Neural Computing and Applications (2022) 34:16717–16738
https://doi.org/10.1007/s00521-022-07307-0

ORIGINAL ARTICLE

Neuro-semantic prediction of user decisions to contribute content to online social networks

Pablo Cleveland[1] · Sebastian A. Rios[2] · Felipe Aguilera[1] · Manuel Graña[3]

Received: 19 February 2021 / Accepted: 13 April 2022 / Published online: 22 June 2022

Abstract

Understanding at microscopic level the generation of contents in an online social network (OSN) is highly desirable for an improved management of the OSN and the prevention of undesirable phenomena, such as online harassment. Content generation, i.e., the decision to post a contributed content in the OSN, can be modeled by neurophysiological approaches on the basis of unbiased semantic analysis of the contents already published in the OSN. This paper proposes a neuro-semantic model composed of (1) an extended leaky competing accumulator (ELCA) as the neural architecture implementing the user concurrent decision process to generate content in a conversation thread of a virtual community of practice, and (2) a semantic modeling based on the topic analysis carried out by a latent Dirichlet allocation (LDA) of both users and conversation threads. We use the similarity between the user and thread semantic representations to built up the model of the interest of the user in the thread contents as the stimulus to contribute content in the thread. The semantic interest of users in discussion threads are the external inputs for the ELCA, i.e., the external value assigned to each choice.. We demonstrate the approach on a dataset extracted from a real life web forum devoted to fans of tinkering with musical instruments and related devices. The neuro-semantic model achieves high performance predicting the content posting decisions (average *F* score 0.61) improving greatly over well known machine learning approaches, namely random forest and support vector machines (average *F* scores 0.19 and 0.21).

Keywords Multi-topic text preferences · Information diffusion · Leaky competing accumulator · Microscopic model of social interaction · Social interaction decision making

1 Introduction

There is a huge amount of research literature dealing with diverse aspects of the analysis of on-line social networks (OSN). Classical research efforts are devoted to identify communities within the network [14, 39, 64], finding influencers or key members of the virtual community [6, 10, 29, 31, 34, 47, 63, 72], or describing the evolution of specific networks [25, 54, 62]. There is, however, very little or no work on the actual decision process conducting a user to publish some content in the OSN, e.g., posting a message in a forum of a virtual community of practice (VCoP). A VCoP implemented as an internet web-based Forum is a virtual place where members interact, discuss ideas, share, and generate knowledge about specific topics organized into sub-forums and discussion threads. Content generation is a radically different process from the propagation effects across the OSN that follow the publication of some new content. For instance, publishing a tweet is radically different from retweeting, sharing, liking, or any other propagation process that spreads the influence of the original tweet content. Synthetic content generation, such as *n*-gram Markov models allowing to generate fake tweets that are difficult to distinguish by humans [66], are out of the scope of the paper.

✉ Manuel Graña
mgrana@ehu.es

Pablo Cleveland
pcleveland@ceine.cl

[1] Business Intelligence Research Center, Universidad de Chile, Beauchef 851, P.O. Box 8370459, Santiago, Chile

[2] Industrial Engineering Department, Universidad de Chile, Beauchef 851, Santiago, Chile

[3] Computational Intelligence Group, University of the Basque Country, San Sebastian, Spain

The decision to contribute a post to a discussion thread of a VCoP is a phenomenon affected by multiple factors like the user's knowledge of the subject, his preferences, other users participating of the discussion, and even the quality of the information presented, among other factors. This decision process can be modeled by the competition of several simultaneously on-going threads to win the attention of the user, i.e., the user selects the winning thread for publishing a contribution. This competition is modeled by a neurophysiological model of choice, the leaky competition accumulator (LCA) [9, 76, 77], where the computational neurons activity is driven by a set of linear differential equations that accumulate inhibitive contributions from other neurons, excitatory input units, and fluctuations from and independent white noise source. LCA has been shown to account successfully for reaction time distribution empirically observed in psychophysical experiments. Specifically, for some combinations of parameter inhibition and decay values, LCA has been shown to reproduce the empirically observed violations of expected value and preference reversals reported in many experiments on value-based preferential choice. These studies focus on the distribution of the decision time for a fixed error ratio after many repetitions of the LCA run trying to mimic the distributions found empirically. LCA parameters are hand tuned (or explored in a grid search) in order to find the values that reproduce the desired response time behavior and the expected choice error ratio understood as choosing the lowest value option. Our work is more akin to machine learning approaches to model the decision process, i.e. we use LCA as decision making model whose performance is measured by the prediction accuracy of the decision made by the users to post a content contribution to a specific conversation thread where the semantic value assigned to the conversation thread is treated as a constant input.

For our specific work, we propose an extended LCA (ELCA) model in several aspects. First, the model includes many simultaneous choices by many users, while classical LCA considers a single agent and a small number of choices. Secondly, we use the semantic modeling of users and threads to compose the input value of each choice, thus linking the abstract valuation of the choices to concrete domain related evidences. Thirdly, we implement a genetic algorithm search for the ELCA model parameter calibration (aka training) using data from the content contribution decisions in a real life VCoP. The recovery of LCA parameters, stated as the induction of model parameters from simulation accumulator trajectories, has been acknowledged as an open difficult problem [49], which has been tackled by exploitation of Lie symmetries for a modified formulation of LCA equations [45]. Contrary to these approaches, we look for the optimal ELCA parameters that reproduce the actual user decisions after convergence of the simulation. However, our work does not try study or reproduce human choice phenomena, such as preference reversal, that are the original domain of study of the LCA model [9, 76, 77].

Semantic analysis of OSN published content is a current hot research area that allows to detect and prevent undesirable uses of the OSN. For instance, the semantic analysis at word level has been reported to allow to detect cyberbullying [30], helps detecting drunken tweets [24], and the age of users [56]. Also, social media posts content analysis allows to predict depression levels [2]. Specifically, we use unsupervised latent Dirichlet allocation (LDA) [8] topic analysis for the semantic modeling of the OSN published content, that allows to build up quantitative vectorial semantic representations of both users and conversation threads, not much unlike the social semantics neurobiological model based on conceptual knowledge [7]. LDA is a powerful tool that has been used to summarize and build network models of contents, such as semantic graphs relating publications about COVID-19 [1].

Paper contributions and contents This paper proposes a neuro-semantic model of the decisions made by the users to contribute contents to a VCoP web forum at the microscopic level. Specific contributions of this work are:

- The semantic characterization of the messages posted in the VCoP web forum is extracted by unsupervised formal topic analysis, namely LDA, allowing the semantic modeling of both users and conversation threads, so that user interest in generating content for a conversation thread can be quantified and assigned as an input value for the neurophysiological model of choice making, namely LCA.
- Ancillary information identifying key members of the social network provided by the online social network (OSN) administrators is used for the stratification of users improving the detail of the model of the content generation decision process.
- An extended LCA neurophysiological model of the user individual decision process to generate and contribute content in three ways: (1) use of semantically grounded value of the various choices, (2) the consideration of many choices and decision agents in a concurrent dynamic process, and (3) the estimation of the model parameters by maximizing prediction accuracy carried out by a genetic algorithm search. to the OSN that uses as input the semantic characterization of the users and the conversation threads.
- Prediction accuracy is based on a graph representation of the user contributions as a bipartite graph where nodes are either users or conversation threads, and edges correspond to the publication of a post by a user

in a thread. Prediction performance measures are based on the distance between the ground truth graph extracted from the dataset and the predicted graph measured in terms of shared edges.

The paper is organized as follows: Sect. 2 presents related works on OSN information diffusion. Section 3 describes the materials and methods, including the description of the dataset, the semantic modeling, and the proposed neuro-semantic model for user content publication decisions. Section 4 reports the details and results of the computational experiments conducted. Finally, Sect. 5 gives our conclusions and future work directions.

2 Related works

A great deal of the literature on OSN dynamic analysis has been focused the propagation of information across the network and the detection of communities and key influencer users. Table 1 gives a non-exhaustive summary of works found in the literature since 2007. There are two main research lines on models of information diffusion in networks [42], namely the explanatory and the predictive models. The first line of research includes modeling inspired in epidemics, while the second includes propagation models such as the cascade [20] or the linear threshold models [23]. This research is of utmost importance to areas like marketing, advertising, epidemiology, and social media analysis [79]. Some approaches to information spread modeling rely only on graph theory results [3, 71] assuming complete knowledge of the network, but they don’t report empirical validation over real data, some are purely speculative [27, 35, 52, 59, 69, 74, 81]. Aggregated predictions of macroscopic or mesoscopic behaviour of information diffusion have been also proposed [18, 26, 78–80]. For example, modeling the spread of information as epidemic propagation predicts the number of users that belong to the infected class [78–80] instead of trying to predict the individual infection. Other works model the density function of the distribution of influenced users [26], the node influence derived from the network topological properties [18], or the macroscopic information dissemination as the propagation of a signal over the network where interference between events is modeled by signal convolution [58]. At the microscopic level, learning from data the payoff of the social agents decisions allows accurate prediction of information diffusion [40]. Machine learning predictors of twitter activity have been developed [55], however data is not always available for confirmation of results. The role of topicality in Twitter adoption has been considered via machine learning predictive models [22] where topics correspond to selected hashtags, discovering that topicality plays a major role at microscopic information propagation. Hashtag topics are also used in the construction of the similarity measure underlying a radiation transfer model for influence prediction [5], but their role is not isolated.

On the other hand, the semantic modeling of the information content published in the OSN is gaining attention. For instance, semantic analysis of social networks weibo and twitter based on single word topics has been applied to study the public perception on vaccines against COVID-19 [46]. It has been shown that semantic modeling of user contents allows for improved community detection [28, 82]. The impact of specific events on the social media can be assessed using semantic modeling. For instance, an approximate model [17] is shown to detect events in the social median, while event summarization on the basis of tweets can be achieved by a deep learning architecture [21]. Specifically, topic analysis by LDA has been used to uncover the meaning of events in social media [44] and the evolution of contents in the social media [15]. Notably, sentiment analysis has been proposed to predict song contest results [16]. For recommender systems, LDA-based topic hybrid recommender system has been proposed [33], and semantic analysis for recommendations has been also used in learning environments [32]. Moreover, semantic modeling of the user interactions with a chatbot allows for personalized interactions [43]. Semantic analysis may be extended in the time domain, allowing to measure changes in contents dynamically. Topic dynamics was applied to track the emergence of influential tweets about Fukushima disaster [53] over a long period of time. The consideration of both time and content allowed to monitor changes is a VCoP where the user exchange information about cosmetics [67].

3 Materials and methods

3.1 Computational pipeline

The computational pipeline of this paper is shown in Fig. 1. It encompasses 5 phases corresponding to the numbered boxes in the figure going from left to right):

(1) Data Mining Process: in this phase we carry out the curation and preprocessing of the raw OSN data described in Sect. 3.2. Section 3.3 describes data curation and preprocessing. Moreover we build a characterization of each forum contribution by LDA semantic unsupervised topic analysis. Section 3.4 gives a short overview of LDA.

(2) Expert Training data Labeling (ETL): in this phase we prepare the user categorization using information

Table 1 Information diffusion modeling approaches found in the literature

Ref./ year	Model description	Results	Data set
[35]/ 2007	SIR model to estimate number of accesses to a site	N/A	“2 channel” web forum. DATA: number of posters per 15 min 9 p.m. Jan 10 2007–6 a.m. Jan 11 2007
[12]/ 2009	Topological properties of OSN graph	N/A	Flickr like data[1]
[3]/ 2010	Game theoretic diffusion of technologies model that allows for competition between agents	N/A	Not applicable to implicit networks
[80]/ 2011	Topic-based SIR model. Applied to violent topic diffusion	*R*-square: 0.57–0.8	Ummah data set Dark Web Forum Portal by AI lab of U. of Arizona. 1,263,724 posts, 76,242 threads, 15,345 authors
[52]/ 2012	Probabilistic generative model of information emergence in networks, capturing internal and external exposures. URL diffusion	N/A	Tested on synthetic data and complete Twitter January 2011 data set. 3 billion tweets, 18,186 URLs
[81]/ 2012	SCIR model	N/A	Tested on synthetic data
[78]/ 2012	Event-driven SIR model	*R*-square: 0.66–0.89	Yahoo! Finance Walmart message board
[71]/ 2013	Deterministic model of competitive information diffusion on the Iterated Local Transitivity	N/A	Not applicable to implicit networks
[27]/ 2014	Evolutionary game theory model for diffusion dynamics	N/A	Twitter hashtag data set. 1000 Twitter hashtags, number of mentions per hour and time series
[74]/ 2014	SIS and SIR models with edge weights	N/A	Synthetic data
[69]/ 2015	Meme propagation model based on network topology	N/A	Tested on Higgs Twitter Network
[22]/ 2015	Adoption probability. Machine learning prediction	$F1 = 0.93$	Twitter hashtags and URLs 2009
[79]/ 2016	Topic-level SIR model	R^2 0.52–0.75 and 0.44–0.79	Yahoo! Finance Walmart message board (139,062 threads, 441,954 messages, 25,500 authors) and US Politics Online Breaking News in Politics (2192 threads, 130,850 messages, 1124 authors)
[59]/ 2016	SIR model with stifling and forgetting mechanisms	N/A	Synthetic data and on OSN Renren (9590 nodes, 89,873 edges)
[26]/ 2017	Hydrodynamic information diffusion prediction model	$\overline{ACC}$: 76.2–88	6500 video tweets from Sina-weibo
[5]/ 2017	Physical radiation transfer	N/A	Twitter dataset about 9000 users
[40]/ 2017	Decision payoff modeling	Avg. precision: 0.7	Sina Weibo and Flickr datasets
[58]	Expectation maximizacion. Monte Carlo simulation	R^2: 0.98	SINA microblogging prediction of diffusion volume
[55]/ 2020	Bayesian logistic regression and random forests predictors	*F*1: 0.89–0.91	Twitter data crawled on informative and trending topics. N/A
[37]/ 2020	Modified forest fire	Num. spreaders	Twitter datasets

http://socialnetworks.mpi-sws.org/datasets.html

N/A not available

from experts (i.e. the network administrators) as described in the Sect. 3.2. This categorization modulates some of the LCA parameters as discussed below.

(3) Neurophysiological Model Setup: in this phase we formulate the LCA neural model that simulates the process of decision making for a content contribution published in some thread of a sub-forum. Our extended LCA (ELCA) is described in Sect. 3.6.

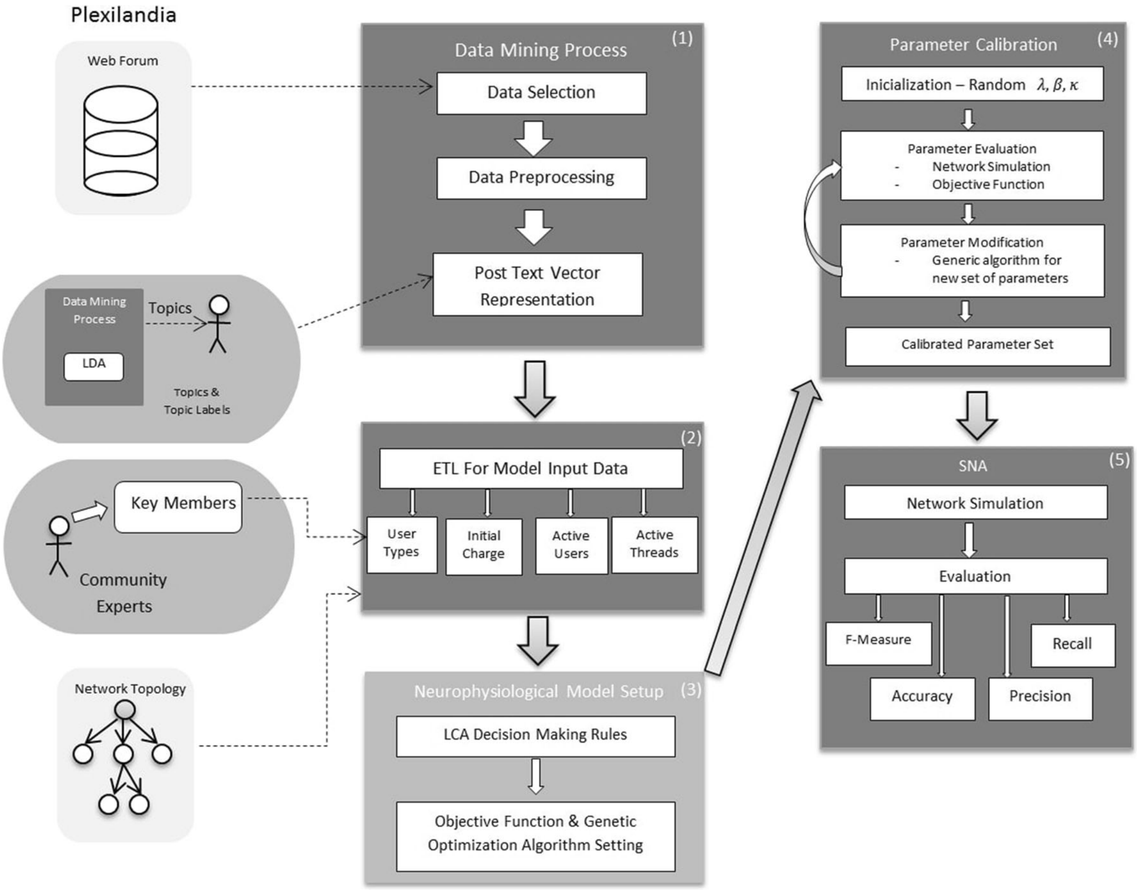


Fig. 1 Study computational pipeline

From the LDA semantic model we construct the value of each conversation thread for each relevant user that will be the input for the ELCA contribution decision prediction. This construction is described in Sect. 3.5.

(4) Parameter Calibration: We set up the genetic algorithm optimization to find the best parameter values of the neural model. The objective function is defined as the predictive performance over a subset of the dataset selected for model calibration. The genetic algorithm searches for the optimal settings of the LCA parameters using the data reserved for training. The genetic algorithm is described in Sect. 3.7.

(5) Social Network Analysis (SNA) computational experiments: we apply LCA to simulate the content contribution decisions made by the users. The results of the simulation are used as prediction of the actual user behavior. The quality of the prediction is evaluated against the actual contributions registered for the time periods designed for validation. The predictive performance is measured by the $F1$ score. Experimental results are presented in Sect. 4.

An algorithmic description of the prediction of posts using the ELCA model is given in Algorithm 1, where the optimal values of the parameters $\hat{\beta}_c$, $\hat{\kappa}_c$, and $\hat{\lambda}_c$ have been already estimated by the genetic algorithm that is described in Algorithm 2.

Algorithm 1 Prediction of post contributions by ELCA on the basis of the valuation of the threads by the users, for each test period after estimation of $\hat{\beta}_c$, $\hat{\kappa}_c$, and $\hat{\lambda}_c$ by the genetic algorithm described in Algorithm 2.

```
Input: collection of posts aggregated in time periods $\mathscr{C} = \{\mathscr{C}_t; t = 1, \ldots, T\}$ where $\mathscr{C}_t = \{p_1^t, \ldots, p_{N_t}^t; p_k \in \mathscr{P}\}$ after data curation; each post is a tuple $p = [u, h, \{v\} \subset \mathscr{V}]$
Semantic preprocessing: Apply LDA to compute semantic representation of the posts as a combination of topics $\left\{\rho_p \subset \mathscr{T}; p \in \mathscr{P}\right\}$.
For each $t \in \{2, \ldots, T\}$
  1. Compute the semantic representation of each thread at each time period $v_h^t = \frac{1}{|\mathscr{P}(h,t)|} \sum_{p \in \mathscr{P}(h,t)} \rho_p$,
  2. Compute the semantic representation of each user at each time period $\mu_{u,s}^t = \frac{1}{|\mathscr{P}(u,s,t)|} \sum_{p \in \mathscr{P}(u,s,t)} \rho_p$,
  3. Compute the utility of a thread for a user $\Psi_1(\mu_{u,s}^t, v_h^t) = \frac{1}{1-\chi(\mu_{u,s}^t, v_h^t)}$, where $\chi(\mu_{u,s}^t, v_h^t) = \frac{\mu_{u,s}^t \cdot v_h^t}{|\mu_{u,s}^t||v_h^t|}$ is the cosine distance,
  4. Compute the normalized utilities $V_{u,s,h}^t = \Psi_2(a, \mu_{u,s}^t, v_h^t) = a \frac{\Psi_1(\mu_{u,s}^t, v_h^t)}{\max\limits_{j \in \mathscr{T}\mathscr{H}_f^t} \Psi_1(\mu_{u,s}^t, v_j^t)}$,
  5. Compute the valuations of each thread by each user $I_{u,s,h}^t = \Omega(\mathbf{V}_{u,s}^t(m), h) = \hat{\beta}_{(c(u))} e^{V_{u,s,h}^t} \left( \sum\limits_{j \in \mathscr{T}\mathscr{H}_f^t(u,m)} e^{V_{u,s,j}^t} \right)^{-1}$.
  6. For all users $u$ and threads $h$ integrate by Euler method the ELCA differential equations
       $$dX_h^{(u)}(\tau) = \left[ I_{u,s,h}^t - \sum_{j \in \mathscr{T}\mathscr{H}_f^t} \hat{\omega}_{hj}^{(c(u))} X_j^{(u)}(\tau) \right] d\tau + \sigma_h^{(u)} dW_h,$$
     until $X_h^{(u)}(\tau^*) > Z$, where $Z$ is the decision threshold, for each user $u$.
  7. The predicted edges of the contribution publication graph are given by $PG_t = \left\{(u,h) \middle| X_h^{(u)}(\tau^*) > Z\right\}$. Performance measures (Section 3.8) are computed comparing against the ground truth post publications $GT_t = \{(u,h) | \exists [u,h,] \in \mathscr{C}_t\}$.
```

Algorithm 2 Genetic algorithm for the estimation of ELCA parameters $\hat{\beta}_c$, $\hat{\kappa}_c$, and $\hat{\lambda}_c$.

```
Input: collection of posts of first time period $\mathscr{C}_1 = \{p_1^1, \ldots, p_{N_1}^1; p_k \in \mathscr{P}\}$ after data curation; each post is a tuple $p = [u, h, \{v\} \subset \mathscr{V}]$; semantic representation of the posts as a combination of topics $\left\{\rho_p \subset \mathscr{T}; p \in \mathscr{P}\right\}$.
  1. Build random initial population $\mathbf{P}(k=0) = \left\{P_g(k); g = 1, \ldots, 100\right\}$, where $P_g(k) = \left\{\left(\hat{\beta}_c(k), \hat{\kappa}_c(k), \hat{\lambda}_c(k)\right), c \in \{A, B, C, X\}\right\}$,
  2. Estimate initial fitness function $f_g(k=0)$ for each individual by
     (a) running the prediction in Algorithm 1 over $\mathscr{C}_1$ using $P_g(k)$ as ELCA parameters.
     (b) fitness $f_g(k=0)$ is the prediction accuracy of $PG_1$ against $GT_1$ after convergence.
  3. For generations $k = 1, \ldots, 1000$
     (a) select by roulette wheel 10 old individuals preserved for the next generation $\mathbf{P}_{old,10}(k)$
     (b) select 90 crossover pairs by roulette wheel over the fitness values $\left\{f_g(k-1)\right\}$ of the parent population $\mathbf{P}(k-1)$
     (c) apply single point crossover to obtain the descendancy $\mathbf{P}_{cross,90}(k)$
     (d) apply real valued mutation to $\mathbf{P}_{cross,90}(k)$ to obtain $\mathbf{P}_{mut,90}(k)$
     (e) compute the fitness function $f_g(k)$ of each individual in $\mathbf{P}_{mut,90}(k)$ as specified in step 2.
     (f) $\mathbf{P}(k) = \mathbf{P}_{old,10}(k) \cup \mathbf{P}_{cross,90}(k)$
  4. Return the individual $P_g^*(k)$ with greatest fitness $f_g^*(k) = \max\limits_{g,k} \left\{f_g(k)\right\}$.
```

3.2 Experimental dataset

The experimental works reported in this article are carried out over the data extracted from a web-based forum called *Plexilandia*, which was implemented as an OSN with more than 2500 members. *Plexilandia* supports a Virtual Community of Practice (VCoP) [6, 14, 62, 63, 65] specifically devoted to tinkering with musical apparatus that has been running for over 15 years. We have access to data from its greatest activity epoch, spanning 9 years. Table 2 contains the number of content publications *per* sub-forum along these 9 years, including the total number of posts. From now on, we may use the word "post" meaning a content contribution to a sub-forum.

The topics treated within Plexilandia's forum are arranged into sub-forums according to the interest of the VCoP members that frequent it, namely Table 2 identifies the following sub-forums: Amplifiers, Effects, Luthiers, General, Audio for professionals, and Synthesizers. Contents published in such sub-forums should be strictly related to the purpose of the community, although spurious topics may emerge from unrestricted user interaction. The forum hierarchical structure of sub-forums is illustrated in Fig. 2.

Content contributions of users are conducted inside conversations that we will be denoting as *threads*. A thread about some discussion begins with a message posted by a user, containing a question or the presentation of an idea for discussion. Then, the different members of the community post their contributions thus increasing the shared knowledge about the central theme of the conversation. Each publication in the thread is composed of elements such as the user identifier (ID); the content contribution, which depending on the forum can be text, images, links to other pages, videos, and the management information of the forum system, such as publication creation date, the thread, and the topic it belongs to. All these elements might be taken into consideration but in this paper only the text content of posts will be exploited to build and analyze the social network.

3.2.1 Experimental training and validation data setup

According to the content structure of the Plexilandia Web Forum, the dataset is partitioned into sub-forums. For the computational experiments five sub-forums are considered. After examination of the distribution of the number of posts for different sizes of time periods (1 week, 2 weeks, 1 month, 2 months, 4 months) and the behavior of the threads during that time, a time period of 1 month has been selected, therefore aggregating the data into 13 time periods. The number of active users, active threads, and posts made during each of these 13 monthly time periods for each of the sub-forums is shown in Table 3. We provide an approximate ratio of imbalance (IBR) of each sub-forum computed as the number of possible content contributions, i.e. number of active users times the number of active threads, divided by the number of actual posts. Figure 3 shows the data partition for the validation experiments, using the data from the first month of 2013 (January) for the ELCA model calibration and the remaining months for testing. In other words, 8% of the data is used for the estimation of the optimal ELCA parameters by a genetic algorithm, and 92% for testing. Thus, model validation is set in the framework of training data scarcity, which is more realistic that training data abundance (such as when using 70% for training, 30% for testing) when trying to predict the online evolution of an OSN.

3.2.2 Categories of users

The OSN administrators provided a stratification of members for the year 2013 into four user categories [63] according to the role that they play in keeping the forum alive:

- Experts Type A: which are the most important key-members that create and sustain meaningful threads in relevant sub-forums. There are 34 such members based on administrators' criteria.
- Experts Type B: which are also very important but to a lesser degree than A-type key-members. They

Table 2 Plexilandia's activity measured in number of content publications *per* relevant sub-forum *per* year

Sub-forum	2006	2007	2008	2009	2010	2011	2012	2013	2014	Total
Amplifiers (SF 2)	392	2165	2884	3940	3444	3361	2398	1252	985	20,821
Effects (SF 3)	184	1432	3362	3718	4268	5995	4738	2317	1331	27,345
Luthier (SF 4)	34	388	849	1373	1340	2140	926	699	633	8382
General (SF 5)	76	403	855	1200	2880	5472	3737	1655	1295	17,573
Pro Audio (SF 6)	–	–	–	–	–	342	624	396	219	1581
Synthesizers (SF 7)	–	–	–	–	–	–	–	104	92	196
Total	**686**	**4388**	**7950**	**10,231**	**11,932**	**17,310**	**12,423**	**6423**	**4555**	**75,898**

Bold values correspond to summary values, either total or first order statistics, mean, min and max values

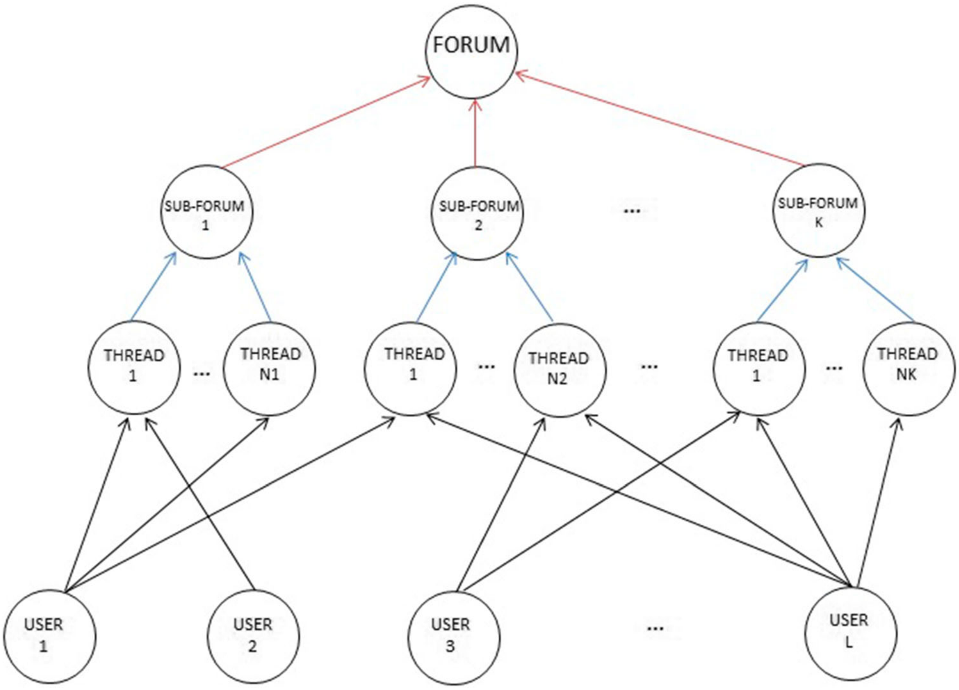


Fig. 2 Hierarchical topology of VCoP web forums

Table 3 Sub-forum statistics (number of active users, number of active threads, number of posts) *per* month

Month	Sub-forums				
	SF 2	SF 3	SF 4	SF 5	SF 6
1	(45, 25, 103)	(49, 43, 145)	(32, 40, 115)	(60, 37, 164)	(14, 11, 49)
2	(19, 10, 51)	(46, 29, 169)	(25, 8, 81)	(47, 27, 131)	(7, 5, 13)
3	(35, 20, 83)	(51, 46, 252)	(20, 13, 60)	(58, 30, 182)	(16, 6, 33)
4	(38, 27, 133)	(53, 43, 196)	(22, 15, 50)	(36, 23, 84)	(6, 5, 13)
5	(32, 22, 55)	(51, 44, 184)	(12, 8, 23)	(55, 28, 145)	(11, 9, 30)
6	(33, 22, 94)	(52, 38, 208)	(5, 3, 7)	(53, 36, 202)	(11, 5, 13)
7	(26, 14, 57)	(49, 32, 173)	(19, 10, 46)	(55, 35, 176)	(10, 7, 52)
8	(38, 24, 127)	(42, 37, 171)	(21, 17, 57)	(45, 29, 116)	(9, 3, 13)
9	(35, 17, 94)	(43, 33, 174)	(19, 10, 52)	(25, 19, 72)	(11, 7, 41)
10	(35, 23, 110)	(44, 29, 138)	(20, 9, 30)	(34, 25, 66)	(15, 5, 27)
11	(38, 22, 121)	(43, 24, 124)	(22, 9, 72)	(25, 13, 41)	(8, 5, 37)
12	(31, 19, 94)	(49, 38, 156)	(12, 8, 33)	(42, 25, 105)	(15, 6, 36)
13	(27, 14, 59)	(31, 30, 102)	(28, 17, 104)	(38, 24, 98)	(11, 6, 27)
Total	**(168, 221, 1181)**	**(174, 351, 2192)**	**(96, 134, 730)**	**(171, 282, 1582)**	**(501, 47, 384)**
IBR	31.43	27.86	17.6	30.48	61.32

Bold values correspond to summary values, either total or first order statistics, mean, min and max values

Last row contains the imbalance ration (IBR) computed as explained in the text

contribute steadily but have less pivotal roles. There are 21 such members.

- Experts Type C: This type corresponds to those that are historic key-members. They have been involved in the social network since its origins, but they are not continuously participating. In this class, there are about 11 members.
- Non-experts or Type X: this class contains all members of the social network which are not key-members. They don't belong to the social network core and usually, they ask questions rather than publish answers or tutorials.

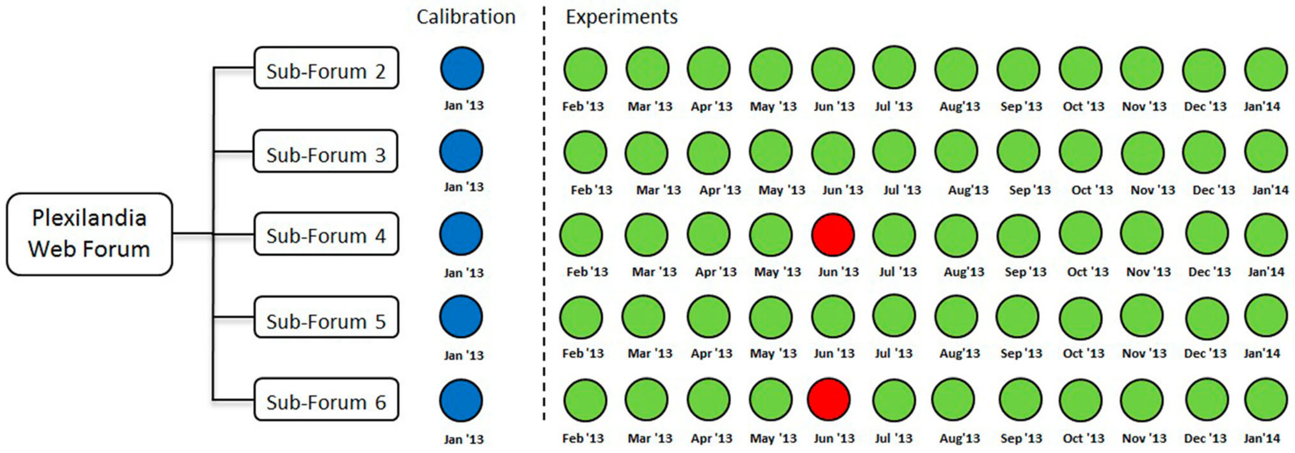


Fig. 3 Experimental setup of data exploitation for model validation. Red dots correspond to months with missed data. Blue dots correspond to months whose data is used for training. Green dots correspond to months whose data is used for testing (color figure online)

We use only the data for the years 2013 and 2014 because we only have the information regarding key-members for these years [63]. We use the data of sub-forums 2 to 6. Discarding sub-forums 1 and 7 because they have not enough posts to contribute to the analysis.

3.3 Data curation and preprocessing

The first step in our computational pipeline is the Plexilandia's data curation and preprocessing [75]. First, we filter out the quotes from previous content contributions posted in the thread. A user can respond to a post by creating a new content contribution including a copy of the cited post plus the additional text of the new contribution. Therefore, it is necessary to delete the replicated part of the new post retaining only the new text input. Next, we transform the acronyms or abbreviations, eliminate spelling errors, and all elements of the posts that make them not comparable. This process is carried out by two natural language processing techniques: stemming and removing stop words. This serves to make posts comparable and to reduce the number of words used to compute post comparison. We apply LDA unsupervised topic modeling described in the next section for the semantic modeling of the content of documents [61].

3.4 LDA topic analysis for semantic modeling

In this section we, give a brief account of the Latent Dirichlet Allocation (LDA) topic analysis used for semantic modeling. Let $\mathscr{V}$ be a vector of size $|\mathscr{V}|$ in which every row represents a different word used in the network, i.e. the vocabulary. Let v_i be the word in place i of vector $\mathscr{V}$. It is possible to represent post p_j as a sequence of S_j words out of $\mathscr{V}$, with $S_j = |p_j|$, where $j \in \{1, \ldots, |\mathscr{P}|\}$ and $\mathscr{P}$ corresponds to the number of posts that have been published in the VCoP forum. A *corpus* is defined as a collection of posts $\mathscr{C} = \{p_1, \ldots, p_N\}$. We can define the matrix $\mathscr{W}$ of size $|\mathscr{V}| \times |\mathscr{P}|$ where each element $w_{i,j}$ of this matrix is defined as the number of times the word v_i appears in post p_j. Then $\sum_{i=1}^{|\mathscr{V}|} w_{i,j} = S_j$. Likewise, we can define $\sum_{j=1}^{|\mathscr{P}|} w_{i,j} = T_i$ which represents the total number of appearances of the term w_i in the corpus.

A corpus can be represented by the product of the term frequency and the inverse document frequency (TF-IDF) matrix $\mathscr{M}$ of size $|\mathscr{V}| \times |\mathscr{P}|$ [68], which is defined as follows: each entry $m_{i,j}$ in the matrix is determined as

$$m_{i,j} = \frac{w_{i,j}}{T_i} \times \log\left[\frac{|\mathscr{P}|}{1 + n_i}\right], \tag{1}$$

where n_i is the number of posts including the word w_i, T_i is the maximum number of appearances of word w_i in any post. The IDF term presented in Eq. (1) contains a correction with respect to the original IDF term $\log\left[\frac{|\mathscr{P}|}{n_i}\right]$ to avoid undefined results when a post does not contain words after data curation. For dimension reduction we employ of an unsupervised topic discovery technique, namely, the LDA [4, 8] using the Gibbs sampling implementation [57]. This implementation does not search for the optimal values of the hyper-parameters α, β, and number of required topics $|\mathscr{T}| = k$, so we have to make an empirical exploration to find them. LDA provides us with the distribution of each word over the discovered topics, the distribution of topics over the posts, and the n most important words that represent each topic together their belonging probabilities. In order to have fixed size probability vectors representing each topic $|\mathscr{V}|$, we pad them with zeros. These vectors are the columns of the semantic matrix (SM) [Terms × Topics]. In order to obtain the semantic description of the posts in a matrix of size [Posts × Topic], we multiply the SM with $\mathscr{M}^t$, the transpose of the TF-IDF matrix defined by Eq. (1). The resulting [Posts × Topic] matrix contains the semantic explanation of each post as a linear combination of the discovered topics via their vector semantic representations given by the rows of the matrix, denoted $\{\rho_p; p \in \mathscr{P}\}$.

3.5 From semantic modeling to valuation

Let us denote $\mathscr{U}$, $\mathscr{T}\mathscr{H}$, and $\mathscr{S}\mathscr{F}$ the set of users, the set of threads, and the set sub-forums in the virtual community, respectively. The results of the LDA semantic analysis, namely the vectors ρ_p, allows to induce each user $(u \in \mathscr{U})$ multi-topic preference vector representation, and each thread $\mathscr{T}\mathscr{H}$ semantic content vector representation. The process to compute these semantic representations is as follows:

1. We aggregate the users content contributions according to the sub-forum $\mathscr{S}\mathscr{F}$ where they are posted.
2. We discretize the time axis into periods of size Δt, thus creating a set of time periods T. Subsequently, we aggregate the content contributions from each sub-forum according to the time $(t \in T)$ period they belong to.
3. We extract the users $(\mathscr{U}_f^t)$ and threads $(\mathscr{T}\mathscr{H}_f^t)$ that are active during each time period. A user u is active in sub-forum f and period t if he makes a content contribution during this period. A thread h in sub-forum f is active if any user makes a content contribution to the thread during period t.
4. The thread semantic content vector representation for a period, denoted v_h^t, is the mean of the semantic vector representations ρ_p for the content contributions that belong to both the thread h and the period t, formally:

$$v_h^t = \frac{1}{|\mathscr{P}(h,t)|} \sum_{p \in \mathscr{P}(h,t)} \rho_p, \tag{2}$$

where

$$\mathscr{P}(h,t) = \{p \in \mathscr{P} : p \,\text{is posted in thread}\, h \,\text{during period}\, t\}$$

.

5. To compute the user semantic representation, we categorize into subgroups, denoted s, the content contributions made by a user during a period according to the thread they were posted in. A user will have as many semantic vector representations for a period as threads that he has contributed to during this period. We denote the collection of these vector representations as S_u^t.
6. A user semantic vector representation for a period t and subgroup of content contributions s, denoted $\mu_{u,s}^t$, is the mean of the semantic vector representations ρ_p for the content contributions made by the user u in this period of time, formally:

$$\mu_{u,s}^t = \frac{1}{|\mathscr{P}(u,s,t)|} \sum_{p \in \mathscr{P}(u,s,t)} \rho_p, \tag{3}$$

where

$$\mathscr{P}(u,s,t) = \{p \in \mathscr{P} : p \,\text{is posted by user}\, u \,\text{in period}\, t \,\text{and belongs to subgroup}\, s\}. \tag{4}$$

Now that we have the multi-topic semantic vector representation of the users and the semantic representation of the threads, we apply the computational pipeline shown in Fig. 4 to obtain the input for the extended LCA that implements the content contribution decision model.

1. First, we select a measure of the similarity χ of two semantic vector representations in the topic space. We use the cosine similarity, given by the cosine of the angle formed between two semantic vector representations. Thus, for a user multi-topic preference vector representation $\mu_{u,s}^t$ and a thread semantic content vector representation v_h^t, the similarity between them is given by

$$\chi(\mu_{u,s}^t, v_h^t) = \cos(\theta) = \frac{\mu_{u,s}^t \cdot v_h^t}{|\mu_{u,s}^t||v_h^t|}, \tag{5}$$

where θ is the angle between $\mu_{u,s}^t$ and v_h^t.
2. Then, we define a function Ψ_1 mapping semantic similarity into user utility. The utility that a user extracts from a thread is the expected number of times he chooses the thread over other threads to make a content contribution. Consider that $\pi = 1 - \chi(\mu_{u,s}^t, v_h^t)$ is the success probability parameter of a geometric distribution. Utility Ψ_1 of the similarity between user and thread semantic representations is defined as follows [11]:

$$\Psi_1(\mu_{u,s}^t, v_h^t) = \frac{1}{1 - \chi(\mu_{u,s}^t, v_h^t)}. \tag{6}$$

Furthermore, the preference of a user for a thread, i.e. the normalized user utility of a thread h, denoted $V_{u,s,h}^t$, takes into account all the threads in the sub-forum, computed by a function Ψ_2 defined as follows:

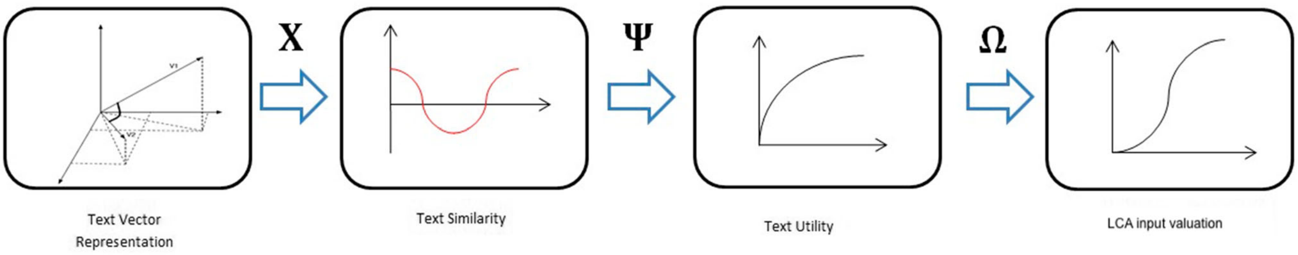


Fig. 4 Transformations applied to the semantic modeling of users and threads to obtain the input values for the extended LCA

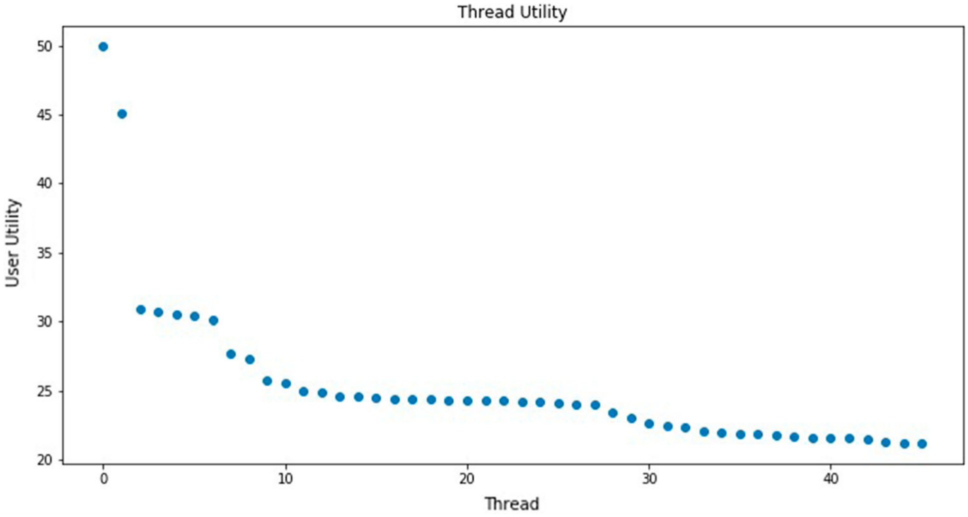


Fig. 5 An instance of thread utility long tail distribution for a user at some specific time period

$$V^t_{u,s,h} = \Psi_2(a, \mu^t_{u,s}, v^t_h) = a\frac{\Psi_1(\mu^t_{u,s}, v^t_h)}{\max_{j\in\mathcal{T}\mathcal{H}^t_f} \Psi_1(\mu^t_{u,s}, v^t_j)}, \quad (7)$$

where parameter a modulates the preference of the users to threads whose topics are similar to the topics covered by the user content contributions. The greater the preference, the greater the satisfaction extracted from the conversation. Figure 5 plots an example of the utility values that a user attributes to the threads that are active at some period in time. Notice that only a few threads are of great interest to the user. Most active threads are stacked at the tail of the plot, meaning that they mostly contribute noise to the decision process. Therefore, we reduction in the number of alternative threads that a user takes into account during his decision-making process to generate content, keeping only the m threads with top utility values. This reduction of alternatives is based on classic research results about working memory and attention span [50].

3. Finally, we define a function Ω that maps the normalized user utility of each thread into the LCA input associated with the decision to make a content contribution to the thread, denoted $I^t_{u,s,h}$. For this purpose, we make use of random utility theory [11]: $I^t_{u,s,h}$ is proportional to the likelihood of choosing between alternative threads. Formally:

$$I^t_{u,s,h} = \Omega(\mathbf{V}^t_{u,s}(m), h) = \beta_{(c(u))}\frac{e^{V^t_{u,s,h}}}{\sum_{j\in\mathcal{T}\mathcal{H}^t_f(u,m)} e^{V^t_{u,s,j}}} \quad (8)$$

where $\beta_{(c(u))}$ is a proportionality parameter of the model that is specific for the category $c(u)$ of the user (defined as A, B, C, or X in Sect. 3.2), and $\mathcal{T}\mathcal{H}^t_f(u,m) = \{h \in \mathcal{T}\mathcal{H}^t_f : h\,\text{utility is one of the top}\,m\,\text{for user}\,u\}$.

3.6 Extended leaky competing accumulator (ELCA)

The decision process leading to the contribution of posts to conversation threads is modeled by an extended leaky competing accumulator (ELCA). The original LCA [9, 65, 76, 77] did only consider a decision carried out by a single agent, while our ECLA carries out simultaneously the decision processes of many users simultaneously, i.e., ECLA extends LCA over a community of users undertaking decisions simultaneously. We consider independent processes for each sub-forum f and each time period t. We

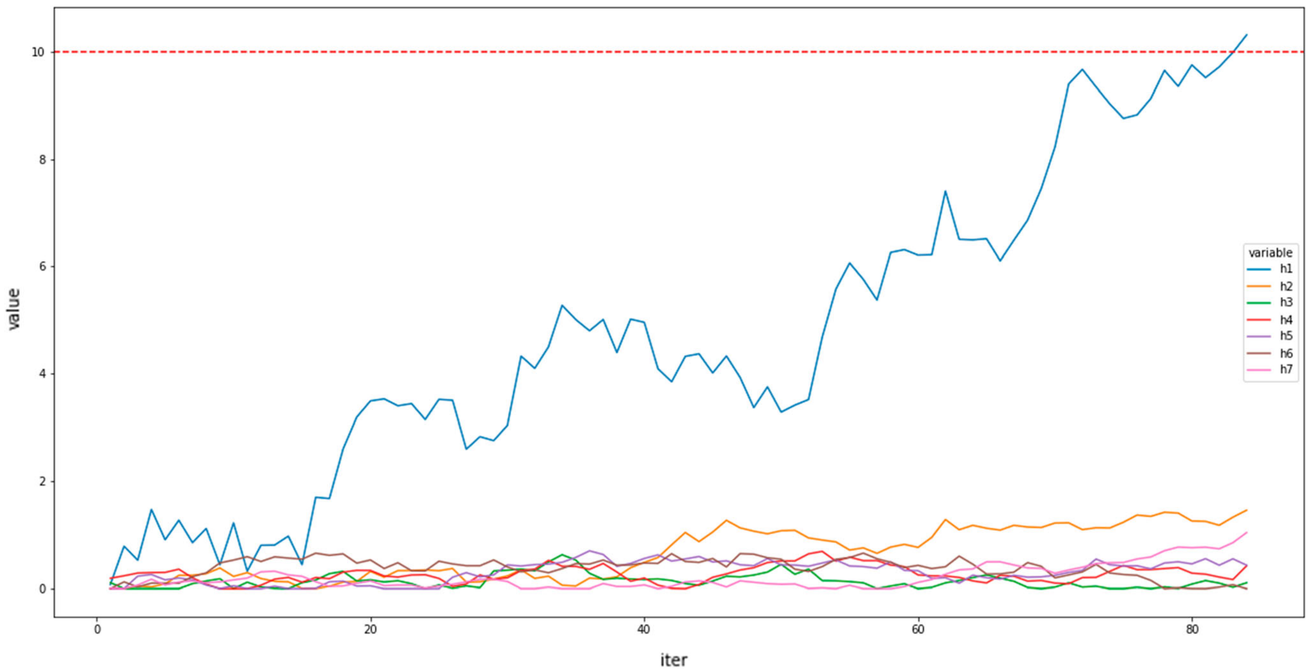


Fig. 6 An instance evolution of the accumulators corresponding to a decision to post by a specific user

define $X_h^{(u)}$ as the (neural) activation associated with the decision by user $u \in \mathscr{U}_f^t$ to publish a post in thread $h \in \mathscr{T}\mathscr{H}_f^t$. The decision process is implemented as dynamic process where the activation units evolve until one of them reaches a given threshold that triggers the corresponding decision. The evolution of the activation units for a user is illustrated in Fig. 6. Moreover, our ELCA has semantically grounded values associated to each choice, the term $I_{u,s,h}^t$ defined in Eq. (8), while classical LCA models have arbitrary values tuned by the researcher intuition. Finally, the provide a procedure to estimate the ELCA optimal parameters to reproduce the actual decisions made by the users, in a way similar to the training of conventional machine learning approaches.

The ELCA model describes the evolution of the joint decision process of all users as the simulation of the following set of dynamic stochastic equations:

$$\mathrm{d}X_h^{(u)}(\tau) = \left[I_{u,s,h}^t - \sum_{j \in \mathscr{T}\mathscr{H}_f^t} \omega_{hj}^{(c(u))} X_j^{(u)}(\tau) \right] \mathrm{d}\tau + \sigma_h^{(u)} \mathrm{d}W_h, \quad h \in \mathscr{T}\mathscr{H}_f^t, u \in \mathscr{U}_f^t, \tag{9}$$

that are integrated applying the Euler method. For each sub-forum f we have as many dynamic equations implementing concurrent decision processes as users and threads that are active during the time period t. The dynamic equations for each user u in Eq. (9) are integrated until time τ^* when user u takes the decision to post a message to thread h^*, i.e. when the corresponding unit overcomes a decision threshold $X_{h^*}^{(u)}(\tau^*) \geq Z$, as illustrated in Fig. 6. We have empirically set $Z = 10$. Parameters $\omega_{hj}^{(c(u))}$ modulate the lateral inhibition by the other ongoing decision processes of user u, where $c(u) \in \{A, B, C, X\}$ denotes the category of the user defined in Sect. 3.2. The term $I_{u,s,h}^t$ in Eq. (9) is an external constant input value in favor of posting a contribution in thread alternative h based on the semantic analysis developed above. Those input values are positive, i.e. $I_{u,s,h}^t \geq 0$. External input values are linearly accumulated in the activation variable $X_h^{(u)}$. It takes different values depending on the relation modeled and the category of the user, as shown in Eq. (10).

$$\omega_{ij}^{(c)} = \begin{cases} \kappa_c & i = j \\ \lambda_c & i \neq j \end{cases}, \quad c \in \{A, B, C, X\}, \tag{10}$$

where the κ_c parameter models the activation decay of each unit [48]. Lateral inhibition between accumulator units is modeled by the λ_c parameter. Equation (10) considers equal effect for all units stratified by the different user category defined by the OSN administrators. Following the biological inspiration, the activation variables are restricted to positive values ($X_h^{(u)} > 0$). This hard limit has some interesting computational properties [9]. This model is in accordance with perceptual decision making [19]. Initial conditions $X_h^{(u)}(\tau = 0)$ are specified by Eq. (11):

$$X_h^{(u)}(\tau = 0) = (1 + \gamma)^l - 1 \tag{11}$$

Parameter l in Eq. (11) denotes the number of times thread alternative h has been chosen previously, and parameter $\gamma \geq 0$ models the effect of repeated choices of the same alternative approaching the asymptotic curve defined in [38]. Recent works have shown convergence to a decision for large number of choices in a modified LCA model [45], but their model is limited to a single agent. They show that it is possible to recover the model parameters by maximum likelihood approach, however, they refer to the reproduction of simulation traces while we deal in the next section with parameter estimation to approximate the user decision behavior extracted from the real OSN data.

3.7 ELCA parameter estimation by genetic algorithm

ELCA parameter estimation was implemented by a genetic algorithm (GA) [73] illustrated in Fig. 7 with the following settings: Each individual $P_g \in \mathbf{P}$ in the GA population is composed of 12 real valued genes, which are estimations of the parameters of the LCA model for each kind of user in the sub-forum, i.e. $P_g = \left\{ \left(\hat{\beta}_c, \hat{\kappa}_c, \hat{\lambda}_c \right), c \in \{A, B, C, X\} \right\}$. The size of the population was 100 individuals. The initial values of the individuals component parameters was generated following a uniform distribution in the [0, 1] interval. The fitness function is the accuracy of content contribution prediction by the LCA model using the individual parameter settings over the first month of the dataset. In other words, in order to compute the fitness of each individual in the population we run an instance of the LCA simulation comparing its track of post publication decision to the data from the first month. The individual selection for crossover is carried out by Baker's linear-ranking algorithm [70] and roulette wheel selection [36]. Reproductive crossover was implemented by a single point crossover algorithm [60]. Mutation operator was a real-valued mutation [51]. Independent GA searches were carried out for each sub-forum. The details of the implementation, such as population size, number of generations computed, and the implementation of elitist selection policies are specified in Algorithm 2.

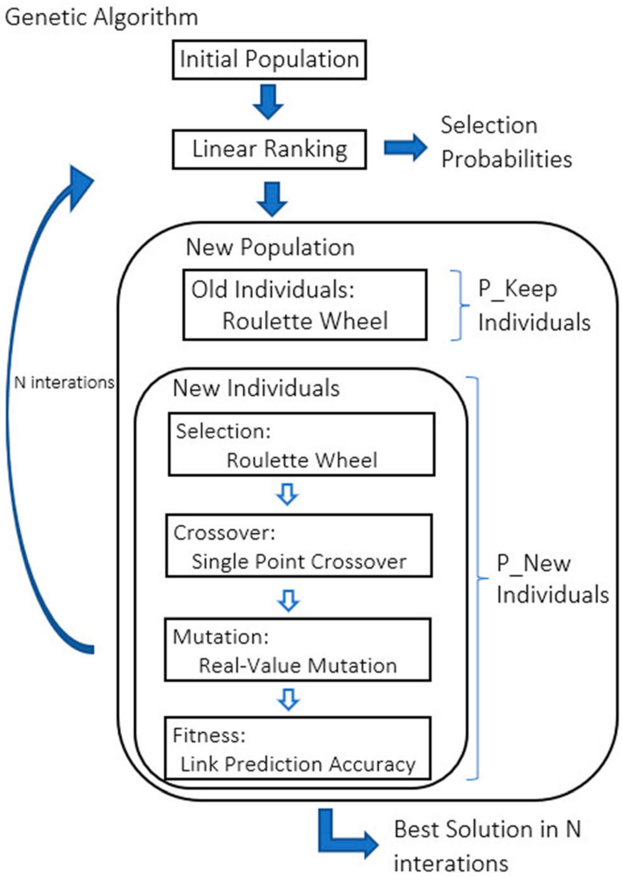

Genetic Algorithm
Initial Population
Linear Ranking
Selection Probabilities
New Population
Old Individuals: Roulette Wheel
P_Keep Individuals
N interations
New Individuals
Selection: Roulette Wheel
Crossover: Single Point Crossover
Mutation: Real-Value Mutation
Fitness: Link Prediction Accuracy
P_New Individuals
Best Solution in N interations

Fig. 7 Flowchart of the GA used for ELCA optimal parameter search

3.8 Performance measures

As specified in Algorithm 1, the result of the ELCA simulation are user-thread pairs $PG_t = \left\{ (u, h) \middle| X_h^{(u)}(\tau^*) > Z \right\}$ that are interpreted as predictors of the actual pairs that can be extracted from the ground truth post publications $GT_t = \{(u, h) | \exists [u, h,] \in \mathscr{C}_t\}$. We make independent predictions for each time period and sub-forum. These pairs can be visualized as the edges of bipartite graphs that are the predicted and the ground truth publication graphs. We can define true positives as the edges that are in both graphs, true negatives as the edges that are absent from the two graphs, false positives are edges that appear in the prediction but are absent in the ground truth, and false negatives edges that are absent in the prediction but appear in the ground truth.

In order to evaluate the quality of the ELCA predictions, we compute 4 performance measures combining these basic measures. Namely: Recall, Accuracy, Precision, and the F measure. Recall is the ratio of true positives over the actual edges in the provided ground truth data:

$$\text{Recall} = \frac{\text{Number of true positive edges}}{\text{Number of ground truth edgess}} \tag{12}$$

Precision is the measure of specificity of the model, i.e. the probability of true positives predictions over all edge predictions made:

$$\text{Precision} = \frac{\text{Number of true positive edges}}{\text{Number of predicted edges}} \tag{13}$$

F measure (aka F_1 score) combines precision and recall measuring the balance between them. It is defined as:

$$F\text{ measure} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \qquad (14)$$

Accuracy is the measure of the overall true positive and negative link predictions:

$$\text{Accuracy} = \frac{\text{Number of true positive edges} + \text{Number of true negative edges}}{\text{Number of possible edges}} \qquad (15)$$

Notice that, in our case study, the number of negative edges is much greater than the positive edges, hence the accuracy will be dominated by the prediction of negative edges, i.e. the absence of positive edge prediction, so that it can be high even if there are many missing actual edges. For this reason, we focus the report of results on the F measure that is a more trustful measure in case of high class imbalance.

4 Results and discussion

4.1 Experimental results

As described in Fig. 3, for each sub-forum we carry out an independent GA search to obtain the optimal parameters for the ELCA model over the data from month 1. The optimal ELCA parameter values obtained by the GA search for each sub-forum are specified in Table 4. The ELCA model with these parameter settings is used to predict the generation of posts from users on specific threads for each sub-forum and for each month between February 2013 and January 2014. The average prediction performance results of the ELCA approach are given in Table 5. In Table 6, we present the detailed results in terms of the F-measure for each sub-forum and for each month considered within the time frame. The overall mean F-measure score of ELCA across all sub-forum experiments is 0.61.

Table 5 Predictive performance results averaged over all test periods of the proposed ELCA approach *per* sub-forum

	Sub-forums				
	SF 2	SF 3	SF 4	SF 5	SF 6
Mean recall	0.55	0.48	0.63	0.51	0.83
Mean accuracy	0.92	0.93	0.89	0.93	0.92
Mean precision	0.57	0.50	0.67	0.53	0.85
Mean F-measure	0.56	0.49	0.65	0.52	0.84

Comparison with machine learning approaches For comparison, we have carried out the training of conventional machine learning approaches. The dataset for training is extracted from the same period (first month) used to calibrate the ELCA model. For each possible pair of active user u and thread h, we define the feature vector concatenating the semantic descriptions of the user and the thread $\mathbf{x}_{u,h} = (\mu^t_{u,s}, v^t_h)$, and the class variable $y_{u,h} \in \{\text{existing}, \text{non-existing}\}$ that signals if there is at least one post by user u in thread h in this time period. The testing

Table 4 Optimal ELCA parameter values for each sub-forum found by independent GA searches over the training data (January 2013)

	$\beta_\mathbf{A}$	$\beta_\mathbf{B}$	$\beta_\mathbf{C}$	$\beta_\mathbf{X}$	$\kappa_\mathbf{A}$	$\kappa_\mathbf{B}$	$\kappa_\mathbf{C}$	$\kappa_\mathbf{X}$	$\lambda_\mathbf{A}$	$\lambda_\mathbf{B}$	$\lambda_\mathbf{C}$	$\lambda_\mathbf{X}$
Sub-forums												
SF 2	0.863	0.148	0.511	0.553	0.174	0.055	0.070	0.965	0.491	0.137	0.399	0.189
SF 3	0.584	0.906	0.389	0.029	0.684	0.340	0.217	0.588	0.146	0.951	0.189	0.949
SF 4	0.586	0.833	0.352	0.476	0.642	0.389	0.866	0.981	0.639	0.478	0.107	0.245
SF 5	0.628	0.184	0.000	0.429	0.707	0.733	0.047	0.623	0.0935	0.864	0.847	0.640
SF 6	0.516	0.126	0.490	0.595	0.287	0.692	0.087	0.401	0.956	0.869	0.044	0.315

Table 6 Detailed F-measure results of the proposed ELCA *per* testing month and sub-forum

Month	2	3	4	5	6	7	8	9	10	11	12	13	Mean	Max	Min
Sub-forums															
SF 2	0.72	0.54	0.45	0.52	0.49	0.65	0.54	0.58	0.58	0.49	0.53	0.68	**0.56**	**0.72**	**0.45**
SF 3	0.44	0.47	0.51	0.44	0.50	0.44	0.47	0.45	0.56	0.43	0.48	0.65	**0.49**	**0.65**	**0.43**
SF 4	0.65	0.48	0.63	0.78	***	0.61	0.67	0.66	0.71	0.68	0.72	0.55	**0.65**	**0.78**	**0.48**
SF 5	0.49	0.44	0.56	0.46	0.39	0.47	0.47	0.68	0.63	0.65	0.51	0.45	**0.52**	**0.68**	**0.39**
SF 6	0.82	0.81	0.86	0.84	***	0.84	0.80	0.92	0.95	0.85	0.84	0.69	**0.84**	**0.95**	**0.69**

Bold values correspond to summary values, either total or first order statistics, mean, min and max values

Table 7 Detailed *F*-measure results of the Random Forest approach *per* testing month and sub-forum

Month	2	3	4	5	6	7	8	9	10	11	12	13	Mean	Max	Min
Sub-forums															
SF 2	0.20	0.16	0.10	0.14	0.13	0.17	0.10	0.15	0.14	0.12	0.13	0.17	**0.14**	**0.20**	**0.10**
SF 3	0.08	0.07	0.09	0.07	0.08	0.12	0.10	0.10	0.11	0.15	0.08	0.14	**0.10**	**0.15**	**0.07**
SF 4	0.22	0.19	0.24	0.40	***	0.30	0.19	0.22	0.27	0.26	0.31	0.17	**0.23**	**0.40**	**0.17**
SF 5	0.13	0.10	0.11	0.11	0.07	0.08	0.09	0.14	0.14	0.22	0.10	0.11	**0.11**	**0.22**	**0.07**
SF 6	0.43	0.30	0.55	0.30	***	0.29	0.59	0.32	0.36	0.35	0.60	0.28	**0.38**	**0.60**	**0.29**

Bold values correspond to summary values, either total or first order statistics, mean, min and max values

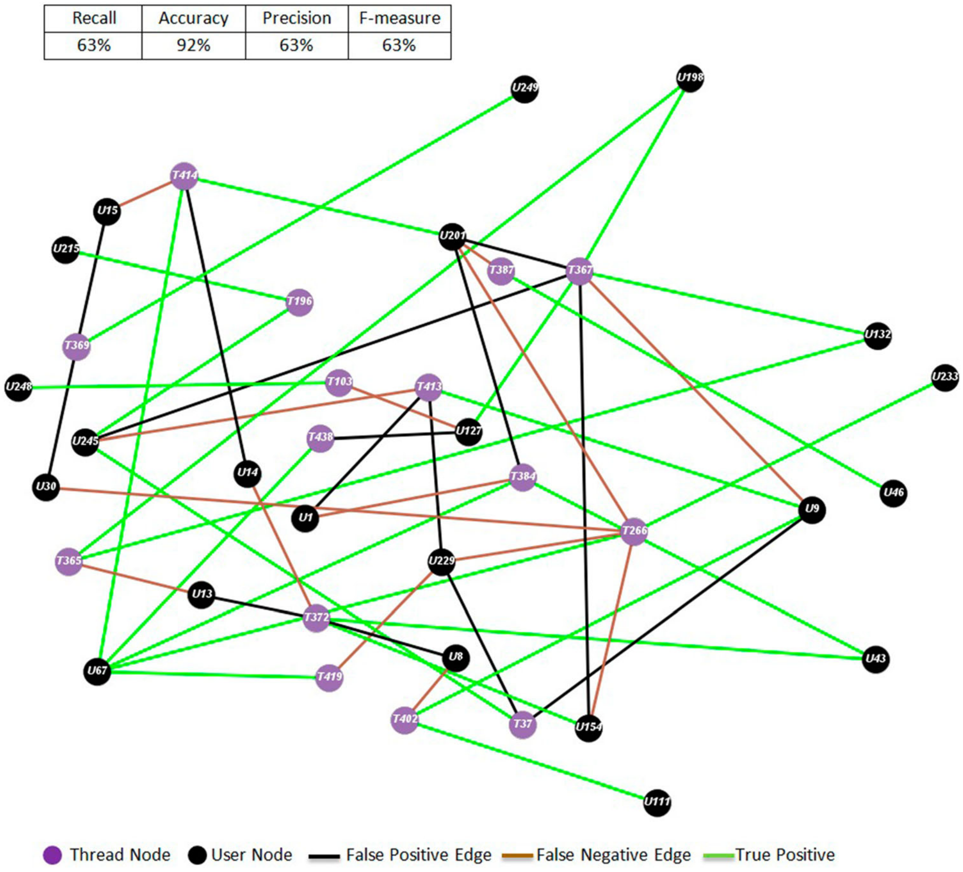
Table 8 Detailed *F*-measure results of the SVM approach *per* testing month and sub-forum

Month	2	3	4	5	6	7	8	9	10	11	12	13	Mean	Max	Min
Sub-forums															
SF 2	0.21	0.19	0.11	0.12	0.11	0.13	0.11	0.17	0.15	0.18	0.11	0.15	**0.145**	**0.21**	**0.11**
SF 3	0.10	0.05	0.11	0.09	0.10	0.15	0.12	0.13	0.16	0.19	0.11	0.13	**0.12**	**0.19**	**0.05**
SF 4	0.18	0.22	0.28	0.38	***	0.33	0.22	0.19	0.23	0.25	0.28	0.18	**0.25**	**0.38**	**0.18**
SF 5	0.11	0.13	0.15	0.11	0.11	0.10	0.11	0.13	0.17	0.26	0.12	0.16	**0.14**	**0.22**	**0.11**
SF 6	0.39	0.31	0.45	0.31	***	0.25	0.61	0.33	0.39	0.33	0.63	0.27	**0.39**	**0.63**	**0.25**

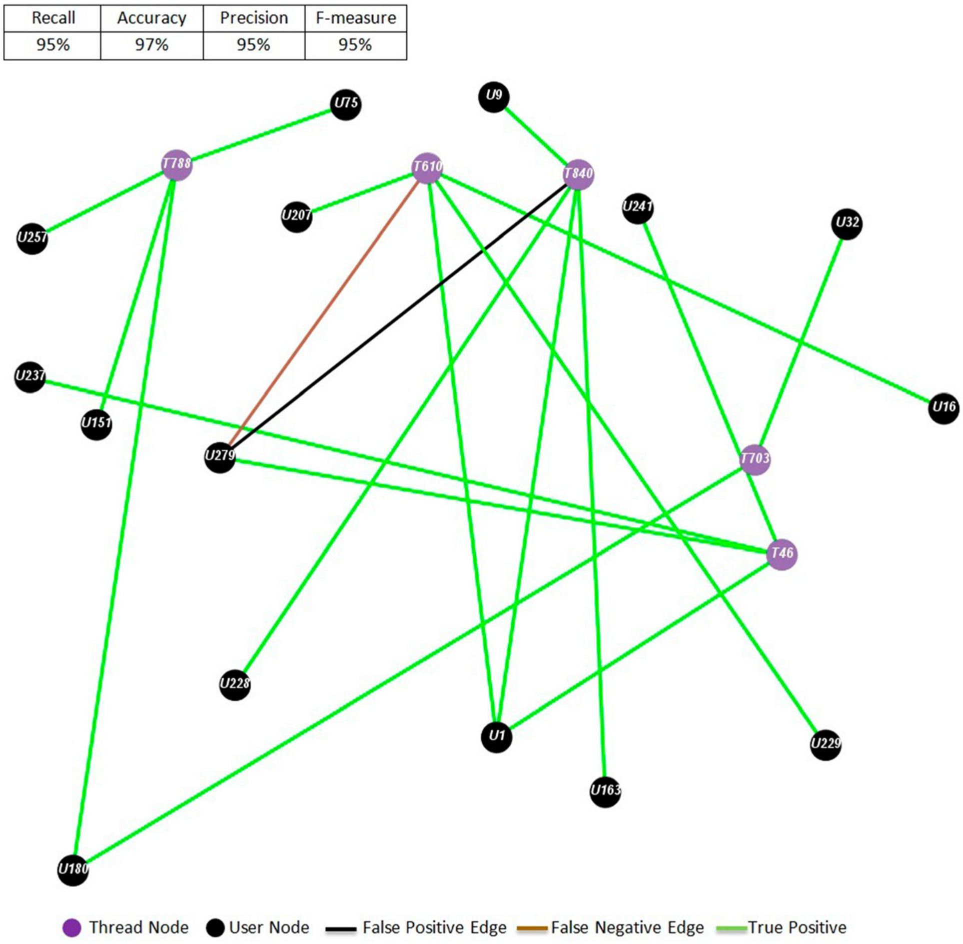
Bold values correspond to summary values, either total or first order statistics, mean, min and max values

****entries correspond to non convergent computation processes, i.e. we do not reach a final value

Fig. 8 Example of middle performance result corresponding to the post publication graph of SF 4 for Month



Recall	Accuracy	Precision	F-measure
63%	92%	63%	63%



Recall	Accuracy	Precision	F-measure
95%	97%	95%	95%

Fig. 9 Best predictive performance corresponding to post publication graph of SF 6 for Month 10

Table 9 Post publication decision rules for SF4-M4

User	Posts in:	User	Posts in:	User	Posts in:
U1	T384, T413	U46	T387	U215	T196
U8	T372, T402	U67	T266, T384, T414, T419, T438	U229	T37, T266, T413, T419
U9	T37, T367, T402, T413	U111	T402	U233	T266
U13	T365, T372	U127	T103, T367, T438	U245	T37, T196, T367, T413
U14	T372, T414	U132	T365, T367	U248	T103
U15	T369, T414	U154	T266, T367, T372	U249	T369
U30	T266, T369	U198	T365		
U43	T372, T384	U201	T266, T367, T384, T387, T414		

User = U**, conversation threads the user has published posts in = T***

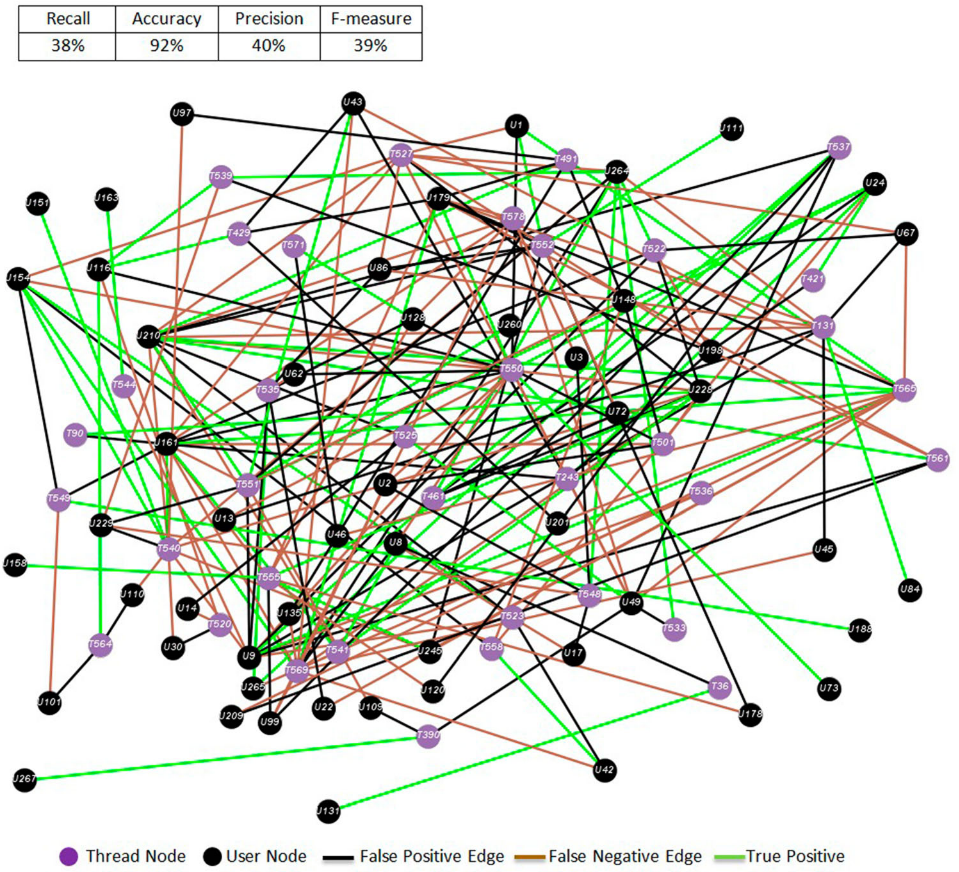
data are composed of similar feature vectors from the remaining time periods. We have tested two well know algorithms using conventional implementations provided in Matlab. First, a random forest (RF) with 101 individual trees. Secondly, a linear support vector machine (SVM).

Tables 7 and 8 give the detailed *F*-measure results for the RF and SVM. The overall average of the *F*-measure of the RF and SVM predictors over all sub-forum experiments are 0.19 and 0.21, respectively, far below the average result achieved by our ELCA approach (0.61). The best *F* score result for a specific month and sub-forum of ELCA (0.95) is far above that of RF (0.60) and SVM (0.63). A one sided Wilkoxon’s rank sum test comparing the entries of Table 6

Table 10 Post publication decision rules for SF6-M10

User	Posts in:	User	Posts in:	User	Posts in:
U1	T46, T610, T840	U151	T788	U229	T610
U9	T840	U163	T840	U237	T46
U16	T610	U180	T703, T788	U241	T46
U32	T703	U207	T610	U257	T788
U75	T788	U228	T840	U279	T46, T610, T840

User = U**, conversation threads the user has published posts in = T***



Recall	Accuracy	Precision	F-measure
38%	92%	40%	39%

Fig. 10 Worst result corresponding to publication graph of sub-forum 5 for Month 6

against Tables 7 and 8 confirms that the superiority of the ELCA model is extremely significative ($p < 1e{-}16$).

4.2 Discussion

For a qualitative appreciation of the results, Figs. 8 and 9 show the graph representations of the content publication predictions for sub-forum 4 at month 4 and sub-forum 6 at month 10, where violet and black nodes correspond to threads and users, respectively. Green edges correspond to the content contributions that the ELCA simulation predicted correctly, black edges are false positives, and brown edges correspond to false negatives. Tables 9 and 10 display the content publishing rules derived from the ELCA simulation. We can notice that most of the network edges are green and that there is approximately the same amount of predicted edges and ground truth edges, which is a very important structural property we must comply with. There are few false positives compared to the large number of non-existing links. This is the reason for the high values of the accuracy performance measure in Table 5 relative to the other measures which only take into account the true positives. We recall from Table 3 that our sub-forum datasets can be considered as very imbalanced two class datasets if we aim to predict the links between users and threads. It is well known, that most classifiers are biased towards the majority class (here the non-existing links). Undersampling the majority class or over-sampling the minority class are proposed as means to improve the

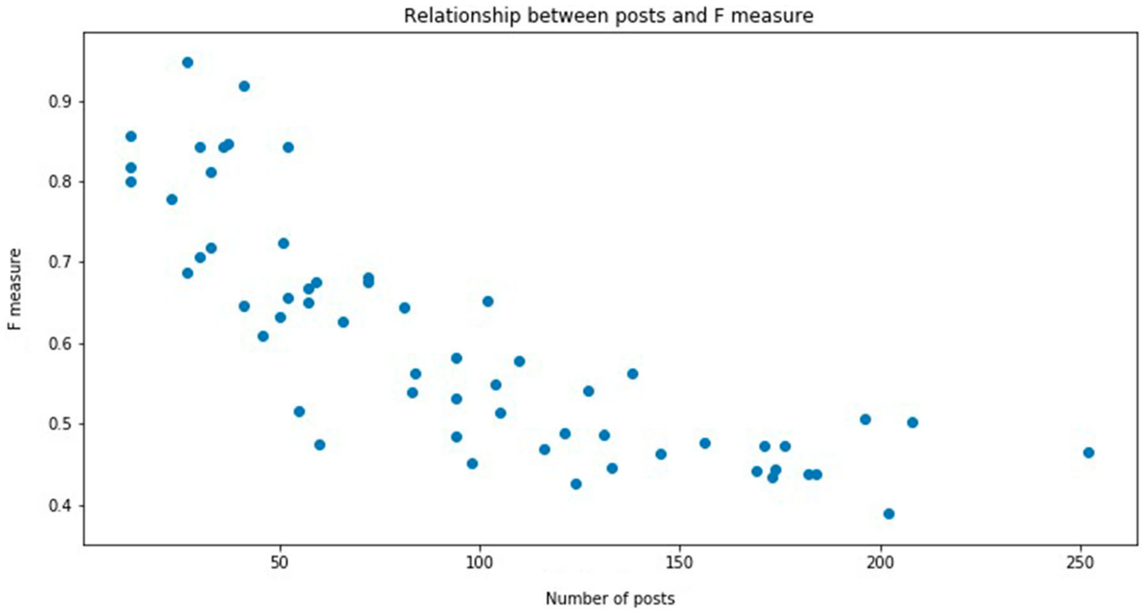


Fig. 11 Relationship between number of posts and *F*-measure score

performance on the minority class, however it is not clear how to carry out these procedures over our sub-forum data.

We get the best results in terms of *F* measure for sub-forum 6. It seems that the lower number of posts allows a more efficient semantic analysis and makes it easier for the model to find the threads a user finds interest in. A relevant observation is that as the number of posts increases in a sub-forum, the predictive results worsen. A qualitative interpretation is that it becomes harder to predict whether a user will post to a thread based on the semantic description of the content because it is contaminated with spurious unfiltered messages. In Fig. 10 we show the network graph corresponding to the month and sub-forum with worst performance results. We notice a large number of false positives. This led us to investigate further, so in Fig. 11 we show the scatter plot of the number of posts made in a unit period of time (month) versus the *F* measure score achieved by the neuro-semantic model in the same period. It appears that as the number of posts increases, the performance of ELCA model prediction decreases. As before, our interpretation is that the cause of this decrease is the increased heterogeneity of the semantic content in the thread, which becomes very noisy.

A way in which we could enhance the neuro-semantic model is to incorporate a discrimination behavior for users that will filter out posts that differ too much with the user semantic preference vector [41]. If we consider the temporal behavior of the *F* measure results within a sub-forum, the scores do not deviate much from the mean value, hence the LCA model is very robust in terms of temporal decay. We associate this behavior with parameter a. In this research, we set the value of $a = 50$ without further search for an optimal setting. However, this parameter could also be optimized by the GA approach.

5 Conclusions

This paper presents a neuro-semantic model of the content publication decisions of users in a web forum OSN at the microscopic level, i.e. the model predicts the specific decision of a user to post a message in a specific conversation thread of a sub-forum. We propose an extended leaky competition accumulator (ELCA) neural model that implements the competition of the diverse threads for the attention of the user as a dynamical process. Model parameter estimation was carried out by a genetic algorithm optimization process. To our knowledge, this is the first work where LCA parameters are estimated from data obtained from a social network content generation prediction in order to achieve optimal predictive performance. The revised literature contains rough qualitative settings of the parameters in order to study the emergent behavior according to theories of value based choice. On the other hand, we have not detected some well known choice phenomena like the preference reversals. More in detail analysis might uncover such phenomena in our problem domain.

Semantic similarity underlaying the attention mechanism is modeled by unsupervised topic analysis, thus it is fully automated. Results over the data extracted from a real life OSN are quite promising. Specifically the ELCA model improves greatly over standard machine learning approaches, namely random forest (RF) and support vector machines (SVM), using the same kind of semantic information as input features. Best and average F score of ELCA was 0.95 and 0.61, respectively, while for the RF and SVM best F score was 0.60 and 0.63, respectively, and the average F score was 0.19 and 0.21, respectively. The fundamental research into the likelihood maximization approaches to LCA parameter estimation is a priority for future works.

Further work will be directed to a deeper exploration into the fundamentals of Natural Language Processing (NLP) algorithms in order to improve the capture of the real meaning of the posted text documents, overcoming frequentist approaches to model the joint occurrence of words in a document [13]. Automatic ontology creation for a specific domain is a promising approach to tackle this problem. We will explore word embeddings as a very powerful modeling approach at the expense of interpretability.

Finally, another quite exciting research area is topic space metrics. Future work could be addressed to the definition of an adequate distance between multi-topic text vector representations allowing the extraction of the most valuable content generated by users. Besides, the approach developed in this work could be combined with other existing methods that capture topological features of the network looking for an improvement in prediction performance by such a hybrid system.

Acknowledgements This work wouldn't be possible without Plexilandia Community, therefore, we would like to thank them; especially, Mr. Jose Ignacio Santa-Cruz, who has provided always the right information; besides, his good observations on the day-to-day community's behavior.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was partially supported by the Business Intelligence Research Center (www.ceine.cl) and CONICYT (CONICYT-PCHA/MagÃsterNacional/2015-22151700, CONICYT-PFCHA/Doctorado Nacional/2019-21190971). This project has received funding from the European Union's Horizon 2020 388 research and innovation programme under the Marie Sklodowska-Curie grant agreement 777720.

Availability of data and material (data transparency) Data used for the computational experiments will be available in zenodo.org after paper acceptance for publication.

Code availability Specific code for the computational pipeline will be publish with the data in zenodo.orf after paper acceptance for publication.

Declarations

Conflict of interest Authors declare that they do not have conflict of interests.

References

1. Al-Obeidat F, Adedugbe O, Hani AB, Benkhelifa E, Majdalawieh M (2020) Cone-KG: a semantic knowledge graph with news content and social context for studying covid-19 news articles on social media. In: 2020 Seventh international conference on social networks analysis, management and security (SNAMS), pp 1–7. https://doi.org/10.1109/SNAMS52053.2020.9336541
2. Aldarwish MM, Ahmad HF (2017) Predicting depression levels using social media posts. In: 2017 IEEE 13th International symposium on autonomous decentralized system (ISADS), pp 277–280. https://doi.org/10.1109/ISADS.2017.41
3. Alon N, Feldman M, Procaccia AD, Tennenholtz M (2010) A note on competitive diffusion through social networks. Inf Process Lett 110(6):221–225. https://doi.org/10.1016/j.ipl.2009.12.009
4. AlSumait L, Barbará D, Domeniconi C (2008) On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Eighth IEEE international conference on data mining, 2008. ICDM'08. IEEE, pp 3–12. https://doi.org/10.1109/ICDM.2008.140
5. AlSuwaidan L, Ykhlef M (2017) A novel information diffusion model for online social networks. In: Proceedings of the 19th international conference on information integration and web-based applications and services, iiWAS '17. Association for Computing Machinery, New York, pp 116–120. https://doi.org/10.1145/3151759.3151812
6. Alvarez H, Ríos SA, Aguilera F, Merlo E, Guerrero L (2010) Enhancing social network analysis with a concept-based text mining approach to discover key members on a virtual community of practice. In: International conference on knowledge-based and intelligent information and engineering systems. Springer, Berlin, pp 591–600. https://doi.org/10.1007/978-3-642-15390-7_61
7. Binney RJ, Ramsey R (2020) Social semantics: the role of conceptual knowledge and cognitive control in a neurobiological model of the social brain. Neurosci Biobehav Rev 112:28–38. https://doi.org/10.1016/j.neubiorev.2020.01.030
8. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
9. Bogacz R, Usher M, Zhang J, McClelland JL (2007) Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. Philos Trans R Soc Lond B Biol Sci 362(1485):1655–1670. https://doi.org/10.1098/rstb.2007.2059

10. Brown PE, Feng J (2011) Measuring user influence on twitter using modified k-shell decomposition. In: Fifth international AAAI conference on weblogs and social media
11. Cascetta E (2009) Random utility theory. Springer US, Boston, pp 89–167. https://doi.org/10.1007/978-0-387-75857-2_3
12. Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the flickr social network. In: ACM (ed) WWW 2009, April 20–24, 2009, Madrid, Spain, pp 721–730
13. Contreras-Piña C, Ríos SA (2016) An empirical comparison of latent sematic models for applications in industry. Neurocomputing 179:176–185. https://doi.org/10.1016/j.neucom.2015.11.080
14. Cuadra L, Rios SA, L'Huillier G (2011) Enhancing community discovery and characterization in VCoP using topic models. In: Proceedings of the 2011 IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology—volume 03. IEEE Computer Society, pp 326–329. https://doi.org/10.1109/WI-IAT.2011.97
15. De Maio C, Fenza G, Loia V, Orciuoli F (2017) Unfolding social content evolution along time and semantics. Future Gener Comput Syst 66:146–159. https://doi.org/10.1016/j.future.2016.05.039
16. Demergis D (2019) Predicting Eurovision song contest results by interpreting the tweets of Eurovision fans. In: 2019 Sixth international conference on social networks analysis, management and security (SNAMS), pp 521–528. https://doi.org/10.1109/SNAMS.2019.8931875
17. Dhiman A, Toshniwal D (2020) An approximate model for event detection from twitter data. IEEE Access 8:122168–122184. https://doi.org/10.1109/ACCESS.2020.3007004
18. Feng Y, Bai B, Chen W (2015) Information diffusion efficiency in online social networks. In: 2015 IEEE International conference on digital signal processing (DSP), pp 1138–1142. https://doi.org/10.1109/ICDSP.2015.7252057
19. Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038
20. Goldenberg J (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth. Mark Lett 12:211–223
21. Goyal P, Kaushik P, Gupta P, Vashisth D, Agarwal S, Goyal N (2020) Multilevel event detection, storyline generation, and summarization for tweet streams. IEEE Trans Comput Soc Syst 7(1):8–23. https://doi.org/10.1109/TCSS.2019.2954116
22. Grabowicz P, Ganguly N, Gummadi K (2015) Microscopic description and prediction of information diffusion in social media: quantifying the impact of topical interests. In: Proceedings of the 24th international conference on world wide web, WWW '15 Companion. Association for Computing Machinery, New York, pp 621–622. https://doi.org/10.1145/2740908.2744106
23. Granovetter M (1978) Threshold models of collective behavior. Am J Sociol 83(6):1420–1443
24. Grzeça M, Becker K, Galante R (2020) Drink2Vec: improving the classification of alcohol-related tweets using distributional semantics and external contextual enrichment. Inf Process Manag 57(6):102369. https://doi.org/10.1016/j.ipm.2020.102369
25. Hu H, Wang X (2009) Evolution of a large online social network. Phys Lett A 373(12–13):1105–1110. https://doi.org/10.1016/j.physleta.2009.02.004
26. Hu Y, Song RJ, Chen M (2017) Modeling for information diffusion in online social networks via hydrodynamics. IEEE Access 5:128–135. https://doi.org/10.1109/ACCESS.2016.2605009
27. Jiang C, Chen Y, Liu KR (2014) Modeling information diffusion dynamics over social networks. In: 2014 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1095–1099. https://doi.org/10.1109/ICASSP.2014.6853766
28. Jiang H, Sun L, Ran J, Bai J, Yang X (2020) Community detection based on individual topics and network topology in social networks. IEEE Access 8:124414–124423. https://doi.org/10.1109/ACCESS.2020.3005935
29. Kao LJ, Huang YP (2015) Mining influential users in social network. In: 2015 IEEE International conference on systems, man, and cybernetics (SMC). IEEE, pp 1209–1214
30. Keitemoge P (2018) Technology threats: impacts of cyberbullying to today's generation. In: 2018 15th International conference on service systems and service management (ICSSSM), pp 1–6. https://doi.org/10.1109/ICSSSM.2018.8464953
31. Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '03. ACM, New York, pp 137–146. https://doi.org/10.1145/956750.956769
32. Khaled A, Ouchani S, Chohra C (2019) Recommendations-based on semantic analysis of social networks in learning environments. Comput Hum Behav 101:435–449. https://doi.org/10.1016/j.chb.2018.08.051
33. Khan Z, Iltaf N, Afzal H, Abbas H (2020) DST-HRS: a topic driven hybrid recommender system based on deep semantics. Comput Commun 156:183–191. https://doi.org/10.1016/j.comcom.2020.02.068
34. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888. https://doi.org/10.1038/nphys1746
35. Kubo M, Naruse K, Sato H, Matubara T (2007) The possibility of an epidemic meme analogy for web community population analysis. In: International conference on intelligent data engineering and automated learning. Springer, Berlin, pp 1073–1080. https://doi.org/10.1007/978-3-540-77226-2_107
36. Kumar R (2012) Blending roulette wheel selection and rank selection in genetic algorithms. Int J Mach Learn Comput 2(4):365. https://doi.org/10.7763/IJMLC.2012.V2.146
37. Kumar S, Saini M, Goel M, Panda BS (2020) Modeling information diffusion in online social networks using a modified forest-fire model. J Intell Inf Syst. https://doi.org/10.1007/s10844-020-00623-8
38. Lally P, Van Jaarsveld CH, Potts HW, Wardle J (2010) How are habits formed: modelling habit formation in the real world. Eur J Soc Psychol 40(6):998–1009. https://doi.org/10.1002/ejsp.674
39. L'Huillier G, Alvarez H, Ríos SA, Aguilera F (2011) Topic-based social network analysis for virtual communities of interests in the dark web. SIGKDD Explor Newsl 12(2):66–73. https://doi.org/10.1145/1964897.1964917
40. Li D, Zhang S, Sun X, Zhou H, Li S, Li X (2017) Modeling information diffusion over social networks for temporal dynamic prediction. IEEE Trans Knowl Data Eng 29(9):1985–1997. https://doi.org/10.1109/TKDE.2017.2702162
41. Li L, Scaglione A, Swami A, Zhao Q (2012) Phase transition in opinion diffusion in social networks. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 3073–3076. https://doi.org/10.1109/ICASSP.2012.6288564
42. Li M, Wang X, Gao K, Zhang S (2017) A survey on information diffusion in online social networks: models and methods. Information 8:118
43. Liu B, Xu Z, Sun C, Wang B, Wang X, Wong DF, Zhang M (2018) Content-oriented user modeling for personalized response ranking in chatbots. IEEE/ACM Trans Audio Speech Lang Process 26(1):122–133. https://doi.org/10.1109/TASLP.2017.2763243

44. Liu W, Luo X, Gong Z, Xuan J, Kou NM, Xu Z (2016) Discovering the core semantics of event from social media. Future Gener Comput Syst 64:175–185. https://doi.org/10.1016/j.future.2015.11.023
45. Lo CF, Ip HY (2021) Modified leaky competing accumulator model of decision making with multiple alternatives: the lie-algebraic approach. Sci Rep 11(1):10923. https://doi.org/10.1038/s41598-021-90356-7
46. Luo C, Chen A, Cui B, Liao W (2021) Exploring public perceptions of the covid-19 vaccine online from a cultural perspective: semantic network analysis of two social media platforms in the United States and China. Telemat Inform 65:101712. https://doi.org/10.1016/j.tele.2021.101712
47. Luo C, Zheng X, Zeng D (2015) Inferring social influence and meme interaction with Hawkes processes. In: 2015 IEEE International conference on intelligence and security informatics (ISI). IEEE, pp 135–137. https://doi.org/10.1109/ISI.2015.7165953
48. McClelland JL (1993) Toward a theory of information processing in graded, random, and interactive networks. The MIT Press, Cambridge, pp 655–688
49. Miletić S, Turner BM, Forstmann BU, van Maanen L (2017) Parameter recovery for the leaky competing accumulator model. J Math Psychol 76:25–50. https://doi.org/10.1016/j.jmp.2016.12.001
50. Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63(2):81. https://doi.org/10.1037/h0043158
51. Mühlenbein H, Schlierkamp-Voosen D (1993) Predictive models for the breeder genetic algorithm I. Continuous parameter optimization. Evol Comput 1(1):25–49. https://doi.org/10.1162/evco.1993.1.1.25
52. Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 33–41. https://doi.org/10.1145/2339530.2339540
53. Nagaya H, Uno K, Torii HA (2019) Tracking topics of influential tweets on fukushima disaster over long periods of time. In: 2019 International conference on data mining workshops (ICDMW), pp 13–16. https://doi.org/10.1109/ICDMW.2019.00010
54. Niu J, Peng J, Shu L, Tong C, Liao W (2013) An empirical study of a Chinese online social network–Renren. Computer 46(9):78–84. https://doi.org/10.1109/MC.2013.1
55. Osho A, Goodman C, Amariucai G (2020) MIDMod-OSN: a microscopic-level information diffusion model for online social networks. arXiv:2002.10522
56. Pandya A, Oussalah M, Monachesi P, Kostakos P (2020) On the use of distributed semantics of tweet metadata for user age prediction. Future Gener Comput Syst 102:437–452. https://doi.org/10.1016/j.future.2019.08.018
57. Phang XH, Nguyen C (2008) Gibbslda++. http://gibbslda.sourceforge.net/
58. Qi J, Liang X, Wang Y, Cheng H (2018) Discrete time information diffusion in online social networks: micro and macro perspectives. Sci Rep 8(1):11872. https://doi.org/10.1038/s41598-018-29733-8
59. Qiu X, Zhao L, Wang J, Wang X, Wang Q (2016) Effects of time-dependent diffusion behaviors on the rumor spreading in social networks. Phys Lett A 380(24):2054–2063. https://doi.org/10.1016/j.physleta.2016.04.025
60. Reeves CR (1994) Genetic algorithms and neighbourhood search. In: AISB workshop on evolutionary computing. Springer, Berlin, pp 115–130. https://doi.org/10.1007/3-540-58483-8_10
61. Ríos SA (2007) A study on web mining techniques for off-line enhancements of web sites. Ph.D. thesis, Tokio Unversity
62. Ríos SA, Aguilera F, Guerrero LA (2009) Virtual communities of practice's purpose evolution analysis using a concept-based mining approach. In: International conference on knowledge-based and intelligent information and engineering systems. Springer, Berlin, pp 480–489. https://doi.org/10.1007/978-3-642-04592-9_60
63. Ríos SA, Aguilera F, Nuñez-Gonzalez JD, Graña M (2017) Semantically enhanced network analysis for influencer identification in online social networks. Neurocomputing. https://doi.org/10.1016/j.neucom.2017.01.123
64. Ríos SA, Muñoz R (2014) Content patterns in topic-based overlapping communities. Sci World J 2014:11. https://doi.org/10.1155/2014/105428
65. Román PE, Gutiérrez ME, Rios SA (2012) A model for content generation in on-line social network. In: KES, pp 756–765. https://doi.org/10.3233/978-1-61499-105-2-756
66. Sagduyu YE, Grushin A, Shi Y (2018) Synthetic social media data generation. IEEE Trans Comput Soc Syst 5(3):605–620. https://doi.org/10.1109/TCSS.2018.2854668
67. Saito K, Ohara K, Kimura M, Motoda H (2013) Detecting changes in content and posting time distributions in social media. In: 2013 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM 2013), pp 572–578. https://doi.org/10.1145/2492517.2492618
68. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620. https://doi.org/10.1145/361219.361220
69. Saxena A, Iyengar S, Gupta Y (2015) Understanding spreading patterns on social networks based on network topology. In: 2015 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 1616–1617. https://doi.org/10.1145/2808797.2809360
70. Shukla A, Pandey HM, Mehrotra D (2015) Comparative review of selection techniques in genetic algorithm. In: 2015 International conference on futuristic trends on computational analysis and knowledge management (ABLAZE). IEEE, pp 515–519. https://doi.org/10.1109/ABLAZE.2015.7154916
71. Small L, Mason O (2013) Information diffusion on the iterated local transitivity model of online social networks. Discrete Appl Math 161(10–11):1338–1344. https://doi.org/10.1016/j.dam.2012.10.029
72. Song X, Chi Y, Hino K, Tseng BL (2007) Information flow modeling based on diffusion rate for prediction and ranking. In: Proceedings of the 16th international conference on world wide web, WWW '07. ACM, New York, pp 191–200. https://doi.org/10.1145/1242572.1242599
73. Srinivas M, Patnaik LM (1994) Genetic algorithms: a survey. Computer 27(6):17–26. https://doi.org/10.1109/2.294849
74. Sun Y, Liu C, Zhang CX, Zhang ZK (2014) Epidemic spreading on weighted complex networks. Phys Lett A 378(7–8):635–640. https://doi.org/10.1016/j.physleta.2014.01.004
75. Tope Omitola Ríos Sebastián JB (2015) Social semantic web intelligence. Morgan & Claypool Publishers, San Rafael
76. Tsetsos K, Gao J, McClelland J, Usher M (2012) Using time-varying evidence to test models of decision dynamics: bounded diffusion versus the leaky competing accumulator model. Front Neurosci 6:79. https://doi.org/10.3389/fnins.2012.00079
77. Usher M, McClelland JL (2001) The time course of perceptual choice: the leaky, competing accumulator model. Psychol Rev 108(3):550. https://doi.org/10.1037/0033-295X.108.3.550
78. Woo J, Chen H (2012) An event-driven sir model for topic diffusion in web forums. In: 2012 IEEE International conference on intelligence and security informatics (ISI). IEEE, pp 108–113. https://doi.org/10.1109/ISI.2012.6284101
79. Woo J, Chen H (2016) Epidemic model for information diffusion in web forums: experiments in marketing exchange and political

dialog. SpringerPlus 5(1):66. https://doi.org/10.1186/s40064-016-1675-x

80. Woo J, Son J, Chen H (2011) An sir model for violent topic diffusion in social media. In: 2011 IEEE International conference on intelligence and security informatics (ISI). IEEE, pp 15–19. https://doi.org/10.1109/ISI.2011.5984043
81. Xiong F, Liu Y, Zhang Zj, Zhu J, Zhang Y (2012) An information diffusion model based on retweeting mechanism for online social media. Phys Lett A 376(30–31):2103–2108. https://doi.org/10.1016/j.physleta.2012.05.021
82. Zhao S, Yu L, Cheng B (2017) Probabilistic community using link and content for social networks. IEEE Access 5:27189–27202. https://doi.org/10.1109/ACCESS.2017.2774798