datasets in a human-understandable structure; this uncovered information can then be used to improve marketing strategies of companies. We focus on unsupervised methods that do not require human-labelled information to operate (Albalate &

Minker, 2011). This is a critical feature for domains where the process of labelling data can be highly time-consuming or even impossible, as happens with (massively gathered) data from SNs. In particular, we rely on the use of *time series clustering* (Liao, 2005) to detect groups of brands with the same temporal social behavior and hidden Markov models (HMMs) (Visser, 2011) to model the dynamics of customer behavior over time. Visualization methods for HMMs state transition patterns are considered to provide the expert with powerful representations of the evolution of user activity with respect to a particular brand over time.

Our case study focused on the wine industry, which is strongly related to SNs nowadays (Dolan, Conduit, Fahy, & Goodman, 2016; Fuentes Fernández, Vriesekoop, & Urbano, 2017; Szolnoki, Dolan, Forbes, Thach, & Goodman, 2018; Wilson & Quinton, 2012). The proposed methodology allowed us to carry out a comparative assessment of different wine-producing regions from a marketing perspective. To do so, a dataset of winery-related tweets from several wine regions of the world (Italy: Veneto, Portugal: Porto and Douro Valley, Spain: La Rioja, and United States: Napa Valley) was gathered first. Once all this information was processed, a set of time series of mentions, representing activity for each winery on Twitter, was generated. Using these time series, collective behavioral models of the regions were created by training some HMMs, from which the (mention-based) time series of the most representative wineries for each region were selected. The selection of those wineries, which were considered to be representatives of their respective areas on Twitter, and the generation of their behavioral models are among the main contributions of the proposed methodology. To address the issue of the selection of those wineries most representative for each region, the application of *a time series clustering method* was performed to detect *groups* of wineries with similar behaviors to later select those most representative. Finally, visualization and analysis of the resulting models on Twitter by region were carried out.

The rest of the paper is structured as follows. Section 2 presents background information by reviewing the basis of HMMs and clustering for the time series. It also presents specific studies related to marketing activities using SN information by wine companies. Section 3 describes the different stages of the proposed methodology. Section 4 shows some experimental results and provides in-depth analysis of these results. Finally, Section 5 draws some general conclusions and discusses possible future research based on this work.

2. Background

Because the proposed methodology considers the use of social collective behavior principles and unsupervised learning techniques (HMMs and time series clustering) for extracting the temporal collective behavior of wine companies on SNs, this section first briefly introduces both techniques. Then, the case study domain is analyzed by providing a brief survey of marketing studies in SNs for wineries.

2.1. Hidden Markov models

HMMs are stochastic models mainly used for modeling and predicting sequences of symbols and time series in general. They are characterized by a set of N discrete (*hidden*) states $S = \{S_1, \cdots, S_N\}$, which can be interpreted as phases in a cognitive process that produces typical behaviors (Visser, 2011). The term *Markov* pertains to the time-dependence between consecutive states S_t, which follows a Markov process. This means that the current state S_t only depends on the previous state S_{t-1} and not on the earlier ones, i.e.,

$$P(s_{t+1} = S_j | s_t = S_i, \cdots, s_1 = k) = P(s_{t+1} = S_j | s_t = S_i) \quad 1 \le i, k \le N$$

The transition probabilities between the states of the model are denoted by a stochastic $N \times N$ square matrix, called a *transition matrix*, with entries

$$a_{ij}(t) := P(s_{t+1} = S_j | s_t = S_i), \quad 1 \le i, j \le N \tag{1}$$

This is a stochastic process, and so we have that $\sum_{j=1}^{N} a_{ij} = 0$ for all $1 \le i \le N$. As in any Markov chain, we need to specify the set of initial state probabilities, Π, which is defined as

$$\Pi_i := P(s_1 = S_i), \quad 1 \le i \le N \tag{2}$$

The term *hidden* in an HMM indicates that underlying states cannot be observed directly during the process; what we see is the *emission* of that state. Although an HMM emission can be both continuous and discrete, this work focuses on social activity, which is measured as a numerical variable. Thus, only continuous observations are considered. The model can emit only one observation in each state at each time step. Continuous observations are generated from a probability distribution function, which is typically a Gaussian distribution. The mean and variance of observations associated with state i are denoted by $\mu_{i=1\cdots N}$ and $\sigma^2_{i=1\cdots N}$, respectively.

In summary, a Gaussian HMM (λ) can be defined as the tuple

$$\lambda := \{S, V, A, \mu, \sigma^2, \pi\} \tag{3}$$

Three main computational issues need to be addressed with HMMs are as follows:

- *Sequence recognition*, which addresses how to compute the probability that a given observation sequence $o = o_1 o_2 \cdots o_T$ is produced by a model λ. This probability, written as $P(o|\lambda)$, is called the *sequence (log-)likelihood*, and it can be computed by the so-called forward-backward algorithm (Rabiner, 1989).
- *Sequence decoding*, which involves determining what sequence of hidden states $s = s_1 s_2 \cdots s_L$ is most likely to produce a given sequence of observation symbols o using a specific model λ. This issue is addressed by using the popular *Viterbi* algorithm (Forney, 1973).

- *Model training*, which addresses the estimation of the parameters of an HMM (see Eq. (3)) based on recorded sequential data. This is usually solved by the so-called *Baum–Welch* algorithm (Baum & Petrie, 1966), which, in brief, is a form of expectation-maximization algorithm that tries to maximize the likelihood of a set of observation sequences $o^1 \cdots o^K$ to be produced by a model λ. Formally, this algorithm computes the optimal model $\hat{\lambda}$ as follows:

$$\hat{\lambda} = argmax_{\lambda}\left(\sum_{K} \log P(o^K|\lambda)\right) \tag{4}$$

One important aspect to consider when fitting an HMM to a given dataset is that the number of hidden states, N, must be known in advance, which is often unrealistic. To choose an optimal number of states without prior knowledge about the model topology, HMMs are usually compared with the Bayesian information criterion (BIC) (Burnham & Anderson, 2004). This metric is defined as follows:

$$\mathrm{BIC}(\lambda) = -2\log\left(\sum_{K} \log P(o^K|\lambda)\right) + P\log(K), \tag{5}$$

where P is the number of parameters in the model and K the number of observations used to train the model. As observed, BIC penalizes the likelihood of a model by a complexity factor proportional to the number of parameters in the model and the number of training observations, so it gives advantage to simple and general models. The lower the BIC score is, the better the model will be.

With regard to the application of these models in the literature, HMMs and derivatives have been proven to generate comprehensive models. They have also achieved good results in predicting abnormal behavior in many fields, including speech recognition (Rabiner, 1989), aviation and pilot performance analysis (Rodríguez-Fernández, Gonzalez-Pardo, & Camacho, 2018), and, what is more relevant for this work, marketing analysis and SNA. On the one hand, HMM can capture the dynamics of customer relationships. In Hassan and Nath (2005), the authors proposed a model that enabled marketers to dynamically segment their customer base and to examine methods by which the firm could alter long-term buying behavior. On the other hand, SNA can be used to represent social relationships, which in turn can be modeled using an HMM. In Vinciarelli and Favre (2007), the structure of social relationships between people involved in broadcast radio news was modeled using an HMM. There has also been research involving HMMs, such as that described in Osotsi (2016), where the authors used an HMM to quantify how conversations in Twitter evolved in response to two major events: an unexpected school shooting and the Super Bowl.

2.2. Clustering methods for time series

Clustering methods (Jain, Murty, & Flynn, 1999) can be described as a blind search on a collection of unlabeled data, where elements with similar features are grouped in sets. Elements included in the same cluster should be similar, and elements included in different clusters should be dissimilar. For this reason, it is necessary to define a similarity measure from which this type of algorithm can assign groupings.

Depending on the technique used for representing data, i.e., how the distance (similarity) between data elements is measured and how grouping the data elements is performed, there is a great variety of clustering methods. However, hierarchical and partitioning approaches can be considered the main methodologies for clustering. Partitioning methods consist of identifying a disjoint division of the data and optimizing a clustering criterion (metric or cost function). Hierarchical approaches nest the clusters based on a criterion (similarity) for merging or splitting them.

K-means (Macqueen, 1967) is one of the most well-known partitioning techniques that, given a fixed number (k) of clusters $C = \{C_1, \cdots, C_k\}$, looks for the best division of the dataset into k groups using a specific measure of distance d. In this algorithm, each cluster is represented by the mean value of the elements in the particular cluster (centroid). It is based on an iterative process, which starts with a random selection of the initial cluster centroids $m_1^{(1)}, \cdots, m_k^{(1)}$. Then, a two-step algorithm is applied. The first step (assignment step) assigns each element x to the nearest cluster centroid, so the clusters at time t are denoted as

$$C_i^{(t)} = \{x: d(x, m_i^{(t)}) \leq d(x, m_j^{(t)}) \quad \forall\, j,\ 1 \leq j \leq k\}$$

where each x is assigned to exactly one cluster, even if it could be assigned to more than one cluster. Then, in the second step (update step), the new cluster centroids are calculated.

$$m_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{x_j \in C_i^{(t)}} x_j$$

The algorithm converges when the cluster assignments no longer change.

When the input data of the clustering algorithm are a time series (as in the methodology proposed in this work), existing classical algorithms, such as *K*-means, cannot be used directly (Liao, 2005). These algorithms should be extended or adapted to handle this type of data using other distance metrics to measure the similarity of the time series. dynamic time warping (DTW) (Sakoe, Chiba, Waibel, & Lee, 1990) is one of the best-known distance metrics for time series. It allows us to compare time series of variable sizes. Also, it is robust to shifts or dilatations across the time dimension. Formally, given two time series $\mathbf{x}$ (with length n) and $\mathbf{y}$ (with length m), representing them as column-wise matrices, we write $\mathcal{A}_{n,m} \subset \{0, 1\}^{n \times m}$ for the set of binary alignment matrices that connect the

upper-left (1, 1) entry to the lower right (n, m) entry using only $\downarrow$, $\searrow$, $\rightarrow$ moves. Given the cost matrix $\Delta(\mathbf{x}, \mathbf{y}) = [\delta(x_i, y_i)]_{ij} \in \mathcal{R}^{n \times m}$, where δ is the Euclidean distance, the inner product $\langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle$ of that matrix with alignment matrix A in $\mathcal{A}_{n,m}$ gives the score of A. The DTW computes the minimum of those scores, i.e.,

$$\mathrm{DTW}(\mathbf{x}, \mathbf{y}) = \min_{A \in \mathcal{A}_{n,m}} \langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle \tag{6}$$

A smoothed formulation of DTW, called Soft-DTW, was recently proposed by Cuturi and Blondel (2017), showing that this new approach is particularly well suited to average and cluster time series.

Finally, one important problem in clustering is to define the number of clusters. Several clustering algorithms do not take into account the fitting of this parameter within their process; it has to be introduced as a fixed parameter. Several internal validation indices can be used to compare and select the best discrimination.

- *silhouette coefficient (SC)* (Rousseeuw, 1987): Given an observation i, the silhouette for that observation, $s(i)$, is defined as

$$s(i) = \frac{b_i - a_i}{\max(b_i, a_i)},$$

where a_i is the average intra-cluster distance for i and b_i the average inter-cluster distance with respect to the nearest cluster to i, i.e.,

$$b_i = \min_{C_k \in C \setminus C(i)} \sum_{j \in C_k} \frac{dist(i, j)}{n(C_k)},$$

where $C(i)$ represents the cluster to which i is assigned and $n(C_k)$ the number of observations contained in cluster C_k. The closer $s(i)$ gets to 1, the more confidence we have that i is well assigned, and *vice versa* if $s(i)$ gets close to -1. Finally, to compute the silhouette coefficient of a discrimination, we simply compute the average silhouette value for each observation.

$$S(C) = \frac{\sum_{C_k \in C} \sum_{i \in C_k} s(i)}{|C|} \tag{7}$$

The result lies in $[-1, 1]$, and it should be maximized to achieve good discrimination.

- *Calinski–Harabasz Index (CHI)* (Caliski & Harabasz, 1974): Given an observation for k clusters, the Calinski–Harabasz score establishes a ratio between the separation and cohesion of the partition.

$$CH(k) = \frac{B_k(N - k)}{W_k(k - 1)} \tag{8}$$

where B_k is the between (separation) group dispersion matrix and W_k is the within-cluster (cohesion) dispersion matrix. A higher CHI score relates to a model with better-defined clusters, so it should be maximized to achieve good discrimination.

- *Davies–Bouldin Index (DBI)* (Davies & Bouldin, 1979): Given an observation for cluster k, this index is defined as the average similarity measure of each cluster with its most similar cluster.

$$DBI = \frac{1}{k} \sum_{j=1}^{k} maxR_{ij} \tag{9}$$

.

Here, R_{ij} represents the ratio between within-cluster distances and inter-cluster distances, defined as

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}, \tag{10}$$

where s_i is the average distance between each point of a cluster and the centroid of that cluster and d_{ij} is the distance between cluster centroids i and j. The minimum value of this index is zero, and the lower values of the index relate to a model with better separation between the clusters. Therefore, it should be minimized to achieve good discrimination.

2.3. Studies on marketing strategies of wine companies using social networks

Marketing professionals have recognized the value of social media platforms, integrating them quickly in the development of marketing strategies, especially in actions focused on communication (Stelzner, 2014). In particular, wine is an experiential product that inherently implies socialization and builds communities around the pleasure of sharing experiences (Bruwer & Wood, 2005; Szolnoki et al., 2018). Social media and communication through them improve that common element of appreciation and consumption of wine by creating communities, and so it is particularly important for those involved in the wine industry to have an active presence on SNs (in particular on Twitter) (Wilson & Quinton, 2012).

The wine industry is progressively recognizing the increasingly relevant role that SNs have as an appropriate and valuable tool to reach consumers. When consumers search for wines and wineries on the Internet, they are bombarded with a massive volume of

brand messages. This implies that content must be creative, polished, and clear if the brand wants to capture the attention of users. Producing a good wine is important, but there is also a need to give it the presentation it deserves, communicating accurately with consumers. Communication activities are in a state of evolutionary development in which new trends continually arise, and wineries should adapt to these new trends if they want to be successful in their campaigns (Stelzner, 2014).

Researchers and marketing specialists insist on the importance of media and SNs as an easy and low-cost service, making them a communication choice that provides an immediate connection with a large number of consumers (Dolan et al., 2016). However, these experts also argue that the wine industry still has a long way to go before SNs become a truly efficient communication and marketing tool (Fuentes Fernández et al., 2017). Without doubt, organizations depend on their communication policies and the image of their products in the media. The goal of wineries is to make good wine at an attractive price, but it is also their job to effectively communicate these characteristics to their present and future interest groups, their consumers, and the general audience. This task must be handled through both traditional media (television and press) and SN platforms and channels. Wineries should consider how much a good communication policy affects the mobilization of the wine consumer, how communication creates and enhances the image of winery brands, and how that process promotes the purchase of wine (Foster, Francescucci, & West, 2010). Hence, in addition to mass information markets, especially television and press, fragmentation is currently taking place on the Internet in numerous "mini-markets, each of them requiring its own communication tools and specific approach to tackle the growing sophistication of consumers (Mariani, Pomarici, & Boatto, 2012).

Traditional media channel their messages about wines through advertising and through the news produced by their editors. It is up to each winery to ensure free dissemination spaces and thus promote their wines, their ideas, and their people to build an attractive and explanatory journalistic account of its activity. However, consumers want to participate more than ever in communication processes. For this reason, the question is no longer just how to reach them but also how they arrive at wineries and how they interact with each other (Bruwer & Wood, 2005).

Some wine brands have already made successful use of SNs, with documented examples showing that both small and large wineries have achieved a positive return of investment through the implementation of successful strategies in these media. Different academic studies have explored the practices of SNs within the wine industry. Among the wineries studied in Australia, Canada, New Zealand, Spain, Italy, South Africa, and the USA, 35% reported using SNs as the main means of communicating with customers about events in the winery and for promoting their wines (Alonso, Bressan, O'Shea, & Krajsic, 2013). Some experts argue that SNs help with wine sales since eWOM (Ferguson, 2008) is particularly effective among wine consumers, and the socialization aspect of these networks is appropriate for wine, allowing consumers to exchange information and encourage others to try different wines (Wilson & Quinton, 2012).

Specifically, Spanish wine brands, one of the wine markets tackled in the current contribution, compete to attract and retain consumers, and many of these companies are adopting SNs to reach their consumers and communicate their brand, quality, and personal experience Vinography. Even if some experts in the wine sector have expressed concern about the ineffective communication policies of most Spanish wineries (despite the significant media presence wines have Castro Galiana), some wineries have actually developed successful communication on social media like Twitter. In Zarco, Santos, and Cordón (2019), the adoption of Twitter by wineries holding the Qualified Denomination of Origin Rioja in Spain was analyzed using a technological watch tool based on SNA and network-based information visualization. Visual representations (maps) of similarity relations with respect to the positioning of the different organizations on Twitter (in terms of engagement and impact) are built from presence and impact data. The Twitter communication model developed, as well as its impact in terms of their degree of engagement, presence, and activity, can be established from the distribution and spatial location of each company on the map.

3. Methodology for fxtracting temporal collective behaviors of wineries on Twitter

This section introduces the proposed methodology to create collective behavior models of wineries on Twitter. A schema of the methodology is depicted in Fig. 1. *Step 1* involves the creation a dataset of winery-related tweets from different wine regions of the world (major regions in terms of wine production).

In terms of the dynamics of communication on SNs, a topic is first generated and then, as the topic spreads within the SN, it gains popularity or interest depending on the contribution of users participating in the discussion (De Choudhury, Diakopoulos, & Naaman, 2012). Several works in the literature state that there are two main factors used to evaluate the interest or value of information (De Choudhury, Mason, Hofman, & Watts, 2010; De Choudhury, Sundaram, John, & Seligmann, 2009; Ghosh, Banerjee, & Yen, 2016): the participation level in the discussion that generates such information and the time of discussion. Therefore, the level of interest in a topic on an SN could be measured based on two factors:

- **Event factor**: The degree of participation on an SN via posts, comments, messages, mentions, likes, or dislikes. The level of activity of participants can be evaluated by measuring these events to check if the topic is motivating or not.
- **Time factor**: Over time, discussions on a topic may pass through several states because the interest of participants on the topic may change, acquiring more or less importance. This factor can be represented as a time series of measured events.

Therefore, to measure and analyze behavior in the communication dynamics of wineries on Twitter from a marketing viewpoint, data extracted are preprocessed to obtain a representation that considers both factors (*Step 2*). In this particular case, the dataset was preprocessed to generate time series that represent the evolution of social activity related to each winery in terms of mentions (*event factor*) made by users over a specific period of time (*time factor*).

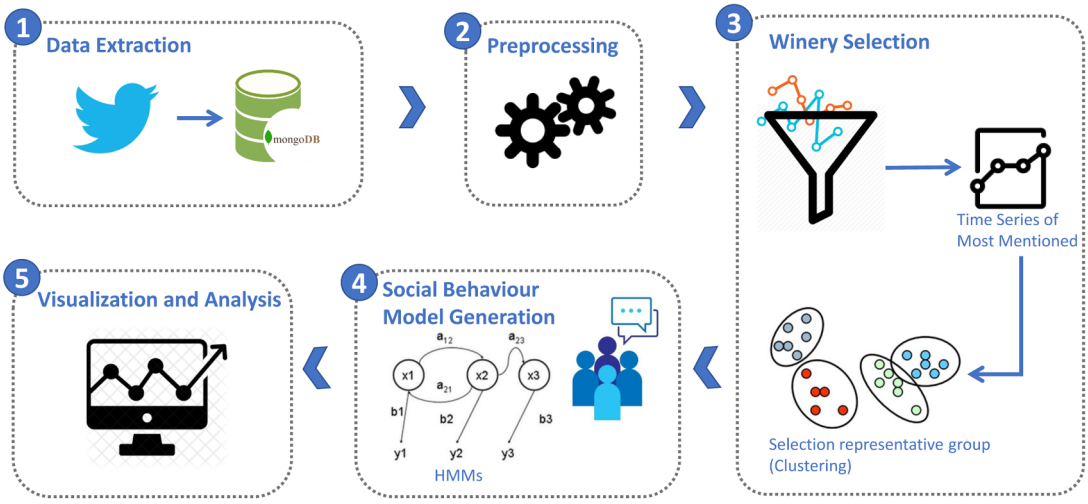

Fig. 1. Diagram of the proposed methodology for collective behavior extraction from users' temporal activity on Twitter.

Another aspect that must be considered when analyzing the dynamics of communication on SNs is that social groups of people have common interests and thus opinions are generated. A research work analyzing correlations in users' activity on SNs reported that correlations in activity occur as short periods of many events that are separated by long periods of no events at the individual level (such patterns can be characterized with the inter-event times) (Rybski et al., 2012). However, the activity of the entire community presents long-term correlations. This result suggests the existence of collective behavior, possibly arising from nontrivial communication patterns through the embedding SN. This fact could be due to social effects, i.e., dynamics on the SN induce persistent fluctuations, such as cascades.

To study the existence of the latter collective behavior in this contribution, Twitter data were extracted from several wineries in different wine-producing regions. Because the wineries may have different levels of activity, the most representative ones for each region were extracted using time-series clustering in the *third step*. In *Step 4*), an HMM-based behavioral model was fitted for each region, and these resulting models were used to identify different states of activity that each winery region usually has on Twitter and to analyze its evolution over time.

As mentioned above, depending on the type of people participating in a conversation on an SN and the type of actions they take, discussions may go through different states. In the work presented by De Choudhury et al. (2009), these states are categorized as "Interesting Motivating", "Interesting Deviating", "Not Interesting Motivating", and "Not Interesting Deviating". An "Interesting Motivating" topic requires influential participants who motivate others to add positive remarks to the topic. An "Interesting Deviating" topic could be interesting but the intervention of users is not very influential or relevant, and thus other participants have deviated away. Similarly, a "Not Interesting Motivating" topic may not be interesting, but due to the effort of some group members, the status of the topic takes a popular turn as other members begin to participate in the conversation. Finally, a "Not Interesting Deviating" topic may not be interesting and no group or member takes the initiative to make it more interesting. Therefore, the new knowledge about the states that provide all the HMMs generated with the proposed methodology will offer valuable information. In the current case, this knowledge allowed us to perform a better marketing analysis on the level of interest or relevance that wineries have for users on Twitter. This is why finally using all these HMMs, a comparative assessment was carried out, analyzing in detail the similarities and differences among social behaviors of each region from a marketing viewpoint (*Step 5*).

The steps of the proposed methodology and the techniques applied in each of them for the current case are shown in detail below:

1. **Data extraction**: Data collected to perform the analysis for different winery regions were extracted from a single source: Twitter application programming interfaces (APIs). Users on Twitter generate over 400 million tweets every day, and these tweets are available through public APIs that provide functionalities for searching by keywords, hashtags, phrases, geographic regions, user names, etc. Using these APIs, each tweet that mentioned one of the winery's Twitter accounts considered in the study was collected for analysis. The extracted tweets were stored in a database (MongoDB), together with all their related information, such as creation date, location, source, retweet count, etc.
2. **Data preprocessing for the generation of activity time series**: Raw data collected from Twitter were prepared for further analysis. To study the activity of wineries on Twitter and the interest generated in their potential clients, two main factors were considered: *event* and *time factors*. Based on both factors, the behavior of each winery on the SN was represented by a time series (time factor) of mentions (event factor). Thus, in this step of the methodology, a time series was created for each winery containing the number of daily mentions that winery received on Twitter, as shown in Fig. 2.
3. **Winery selection using time series clustering**: The data extracted from several wineries of each region had different levels and types of activity or interest within the SN. The wine industry does not have a very active presence on SNs, so some of the Twitter accounts identified had very low activity and were not included in this study. To generate models of the collective behavior of

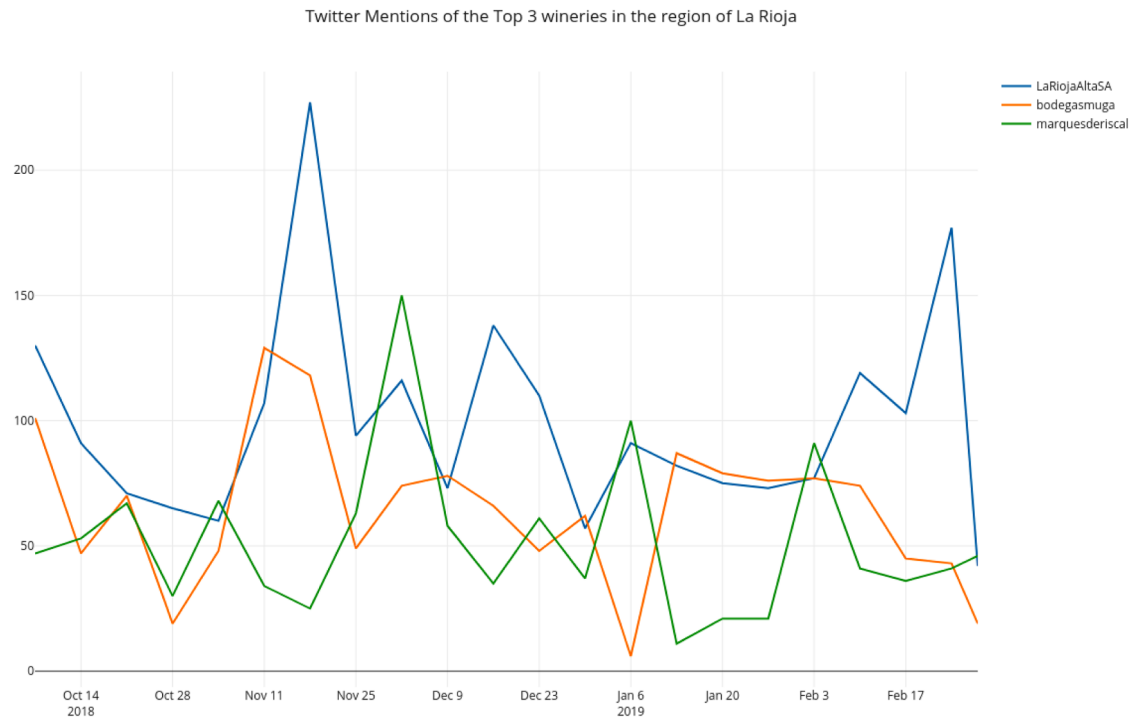

Fig. 2. Example of the time series of Twitter mentions generated for the top three most-mentioned wineries in La Rioja.

regions according to the activity of their wineries on Twitter, it was important to select only those wineries that had some activity and could thus represent general behaviors. Therefore, this winery selection process consisted of a double filtering that applied the following criteria:

- *Activity filter by region*: Only those wineries with a number of mentions higher than the average of the total number of mentions (wineries with high or medium Twitter activity) were selected to create collective behavior models of the region's activity.
- *Selection of a representative group of wineries for each region using clustering methods*: Those wineries having similar and representative behavior in their region needed to be identified. This task was addressed using *time series clustering methods* on their daily sequences of mentions.

  In particular, in this work, the `tslearn` Python package (Tavenard, Faouzi, & Vandewiele, 2017) was used to apply time series clustering. This package provides ML tools for time series analysis. Specifically, its clustering module contains three partitioning methods adapted to handle time series:

  (a) *K-means* (Liao, 2005): Given a fixed number (k) of clusters, the module looks for the best partition of the dataset into k groups using a specific distance metric d. This clustering algorithm has two main input parameters to define the k number of the resulting clusters and the distance metric used (d). Tslearn provides three well-known distance metrics for time series: Euclidean, DTW (Sakoe et al., 1990), and Soft-DTW (Cuturi & Blondel, 2017)).

  (b) *Global alignment kernel K-means* (Cuturi, 2011; Dhillon, Guan, & Kulis, 2004): This is an improved version of k-means, where elements (points) of the original (input) space are mapped to a higher-dimensional feature space using a nonlinear function before performing the grouping. Then, kernel K-means partitions the points by linear separators in the new space.

  (c) *KShape* (Paparrizos & Gravano, 2015): This is an iterative method that uses a normalized version of the cross-correlation measure as a distance metric to consider the shapes of time series in the computation of cluster centroids and the assignment of time series to clusters.

  Choosing the best method for a given dataset is a difficult task. Another critical point is the choice of the optimal number of clusters. To address these issues for the current case, an iterative process was carried out, testing a range of k values for each algorithm and computing some cluster validation indices measuring the quality of the obtained partitions to compare them. The three indices introduced in the background section were used: SC (Rousseeuw, 1987), CHI (Caliski & Harabasz, 1974), and DBI (Davies & Bouldin, 1979). To choose the best discrimination based on these three validation indices, a final validation rating (VR) was defined that balanced the scores obtained for each of them.

$$VR(k) = \frac{SC(k) - min_k SC}{max_k SC - min_k SC} + \frac{CHI(k) - min_k CHI}{max_k CHI - min_k CHI} + \left(1 - \frac{DBI(k) - min_k DBI}{max_k DBI - min_k DBI}\right) \tag{11}$$

The use of this criteria allowed choosing the partition that guaranteed reasonable values for all the quality indices. Because the

three indices have different scales, their values were first normalized in [0,1]. Because the first two coefficients showed higher values for better partitions and the third one was the opposite, its value was inverted. The aggregated index value thus ranged from 0 to 3, with a high value indicating a good partition.

Finally, to conclude the selection process, the best clustering algorithm found was applied with its best settings to obtain groups of wineries for each region. Then, the Soft-DTW distance metric was used to identify which group was the most representative for the region. The group whose elements had the lowest average intra-distance was selected as the final result of this process. This means that its elements had more similar behavior based on their activity time series.

4. **Social behavior model generation by applying HMMs**: A social behavioral model was trained for each region using the activity time series of the wineries selected as the most representative in the previous step. The `hmmlearn` Python package ((2010), BSD License), which implements a set of algorithms for unsupervised learning and inference of HMMs, was used to train the activity models.

   The reason why we used HMMs instead of any other approach for marketing analysis is because HMMs provide a unique framework with both predictive and interpretable characteristics. On the one hand, this allowed us to perform a cross-comparison among wineries by checking how well the model from one specific winery recognized the data from any other one. On the other hand, we could perform individual analysis of patterns exploited by each winery model by looking at the states of its HMM through the *Viterbi* algorithm (Forney, 1973).

   Gaussian HMMs are used because they are the most widely used type of HMMs for continuous data, such as the activity time series. Thus, each state of the model will contain a different Gaussian distribution function (i.e., different values of mean and variance), which represent different phases in the evolution of the winery in terms of its Twitter activity.

   Because *the optimal number of states* of each HMM is not known *a priori* in this context, models with different numbers of states were trained in an iterative process and then the one that scored the lowest value in the BIC measure (see Eq. (5)) was selected as the optimal number of hidden states (the lower the BIC score, the better the model).

   Another main problem to generating an HMM is the estimation of its parameters. In this work, this task was performed by means of the *Baum–Welch* algorithm (Baum & Petrie, 1966) (see Eq. (4)) implemented in the `hmmlearn` package, which is an expectation-maximization algorithm that tries to maximize the likelihood of a set of observation sequences. In this particular case, these observation sequences were the time series of activity of the wineries selected for each specific region. Furthermore, to avoid falling into a local optimum when training the HMMs, each training process was run several times using random starting parameters and the result with higher (log-)likelihood was used.

5. **Visualization and analysis of the social behavior models**: The *visualization of the state transition patterns of a HMM* was used to identify different states of activity that each winery region usually has on Twitter, as well as its evolution over time. Given a time series representing the evolution of the social activity of a specific winery and an HMM, the *Viterbi* algorithm, also included in the `hmmlearn` package, can decode the series to its corresponding sequence of states. Therefore, it allowed us to show the evolution of the activity of a particular winery over time through the changes of states produced.

   In this work, to study the different social behaviors across regions, *the most representative winery of a region was visualized* with this method. The criterion used to choose that representative winery was based on its distance to the centroid of the cluster to which it belonged. The shorter that distance, the more representative the winery was considered within its group.

   In addition, to carry out a comparative assessment of social behavior models by winery regions, *a cross-similarity matrix by region* was created and visualized. This allowed our marketing expert to perform an in-depth study of the differences and similarities detected in the social behaviors on Twitter for the different winery regions.

4. Experimental results

This section presents the results obtained in each step of the proposed methodology when applied to Twitter data from wineries belonging to four of the main wine-producing regions in the world. First, the collected dataset is described in detail, as well as how it was preprocessed to obtain the time series of Twitter activity per winery. Then, we show the results obtained in the process of selecting the most representative wineries for each region using time series clustering. In this step of the methodology, a comparative evaluation of clustering algorithms was carried out to choose the most suitable for the given dataset. Finally, the HMM of social behavior for each region was trained and analyzed from a marketing point of view. All processes executed in this experimental phase were run on a PC (Intel Core i7) with a 3.4 GHz CPU, a 15.5 GB memory, and an Ubuntu 16.04 operating system.

4.1. Dataset description

Data gathered for this work were extracted from Twitter (*Step 1 of the proposed methodology*). In particular, information collected from its APIs comprised comments that mentioned the hashtags of wineries located in Porto and Douro Valley (Portugal), Napa Valley (USA), Veneto (Italy), and La Rioja (Spain). Table 1 presents some details of the regions studied, including the number of wineries with any level of Twitter activity, the total number of tweets mentioning the wineries during the considered period (ranging from *10-01-2018* to *03-01-2019*), the average number of mentions, and finally the top three most-mentioned wineries of the region. **The period of data collection is especially interesting because it included one of the most important seasonal periods for the wine industry, the Christmas season**. As shown in this table, the number of wineries considered in La Rioja was higher than the rest. This is a consequence of the larger availability of information associated to this region (taken from the research work in Zarco et al., 2019), whereas wineries from the remaining regions were the result of a manual search on winery rankings on the Internet.

Table 1
Details about the information considered for each of the regions participating in the experiment.

Region name	N. Win	N. Men.	Avg Men.	Top 3 Wineries
Porto and Douro Valley	13	5581	142.15	TaylorsPortWine, Cockburns_Port, grahams_port
Napa Valley	40	7276	150.51	SilverOak, ShaferVineyards, RobertMondavi
Veneto	33	6116	152.26	VillaSandi_it, MrAmaroneMasi, Tommasiwine
La Rioja	193	37,273	171.86	LaRiojaAltaSA, bodegasmuga, marquesderiscal

However, the average number of mentions for each of the four regions was close, which suggests that their level of activity on Twitter was similar.

By preprocessing all tweets extracted from the wineries of the four regions, a *time series of mentions* for each winery was generated (*Step 2 of the methodology*). Each time series represented the sequence of mentions per day (activity level) received by each winery on Twitter during the period fixed for the study, as shown in Fig. 2 for the top three wineries in La Rioja. Because the period of analysis covered 155 days, each time series was composed of 155 values. Because we were interested in the shape of the time series more than in the absolute number of mentions reached by a region, the time series values were normalized to the interval [0, 100].

4.2. Winery selection per region using time series clustering

To select the wineries that would be part of a behavioral model, a double filtering process was carried out (*Step 3 of the methodology*). First, those wineries having a number of mentions lower than the average of total mentions were discarded. Table 2 shows the number of selected wineries of each region after applying this filter.

Second, a comparative evaluation of clustering algorithms was performed to choose the most appropriate one for the data given. For each region, three clustering algorithms (*k*-means, kernel *k*-means, and Kshape) were applied to group wineries by their similar Twitter activity. *K*-means was run using several distance metrics. The three algorithms were also tested with different values for the number of clusters ($k \in \{2, \ldots, N\}$) for each region, with N varying depending on the specific region. To choose the best partition obtained by the three clustering algorithms, the VR score defined in Step 3 was used.

Because of the large number of parameter combinations tested to find the best cluster partition for the given dataset, only the best results from each algorithm and region are shown in Table 3. Analyzing these results, it can be seen that for three of the four regions, the *k*-means algorithm with the Soft-DTW distance metric obtained the best results (higher values in VR). Therefore, this clustering algorithm was used to group the wineries.

After performing the iterative process to choose the value of k, the best number of clusters to group the wineries was 2 for the regions of Porto and Douro Valley and La Rioja, 3 for the region of Veneto, and 4 for the region of Napa Valley (see Table 4).

Once the optimal value of k was tuned, the clustering algorithm was applied to group the wineries according to their activity on Twitter (time series of mentions). Fig. 3 shows the time series of mentions belonging to each cluster for each region. Gray lines represent all the time series belonging to the cluster, whereas the red line shows the one closest to the cluster centroid and thus the most representative one.

By analyzing this figure, we can conclude that there were broadly two different types of Twitter behavior in all regions: one where the time series tended to have more frequent variations in their mentions (cluster 1 on Porto and Douro Valley; clusters 2 and 3 on Napa Valley; cluster 1 on Veneto; and cluster 0 on La Rioja) and another where these variations were less frequent (the rest of the clusters of each region). The time series belonging to the clusters with the most frequent variations tended to have a higher average number of mentions. Therefore, we can classify two different behaviors: one with high and variable Twitter activity and another with medium–low activity and a smoother evolution over time (cluster 0 on Porto and Douro Valley; clusters 0 and 1 on Napa Valley; cluster 0 on Veneto; and cluster 1 on La Rioja).

Finally, to conclude the winery selection process, a single cluster (a group of wineries) needed to be selected as the most representative for each region. For this purpose, the Soft-DTW distance was computed among the elements of the same cluster, and the cluster with the smallest average intra-distance (its wineries having more similar behaviors) was selected. Wineries belonging to these clusters were considered to have the most characteristic collective behavior of the region in question, and therefore these wineries were used to create the behavioral models of the region. Because each activity model represented the general behavior of a region, only those clusters containing more than two wineries were considered at this final stage of the selection process. Table 5 shows the

Table 2
Wineries selected in the first filtering process. Only those wineries with a Twitter activity higher than the average of total mentions were selected for further analysis.

Region name	Total num.	Selected num.
Porto and Douro Valley	13	4
Napa Valley	40	12
Veneto	33	11
La Rioja	193	41

Table 3
Summary of the best results of time series clustering for each winery region and clustering algorithm according to the VR score.

Region name	Algorithm	Distance	SC	CHI	DBI	VR
Porto and Douro Valley	kernel k-means		−0.07	0.71	1.66	0
	k-means	euclidean	0.16	1.94	0.51	2.72
	k-means	dtw	0.14	1.94	0.51	2.66
	k-means	**softdtw**	**0.25**	**1.94**	**0.51**	**3**
	Kshape		0.09	1.60	1.11	1.70
Napa Valley	kernel k-means		0.05	1.21	2.44	0.89
	k-means	euclidean	0.05	2.14	0.63	2.39
	k-means	dtw	0.04	1.93	1.28	1.91
	k-means	softdtw	0.07	1.93	1.28	1.98
	Kshape		**0.29**	**2.10**	**0.62**	**2.93**
Veneto	kernel k-means		0.06	1.12	2.06	1.07
	k-means	euclidean	0.05	1.90	1.15	2.14
	k-means	dtw	0.20	1.96	1.20	2.40
	k-means	**softdtw**	**0.33**	**1.96**	**1.20**	**2.62**
	Kshape		0.16	1.62	0.70	2.34
La Rioja	kernel k-means		−0.18	0.98	1.73	0.97
	k-means	euclidean	0.09	3.18	2.40	2.14
	k-means	dtw	0.19	3.06	3.52	2.05
	k-means	**softdtw**	**0.31**	**3.60**	**3.22**	**2.54**
	Kshape		−0.03	1.90	1.73	1.59

Table 4
Optimal *k* applying k-means with Soft-DTW for each winery region, according to the VR score.

Region name	*k*	SC	CHI	DBI	VR	Time(s)
Porto and Douro Valley	**2**	**0.25**	**1.94**	**0.51**	–	**2.38**
Napa Valley	2	0.13	1.03	2.64	1	8.10
	3	0.10	1.26	2.42	1.05	9.30
	4	**0.12**	**1.65**	**1.77**	**2.24**	**10.12**
	5	0.05	1.45	1.66	1.19	11.26
	6	0.07	1.93	1.28	2.17	11.91
Veneto	2	0.38	1.79	2.00	1.59	8.41
	3	**0.33**	**1.96**	**1.20**	**2.75**	**8.46**
	4	0.28	1.73	1.56	1.47	9.73
	5	0.19	1.55	1.29	0.89	9.71
La Rioja	**2**	**0.31**	**3.60**	**3.22**	**2.48**	**31.99**
	3	0.17	2.51	3.99	0.61	32.88
	4	0.14	2.38	3.70	0.55	34.29
	5	0.15	2.22	3.27	0.74	38.73
	6	0.15	2.10	2.99	0.86	41.92
	7	0.15	2.19	2.61	1.16	44.68
	8	0.10	2.09	2.36	1	47.26

most representative clusters selected for each region.

4.3. Training and selecting best HMM for each region

Using the time series of mentions from the wineries selected in the previous process, the next step was to fit an HMM that represents the collective behavior of each region. For this purpose, an iterative process was first performed to choose the optimal number of states of each model based on the BIC score. The range of the number of states to test was fixed at {2, ..., 20}. Also, because the training of an HMM is a stochastic process that largely depends on the initial parameters of the model, 10 runs of the training process were performed for each number of states tested. The value of the number of states that most frequently achieved the lowest BIC value across the 10 runs was selected as the optimum value for each region.

As shown in Table 6, the optimal number of states usually as the number of wineries used to train the model increased (see the difference between the first three regions with a low number of wineries and the last one with a much higher number). In these cases, it is more likely to find higher diversity in the training data, and thus a complex model with a higher number of states may fit that diversity better, even though the BIC is meant to penalize the complexity of a model. A similar effect can be seen in the computational time for model training, which increases as the complexity of the model increases.

The second to the last column of Table 6 shows the average (log-)likelihood of the best-fitted model for each region, which measures how well the model recognized its training sequences. This was used to score the quality of the models, in the sense that a higher likelihood implies that the behavioral patterns of a region are better captured. Analyzing the results shown in Table 6, it can be

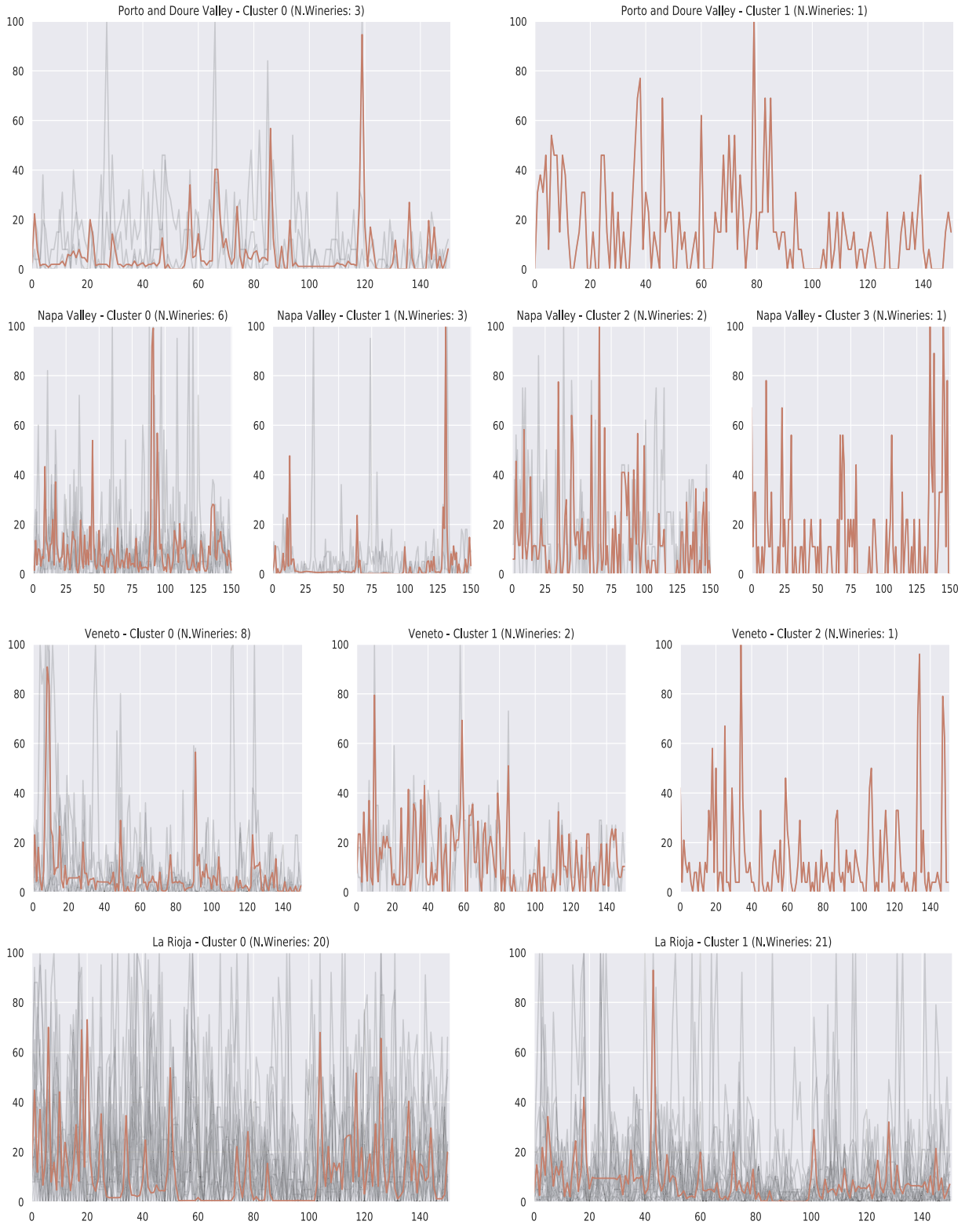

Fig. 3. Time series of Twitter mentions for each of the resulting clusters in the regions of Porto and Douro Valley. Gray lines represent all the time series belonging to the cluster, whereas the red line represents the one closest to the center of the cluster (the centroid). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Clusters selected as more representative by region based on the Soft-TWD measure (marked in bold). In addition, the winery selected as the most representative of the region is marked in italics. Only those clusters with more than two elements were considered to build general models of behavior by region

Region	Clu.	Soft-DTW	Wineries
Porto and Douro Valley	**0**	**9790.80**	***Cockburns_Port*, Quintadelarosa1, TaylorsPortWine**
	1	–	grahams_port
Napa Valley	0	8236.06	JeanEdwardsWine, RobertMondavi, ShaferVineyards, SilverOak, SpottswoodeWine, trefethenfamily
	1	**5658.55**	**GrgichHills, pineridgewine, *stagsleapwines***
	2	–	ClosDuValNapa, TheCastello
	3	–	AOwinery
Veneto	**0**	**9549.96**	**AsoloMontello, C_Valpolicella, RizzardiEstates, SoaveWine, Tommasiwine, VillaSandi_it, VinoLuganaDoc, zonin1821**
	1	–	ProseccoCV, zenatowinery
	2	–	MrAmaroneMasi
La Rioja	0	25282.16	BDavidMoreno, BodegasLC, BodegasLan, BodegasLecea, CampoViejoRioja, CarlosMoroRioja, LaRiojaAltaSA, RamonBilbao, Vina_Ijalba, Vivanco_es, ZuazoGaston, bodegariojavega, bodegasValduero, bodegascorral, bodegasfaustino, bodegasizadi, bodegasmuga, marquesderiscal, ontanonbodegas, torredeona
	1	**13203.43**	**BodegaValdelana, BodegasBaigorri, BodegasBeronia, BodegasCampillo, BodegasValdemar, Bohedal, CondedelosAndes, Cvne, MarquesMurrieta, Marques_Caceres, *MrtnzLacuesta*, PagosdelRey, Pernod_RicardES, RemirezdeGanuza, RiojaBordon, b_gomez_cruzado, bodegasriojanas, bodegastorres, garcia_carrion, maset, vinomontecillo**

Table 6
Optimal number of states for the HMMs trained in each region, based on the minimum BIC score achieved.

Region	N. Win.	N. States	BIC	(Log-)likelihood	*Time(s)*
Porto and Douro	3	9	952.624 ± 109.39	−85.62 ± 0.16	0.32
Napa Valley	3	12	−58.58 ± 85.85	168.62 ± 2.04	0.39
Veneto	8	10	−1606.95 ± 0.64	153.24 ± 0.03	1.16
La Rioja	21	15	3786.93 ± 217.73	−66.44 ± 27.32	18.99

noticed that, in terms of training likelihood, the best models corresponded to the regions of Napa Valley and Veneto, whereas poor models were obtained for Porto and Douro and La Rioja. These results are in accordance with those obtained for the average intra-distance computed among the elements of the most representative cluster for each region (see Table 5). Those wineries having a more similar behavior will have smaller average intra-distances and, as expected, the greater the similarity of the clusters of wineries, the better the HMMs fitted to them.

4.4. Visualization and marketing analysis of the activity evolution on Twitter for the winery regions

4.4.1. Visualization of collective behavioral models

Given a specific time series of mentions of a winery and a particular HMM, it is possible to decode the most likely state sequence that may generate the input data by applying the *Viterbi* algorithm. This means that it is possible to observe the activity evolution of a particular winery over time in terms of changes in the hidden states.

To study the state transition patterns on Twitter activity of the different regions, the sequence of hidden states returned by *Viterbi* algorithm and the actual observation sequences (the time series) were visualized one on top of the other for the most representative winery of each region, which in turn is the closest to the center of the most representative cluster found in the described methodology (see Table 5).

Considering this, Figs. 4–7 show the Twitter activity evolution of the representative wineries of each region. The first conclusion that can be drawn from the analysis of the four figures is that states corresponding to low and high levels of activity are less frequent than those corresponding to a medium level. This means that levels of low and high activity are easier to detect than those of medium level that have a greater variation of mentions. Therefore, a greater number of states is required to fine-tune its model.

However, studying the sojourn time and frequency of states, it can be observed that states modeling low levels of activity usually last longer and happen more frequently (e.g., states 3 and 7 of Porto and Douro Valley; 5 and 9 of Napa Valley; 0 and 2 of Veneto; and 2 and 6 of La Rioja). However, as the level of activity increases, frequency and sojourn time of the corresponding hidden states decreases (states 1 and 4 of Porto and Douro Valley; 1 and 2 of Napa Valley; 6 and 7 of Veneto; and 1 and 4 of La Rioja). Furthermore, for medium and high levels of activity, it seems that each peak of activity is modeled by its single state, which may suggest that the model is overfitted.

4.4.2. Cross-comparison of likelihood among regions

The second part of this study focused on carrying out an analysis of similarities or differences in social behavior among the different winery regions. For this purpose, a matrix indicating the cross-likelihood among HMMs and time series from each region

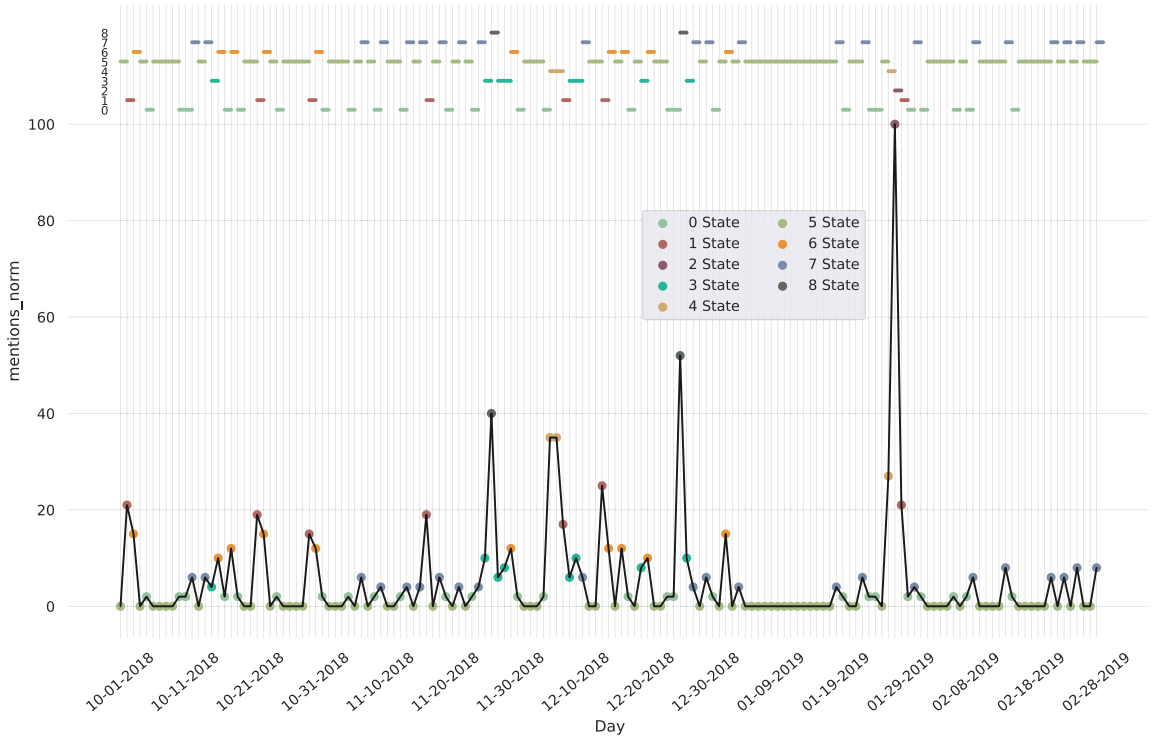

Fig. 4. Visualization of the evolution of Twitter activity for the most representative winery of Porto and Douro Valley (Cockburns_Port).

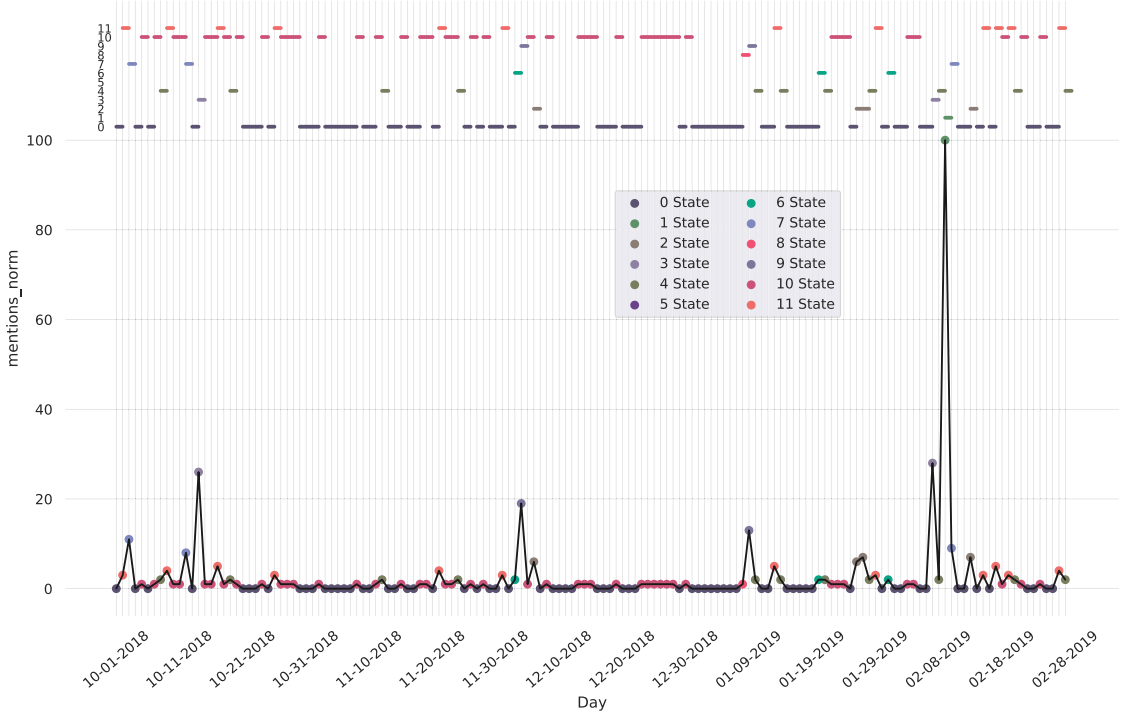

Fig. 5. Visualization of the evolution of Twitter activity for the most representative winery of Napa Valley (stagsleapwines).

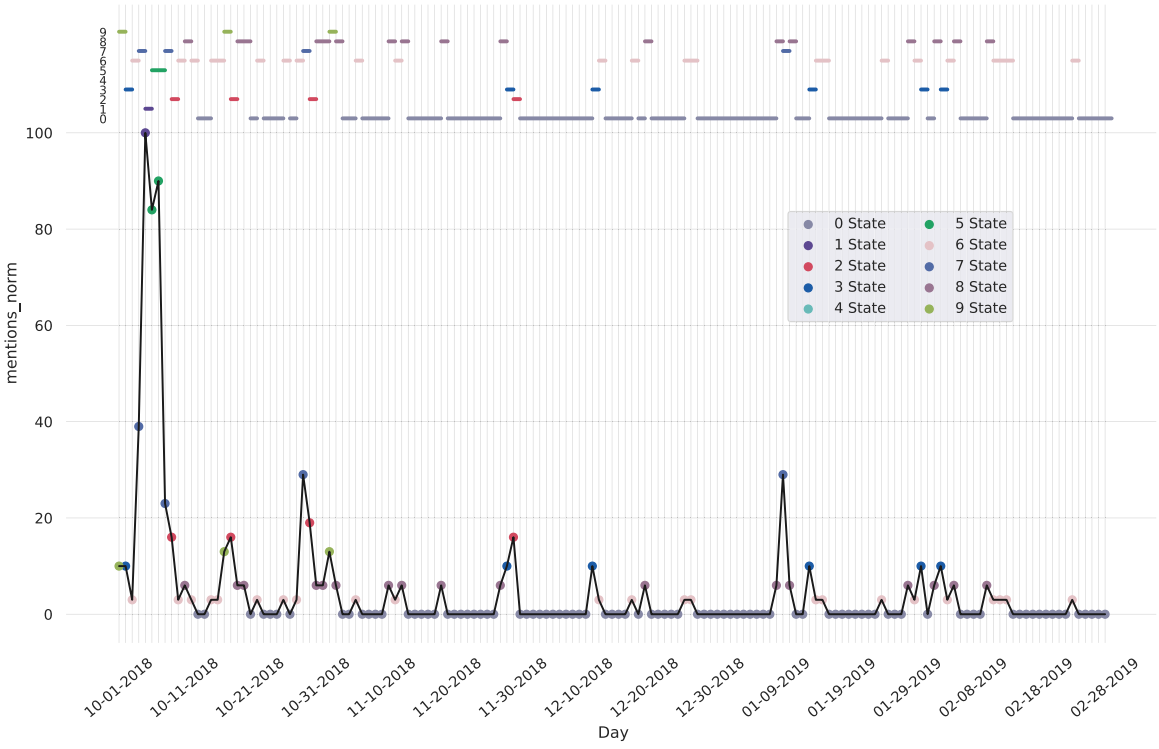

Fig. 6. Visualization of the evolution of Twitter activity for the most representative winery of Veneto (VinoLuganaDoc).

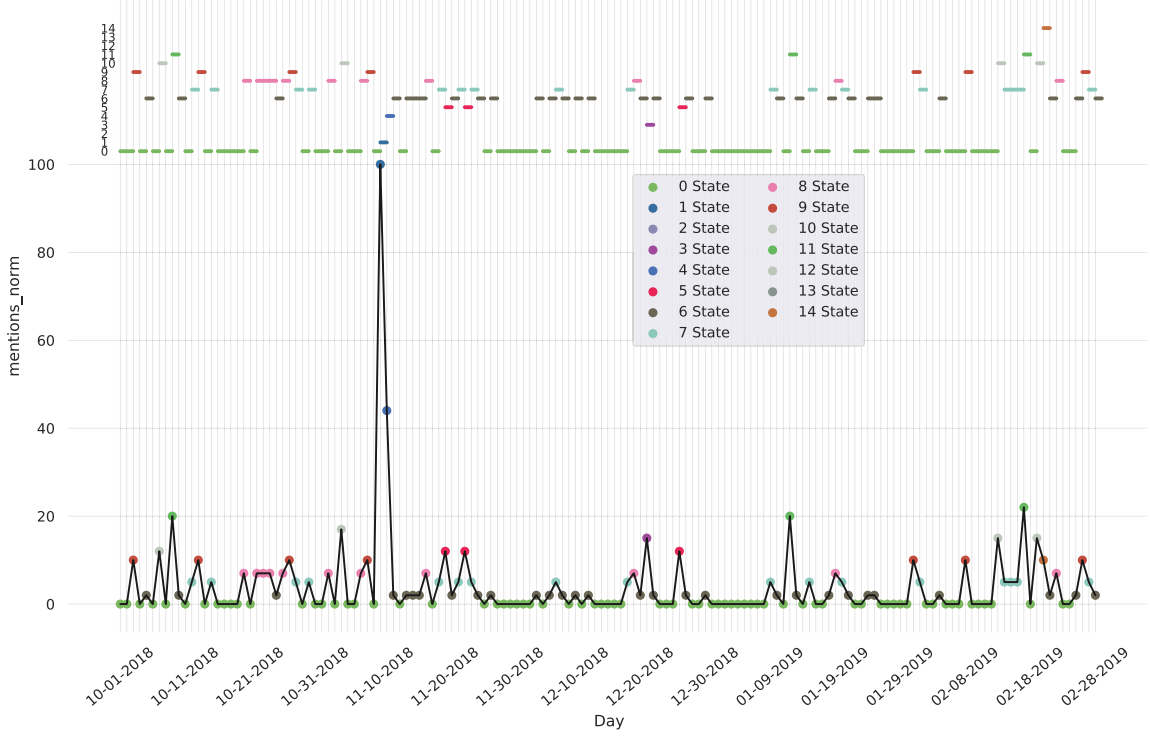

Fig. 7. Visualization of the evolution of Twitter activity for the most representative winery of La Rioja (MrtnzLacuesta).

Table 7
Cross-likelihood matrix by winery region based on the (log-)likelihood between the HMM (rows) and time series training data (columns).

	Porto and Douro	Napa	Veneto	La Rioja
Porto and Douro HMM	−**85.62**	−339.32	−319.95	−440.26
Napa HMM	−1341.45	**163.98**	−1112.93	−1177.79
Veneto HMM	−201.45	−62.78	**153.17**	−180.60
La Rioja HMM	−105.48	**85.94**	58.58	−57.80

was created. Formally, the (log-)likelihood can be expressed as $\log P(o|\lambda)$, where o is an observation sequence and λ is an HMM. In the context of this experiment, the (log-)likelihood was applied as a non-symmetric distance between the regions, calculated by averaging the probability that a time series from a specific region was generated by the model of a different one. Thus, this score can be used to measure how well the model of one region recognizes sequences of different regions and thus how similar the collective behaviors of wineries in the different regions are. Table 7 shows the (log-)likelihood values obtained. The same results are visualized as a heat map in Fig. 8.

As expected, all models recognized their sequences better than the rest except in the case of the region of La Rioja (see the values in bold font in Table 7, corresponding to the maximum likelihood values), which may mean that wineries comprising that region do not have a proper identity in terms of social activity on Twitter. In addition, the row that represents how well the model of La Rioja recognizes the rest of the regions (fourth row) shows high values of likelihood in all cases. These facts indicate that this model is too generic and does not have much differential identity from the rest.

Just the opposite effect happened for the region of Napa Valley (second row of the matrix). This model achieved high values of likelihood with itself and very low values against the remainder. Therefore, it can be concluded that this region has more differentiated activity patterns on Twitter than the rest. Thus, it is the region with a greater identity. In case of the Porto and Douro Valley and Veneto (second row of the matrix), its model exhibited intermediate likelihood values when recognizing other regions.

Taking all these results into account, it could be concluded that the most discriminant model is Napa Valley's HMM, showing more differentiated social collective behavior of the users concerning the wineries of the region on Twitter.

Finally, we compared the collective behaviors of the different regions by following a simple computation over the (log-)likelihood values obtained by the region HMMs in Table 7. Those values correspond to a non-symmetric distance between the regions, calculated by averaging the probability that a time series of a specific region has been generated by the model of a different one. A symmetric similarity metric $Norm - Sim \in [0, 1]$ was defined for each pair of HMMs by averaging the pairwise values and normalizing them as follows:

$$N - Sim(HMM_i, HMM_j) = \frac{Sim(HMM_i, HMM_j) - \min_{k,l} Sim(HMM_k, HMM_l)}{\max_{k,l} Sim(HMM_k, HMM_l) - \min_{k,l} Sim(HMM_k, HMM_l)} \quad (12)$$

where $Sim(HMM_i, HMM_j) = \frac{Cross-likelihood(i,j) + Cross-likelihood(j,i)}{2}$, and $\max_{k,l} Sim(HMM_k, HMM_l)$ and $\min_{k,l} Sim(HMM_k, HMM_l)$ are respectively the maximum and minimum values of non-normalized similarity Sim along all the HMMs.

This similarity metric allowed us to directly compare the temporal collective behaviors of the different winery regions pairwise. Given its values, reported in Table 8, the most similar regions according to their activity patterns were Veneto and La Rioja, followed by Porto and Douro and Veneto. As expected, the Napa Valley region showed less similar behavior to the rest of the regions, represented by the lowest values in the table.

4.4.3. Marketing analysis of the activity evolution on Twitter

A relevant aspect from the viewpoint of marketing would be to identify if the causes of the wineries' peaks of activity are related to some marketing actions carried out by the wineries (i.e., attribution models). To check this, each of the highest peaks of activity

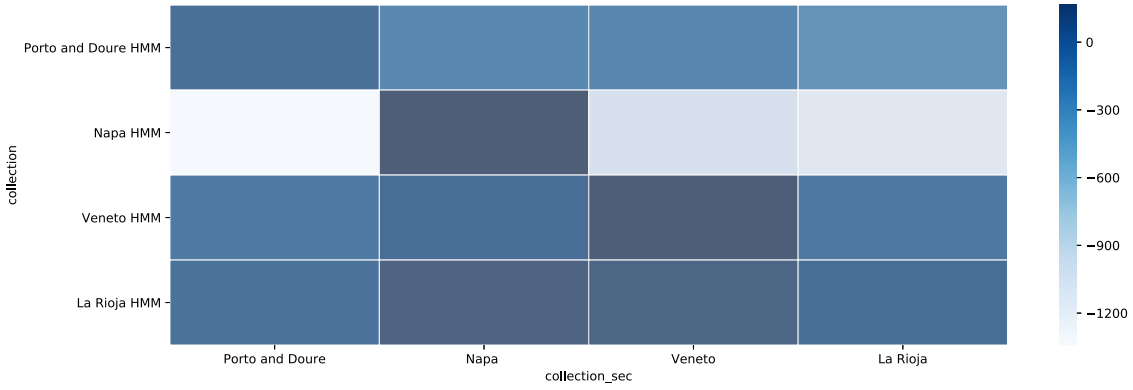

Fig. 8. Heat-map of cross-likelihood between regions.

Table 8
Similarity matrix by winery region computed from the cross-likelihood values of the HMM.

	Napa HMM	Veneto HMM	La Rioja HMM
Porto and Douro HMM	0.00	0.74	0.73
Napa HMM	—	0.33	0.38
Veneto HMM	—	—	**1**

related to the four selected wineries was analyzed in detail to identify its possible cause, as shown below.

1. *Cockburns_Port* account (Porto Douro Valley): This winerys highest peak corresponds to the date 01-28-2019. In the timeline of its account, it can be seen that the day before this, the winery promoted a giveaway for the International Port Wine Day (text of the tweet: "International Port Wine Day GIVEAWAY How to participate: Like this post Follow our page @Cockburns_Port Tag a friend in the comments Repeat as many times as you like. Each winner will win one bottle. Five winners will be selected randomly. Best of luck!"). In addition, studying the evolution of its Twitter activity, it can be observed that this was the winery with the highest level of activity during all the period analyzed in the four regions. As Fig. 4 shows, specifically in the period between 11-20-2018 to 12-30-2018, the level of mentions was significantly higher than the average. This period corresponds to the Christmas campaign that the winery carried out to promote its wines. In its Twitter timeline, it can be seen how on 11-22-2018, a Christmas campaign was started by publishing tweets with references to Christmas and mentioning its wines, and the campaign remained active until the end of Christmas. An example of these tweets is "Christmas is almost a month away, which means it's time to discover magic in every corner. Christmas means beautiful lights and decorations! #cockburns #cockburnsport #specialreserve #myport #xmas #christmas #xmastim."
2. *stagsleapwines* account (Napa Valley): For this winery, the day 02-09-2019 had its highest peak of mentions. Its timeline on Twitter during the studied period shows that this winery generally publishes very few tweets. Even on the days of highest activity, it published nothing. As shown in Fig. 5, its overall activity level of mentions is also usually quite low. The peak of activity was caused by a tweet written by a user who usually publishes wine tasting notes. Specifically the day 02-07-2019, this user published a note about the winery (text of the tweet: "#HumpDayTreat continues w/ a gorgeous limited edition @stagsleapwines #Chard Tasting notes https://www.instagram.com/p/Btj7AtHH084 #WineWednesday #wineoclock...").
3. *VinoLuganaDoc* account (Veneto): This winery usually publishes tweets frequently in its timeline. However, during the days closer to its greater activity, no tweets were published (see Fig. 6). In this case, the reason for the increase in activity was due to The 2018 Wine Bloggers Conference (#wbc18) W. M. Conference (2018) that took place in October 4–7 in Walla Walla, Washington, and where users who attended mentioned the winery on Twitter (text example of tweet: "@VinoLuganaDoc excited to taste Lugana wines! @WineBloggersCon #wbc18 #vino https://t.co/CW6M7LFTAb ").
4. *MrtnzLacuesta* account (La Rioja): The timeline of this winery has a rather low frequency of tweet publications. Nevertheless, Fig. 7 shows that it has a medium level of mentions. In this case, the peak of activity produced on the day 11-10-2018 was caused by the holding of a meeting of Spanish political formation in the winery. This meeting was mentioned by the winery in its particular Twitter account (translation of the tweet text: "Today we are having the I Lunch and Coexistence of @cslarioja in the winery @MrtnzLacuesta. We tell you now..."[1]).

Taking into account the detailed analysis carried out on each of the representative wineries of each region, one can conclude that only the wineries from the region of Porto and Douro Valley carried out campaigns and marketing strategies using Twitter. In addition, a conclusion can also be drawn that these marketing actions affected their activity level and media impact within the SN.

The regions of Veneto and La Rioja were not applying marketing strategies through their Twitter accounts during the analyzed period. However, they had an average level of activity on the SN that tended to increase as other relevant users on Twitter mentioned them when participating in common events together (conference, meetings, etc...).

In contrast, the region of Napa Valley showed a rather low level of mentions, being the region with the most differentiated behavioral patterns on activity as compared to the other regions. This fact is in accordance with the results presented in the previous section that showed the Napa Valley's HMM as the model showing the most differential social behavior.

5. Conclusions and discussion

This work shows practical application of unsupervised ML techniques to extract, model, and analyze collective behavior on the Twitter activity of four wine-producing regions with important wineries at the international scale. The collective social behavior was defined from the time series of mentions to each company's Twitter account, reflecting the responses of users to both the brand's and other users' actions. **The methodology proposed for this aim generated a model of general behavior for each region, allowing us to comparatively analyze them from a marketing perspective**. The considered dataset was composed of tweets mentioning the main wineries in each region over a 5-month period. Using this dataset, the proposed methodology was applied to, on the one hand, identify the most representative wineries by region through the use of time series clustering and, on the other hand, create HMMs

[1] original text of the tweet:"Hoy estamos teniendo la I Comida y convivencia de @cslarioja en la bodega @MrtnzLacuesta. Os lo contamos ahora..."

based on these selected wineries that capture the behavioral patterns on the social activity of each particular region.

Regarding the results obtained by applying the clustering algorithms to group the wineries, there were broadly two different types of Twitter collective behaviors according to the activity (based on the number of mentions) in all regions. One was related to a high and variable Twitter activity, whereas the other involved medium–low activity and more constant evolution. However, analyzing the quality of the identified groups of wineries, it was concluded that although Napa Valley featured stronger differences among their wineries, the wine companies in Veneto had the weakest distinctions.

Concerning the results obtained for the HMMs trained in each region, the best models, in terms of training likelihood, corresponded to the regions of Napa Valley and Veneto, whereas the worst models were obtained for Porto and Douro and La Rioja. These results were in accordance with those obtained for the most representative clusters found for each region, i.e., **the greater the similarity of the group of wineries is (wineries having a smallest average intra-distance between them), the better the HMMs fitted to them**.

Visualization of the activity evolution of the most representative wineries of each region in terms of changes on the hidden states of their HMMs allowed us to study the state transition patterns on their Twitter activity. It was observed that **states modeling low levels of activity usually lasted longer and occurred more frequently, whereas as the level of activity increased, the frequency and sojourn time of the corresponding hidden states decreased**.

The final marketing analysis on the evolution of the activity of the winery regions on Twitter drew the conclusion that only the wineries from the region of Porto and Douro Valley carried out distinct campaigns and marketing strategies using Twitter. In contrast, the rest of the regions (Napa Valley, Veneto, and La Rioja) did not apply marketing strategies through their Twitter accounts during the analyzed period. In addition, the region of Napa Valley had a rather low activity level of mentions, being the region with the most differentiated behavioral patterns on activity from the rest.

Taking into account all the experimental results presented, it can be concluded that the application of unsupervised ML techniques is useful for this type of analysis. This combination of techniques provides winery companies with new collective knowledge that can be very valuable. Therefore, it can be considered that the proposed methodology is a good technological watch tool for the companies, allowing them to check users' behavior regarding mentions of the wineries' Twitter accounts.

Regarding future work, it may be interesting to study how to achieve simpler and more general models that get fair values of likelihood without the need for a high number of states. This goal could be achieved by considering additional measures to the model selection process to encourage model interpretability, such as the number of rare states, to avoid model states with very low frequencies of occurrence. In addition, a possible future line of research of this work could include new features for modeling users' behavior with respect to wineries on social media, such as sentiment analysis of tweets mentioning wineries or a score that measures the Google searches targeting them (using Google Trends API).

Acknowledgments

This work has been co-funded by the following grants: Spanish Agencia Estatal de Investigación (AEI) and European Regional Development Funds (EDRF) under grants TIN2017-85727-C4-3-P (DeepBio) and PGC2018-101216-B-I00 (EXASOCO) and by Comunidad Autonóma de Madrid under grant P2018/ TCS-4566 (CYNAMON).

References

Alamaki, A., Pesonen, J., & Dirin, A. (2019). Triggering effects of mobile video marketing in nature tourism: Media richness perspective. *Information Processing & Management, 56*(3), 756–770. https://doi.org/10.1016/j.ipm.2019.01.003.

Albalate, A., & Minker, W. (2011). *Semi-supervised and unsupervised machine learning: novel strategies.* Wiley-ISTE.

Alonso, A., Bressan, A., O'Shea, M., & Krajsic, V. (2013). Website and social media usage: Developments of wine tourison, hospitality and the wine sector. *Tourism Planning and Development, 25*(3), 229–248.

Araniti, G., Orsino, A., Militano, L., Wang, L., & Iera, A. (2016). Context-aware information diffusion for alerting messages in 5g mobile social networks. *IEEE Internet of Things Journal, 4*(2), 427–436.

Asur, S., Huberman, B., et al. (2010). *Predicting the future with social media. Web intelligence and intelligent agent technology (WI-IAT), 2010 IEEE/WIC/ACM international conference onVol. 1. Web intelligence and intelligent agent technology (WI-IAT), 2010 IEEE/WIC/ACM international conference on* IEEE492–499.

Atzori, L., Iera, A., Morabito, G., & Nitti, M. (2012). The social internet of things (SIOT)–when social networks meet the internet of things: Concept, architecture and network characterization. *Computer Networks, 56*(16), 3594–3608.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics,* 1554–1563.

Bello-Orgaz, G., Hernandez-Castro, J., & Camacho, D. (2016). Detecting discussion communities on vaccination in Twitter. *Future Generation Computer Systems, 66*, 125–136.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59.

Bello-Orgaz, G., Menéndez, H., Okazaki, S., & Camacho, D. (2014). Combining social-based data mining techniques to extract collective trends from twitter. *Malaysian Journal of Computer Science, 27*(2), 95–111.

Bruckhaus, T. (2010). *Collective intelligence in marketing. Marketing intelligent systems using soft computing.* Springer131–154.

Bruwer, J., & Wood, G. (2005). The australian online wine buying consumer: motivation and behaviour perspectives. *Journal of Wine Research, 6*(3), 193–211.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304.

Caliski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1–27. https://doi.org/10.1080/03610927408827101.

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems, 31*(2), 102–107. https://doi.org/10.1109/MIS.2016.31.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). *A practical guide to sentiment analysis.* Cham, Switzerland: Springer.

Cambria, E., Grassi, M., Hussain, A., & Havasi, C. (2012). Sentic computing for social media marketing. *Multimedia Tools and Applications, 59*(2), 557–577. https://doi.org/10.1007/s11042-011-0815-0.

Castro Galiana, R. Las bodegas, los vinos y las políticas de comunicación. http://castrogaliana.com/las-bodegas-los-vinos-y-las-politicas-de-comunicacion/.

Cavallari, S., Zheng, V. W., Cai, H., Chang, K. C.-C., & Cambria, E. (2017). *Learning community embedding with community detection and node embedding on graphs. Proceedings of the 2017 ACM on conference on information and knowledge management.* ACM377–386.

Chen, W., Wang, C., & Wang, Y. (2010). *Scalable influence maximization for prevalent viral marketing in large-scale social networks. Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining.* ACM1029–1038.
Collier, N. (2012). Uncovering text mining: A survey of current work on web-based epidemic intelligence. *Global Public Health, 7*(7), 731–749.
Cuturi, M. (2011). *Fast global alignment kernels. Proceedings of the 28th international conference on international conference on machine learningICML'11*USA: Omnipress929–936.
Cuturi, M., & Blondel, M. (2017). *Soft-DTW: A differentiable loss function for time-series. Proceedings of the 34th international conference on machine learning-volume 70.* JMLR. org894–903.
Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2),* 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.
De Choudhury, M., Diakopoulos, N., & Naaman, M. (2012). *Unfolding the event landscape on Twitter: Classification and exploration of user categories. Proceedings of the ACM 2012 conference on computer supported cooperative workCSCW '12*New York, NY, USA: ACMhttps://doi.org/10.1145/2145204.2145242 241–244
De Choudhury, M., Mason, W. A., Hofman, J. M., & Watts, D. J. (2010). *Inferring relevant social networks from interpersonal communication. Proceedings of the 19th international conference on World Wide WebWWW '10*New York, NY, USA: ACM301–310. https://doi.org/10.1145/1772690.1772722.
De Choudhury, M., Sundaram, H., John, A., & Seligmann, D. D. (2009). *What makes conversations interesting?: Themes, participants and consequences of conversations in online social media. Proceedings of the 18th international conference on World Wide WebWWW '09*New York, NY, USA: ACM331–340. https://doi.org/10.1145/1526709.1526754.
Dhillon, I. S., Guan, Y., & Kulis, B. (2004). *Kernel k-means: Spectral clustering and normalized cuts. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining.* ACM551–556.
Dolan, R., Conduit, J., Fahy, J., & Goodman, S. (2016). *Facebook for wine brands: An analysis of strategies for Facebook posts and user engagement actions. Proceedings of the 9th academy of wine business research conference17. Proceedings of the 9th academy of wine business research conference* 457–465.
Fang, Y., Lin, W., Zheng, V. W., Wu, M., Chang, K. C.-C., & Li, X.-L. (2016). *Semantic proximity search on graphs with metagraph-based learning. 2016 IEEE 32nd International conference on data engineering (ICDE).* IEEE277–288.
Fang, E. The 5G revolution: Why the next generation of mobile internet will force advertisers and marketers to change the way they think. Digital Marketing Magazinehttps://digitalmarketingmagazine.co.uk/articles/the-5g-revolution-why-the-next-generation-of-mobile-internet-will-force-advertisers-and-marketers-to-change-the-way-they-think/5072.
Ferguson, R. (2008). Word of mouth and viral marketing: taking the temperature of the hottest trends in marketing. *Journal of Consumer Marketing, 25*(3), 179–182.
Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE, 61*(3), 268–278.
Foster, M., Francescucci, A., & West, B. (2010). Why users participate in online social networks. *International Journal of e-Business Management, 4*(1), 3–19.
Fuentes Fernández, R., Vriesekoop, R., & Urbano, B. (2017). Social media as a means to access millennial wine consumers. *International Journal of Wine Business Research, 29*(3), 269–284.
Ghosh, G., Banerjee, S., & Yen, N. Y. (2016). State transition in communication under social network: An analysis using fuzzy logic and density based clustering towards big data paradigm. *Future Generation Computer Systems, 65*, 207–220. https://doi.org/10.1016/j.future.2016.02.017 Special Issue on Big Data in the Cloud
Guan, P., & Wu, J. (2019). Effective data communication based on social community in social opportunistic networks. *IEEE Access, 7*, 12405–12414.
Hassan, M. R., & Nath, B. (2005). *Stock market forecasting using hidden Markov model: A new approach. 5th International conference on intelligent systems design and applications (ISDA'05).* IEEE192–196.
hmmlearn developers (BSD License) (2010). hmmlearn: Unsupervised learning and inference of hidden Markov models. https://github.com/hmmlearn/hmmlearn.
Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR), 31*(3), 264–323.
Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology, 60*(11), 2169–2188.
Javed, A., Burnap, P., & Rana, O. (2019). Prediction of drive-by download attacks on Twitter. *Information Processing & Management, 56*(3), 1133–1145. https://doi.org/10.1016/j.ipm.2018.02.003.
Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining.* John Wiley & Sons.
Liao, T. W. (2005). Clustering of time series data–a survey. *Pattern Recognition, 38*(11), 1857–1874.
Liu, Z., Zheng, V. W., Zhao, Z., Zhu, F., Chang, K. C.-C., Wu, M., & Ying, J. (2017). *Semantic proximity search on heterogeneous graph by proximity embedding. Thirty-first AAAI conference on artificial intelligence*154–160.
Macqueen, J. B. (1967). *Some methods of classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*281–297.
Mariani, A., Pomarici, E., & Boatto, V. (2012). The international wine trade: Recent trends and critical issues. *Wine Economics and Policy, 1*(1), 24–40.
Maurer, C., & Wiegmann, R. (2011). Effectiveness of advertising on social network sites: A case study on Facebook. In R. Law, M. Fuchs, & F. Ricci (Eds.). *Information and communication technologies in tourism 2011* (pp. 485–498). Vienna: Springer Vienna.
Osotsi, A. J. (2016). Event detection in Twitter data: A hidden Markov model-based change point algorithm. Master's thesisPennsylvania (USA) Penn State University.
Owyang, J. The future of the social web: In five eras. https://web-strategist.com/blog/.
Paparrizos, J., & Gravano, L. (2015). *k-shape: Efficient and accurate clustering of time series. Proceedings of the 2015 ACM SIGMOD international conference on management of data.* ACM1855–1870.
Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.
Rodríguez-Fernández, V., Gonzalez-Pardo, A., & Camacho, D. (2018). Modelling behaviour in UAV operations using higher order double chain Markov models. *IEEE Computational Intelligence Magazine, 12*(4), 28–37. https://doi.org/10.1109/MCI.2017.2742738.
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.
Rybski, D., Buldyrev, S. V., Havlin, S., Liljeros, F., & Makse, H. A. (2009). Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences U.S.A. 106*, 12640–12645.
Rybski, D., Buldyrev, S. V., Havlin, S., Liljeros, F., & Makse, H. A. (2012). Communication activity in a social network: relation between long-term correlations and inter-event clustering. *Scientific Reports, 2:560*, 1–6.
Sakoe, H., Chiba, S., Waibel, A., & Lee, K. (1990). Dynamic programming algorithm optimization for spoken word recognition. *Readings in Speech Recognition, 159*, 224.
Singh, S., Saxena, N., Roy, A., & Kim, H. (2017). A survey on 5g network technologies from social perspective. *IETE Technical Review, 34*(1), 30–39.
Stelzner, M. (2014). 2014 Social media marketing industry report. Available at: https://www.socialmediaexaminer.com/social-media-marketing-industry-report-2014/.
Szolnoki, G., Dolan, R., Forbes, S., Thach, L., & Goodman, S. (2018). Using social media for consumer interaction: An international comparison of winery adoption and activity. *Wine Economics and Policy, 7*(2), 109–119.
Tavenard, R., Faouzi, J., & Vandewiele, G. (2017). tslearn: A machine learning toolkit dedicated to time-series data. https://github.com/rtavenar/tslearn.
van Esch, P. (2020). *Disruptive technologies impact on digital & social media marketing: A new frontier. Australasian marketing journal* Special issue, in preparation
Vinciarelli, A., & Favre, S. (2007). *Broadcast news story segmentation using social network analysis and hidden Markov models. Proceedings of the 15th ACM international conference on multimedia.* ACM261–264.
Vinography Social media and the wine industry: A new era. http://www.vinography.com/archives/2012/02/social_media_and_the_wine_indu.html.
Visser, I. (2011). Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. *Journal of Mathematical Psychology, 55*(6), 403–415.
W. M. Conference (2018) The 2018 wine bloggers conference. https://www.winemediaconference.org/tag/2018/.
Wilson, D., & Quinton, S. (2012). Let's talk about wine: does twitter have value? *International Journal of Wine Business Research, 24*(4), 271–286.
Wu, J., Chen, Z., & Zhao, M. (Chen, Zhao, 2019a). Information cache management and data transmission algorithm in opportunistic social networks. *Wireless Networks,*

25(6), 2977–2988.
Wu, J., Chen, Z., & Zhao, M. (Chen, Zhao, 2019b). Weight distribution and community reconstitution based on communities communications in social opportunistic networks. *Peer-to-Peer Networking and Applications, 12*(1), 158–166.
Zarco, C., Santos, E., & Cordón, O. (2019). Advanced visualization of twitter data for its analysis as a communication channel in traditional companies. *Progress in Artificial Intelligence, 8*(3), 307–323.