Future Generation Computer Systems 106 (2020) 92–104

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers

Gonzalo A. Ruz [a,b,*], Pablo A. Henríquez [a], Aldo Mascareño [c,d]

[a] Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Santiago, Chile
[b] Center of Applied Ecology and Sustainability (CAPES), Santiago, Chile
[c] Centro de Estudios Públicos, Santiago, Chile
[d] Escuela de Gobierno, Universidad Adolfo Ibáñez, Santiago, Chile

ARTICLE INFO

Article history:
Received 4 February 2019
Received in revised form 19 December 2019
Accepted 4 January 2020
Available online 8 January 2020

Keywords:
Bayesian network classifiers
Twitter data
Sentiment analysis
Bayes factor
Support vector machines
Random forests

ABSTRACT

Sentiment analysis through machine learning using Twitter data has become a popular topic in recent years. Here we address the problem of sentiment analysis during critical events such as natural disasters or social movements. We consider Bayesian network classifiers to perform sentiment analysis on two datasets in Spanish: the 2010 Chilean earthquake and the 2017 Catalan independence referendum. In order to automatically control the number of edges that are supported by the training examples in the Bayesian network classifier, we adopt a Bayes factor approach for this purpose, yielding more realistic networks. The results show the effectiveness of using the Bayes factor measure as well as its competitive predictive results when compared to support vector machines and random forests, given a sufficient number of training examples. Also, the resulting networks allow to identify the relations amongst words, offering interesting qualitative information to historically and socially comprehend the main features of the event dynamics.

1. Introduction

In recent years, the production of textual documents in social media has increased exponentially. For instance, up to July 2018, Twitter had 326 million active users sending more than 500 million tweets per day.[1] Social media creates virtual bonds between users, where people express opinions and develop relationships through posts, comments, messages, and likes. Social media allow people to share their thoughts, feelings, and opinions with others instantly and easily [1].

Twitter is a fast growing online platform where people can create, post, update, and read short text messages called *tweets*. The Twitter platform may even indirectly influence traditional media agenda setting particularly in critical events, as journalists gather information from tweets and retweet valuable messages shared by users [2]. The scientific study of the semantic content of these tweets is called *sentiment analysis* [3]. In general, sentiment analysis is a method for identifying and categorizing the polarity of a given text [4], where the goal is to determine whether a particular document has either a positive or a negative value according to a standard categorization [5].

Sentiment classification is not a recent development. The seminal works on sentiment analysis were carried out by Pang et al. [6] and Turney [7]. They proposed the approaches typically used in sentiment analysis implementations, whatever the source of the textual data.

Sentiment analysis has been traditionally tackled as a classification task (supervised learning) where the user decides which classification algorithm to use. Support Vector Machines (SVM) is one of the most popular classifiers. In Table 1 we present a review of the most commonly used algorithms for sentiment classification. From this review, we notice that most algorithms can be considered as black box models that make it difficult to understand how the words (features) interact during the classification process. In this paper, we analyze the performance of Bayesian network classifiers [8], which are probabilistic graphical models that effectively combine the quantitative aspect of the classification task with a qualitative dimension constructed by the probabilistic relationships among the attributes.

Understanding the qualitative dimension of Twitter communication is particularly useful in critical situations not only for scientific or methodological purposes. Major critical events such as natural disasters (earthquakes, floods, hurricanes, wildfires)

* Corresponding author at: Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Santiago, Chile.
E-mail addresses: gonzalo.ruz@uai.cl (G.A. Ruz), pablo.henriquez.v@edu.uai.cl (P.A. Henríquez), amascareno@cepchile.cl (A. Mascareño).

[1] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

https://doi.org/10.1016/j.future.2020.01.005
0167-739X/

Table 1
Comparison of sentiment analysis approaches.

Paper	Description of datasets	Machine learning techniques	Languages	Reported accuracy (%)
Singh et al. [9]	The dataset 1 comprises of 1000 positive and 1000 negative processed reviews. The dataset 2 comprises of 700 positive and 700 negative processed reviews. The third dataset is of 1000 reviews of Hindi movies.	NB, SVM	English	81.14
Zhu et al. [10]	Positive blogs are 100, which is equal to the size of negative blogs.	SVM	Chinese	62.90
Tan and Zhang [11]	The total size is 1021 documents that consist of three domains: education, movie, and house.	SVM, NB, k-NN	Chinese	82
Henríquez and Ruz [12]	The dataset 1 contains a collection of 10 000 tweets from the Catalan referendum of 2017. The dataset 2 contains a collection of 2187 tweets from the Chilean earthquake of 2010.	RVFL	Spanish	82.90
Al-Ayyoub et al. [13]	The dataset contains 300 tweets positive, 300 tweets negative and 300 tweets neutral.	SVM	Arabic	86.89
Ankit and Saleena [14]	The dataset consists of 43 532 negative and 56 457 positive tweets.	NB, RF, SVM	English	75.81
Boiy and Moens [15]	The corpus contains 2000 movie reviews, that are characterized by a varied vocabulary with a typical mix of words that describe the storyline and words that describe the evaluation of that particular movie.	SVM, NB	English	86.35
Ghorbel and Jacot [16]	The corpus has 2000 French movie reviews, 1000 positive and 1000 negative, from 10 movies, 1600 were used for training and 400 for testing.	SVM	French	93.25
Melville et al. [17]	The data consists of 1000 positive and 1000 negative reviews from the internet movie database.	NB	English	81.42
Wang et al. [18]	The dataset contains about 0.6 million tweets which were collected in one week period from Twitter.	SVM	English	84.13
Gamon [19]	The dataset consists of 11 399 feedback items from a Global Support Services survey, and 29 485 feedback items from a Knowledge Base survey for a total of 40 884 items.	SVM	English	77.5
Pang and Lee [20]	The dataset contains four corpora of movie reviews. All of the 1770, 902, 1307, or 1027 documents in a given corpus were written by the same author.	SVM, regression	English	66.3
Pang et al. [6]	The corpus has 752 negative and 1301 positive reviews (movie reviews.), with a total of 144 reviewers represented.	NB, SVM, maximum entropy	English	82.9
Prabowo and Thelwall [21]	One of the datasets consists of 1000 positive and 1000 negative reviews from the internet movie database. The authors analyzed three movie datasets. The last dataset was extracted from MySpace.	SVM	English	87.30
Annett and Kondrak [22]	Cornell Movie Review Dataset of Tagged Blogs (1000 positive and 1000 negative).	SVM, NB	English	77.5
Mullen and Collier [23]	The first dataset consisted of a total of 1380 Epinions.com movie reviews, approximately half positive and half negative. The second dataset consists of 100 record reviews.	Hybrid SVM	English	89

or political transitions (independence movements, coups, revolutions, terrorist attacks) are certainly disturbing for individuals, communities or large regions, especially during the core of the event. Despite opinions against [24], the high volume, ubiquity, and real-time nature of Twitter communication in those events, turn Twitter into a powerful source of information that may contribute to organize the allocation of human and material resources, alleviate subsequent impacts, and prevent other people from accessing to damaged infrastructure and affected zones [25].

Analysis of emotions during crisis situations is a complicated matter. Critical events are characterized by the experience of trespassing a threshold, either individually or socially, to a new, yet unknown state, [26] thereby producing uncertainty and mixed emotional reactions [27–29]. Analyzing expressions in pager text messages sent during September 11 attacks, [30] have found that positive communications (expressions of solidarity, compassion, and other pro-social feelings) emerge first, followed by negative communications of distress, rejection, and astonishment. Religious sentiments aimed at making sense of the situation (to that extent, also positive feelings) emerge in the third place. As argued by [31], personal orientation towards entities or topics triggers emotional responses expressed in positive or negative opinions that polarize individuals or groups. On that basis, lexicons of emotional words for polarity prediction can be generated, being the manually construed ones the most useful. Further, neutral feelings informing about plain facts can also be detected and analyzed [32]. Sentiment analysis has thus developed into a relevant technique to detect emotional responses to societal events on an aggregate scale, recently contributing to the measurement of collective happiness. Considering happiness (or positivity) as a crucial measure of societal well being, besides classical economic metrics such as income distribution or GDP, [33] develop a 'hedonometer' using Twitter as a data source and an 10 222 word lexicon calibrated by users of Amazon's Mechanical Turk with

50 independent evaluations per word. With this tool, authors are able to identify the dynamics of positive (and negative) emotions regarding different facts as a function of time, space, demographic distribution, and network structure over a given period. The approach has been also applied to climate change sentiments on Twitter [34], public health monitoring [35], the impact of weather dynamics on societal sentiments [36], and the assessment of national stereotypes though linguistic behavior [37].

Critical events, however, are not circumscribed to the distressful core of the situation. We can distinguish between three relevant moments: the incubation or early phases of the problem, the critical event itself or critical transition, and the restructuration after the crisis or recovery [26,38,39]. Combining quantitative and qualitative analysis in the classification of Twitter communication contribute to a more robust knowledge about those three moments and to probabilistic predictions of future developments, detection of emerging critical events, and improvement of disaster management strategies [40,41]. In early phases, some studies argue that critical events such as earthquakes or hurricanes can be detected with high probability and promptly notified by monitoring Twitter [42]. A combination of social media data and geo-location is crucial for these purposes [43]. Twitter is also useful for creating awareness about risks regarding climate change [44], for designing tsunami warning systems and response strategies when coordinated with official Twitter sites and *in situ* Twitter leaders [45,46].

During the core of the crisis, studies have argued that by qualitatively analyzing Twitter communication, we can obtain relevant information about peoples' perceptions, the nature of the critical event, and its visibility. This information is useful for improving disaster response technology [47]. Situational awareness, coordination of efforts, and sentimental trajectories of user's emotions during critical events are also mentioned as relevant contributions of qualitative sentiment classification in Twitter communication [48–50].

In the restructuration phase, the interactivity of Twitter communication and sharing experiences of crisis become crucial for enhancing the actual response to disasters [51]. Expressions of solidarity and the work through emotions that characterizes Twitter communication may also increase the pool of donors and volunteers for critical events turned into humanitarian crises [52]. Sentiment analysis also contributes to detect emotional patterns during crises that may help to develop mental health risk approaches for confronting the psychological consequences after the crisis [53]. Also perceptions about damages to community assets and faith-based related issues communicated via Twitter, such as interactions of mutual support during crises, are relevant to design recovery strategies [54]. Research shows that evaluations performed by representatives of government and local officers regarding their ability in controlling a crisis and the strength of their responses are positively related to the extent of social media being involved [55]. People also expect fast arriving of help after posting a request on a social media site [56].

Considering this, the motivation of our research is twofold. On the one hand, we aim to know whether usual approaches that have been proved to be effective with English texts behave similarly with Spanish tweets, in particular, the use of SVM and random forests (RF). On the other hand, we aim to identify the relations amongst words for sentiment classification in Twitter using Bayesian networks. For this application, we assessed the performance of the algorithms using two Twitter datasets in Spanish: the Chilean earthquake of 2010 and the Catalan independence referendum of 2017.

The remainder of this paper is organized as follows. Section 2 presents the background on the use of classifiers for sentiment analysis. Bayesian network classifiers are introduced in Section 3. A description of the methodology which includes data collection, pre-processing for sentiment analysis, and simulation setup is presented in Section 4. The results are shown in Section 5, while we analyze and discuss these results, as well as pointing out some limitations, in Section 6. Finally, overall conclusions are presented in Section 7.

2. Background

Several methods are available in the literature that use classifiers for Twitter sentiment analysis. The work presented in [57] proposes an approach for sentiment analysis which combines a SVM classifier and a wide range of features like bag-of-word (1-words, 2-words) and part-of-speech features, as well as votes derived from character n-words language models to achieve the final result. The authors concluded that lexical features (1-words, 2-words) produce the best contributions. In [45], the tsunami warnings in Padang Indonesia and reactions among Twitter users have been examined. The authors in [58] showed that some preprocessing techniques improved the classification performance for Twitter sentiment analysis. In [59], a deep learning based sentiment classifier was developed using a word embedding model and a linear machine learning algorithm. Whereas in [60], a topic-enhanced word embedding for Twitter sentiment classification is presented. Improvements to semi-supervised tweet sentiment classification by using unsupervised information was developed in [61]. In [62], the authors proposed a multi-level sentiment-enriched word embedding learning method, which employs a parallel asymmetric neural network to model n-gram, word-level sentiment, and tweet-level sentiment in the learning process. In [63], a probabilistic approach is developed for multi-class sentiment classification by incorporating lexical information and specific grammatical characteristics into the naive Bayes (NB) classifier. Fake news detection is addressed in [64], where a new set of features are proposed and the performance of 5 classifiers are evaluated with random forest and XGBoost (XGB), obtaining the best results based on the area under the ROC curve (AUC) and the F1 score. An unsupervised learning approach called multiclustering logistic approximation (MLA) is presented in [65] to adapt the source-domain training data to the target domain, thereby making better use of massive amount of labeled data from different domains.

Deep learning approaches have become popular recently. Rumors on Twitter are investigated in [66], where a two stage methodology is presented. First, for rumor stance detection a hierarchical long short term memory (LSTM) network is proposed, where the first layer LTSM is used to learn sentence embedding of each tweet, then the output is used into the second layer LSTM network to exploit the sequence information of the reply chain. In the second stage, for rumor veracity prediction, a supervised feature-driven approach is used with the following classifiers: SVM, NB, DT, and MLP. In [67], a stacked ensemble of shallow convolutional neural networks (CNNs) is proposed for tracking moods, emotions, and sentiments of patients expressing intake of medicine in Twitter. Sentiment and sarcasm classification is tackled in [68] via a multitask learning-based framework using a deep neural network, outperforming a state-of-the-art method based on a CNN. Sentiment analysis combining text and video is known as multimodal sentiment analysis. In [69], different deep-learning-based architectures are explored for multimodal sentiment classification, with the bc-LSTM model obtaining the best results. Capsule networks [70], which represent recent improvements to deep learning architectures, in particular CNN, although focused mainly for image classification tasks, have also been used for challenging NLP applications, such as: multi-label

text classification and question answering [71], sentiment classification [72], sentence classification [73], slot filling and intent detection [74], and text classification [75].

While in this work we use words directly in order to interpret the resulting Bayesian networks, word representations or embeddings are common in sentiment analysis. Deep learning approaches described previously are used for this task, for example, an adaptive embedding learning via LSTM for Chinese sentiment analysis [76]. Common word embedding also include: neural network language model (NNLM), log-bilinear language (LBL) model, Collobert and Weston (C&W) model, continuous bag of words (CBOW) model, skip-gram model, order model, and global vectors (GloVe) model. All of these are described and compared in [77].

Recently, researchers have focused on applying sentiment analysis techniques to crises events. [42] investigated the real-time interaction of events such as earthquakes in Twitter and proposed an algorithm to monitor tweets and to detect a target event. To do this, they devised a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. The work presented in [78] consisted in an automated text classification system to effectively classify the data. For this, a manual vocabulary was created taking into consideration the nature of the disaster data. The work in [12] presented Twitter sentiment classification using a non-iterative neural network. The authors considered the emotional polarity of a text as a two-class classification problem (positive and negative). In another study, [79] proposed a hybrid method to classify and segregate the crisis related tweets from the people who were trapped and struggling for survival during disaster situations. The work presented in [48] performed sentiment classification of user posts in Twitter during the Hurricane Sandy and visualized these sentiments on a geographical map centered around the hurricane. The authors showed how users' sentiments change according not only to users locations, but also based on the distance from the disaster.

3. Bayesian network classifiers

Bayesian networks (BNs) were introduced by Judea Pearl [80] as a graphical model that encodes the joint probability distribution of a set of discrete random variables. In particular, a BN is a directed acyclic graph (DAG), where the nodes represent discrete random variables and the probabilistic relations amongst them are represented by the edges of the graph. In the context of classification (supervised learning), BNs can be used together with the Bayes theorem to obtain the posterior probability of the class variable C given an input data point. Formally, let us consider a set of n discrete random variables $\{X_1, X_2, \ldots, X_n\}$ and N training examples $\{x_{1,i}, x_{2,i}, \ldots, x_{n,i}, c_i\}$ with $i = 1, \ldots, N$ and $c_i \in \{1, \ldots, k\}$. A new input data point $\{g_1, g_2, \ldots, g_n\}$ is classified using the following rule

$$C = \underset{c}{\operatorname{argmax}} \quad P(C = c|X_1 = g_1, \ldots, X_n = g_n). \tag{1}$$

The posterior probability can be obtained using the Bayes theorem,

$$P(C = c|X_1 = g_1, \ldots, X_n = g_n) \propto P(C = c)P(X_1 = g_1, \ldots, X_n = g_n|C = c). \tag{2}$$

The first term on the r.h.s. of (2) is called the *a priori* probability, and can be easily estimated by relative frequencies of the class variable values. The second term on the r.h.s. of (2) is called the *likelihood*. This joint probability, conditioned to the class, is more difficult to compute, and therefore, several approaches have been proposed. The most simple approach considers a rather naive assumption that each attribute (variable) is conditionally independent of the rest of the attributes. By adopting this strong assumption, the resulting model is known as the naive Bayes (NB) classifier [81],

$$P(C = c|X_1 = g_1, \ldots, X_n = g_n) \propto P(C = c)\prod_{i=1}^{n} P(X_i = g_i|\pi_i), \tag{3}$$

with π_i being the set of parent nodes of X_i. Under the NB classifier, $\pi_i = \{C\}$.

Instead of imposing this strong independence assumption, an alternative model called the tree augmented naive Bayes (TAN) [8], allows each node to have at the most one parent node in addition to the class variable node. The resulting DAG is a tree, with $n - 1$ edges (without counting the edges from the class node to each attribute). In this way, the conditional independence assumption is dropped, and the posterior probability can be computed by (3), but considering $\pi_i = \{X_j, C\}$ with X_j $(j \neq i)$ the parent node of X_i given by the tree structure. Also, one of the attributes X_{i^*} acts as the root node, therefore $\pi_{i^*} = \{C\}$. The NB and the TAN classifier are the two most popular BN classifiers. Of course, there are many other variations[2] such as the *Markov blanket* of the class variable [83], the K2-attribute selection (K2-AS) algorithm [84], the seminaive Bayes model [85], and the k-dependence Bayesian classifier [86].

The TAN classifier is restricted to learn a tree structure amongst the attributes. In some cases, there is not enough training data to support all the edges in the tree, impacting negatively in the classifier's generalization power. To overcome this restriction, in [87] the *Bayes factor* is employed to explore midway structures between NB and TAN. In particular, given the decomposability of a Bayesian network and the Bayes factor for model selection, the following measure h captures the effect of adding an extra edge (from X_q to X_p) to a NB classifier,

$$h = 2\log_2(n+1) + \sum_{i=1}^{N} \log_2 P(X_p = x_{p,i}|C = c_i) - \log_2 P(X_p = x_{p,i}|X_q = x_{q,i}, C = c_i), \tag{4}$$

with negative values of h indicating that there is enough data to support that extra edge. The cumulative value H_e indicates whether there is enough data to support e edges compared with 0 edges (naive Bayes), given by

$$H_e = \sum_{i=1}^{e} h_i, \tag{5}$$

where h_i is the h value for the ith edge being considered. Therefore, we consider an incremental construction of the TAN classifier, where for each edge that is used for the tree structure in a descending order with respect to its weight (conditional mutual information) we evaluate its effect by computing the h measure. We continue adding edges while $H_e < 0$. If the previous condition holds until $e = n - 1$, then there is sufficient data to support the tree structure, thus resulting in the construction of the TAN classifier. On the other hand, if the condition holds until $e = l$ with $l < n - 1$, then the resulting structure is a forest.

4. Methodology

4.1. Data collection

For this application, we assessed the performance of the proposed algorithm using two Twitter datasets in Spanish: the 2010

[2] For a complete review of BN classifiers please refer to [82].

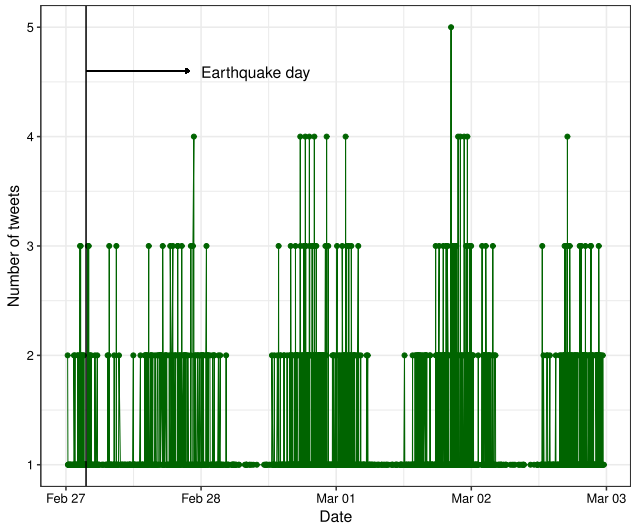

Fig. 1. Time series representing the total number of tweets over the period in the 2010 Chilean earthquake.

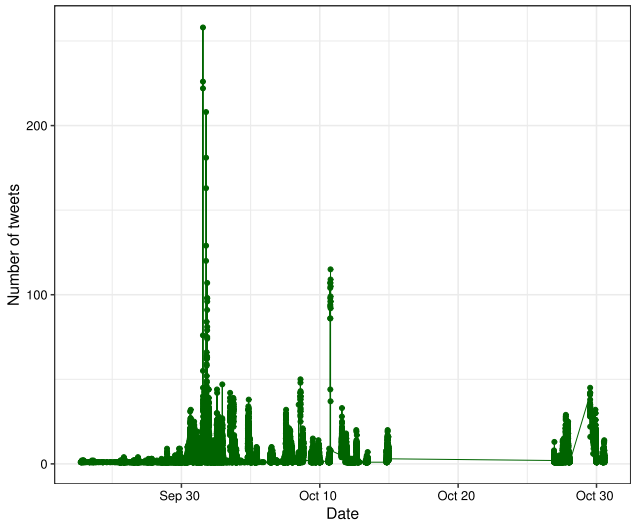

Fig. 2. Time series representing the total number of tweets over the period in the 2017 Catalan independence referendum.

Chilean earthquake and the 2017 Catalan independence referendum. All the tweets were programmatically searched and extracted from Twitter by using twitteR package, written in R programming language. Dataset 1 contains a collection of 2187 tweets from the 2010 Chilean earthquake, which were posted before and after the critical event (2010-02-27 03:34:08). They started at midnight of February 27th and ended at midnight of March 2nd. This dataset was obtained from [88]. For Dataset 2, we collected 60 000 tweets from the 2017 Catalan independence referendum. For this, we used the corresponding keywords for the event: #cataluña, #IndependenciaCatalunya, #2Oct, #CatalanReferedendum, #L6Nenlaencrucijada, and others. This comprehensive tweet search was conducted between 22-09-2017 and 30-10-2017. Figs. 1 and 2 show the tweet count per minute for each dataset.

4.2. Pre-processing

A tweet may contain a lot of opinions, expressed in different ways by different users. The raw data are highly susceptible of inconsistency and redundancy. Preprocessing of tweets is fundamental. In this work, we consider the following procedures:

- Removed all URLs, hashtags (#topic) and targets (@username).
- Removed all punctuations, symbols, numbers.
- Common stop words are removed.
- Removed non-Spanish tweets.
- All the tweets were converted to lower case to make the dataset uniform.
- Repeated characters in a word were removed, e.g. *ayudaaaa* is replaced with *ayuda* (Spanish word for 'help').

Although we do not consider symbols, there are works that have included emotion symbols for sentiment analysis, such as emoticons and emoji ideograms [89,90]. Also, in relation to repeated characters, these may contain sentiment information. Nevertheless, since our source is Twitter, further analysis during crisis situations is required to validate whether these repeated characters are intentional or not. An example where repeated characters have been used can be found in [91].

4.3. Feature representation

This module is responsible for extracting features from preprocessed tweets. In this paper, bag-of-words (BOW) technique is used to convert training tweets into a numeric representation resulting in a term document matrix (TDM). After learning the vocabulary, BOW describes the presence of known words within a tweet. This method creates an indicator vector signaling whether words in key-words-dictionary of a text are in the text. For example, consider the following two tweets: tweet1: *yesterday is past* and tweet2: *today is present*. The vocabulary is {*yesterday, is, past, today, present*}. Now, the above tweets are represented as: $tweet1vector = [1, 1, 1, 0, 0]$ and $tweet2vector = [0, 1, 0, 1, 1]$.

4.4. Sentiment analysis

Each tweet is then labeled with a sentiment with two possible values: negative or positive. We used a list of English positive and negative opinion words or sentiment words (around 7000 words). This list of words was translated into Spanish.[3] There exists a variety of sentiment analysis algorithms able to capture positive and negative sentiment, some specifically designed for short, informal texts [92]. In this work, we first determined the sentiment polarity of each tweet by adapting the following measure, which determines the direction of the sentiment as well as its strength [93,94]:

$$Sentiment\ score = \frac{positive - negative}{positive + negative + 2}, \tag{6}$$

where *positive* represents the positive words count and *negative* the negative words count in the tweet. Formally, we represent it by a discrete 2-valued variable C, which denotes the sentiment class:

$$C \in \{-1, 1\}. \tag{7}$$

This variable captures well our assumptions about the ordering of the sentiment values and the distances between them.

[3] This file and the papers can all be downloaded from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

In some cases, the measure polarity fails to capture the degree of emotionality of the tweet because the positive and negative sentiment scores cancel out each other (*Sentiment score* $= 0$, although the tweet is actually heavily emotional and not neutral as the measure might indicate). Therefore, we introduced the following definition:

$$C = \begin{cases} 1 \text{ (positive tweet)} & \text{if } \textit{Sentiment score} \geq 0.1 \\ -1 \text{ (negative tweet)} & \text{if } \textit{Sentiment score} < 0.1. \end{cases} \quad (8)$$

The use of one or more thresholds is common, for example [95]. In our case, the value of 0.1 was found empirically after several runs using the training sets, where the effect of not real neutrals was reduced.

4.5. Simulation setup

To explore the performances of NB, TAN, and BF TAN for sentiment analysis considering both previously described datasets, we randomly sample 70% of the examples of each dataset to generate the training set, and use the remaining 30% to generate the test set. Additionally, we consider two black box classifiers [96] such as SVM and RF for comparison purposes.

We train the five classifiers on the same training set and then compute, for each classifier, the confusion matrix (using the test set) where,

- True Positives (TP): The number of correctly classified positive tweets
- True Negatives (TN): The number of correctly classified negative tweets
- False Positives (FP): The number of incorrectly classified positive tweets
- False Negatives (FN): The number of incorrectly classified negative tweets

Then, the following performance measures are computed:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \quad (9)$$

$$Precision = \frac{TP}{TP + FP}, \quad (10)$$

$$Recall = \frac{TP}{TP + FN}, \quad (11)$$

$$F_1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \quad (12)$$

For each dataset, we run 30 times the data splitting procedure, 70% for training and 30% for testing (the splitting was carried out randomly). For each run, we computed the above classification performance measures on the test set, then the average and the standard deviation of each measure was reported.

4.6. Imbalanced sentiment analysis

The class imbalance problem in binary classification occurs when the sizes of the classes differ greatly. In this case, any classifier is biased toward the majority class. In this work, the class distribution for each dataset is shown in Table 2, where we notice that more than 80% of the tweets are negative.

We applied SMOTE [97] to the training sets to handle the class imbalance problem. SMOTE over-samples the minority class by adding synthetic samples based on feature-space similarities between existing minority examples. We used the following parameters for SMOTE: $k = 5$ for the k-nearest neighbors algorithm used by SMOTE to generate synthetic training data, over-sampling rate of 100% (for the minority class), and under-sampling rate of 200% (for the majority class). With these parameters we obtained fully balanced datasets. SMOTE was applied before each training round.

5. Results

The results of the five classifiers using dataset 1 are shown in Table 3. Overall, the best performance was obtained by the SVM classifier followed by the BF TAN model with 19 edges. The corresponding networks for TAN and BF TAN are shown in Figs. 3 and 4, respectively. Edges from the CLASS node to all the other nodes have been omitted to simplify the visualization.

For dataset 2, the results are shown in Table 4. In this case, BF TAN allowed all the edges, thus obtaining the TAN model. Hence, both classifiers show the same performances (see Table 4). Here, RF obtained the best predictive performance, but now with the TAN model showing a competitive result (compared to RF and SVM) passing the 80% accuracy threshold. The network for the TAN model is shown in Fig. 5. Edges from the CLASS node to all the other nodes have been omitted to simplify the visualization.

6. Discussion

The SVM was the only classifier obtaining above 80% of accuracy in both datasets, which is consistent with previous works shown in Table 1. Nevertheless, as discussed previously, the SVM and RF are black box models, which makes it difficult to interpret the relations amongst the attributes during the classification process. This limitation is partially addressed through the use of Bayesian network classifiers. In the case of dataset 1, we noticed that there was not enough training examples to support the (complete) TAN model. In fact, by forcing the tree structure, the performance of the resulting classifier is lower than the NB model. In this case, the usefulness of the BF TAN approach is twofold. On the one hand, by using (5) it was found that only 19 of the total 144 edges are supported by the data, yielding in a forest structure with better classification performance when compared to NB, TAN, and RF. On the other hand, the visualization of the TAN model (Fig. 3) could be misleading since, from the classification perspective, there are several edges which negatively contribute to the classification performance, and therefore, should be omitted. This is exactly what BF TAN does: it shows that for this sentiment analysis application only a few words are conditioned to another word (in addition to the CLASS variable) as can be seen in Fig. 4. When analyzing the results using dataset 2, we noticed that for this case there were sufficient training examples to support the TAN model, i.e., the tree structure with 206 edges. Therefore, BF TAN and TAN are equivalent for this application. Also, the generalization performance of the resulting TAN model is competitive with the SVM and less competitive when compared to RF, nevertheless offering a model with good quantitative performance and also complementary qualitative information obtained through Fig. 5.

Even though BF TAN shows a few conditioned words for dataset 1 in Fig. 4, they are, however, qualitatively relevant regarding central aspects of the critical event and the restructuration or recovery phase. The directed path Hawaii → tsunami → alerta refers to the early warning system for tsunamis in Chilean coasts. The 2010 Chilean earthquake (8.8 magnitude) produced several tsunami waves in Chilean coasts and was particularly destructive in Constitución (a city in Southern Chile also named in the graph). Chilean maritime authorities misunderstood the information on the tsunami provided by the Pacific Tsunami Warning Center, located in Ewa Beach, Hawaii. The tsunami

Table 2
Datasets structure.

Dataset	Positive	Negative	Total	Words
Dataset 1: Chilean earthquake	298	1889	2187	145
Dataset 2: Catalan independence referendum	10 816	49 184	60 000	207

Table 3
Performance measures for each classifier using dataset 1.

Algorithm	Accuracy	Precision	Recall	F_1-score	N° Edges
NB	0.742 ± 0.027	0.895 ± 0.004	0.790 ± 0.009	0.841 ± 0.020	0
TAN	0.721 ± 0.029	0.896 ± 0.002	0.765 ± 0.032	0.825 ± 0.019	144
BF TAN	0.764 ± 0.007	0.898 ± 0.003	0.809 ± 0.034	0.849 ± 0.021	19
SVM	0.812 ± 0.067	0.867 ± 0.009	0.936 ± 0.081	0.899 ± 0.042	–
RF	0.725 ± 0.061	0.892 ± 0.011	0.776 ± 0.079	0.828 ± 0.054	–

Table 4
Performance measures for each classifier using dataset 2.

Algorithm	Accuracy	Precision	Recall	F_1-score	N° Edges
NB	0.781 ± 0.013	0.885 ± 0.005	0.852 ± 0.004	0.868 ± 0.000	0
TAN	0.808 ± 0.004	0.906 ± 0.005	0.854 ± 0.009	0.879 ± 0.007	206
BF TAN	0.808 ± 0.004	0.906 ± 0.005	0.854 ± 0.009	0.879 ± 0.007	206
SVM	0.829 ± 0.005	0.841 ± 0.011	0.985 ± 0.008	0.907 ± 0.007	–
RF	0.858 ± 0.008	0.922 ± 0.002	0.895 ± 0.010	0.908 ± 0.005	–

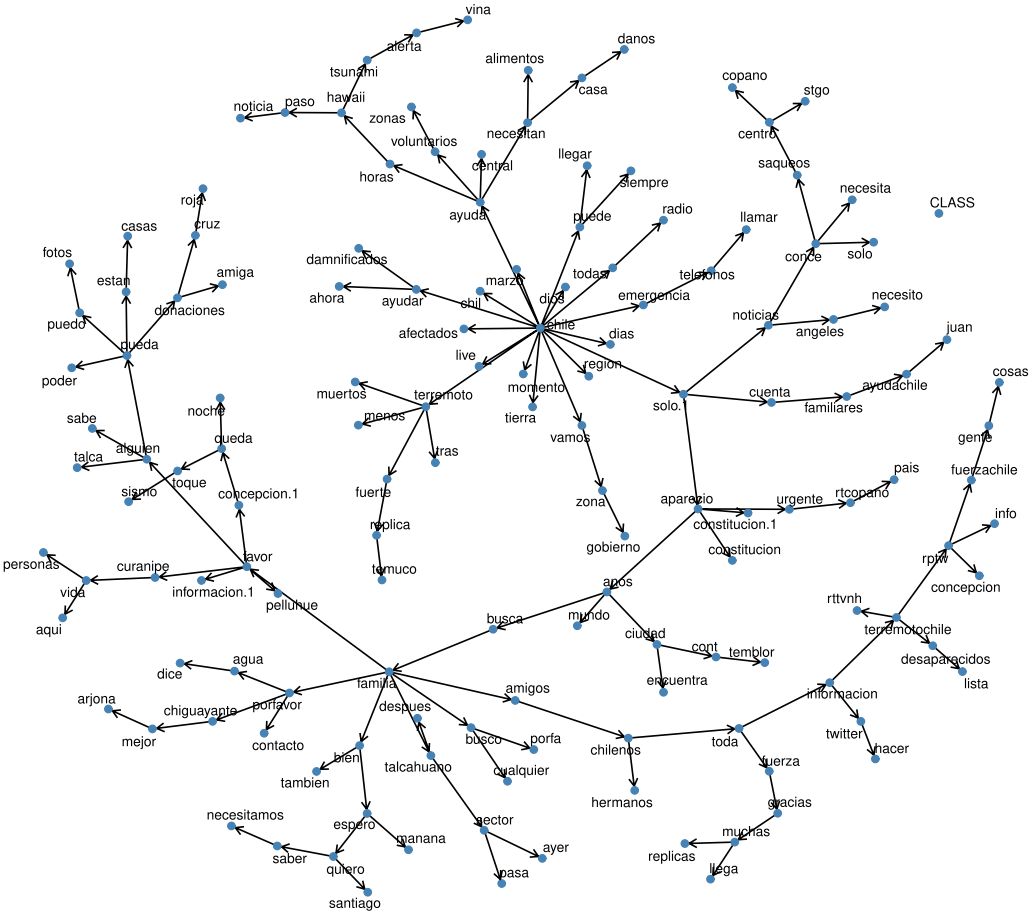

Fig. 3. TAN classifier for the 2010 Chilean earthquake dataset.

hit the coast unannounced, thereby killing about 156 persons and producing substantial material loses [98]. The directed edge emergencia → telefono signals another relevant topic regarding the unpreparedness of critical infrastructure for the crisis. Regular communication systems collapsed after the earthquake. Even the government agencies did not have an alternative satellite network (emergency line) to communicate with the affected zones and coordinate the assistance, thus delaying the government reaction to the emergency and unleashing uncertainty and disturbs [99]. The words of the directed edge queda → noche and queda → toque refers to the curfew ('toque de queda') that went into effect during the days after the earthquake. Riots and looting in the area led the government to declare a state of emergency in the most affected cities. Politically, the measure

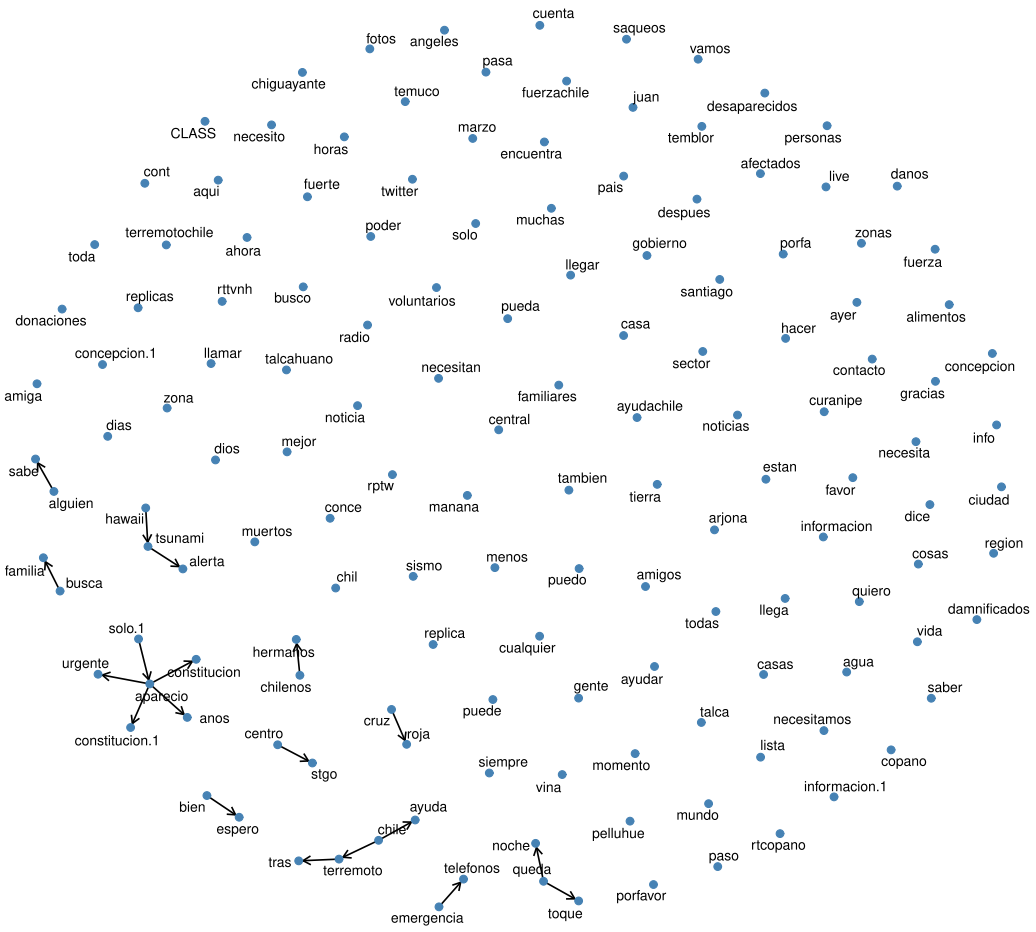

Fig. 4. Bayes factor TAN classifier for the 2010 Chilean earthquake dataset.

was highly disputed in public opinion because it implied to hand over the control of major cities to military forces [100]. Finally, the directed path chile → terremoto → tras and the directed edge chile → ayuda refer to public appeals calling to provide assistance and organize the recovery of the affected zones after the earthquake.

The tree structure of the TAN model for dataset 2 is much more diverse in qualitative information as shown in Fig. 5. Overall, the graph presents the strong controversy around the referendum for independence of Catalunya from Spain in October 2017. For instance, three links appear from the node 'catalan.1' (middle left), referring to October 1 2017, the day of the Catalan independence referendum. The directed paths catalan.1 → quieren → votar; catalan.1 → democracia → legalidad; and catalan.1 → defender → elecciones reflect the controversy on the democratic legitimacy of the Catalan referendum. The referendum of independence of Catalunya took place in spite of the decision of the Spanish Constitutional Court about the illegality of the act. In the morning of October 1 2017, the Generalitat (Catalan parliament and government) announces the opening of polling sites protected by voters, while State security forces intervene to prevent the illegally declared election [101]. The directed path nazis → alertaultra → vamos → represion → violencia → policial → imagenes (lower right) informs about the State police repression against the independence movement in the day of the referendum and in the following days. While, the directed path gobierno → espanol → catalan → pueblo → apoyo → nunca (lower left) express the opinion that the Spanish government shall never concede support to Catalan people. Complementarily, the directed path espana → unidad → barcelona → totssomcatalunya → borrell → colau (middle right) reflects the opposition to the independence movement by highlighting: (a) the unity of Spain; (b) the slogan 'we all are Catalunya' that represents the manifestations in Barcelona in October 2017 against the independence declaration; and (c) the relevance of two recognized Catalan politicians opposed to the independence movement: Josep Borrell and Ada Colau.

As seen, the TAN is an alternative model that unpacks the black box of the SVM (or RF) classifier offering (when data availability is enough to train the model) a detailed view on the qualitative content of Twitter communication. And even if there are no sufficient examples, available words form the directed paths (obtained via BF TAN) may help to reconstruct the milestones of the critical event, as in the case of the Chilean earthquake 2010.

Compared with traditional qualitative methods in social sciences (in-depth interviews, questionnaires, focus groups), big data analysis on Twitter communication and social media provides a wider and far more dynamic picture of critical situations, as historical as up-to-date, although challenges in codification and classification of Twitter data persist [102]. Considering the increase of tweets and retweets as the critical event develops, relevant information loses centrality, making it difficult to identify qualitatively meaningful communication patterns [103]. On the other hand, Twitter information coincides in some cases with the ground truth obtained by technical devices in natural locations, while in other it does not, because emotions construe an alternative reality of the critical event that may lead authorities to bad decisions [104]. There are several challenges in this area that still need to be addressed, such as strategies to reduce the errors from subjective information about critical events, to increase participation of users, and to improve geo-location techniques [105].

Fig. 5. TAN classifier for the 2017 Catalan independence referendum dataset.

Nonetheless, as seen in our analysis, the work on classifiers of Twitter data improves our comprehension of the dynamics of communication in critical events.

6.1. Limitations of this study

Despite its widespread use in scientific research, a methodological discussion has grown in recent years regarding the representativeness and validity of the samples publicly delivered by Twitter. Two main questions arise: whether Twitter users represent society and whether Twitter samples represent Twitter communication in general [106]. Aligned with the purposes of our analysis, we focus on the second question.

There are different forms to have access to Twitter data. The most popular is the Application Programming Interface (API), including the Streaming API divided into two sources: Filter API (search by parameters, e.g. keywords, user accounts, hashtags, geographic areas), and Sample API (it delivers 1% of all tweets without parameters). Another source is the Representational State Transfer or REST API, which provides timelines of individual users (limited to the 3200 most recent tweets by user) and several tools for data interaction. Additionally, paid services include a 10% of the Sample API (known as Decahose or Gardenhose) and the most expensive and technologically demanding option for the 100% of tweets (known as Firehose) [107,108].

Since Twitter does not share information about its sampling method, a crucial methodological question arises whether the 1% of public Twitter data is a valid representation of Twitter communication. Sample bias has been detected by independent researchers when analyzing different API sources or comparing API data with Twitter Firehose. [109] have found that for a large size Sample API works well, but for low coverage it shows negative correlations whereas random sampled data from Firehose exhibit positive correlations. [110] find that the application of filters marginalizes relevant information regarding the network structure. Because of snowballing effects, particularly peripheral activity is misrepresented. Without parameters, Sample API (Streaming API after the authors) delivers a more realistic depiction of the network activity. [111,112] show that it is possible to detect bias in data from Streaming API by either contrasting data with other API sources or replicating the sampling procedure at different times and from different access points in a given period. A significant amount of overlap between samples (around 96%) supports this procedure and avoids having to resort to the for most researchers costly prohibitive Decahose or Firehose.

As shown by [113], an additional problem arises, namely, hashtags can be individually (same user) or socially (other users) reused, which indicates confirmation bias and bubble effects in time. Since Sample API (data returned in streaming without parameters) is an unbiased section of all tweets, [114] have found that bias in hashtag trends in the Streaming API with parameters (obtained through the filter API endpoint) can be detected by extracting bootstrapped samples from the Sample API and obtaining a confidence interval for hashtag activity at a given time period. In any case, [107] hold that even Sample API can be biased either by design, thereby fostering preferences by specific sociodemographic groups, or by purpose, thus influencing accounts or topics via bots intentionally (sample cheaters, corporate spammers) or unintentionally (in-syncs and tweet scheduling tools). Particularly considering sentiment analysis, the authors demonstrate that the time window of 10 millisecond of the 1% Sample API can be influenced by injecting positive tweets into the

communication stream, thereby radically changing the emotion of the sample (a ten-fold increase) from negative to positive. After experimental analyses, the authors conclude that Twitter cannot provide random samples via Sample API and offer tools to remove over-sampling of accounts and topics.

Considering this, [115] have designed a data collection strategy to increase the reliability of Twitter data by sampling user accounts from REST API (3200 tweets of different users), thus developing a comprehensive data set of 1.6 m of accounts and 500 m of tweets for a given period of time (months). The data set was constructed with Japanese language users, the most active population in Twitter. Depending on storage systems (rather inexpensive Cloud services), data can be refreshed every month. Comparing the sample with the historical archive of tweets from Crimson Hexagon complete dataset (a company for data analytics) acquired from Twitter Firehose, results showed a correlation of 0.97 in keyword trends and other patterns, which is significantly higher than the representativeness obtained from other API-based strategies.

As pointed out previously, the emotion of the sample can be changed, i.e., the class distribution for sentiment analysis could be altered due to the Twitter data sampling mechanism employed. Although during critical events, one would expect more negative tweets than positive tweets, it is not clear whether the proportion shown in Table 2 approximates the actual class distribution of the sentiments during the events. Nevertheless, the class imbalance problem, regardless of the proportion, can be handled to avoid introducing a bias towards the majority class, during the training phase of the machine learning classifier, either adopting some weighting scheme like in [116,117], or using some oversampling from the minority class, like SMOTE (used in this work) and others [118]. Of course, the evaluations are reported on the unbalanced test sets, which resembles better the situation during the critical events. In fact, from Tables 3 and 4 it is shown that precision, recall, and F_1-score, which are measured based on the positive class (positive tweets), achieve competitive results, meaning that the classifiers are able to predict effectively positive tweets even though the majority of the examples in the test sets are negative, where the classifiers are also effective as shown by the accuracy. This is an important topic, and although in this work we only considered the use of SMOTE, other techniques for the class imbalance problem should be analyzed in this context.

While the focus of this paper is the use of Bayesian network classifiers and a Bayes factor approach for sentiment analysis during critical events, we have only considered the straightforward tweet-level sentiment analysis, in particular, only the text of the tweet. The performance of this approach relies heavily in the quality of the training data, and as previously mentioned, this data can be biased and noisy. Thus, performing text-level sentiment analysis through machine learning, in particular, black box models, can be rather unreliable as pointed out by [18]. One approach that has been considered to reduce the effect of the biased and noisy data, is to use the information of the hashtags [18,119]. For example, [119] used sentiment polarity of hashtags as the features in classification in the political domain. In that work, the number of positive and negative hashtags for each tweet are considered as input features. In our case, for future research, an initial stage of hashtags classification (hashtag-level sentiment classification) could be explored for the 2017 Catalan independence referendum, i.e., hashtags that support the independence referendum and those that do not in the recently reactivated conflict in October 2019 [120]. Or in the case of natural disasters like an earthquake, hashtags that are related to the negative effects of the earthquake and those related to solidarity, compassion, and other pro-social feelings could be analyzed. Then in a second stage, a tweet-level sentiment analysis can be conducted. Finally, it could be possible to explore different mechanisms to combine the hashtag-level sentiment classification with the tweet-level sentiment analysis, to enhance and generate more reliable predictions.

Another limitation is that the proposed sentiment analysis approach is heavily time and event dependent, meaning that the applicability and predictive validity of the model is only for a short time window around the event. Even if we consider a specific type of critical event, and train a Bayesian network classifier, then for a future event of the same nature, new hashtags will emerge, and people may use other or additional words to express emotions that are not present in the trained Bayesian network classifier, thus limiting or disabling its use. On the other hand, the fact that the resulting model has a network structure relating the words and how these interact to perform the classification during the event, allows a qualitative assessment by the user in order to ensure the conceptual validity of the model. This qualitative assessment can be complemented by what other sources of information (written press, radio, television, etc.) are reporting on the same event. In relation to the use of the predictive model for future events, where the context that was originally modeled by the Bayesian network classifier changes, either slowly or rapidly and unexpected, in time or in characteristic, an online learning approach could be adopted, as in [121–123]. Alternatively, an unsupervised learning algorithm for Bayesian networks [124] could be used to learn from the new data coming from a new event with new words, and then combine this new Bayesian network learned in a unsupervised manner with a pruned version of the original Bayesian network, leaving only words that are appearing in both events. The flexibility of different learning schemes for Bayesian networks as well as missing data handling, makes this type of machine learning technique promising for sentiment analysis during critical events.

7. Conclusions

The use of Twitter has many potentialities and applications before, during, and after critical events. The accuracy of classification of tweets' semantic content is highly important to reduce the risks of misinformation in those situations. In this paper, we have reviewed five classifiers (one being a variant of the TAN model) and assessed their performance with two Twitter datasets from two different critical events, the 2010 Chile earthquake and the 2017 Catalan independence referendum. We conclude that there is no difference as to how SVM and RF behave in English or Spanish. Regarding the accuracy of the models, while SVM obtained the best performance in dataset 1 and RF in dataset 2, the Bayesian network classifier TAN obtained competitive results when there was sufficient data to support the tree structure.

In addition, we conclude that TAN and BF TAN offer interesting qualitative information to historically and socially comprehend the main features of the event dynamics, even if there are no sufficient training examples. Moreover, the resulting networks allow for the construction of a narrative or storytelling of the critical event been analyzed. For future research, other qualitative approaches like grounded theory will be explored in combination with the resulting Bayesian network classifier, as proposed in [125].

This may motivate traditional social scientists to integrate machine learning techniques into their regular research toolbox and consider Twitter communication as a privileged platform in which social facts and emotions are produced and reflected.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by CONICYT-Chile under grant Fondecyt 1180706 (G.A.R), Fondecyt 1190265 (A.M., G.A.R.), Basal (CONICYT)-CMM, Chile (G.A.R), and the Doctoral scholarship, Chile (2015- 21150790) (P.A.H.).

References

[1] N. Öztürk, S. Ayvaz, Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis, Telemat. Inform. 35 (1) (2018) 136–147.
[2] S. Valenzuela, S. Puente, P.M. Flores, Comparing disaster news on twitter and television: an intermedia agenda setting perspective, J. Broadcast. Electron. Media 61 (4) (2017) 615–637.
[3] C. Diamantini, A. Mircoli, D. Potena, E. Storti, Social information discovery enhanced by sentiment analysis techniques, Future Gener. Comput. Syst. 95 (2019) 816–828.
[4] L.F.S. Coletta, N.F.F. da Silva, E.R. Hruschka, E.R. Hruschka, Combining classification and clustering for tweet sentiment analysis, in: 2014 Brazilian Conference on Intelligent Systems, IEEE, 2014.
[5] E.S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O.S. Siordia, E.A. Villaseñor, A case study of spanish text transformations for twitter sentiment analysis, Expert Syst. Appl. 81 (2017) 457–471.
[6] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2002, pp. 79–86.
[7] P.D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.
[8] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2) (1997) 131–163.
[9] V.K. Singh, R. Piryani, A. Uddin, P. Waila, Marisha, Sentiment analysis of textual reviews; evaluating machine learning, unsupervised and sentiwordnet approaches, in: 2013 5th International Conference on Knowledge and Smart Technology (KST), 2013, pp. 122–127.
[10] S. Zhu, B. Xu, D. Zheng, T. Zhao, Chinese microblog sentiment analysis based on semi-supervised learning, in: Semantic Web and Web Science, Springer New York, New York, NY, 2013, pp. 325–331.
[11] S. Tan, J. Zhang, An empirical study of sentiment analysis for chinese documents, Expert Syst. Appl. 34 (4) (2008) 2622–2629.
[12] P.A. Henríquez, G.A. Ruz, Twitter sentiment classification based on deep random vector functional link, in: International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–6.
[13] M. Al-Ayyoub, S.B. Essa, I. Alsmadi, Lexicon-based sentiment analysis of arabic tweets, Int. J. Soc. Netw. Min. 2 (2) (2015) 101.
[14] Ankit, N. Saleena, An ensemble classification system for twitter sentiment analysis, Procedia Comput. Sci. 132 (2018) 937–946, International Conference on Computational Intelligence and Data Science.
[15] E. Boiy, M.-F. Moens, A machine learning approach to sentiment analysis in multilingual web texts, Inf. Retr. 12 (5) (2008) 526–558.
[16] H. Ghorbel, D. Jacot, Further experiments in sentiment analysis of french movie reviews, in: Advances in Intelligent Web Mastering – Vol. 3, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 19–28.
[17] P. Melville, W. Gryc, R.D. Lawrence, Sentiment analysis of blogs by combining lexical knowledge with text classification, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 1275–1284.
[18] X. Wang, F. Wei, X. Liu, M. Zhou, M. Zhang, Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 1031–1040.
[19] M. Gamon, Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis, in: Proceedings of the 20th International Conference on Computational Linguistics, Association for Computational Linguistics, 2004.
[20] B. Pang, L. Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2005, pp. 115–124.
[21] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach, J. Informetr. 3 (2) (2009) 143–157.
[22] M. Annett, G. Kondrak, A comparison of sentiment analysis techniques: Polarizing movie blogs, in: Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 25–35.
[23] T. Mullen, N. Collier, Sentiment analysis using support vector machines with diverse information sources, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2004.
[24] I. Helsloot, J. Groenendaal, Twitter: An underutilized potential during sudden crises?, J. Contingencies Crisis Manag. 21 (3) (2013) 178–183.
[25] M. Martínez-Rojas, M. del Carmen Pardo-Ferreira, J.C. Rubio-Romero, Twitter as a tool for the management and analysis of emergency situations: A systematic literature review, Int. J. Inf. Manage. 43 (2018) 196–208.
[26] M. Scheffer, Critical Transitions in Nature and Society, Princeton University Press, Princeton, 2009.
[27] K. Weick, K. Sutcliffe, Managing the Unexpected: Sustained Performance in a Complex World, Wiley, Hoboke, NJ, 2015.
[28] B. Fredrickson, M. Tugade, C. Waugh, G. Larkin, What good are positive emotions in crises? a prospective study of resilience and emotions following the terrorist attacks on the united states on september 11th, 2001, J. Personal. Soc. Psychol. 84 (2003) (2001) 365–376.
[29] S. Folkman, J. Moskowitz, Positive affect and the other side of coping, Amer. Psychol. 55 (2000) 647–654.
[30] D. Savage, B. Torgler, The emergence of emotions and religious sentiments during the september 11 disaster, Motiv. Emot. 37 (2013) 586–599.
[31] F. Bravo-Marquez, M. Mendoza, B. Poblete, Meta-level sentiment models for big social data analysis, Knowl.-Based Syst. 69 (2014) 86–99.
[32] C. Caragea, A. Squicciarini, S. Stehle, K. Neppalli, A. Tapia, Mapping moods: Geo-mapped sentiment analysis during hurricane sandy, in: ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management, The Pennsylvania State University, 2014, pp. 642–651.
[33] P.S. Dodds, K.D. Harris, I.M. Kloumann, C.A. Bliss, C.M. Danforth, Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter, PLoS One 6 (2011) e26752.
[34] E.M. Cody, A.J. Reagan, L. Mitchell, P.S. Dodds, C.M. Danforth, Climate change sentiment on twitter: An unsolicited public opinion poll, PLoS One 10 (2015) e0136092.
[35] S.E. Alajajian, J.R. Williams, A.J. Reagan, S.C. Alajajian, M.R. Frank, L. Mitchell, J. Lahne, C.M. Danforth, P.S. Dodds, The lexicocalorimeter: Gauging public health through caloric input and output on social media, PLoS One 12 (2017) e0168893.
[36] P. Baylis, N. Obradovich, Y. Kryvasheyeu, H. Chen, L. Coviello, E. Moro, M. Cebrian, J.H. Fowler, Weather impacts expressed sentiment, PLoS One 13 (2018) e0195750.
[37] B. Snefjella, D. Schmidtke, V. Kuperman, National character stereotypes mirror language use: A study of Canadian and american tweets, PLoS One 13 (2018) e0206188.
[38] M. Stein, The critical period of disasters: Insights from sense-making and psychoanalytic theory, Hum. Relat. 57 (10) (2004) 1243–1261.
[39] A. Mascareño, E. Goles, G.A. Ruz, Crisis in complex social systems: A social theory view illustrated with the chilean case, Complexity 21 (S2) (2016) 13–23.
[40] S. Goswami, S. Chakraborty, S. Ghosh, A. Chakrabarti, B. Chakraborty, A review on application of data mining techniques to combat natural disasters, Ain Shams Eng. J. 9 (3) (2018) 365–378.
[41] J. Son, H.K. Lee, S. Jin, J. Lee, Content features of tweets for effective communication during disasters: A media synchronicity theory perspective, Int. J. Inf. Manage. 45 (2019) 56–68.
[42] T. Sakaki, M. Okazaki, Y. Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development, IEEE Trans. Knowl. Data Eng. 25 (4) (2013) 919–931.
[43] D. Wu, Y. Cui, Disaster early warning and damage assessment analysis using social media data and geo-location information, Decis. Support Syst. 111 (2018) 48–59.
[44] E.M. Cody, A.J. Reagan, L. Mitchell, P.S. Dodds, C.M. Danforth, Climate change sentiment on twitter: An unsolicited public opinion poll, PLoS One 10 (8) (2015) 1–18.
[45] K.M. Carley, M. Malik, P.M. Landwehr, J. Pfeffer, M. Kowalchuck, Crowd sourcing disaster management: The complex nature of twitter usage in padang indonesia, Saf. Sci. 90 (2016) 48–61.
[46] P.M. Landwehr, W. Wei, M. Kowalchuck, K.M. Carley, Using tweets to support disaster planning, warning and response, Saf. Sci. 90 (2016) 33–47.
[47] H.M. Saleem, Y. Xu, D. Ruths, Effects of disaster characteristics on twitter event signature, Procedia Eng. 78 (2014) 165–172.
[48] C. Caragea, A. Squicciarini, S. Stehle, K. Neppalli, A. Tapia, Mapping moods: Geo-mapped sentiment analysis during hurricane sandy, in: 11th International Conference on Information Systems for Crisis Response and Management, The Pennsylvania State University, 2014, pp. 642–651.
[49] V.K. Neppalli, C. Caragea, A. Squicciarini, A. Tapia, S. Stehle, Sentiment analysis during hurricane sandy in emergency response, Int. J. Disaster Risk Reduct. 21 (2017) 213–222.
[50] S. Yoo, J. Song, O. Jeong, Social media contents based sentiment analysis and prediction system, Expert Syst. Appl. 105 (2018) 102–111.

[51] T. Gurman, N. Ellenberger, Reaching the global community during disasters: Findings from a content analysis of the organizational use of Twitter after the 2010 Haiti earthquake, J. Health Commun. 20 (2010) 687–696.
[52] C.C. David, J.C. Ong, E.F.T. Legara, Tweeting supertyphoon haiyan: Evolving functions of twitter during and after a disaster event, PLoS One 11 (3) (2016) 1–19.
[53] O. Gruebner, S.R. Lowe, M. Sykora, K. Shankardass, S.V. Subramanian, S. Galea, A novel surveillance approach for disaster mental health, PLoS One 12 (7) (2017) 1–15.
[54] M. Jamali, A. Nejat, S. Ghosh, F. Jin, G. Cao, Social media data and post-disaster recovery, Int. J. Inf. Manage. 44 (2019) 25–37.
[55] M.W. Graham, E.J. Avery, S. Park, The role of social media in local government crisis communications, Public Relat. Rev. 41 (3) (2015) 386–394.
[56] N. Pourebrahim, S. Sultana, J. Edwards, A. Gochanour, S. Mohanty, Understanding communication dynamics on twitter during natural disasters: A case study of hurricane sandy, Int. J. Disaster Risk Reduct. 37 (2019) 101176.
[57] Q. Han, J. Guo, H. Schütze, Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text, in: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, 2013.
[58] S. Symeonidis, D. Effrosynidis, A. Arampatzis, A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis, Expert Syst. Appl. 110 (2018) 298–310.
[59] O. Araque, I. Corcuera-Platas, J.F. Sánchez-Rada, C.A. Iglesias, Enhancing deep learning sentiment analysis with ensemble techniques in social applications, Expert Syst. Appl. 77 (2017) 236–246.
[60] Y. Ren, R. Wang, D. Ji, A topic-enhanced word embedding for twitter sentiment classification, Inform. Sci. 369 (2016) 188–198.
[61] N.F.F. da Silva, L.F. Coletta, E.R. Hruschka, E.R.H. Jr, Using unsupervised information to improve semi-supervised tweet sentiment classification, Inform. Sci. 355–356 (2016) 348–365.
[62] S. Xiong, H. Lv, W. Zhao, D. Ji, Towards twitter sentiment classification by multi-level sentiment-enriched word embeddings, Neurocomputing 275 (2018) 2459–2466.
[63] M.D. Cao, I. Zukerman, Experimental evaluation of a lexicon-and corpus-based ensemble for multi-way sentiment analysis, in: Proceedings of the Australasian Language Technology Association Workshop, 2012, pp. 52–60.
[64] J.C.S. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, Supervised learning for fake news detection, IEEE Intell. Syst. 34 (2) (2019) 76–81.
[65] F. Xu, J. Yu, R. Xia, Instance-based domain adaptation via multiclustering logistic approximation, IEEE Intell. Syst. 33 (1) (2018) 78–88.
[66] M.S. Akhtar, A. Ekbal, S. Narayan, V. Singh, No, that never happened!! investigating rumors on twitter, IEEE Intell. Syst. 33 (5) (2018) 8–15.
[67] D. Mahata, J. Friedrichs, R.R. Shah, J. Jiang, Detecting personal intake of medicine from twitter, IEEE Intell. Syst. 33 (4) (2018) 87–95.
[68] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, Sentiment and sarcasm classification with multitask learning, IEEE Intell. Syst. 34 (3) (2019) 38–43.
[69] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, A. Hussain, Multimodal sentiment analysis: Addressing key issues and setting up the baselines, IEEE Intell. Syst. 33 (6) (2018) 17–25.
[70] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, pp. 3856–3866.
[71] W. Zhao, H. Peng, S. Eger, E. Cambria, M. Yang, Towards scalable and reliable capsule networks for challenging nlp applications, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1549–1559.
[72] Y. Du, X. Zhao, M. He, W. Guo, A novel capsule based hybrid neural network for sentiment classification, IEEE Access 7 (2019) 39321–39328.
[73] H.W. Fentaw, T.-H. Kim, Design and investigation of capsule networks for sentence classification, Appl. Sci. 9 (11) (2019).
[74] C. Zhang, Y. Li, N. Du, W. Fan, P. Yu, Joint slot filling and intent detection via capsule neural networks, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5259–5267.
[75] H. Ren, H. Lu, Compositional coding capsule network with k-means routing for text classification, 2018, arXiv e-prints arXiv:1810.09177 arXiv:1810.09177.
[76] H. Peng, Y. Ma, Y. Li, E. Cambria, Learning multi-grained aspect target sequence for chinese sentiment analysis, Knowl.-Based Syst. 148 (2018) 167–176.
[77] S. Lai, K. Liu, S. He, J. Zhao, How to generate a good word embedding, IEEE Intell. Syst. 31 (6) (2016) 5–14.
[78] J.R. Ragini, P.M.R. Anand, An empirical analysis and classification of crisis related tweets, in: 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), IEEE, 2016.
[79] J.R. Ragini, P.R. Anand, V. Bhaskar, Mining crisis information: A strategic approach for detection of people at risk through social media analysis, Int. J. Disaster Risk Reduct. 27 (2018) 556–566.
[80] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco, California, 1988.
[81] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons, New York, 1973.
[82] C. Bielza, P. Larrañaga, Discrete bayesian network classifiers: A survey, ACM Comput. Surv. 47 (2014) 5:1–5:43.
[83] D. Margaritis, S. Thrun, Bayesian network induction via local neighborhoods, in: S.A. Solla, T.K. Leen, K. Müller (Eds.), Advances in Neural Information Processing Systems, Vol. 12, MIT Press, 2000, pp. 505–511.
[84] G.M. Provan, M. Singh, Learning Bayesian Networks using Feature Selection, Springer New York, New York, NY, 1996, pp. 291–300.
[85] M.J. Pazzani, Constructive Induction of Cartesian Product Attributes, Springer US, Boston, MA, 1998, pp. 341–354.
[86] M. Sahami, Learning limited dependence bayesian classifiers, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, 1996, pp. 335–338.
[87] G.A. Ruz, D.T. Pham, Building bayesian network classifiers through a bayesian complexity monitoring system, Proc. Inst. Mech. Eng. Part C 223 (2009) 743–755.
[88] A. Cobo, D. Parra, J. Navón, Identifying relevant messages in a twitter-based citizen channel for natural disaster situations, in: Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion, ACM, New York, NY, USA, 2015, pp. 1189–1194.
[89] D. Vilares, M.A. Alonso, C. Gómez-Rodríguez, Supervised polarity classification of spanish tweets based on linguistic knowledge, in: Proceedings of the 2013 ACM Symposium on Document Engineering, DocEng '13, ACM, New York, NY, USA, 2013, pp. 169–172.
[90] W. Wolny, Sentiment analysis of twitter data using emoticons and emoji ideograms, Wydaw. Uniw. Ekon. Katowicach. (2016).
[91] V. Hangya, R. Farkas, Target-oriented opinion mining from tweets, in: IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom), 2013, pp. 251–254.
[92] C. Akkaya, J. Wiebe, R. Mihalcea, Subjectivity word sense disambiguation, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 190–199.
[93] M. Grčar, D. Cherepnalkoski, I. Mozetič, P. Kralj Novak, Stance and influence of twitter users regarding the brexit referendum, Comput. Soc. Netw. 4 (1) (2017) 6.
[94] P. Gabrovšek, D. Aleksovski, I. Mozetič, M. Grčar, Twitter sentiment around the earnings announcement events, PLoS One 12 (2) (2017) e0173151.
[95] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčr, M. Igor, The effects of twitter sentiment on stock price returns, PLoS One 10 (9) (2015) 1–21.
[96] B. Lantz, Machine Learning with R, second ed., Packt Publishing, 2015.
[97] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.
[98] B. Soulé, Post-crisis analysis of an ineffective tsunami alert: the 2010 earthquake in maule chile, Disasters 38 (2) (2010) 375–397.
[99] J. Moreno, The role of communities in coping with natural disasters: Lessons from the 2010 chile earthquake and tsunami, Procedia Eng. 212 (2018) 1040–1045.
[100] M. Salgado, E. Marchione, A. Gill, The calm after the storm? looting in the context of disasters, in: 3rd World Congress on Social Simulation, 2010, pp. 1–8.
[101] A. Boix, El conflicto catalán y la crisis constitucional española: una cronología, Croni. Estado Soc. Democr. Derecho 71–72 (2017) 172–181.
[102] D. Karamshuk, F. Shaw, J. Brownlie, N. Sastry, Bridging big data and qualitative methods in the social sciences: A case study of twitter responses to high profile deaths by suicide, Online Soc. Netw. Media 1 (2017) 33–43.
[103] P.R. Spence, K.A. Lachlan, X. Lin, M. del Greco, Variability in twitter content across the stages of a natural disaster: Implications for crisis communication, Commun. Quart. 63 (2) (2015) 171–186.
[104] P. Lei, G. Marfia, G. Pau, R. Tse, Can we monitor the natural environment analyzing online social network posts? a literature review, Online Soc. Netw. Media 5 (2018) 51–60.
[105] R. Ogie, R. Clarke, H. Forehead, P. Perez, Crowdsourced social media data for disaster management: Lessons from the petajakarta.org project, Comput. Environ. Urban Syst. 73 (2019) 108–117.
[106] K. Weller, What do we get from twitter—and what not? a close look at twitter research in the social sciences, Knowl. Organ. 41 (3) (2014) 238–248.
[107] J. Pfeffer, K. Mayer, F. Morstatter, Tampering with twitter's sample API, EPJ Data Sci. 7 (1) (2018).
[108] D. Gaffney, C. Puschmann, Data Collection on Twitter, Peter Lang, New York, 2013.

[109] F. Morstatter, J. Pfeffer, H. Liu, K. Carley, Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose, in: Proceedings of the 7th International Conference on Weblogs and Social Media, AAAI press, 2013, pp. 400–408.
[110] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, Y. Moreno, Assessing the bias in samples of large online networks, Social Networks 38 (2014) 16–27.
[111] K. Joseph, P.M. Landwehr, K.M. Carley, Two 1%s don't make a whole: Comparing simultaneous samples from twitter's streaming API, in: W.G. Kennedy, N. Agarwal, S.J. Yang (Eds.), Social Computing, Behavioral-Cultural Modeling and Prediction, Springer International Publishing, Cham, 2014, pp. 75–83.
[112] F. Morstatter, J. Pfeffer, H. Liu, When is it biased? assessing the representativeness of twitter's streaming api, in: WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web, Association for Computing Machinery, Inc., 2014, pp. 555–556.
[113] D. Kowald, E. Lex, Studying confirmation bias in hashtag usage on twitter, CoRR abs/1809.03203. arXiv:1809.03203.
[114] F. Morstatter, H. Liu, Discovering, assessing, and mitigating data bias in social media, Online Soc. Netw. Media 1 (2017) 1–13.
[115] A. Hino, R.A. Fahey, Representing the twittersphere: Archiving a representative sample of twitter data under resource constraints, Int. J. Inf. Manage. 48 (2019) 175–184.
[116] H. Hamdan, P. Bellot, F. Bechet, Lsislif: Feature extraction and label weighting for sentiment analysis in twitter, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEVal 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 568–573.
[117] Y. Miura, S. Sakaki, K. Hattori, T. Ohkuma, TeamX: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEVal 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 628–632.
[118] J. Ah-Pine, E.-P. Soriano-Morales, A study of synthetic oversampling for twitter imbalanced sentiment analysis, in: DMNLP@PKDD/ECML, 2016, pp. 17–24.
[119] I. Alfina, D. Sigmawaty, F. Nurhidayati, A.N. Hidayanto, Utilizing hashtags for sentiment analysis of tweets in the political domain, in: Proceedings of the 9th International Conference on Machine Learning and Computing, ICMLC 2017, ACM, New York, NY, USA, 2017, pp. 43–47.
[120] L. Clarke, Catalonia has created a new kind of online activism. everyone should pay attention, WIRED. URL https://www.wired.co.uk/article/barcelonia-riots-catalonia-protests-news.
[121] I. Cohen, A. Bronstein, F.G. Cozman, Online Learning of Bayesian Network Parameters, In Report No. HPL-2001-55, HP Labs, 2001, pp. 1–8.
[122] S. Lim, S.-B. Cho, Online learning of bayesian network parameters with incomplete data, in: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 309–314.
[123] J. Liu, Q. Liao, Online learning of bayesian network parameters, in: 2008 Fourth International Conference on Natural Computation, Vol. 3, 2008, pp. 267–271.
[124] D.T. Pham, G.A. Ruz, Unsupervised training of bayesian networks for data clustering, Proc. R. Soc. A 465 (2109) (2009) 2927–2948.
[125] M. Muller, S. Guha, E.P. Baumer, D. Mimno, N.S. Shami, Machine learning and grounded theory method: Convergence, divergence, and combination, in: Proceedings of the 19th International Conference on Supporting Group Work, GROUP '16, ACM, New York, NY, USA, 2016, pp. 3–8.

Gonzalo A. Ruz received his B.Sc. (2002), P.E. and M.Sc. (2003) degrees in Electrical Engineering from Universidad de Chile, Santiago, Chile. He then completed his Ph.D. degree (2008) at Cardiff University, UK. Currently, he is a Professor and Director of the Complexity Research Center, at the Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Santiago, Chile. His research interests include machine learning, evolutionary computation, data mining, gene regulatory network modeling, and complex systems.

Pablo A. Henríquez received his M.Sc. degree (2015) in physics from Universidad de Concepción, Chile. He then completed his Ph.D. degree (2019) in Complex Systems Engineering from Universidad Adolfo Ibáñez, Chile. His research interests include neural networks, data mining, and data science.

Aldo Mascareño received his B.Sc. degree (1992) in Social Anthropology from Universidad Austral de Chile, Chile, and his M.Sc. degree (1996) in Sociology from Universidad Católica de Chile, Chile. He then completed his Ph.D. degree (2001) at University of Bielefeld, Germany. Currently, he is a Professor at the Escuela de Gobierno, Universidad Adolfo Ibáñez, Santiago, Chile, and researcher at Centro de Estudios Públicos, Chile. His research interests include systems theory, complexity theories, socialecological systems, and complex systems.