EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 152 No. 4
April 2023

Research in Computing Science

Research in Computing Science, Año 22, Volumen 152, No. 4, abril de 2023, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 01 de abril de 2023.

Research in Computing Science, year 22, Volume 152, No. 4, April 2023, is published monthly by the Center for Computing Research of IPN.

Intelligent Learning Environments

María Lucía Barrón Estrada
Ramón Zatarain Cabada
María Yasmín Hernández Pérez
Carlos Alberto Reyes García (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2023

ISSN: in process

Formerly ISSNs: 1870-4069, 1665-9899

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Editorial

Nowadays, students are used to utilize virtual learning environments with several kinds of activities to learn new topics, collaborate with peers, solve exercises, do homework and much more. The pandemic of the year 2020 make all of us get used to participate virtually in all kinds of activities, for students the most important is to continue their learning process. Today many learning activities are performed in virtual environments and students of all educational levels use electronic and mobile devices to access educational technology.

Artificial intelligence (AI) has influenced a wide range of activities, and as a result, people is doing things differently. Many activities that people used to do manually or plan ahead now are solved dynamically as needed, like getting a route in a map to go from one place to another.

IA is part of new platforms and applications that are emerging in education to enable students to study in different setting such as home or schools. A very important example today is the use in home or schools of the ChatGPT application to produce texts such as essays, papers or source code using several programming languages.

On the other hand, to create modern learning environments that capture students' attention, researchers need to add such technologies to tailor the material presented to students according to their needs.

In this volume, we present extended versions of seven research works in some fields of intelligent learning systems. In WILE 2022, our goal is to offer researchers an opportunity to show their work exploring new ways of applying AI techniques in the development of educational systems. The papers presented in WILE 2022 were carefully chosen by the editorial board based on three reviews by the specialists of the Technical Committee. The extended versions were additionally reviewed. To select the papers to be published the reviewers considered the originality, scientific contribution to the field, soundness, and technical quality of the papers.

Many people participated in WILE 2022, we appreciate the support of RedICA (Conacyt Thematic Network in Applied Computational Intelligence) members that were part of the Technical Committee as well as members of Mexican Society for Artificial Intelligence (SMIA Sociedad Mexicana de Inteligencia Artificial). As every year MICAI 2022 was the appropriate host for this event. Last, but not least, we thank Centro de Investigación en Computación-Instituto Politécnico Nacional (CIC-IPN) for their support in preparation and publication of this volume.

María Lucía Barrón Estrada
Ramón Zatarain Cabada
María Yasmín Hernández Pérez
Carlos Alberto Reyes García
October 2022

ISSN 1870-4069

An Overview of Gender Recognition Systems Using Acoustic Voice Analysis

Enrique Díaz-Ocampo, Andrea Magadán-Salazar,
Raúl Pinto-Elías, Máximo López-Sánchez

Tecnológico Nacional de México CENIDET,
Cuernavaca, Morelos,
Mexico

{m22ce002,andrea.ms,raul.pe,maximo.ls}@cenidet.tecnm.mx

Abstract. Intelligent learning environments are a category of academic software that implements artificial intelligence techniques to assist a student in learning. However, to have a greater impact on their learning, it is required to personalize the learning process. For this it is necessary to determine certain characteristics and traits of the student. One example of these characteristics that can be inferred from the learner's speech is gender. Gender differences in voice presents the essentialist and the constructivist view. The essentialist view explains the differences between male and female voices from anatomical differences, and the second one from socially learned behavior. Voice gender recognition systems is a term that refers the automatization of gender detection by an acoustic signal of voice. Most voice gender recognition systems only employ the essentialist approach of gender. The present article is an overview of the voice gender recognition system's development. From some classic models developed in the nineties, up to those presented in recent years. These systems developed five characteristics over the years, that will be exposed in this work and could be implemented as auxiliary tools in intelligent learning environments in order to extract features from students.

Keywords: Voice gender recognition, voice activity detector, ensembled methods, pitch, fundamental frequency.

1 Introduction

Humans developed a biological and complex system, that allow them to obtain paralinguistic information (emotion, gender, age, health, pitch, etc.) from voice. However, the process of automating this task has required several years of biological, mathematical, and computational research.

From the earliest works [1, 2] in the 1990s to the present year, there have been multiple designs of speech-based gender recognition systems that implement artificial intelligences. Such is the case that intelligent assistants personalize themselves based on the tone of the speaker. But, there have been few overviews of these systems (see [3, 4]).

In particular, overviews of these systems are required that expose the type of features and computational tools that can be used to extract them and thus can be used in different environments like Intelligent learning environments.

In this paper, an overview of gender recognition systems using acoustic voice analysis was carried out. The methodology for summarizing is based on [5]. The relevance of this topic lies in the fact that the extraction of paralinguistic information from the user is necessary to be able to design systems that achieve a higher degree of affinity with the user such as those used in intelligent learning environments. In this way, new academic proposals can be explored and those already in use can be evaluated according to the characteristics of the learner.

This paper is organized as follows. In Section 2 the description of the research protocol will be presented, where 4 questions relevant to the research of voice gender recognition systems will be presented. Subsequently, the selection and extraction of documents in two search engines: Google Scholar and Science Direct is presented. In Section 3 the results obtained are discussed and answers to each of the questions are provided by means of a detailed explanation based on the articles found. Finally, in section 4 will provide the most relevant aspects of this work and future work.

2 Description of the Research Protocol

The research protocol is based on the search for information to answer the following guiding questions:

- What are the biological features used in gender recognition systems by voice?
- What characteristics determine the design of a voice gender recognition system?
- What phases are implemented voice gender recognition systems?
- What are the current trends in voice gender recognition systems and what aspects characterize them?

The first was conceived because there are various combinations of characteristics for gender detection. However, if biological characteristics are used, certain traits of a person can be determined beyond their gender. The second question is about recent architectures in gender recognition systems. Finally, the third question is about the breakdown of the phases implemented in this type of systems.

2.1 Document Selection and Data Extraction Process

This section describes the search criteria and tools used in the research. Two search engines were used: Google Academic and Science Direct. The breakdown of the class words used is presented below:

- (“Acoustic analysis” OR “análisis acústico”) AND (“voice analysis” OR “análisis de voz”) AND (“gender” OR “género”).
- ((“Acoustic analysis” OR “análisis acústico”) OR (“voice analysis” OR “análisis de voz”)) AND (“gender” OR “género ”).
- “Deep Learning Gender Classification” OR “voice gender classification”.

– (“voice” AND “gender” AND “score”) AND “fuzzy”.
– (“voice” AND “gender” AND “score”) AND “machine learning”.
– “Voice” AND “background noise” AND “gender recognition” AND “analysis”.
– “Voice” AND “background noise” AND “speaker gender recognition”.
– “Voice” AND “noise” AND “gender recognition” AND “uncontrolled”.
– “Voice” AND “noise” AND “speaker gender recognition”.

The criteria for elimination were as follows:

– Papers published from 2018 to 2022.
– Only peer-reviewed articles.
– Only articles in Spanish and English.
– Available in pdf format.
– Articles that will show the architecture of their systems.
– Articles using public or accessible databases.
– Articles that will use free software or similar for its processing.

3 Results

Using the criteria established in the present work, 31 articles relevant to the present work were found. They were subsequently reviewed in terms of both their contributions and their references in order to provide information for the research questions posed in this study.

3.1 Biological Features in Voice Gender Recognition Systems

During the 1990s, the first investigations consisted of investigating the correlation of acoustic measures of laryngeal activity and the phonetic structure of healthy people. In [6], the latter was investigated by choosing various vowels of the English language pronounced by a group of 8 people (4 women and 4 men) between 20 and 45 years old.

Acoustic measurements: Fundamental Frecuency or F_0 (frequency of vibration of the vocal folds), Jitter (Frequency variation of F_0), Shimmmer (mean percent change in waveform amplitude amount period), Signal-to-noise ratio (decibel ratio of total energy in the acoustic speech signal to the energy in the aperiodic or noise component), Harmoninc-to-noise-rato (ratio between periodic and non-periodic components of a speech signal, articulate rate (number of syllabes per second), number and duration of speech pauses. The most representative results to distinguish the gender were F_0 and jitter.

Despite the studies focused on the fundamental frequency for the detection of gender, in general, it is not a decisive factor. This is due the characteristics of flexibility of the vocal cords and dimensions of the vocal tract. In the 2000s, several parameters were defined to be able to make the gender distinction between people of different ages and health conditions [7]. Here are some examples:

– Prosodic: accent, stress, rhythm, tone, pitch, and intonation.
– Acoustic: Fundamental frecuency, mean of frequency spectrum, standard deviation of frequency, duration, formant frequencies, etc.

– Cepstral: Coefficients of linear, Mel, Mel inverse and rectangular frequencies, among others.

In the 2010's, more sophisticated features like formant frequencies, aperiodicity, and spectrum level were used to study voice gender perception [8]. Nowadays, the acoustic voice analysis area is in charge of developing non-invasive techniques to study multiple acoustic parameters obtained by special microphones [9].

3.2 Systems of Gender Recognition through Voice Analysis

One way to understand the development of gender recognition systems through acoustic voice analysis is from the following points:

– Characteristics used in the classifier.
– Voice recording and multi-label environment.

The acoustic characteristics are analyzed with classical statistical techniques. For example, scatter plots, probability distributions and correlation matrices are used for each gender feature set [10]. Also, each acoustic parameter is given a weight w obtained by the implemented classifiers [10]. Futhermore, papers focused on acoustic features use multiple classifiers and ensemble techniques to solve the classification problem [11]. This is a heuristic method to avoid overfitting.

On the other hand, cepstral features are common in deep learning algorithms. In [12], multiple combinations were explored in a multilayer perceptron neural network. While in [13], multiple normalization techniques for cepstral coefficients were compared to improve classifier metrics. In addition, due to their vector form, they have been implemented with tensor analysis [14]. However, the trends in recent years have been to design more sophisticated networks (e.g., temporally convolutional) and using several thousand coefficients to improve the accuracy of gender recognition [15].

The voices used by a gender recognition system can be in controlled and uncontrolled environments. In the first case, specialized microphones are used, each audio lasts the same length and each voice reproduces a specific text [9, 16]. In the second case, the voices come from public voice sets, where each user can upload his own recorded audio. An example of such databases are Mozilla Common Voice Dataset [17].

3.3 Phases of a Voice Gender Recognition Systems

The phases of a classical voice gender recognition system consist of the choice of a dataset, a preprocessing phase, a data extraction phase, obtaining the feature vectors, the implementation of the classifier and the evaluation of the metrics. The first one process consists of removing as much noise as possible from our samples in order to obtain clean signal for further processing.

Then, feature extraction is performed. Among the most popular are the mel frequency cepstral coefficients (MFCC) [16] and acoustic parameters [18]. Since some features has more relevant information that the others, it is necessary to implement a feature selection [19]. The feature vectors feed the classifier or assembly techniques [20]. Finally, the metrics will depend on the condition of balanced or unbalanced dataset.

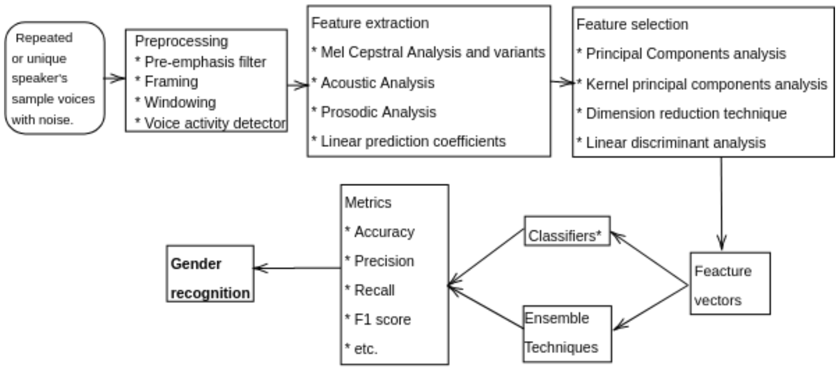


Fig. 1. General design of a gender recognition system.

3.4 General Trends in Gender Recognition Systems

A basic gender recognition system is compound by two elements [21]. The first element is a front-end system, that is designed for extract the relevant features in order to create a template model of voice's gender. The second element is a back-end system, which is the classifier algorithm implementation.

The first models for gender recognition in the 1990's, were based on vowel's analysis via statistical, probabilistic, and multilayer perceptrons models [22]. In 2000's, the trend was developing in systems with multiple probabilistic classifiers, mixed parameters and specific acoustic features. For the 2010s, parameters were built that allowed to distinguish the gender of a person but at the same time were not conditioned to language [23].

Current voice gender has multiple and complex modules. Figure 1 presents in general terms the current designs of voice gender recognition systems. In the following subsections it is discussed in depth and how to meet these requirements.

3.5 Five Main aspects in a Voice Gender Recognition System

Sample Voices. The datasets to be used for training and learning validation require handling the scenarios of repeated voices and unique voices. Since a database with multiple audios of a single person will present a bias towards the gender of that person, but a database of unique voices is more complicated to analyze since they are not common or free to use.

Existing ones generally have more voices of one gender than another (e.g. Mozilla Common Voice). So an analysis in these environments will better expose the strengths and weaknesses of both the features and the classifier algorithm. On the other hand, English language has been one of the most studied in terms of gender detection.

This is due to the availability of databases in that language. However, voice gender detection results have been obtained in multiple languages [24], and in dialects of the same language. Therefore, it is necessary to use databases with a variety of languages used in order to find patterns in them.

Preprocessing. The speech signal can be divided as the sum of a pseudo-periodic signal and a noise signal. It is usually assumed that such noise is stationary (i.e. statistical parameters are constant), so that various techniques can be employed to attenuate it [25]. One example is the use of different filters like pre-emphasis [26, 24]. Which is a time-domain finite impulse response filter with one free parameter.

Framing consists of dividing our signal into parts (intervals from 20 ms to 30 ms), so that the pseudo-periodic part of the speech signal and the stationary property of the noise signal are preserved. These segments are known as frames. For practical purposes, the pre-emphasized signal is blocked into 200 frames each 25 ms long with 10ms frame shift [26].

Windowing is the process of applying a window function to each frame obtained. A more sophisticated way to remove noise is by detecting audio segments where there is a human voice. This technique is known as voice activity detector [16, 27]. It consists of assuming that noise is a random variable that is added to our signal. Therefore, by means of hypothesis tests in each of the frames of our signal, the voice is detected.

Feature Extraction and Feature Selection. Feature extraction is associated with the type of voice emitted by the speakers. If the audios come from controlled environments and what was recorded consists of phonemes, prosodic features are often used. It is also possible to construct multiple statistics from countour pitch [28].

While in uncontrolled environments where there is a greater influence of noise, cepstral features are mostly used because of their robustness to noise [29]. Once the characteristics of interest have been obtained, a classic way of choosing the most relevant ones is from a correlation matrix. However, techniques such as principal component analysis and its variant with kernel modification are also used.

Classifiers and Metrics. Due to the flexibility of the vocal cords, as well as the dimensions of the vocal tract of each person, it is possible that there are voices of different genders but that share a similarity in their vector of characteristics. This causes the classifier to be wrong. This can be made worse if the classifier used assumes that the feature space is linearly separable when it is not.

Therefore, for any machine learning algorithm or deep learning algorithm to be used, it is essential to assume a priori that any set of features derived or associated to F_0 will not be linearly separable. With respect to metrics,since is binary classification problem, precision is mainly used in balanced data and F1-measure is used for unbalanced dataset.

Capable of Recognizing Multi-labels. The voice is subject to change over time. This is visualized in the maturation of the voice in teenagers and its aging in older adults. Therefore gender detection is affected by these conditions. One way to solve the problem is the implementation of systems that detect age and gender independently [30].

A second way is by implementing multiple tags indicating gender and age range. In [31] a convolutional neural network was implemented for the detection of gender and age range and the labels young male, young female, adult male, adult female, senior male, and senior female were implemented.

In the case of [32], a convolutional neural network with a Specially Designed Multi-Attention Module through Speech Spectrograms was implemented for the detection of gender and age in English and Korean. The labels teens, twenties, thirties, fourties, fifties, and sixties were used. Each of these was associated with the male or female gender.

4 Conclusion

In this study, a brief discussion of the development of gender recognition systems was presented. The types of features for such recognition were presented. This work showed the methodology for the search and selection of articles, as well as an exposition of the different characteristics of these systems.

In addition, an overview of the general panorama was provided in the computational study of gender detection by voice. On the other hand, five characteristics of the current design trend of these systems to improve their performance in various contexts were exposed. Future work would include systematic reviews of multi-paralinguistic recognition systems for gender, age, accent and emotion. Based on these four features, it is posible to obtain a broader picture of the speaker interacting with the system.

It is expected that this work will facilitate the implementation of the recognizer in systems associated with intelligent learning environments with the objective of obtaining features that allow the personalization of student learning.

References

1. Childers, D. G., Wu, K.: Gender recognition from speech. Part I: Coarse analysis. The Journal of the Acoustical Society of America, vol. 90, no. 4, pp. 1828–1840 (1991)
2. Childers, D. G., Wu, K.: Gender recognition from speech. Part II: Fine analysis. The Journal of the Acoustical Society of America, vol. 90, no. 4, pp. 1841–1856 (1991)
3. Zhao, H., Wang, P.: A short review of age and gender recognition based on speech. In: 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data an Security (IDS), pp. 183–185 (2019)
4. La Mura, M., Lamberti, P.: Human-machine interaction personalization: A review on gender and emotion recognition through speech analysis. In: 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT, pp. 319–323 (2020)
5. Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., Khalil, M.: Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software, vol. 80, no. 4, pp. 571–583 (2007)
6. Nittrouer, S., McGowan, R. S., Milenkovic, P. H., Beehler, D.: Acoustic measurements of men's and women's voices: A study of context effects and covariation. Journal of Speech, Language, and Hearing Research, vol. 30, no. 4, pp. 761–775 (1990)
7. Bellandese, M. H.: Fundamental frequency and gender identification in standard esophageal and tracheoesophageal speakers. Journal of Communication Disorders, vol. 42, no. 2, pp. 89–99 (2009)

8. Skuk, V. G., Schweinberger, S. R.: Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender (2014)
9. Bensoussan, Y., Pinto, J., Crowson, M., Walden, P. R., Rudzicz, F., Johns III, M.: Deep learning for voice gender identification: Proof-of-concept for gender-affirming voice care. The Laryngoscope, vol. 131, no. 5, pp. E1611–E1615 (2021)
10. Ertam, F.: An effective gender recognition approach using voice data via deeper LSTM networks. Applied Acoustics, vol. 156, pp. 351–358 (2019) doi: 10.1016/j.apacoust.2019.07.033
11. Sahar, R., Rao, T., Anuradha, S., Rao, B.: Performance analysis of ML algorithms to detect gender based on voice. Recent Trends in Intensive Computing, pp. 163–171 (2021) doi: 10.3233/APC210192
12. Kanani, I., Shah, H., Mankad, S. H.: On the performance of cepstral features for voice-based gender recognition. Information and Communication Technology for Intelligent Systems, pp. 327–333 (2018)
13. İleri, S., Karabina, A., Kılıç, E.: Comparison of different normalization techniques on speakers' gender detection. Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi, vol. 2, no. 2, pp. 1-12 (2018) doi: 10.31200/makuubd.410625
14. Roy, P., Bhagath, P., Das, P.: Gender detection from human voice using tensor analysis. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pp. 211–217 (2020)
15. Nandan, V. G., Shivakumar, S., Sangeetha, J., Nayak, M. P., SK, N.: A comparative study of deep learning and machine learning approaches in speech emotion and gender recognition system. Natural Volatiles and Essential Oils, vol. 8, no. 5, pp. 12261–12273 (2021)
16. Alkhawaldeh, R. S.: DGR: Gender recognition of human speech using one-dimensional conventional neural network. Scientific Programming, vol. 2019 (2019)
17. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4218–4222 (2020)
18. Kacamarga, M. F., Cenggoro, T. W., Budiarto, A., Rahutomo, R., Pardamean, B.: Analysis of acoustic features in gender identification model for english and bahasa indonesia telephone speeches. Procedia Computer Science, vol. 157, pp. 199–204 (2019)
19. Badr, A. A., Abdul-Hassan, A. K.: Speaker gender identification in matched and mismatched conditions based on stacking ensemble method. Journal of Engineering Science and Technology, vol. 17, no. 2, pp. 1119–1134 (2022)
20. Kushwah, S., Singh, S., Vats, K., Nemade, V.: Gender identification via voice analysis. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 746–753 (2019)
21. Kaur, S., Garg, D., Arora, D.: Comparative analysis of speech processing techniques for gender recognition. International Journal of Advances in Electrical and Electronics Engineering, pp. 278–283 (2012)
22. Konig, Y., Morgan, N., Chandra, C.: GDNN: A gender-dependent neural network for continuous speech recognition. International Computer Science Institute (1991)
23. Alsulaiman, M., Ali, Z., Muhammad, G.: Voice intensity based gender classification by using Simpson's rule with SVM. In: 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 552–555 (2012)
24. Ponraj, A. S., Naveen M.: Speech recognition with gender identification and speaker diarization. In: 2020 IEEE International Conference for Innovation in Technology (INOCON), pp. 1–4 (2020)

25. Bawa, P., Kumar, V., Kadyan, V., Singh, A.: Noise-robust gender classification system through optimal selection of acoustic features. Deep Learning Approaches for Spoken and Natural Language Processing, pp. 147–159 (2021)
26. Guha, S., Das, A., Singh, P. K., Ahmadian, A., Senu, N., Sarkar, R.: Hybrid feature selection method based on harmony search and naked mole-rat algorithms for spoken language identification from audio signals. IEEE Access, vol. 8, pp. 182868–182887 (2020)
27. Ma, Y., Nishihara, A.: Efficient voice activity detection algorithm using long-term spectral flatness measure. EURASIP Journal on Audio, Speech, and Music Processing, vol. 2013, no. 1, pp. 1–18 (2013)
28. Shagi, G. U., Aji, S.: A machine learning approach for gender identification using statistical features of pitch in speeches. Applied Acoustics, vol. 185, pp. 108392 (2022)
29. Nasef, M. M., Sauber, A. M., Nabil, M. M.: Voice gender recognition under unconstrained environments using self-attention. Applied Acoustics, vol. 175, pp. 107823 (2021)
30. Sánchez-Hevia, H. A., Gil-Pita, R., Utrilla-Manso, M., Rosa-Zurera, M.: Convolutional-recurrent neural network for age and gender prediction from speech. In: 2019 Signal Processing Symposium (SPSympo), pp. 242–245 (2019)
31. Sánchez-Hevia, H. A., Gil-Pita, R., Utrilla-Manso, M., Rosa-Zurera, M.: Age group classification and gender recognition from speech with temporal convolutional neural networks. Multimedia Tools and Applications, vol. 81, no. 1, pp. 3535–3552 (2022)
32. Tursunov, A., Choeh, J. Y., Kwon, S.: Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. Sensors, vol. 21, no. 17, pp. 5892 (2021)

ISSN 1870-4069

Conversational System based on Socratic Dialectics in a Virtual Educational Environment

Bárbara María Esther García-Morales,
María Lucila Morales-Rodríguez, Nelson Rangel-Valdez

Tecnológico Nacional de México,
Instituto Tecnológico de Ciudad Madero,
Mexico

{G10070255, lucila.mr, nelson.rv}@cdmadero.tecnm.mx

Abstract. In this work, a conversational system is presented through a credible socio-emotional agent with the role of a coach. Which involves dialectics as a self-learning process for the coachee, using the types of questions that characterize the technique of the dialectic for the selection of their sentences. The personality type of the coach will be based on the personality of the coachee, which influences the preferences of how it wants to perceive the information within the dialogue and the impact criteria of the information associated with the importance and relevance of the sentences in a Socratic dialogue. In this paper, the ELECTRE III overclassification method is proposed as a strategy to obtain the direct powerful question closest to the agent's preferences, which are associated simultaneously with the coachee's answers. A conversational system with these characteristics can contribute to the development of serious games, modeling virtual entities with the ability to guide users through consistent dialogue, making users find solutions to their problems based on their cognitive possibilities.

Keywords: Conversational educational system, Socratic dialectic, socio-emotional agent with personality, virtual coach.

1 Introduction

In education, various teaching-learning models can include reflection methodologies such as the one based on thinking. This methodology makes use of inductive and deductive methods to carry out learning [1]. In the deductive method, concepts or clues are transmitted to determine an answer within a set of possibilities. The inductive method establishes clear and specific examples or questions for the adviser to identify and reflect on globally until reaching their conclusion [2].

To obtain self-knowledge through inductive learning, the technique of dialectics can be used, which is a rhetorical technique to reason and argue through a dialogue between two individuals. In this type of Socratic dialogue, the coach (advised) uses powerful direct questions based on maieutics or irony. One of the application areas where Socratic dialectics is used as a teaching-learning technique is in coaching processes [3].

Currently, it is possible to emulate coaching processes through a conversational agent based on dialectics, which will allow the creation of a virtual coach, simulating events that in this conversational process could offer the user (coachee) possible solutions to make decisions and solve the problems raised [4].

To give realism to the behavior of the virtual coach based on dialectics, it is necessary to incorporate artificial intelligence techniques that support the behavioral purposes of the coach within the coaching processes, these behaviors can be divided into motivated by emotions [5] and motivated by belief or reasoning [6].

In this case, the modeling of functions that integrate emotions, personality traits, preferences, and motivations, take on great relevance in the selection of powerful questions based on Socratic dialectics.

This research work proposes a conversational system that integrates the technique of dialectics as a strategy for the design of a virtual coach. To incorporate the characteristics of dialectics, it is proposed to ask rhetorical questions in the function of the mirror so that the coachee can reach their solutions, it is proposed that the personality of the coach reflects the personality of the coachee.

That is, the personality of the coachee will define what the personality of the coach will be. Its influence on the selection of questions will be based on the profile of this personality, reflecting a relaxed or strict coach.

2 Related Works

This section presents a brief description of some works related to intelligent virtual agents that use personality traits to give realism and credibility to their sentences through their dialogue management system. These agents are developed based on expert knowledge according to the context of the application area for which their use is intended.

In the educational context the work of Gómez-Róspide [7], which focuses on developing an intelligent virtual agent to help with learning. This agent is intended to resolve academic doubts about a subject through a conversation with a human, being able to understand the questions asked by the student and prepare an answer according to what was questioned.

However, its vocabulary is based on a single context, for example, a topic-specific vocabulary related to the topic of linear algebra. This represents a limitation for the agent if the student questions issue not specified in the context of algebra.

In [8] and [9] a socio-emotional tutor agent is created, that is, through the dialogue they try to guide the user through dialogues that express emotions and personality to guide the interaction.

In the work of Morales-Rodríguez [9], an intelligent emotional character is proposed in a virtual therapy context capable of modifying their emotional state based on the evaluation of it interlocutor's actions and attitudes, that is, It could react positively or negatively to certain stimuli.

It intends to generate a feeling of immersion in the patient through the emotional expression of a virtual social presence, stimulating the patient's participation during therapy. This work involves the approach of personality traits with the FFM model (Five-Factor Model).

Another example of the selection processes that involve taking into consideration emotional characteristics in conversational agents can be seen in the work of Florencia-Juárez [10], in this project, credible dialogues are produced capable of creating real scenarios for the student in psychology and psychiatry diagnose the condition of generalized anxiety disorder simulated by the conversational agent. Florencia uses a deliberative architecture and for personality modeling.

He uses the FFM model. The works of Morales-Rodríguez [9] and Florencia-Juárez [10] present dialogues that are reactively established between the user and the agent. That is, they are previously stored in a knowledge base and selected according to the interaction with the user. This can produce a repetitive dialogue and the person does not reflect on his problem.

To give realism and credibility to each of the sentences of a socio-emotional agent using personality traits, Delgado-Hernández et al. [11] propose a sentence selection model which is responsible for choosing the most appropriate sentence in response to the perceived context within the dialogue.

This model requires as input the corpus and the context characterized according to the personality profile of the agent. The author uses the MBTI model (Myers-Briggs Type Indicator) and the values of the individual's preference thresholds that are integrated into an overclassification strategy called ELECTRE. This work is the closest to the work proposed in this document.

However, the dialogues are characterized manually and are characterized according to the vocabulary of the authors. Therefore, it cannot answer questions that are not stored in its knowledge base, nor can it use more than one personality and preferences profile.

Following the line of investigation of personality models that can be used in socio-emotional agents to show congruent dialogues related to the preferences of how information is intended to be perceived in a particular context, is the Castro-Rivera model [12].

This personality model is known as MPBCD (Decision Context Personality Model) and its decision profiles are: optimistic/relaxed, collaborative, inquiring, and strict. These personality profiles can reflect a relaxed or strict coach according to the perception of information received from the coach.

This work does not contain the construction of dialogs to guide the user. It is worth mentioning that the MPBCD personality model is useful to guide the process of selecting an answer in the Socratic dialogue through the influence of personality on preferences in decision processes.

The works presented in this Section involve conversational social-emotional virtual agents with personality. However, they present limitations in the way in which the dialogue between the user and the agent is carried out. This article proposes a response dialogue generated from a user's phrase, which is displayed as a direct question based on Socratic dialectic.

These dialogues are not stored in a knowledge base, that is, they are automatically generated through the perception of characteristics related to speech acts and Socratic dialectic. In addition, the selection of the answer or question is carried out using the ELECTRE-III strategy, which uses preferential thresholds influenced by personality.

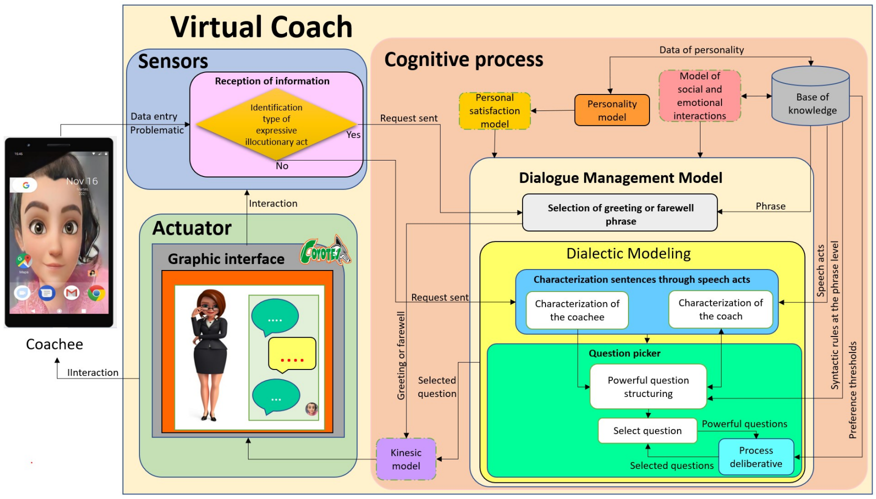


Fig. 1. Proposed architecture of a virtual coach based on dialectics.

These thresholds have been taken up from the work of Castro-Rivera [12] , as well as from personality profiles. These profiles will allow the selection of the direct answer or question to be adapted based on the user's personality detected in their sentence through speech acts.

3 Influence of Personality on the Selection of Questions based on Dialectics in Credible Socio-Emotional Agents

To understand how to give certainty to the behavior of a virtual entity based on the Socratic dialectic, and the influence on the selection of powerful questions through a personality profile, it is necessary to know some concepts related to intelligent virtual agents and the mechanism they use to make decisions based on emotions, personality traits and preferences under a decision context using the overclassification strategy known as ELECTRE III.

3.1 Deliberative Conversational Agent

A virtual agent that emulates human behavior is a deliberative conversational agent. This type of agent incorporates in its architecture the basic principles of most intelligent systems. These are reactivity, internal state, principles, and goals, autonomy, sociability, and reasoning [13]. These principles are related to the symbology of the real world, through beliefs, desires, and intentions.

Where beliefs are related to personality, that is, the agent knows who he is and the context or environment in which he operates. The desires are related to the objectives that pursue the agent, in this case, it is the formulation of questions based on the dialectic influenced by the personality.

The intention is to deliver a suggestion as a question to the coachee to create a reflective environment that allows them to create their solution to their problem. The

Table 1. Example of the phrases said in a Socratic dialogue.

	Phrases		
I	**Coachee**	**Virtual coach**	**TIA**
1	Hi good morning. I have problems with a student in the subject of integral calculus	Hello, good day. What do you think is the reason why you have problems with a student in the subject of integral calculus?	Assertive
2	Possibly a lack of attention in the classroom	How do I possibly think of the lack of attention in the classroom?	Expressive, Commitment

selection of powerful questions in this virtual agent is based on an overclassification method called ELECTRE III.

3.2 ELECTRE III

The ELECTRE model (Elimination and (et) Choice Translating Algorithm) is the best known among the methods developed for multi-criteria attendance problems and is used as a procedure to reduce the size of the set of efficient solutions [13]. ELECTRE III orders different solution alternatives, for which different criteria are generated with a certain weighting [14].

It works with pairs of alternatives and in these, there is a measure of agreement and disagreement to produce the degree of overclassification, this model aims to combine the credibility index that validates the decision-making process when choosing one alternative over another in a problem [13].

To work with the ELECTRE method, the development of a performance matrix is needed. In this matrix, the linguistic variable and the weight of importance must be indicated. After establishing the values of these variables, a matrix must be made to compare the values of the alternatives against the criteria that were previously established for the problem, and to conclude in the selection, the level of agreement between the variables must be taken into consideration [14].

3.3 Personality

In this work, it is proposed to integrate the Castro-Rivera personality model [12] that calculates the parameters that will reflect the behavior and preferences of the virtual coach, which will influence the formulation of the questions based on the Socratic dialectic using the four profiles of decisions proposed in this model. Personality is commonly seen as the set of behaviors that constitute a person's individuality and is used to describe and classify attitude.

Attitude is defined by two inseparable aspects, the affective aspect, and the cognitive aspect. Attitude is defined as the degree of positive or negative effect consistently associated with a person's response [15]. There are two major theories of personality on human behavior, these are the trait-based theory, The Five-Factor Model of personality (FFM-OCEAN), and the type-based theory, Myers-Briggs Type Indicator (MBTI), which are very recurrent in the modeling of behavior in virtual agents [16].

4 Conversational System for a Virtual Coach Focused on Socratic Dialectics

The modeling of a virtual coach requires an architecture supported by agent theory, due to the complexity of the elements involved in its operation, such as the dialogue management model, decision making, and the personality model. The architecture of this work has the function of representing virtual humans with the possibility of performing various roles in education. Figure 1 shows the general architecture of a virtual coach based on Socratic dialectics, which is generally composed of three main components: sensors, actuators, and cognitive process.

The information reception module is located in the sensor component. In actuator components, there is the graphical user interface module and in the cognitive process component, there is the deliberative process associated with the dialogue management model, which is made up of dialectic modeling.

The information reception module receives the problem, that is, the user's sentence. From the problem, it is possible to identify the intention and the type of illocutionary class of this, as well as the phase of the dialogue.

If the type of illocutionary class is of the expressive type, the answer is taken from the knowledge base obtaining phrases of greetings or farewells. Otherwise, a dialectic-based sentence must be determined which will be a powerful question in response to the coachee. The knowledge base module contains the welcome or greeting phrases for the user, the grammar rules at the phrase level, information criteria, and preference thresholds of the virtual agent.

In the modeling of the dialectic, there are the speech and question selector modules. The latter is responsible for generating a reflective conversational environment influenced by the preferences and personality of a virtual agent through direct questions.

The sentence characterization module through speech acts characterizes the sentence of the user (problematic) and the virtual coach through the theory of speech acts. Through this theory, the intention of the sentence is established and the personality type of the coachee is determined. In this way, it generates a powerful question as a response from the coachee, which is structured under the same personality profile.

The question selector module structures powerful questions at the level of phrases analyzing the problems derived from speech acts. Through the phrases and the user's problems, it generates a set of powerful questions.

The powerful questions were born given to the deliberative process so that they select those questions that are closest to the interests of the coach according to their preferences. According to the personality profile of the coach, it is possible to choose from the knowledge base the preferential parameters that guide the deliberative process to select the powerful questions. That is, if the personality is of the optimistic type, the preference parameters or thresholds correspond to an optimistic person.

It is worth mentioning that these preference thresholds are influenced by personality. The calculation of these thresholds is based on the work of Castro-Rivera [12].

Table 2. Comparison of 8 interactions in a Socratic dialogue under the influence of personality.

	Coachee			Virtual coach			Real coach		
I	**TIA**	**TDQ**	**PP**	**PN**	**TDQ**	**PP**	**PN**	**TDQ**	**PP**
1	Assertive	M	E	**How?**	M	E	Which?	M	E
2	Expressive, Commitment	M	O	How?	**M**	O	Why?	**I**	O
3	Commitment, assertive	M	C	How?	**M**	C	¿What?	**I**	C
4	Expressive, Commitment, Assertive	M	O	**How?**	M	O	Which?	M	O
5	Assertive	M	E	Which?	**M**	E	¿What?	**I**	E
6	Assertive	M	C	How?	**M**	C	¿What?	**I**	C
7	Assertive	I	E	What?	I	E	¿What?	I	E
8	Assertive	I	E	Why?	**I**	E	Which?	**M**	E

In the deliberative process module, the questions provided by the question selector module are provided, as well as the preferential thresholds influenced by the personality of the coach. In this module, the preference-based solution strategy corresponds to the ELECTRE III method.

This method obtains the solution alternatives, that is, the powerful questions closest to the preferences of the virtual tutor. A powerful question can use maieutics or Socratic irony according to the perception of the coachee's personality type and the existing relationship with his illocutionary act.

To structure powerful questions at the level of phrases, a previous study of each of the phrases must be carried out to make a list of all the rules that can be formed as a result of what the coachee said, that is, all the possible combinations that can be built through syntactic trees by structuring questions with the four proposed interrogative pronouns (What? How? Which? and Why?), which define the type of question-based on dialectics. Next, algorithm 1 shows the general procedure of the architecture of a dialectic-based virtual coach.

Algorithm1. General procedure of the virtual coach based on dialectics.

```
function COACH (phrase)
  greeting_farewell=es_ Greeting _ Farewell (Perception(phrase))
  if greeting_farewell==null
     questions_characterized_dialectic=Speech_Acts(phrase)
     personality=Get_Personality(questions_characterized_dialectic)
     thresholds=select_preference thresholds _ELECTRE_III(personality))
     return select_phrase(questions_characterized_dialectic, thresholds)
  else:
     return greeting_farewell
end function
```

5 Evaluation Results of the Selected Sentences in a Socratic Dialogue

The result of the agent's execution will generate powerful questions based on Socratic dialectics. To contrast the effectiveness of the simulation process, a comparison is made between the result of the simulation and that of the interaction of a real coach in a Socratic dialogue focused on the problems of a teacher with a student due to the lack of interest in his class. Table 1 shows the first two sentences of the interaction and its associated illocutionary act (TIA).

Next, a brief description of the results obtained through the simulation of the coaching processes is presented. The experimentation seeks to compare whether the responses of a virtual coach correspond to the responses of a real context under the same decision profile in the argumentation phase of the proposed Socratic dialogue.

Table 2 shows the comparison between the response of the virtual coach using the ELECTRE III method and the response of each of the phrases according to each personality profile of the coach. A total of 8 interactions belonging to the argumentation phase are observed. The evaluation validates if the type of dialectical question (TDQ), that is, maieutics and irony, corresponds to the one selected by the real coach.

The characteristics considered in this table are the type of illocutionary act (TIA), TDQ, personality profile (PP), and interrogative pronouns (PN). The PP is defined in terms of the TIA, that is, if the TIA is Assertive, then the coach's PP is assumed to be of type Strict (E). If the TIA is Commitment, the PP will be Collaborative (C). If the TIA is Directive, the PP will be Inquiring (Id) and if the TIA is Expressive or Declarative, the PP will be Optimistic (O).

Based on the type of TIA, the TDQ can be defined, which are Maieutics (M) and Irony (I). These are defined by the PN, those that correspond to M are How? and Which? and for I What? and Why? The final definition for each TDQ is given randomly.

The first aspect to be evaluated is the results in the choice of the TDQ, in which a total of 100% correspondence is obtained between the TDQ of the phrases selected by the virtual coach concerning that expressed by the coachee, which corresponds to the dialectic methodology.

On the other hand, the results of the evaluation of the TDQ selected between the virtual coach and the real coach differ by 50%, the difference percentage may be because the real person does not have adequate preparation to use dialectics in an appropriate way professional.

6 Conclusions and Future Work

In this article, a conversational system based on Socratic dialectics was presented, which uses the overclassification method called ELECTRE III for the selection of the phrase closest to the coach's preferences within the coaching process.

For the selection of the type of question-based on dialectics, preference thresholds were used according to the type of personality profile of the coachee perceived in the characterization phase.

This is to emulate the characteristics exposed in the literature related to the Socratic dialectic, where it is observed that the coach performs the function of a mirror, thus achieving that the coachee can reflect on their cognitive scenarios through their responses.

In future works, the integration of more grammatical rules exposed in the Spanish language is expected so that the agent can have greater reasoning not only limited to the rules within the possible phrases that can be developed from a sentence.

Another aspect that can be improved is the assignation of the interrogative pronoun associated with the type of powerful question since currently the selection of the pronoun is randomly assigned among the possible values of pronouns related to perlocutionary verbs.

It is expected that the phrases selected with this strategy contribute to inductive advice based on Socratic dialectics so that the coachee can find and improve the integration of their skills in their work environment, finding solutions to their problems through reflection on their behavior.

Acknowledgments. The authors thanks CONACyT for the support obtained through the Scholarship for Postgraduate Studies with CVU 464864, and the chair project number 3058. They also thank the support of the TecNM/Instituto Tecnológico de Ciudad Madero for the support provided through their installations.

References

1. Prieto, A., Díaz, D., Santiago, R.: Metodologías inductivas: El desafío de enseñar mediante el cuestionamiento y los retos. Digital-Text (2014)
2. Rodríguez-Jiménez, A., Pérez-Jacinto, A. O.: Métodos científicos de indagación y de construcción del conocimiento. Revista Escuela de Administración de Negocios, vol. 82, pp. 175–195 (2017) doi: 10.21158/01208160.n82.2017.1647
3. Manga-Beltrán, B., Charris-Charris, L.: La pregunta socrática como estrategia pedagógica para favorecer el desarrollo del pensamiento crítico (2018)
4. Marcano-Lárez, B. E.: Juegos serios y entrenamiento en la sociedad digital. Teoría de la Educación: Educación y Cultura en la Sociedad de la Información, vol. 9, no. 3, pp. 93–107 (2008)
5. Luzardo, G., Hernández, J.: Inteligencia artificial en ambientes virtuales: Humanos virtuales autónomos (HVA) como agentes virtuales inteligentes (3diva) (2010) blog.espol.edu.ec/gluzardo/files/2009/03/trabajo-final.pdf
6. Pérez-Pinillos, D., Fernández, S., Borrajo, D.: Modeling motivations, personality traits and emotional states in deliberative agents based on automated planning. In: Agents and Artificial Intelligence: Third International Conference, pp. 146–160 (2013) doi: 10.1007/978-3-642-29966-7_10
7. Gómez-Róspide, C., Puente, C.: Agente virtual inteligente de ayuda al aprendizaje. Doctoral dissertation, Tesis de Pre-Grado: Universidad Pontificia Comillas (2012)
8. Heylen, D., Nijhot, A., op den Akker, R., Vissers, M.: Socially intelligent tutor agents. Lecture notes in computer science, pp. 341–347 (2003)
9. Morales-Rodriguez, M. L.: Modèle d'interaction sociale pour des agents conversationnels animés: Application à la rééducation de patients cérébro-lésés. Doctoral dissertation, Université Paul Sabatier-Toulouse III (2007)

10. Florencia-Juárez, R.: Agente conversacional corpóreo que utiliza AIML para integrar procesos de personalidad. Doctoral dissertation, Tesis de Maestría (2010)
11. Delgado-Hernández, X. S., Morales-Rodríguez, M. L., Rangel-Valdez, N., Reyes, L. C., Santillán, C. G., Barbosa, J. J. G.: Analysis of speech acts for the design of a corpus of phrases used in an intelligent learning environment. Research in Computing Science, vol. 148, no. 5, pp. 33–41 (2019)
12. Castro-Rivera, J. A: Modelado de la personalidad en modelos preferenciales multicriterio a través de agentes virtuales inteligentes (2018)
13. Iglesias-Fernández, C. Á.: Definición de una metodología para el desarrollo de sistemas multiagente. Enfoques, vol. 35, pp. 57–190 (1998)
14. Molica de Mendonça, F., Durange-de Carvalho Infante, C. E., Bastos-do Valle, R. D. A.: Aplicación del método Electre III en la clasificación de clústeres de artesanías. INGE CUC, vol. 7, no. 1, pp. 97–112 (2011)
15. Polaino-Lorente, A., Cabanyes Truffino, J., Pozo Armentia, A. D.: Fundamentos de psicología de la personalidad. Rialp (2003)
16. Ahrndt, S., Fähndrich, J., Albayrak, S.: Modelling of personality in agents: From psychology to implementation. In: Proceedings in Fourth International Workshop on Human-Agent Interaction Design and Models, pp. 1–16 (2015)

ISSN 1870-4069

Da Capo: A Proposal for a BDI-Based Tutorial for Teaching Music

Alberto Nieto-Rocha, Wulfrano Arturo Luna-Ramírez

Universidad Autónoma Metropolitana,
Cuajimalpa,
Mexico

2163071736@cua.uam.mx, wluna@correo.cua.uam.mx

Abstract. The use of digital educational tools allows more people to have access to educational content, they have a wide range of resources and help to meet the demand for education. Learning music can provide various health benefits, such as improving mood. One of the ways to create a tool to help teach music is thanks to AI, in particular with tutoring agents. This paper presents a proposal of a multi-agent system making use of cognitive agents under the well known Belief-Desire-Intention paradigm based on a successful integration of web development tools such as *Flask* to create a web-based system for teaching basic music concepts.

Keywords: E-Learning, AgentSpeak, music tutor agents, BDI agents, multiagent system design, python, flask.

1 Introduction

Technology tools have solved some issues generated by social distance, noteworthy is teaching and e-learning. Most of the schools in the world have opted for virtual activities due to the increase in the use of internet and digital platforms that allow contents to be taught, homework to be assigned, homework to be reviewed, among other educational tasks.

Online activities proved to be very attractive to the population was those related with arts. Is in this context where we identify the need to create a music tutor agent for both to exploit the digital resources and connectivity, and to overcome the limited capacity of live human interaction.

Additionally, learning music can be beneficial for health, in a study done by [13], basic music theory and piano lesson were given for four months to a group of older adults who had no knowledge of the subject. The research determined that learning music and the cognitive processes involved can help combat depression and promote a better mood.

This paper presents one way to make use of AI and in particular tutor agents to support music students who would like to reinforce their basic music knowledge and skills. Using a web-based system, this present concepts and practical exercises. This tool is an intelligent music tutor called *Da Capo*.

2 Preliminaries and Related Work

Musical formal education has been oriented to traditional teaching, where teachers and students meet to realize the act of teaching/learning. Technology has let innovation on this type of act. For example, some music school has given formal and informal course of music during summer. Activities pursued in this course imply to make video and audio recording while students do some exercises. With this file, the teacher can do the evaluation and provide feedback to their students.

One of the problems of this kind of course, as mentioned in [2], arise with the high demand and the difficult to serve so many students, teachers can be overwhelmed by the workload. This can result in not providing adequate feedback to each student. One solution to this difficulty is the use of software that allows students to make use of tools that are capable of acting as tutors [1]. These tools have the particularity of being able to provide personalized and instant attention to each student, which allows students to be able to advance at their pace and helps to prevent gaps in the understanding of any subject.

Considering this, we can contemplate for a first version of the Da Capo system, the possibility of showing different types of content such as texts, audios, images, and videos. With these, the student can be presented with various exercises such as identifying concepts, improving his musical ear and learning about the history of music. It should be clarified that this article does not contemplate the design of exercises to teach music theory, since to do so, at least one person with expertise in music pedagogy is needed.

In this article, we present a system which, with adequate music content, can be used by people who wish to reinforce their musical knowledge and skills. In this section, we discuss the technologies that take on the role of the teacher and the methods used to develop them [1, 4, 11, 7].

2.1 Agents

Agents are informatics system that are located in an environment, they are capable of perceiving changes and interacting autonomously and persistently according to these changes. Their existence summarizes in the following characteristics:

Reactivity can be considered as the perpetual interaction with its environment, which forces it to respond to the changes that arise. To be able to respond adequately to changes, it needs to be endowed with initiative; this characteristic is known as *proactivity*. Since their behavior cannot be a product of chance, agents must be *autonomous*, i.e., they must base their behavior on the measure of their experiences. This implies that the agent does not rely solely on the knowledge with which it was created; the more it bases its behavior on learning and on the knowledge incorporated, the more autonomous it will be.

Social ability in agents is described as their ability to interact with other agents and even humans, they must use certain languages and protocols to establish accurate communication, since with this, they can achieve certain goals that can only be achieved with their peers, sets of these are known as multi-agent systems.

Persistence is expressed as the constant functioning of the system. There are different perspectives for designing agents. In the following, we will discuss two approaches in particular, tutor agents and BDI agents [9].

2.2 Tutor Agents

The intelligent tutor agents guide the student in the learning process; providing constant feedback, through evaluations the intelligent tutor agent chooses the topics in which the student presents some learning deficit. Some of the research that has been carried out on the subject indicates that intelligent tutor agents have four components: the knowledge base, the student model, the tutor model and the user interface [4, 5, 1]:

– **Knowledge Base**. Stores the materials that are taught in the agent, these can be: sounds, images, texts, among others. It also contains the solutions and explanations of the topics.
– **Student Model**. Manages the information about students, such as the access credentials, the activities that the student performs, the errors and successes about the topics seen.
– **Tutor Model**. It is responsible for deciding, based on the information in the student's model, the topics that will be presented to the student. These topics are the ones found in the knowledge base.
– **User Interface**. It is the connection between the tutor system and the student. It shows the exercises and materials pleasantly. It is necessary that this component is refined to provide adequate feedback to the student, as this will help the student to understand the failures he has. Ease of use should also be considered, since the user should concentrate on the topics presented, not on how to use the platform.

To create an intelligent music tutor agent, the tutor agent components should be taken as a base, to which other components that the developer considers necessary can be added. The knowledge base should be endowed with contents related to the musical field. For example, in the article related to WMITS, [11] discusses that, in addition to the tutor agent components, they use third-party packages that interface directly with the tutor model.

2.3 BDI Agents

The BDI model, in the words of García et al. [6], is a promising architecture. This is thanks to their capacity for practical reasoning, which in other words is the process of deciding what action to take to achieve goals. The BDI architecture is based on beliefs, desires, and intentions[14], and integrates Dennett's and Bratman's positions. They are: intentional approach and Plans and interactions and practical reasoning, respectively. This architecture has been extensively tested in a diversity of systems ranging from manufacturing to space navigation.

Beliefs are the information that the agent knows about its environment; they are actualized by perceptions and intentions. Desires are the agent's goals. Intentions are stacks (like data structures) of executing plans, the latter are stored in a plan library.

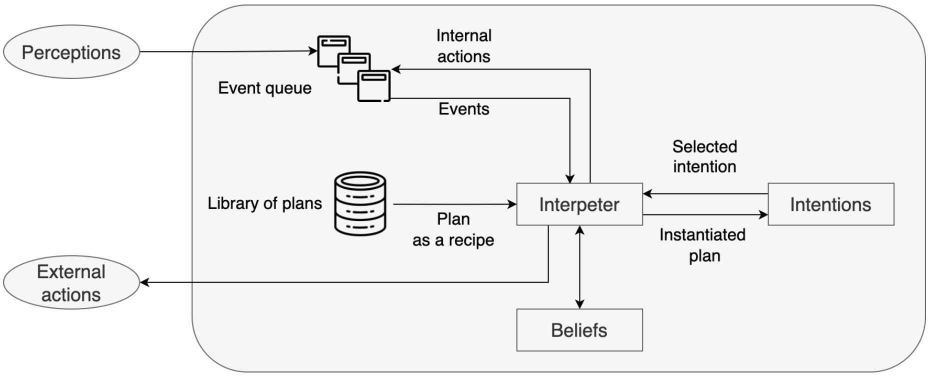


Fig. 1. Architecture of a BDI agent. Adapted from [8].

Plans are skills or procedural knowledge that the BDI agent has at its disposal to achieve certain states. Within plans are actions, which can update beliefs or modify the environment [8, 10]. In Figure 1, the above is shown.

2.4 Related Work

Some related work with agents and education has in common that used agents of type cognitive or reactive. In this section, we present some of the systems that make use of agents to teach (*Slang*, *Maestoso*, *WMITS*) or whose teaching area is music (*EarMaster*).

Maestoso is a tool focused on people without knowledge about music, it's an intelligent tutoring system that using an optical pencil that allows to make different exercises of writing notes. These are automatically graded by the system, which provides feedback and hints on what to do in the exercises [3].

Slang is an application focused on teaching English, created at MIT. It makes use of the tutor agent components described in section 2.2, to present an education with constant, unique, and customized feedback. The system allows the learner to perform activities such as recording sentences, identifying objects by name, spelling words, matching words with their meaning, etc. [12]

WMITS(Web-based Music Intelligent Tutoring Systems) is a web-based system for teaching music. This system teaches different music concepts such as intervals, chords, and harmony. According to the author of the system, to display the content, intelligent tutor agents are implemented to define the content to be displayed to the user. Since the research was conducted in 2007, the system was somewhat limited by the Web development technologies of the time. The development of the system seems not to have continued, since it is not possible to find it [11].

EarMaster is a paid music theory trainer with more than 2500 interactive lessons, covering the fundamental aspects of ear training, sight-singing and rhythmic training. To use the tool, you make use of the device's peripherals such as the keyboard, headphones, microphone, touch screen and touch screen to carry out the different activities it proposes.

The activities can include saying a note out loud and the system detecting if it is the correct note, identifying chords, identifying rhythms, etc. It should be noted that the system shows constant feedback, which lets you know exactly what you are doing. It is not specified if the system uses some kind of intelligent tutor to select the content that the user will see.

In a test performed I could detect that when doing several exercises wrong, the system put them again, also when doing several exercises correctly the system finished the lesson. This could indicate that there is a certain component that may or may not be an intelligent tutor [5].

Those systems mention before are tools designed to teach, and they use agents to make that, but the principal difference, in respect to Da Capo, is the use of BDI agents for improving the learning process. With BDI agents, Da Capo can display different types of content depending on the user's needs, i.e., it is not just a software to show exactly the same lesson to all students. Other difference it's that Da Capo it's a free software, and use more modern technologies like AgentSpeak or Flask.

3 Da Capo System Overview

This section describes the architecture of the system *Da Capo* and the three agents that conform it, and also explains how the interaction between the components is. For the development of the system, different technologies were used: Python, *Flask*, *MySQL*, *Python-Agentspeak* (alternative to *Jason*), *WSGI*, *HTML*, *CSS*, *JavaScript*.

We use *Prometheus* methodology [10] to design and develop the system *Da Capo* as it stands out among other software development methodologies for supporting the creation of intelligent and BDI agents. Figure 2, shows a diagram elaborated as part of the *Prometheus* methodology.

It shows the three agents that compose the system (`Teacher`, `Librarian`, and `School Control`) with their respective actions (rectangular trapezoid) and perceptions (star), according to [10], the actions are ways in which the agent can operate in the environment (in *Da Capo*, the environment is *Flask*) and the perceptions are the relevant information coming from web requests. A brief description of Da Capo agents is presented as follows:

- **Teacher**: It is responsible for interacting with students, providing content according to their needs. It constantly applies evaluations on the academic achievement of students. He also communicates with $SchoolControl$ to send a record of the student. It establishes a connection with $Librarian$ to request the contents to be shown to the student.
- **Librarian**: It consults the database, but acts under the orders of the teacher agent. When changes in the content needs to be done, this agent will be modified in the database.
- **School Control**: It has the academic information about the students, it will identify and indicate to the teacher agent which student arrive and will provide relevant data to this for the request of a lesson.

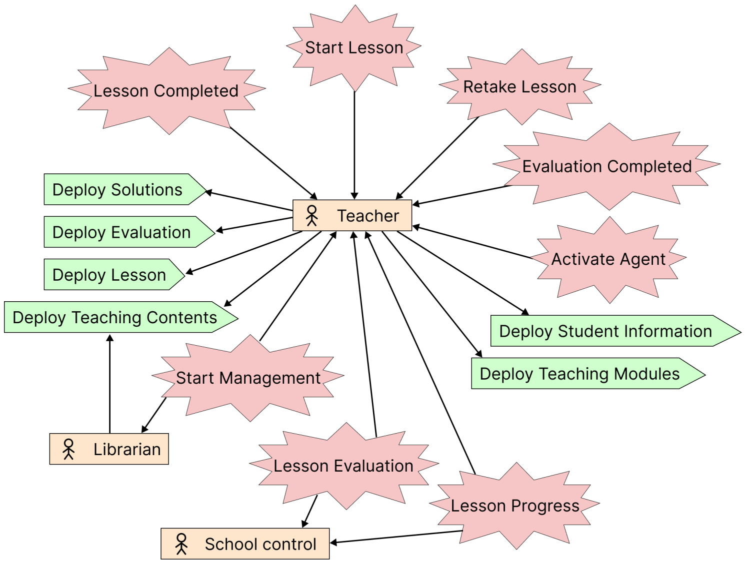


Fig. 2. System Overview Diagram from Prometheus methodology [10].

Fig. 3. Main page of *Da Capo* displayed into a Flask environment.

Flask, a framework developed in Python, was used to build the website. It uses *Jinja2* as a web template engine and the *WSGI* interface for the web server to call a Python program.

Three Python codes were written, app.py for *Flask* related, $agent.py$ for *Agentspeak* related, and $database.py$ for database queries. An *ASL* file was also written, which contains the agent procedures. In addition to the three mentioned files, there is a fourth one called variables.py, this one contains global variable declarations, which are used to communicate *Flask* with *Agentspeak*. In Figure 3 the main page of Da Capo system is displayed.

Fig. 4. Student's profile, with their teaching modules.

Figure 4, shows a user's profile, in which you can see the educational modules that the user can access. When entering these, the lesson (in Spanish) is presented.

Figure 5, shows how the different parts of the system interact to display a complete lesson. In general terms, the user must choose the module to be studied, the agent queries it and displays the corresponding lesson. Once the user has finished the lesson they should request the evaluation, again the agent queries and displays the request. At the end of the evaluation, the agent receives the user's answers, grades them and updates the user's profile.

Figure 5, shows the interactions between all the components of the system. But in terms of the BDI agent, the process to consult a lesson is the following. Using as a reference, Figure 1. The agent perceives an event (generate by the user in the web page) this event enters to the event queue, the interpreter deliberates the best option (plan) to achieve the activity (desire), this process is called intention. Once the agent chose what to do, the plan begun to execute. In this case, the plan is called "lessonConsultation", this is conformed for two steps: consult lesson, and show lesson.

The first step executed an internal action in *AgentSpeak* query *"x"* lesson. As mentioned in section 2.2, one of the components of Tutor Agent is *the knowledge base*, this stores the contents that the agent will teach, in our case materials related to the field of music.

This knowledge base are the beliefs of our agent. Once the results are ready, the agent makes the step two *show lesson*, in this case it's an external action. The agent sends the lesson to *Flask* and this last one is responsible for displayed the lesson to the user. The latter is the logic behind the BDI agent, which can be applied to different activities such as evaluating the lesson or showing the correct answer of the evaluation, as long as there are plans for them.

4 Conclusions and Future Work

Technology has revolutionized the way that people learn. This article presents the design of a multi-agent system that takes the activities that a teacher performs and makes a system do them.

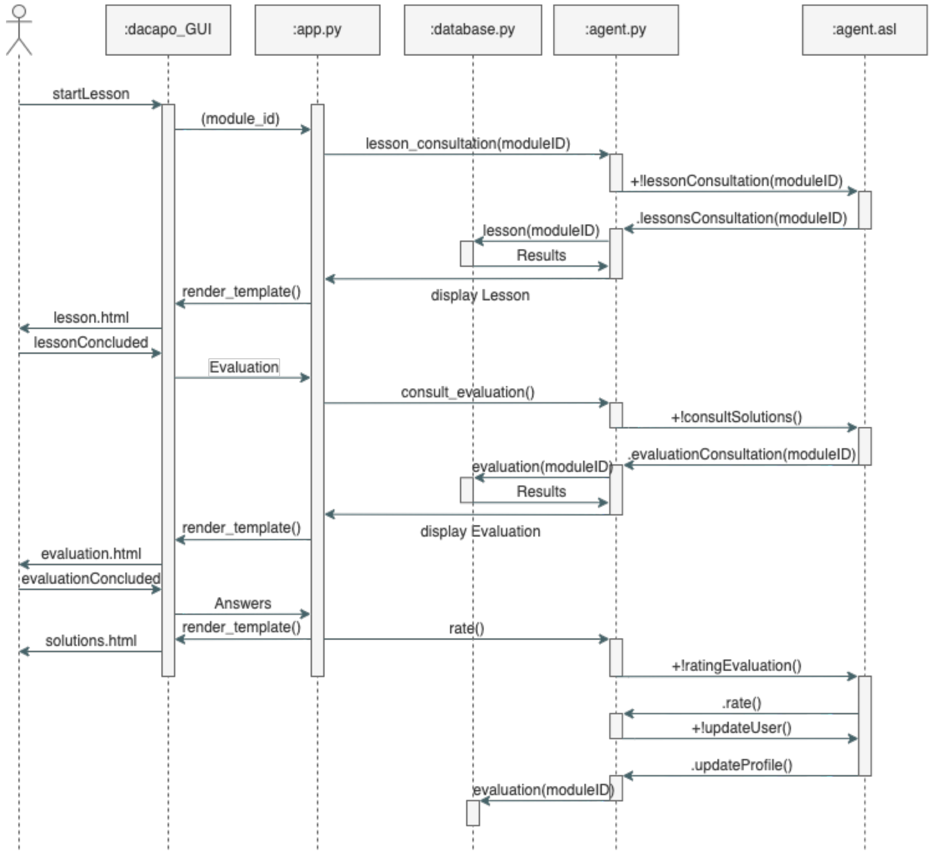


Fig. 5. Interaction Diagram for *Da Capo* Lesson Consultation.

This is not a replacement for a music teacher, it is a tool to help reinforce knowledge. Creating this tool had its difficulties, since there was no precedent that integrated *Flask* and *AgentSpeak*, although both are created in the same language, plus what it means to integrate *MySQL*, *HTML*, *CSS*, *JavaScript*, *WSGI* and the design of the agent.

For future work it is necessary to contemplate other tools that allow to communicate to the agents since *AgentSpeak* does not solve adequately this part, for this reason, the prototype of the system only counts with one agent that pursue all the functionalities.

It is also necessary to add more predicates to the agent to exploit the potential of the BDI architecture. It would also be interesting to enhance agent capabilities for retrieving musical information from dedicated storing sites to create the lessons.

Finally, it is essential to design and create musical content with assistance of professionals in the area, this will allow us to carry out tests with users with which we can pedagogically evaluate Da Capo and what is related to human-computer interaction.

Aknowledgments. Supported by UAM-Cuajimalpa, División de Ciencias de la Comunicación y Diseño, Licenciatura en Tecnologías y Sistemas de Información.

References

1. Apaydinli, K.: Intelligent tutoring systems in music education. In: 2nd International Conference on Future of Teaching and Education (2020)
2. Atiaja-Atiaja, L. N., Guerrero-Proenza, R. S.: MOOCS: Problems and challenges in higher education. In: International Conference on Advances in Education, Teaching & Technology, pp. 82–88 (2016)
3. Barreto, L., Taele, P., Hammond, T.: A stylus-driven intelligent tutoring system for music education instruction. Revolutionizing Education with Digital Ink: The Impact of Pen and Touch Technology on Education, pp. 141–161 (2016) doi: 10.1007/978-3-319-31193-7_10
4. Da, S.: Research on the design of online intelligent tutoring agents. In: 2010 International Conference on Electrical and Control Engineering, pp. 4755–4758 (2010) doi: 10.1109/iCECE.2010.1151
5. Earmaster-music theory and ear training on PC, Mac, iPad and iPhone. https://www.earmaster.com/es/company/press-material.html
6. García, D., Simari, G., García, A.: Planificacion en agentes BDI. In: VI Workshop de Investigadores en Ciencias de la Computación (2004)
7. Latham, A. M., Crockett, K. A., McLean, D. A., Edmonds, B., O'Shea, K.: Oscar: An intelligent conversational agent tutor to estimate learning styles. In: International Conference on Fuzzy Systems. pp. 1–8 (2010) doi: 10.1109/FUZZY.2010.5584064
8. Luna Ramírez, W. A.: Agentes cognitivos, intencionales y BDI (2021)
9. Luna Ramírez, W. A.: Botzoologia: Un bestiario de agentes artificiales (2021) https://www.uv.mx/cienciauv/blog/botzoologiaunbestiariodeagentesartificiales/
10. Padgham, L., Winikoff, M.: Developing intelligent agent systems: A practical guide (2004) doi: 10.1002/0470861223
11. Phon-Amnuaisuk, S., Siong, C.: Web-bases music intelligent tutoring systems. Interactive Multimedia Music Technologies, pp. 231–248 (2007) doi: 10.4018/978-1-59904-150-6.ch011
12. Rojas, N.: Queremos contarte en qué consiste el adaptive e-learning de Slang y cuáles son sus ventajas. https://es.blog.slangapp.com/queremos-contarte-en-qu\%C3\%A9-consiste-el-adaptive-e-learning-de-slang-y-cu\%C3\%A1les-son-sus-ventajas-f62138d69690
13. Seinfeld, S., Figueroa, H., Ortiz-Gil, J., Sanchez-Vives, M. V.: Effects of music learning and piano practice on cognitive function, mood and quality of life in older adults. Frontiers in Psychology, vol. 4 (2013) doi: 10.3389/fpsyg.2013.00810
14. Wooldridge, M. J.: Reasoning about rational agents. MIT Press (2000)

ISSN 1870-4069

Machine Learning for Intent Classification in an Educational Chatbot

Carlos Natanael Lecona-Valdespino[1], José Moisés Aguilar-Ibarra[2], Guillermo Santamaría-Bonfil[3]

[1] Universidad Panamericana, Mexico

[2] Instituto Politécnico Nacional, Escuela Superior de Cómputo, Mexico

[3] Consejo Nacional de Ciencia y Tecnología, Mexico

0245614@up.edu.mx, jaguilari1800@alumno.ipn.mx, gsantamaria@conacyt.mx

Abstract. Nowadays, intelligent learning environments use advanced human-computer interfaces such as voice-commanded chatbots. A fundamental aspect of such chatbots is that the dialogue system is able to correctly identify the intention of the users. In this work, an user intention classifier for an educational chatbot is built using different NLP techniques as well as different ML algorithms. A corpus of phrases belonging to intentions from an educational chatbot is used to train and compare support vector machines, decision trees, random forest, and extreme gradient boosting. A key step in this analysis is the inclusion of the parts of speech filtering to reduce the number of features required for intent classification, hence, reducing the computational burden. Results showed that the random forest algorithm using the parts of speech filtering identifies user intentions accurately and with less computational effort, in comparison to the rest of the algorithms combinations.

Keywords: Educational conversational agent, machine learning, intent classification, natural language processing, recommender system.

1 Introduction

Chatbots are considered to be one of the key technologies powered by artificial intelligence to improve the communication and interaction of users with computers [4]. This technology reduces the workload required from users by providing them with friendly and human-like interfaces. For this, chatbots employ Natural Language Processing (NLP) techniques and Artificial Intelligence (AI) tools to process and understand speech commands or written text. These systems range from simple frequent ask questions answering machines to more sophisticated systems that can sustain a conversation with a human about a specific topic.

Chatbots are used broadly in several industries such as social media platforms, customers services, medicine, and even as educational assistants [3, 8]. Regarding the application of chatbots to educational settings, studies revealed that chatbots can enhance the learning settings by allowing personalized real-time interactions, improving the communication among peers, and increasing the efficiency of the learning procedure [4].

One of the main abilities of any chatbot is the identification of users intentions. An intent refers to the action or goal a user looks forward to achieve. Hence, intent classification tries to identify what the user wants by applying Natural Language Processing (NLP) techniques along with Machine Learning (ML) algorithms [9]. By identifying the user intent from a predefined intention list, the chatbot dialogue manager can redirect the conversation to specific dialogue paths [2].

Several works have been developed for intent classification using a multitudes of NLP and ML techniques. In the case of the former, tokenization, stop-words removal, lemmatization, parts of speech (POS) tagging and filtering have been employed [9, 2, 8]. In the case of the latter, naïve bayes, logistic regression, Decision Trees (DT), artificial neural networks (multilayer perceptron, convolutional neural networks, long-short term memory, BERT), Support Vector Machines (SVM), naïve bayes using bag of words and word embeddings, majority voting, FastText, stand among the most recently employed [3, 9, 6, 2, 8].

While complex neural networks such as BERT have shown outstanding performance in NLP problems where there are large amounts of available data, intent classification datasets are, in average, around a few thousand samples [5]. Under such circumstances traditional or ensemble algorithms perform better than complex deep learning structures. A key NLP procedure is identifying the POS. Such procedure consists in tagging words in accordance to their grammatical function. Although, POS tags are different for each language, within the same language, words tagged with the same POS can be considered to play a similar grammatical function in the sentences. Such POS property can be employed to filter the less important features for each sentence. In fact, in [1] considers this preprocessing step key for improving the classification accuracy.

Therefore, in this work several NLP procedures are used to preprocess user utterances from a corpus of phrases of an educational chatbot. By cleaning and filtering important features using POS tagging, support vector machines, decision trees, random forest, and extreme gradient boosting are trained for intent classification. The results showed that the random forest algorithm using POS tagging filtering identifies user intentions with high accuracy and with less computational effort, by reducing the number of features employed for classification.

The rest of the paper is organized as follows: section 2 presents the materials and methods; section 3 presents the experimental setup and the results of the experimentation; section 4 presents conclusions and discusses future work.

2 Materials and Methods

In Fig. 1 the methodology to build a model for intent classification is presented.

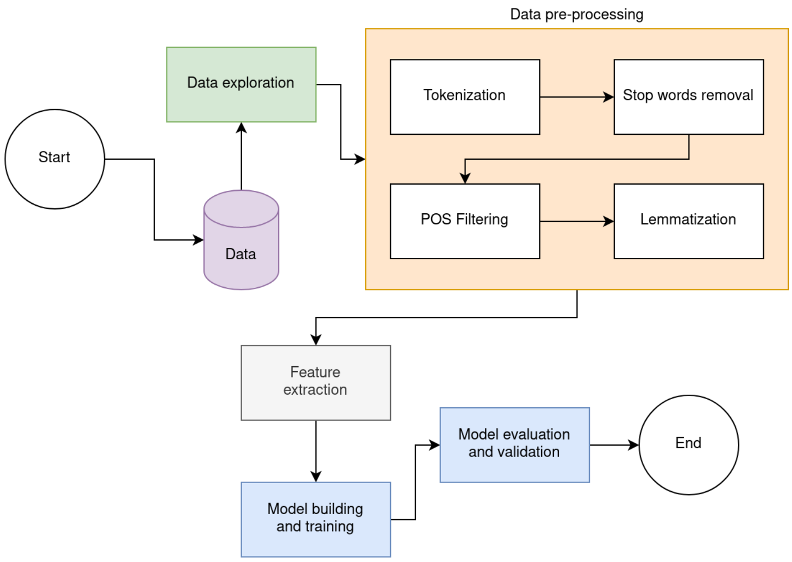


Fig. 1. Intent classification pipeline.

First, data corresponding to the intents of an educational chatbot used for training in energy topics [7] is collected and classified by intent.

Next, an exploratory data analysis is carried out to visualize i) words and ii) the POS that compose each intent. The dataset is then pre-processed using four NLP processes namely tokenization, stop words removal, POS (Part-of-Speech) filtering, and lemmatization.

From the pre-processed phrases features are extracted by building n-grams and transforming the text into TF-IDF vectors. Once the vectors are obtained, the dataset is separated into training and test sets. Using these, four ML models are trained and evaluated using three metrics: Balanced Accuracy (BA), Matthews Correlation Coefficient (MCC), and Cohen Kappa Score (KCS).

2.1 Data

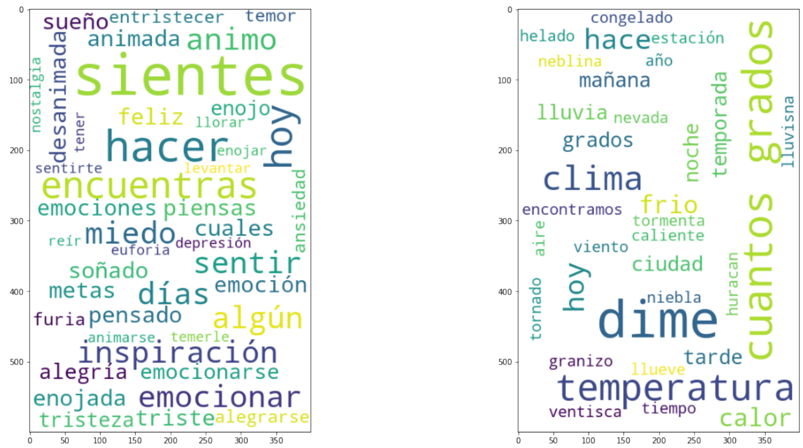
Initially, a corpus of 1850 phrases labeled their respective intention was developed for an educational chatbot used for virtual reality training environments [7]. The phrases are in Spanish and correspond to user requests to the chatbot to achieve one of the available actions.

The list of the eleven (11) available intents in the chatbot is shown in Table 1. In this table, first and second columns corresponds to the user intention and the name assigned in the dialogue manager of the chatbot, respectively. The third and four column presents the frequency of phrases and their corresponding proportion in the corpus.

Observe that, the corpus is highly imbalanced since the search intention corresponds to more than 60% of it. This is due to the chatbot main duty is to be capable of answering any knowledge question requested by the user as accurately as possible.

Table 1. Description of chatbot intentions.

User intention	Name	Frequency	Proportion
Lookup for a concept	Search	1200	0.65%
Teleportation within the virtual world	Navigation_microred_caseta	200	0.11%
Ask the chatbot to tell an interesting fact	Utility_check	50	0.03%
Ask for the date	Date_check	50	0.03%
Ask the chatbot to tell a joke	Joke_check	50	0.03%
Ask for the emotional state of the chatbot	Feel_check	50	0.03%
Ask for the time	Time_check	50	0.03%
Ask about the weather	Weather_check	50	0.03%
Thanking the chatbot	Gratitude_check	25	0.013%
End the conversation	End_conversation	25	0.013%



(a) (b)

Fig. 2. Word clouds for two different intents. On (a) the word cloud for Feel_check intention is shown, whereas on (b) the Weather_check is presented.

2.2 Data Exploration

To qualitatively understand differences between the available intentions a word cloud for each of these was generated. These word clouds allow to visualize the words that are most frequently used depending on each intent and the type of words that should be included for training.

Figs. 2a and 2b show the word clouds corresponding to feel_check and weather_check, respectively. Observe that, the feel_check intention employs words such as feel (sentir), emotions (emociones), joy (alegría), inspiration (inspiración), and so on. In contrast, the weather_check intention uses words such as temperature (temperatura), degrees (grados), weather (clima), and so on.

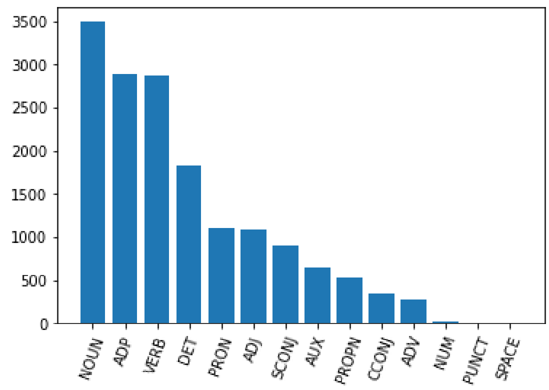


Fig. 3. POS Exploration.

2.3 Pre-Processing

Pre-processing consists in tokenization, stop words removal, POS (Part-of-Speech) filtering, and lemmatization. At the beginning, tokenization is performed to divide the sentences into words and apply the next pre-processing steps. Next, all stop-words are removed using a Spanish stop-words dictionary.

Afterwards, the POS filtering is applied to select only relevant words based on the data exploration and finally all words are normalized by lemmatization.

POS Filtering

Fig. 3 shows how many words belong to each part of speech present in the dataset. Based on this, it is possible to directly discard parts of speech that are insignificant and are frequent in the dataset, such as adpositions, which in Spanish are generally only prepositions, and determiners, which include articles, demonstratives, possessives, and quantifiers.

Different combinations of parts of speech may improve or worse the performance of the ML models. Therefore, in this work the POS considered are Adjective (ADJ), Adverb (ADV), Auxiliary (AUX), Noun (NOUN), Proper Noun (PN), Verb (VERB), Pronoun (PRON), and Subordinating conjunction (Sc). The reason for this is that by testing various combinations of POS, it was determined that the mentioned POS elements were the most relevant for the dataset used.

The combinations selected were i) {NOUN, VERB, PN}, ii) {NOUN, VERB, ADJ, PRON} , iii) {NOUN, VERB, ADJ, AUX, PN}, iv) {NOUN, VERB, ADJ, ADV, Sc}, and v) {NOUN, VERB, ADJ, PN}. In addition, the performance of the ML algorithms are compared with and without the proposed POS combinations.

2.4 Feature Extraction

To extract the features on which the models are trained on, uni-grams and bi-grams are built. Then, these are transformed into TF-IDF vectors.

A different feature space is generated for all combinations of POS filtering and No-POS filtering. In general, by including a POS filtering, the total number of features is decreased (ranges from 97 to 111) in comparison with No-POS filtering (114 features).

2.5 Machine Learning Classifiers

Once data is pre-processed, the dataset is splitted into training and testing subsets. The proportions used for these subsets were 80% and 20%, respectively. It is worth to note that, due to the classes imbalance, an stratification approach must be consider in the splitting of the data. In this work, sample proportions are maintained for both, the training and testing subsets.

On the other hand, classifiers can be grouped into single (DT or SVM) or ensemble (RF or XGBoost) algorithms. The main difference between these is that, the first looks forward to obtain a robust model with a good generalization, while the second achieves a good generalization by combining several classifiers by specific strategies such as bagging or boosting.

Therefore, in this work DT, SVM, RF, and XGBoost are employed for intent classification. From these, DT, RF, and XGBoost are tree-based classifiers. A DT partition the feature space into regions defined by thresholding each feature values. This algorithm is the most interpretable and the most prone to overfitting. RF and XGBoost combine several DTs to classify, improving the variance and the accuracy of the model using clever strategies. In the case of RF, DTs are trained in parallel, each receiving a random subset of data sampled with replacement.

In the case of XGBoost, a series of DT are stacked one after the other, each focusing upon the samples misclassified by the previous learner. Finally, SVM is a robust single ML model that is built upon the structural error minimization principle. SVM can handle noisy non-linear relationships by penalizing errors and mapping to larger feature spaces through the kernel functions.

Furthermore, each ML model was tuned by a grid search cross validation procedure for obtaining the best hyperparameter values. In this work, a 3-fold cross validation procedure was carried out. The parameters tested where: i) for SVM, penalization C (0.1-10), γ ($1e^{-3}$-1), and different kernels (liner, radial, and polynomial); ii) for DT and RF, parameters max_features (log_2 and the square root), max_depth (1-100), min_samples_split (2, 5, 10), min_samples_leaf (1, 2, 4), and in particular for RF, class_weight (stratified or sub-sampling) and number of estimators (10-200); iii) for XGBoost, learning rate (0-1), max_depth (10-100), and number of estimators (10-2000).

2.6 Model Evaluation

To compare the performance of the ML algorithms for classifying intentions from an educational chatbot, multi-classification measures are employed. In this work the Balanced Accuracy (BA), Cohen's Kappa Score (KCS), and Matthews Correlation Coefficient (MCC) are employed.

These measures are useful for assessing the overall performance of a classifier in a multi-class problem. Further, MCC and KCS provide robust measures in the presence of classes imbalance, such as the corpus used in this work.

Table 2. Performance results for intent classification of the educational chatbot corpus.

POS Filtering	SVM			DT			RF			XGBoost		
	MCC	BA	CKS	MCC	BA	CKS	MCC	BA	CKS	MCC	BA	CKS
NOUN, VERB, PN	76.1	53.7	75.2	75.2	**51.7**	74.3	76.9	53.6	75.9	75	**50.6**	74
NOUN, VERB, ADJ, PROP	78.5	54	**77.7**	74.5	48.5	73.4	79.4	53.4	78.5	66.6	39	61.5
NOUN, VERB, ADJ, AUX, PN	76.5	**54.5**	75.7	76.4	51.3	75.3	**80.4**	**55.2**	**79.5**	77	49.5	76
NOUN, VERB, ADJ, ADV, Sc	77.6	52.4	76.6	74.2	48.5	73.1	79.9	54.3	79.1	68.1	39.7	64.3
NOUN, VERB, ADJ, PN	**78.7**	54.3	77.6	**76.8**	51.3	**75.7**	**80.4**	**55.2**	**79.5**	**77.6**	50.4	**76.6**
No-POS filtering	76.5	**54.5**	75.7	**76.8**	51.3	**75.7**	**80.4**	**55.2**	**79.5**	77.5	50.4	76.5

2.7 Software

To implement the machine learning pipeline, python is used. SpaCy, a free open-source Natural Language Processing library, is used for tokenization, stop words removal, POS filtering, and lemmatization. On the other hand, Scikit-learn, a machine learning library, is used for feature extraction and all subsequent steps in the pipeline [4].

3 Experimentation

In this section the results for intents classification for the educational chatbot are presented. Table 2 shows the performance of training each ML classifier with and without POS filtering combinations.

The overall best performing model was RF as shown by BA, MCC, and CKS metrics, followed by SVM, XGBoost, and DT. Regarding, POS filtering we can observe that, the best performing POS combination is {NOUN, VERB, ADJ, PN}. Further, by reducing the number of features employed in the classification the performance of models is not only unaffected, it can even increase algorithms performance.

For instance, in the case of XGBoost and SVM, by reducing the number of features, the performance of the model is increased. In comparison, for DT and RF, reducing the number of features does not improves the accuracy of the models, however, it reduces the computational burden.

Furthermore, Figure 4 presents the confusion matrix for RF. On the x-axis the predicted values are shown, whereas on the y-axis the true labels are presented. Observe that intents corresponding to feel_check, navigation_menu, navigation_microred_caseta, search, and utility_check have in most cases, a perfect classification.

In contrast, intentions such as date_check, end_conversation, gratitude_check, joke_check, time_check, and weather_check have poor identification performance. An interesting fact is that, for weather_check, time_check, joke_check, gratitude_check, and end_conversation are misclassified as a feel_check intention. Similarly, date_check is misclassified with end_conversation, feel_check, joke_check. These points out that, phrases used for all these intentions have many common words worsening classes discrimination.

[4] https://colab.research.google.com/drive/1v3TPPkoyilhwTgXsmw-sm4BKdHIhIj7q?usp=sharing

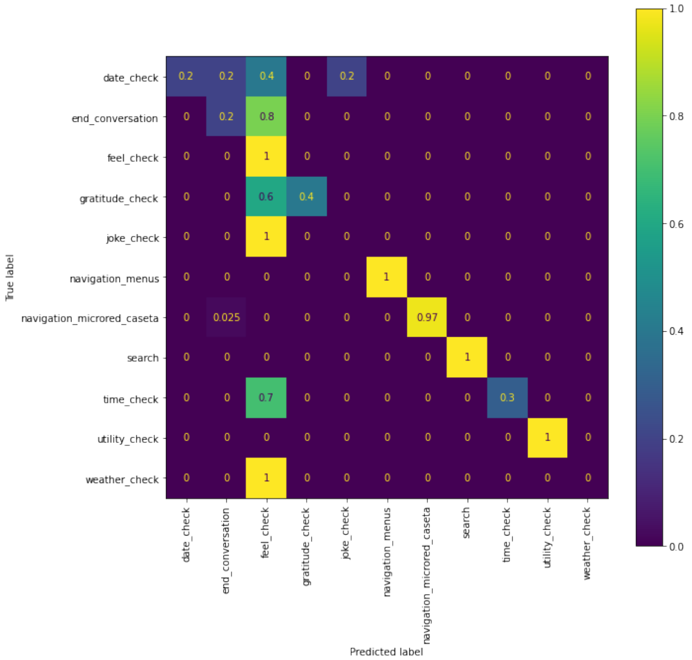


Fig. 4. Random forest confusion matrix.

4 Discussion and Conclusion

Based on a corpus consisting of 1850 phrases from an educational chatbot, a model for intent classification was developed using four different ML algorithms: SVM, DT, RF, and XGBoost. These are trained upon a common pipeline consisting in a pre-processing NLP stage (tokenization, stop–words removal, POS filtering, and lemmatization), feature creation using POS filtering and n-grams, and finally training and evaluating the models using three different metrics (BA, MCC, and CKS).

Results show that RF performs better than all other ML algorithms tested, achieving the highest score in every metric. Applying POS filtering in some algorithms improved their performance, and in others, the same performance is obtained. By applying the POS filtering an improvement in the computational burden of training ML models in larger dimensional spaces can be obtained, since less features are applied to obtain the best results. In particular, POS filtering has the greatest effect when applying it to XGBoost, where it produced the best result for all metrics, and SVM, where it improves the score in MCC and CKS.

Future work considers improving the dataset by increasing the number of phrases for each intent, in particular, those unrepresented in the dataset. Another venue of future work is the usage of word embeddings for feature extraction, and evaluating Deep Learning algorithms, such as DIET or BERT. Finally, the best resulting intent classifier will be implemented in the dialogue manager of an educational chatbot [7].

References

1. Chotirat, S., Meesad, P.: Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning. Heliyon, vol. 7, no. 10 (2021) doi: 10.1016/j.heliyon.2021.e08216
2. Hefny, A. H., Dafoulas, G. A., Ismail, M. A.: Intent classification for a management conversational assistant. In: 2020 15th International Conference on Computer Engineering and Systems (ICCES), pp. 1–6 (2020) doi: 10.1109/ICCES51560.2020.9334685
3. Helmi-Setyawan, M. Y., Maulana-Awangga, R., Rafi-Efendi, S.: Comparison of multinomial naive bayes algorithm and logistic regression for intent classification in chatbot. In: 2018 International Conference on Applied Engineering (ICAE), pp. 1–5 (2018) doi: 10.1109/INCAE.2018.8579372
4. Hwang, G. J., Chang, C. Y.: A review of opportunities and challenges of chatbots in education. Interactive Learning Environments, pp. 1–14 (2021) doi: 10.1080/10494820.2021.1952615
5. Larson, S., Leach, K.: A survey of intent classification and slot-filling datasets for task-oriented dialog (2022) doi: 10.48550/arXiv.2207.13211
6. Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., Mars, J.: An evaluation dataset for intent classification and out-of-scope prediction. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 1311–1316 (2019) doi: 10.18653/v1/D19-1131
7. Macias-Huerta, P. I., Santamarıa-Bonfil, G., Blanca-Ibañez, M.: CARLA: Conversational agent in virtual reality with analytics. Research in Computing Science, vol. 149, no. 12, pp. 15–23 (2020)
8. Santana, R., Ferreira, S., Rolim, V., de Miranda, P. B., Nascimento, A. C., Mello, R. F.: A chatbot to support basic students questions. In: IV Latin American Conference on Learning Analytics, pp. 58–67 (2021)
9. Schuurmans, J., Frasincar, F.: Intent classification for dialogue utterances. IEEE Intelligent Systems, vol. 35, no. 1, pp. 82–88 (2019) doi: 10.1109/MIS.2019.2954966

ISSN 1870-4069

Physiological and Behavioral Dataset Captured in a Learning Environment

Yesenia Nohemí González-Meneses, Blanca Estela Pedroza-Méndez,
José Federico Ramírez-Cruz, José Crispín Hernández-Hernández,
Edmundo Bonilla-Huerta

Tecnológico Nacional de México,
Tlaxcala,
Mexico

{yesenia.gm, blanca.pm, federico.rc,
crispin.hh, edmundo.bh} @apizaco.tecnm.mx

Abstract. In this article, we describe the experiment for the capture of physiological and behavioral signals of college students while doing a learning activity. This is a fundamental task for the recognition of emotions in specific contexts; it is part of the data processing stage in the methodology of the research project in which it is intended to automatically recognize emotions in educational environments using machine-learning algorithms and acquisition technologies of physiological and behavioral signals to identify the connections between emotions and learning. Data capture is guided by a formal procedure designed for a specific educational context. In this, the characteristics of the learning environment (of the physical place and of the software), of the participants and of the capture devices used were considered. 82 college students in engineering careers participated in the data capture. The learning activity was through interaction with a basic algebra MOOC with which they learned basic concepts by watching some videos containing the theory and practical exercises, in the end, they answered an evaluation of the subjects studied. We use 2 physiological signal capture devices and 2 video capture devices.

Keyword: Physiological and behavioral signals dataset, emotion recognition, learning-centered emotions, educational innovation, learning environments.

1 Introduction

In the learning process of a student, emotions have a determining influence on the acquisition of knowledge. In people, emotions awaken curiosity, interest, and, therefore, the focus of attention.

The learning that is associated with emotions generated in everyday life or in a classroom - whether positive (such as joy or interest) or negative (such as fear or sadness) - will remain in our memory. The identification of student emotions during learning processes helps to select teaching strategies more suited to the student's model.

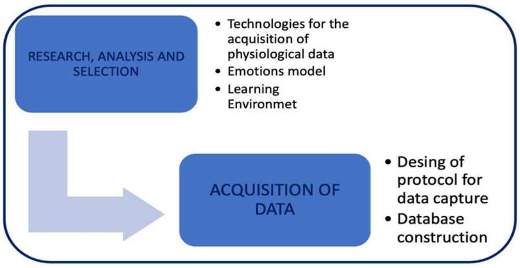


Fig. 1. Methodological stages.

This model is usually contained in educational software such as intelligent tutorial Systems (ITS). These types of tutorials consider students' emotions, their ways of learning, their intelligence, cognitive abilities, and other information that form the student's model [1].

The process of emotional recognition is less invasive when done automatically using devices that give us information about the emotional state of the student. The devices using for this work are thermal camera, video camera, heart rate sensor and Kinect 360 for Windows.

The problems of automatic recognition of emotions have been actively investigated in recent years. Despite the amount of research and attention, a clear solution which could be utilized by most people is still in the distant future. Several drawbacks have hindered the construction of an appropriate solution from a computational point of view. A big challenge to the effective performance of software and hardware that recognize emotions in real contexts is the difficulty in generating databases with spontaneous emotions.

Other aspects to consider are the areas of application, which have a significant effect on the accuracy of emotion recognition, and the degree of intrusion of the tools used [2]. The main objective in this work is to provide an adequate dataset for after processing them to build a database for the study of learning-centered emotions.

This is a fundamental task for the recognition of emotions in specific contexts. To justify the creation of the dataset, we did a review of state of the art on automatic recognition of emotions in learning environments and on databases for the identification of emotions. Considering the revision of the state of the art we identify the problems in the creation of a robust dataset for the automatic recognition of emotions.

Later, we set out the stages of procedure design for the capture of physiological and behavioral signals of students while they did a learning activity, and we present statistics of the dataset created.

After the analysis of recent related works, we can identify that in learning environments, there is a lack of adequate methodologies to recognize the emotional state of students during learning processes through interactions with computers. These methodologies preferably would need to integrate diversified characteristics obtained even from the fusion of data coming from the use of different technologies that acquire physiological and behavioral signals.

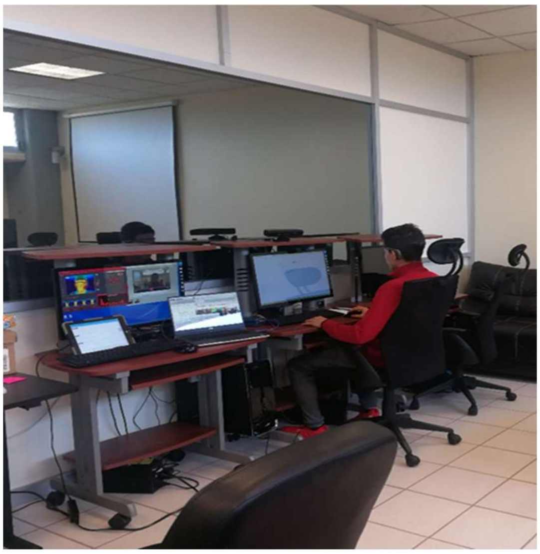


Fig. 2 User experience laboratory of the UPPue.

They should be based on a model of emotions that is realistic and allows becoming closer, overall, to the cognitive learning process that contributes to the analysis of the emotion-learning relation (i.e., the connections between emotions and learning).

We also identify that the main factor that affects the performance of emotion-recognizers in real contexts is the difficulty in generating databases of spontaneous emotions.

Generally, works are done with acted databases, which provide portraits of emotions to represent prototypical and intense emotions that facilitate the search of correlations and the subsequent automatic classifications.

The problems that are usually faced in the capture of physiological and behavioral signals data in real contexts are: Use of invasive technologies that cause discomfort in the capture of spontaneous expressions, synchronization in the different capture devices, the environmental conditions where the capture takes place, the correct sensing of the devices, the previous emotional state of the participants, the availability of the participants, the physical appearance of the face (amount of hair or artifacts on the face), the necessary capture time.

Due to the above mentioned, the formal design of a procedure for the capture of data (either physiological, behavioral or both) is necessary. The procedure will guide the execution of a controlled experiment with the objective of avoiding errors in the acquisition of data, describing the technologies to be used, the capture process, the environmental conditions, and the preparation of the participants.

To address the lack of methodologies owed to the weaknesses of existing data bases, we made a priority the creation of a more robust dataset whose construction is guided by the steps of a formal procedure for data capture (either physiological, behavioral, or both). This priority is important because data are the basis for the implementation of

Table 1. Sample characteristics statistics.

Total Students	Average Age (years)	Men	Women	Engineering	Master	PhD	High School
82	21	50	32	69	8	4	1

a model of emotion identification with an acceptable level of accuracy that is more in line with reality in an educational context.

2 State of the Art

To address the state of the art, we first reviewed the research on automatic recognition of learning-centered emotions. After this analysis, we identified the lack of databases with emotional expressions captured in real learning environments as the main problem.

From this, we made an analysis of the databases constructed in various research works and captured both in generic environments and in specific educational contexts. In this way we identify that the creation of a physiological and behavioral database as an initial stage in the recognition of learning-centered emotions and the problems that must be addressed in the construction of the dataset is necessarily.

A database of emotional expressions is a collection of images or video clips, speech, and physiological signals related to a wide range of emotions. Its content corresponds to emotional expressions related to the context in which they were captured and then tagged; this is essential for the training, testing, and validation of algorithms to develop expression-recognition systems.

The labeling of emotions can be done on a discrete or continuous scale. Many of the databases are usually based on the theory of emotions [3] that assumes that there are six basic emotions and an approximate variety of 22 secondary emotions on a discrete scale. However, in some databases, emotions are labeled on the continuous arousal-valence scale [4]. And other databases include the AUs (Action Units) based on the FACS (Facial Action Coding System) [5].

The databases of expressions of emotions are mostly formed only by facial expressions and these are classified into posed and spontaneous. In posed expression databases, the participants are asked to display different emotional expressions, while in spontaneous expression database, the expressions are natural. Spontaneous expressions differ from posed ones remarkably in terms of intensity, configuration, and duration. In most cases, the posed expressions are exaggerated, while the spontaneous ones are subtle and differ in appearance.

Apart from this, synthesis of some AUs are barely achievable without undergoing the associated emotional state so it is not possible to capture physiological data since these cannot be controlled by people and therefore do not correspond to the emotion that is acted out.

For our research we first made a compilation of databases of facial expressions (approximately 18 databases of works from 2010 until June 2021), although our interest lies in hybrid databases that contain facial images and physiological data, so far, we have only found one database of this type publicly available. The DEAP database [6]

Table 2. Data obtained from each of the experiments.

# Experiment	**1**	**Pulse**	**YES, range: 513-553, Mean: 546**
Date	10/10/2018	% Pulse Data	83%
Name	Student´s Name	Total Captures	21672
Sex	M	Sample Range	9.505263158
Age	20	Captures Per Minute	571
Engineering	Computation Science	Files CSV/JPG	2074 / 0
Set Time (min)	38	Time AVI (min)	03:48
Set Time (seg)	2280	Video (min)	39.00
Start Time	08:40	Kinect (in)	39
Ending Time (real)	09:19	Observations	none
Real Time	39		

Table 3. Statistics of raw data files.

Pulse Sensor "txt"	**Thermal Camera "csv/jpg/avi"**	**Webcam "wmv"**	**Kinect "xed"**	**Files Per Experiment**	**Total Files (82 Experiments)**
1	4801	1	1	4804	393,928

contains physiological recordings (of EEG) and facial video of an experiment where 32 volunteers watched a subset of 40 music videos.

In this experiment participants are also asked to rate each video according to the emotion it caused in them. These databases correspond to facial expressions of basic emotions, which have been captured while people perform different activities.

Most databases are comprised of viewers watching media content (i.e. ads, movie trailers, television shows, animated gifs and online viral campaigns) and of posed facial expressions like the database DEAP [6], CK++ [7] and Affective (the most robust database for the amount for data stored) [8].

Subsequently, we also compiled databases of facial expressions that correspond to learning-centered emotions (approximately 13 databases) like the database of [4, 9] that have been captured while students perform some learning activity in a wild setting with spontaneous facial expressions, so it is possible to capture physiological data to complement them.

These criteria are what we consider for the selection of the databases analyzed. These correspond to learning-centered emotions that have been captured while students perform some learning activity in a wild setting.

Some of them contain physiological data but unfortunately, they do not make them available to the public or provide information about the characteristics of these data, let alone details of the processing they do with them. Of these, four databases include physiological data. The most complete is the one mentioned in [10] in which 67 students participate and use three physiological data sensors and a video camera.

In general, they have been created with data from very few participants. From this analysis, we have identified as an area of opportunity the creation of a physiological and behavioral database supported by a formal procedure for data collection. The definition of the procedure allows us to run a controlled experiment in a natural environment and specific that can be replicated the number of times necessary to create a robust dataset.

3 Methodology

In this section, we describe the methodological stages, which can be seen in Figure 1. First, we will describe the activities of research, analysis, and selection of critical elements for the creation of a spontaneous dataset captured in real educational environments through the execution of a controlled experiment.

This description includes the physiological and behavioral signals to be captured, the data-acquisition technologies used to capture the signals, the emotions model on which the identification of emotions is based, and the activity in which the students participate while their data is captured.

The next stage includes the creation of the dataset. We will create a dataset of student features obtained in real time while doing a learning activity. Such a database will be formed from two physiological characteristics (the temperature of the face and the heart rate) and two behavioral characteristics (the facial image and the movements of the head and of the superior limbs).

3.1 Investigation, Analysis and Selection

After the corresponding investigation and analysis, the first three elements of this stage have been selected. The physiological and behavioral data acquisition technologies that were chosen are the following:

1. Logitech Webcam®: to capture continuous video during the 35 minutes that the experiment lasts in "MP4" format (1920X1080, 28 f / s).
2. ICI 9320P thermal camera: to capture arrays of face temperatures, in "CSV" files (240X320), generating 1 file per second. An "AVI" video (600 images per minute) of approximately 3.5 minutes is also generated.
3. Cardiac pulse sensor implemented with Arduino: a "TXT" file of the sensor signal power readings is generated with a range of approximately 9.5 captures per second (10Hz sampling rate).
4. Kinect for Windows®: continuous video is recorded during the minutes that the experiment lasts in an "XED" file.

For the emotion model, for now both the discrete and the continuous model will be considered. The types of learning-centered emotions that are tried to recognize are: interest (pleasure), boredom, confusion, and frustration.

The educational environment with which students will interact for the teaching-learning process will be a MOOC of basic Coursera algebra working only with the first subject (approximate duration of 36 min.). The young people of the sample will be college students of careers in the engineering area, whose ages range between 18 and 24 years.

Table 4. Number of experiments with complete files per device used.

Thermal Camera	Pulse Sensor	Video	Kinect	Complete Experiments	Thermal Camera, Video and Pulse
64	48	79	73	44	34

3.2 Procedure for the Capture of Physiological and Behavioral Data

The procedure guides the execution of a controlled experiment with the objective of avoiding errors in the acquisition of data, describing the technologies to be used, the capture process, the environmental conditions, and the preparation of the participants.

In the first part of the procedure, the stages of research, analysis, and selection of key elements for the creation of a spontaneous dataset captured in real educational environments are described. These were already described in the previous section.

The steps to follow to carry out the data capture experiment are listed below:

1. Upon arrival the student is cordially greeted.
2. You explain the experiment.
3. You explain the definitions and characteristics of the emotions.
4. You ask him/her if he/she has any questions.
5. You are given the *Informed Consent Format* to sign the authorization to record your data.
6. You ask to sit in front of the computer.
7. The use of the tutorial and the self-assessment is explained.
8. The heart rate sensor is placed on the left ring finger.
9. The Kinect, the video camera, the thermal camera, and the heart rate sensor are initialized.
10. Data recording starts.
11. Once the MOOC session is over, device recordings stop.
12. Captured data are saved in storage units.
13. Student is asked how he feels and is grateful for his participation by offering some candy, cookies, or a bottle of water.

In the second part, we explain legal, ethical, and environmental (physical place) aspects that should be considered in the design of a formal procedure. Regarding legal and ethical aspects, in Mexico, we will be respectful of the Federal Law on the Protection of Personal Data in Possession of Individuals DOF: 05/07/2010, published in the Official Gazette of the Federation on July 5th, 2010, this protects the privacy of individuals with respect to the treatment we give their personal information.

Its provisions are applicable to all individuals or corporations, public and private sector, both at the federal and state level, which carry out the processing of personal data in the exercise of their activities, therefore, companies such as banks, insurers, hospitals, schools, telecommunications companies, religious associations, and professionals such as lawyers, doctors, among others, are obliged to comply with what is established by this law.

Also, part of the procedure is the design of an informed consent letter with which the student authorizes their participation and the use of their data for research purposes. In this letter they are asked for their personal data, and they are explained what the experiment consists of.

In addition, they are informed that there is no consequence of deciding whether to participate in the experiment. They are informed that full confidentiality of their data will be kept and that they will only be used for research purposes. Finally, they are asked to authorize their participation by signing the letter.

4 Implementation of the Experiment

For the execution of the experiment, the voluntary participation of students in a learning activity while their behavior and physiological signals are being recorded through different data acquisition devices is necessary. In this section, we will describe the details about the execution of the data capture experiment based on the design of the procedure described above.

Information such as characteristics of the selected sample, place for data capture, devices used and their location, the way in which the data was collected, and creation of files with the captured data and their storage will be described. Finally, we will show a summary of the results obtained, the dataset created, and the statistical analysis of the information collected.

The following captures were carried out at the *Universidad Politécnica de Puebla* (UPPue) in its User Experience Laboratory, with the participation of engineering, masters, and doctoral students. In figure 2 we can observe the physical characteristics of the place.

At UPPue, 82 students from different engineering (information technologies, biotechnology, computer science and financial sciences) were captured. Statistics of the characteristics of the total experiments carried out so far as number of participants, age, sex, and level of studies, are shown in Table 1. Of the 82 participants, 50 are men and 32 women with an average age of 21 years.

The information shown in Table 2 was collected from each experiment. The data from experiment 1 is taken as an example. The first part is general data such as experiment number, date of capture, start time, end time, participant data. The second part corresponds to the emotions test responses, statistics of each of the sensors used and the video cameras.

At the end there is an observation section to describe some special situation that arose during the capture and needs to be documented. So far, the dataset contains the raw data sensed by each device used, stored in their corresponding files. Pulse sensor data is stored in 1 data file, *txt*. For thermal camera data the number of files varies depending on the scheduled capture time, if the programmed time is 40 minutes, 2400 *csv* files, 2400 *jpg* files and 1 *avi* file (4801 files) are stored.

Kinect video is stored in 1 *xed* file. And the webcam video is stored in 1 *wmv* file. Thus, by experiment, they are generated with a total of 4804 files, stored on an external 4TB hard drive. The summary of stored raw data statistics is shown in Table 3.

For the statistical analysis of the thermal camera temperature matrices, 360 more files are generated per minute to be analyzed, since the complete files of the face per minute are divided into 6 zones (front, nose, mouth, chin, right cheek, and left cheek) thus generating 60 new files per minute (for each area of the face) per 6 zones, in total 360 files per minute. Thus, the number of new files generated is increased in proportion to the number of minutes to be analyzed.

In the case of the analysis of the heart pulse files, 2 new files are generated for each minute to be analyzed, one to store the beats per minute and the other to store the signal power readings of the minute to be analyzed, if we analyze 6 minutes it gives us a total of 12 more files.

A summary of the number of experiments that have complete data files for each of the devices used is shown in Table 4. In this we can also see the number of experiments considered as complete, these are those experiments that are counted on with information on at least 4 of the devices used.

Thus, we have 78% of experiments with complete thermal camera files, 59% with useful pulse sensor data, 96% with full video and 89% with Kinect video. This represents 54% of experiments with complete data files from at least 4 devices.

5 Conclusion

The dataset contains the raw data counted by each device used and stored in their corresponding files. All these files make up raw data dataset. We also performed a summary of the number of experiments that have complete data files for each of the devices used.

The experiments considered complete are those experiments with which there is information on at least 4 of the devices used. So, we have 78% of experiments with complete files of the thermal camera, 59% with useful data from the pulse sensor, 96% with full video, 89% with video from the Kinect. This represents 54% of experiments with complete data files from at least 4 devices.

Constructing a dataset with physiological and behavioral signals of humans performing activities in real environments, both controlled and uncontrolled, is a complex, laborious and time-consuming task, especially when the challenge is to integrate data from various devices to capture different signs. Regarding the scope, a dataset was developed with captures from 2 physiological signal acquisition devices, heart rate and temperature.

Behavioral data of 82 students were obtained through video recordings. This fulfilled the objective of creating a dataset of physiological and behavioral information obtained from students learning in real online environments.

Regarding the contributions to the state of the art of automatic recognition of learning-centered emotions, they can be identified first because of the scarcity of dataset of physiological and behavioral signals of individuals performing learning activities, the creation of a formal procedure for the capture of data associated with learning-centered emotions.

This can be generalized for application in different virtual learning environments. In addition, a dataset with physiological signals of heart rate and facial temperature data is created, with behavioral images obtained from video recordings of the face of students performing learning activities. This type of data was not jointly identified in any of the works analyzed in the state of the art.

Most of the dataset found in the state of the art are created with data from a few students and consist only of video images, with a maximum of 9068 images. From this dataset, different research proposals may arise that complement the results obtained.

References

1. González, C., Mora, A., Toledo, P.: Gamification in intelligent tutoring systems. In: Proceedings of the second international conference on technological ecosystems for enhancing multiculturality, pp. 221–225 (2014) doi: 10.1145/2669711.2669903
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G.: Emotion recognition in human computer interaction. IEEE Signal processing magazine, vol. 18, no 1, pp. 32–80 (2001) doi: 10.1109/79.911197
3. Ekman, P.: Emotions revealed: recognizing faces and feelings to improve communication and emotional life. New York: Times books (2003)
4. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learning and Instruction, vol. 22, no. 2, pp. 145–157 (2012)
5. Ekman, P., Friesen, W. V., Hager, J. C.: Facial action coding system. The manual, Salt Lake City, Network Information Research Corporation (2002)
6. Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. Patras, I.: Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing, vol. 3, no. 1, pp. 18–31 (2012) doi: 10.1109/ TAFFC.2011.15
7. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: IEEE computer society conference on computer vision and pattern recognition-workshops, pp. 94–101 (2010) doi: 10.1109/CVPRW.2010.5543262
8. El Kaliouby, R., Picard, R. W.: Affectiva Database. MIT Media Laboratory (2019) https://www.affectiva.com
9. González-Hernández, F., Zatarain-Cabada, R., Barrón-Estrada, M. L., Rodríguez-Rangel, H.: Recognition of learning-centered emotions using a convolutional neural network, Journal of Intelligent & Fuzzy Systems, vol. 34, no. 5, pp 3325–33636 (2018)
10. Botelho, A. F., Baker, R. S., Heffernan, N. T.: Improving sensor-free affect detection using deep learning. In: Artificial Intelligence in Education: 18th International Conference, pp. 40–51 (2017) doi: 10.1007/978-3-319-61425-0_4

ISSN 1870-4069

Building a Dataset for Professional Interest Recognition

María Lucia Barrón-Estrada, Ramón Zatarain-Cabada,
Abel Robles-Montoya, Héctor Manuel Cárdenas-López,
Arcelia Judith Bustillos-Martínez, Néstor Leyva-López

Tecnológico Nacional de México,
Instituto Tecnológico de Culiacán,
Mexico

{lucia.be, ramon.zc, arcelia.bm}@culiacan.tecnm.mx

Abstract. The personality of the student is a key factor influencing the choice of career as students choose their career according to their personality. However, young people do not always know which options are the most suitable for their own characteristics. This paper presents the design and implementation of an intelligent system to help the student to better choose the professional career to study using a video of the same student as input. For the development of this system, a dataset was used that was created with personality and professional interest questionnaires, videos, and information related to the academic and professional training of the participants. This dataset was used to train machine and deep learning models that can predict personality and career interests from video. The results of the personality prediction models were in the order of 0.18, while for the model of professional interests it was 0.16.

Keywords: Automatic personality recognition, professional interests, RIASEC, HEXACO, machine learning.

1 Introduction

The term "dropout" is used to describe students who stop attending school for a variety of reasons, including personal obligations, academic failure, or the expiration of their enrollment [1]. In Mexico, however, the Secretaría de Educación Public (SEP) uses the term to describe students who stop attending school without considering their academic performance, choice of careers, or other factors.

We refer to school dropouts throughout this essay using the first term. Only 6 out of 10 Mexican students enrolled in higher education (university and technology bachelor's degree) can complete their studies, according to data released by SEP (2019-2020) [2]. The socioeconomic and personal factors are the most frequently cited causes for the 40% of students who drop out of university.

However, other significant factors include academic dissatisfaction, poor school performance, and career choices [3]. According to Holland's theory, students choose academic environments in accordance with their personalities [4], but young people don't always know which options are suitable for them considering their own interests.

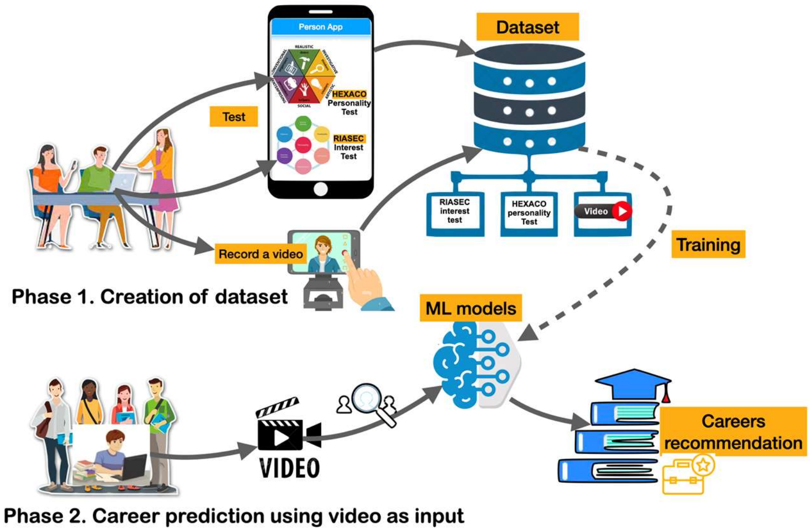


Fig. 1. General system to predict career based on personality and interest.

The personality of the student has been identified as a factor that influences its career choice. Therefore, career counseling using instruments (such psychometric tests) can help students discover a career appropriate to their personality and therefore prevent school dropout.

This can lead to a number of circumstances, such as not obtaining an adequate academic performance in that career or losing interest. Numerous academic fields, including psychology, philosophy, and others, have extensively examined personality analysis [5]. The computer vision community has recently showed interest in identifying personalities, starting with physical information expressed through the face, postures, or behaviors of an individual [6].

The personality analysis can benefit society in a variety of other ways as well. For instance, it can be used in education to determine the best method of instruction; in business when hiring employees, it can be used to choose candidates who have the right personality for the job; and in educational guidance, it may be used to assist young people in choosing their field of study to reduce the number of school dropouts.

Vocational guidance is a collection of procedures focused on the vocational dilemma that tries to give each subject the resources necessary to make the best decision possible given their circumstances [7]. To help people discover more about themselves and choose a job they enjoy while also achieving a gratifying performance, vocational counselling is now backed by approaches and tools like interviews, surveys, or psychometric testing [8].

The Kuder Vocational Interest Scale, the Explora Test, and the Basic Academic and Professional Interest Areas Questionnaire are some of the vocational instruments suggested by various authors; each one uses a different methodology to identify the area or profession that best suits each person [9]. A software system was created that, through personality recognition using videos, can predict which professional fields of study are most suitable for each student.

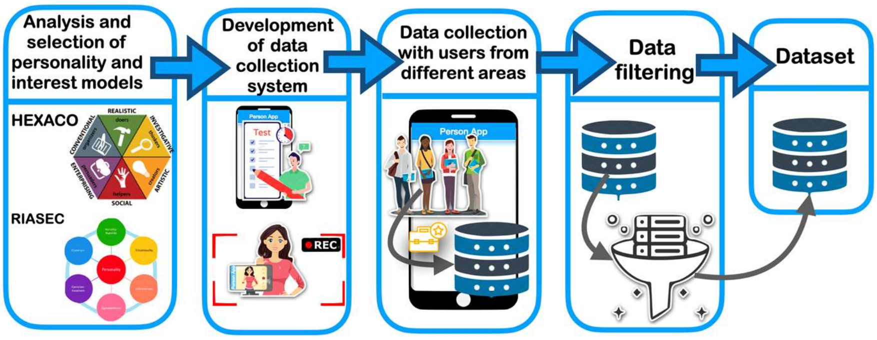


Fig. 2. Methodology for the creation of the personality-interest corpus.

This is because school dropout is a problem, and because academic dissatisfaction and poor school performance are linked to a poor career choice. Two artificial intelligence (ML) models are required for this process to make predictions.

The first model was trained using user-recorded videos and personality questionnaires in such a way that the model received the videos as input and generated as output the user's personality attributes; the second model received the personality attributes and responded with the user's vocational interests.

By combining the two models, it is possible to predict the user's vocational interests based on the personality attributes. The key obstacle to attaining it is the need for a large dataset of user videos, personality tests, and career interests. This dataset needs to have enough records and high-quality data to enable the models to train.

In this paper, a mechanism for predicting a user's associated vocations by personality detection in videos is presented. The paper structure is as follows: section 2 shows related works that were analyzed in this research; in section 3 we explain the method to develop this research work; section 4 presents the results that were obtained; and finally, section 5 shows conclusions and limitations.

2 Related Work

Interest has been the subject of studies in vocational psychology that have been applied in actual situations, including the First World War. Academics had known this for most of the previous century, and American industry swiftly embraced its significance in choosing employment. The first instrument to measure people's interests was created in 1919 by Clarence S. Yoakum at Carnegie Institute of Technology. He used a logical sampling strategy meant to represent the domain of interests [4, 10].

Holland's theory, which states that it is possible to categorize each person in one of six different dimensions (RIASEC) —Realistic, Investigative, Artistic, Social, Enterprising and Conventional—or a combination of these—has been one of the most influential for the development of various questionnaires to assess interests [4, 11]. This hypothesis has inspired the creation of numerous tools throughout the years, including questionnaires that assess each person's interests.

Table 1. PersonApp 2.0 Functional requirements.

Requirement	Description
RF1	The system will allow users to take a standardized vocational test.
RF2	The system will allow users to take a standardized personality test.
RF3	The system shall store user information.
RF4	The system will show the user a description of his or her vocational interests.
RF5	The system will show the user a list of careers according to their interests.
RF5	The system will show the user a list of careers according to their interests.
RF6	The system will allow downloading a report of the results in PDF format.

The Department of Labor in the United States created O*NET, a career exploration and planning tool, in 1999 using Holland's six kinds as a benchmark. This platform provides a condensed version of the interest questionnaire that only accounts for 10 questions for each RIASEC category and totals 60 questions [12, 13].

Hansen [13] then examined the psychometric properties of a condensed form of the interest questionnaire and discovered evidence of its reliability, validity, and excellent concordance with the full version.

The U.S. Department of Labor uses this survey under a free-use license.There have been several, long running, and diverse attempts to identify a model that might accurately capture human personality on a global scale [14]. The Big Five model, which consists of five dimensions that each measure a different personality attribute, is one of the most extensively used models to describe how people differ from one another. The five dimensions—Openness, Consciousness, Extraversion, Agreeableness, and Neuroticism—are collectively referred to as the OCEAN model.

Each dimension is symbolized by its first letter. K. Lee and M. Ashton created the HEXACO-PI-R instrument, which Roncero Fornés and Belloch translated into Spanish with the authors' approval [14]. The validity of the Spanish version of this instrument was examined, and it was determined that the internal consistency and stability were very satisfying.

They came to the conclusion that HEXACO is a suitable tool to evaluate personality. Since the authors claim that this model adds a new dimension called Honesty that completes all the features of a person and increases interpretability, it has lately been researched for its theoretical advantages over the most widely used model of the big five, OCEAN [15].

The six dimensions of the HEXACO model are as follows: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness and Openness to Experience [14, 16]. In two separate studies, verko & Babarovi [17] describe the relationship between personality and interests.

In the first study, 602 high school graduates completed the HEXACO model using the HEXACO-PI-(R)-100 questionnaire and the OCEAN model using the IPIP-50 questionnaire, and in the second study, 981 participants completed the PGI-Short and HEXACO-60 questionnaires.

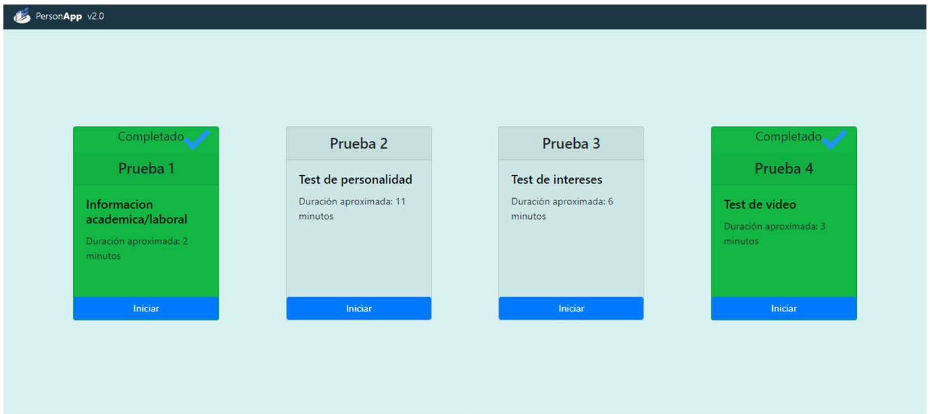


Fig. 3. PersonApp main menu.

Positive or negative correlations between interests and personality were found in the results of both investigations; also, the HEXACO model shown predictive advantages in describing interests compared to the Big Five.

In a recent study, Barmanpek [18] discovered that personality and interests are generalizable and do not vary across time, groups, or cultures. He also discovered some direct relationships on some RIASEC and HEXACO model variables. Tandera, et al. [19] describe a technique to estimate personality using the user's Facebook information and the OCEAN model.

In this study, they evaluate the outcomes of using machine learning and deep learning and concluded that the latter is superior to the former since deep learning can increase the accuracy of the majority of the features.

By categorizing participants of a video social network and then utilizing the Big Five model to identify, frame by frame, the emotion present in each facial expression, Biel, et al. [20] take advantage of state-of-the-art technologies for facial expression identification.

They concluded that these can be helpful for predicting personality after discovering that particular facial expressions are closely associated to some personality traits. Using text mining algorithms, Pavithran and Ashraf [21] try to determine career affinity. The theoretical foundation for this research, which makes use of the O*NET career database, is Holland's RIASEC model.

The method entails removing text from Twitter, preprocessing the words, labeling them in an RIASEC dimension, counting the words, and so on. Walker, et al. [22] present the "Basel Face Database", which integrates three various image gathering methodologies for personality recognition. To display variations in the dimensions of the big two (communion and agency) and the big five, images of real faces were used to create this database.

The procedure involved selecting participants from Amazon Mechanical Turk and obtaining their agreement for data collection before displaying a random selection of both original and changed photographs and asking them to evaluate the images by ranking them according to each personality dimension.

3 Methodology

A new approach was adopted to carry out this project that made use of an intelligent system. Like a standardized vocational orientation test, it uses video analysis to identify the user's personality and traits before recommending a career. The benefit of this technique is that students simply need to film a little video instead of filling out a lengthy questionnaire.

Fig. 1 provides a general overview of the system and clearly illustrates the system's two phases: phase 1 involves the creation of a dataset for training machine learning models, and phase 2 involves making career recommendations using user videos as input. The HEXACO model's and the RIASEC interest model's personality traits must be used in the system's implementation.

A platform is utilized to gather data from many users by distributing personality and interest questionnaires and capturing videos through a few trials, as shown in Fig. 1, phase 1. A dataset will be created from the data collected using that platform and used to train neural networks. One of them will analyze video to predict personality using questionnaires, while the other will use personality to predict career preferences.

This method is suggested rather than predicting interests directly from video due to the vast research that already exists on automatic personality detection. By combining these models, it is possible to predict career interests from videos (see Fig. 1, phase 2).

A four-stage methodology was created to develop a dataset for personality and interest recognition based on the research examined on the generation of corpus or datasets for attribute prediction, as illustrated in Fig. 2. The following sections detail each stage of the methodology.

a) Analysis and Selection of Personality and Interest Models.

The background information on personality and interest models was examined to determine which model would be most useful for the proposed situation. We chose the most well-known and extensively researched psychological models. Finally, it was confirmed that there were questionnaires with research aims for each model since the goal is to predict professional interests using personality. Both models needed to have direct links between their traits or the dimensions they cover.

In light of the aforementioned requirements and the background data, it was discovered that Holland's theory can accurately represent interests and also has connections with several aspects of other personality models. In addition, the O*NET database uses the RIASEC model that Holland proposed, and for these reasons, this model was chosen to assess professional interests.

The HEXACO model was chosen to assess personality because to advancements over the conventional Big Five model as discussed in the background and, more significantly, due to evidence of a greater association between the traits represented in HEXACO and the RIASEC interest model.

b) Development of the Data Collection System (videos and questionnaires).

With the use of the OCEAN model, a software program called PersonApp [23] was created to automate personality questionnaires and collect data while also producing a video dataset.

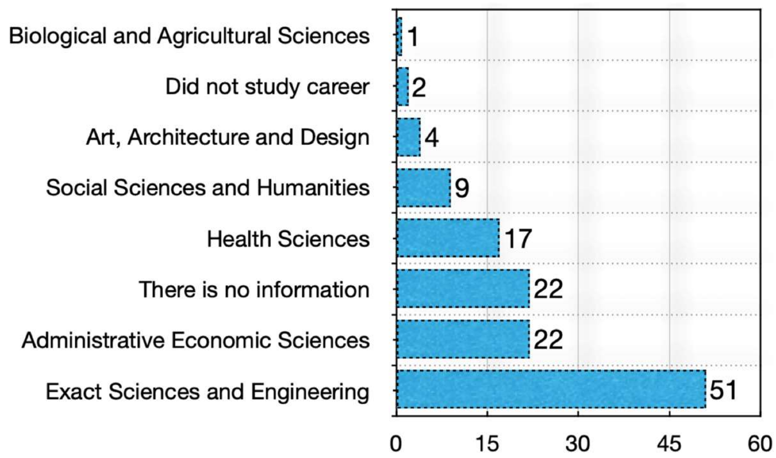


Fig. 4. Number of records by area.

The system specifies three experiments for the filming of the videos, each of which requires the user to discuss a different subject for the purpose of four quick 15-second recordings. The IPIP-50 is the used personality test. A client-server architecture was used to construct this system utilizing a layered architectural paradigm.

The PersonApp program served as the foundation for the creation of a new version that considers the updated specifications created to produce the datasets of videos, personalities, and professional interests. The updated version, known as PersonApp 2.0, integrates the functional specifications listed in Table 1 that were discovered during the analysis and model selection stage.

The following modifications were considered for PersonApp 2.0. As shown in Fig. 3, a monitoring interface was first added so that the user could see his/her progress as he/she finished each test.

The RIASEC interest test, which consists of 60 multiple-choice questions on users' interests and 2 control questions that direct users to choose a certain response to ensure that the results aren't random, was also added. For the responses, we utilized a Likert scale with five possibilities, where the first one has a value of 1 and goes up to 5 for the last option (1: I like it very much, 2: I dislike it, 3: I am not sure, 4: I like it, and 5: I like it very much). Each question in the questionnaire has five possible answers.

The third modification to the personality test replaced the IPIP test that was previously used to determine the users' OCEAN model values. This test was replaced by HEXACO, which has 72 questions with 5 possible answers (1: Totally disagree, 2: Partially disagree, 3: Neither agree nor disagree, 4: Partially agree, 5: Totally agree), of which 70 are about the test and 2 are control questions.

The fourth update was the addition of many questions to gather information about some of the participants' academic backgrounds, including their schooling and professions, positions held, and others. The gathering of data was performed in the non-relational Google Firebase database. The values that correspond to each model are determined when the user completes each test, and these values are then translated into a range from 0 to 1 to be saved in the database.

Table 2. Number of records by gender.

Gender	Number of records
Female	36
Male	70
Prefer not to say	2
No information	20

Additionally, information regarding the user's response time, how many of the control questions they successfully answered, user reviews of the program, and videos of the three trials are all preserved. Tests with several individuals were conducted before the application for database generation was released to ensure proper performance.

c) Data Collection with Users from Different Areas.

A group of 30 experts from various fields were given the app to test how well it worked across various platforms. To fill out the tests, certain individuals from various backgrounds (mainly students and professionals) were chosen, resulting in a sizable percentage of individuals with comparable personality traits in the database. For the prediction of interests to be possible, this is crucial.

Participants were chosen from among classmates, friends, and family members for the initial stage of data collection, which started with gathering as much information as possible without paying close attention to maintaining a balanced number of entries for each subject area.

To adopt a procedure centered on data collecting and expand the number of entries, undergraduate students from the Instituto Tecnológico de Culiacán joined the project in a later stage. For this iteration, participants were sought from fields with less items in the database.

d) Data Filtering.

To ensure that it can be used to train artificial neural networks, all data gathered into the dataset must pass a filtering process. They must accurately reflect the professional interests and personality of the users; otherwise, no neural network or other type of technique will be able to find the relationship between these variables. It is crucial to remove the inconsistent data that indicate random replies for the dataset to satisfy the needs of this job.

Two mechanisms were used to identify these cases:

- Categorize the amount of time the user spends responding to the questionnaire's questions. Outliers were eliminated, and a normal curve was plotted to determine that, in the personality and interest tests, 210 and 120 seconds, respectively, are the minimal times at which it is assumed that the user completed the questions diligently.
- Verify control question responses. The user is given instructions to select a specific response, such as "I really dislike it," and if they select any other response, it is assumed that they answered the questionnaire at random.

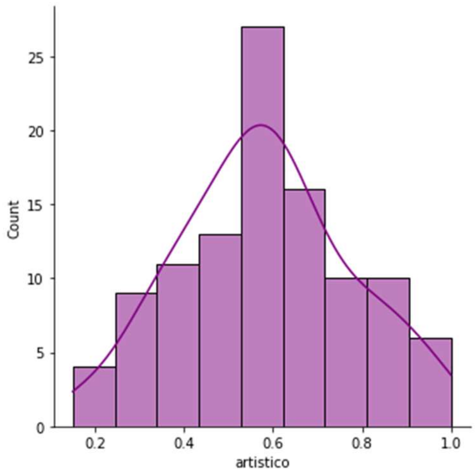


Fig. 5. Graph of the distribution of aristic attibute.

4 Results

The launch of the app gave participants a chance to voice any feedback or issues they had with the setup. The following major adjustments were suggested by users who gave us feedback on the platform, we also found areas for improvement, as follows:

- A warning note was added at the start to inform the user that he/she must agree to participate and that their information would only be used for research purposes.
- Users of the app can view their progress on a menu.
- The software has a feature that allows you to pause video recording and restart it.
- Visual components were added to help the user know when an experiment's recording is complete.
- An additional counter was included to show the recording time.
- Additionally, email validation was added, which the user must confirm in order to access their account.
- Also included is a screen that instructs consumers on which option to select based on their circumstances.

The data on the users' academic backgrounds was used to create the graph in Fig. 4, where it is possible to see the number of records for each area of knowledge. Given that 50% of the data came from the sciences and engineering, these fields attracted most participants. In contrast, only a small number of people engaged in the fields of biological and agricultural sciences, as well as art, architecture, and design.

It will be crucial to attempt and collect additional data from the areas with less records because the uneven data could make it challenging to develop a model that can generalize effectively. Table 2 lists the number of records broken down by the gender of those who responded to the app's questionnaires.

Table 3. Descriptive statistics for each attribute in HEXACO and RIASEC.

Attribute	Mean	St. dev.	Min	Max
Neuroticism	0.44	0.11	0.16	0.7
Honesty-Humility	0.63	0.07	0.48	0.76
Emotionality	0.67	0.09	0.46	0.9
Extroversion	0.68	0.09	0.42	0.9
Agreeableness	0.68	0.08	0.52	0.88
Conscientiousness	0.65	0.07	0.44	0.76
Openness to experience	0.68	0.08	0.48	0.84
Realistic (doers)	0.51	0.18	0	1
Investigative (thinkers)	0.61	0.2	0.2	1
Artistic (creators)	0.58	0.19	0.15	1
Social (helpers)	0.6	0.18	0.1	0.97
Enterprising (persuaders)	0.58	0.2	0	1
Conventional (organizers)	0.54	0.2	0	1

Like the knowledge domains, there is a significant disparity in the data; in this case, there are approximately twice as many men as women. Figure 5 is a graph that represents the results of the personality test's Artistic attribute, all other attributes follow a similar pattern. These graphs demonstrate how the data fit into a normal distribution quite well, however the ranges for the personality tests are not comprehensive.

This can have a severe impact on how these data are used to train machine learning models because, in the specific case of neuroticism, it has a range from 0.16 to 0.7. This pattern is reproduced in the other traits of the HEXACO model, where there are also incomplete ranges, but in different measures.

The PersonApp 2.0 system created 215 accounts, of whom 193 gave information on their education and employment, 174 completed personality and interest questionnaires, and 98 participated in video experimentation. The appropriate filters were applied to each questionnaire, yielding 127 records for the final database.

Some data about the dataset are included in Table 3. The first column includes all the HEXACO and RIASEC model features, as well as neuroticism, which is an OCEAN model attribute. The table only considers filtered data from 128 participants, with maximum and lowest values of 1 and 0 for each attribute, respectively.

5 Conclusions

In this study, the issue of school dropout was discussed, along with how computing may help to enhance current solutions. The work in this paper, presents the creation of a platform where users can take personality and career interest tests and produce videos.

Using this platform, information was gathered by using specific filters to eliminate any incomplete or incorrect information, and ultimately, a database was created that had 127 questionnaires and 98 videos.

When conducting these types of experiments, using a digital platform can be much more helpful than using other, more rudimentary methods (such as filling out questionnaires and filming videos in person) because the data can be collected, stored, and organized automatically. However, it would be important to have more mechanisms in place to ensure the accuracy of the data.

The dataset has some drawbacks, including an uneven distribution of people by sex and by various types of expertise. As a result, the number of records is not equal for each area and for each gender. Another issue resulting from the unbalanced data is that since most people have similar traits, such as gender and career, the values for each attribute fall within a narrower range.

Since the data are unbalanced, there are no records for the whole range of the data, even if each characteristic measured has values ranging from 0 to 1. Because they require balanced data to explore for links between variables, rather than just one or two, this issue makes prediction models ineffective.

Future work will continue to expand the database to achieve a better balance of the experiment participants' characteristics. Additionally, the project's next phases will see the development and optimization of prediction models for both personality and professional interests in order to use them in a system that can take user videos as input and recommend products to them.

References

1. Aldaco-Linares, R. N., Carpio-Hernández, M. M.: Las estrategias para abordar el abandono escolar en una institución de educación superior tecnológica en México. In: Congresos CLABES (2017)
2. DGPPYEE-SEP. (s/f). Gob.Mx. Recuperado el 14 de marzo de 2022, de http://www.planeacion.sep.gob.mx/estadisticaeindicadores.aspx.
3. Reyes, J.: Modelo de decisión multicriterio para la selección de carrera universitaria. Investigación y Desarrollo, vol. 6, no. 1, pp. 25–32 (2013)
4. Hansen, J. C.: Interest inventories. Handbook of psychological assessment, pp. 169–190 (2019) doi: 10.1016/B978-0-12-802203-0.00006-7
5. Sinisterra, M. M., Cruz, J. P., Gantiva, C.: Teorías de la personalidad. Un análisis histórico del concepto y su medición. Psychologia. Avances de la disciplina, vol. 3, no. 2, pp. 81–107 (2009)
6. Jacques-Junior, J. C. J., Güçlütürk, Y., Pérez, M., Güçlü, U., Andujar, C., Baró, X., Jair-Escalante, H., Guyon, I., van-Gerven, M. A. J., van-Lier, R., Escalera, S.: First impressions: A survey on vision-based apparent personality trait analysis. IEEE Transactions on Affective Computing, vol. 13, no 1, pp. 75–95 (2019) doi: 10.1109/ TAFFC.2019.2930058
7. Reyes-Campos, I. M., Novoa-Cely, A. M.: Orientación vocacional. Universidad Central (2015)
8. Ramos-Monsivais, C. L., González, B. A.: Orientación vocacional, aprendizaje socio-emocional y sentido de vida en la educación superior. Dilemas contemporáneos: educación, política y valores, vol. 8, no. SPE5 (2020) doi: 10.46377/dilemas.v8i.2500

9. Alfaro-Barquero, A., Chinchilla-Brenes, S.: Diseño de un instrumento de preferencias vocacionales en administración, materiales y biotecnología. Revista Costarricense de Psicología, vol. 38, no. 2, pp. 99–124 (2019) doi: 10.22544/rcps.v38i02.01
10. Holland, J. L.: Making vocational choices: A theory of vocational personalities and work environments. Psychological Assessment Resources (1997)
11. Acosta-Amaya, M. M. La orientación profesional: Revisión del modelo RIASEC y la teoría social cognitiva del desarrollo de carrera. 2 cuadernos de Ciencias Sociales: Investigación en Psicología, vol. 47 (2018)
12. Rounds, J., Su, R., Lewis, P., Rivkin, D.: O* NET interest profiler short form psychometric characteristics: Summary. Raleigh, NC: National Center for O* NET Development, pp. 1–43 (2010)
13. Hansen, J. I. C.: Interest inventories. Handbook of psychological assessment, pp. 169–190 (2019) doi: 10.1016/B978-0-12-802203-0.00006-7
14. Roncero, M., Fornés, G., Belloch, A.: Hexaco: Una nueva aproximación a la evaluación de la personalidad en español. Revista argentina de clínica psicológica, vol. 22, no. 3, pp. 205–217 (2013)
15. Ashton, M. C., Lee, K.: Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. Personality and social psychology review, vol. 11, no.2, pp. 150–166 (2007) doi: 10.1177/1088868306294907
16. Lee, K., Ashton, M. C.: Psychometric properties of the HEXACO personality inventory. Multivariate behavioral research, vol. 39, no. 2, pp. 329–358 (2004) doi: 10.1207/s15327906mbr3902_8
17. Šverko, I., Babarović, T.: Integrating personality and career adaptability into vocational interest space. Journal of Vocational Behavior, vol. 94, pp. 89–103 (2016)
18. Barmanpek, U.: Relationship between personality traits and vocational interests: A multi-faceted study. Doctoral dissertation, University of Leicester (2019)
19. Tandera, T., Suhartono, D., Wongso, R., Prasetio, Y. L.: Personality prediction system from facebook users. Procedia computer science, vol. 116, pp. 604–611 (2017) doi: 10.1016/j.procs.2017.10.016
20. Biel, J. I., Teijeiro-Mosquera, L., Gatica-Perez, D.: Facetube: predicting personality from facial expressions of emotion in online conversational video. In: Proceedings of the 14th ACM international conference on Multimodal interaction, pp. 53–56 (2012)
21. Walker, M., Schönborn, S., Greifeneder, R., Vetter, T.: The basel face database: A validated set of photographs reflecting systematic differences in big two and big five personality dimensions. PloS one, vol. 13, no. 3, pp. e0193190 (2018)
22. Bátiz Beltrán, V. M. (2021). Sistema de evaluación de la personalidad y de las emociones en el proceso cognitivo. Tesis de Maestría en Ciencias de la Computación. Instituto Tecnológico de Culiacán.

ISSN 1870-4069

Gathering, Integration and Preprocessing of Data in Educational Data Mining: Towards a Study of Self-efficacy in Learning

Mayra Mendoza, Yasmín Hernández,
Javier Ortiz, Alicia Martínez, Hugo Estrada

Tecnológico Nacional de México, Cenidet,
Mexico

{m21ce017, yasmin.hp, javier.oh,
alicia.mr, hugo.ee}@cenidet.tecnm.mx

Abstract. Educational Data Mining emerged to take advantage of growing educational data and it has been extensively applied to improve Learning Environments, such as Intelligent Tutoring Systems. We are developing an intelligent tutoring system to teach mathematical logic, and we want to model self-efficacy in the modelling of students driven by data. We conducted a study to gather data from several sources. To have a precise learning model, we have conducted several visualization and preprocessing tasks to prepare data for machine learning algorithms, such as data integration, data exploratory analysis, data normalization, data standardization, resampling, management of outliers, discretization, and annotating. We built several versions of the dataset to prepare data for machine learning algorithms. The results of the preparation data process are presented.

Keywords: Data mining, educational data mining, intelligent tutoring systems, intelligent learning environments, preprocessing data, self-efficacy.

1 Introduction

Electronic devices in daily life generates cumulus of data every second. The analysis of this data allows us to obtain information about processes, people, relationships, behaviors, and about ourselves, which in turn allows us to make decisions and take actions. Data mining emerged to take advantage of growing data. Data mining seeks to discover patterns in large volumes of data, to extract information and to transform it into an understandable structure for later use [1].

The advancement of technology has made possible to store and process a huge amount of data, and it allows having successful data mining applications in different fields, for example, commerce, banking, and health. However, education is one of the fields where data mining has arisen the most interest and research. Education is one of the fields that has benefited the most from computers, therefore there are an incredible volume of data on the interaction of students with learning environments.

This is result of the increasing use of learning environments, such as intelligent tutoring systems (ITS), e-learning systems, educational games, learning management systems (LMS) and massive open online courses (MOOC), in addition to administrative computer-based systems. With these data, we could know the students, understand different aspects of the interaction of the students with the systems and comprehend the learning process itself.

Educational Data Mining (EDM) is an emerging discipline, interested in the development of methods to explore the exceptional data that comes from educational environments, and concerned in the use of these methods to understand students and the environments in which they learn [2]. The research on intelligent tutoring systems has taken advantage of EDM. These educational programs simulate the behavior of human tutors.

Namely, ITS teach students in the same way that a human tutor does [3]. To have a more precise and adaptive behavior, several elements have been integrated to ITS behaviors such as the recognition of emotions and personality, as well as the modeling of different cognitive states such as motivation, self-efficacy, and self-regulated learning. These elements allow a more personal and motivating interaction with ITS.

As known, some subjects are perceived as difficult; thus, students are apathetic, do not do homework, and do not study in their account, and consequently they have a low achievement in their studies. We want to promote motivation and self-efficacy in students, and to provide them with tools for self-regulated learning. These constructs have been identified as important players in learning [4].

With this aim, we are developing MateLog, an ITS to teach mathematical logic. Towards including self-efficacy in the student model, we conducted a study to gather data about performance of the students and about personal traits such as self-efficacy and procrastination. We are going to analyze this educational dataset to know the relationship of self-efficacy and learning. Before to apply machine learning algorithms, we conducted an exploratory analysis and preprocessing of data.

In this paper, we present the study we conducted to gather data and the results of the exploratory analysis and preprocessing of data. The paper is organized as following: Section 2 presents background and a brief review of literature on educational data mining; Section 3 describes the study we conducted to gather data; Section 4 depicts the tasks for visualization and preprocessing data. Finally, Section 5 outlines conclusions and future work.

2 Background and Related Work

Educational data can be used to improve the understanding of learning, of students, and to create a better, smarter, more interactive, engaging, and effective education. This requires advances in artificial intelligence and machine learning, human intelligence understanding and learning theories [5].

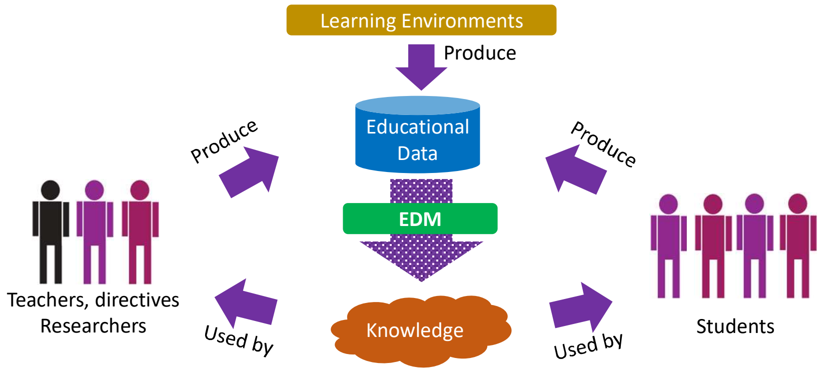


Fig. 1. Educational Data Mining knowledge discovery cycle [2].

Educational data mining is an emerging discipline that still has many pending solutions; but it has a potential to support the development of other fields related to education. EDM and, its technical basis, the machine learning techniques play an important role in augmenting and improving learning environments.

Machine learning is concerned with the ability of a system to acquire and integrate new knowledge through observations of users and with improving and extend itself by learning rather than by being programmed with knowledge [1]. These techniques organize existing knowledge and acquire new knowledge by intelligently recording and reasoning about data.

For example, observations of the previous behavior of students will be used to provide training examples that will form a model designed to predict future behavior. As in data mining, in EDM several computing paradigms and algorithms converge, such as decision trees, artificial neural networks, machine learning, Bayesian learning, logic programming, statistical algorithms, among others.

However, traditional mining algorithms need to consider the characteristics of the educational context to support instructional design and pedagogical decisions [2]. Educational data have meanings with multiple levels of hierarchy, which need to be determined by means of the properties of the data itself. Time, sequence, and context play an important role in the study of educational data [6].

EDM supports the development of research on many problems in education, since it not only allows to see the unique learning trajectories of individuals, but it also allows to build increasingly complex and sophisticated learning models [7].

The knowledge uncovered by EDM algorithms can be used not only to help teachers manage their classes, understand learning processes of their students, and reflect it in their own teaching methods, but also to support reflections of the student about the situation and give feedback to them [8]. Although one might think that there are only these two stakeholders in EDM, there are other groups of users, who see EDM from different points of view, according to their own objectives [2].

For example, education researchers, universities, course developers, training companies, school supervisors, school administrators, could also benefit from the knowledge generated by EDM [5]. Fig. 1 shows the interrelationships of educational environments, stakeholders and the EDM process.

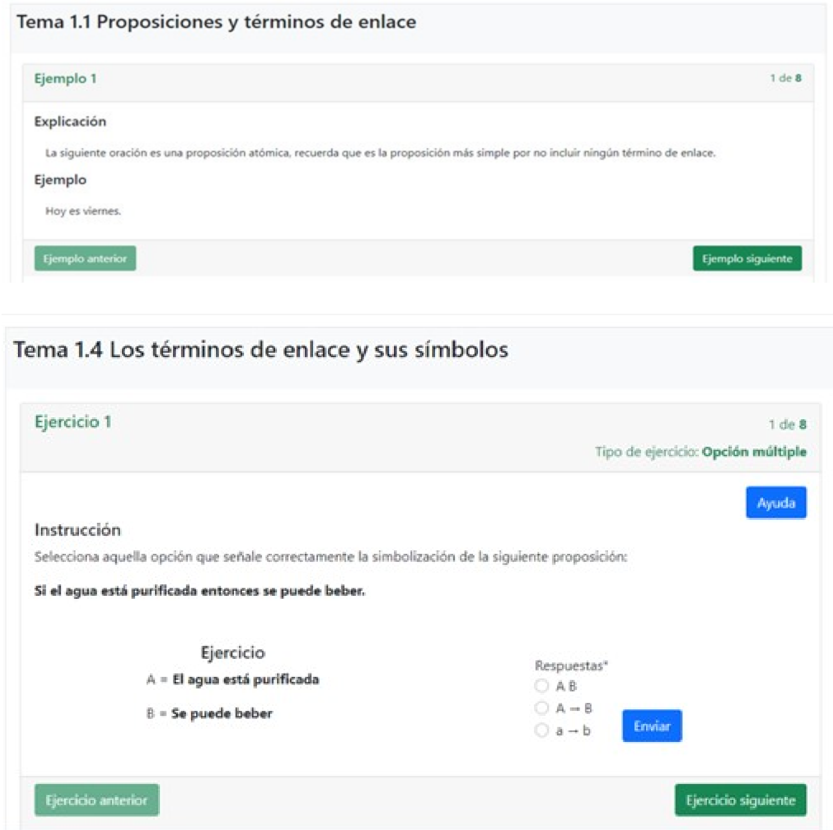


Fig. 2. Topics of mathematical logic included in MateLog, an example (left) and an exercise (right) are shown.

2.1 MateLog

MateLog is an intelligent tutoring system to learn mathematical logic for undergraduate and graduate students. MateLog consists of two main modules: the expert module and the student model. The expert module contains the knowledge about mathematical logic in the form of lessons, topics, examples, and exercises. The student model stores personal and academic information, and performance of the student in the course. In the Fig. 2 an example and an exercise are shown.

2.2 Self-Efficacy and Procrastination

Self-efficacy is a personal judgment of how well or poorly a person can cope with a given situation based on the skills they have and the circumstances they face. Self-efficacy affects every area of human endeavor. By determining the beliefs, a person holds regarding their power to affect situations, self-efficacy strongly influences both

Table 1. Extracts from scales for general self-efficacy, academic self-efficacy, and procrastination.

Scale	Items
General self-efficacy	I can find a way to get what I want, even if someone opposes me Puedo encontrar la forma de obtener lo que quiero, aunque alguien se me oponga
	I can solve difficult problems if I try hard enough Puedo resolver problemas difíciles si me esfuerzo lo suficiente
Academic self-efficacy	I consider myself capable enough to successfully face any academic task Me considero lo suficientemente capacitado para enfrentarme con éxito a cualquier tarea académica
	I think I have enough ability to understand a subject well and quickly Pienso que tengo bastante capacidad para comprender bien y con rapidez una materia
Procrastination	When I must do a task, I usually leave it until the last minute. Cuando tengo que hacer una tarea, normalmente la dejo para el último minuto
	I usually prepare in advance for exams Generalmente me preparo por adelantado para los exámenes

the power a person actually has to face challenges competently and the choices a person is most likely to make [9].

Also, self-efficacy has been recognized as a key trait of self-regulated learners [10]. Self-efficacy for self-regulated learning refers to the beliefs that individuals hold in their capabilities to think and behave in ways that are systematically oriented toward or associated with their learning goals.

Students with a robust sense of efficacy in their self-regulatory capabilities believe they can manage their time effectively, organize their work, minimize distractions, set goals for themselves, monitor their comprehension, ask for help when necessary, and maintain an effective work environment [11]. This action of avoiding, promising to do homework later, excusing or justifying delays, and avoiding guilt in the face of an academic task, refers to Academic Procrastination [12, 13].

In this situation, the student displays behaviors to voluntarily postpone activities that must be completed at a set time, either due to early family influence that has affected their self-esteem and tolerance to frustration; by the current choice of activities that guarantee an immediate achievement [13]; due to inadequate information processing [14] and of an irrational nature, or due to carrying out activities with more rewarding consequences in the short term than in the long term [13].

Evidence indicates that in the medium and long term, procrastination affects the academic life of people [15], since it is the first step towards other academic difficulties such as failures in the process of regulating academic behavior, Procrastination is related to less effective learning strategies which can affect the academic training and subsequently the professional performance of the person.

There are several instruments to know the self-efficacy and procrastination. A recognized scale for general self-efficacy is the proposed by Baessler y Schwarcer [16]. An extract of a Spanish scale for academic self-efficacy is proposed by Del

Table 2. Attributes in the dataset.

	Name	Description
1	Sex	Gender
2	Group	Group (A, B, C)
3	SP_Recoursing	Student is retaking the subject
4	SC_GeneralSelfEfficacy	Ten answers for each item of general self-efficacy scale
5	SC_AcademicSelfEfficacy	Ten answers for each item of academic self-efficacy scale
6	SC_Procrastination	Twelve answers for each item of procrastination scale
7	SP_Attendance	Attendances to class
8	SP_PretestGrade	Pretest exam grade
9	SP_Exercise1Grade	History of classical physics exercise grade
10	SP_Exercise2Grade	Unit conversion exercise grade
11	SP_Exercise3Grade	Introduction to vectors exercise grade
12	SP_Quiz1Grade	First term exam grade
13	SP_Unit1Grade	First term grade
14	SP_Quiz2Grade	Second term exam grade
15	SP_Unit2Grade	Second partial grade
16	SP_Exercise4Grade	Video exposition exercise grade
17	SP_Quiz3Grade	Third term exam grade
18	SP_FinalProject	Final project grade
19	SP_Unit3Grade	Third partial grade
20	SP_FinalGrade	Final grade in the subject.
21	ML_PreviousKnowledge	Previous knowledge about mathematical logic
22	ML_PreviousCourse	Student attended a previous course in mathematical logic
23	ML_CourseProgress	Progress in the mathematical logic course
24	ML_Topics	Topics completed
24	ML_Examples	Examples studied
28	ML_ExamplesSeconds	Time spent in examples
29	ML_Exercises	Exercises solved
30	ML_ExercisesSeconds	Time spent in exercises
31	ML_CorrectExercises	Correct exercises

Valle [17]. Dominguez [18] proposes an adaptation of the General and Academic Procrastination Scale proposed by Busko. Table 1 presents extracts of these scales, and its adaptation to Spanish, only two items are presented for each scale.

2.3 Related Work

The collection and preprocessing of data are an important part of the data mining process.

Educational data have unique characteristics that must be considered when preprocessing them. Research have been carried out to propose diverse methods for the treatment of educational data.

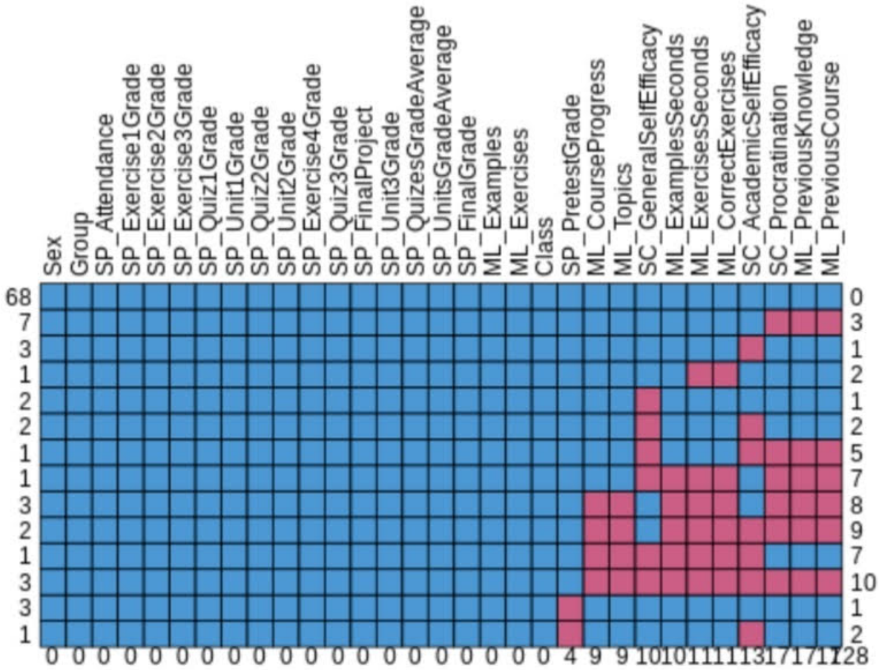


Fig. 3. Patterns of missing data in the dataset.

Romero, Ventura, and Romero [2] provides a survey of research related to the preprocessing of educational data. They state data pre-processing is one of the most important but less studied tasks in educational data mining research and shows the main tasks and issues in the pre-processing of educational data. They pay attention to types of data.

Romero, Ventura and García [3] expose the need to transform data from relational databases to transactional databases to have them in a format that can be read by data analysis tools and machine learning algorithms. To access this data stored in relational databases, it is necessary to use database queries, such as in SQL.

Subsequently, with the data obtained, a transactional summary table can be created. They also describe the process of collecting and preprocessing data from a course taught through the Moodle platform, which stores the data in a relational database (MySQL and PostgreSQL), to subsequently apply machine learning algorithms.

The stages they present are data selection, creation of summary tables, data discretization, and data transformation. They also present statistical data analysis tools. Chango and colleagues [5] describe the preprocessing and integration of data from different sources.

Tasks such as anonymization, normalization of attributes, discretization and data transformation are included. Hanna [6] describes a study about self-efficacy based on

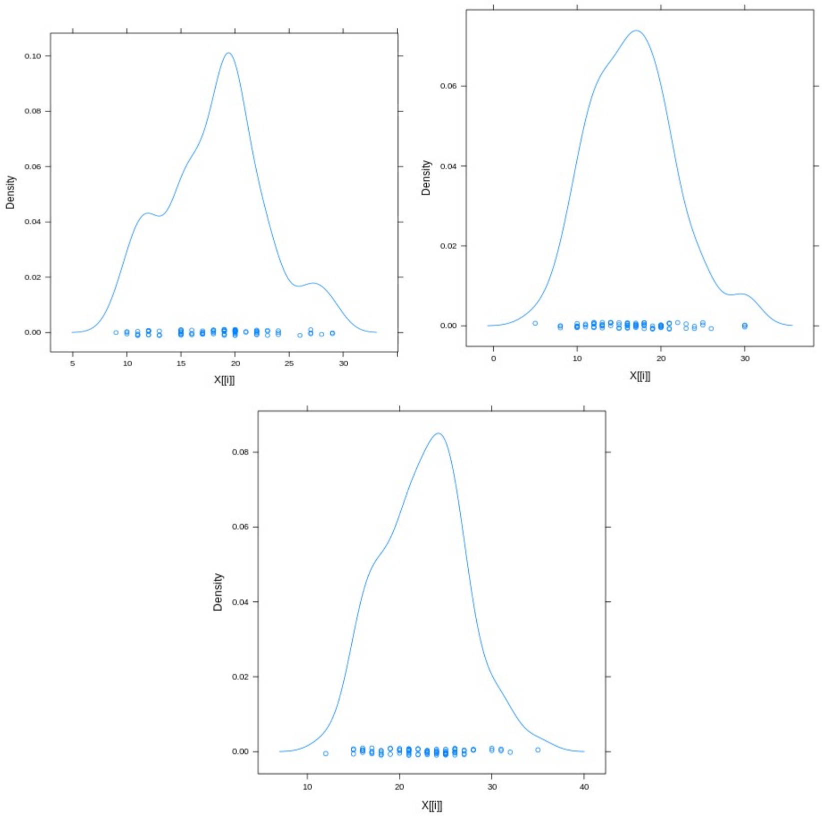


Fig. 4. Distribution of the scales of general self-efficacy (left), academic self-efficacy (center) and procrastination (right).

data mining to include this construct in the student model of an intelligent tutoring system. They state the feature and data selection is a key aspect of the educational data mining to understand the learning process.

3 Data Gathering

We conducted a study to gather data to understand the relationship between self-efficacy and learning. We analyze the academic performance of a group of undergraduate students group enrolled in a university course. Participants are 98 students, 25 female and 73 males, between 19 and 30 years old.

The study consists in three parts. Firstly, we obtained the data about the academic performance during the term, recorded by the professor. These data consist of attendances and grades in quizzes, exercises, and projects.

Then, we introduced MateLog to the students. They were asked to study mathematical logic with this ITS. The interaction of students with MateLog was

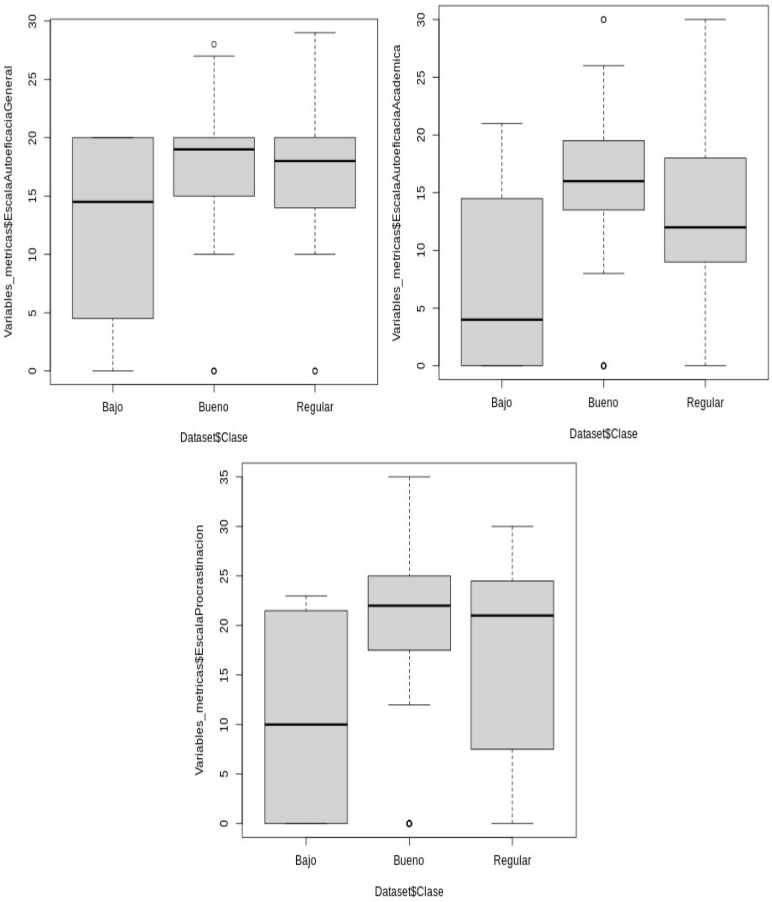


Fig. 5. Identification of outliers in the scales of general self-efficacy (left), academic self-efficacy (center) and procrastination (right).

recorded, these data consist of studied lessons and examples, correct and incorrect exercises. In the third part students filled in three scales about self-efficacy, academic self-efficacy, and procrastination. The scales are presented in section 2 (Table 1). We obtained a dataset with 98 observations. The description of attributes is presented in Table 2.

4 Data Visualization and Preprocessing

After the integration of the three sources of data, we got a dataset with 29 attributes and 98 examples.

The data were analyzed for identification of missing values. We find missing values in age and sex; these were filled looking in other databases. The data were analyzed to find missing data patterns. The result is shown in Fig. 3. There are 69 complete examples and 29 examples with missing data. There are 14 patterns with 1, 2, 3, 4, 6, 7, 8, 9 and 10 missing data.

Irrelevant attributes were eliminated, for example, the school grade since all participants were in the same grade. We build diverse versions of the dataset. In the case of scales answers, we have a version with all the individual answers to each item and other version with the result of the scale.

We analyze data to whether know data distribution. In Fig. 4 the distribution of general self-efficacy, academic self-efficacy, and general procrastination. As can be seen, they have a normal distribution. There are not classes in the dataset, therefore we annotated every example with based on the course grade (*low*, *regular*, *good*), the general self-efficacy (*low*, *medium*, *high*), academic self-efficacy (*low*, *medium*, *high*), and procrastination (*low*, *medium*, *high*).

In this way we have four versions of the dataset with different objective variables. To identify outliers, we used the box and whiskers chart. In Fig. 5 the comparison of scales variables versus performance is presented. As can be observed there are few outliers. After annotating, we have four unbalanced versions of the dataset, for example, for *performance* we have 8 *low* examples, 13 *regular* examples, and 57 *good* examples. Therefore, we must balance classes before to apply the classification algorithms.

5 Conclusions and Future Work

Educational data can be used to improve our understanding of learning and students, and which in turn allows having better educational technologies and a better education. These objectives require further advances in artificial intelligence and in human learning theories. Educational data mining is an emerging discipline that can be useful towards these aims due its potential to support the development of fields related to education.

In this paper, we present a study to gather data to model self-efficacy, to improve learning. The knowledge will be applied in the designing of MateLog, an ITS to teach mathematical logic. The ITS is building by means of applying several educational data mining techniques. The future steps include applying classification with machine learning algorithms to build prediction models, and we want to do clustering to find patterns and relationships in these data.

Acknowledgments. This research was supported by projects TecNM 13869.22-P and TecNM 14923.22-P.

References

1. Witten, I. H., Frank, E., Hall, M. A., Pal, C. J.: Practical machine learning tools and techniques. Data Mining, vol. 2, no. 4 (2005)
2. Romero, C., Ventura, S.: Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, pp. e1355 (2020) doi: 10.1002/widm.1355

3. Woolf, B. P.: Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning (2010)
4. Lavasani, M. G., Mirhosseini, F. S., Hejazi, E., Davoodi, M.: The effect of self-regulation learning strategies training on the academic motivation and self-efficacy. Procedia-Social and Behavioral Sciences, vol. 29, pp. 627–632 (2011) doi: 10.1016/j.sbspro.2011.11.285
5. Koedinger, K. R., Brunskill, E., Baker, R. S., McLaughlin, E. A., Stamper, J.: New potentials for data-driven intelligent tutoring system development and optimization. AI Magazine, vol. 34, no. 3, pp. 27–41 (2013) doi: 10.1609/aimag.v34i3.2484
6. International Educational Data Mining Society: Educational Data Mining. http://www.educationaldatamining.org/
7. Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., Warschauer, M.: Mining big data in education: Affordances and challenges. Review of Research in Education, vol. 44, no. 1, pp. 130–160 (2020) doi: 10.3102/0091732X20903304
8. Merceron, A., Blikstein, P., Siemens, G.: Learning analytics: From big data to meaningful data. Journal of Learning Analytics, vol. 2, no. 3, pp. 4–8 (2016) doi: 10.18608/jla.2015.23.2
9. Bandura, A.: Guide for constructing self-efficacy scales. Self-efficacy beliefs of adolescents, vol. 5, no. 1, pp. 307–337 (2006) doi: 10.1017/CBO9781107415324.004
10. Panadero, E.: A review of self-regulated learning: Six models and four directions for research. Frontiers in psychology, vol. 8, p. 422 (2017) doi: 10.3389/fpsyg.2017.00422
11. Usher, E. L.: Self-efficacy for self-regulated learning. Encyclopedia of the Sciences of Learning (2012) doi: https: 10.1007/978-1-4419-1428-6_835
12. Onwuegbuzie, A. J: Academic procrastination and statistics anxiety. Assessment & Evaluation in Higher Education, vol. 29, no. 1, pp. 3–19 (2004) doi: 10.1080/0260293042000160384
13. Quant, D. M., Sánchez, A.: Procrastinación, procrastinación académica: Concepto e implicaciones. Revista vanguardia psicológica clínica teórica y práctica, vol. 3, no. 1, pp. 45–59 (2012)
14. Stainton, M., Lay, C. H., Flett, G. L.: Trait procrastinators and behavior/trait-specific cognitions. Journal of Social Behavior and Personality, vol. 15, no. 5, pp. 297–312 (2015)
15. Clariana-Muntada, M., i Pros, R. C., Badia-Martín, M. D. M., Gotzens-Busquets, C.: La influencia del género en variables de la personalidad que condicionan el aprendizaje: inteligencia emocional y procrastinación académica. Revista electrónica interuniversitaria de formación del profesorado, vol. 14, no. 3, pp. 87–96 (2011)
16. Bueno-Pacheco, A., Lima-Castro, S., Peña-Contreras, E., Cedillo-Quizhpe, C., Aguilar-Sizer, M.: Adaptación al español de la escala de autoeficacia general para su uso en el contexto ecuatoriano. Revista Iberoamericana de Diagnóstico y Evaluación-e Avaliação Psicológica, vol. 3, no. 48 pp. 5–17 (2018)
17. Del Valle, M., Díaz, A., Pérez, M. V., Vergara, J.: Análisis factorial confirmatorio escala autoeficacia percibida en situaciones académicas (EAPESA) en universitarios chilenos. Revista Iberoamericana de Diagnóstico y Evaluación-e Avaliação Psicológica, vol. 4, no. 41, pp. 97–106 (2018)
18. Dominguez, S., Villegas, G., Centeno, S.: Academic procrastination: validation of a scale in a sample of students from a private university. vol. 20, no. 2, pp. 293–304 (2014)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación