Journal of Autonomous Intelligence (2022) Volume 5 Issue 1.
doi: 10.32629/jai.v5i1.504

Original Research Article

Perspectives and Challenges of AI Techniques in the Field of Social Sciences and Communication

Raul Ramos Pollán*

Department of Systems Engineering, Antioquia University, Colombia

ABSTRACT

In the past decade, the methods and technologies of artificial intelligence (AI) have made great progress. In many cases, they have become part of the usual landscape of solving new or old problems in different fields of human knowledge. In this progress, there are several aspects, especially three aspects: the availability and universality of data in many fields of human activities; a deeper understanding of the mathematics of the basic control algorithm; and the availability and capability of hardware and computing which allows a wide range and a large number of data experiments. Considering these aspects, the key challenge for each problem and application area is to understand how to use these technologies, to what extent they may reach, and what constraints need to be overcome in order to obtain beneficial results (in terms of production cost, value, etc.). This challenge includes identifying data sources and their integration and recovery requirements, the necessity and cost of acquiring or constructing tag data sets, volume data required for measurement, verifying its feasibility, technical method of data analysis task and its consistency with the final application goal, and social and communication sciences are no exception. The knowledge in these fields is related to artificial intelligence, but they do have particularities that define the most appropriate type of artificial intelligence technology and method (i.e. natural language processing). The successful use of AI technology in these disciplines involves not only technical knowledge, but also the establishment of a viable application environment, including the availability of data, the appropriate complexity of tasks to be performed, and verification procedures with experts in the field. This paper introduces the methodology of generating artificial intelligence model, summarizes the artificial intelligence methods and services most likely to be used in social and communication sciences, and finally gives some application examples to illustrate the practical and technical considerations in this regard.

Keywords: *Artificial Intelligence; Machine Learning; Social Science; Data Science*

ARTICLE INFO
Received: Dec 21, 2021
Accepted: Feb 19, 2022
Available online: Mar 2, 2022

***CORRESPONDING AUTHOR**
Raul Ramos Pollán
raul.ramos@udea.edu.co;

CITATION
Ramos Pollán R. Perspectives and challenges of ai techniques in the field of social sciences and communication. Journal of Autonomous Intelligence 2022; 5(1): 34-42. doi: 10.32629/jai.v5i1.504

1. Introduction

The main purpose of this paper is to provide judgment basis for evaluating the application and possible scope of machine learning methods in the field of social science and communication. It is generally believed that the term artificial intelligence was born in 1956. In the Dartmouth artificial intelligence summer research project[1], many people believe that this is a basic event in this field. Since then, its meaning and all the methods and technologies that make up it have developed in different ways and are occasionally encountered in other fields (computing, statistics, linguistics, biology). Therefore, it has experienced various changes in public opinion: from doubt to overestimation of its possibility, to rejection, panic and admiration.

At present, most of the efforts and applications of artificial intelligence are based on the field of machine learning (we often call it ML). ML is a set of techniques and methods, mainly mathematics and statistics, which enable us to use observed data to calibrate the model. These models can be predictive or descriptive.

Many ML technologies have been known for decades, especially since the 1940s (neural network, regression technology, etc.), and many other technologies have been developed in recent years. Three aspects determine the wide and successful application of ML in different knowledge fields in the past decade:

- The availability and diversity of data widely exist in many fields of human activities.
- A deeper understanding of the mathematics of basic control algorithms.
- The availability and capability of hardware and computing allow extensive experiments on a large amount of data.

In each application case, the usefulness of these models depends on various factors, but ultimately, they all depend on our ability to accurately determine what we want from our prediction model, design effective data collection and labeling activities, and form a comprehensive multidisciplinary team.

For example, when the perceptron was invented in 1957[2], people believed that "it will be the embryo of an electronic computer, and people want it to be able to walk, speak, see, write, copy and be aware of its own existence"[3]. At present, compared with other algorithms, perceptron is a relatively simple binary classifier. However, in the long run, we may see now that these statements are too naive, bold or unnecessarily shocking. Similarly, now we must carefully judge the scope of artificial intelligence we envisage, and adhere to the effective and strict use of the mathematical and technical tools we have.

2. Machine Learning

2.1 Data derived model

Mankind has been developing the mode of world operation. Many models are only intuitive (i.e. if I throw this mastodon into a trap, it will fall down and we can have dinner today). Many others have reached a higher level of formalism (i.e. classical mechanics or species evolution theory). These models are more or less derived from the deep understanding gained by a person or a group of people through experiment, learning or intuitive development.

The model is implemented by a process through which prediction is made from an input. For example, given the angle of the barrel, the weight and wind of the bullet, if the actual situation is sufficiently similar to the ideal situation assumed in the model, classical mechanics will predict where the bullet will fall. In the case of classical mechanics, this process is based on the use of differential equations to simulate all aspects of physical processes considered to occur (gravity, resistance). In the case of thermodynamics, this process is based on probability distribution, etc.

In the case of ML, data is used to obtain a model that also attempts to capture and understand the functions of some aspect of the world as shown in **Figure 1**. For example, data is used to predict what the customer will buy next, or estimate the deviation in the text expression. In fact, ML does not generate models in an expert way (i.e. as differential equations). On the contrary, in ML, a kind of model (i.e. neural network, decision tree or support vector machine with given layer structure) is selected and calibrated with data and computing power. This is a fundamental difference involving the following aspects:

- The experimenter must choose the model he wants to calibrate with the data.
- The data are treated as samples to illustrate how the model we want to calibrate responds to the input.
- The calibration process (model training or fitting) is limited to the model category selected by the experimenter.
- The complexity of the selected model class must be similar to the task to be solved. The experimenter must generate experimental evidence in this regard and take corresponding actions (i.e. increase or reduce the complexity of the selected model).
- If the data does not represent the problem to be solved (because they are few, of poor quality or highly mixed, etc.), the experimenter's goal is to generate evidence to

decide how to continue his project (i.e. collect more data, collect more information from each data, or improve the label of the data).

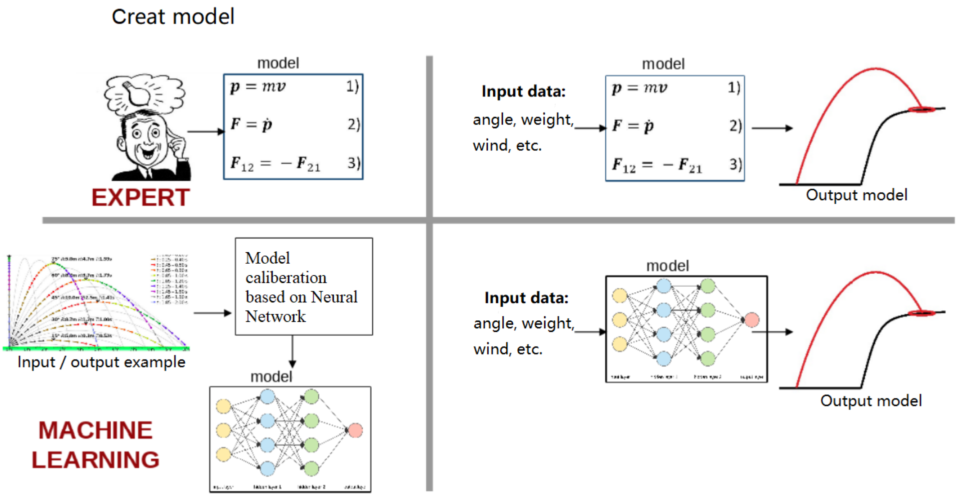

Figure 1. Create models using ML and experts (source: self-developed).

It is in this sense that explains the concept of learning, part of the word machine learning: calibrating algorithms with data. This is quite different from the concept of learning in the sense of human learning. It involves the emergence of a goal, learning plan, theory and practice.

In any case, it should be noted that, like scientific methods, establishing data-based models through machine learning methods is a basically experimental process, which requires observation (data capture), hypothesis creation (model calibration), verification (using new data), hypothesis recalibration (selecting more or less complex model classes), etc., as shown in **Figure 2**.

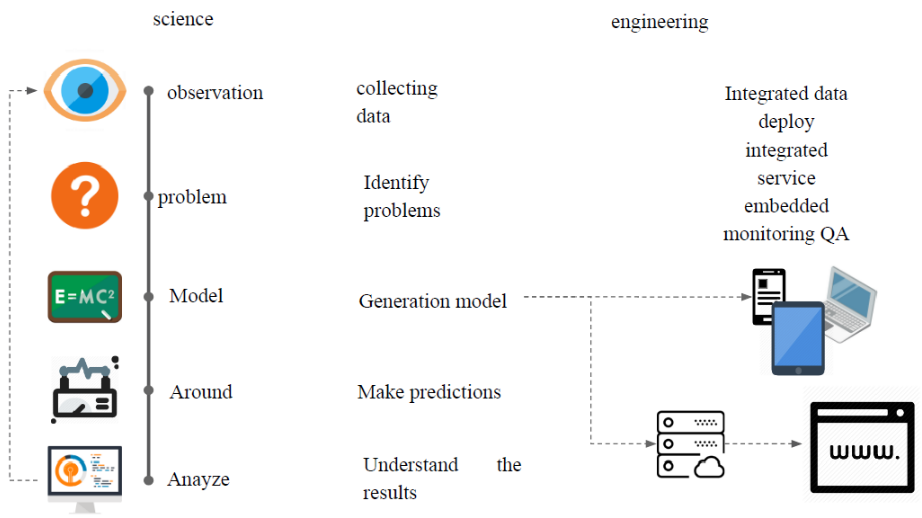

Figure 2. Experimental process model generation (source: self compiled).

Once we have a calibrated model and a proper assessment of the prediction level, it will be expanded in production (**Figure 2**, right). For example, in a web application where customers buy new products, or in a mobile application where they buy new data.

2.2 Machine learning task type

There are three main branches of computational learning, which distinguish the calibration of models according to the way data are provided. These models are described below.

Supervised learning: each input data is associated with a tag or comment that represents the expected output you want from any model. Its purpose is to generalize a mapping function, which can predict the correct labels of new input data when they are received. Supervised learning can be divided into two categories:

- **Classification:** in this kind of problem, the output variables are discretized into different categories existing in the input data. For example, given a text, we need a model to predict whether the text is biased against topic A or topic B, or given an image of a street, we need a model to tell us whether it contains one or another business type.
- **Regression:** in these problems, the output variables are continuous. For example, the possibility of customer default, real estate price or the height of shell hit.

The most commonly used monitoring algorithms include K-neighborhood, artificial neural network, support vector machine, Gaussian classifier and decision tree.

Unsupervised learning: the data does not contain corresponding tags, and patterns or similarities are searched among the data. In this case, interpreting and using the model output is much more difficult because there is no comparable reference. In addition, the experimenter must provide some standards to guide the calibration process of the model, for example, similarity measurement between data, methods of identifying anomalies, etc. Clustering, correlation or estimating probability distribution are some of the techniques used.

Some examples of unsupervised learning algorithms are as follows: principal component, singular value decomposition, clustering, Gaussian mixture, self-organizing mapping.

Reinforcement learning: an agent learns to behave in an environment when he takes actions and sees the results of these actions, from measurement to reward. The data does not contain labels, and the experimenter must define a mechanism to evaluate the applicability of the model output.

Dynamic programming or Q-learning are some of these algorithms.

2.3 Construction of tag dataset

Generally speaking, the calibration model can be used more directly and pertinently through supervised learning technology. By specifying the expected output, our expectations for the model are not ambiguous, so the training or calibration process will better meet our needs.

However, the cost of establishing tag data is considerable. Experts must examine the data one by one to define the labels they want the prediction model to emit. From a certain point of view, we are establishing a complementary method to obtain expert knowledge. Instead of waiting for this formal prediction model (differential equation), we try to express our knowledge with examples: labeled data.

In fact, the publication and availability of open label datasets is a major breakthrough in any field, because it allows research and experiments on datasets, and it is possible to show the results to the community, which will produce broader common knowledge in the long run.

In any case, before starting to solve any problems through data-driven gene models, the following questions must be raised:

- Is there a marked dataset that we can use? Do these datasets represent the problem I want to solve?
- If there is no representative dataset, can I create my own tag dataset?
- How much does it cost?
- Finally, how many data tags do I need? What information should I collect for each data?

There are no absolute answers to these questions. Only through the experiment and iterative process shown in **Figure 2** (left), can we gradually answer these questions i.e. start with an initial data set. The ML method allows us to generate evidence about the initial performance of our calibrated model (i.e. 76% ± 5% classification hit rate) and allows us to verify some performance assumptions if the data increases or enriches the same data we already have. On the basis of this process, a second data collection campaign is usually proposed, which may be more targeted (i.e. we will collect more text on Topic A, because topic A seems to be a topic with low model predictability) and have a better understanding of the cost. We will experiment again with ML method, which may improve the reliability of our results and adjust our cost estimation.

3. Application of ML in Social Science and Communication

The advanced culture of ML community is based on the high attention paid to data publishing and participation in competition. In some areas, there are established competitions dedicated to activities, such as:

- Imagenet Large-scale Visual Recognition Challenge[4] for general image analysis tasks (positioning, detection, segmentation, etc.).
- Speech Recognition and Separation Challenges1 in speech recognition and interpretation tasks.
- ImageCLEF, The Cross Language Evaluation Forum Image Retrieval Track. It contains several challenges to support visual media analysis, indexing, classification and retrieval.
- The challenges of medical image analysis range from the detection of various cancers to histopathology to neurology.

Similarly, Kaggle4 is an ML competition library, which brings together a wide range of fields (finance, satellite images, marketing, medicine, physics, etc.), and has been widely welcomed in recent years. Each competition establishes: (1) the task goal or prediction model to be developed; (2) data sets marked for calibration of our own models; (3) a mechanism through which participants send prediction data to the owners of each competition so that they know the expected actual output; (4) the predicted performance indicators submitted by the participants.

There are various competitions, like the complexity of data, the need of preprocessing, and, similarly, performance indicators. Sometimes, capabilities produce relatively clean data (a few csv tables) or standard measures (prediction success rate, average quadratic error). In other cases, due to the nature of the problem, the data are complex (i.e. images with relevant metadata and time series), and the indicators are very specific (i.e. the percentage of customers and products weighted by economic sector in the competition of demand forecast for some products).

For example, here are some existing (still open or past) competitions related to the dissemination of social sciences:

Two sigma: use news to predict stock movements[1]: this challenge proposes to use natural language processing technology to analyze text news flow to generate prediction signals in financial instruments. The data includes the historical value of a series of financial instruments (stock quotations) and the historical financial news released by Thomson Reuters, as well as various indicators extracted from each news (companies referred to in the news, positive or negative trends, etc.). The objective of the challenge is to publish value and price trend forecasts for each financial instrument. The data is naturally marked from the history itself, because it is a time series, and its purpose is to predict the price of the next day.

Quora insincere questions classification[2]: this challenge has led to a model for identifying insincerity in the forum. According to the description of the challenge, the goal is to identify toxic and fragmented content, build trust among users, and share doubt and knowledge. The training tag dataset consists of about 1.3 million questions. If the question is considered to be insincere, it is marked as "1", otherwise it is marked as "0". Quora defines its concept of indirect thinking by generating 1.3 million tags. Note the cost of organizing tag datasets and organizing competitions.

Costa Rical household poverty level prediction[3]**:** the challenges faced by the Inter-American Development Bank (IADB) are used to establish models that can predict the level of poverty in each family. The goal of this challenge is to improve the pattern of distribution of development assistance to populations in real need. The predictions provided by classical statistical methods are not completely

---

[1] For more information, visit: https://www.kaggle.com/c/two-sigma-financial-news

[2] For more information, visit: https://www.kaggle.com/c/quora-insincere-questions-classification

[3] For more information, visit: https://www.kaggle.com/c/costa-rican-household-poverty-prediction

satisfactory, especially considering population growth and wider data availability. The labeled data set that can be used for model training consists of about attributes of 10,000 families and 140 families of individuals. The poverty level of these families and individuals is between 1 and 4. All challenge information comes from the Inter-American Development Bank's own database.

DonorsChoose.org application screening[4]: challenges posed by DonorChoose.org received hundreds of thousands of applications to support primary and secondary school teachers to provide materials and funds for classroom projects. The purpose of the challenge is to generate a model to predict whether project requests are approved or rejected. The dataset is built from the organization's historical database and stores 500,000 projects. Each project is accompanied by the evaluation results of each project, which constitute the data label.

Spooky author identification[5]: in order to construct this challenge, a data set was also created, including about 20,000 excerpts from Edgar Allan Poe, H.P. Lovecraft and M.W. Shelley's novels. In view of this challenge, this paper puts forward the suggestion of establishing prediction model, and determines the corresponding author of terrorists through text excerpts.

Instacart market basket analysis[6]: in this competition, we propose to generate a model to predict what products consumers buy in supermarkets based on recent consumption.

Outbrain click prediction[7]: from this competition, we propose to generate a model to predict what users will click based on their recent internet browsing history. The purpose is to fine-tune the suggestions provided to users to increase their possibility of selecting suggestions in navigation.

StateFarm distracted driver detection[8]: to solve this challenge, a method is proposed to detect distracted drivers from their own images when driving a vehicle. Its purpose is to detect these drivers through computer vision technology so that they can send out corresponding alarms. The dataset contains approximately 22,400 images, each linked to a label in 10 possible categories (normal driving, telephone conversation, drinking, turning on the radio, makeup).

In particular, it is worth noting that there are different ways to construct tag datasets in different projects. In some cases, experts or minimum qualified personnel (Quora, StateFarm) are required to carry out explicit work of labeling. In other cases, labels are generated automatically or semi-automatically by checking historical data (DonorsChoice) or crossing data with other information sources (IADB).

Social Sciences and communication may benefit more from certain types of ML tasks. These examples are far from exhaustive:

Emotional analysis: through this analysis, the text information is related to some kind of intentionality (positive/negative, political tendency, deception or toxicity, etc.). Different studies have explored the application of this analysis in different fields: citizen judgment[5], tourism[6], gender bias[7], social media[8], political trends[9,10], etc.

Keyword extraction: given a text, find and extract the key terms. Some applications include public opinion analysis tools[8], Twitter speech analysis[11], news classification[12], etc.

Image analysis: for example, it is used to infer the dynamics of population and economic activities from satellite images[13], socio-economic strata[14], crime and crime risk detection[15].

In this regard, it is worth noting that the Spanish Society for Natural Language Processing (SEPLN) has carried out its work through seminars such as TASS[16].

4. Conclusion

From the above examples, we can see that there are many potential applications of computational learning technology in the field of social science and communication. In general, from the perspective of

[4] For more information, visit: https://www.kaggle.com/c/donorschoose-application-screening
[5] For more information, visit: https://www.kaggle.com/c/spooky-author-identification
[6] For more information, visit: https://www.kaggle.com/c/instacart-market-basket-analysis
[7] For more information, visit: https://www.kaggle.com/c/outbrain-click-prediction
[8] For more information, visit: https://www.kaggle.com/c/state-farm-distracted-driver-detectio n

data with quality labels and an excellent professional team, the possibility of quality model based on data and almost any knowledge field is very high. The author believes that the factors restricting the success and feasibility of these models are a series of factors related to organizational ability and the rigor of process methods. In these areas, the following are the key:

Purpose of task definition and prediction model: the degree of specificity of the prediction model is the first contributing factor to the effective use of the model. For example, if we want to generate a model that detects unsafe driving patterns, we need to define what is an unsafe driving demonstration and what parameters we use to describe it. If we need a model to detect which social network users have the greatest impact on other users, we need to define indicators to measure this impact. Increasingly, these definitions are implicitly implemented by providing a tagged dataset (see Quora's challenge example, the previous name). This process simplifies the task of defining itself, but increases the project cost because of the marking process. In most cases, this degree of concretization cannot be achieved immediately at the beginning of the project, because it may be based on the intuition of experts or people familiar with the field (i.e. "I have an intuition that most opinions are put forward by a few people"), but we are not sure how to parameterize it. It is important to be aware of this and propose the project iteratively in order to materialize in the gradually raised questions and assumptions. Initially, through exploratory analysis of these data, we can determine the problem according to our understanding of the data, and gradually produce evidence about the prediction level that can be achieved by the automatic model, which can be realized from a small number of data. Otherwise, we may invest disproportionate resources to generate data sets with labels, which do not fully answer our questions, and only generate prediction models that we do not know how to interpret their output, and so on.

Availability and cost of data acquisition and tagging: as mentioned above, without high-quality data, it is difficult to establish a useful prediction model. The availability of such data and its repair cost is a factor that needs to be evaluated at the beginning of the project. Specifically, this task needs to be performed if data collection is dependent on a third party (i.e., because the data is stored in another organization, and protocols or access rights are required, or simply because communication with the data administrator fails). In many cases, the cost of acquiring and tagging data may account for a large part of project funds and eventually make it infeasible.

Multidisciplinary team building: effective multi-disciplinary is also essential for successful application data analysis projects. However, multi-disciplinary is difficult to achieve, which depends on the background and culture of each specialty and knowledge area in the project. Usually, in order to solve such problems, a team needs: (1) experts from the field of the project (sociology, criminology, politics, biology, etc.), who may have different levels of experience to provide insights and interpret results, and can play a role in data collection and management; (2) scientists and data engineers who have the ability to generate models and interpret them together with domain experts and generate statistical evidence, so as to make decisions on the evolution of models and data acquisition equipment; (3) computer engineers to integrate systems and support other research products (web applications, mobile applications, servers). It is essential that everyone on the team be sensitive and get involved in other people's areas of knowledge. That means experts in this field must acquire statistical concepts and at least understand the process of deriving genetic ran models from the above data. On the other hand, data scientists and engineers must understand the principles of the field they are studying. The purpose is to have a common language so that the standards and decisions of the project, even if made independently, can be understood by all members, so as to avoid being regarded as a black box. At the beginning of any project, it is essential to evaluate these aspects through honest analysis of team communication skills. The communication factor seems insignificant, but it is complex in terms of its composition and must be monitored during the implementation of any project.

Establish performance indicators: performance indicators objectify the usefulness of model output and ultimately enable us to make decisions on data collection, model generation, result analysis and other processes. There are two types of performance indicators: performance indicators that directly measure and predict the performance of the model (i.e. predict the success rate of the next product that customers will buy), and those indicators that measure the impact on the overall goal pursued (i.e. by using a model to predict what economic benefits do I get when I use a model to predict that the next customer will buy products with an 87% success rate). Both measurement types need to be aligned, and in many cases, their definitions are also iteratively built in ML projects. It is important to realize that we have the definition level of performance indicators at every moment. At the beginning of the project, we may have an incomplete intuition on how to measure the performance of the model and its impact on the project. This view is normal, but specific targets need to be set to adjust these indicators. Similarly, although this factor seems insignificant, few projects start with a very specific definition of these indicators and knowledge is growing.

Calculate the availability and cost of infrastructure: this factor also includes storage, computing, and communications infrastructure (i.e., transferring several gigabytes of data sets from one place to another), whether in the cloud or locally.

In short, it is the integration management of these aspects that ultimately determines the success of ML project. Today, technology enables us to develop prediction systems with amazing or unimaginable results a few years ago. The systems and methods related to artificial intelligence (especially those related to computational learning) are basically statistical systems that use data in a very specific way. These systems themselves do not discover patterns, nor do they detect any objects in any view or image in the text. The operation must be guided by careful design of data sets and experiments. The practical use of these systems will depend on our ability to determine in sufficient detail what we want from them, and whether we have and coordinate the necessary resources, including data (possibly labels) and professionals from various disciplines working in a common and agreed language.

Conflict of interest

The authors declared no conflict of interest.

References

1. Moor J. The Dartmouth College artificial intelligence conference: The next fifty years. AI Journal 2006; 27(4): 87-91. doi: 10.1609/aimag.v27i4.1911.
2. Rosenblatt F. The perceptron—A perceiving and recognizing automaton. Presentation. Cornell Aeronautical Laboratory. 1957. doi: 10.1002/9780470694077.ch9.
3. Olazaran M. A sociological study of the official history of perceptron controversy. Social Studies of Science 1996; 26(3): 611-659. doi: 10.1177/030631296026003005.
4. Russakovsky O, Deng J, Su H, *et al.* ImageNet large scale visual recognition challenge. International Journal of Computer Vision 2015; 115(3): 211-252. doi: 10.1007/s11263-015-0816-y.
5. Etter M, Colleoni E, Lllia L, *et al.* Measuring organizational legitimacy in social media: Assessing citizens' judgments with sentiment analysis. Business & Society 2018; 57(1): 60-97. doi: 10.1177/0007650316683926.
6. Alaei A, Becken S, Stantic, B. Sentiment analysis in tourism: Capitalizing on big data. Journal of Travel Research 2019; 58(2): 175-191. doi: 10.1177/0047287517747753.
7. Thelwall M. Gender bias in sentiment analysis. Online Information Review 2018; 42(1): 45-57. doi: 10.1108/OIR-05-2017-0139.
8. Chang YC, Lee FY, Chen CH. A public opinion keyword vector for social sentiment analysis research. 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI); 2018 Mar 29–31; Xiamen. IEEE; 2018. p. 752–757.
9. Gómez-Torres E, Jaimes R, Hidalgo O, *et al.* (2018). Influence of social networks on the analysis of sentiment applied to the political situation in Ecuador. Enfoque UTE 2018; 9(1): 67-78.
10. Arcila-Calderón C, Ortega-Mohedano F, Jiménez-Amores J, *et al.* Análisis supervisado de sentimientos políticos en español: Clasificación en tiem-

po real de tweets basada en aprendizaje automático [Supervised political sentiment analysis in Spanish: Real-time twitter classification based on automatic learning]. El profesional de la información 2017; 26(5): 1699-2407. doi: 10.3145/epi.2017.sep.18.

11. Stegmeier J, Schünemann WJ, Müller M, *et al.* Multi method discourse analysis of twitter communication: A comparison of two global political issues. In: Scholz R (Editor). Quantifying approaches to discourse for social scientists. Palgrave Macmillan, Cham; 2019. p. 285-314.
12. Pranav A, Sukiennik N, Hui P. Inflo: News categorization and keyphrase extraction for implementation in an aggregation system. 2018.
13. Ma T. Multi-level relationships between satellite-derived nighttime lighting signals and social media-derived human population dynamics. Remote sensing 2018; 10(7): 1128. doi: 10.3390/RS10071128.
14. Diou C，Lelekas P, Delopoulos A. Image-based surrogates of socio-economic status in urban neighborhoods using deep multiple instance learning. Journal of imaging 2018; 4(11): 125. doi: 10.3390/Jimaging4110125.
15. Bachhety S, Singhal R, Rawat K, *et al.* Crime detection using text recognition and face recognition. International Journal of pure and applied mathematics 2018; 119(15): 2797-2807.
16. Martínez-Camera E, Díaz-Galiano MC, García-Cumbreras A, *et al.* (2017). Overview of TASS 2017. TASS 2017: Workshop on Semantic Analysis at SEPLN; 2017. p. 13-21.