Comunicar, n. 71, v. XXX, 2022 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478
www.comunicarjournal.com

Detection of traits in students with suicidal tendencies on Internet applying Web Mining

Detección de rasgos en estudiantes con tendencia suicida en Internet aplicando Minería Web

Dr. Iván Castillo-Zúñiga. Professor, Systems and Computing Department, National Technological Institute of Mexico, Llano Aguascalientes Campus (Mexico) (ivan.cz@llano.tecnm.mx) (https://orcid.org/0000-0001-8017-5908)
Dr. Francisco-Javier Luna-Rosas. Professor, Systems and Computing Department, National Technological Institute of Mexico, Aguascalientes Campus (Mexico) (fcoluna2000@yahoo.com.mx) (https://orcid.org/0000-0001-6821-4046)
Dr. Jaime-Iván López-Veyna. Professor, Systems and Computing Department, National Technological Institute of Mexico, Zacatecas Campus (Mexico) (ivanlopezveyna@zacatecas.tecnm.mx) (https://orcid.org/0000-0002-9225-3202)

ABSTRACT

This article presents an Internet data analysis model based on Web Mining with the aim to find knowledge about large amounts of data in cyberspace. To test the proposed method, suicide web pages were analyzed as a study case to identify and detect traits in students with suicidal tendencies. The procedure considers a Web Scraper to locate and download information from the Internet, as well as Natural Language Processing techniques to retrieve the words. To explore the information, a dataset based on Dynamic Tables and Semantic Ontologies was constructed, specifying the predictive variables in young people with suicidal inclination. Finally, to evaluate the efficiency of the model, Machine Learning and Deep Learning algorithms were used. It should be noticed that the procedures for the construction of the dataset (using Genetic Algorithms) and obtaining the knowledge (using Parallel Computing and Acceleration with GPU) were optimized. The results reveal an accuracy of 96.28% on the detection of characteristics in adolescents with suicidal tendencies, reaching the best result through a Recurrent Neural Network with 98% accuracy. It is inferred that the model is viable to establish bases on mechanisms of action and prevention of suicidal behaviors, which can be implemented in educational institutions or different social actors.

RESUMEN

Este artículo presenta un modelo de análisis de datos en Internet basado en Minería Web con el objetivo de encontrar conocimiento sobre grandes cantidades de datos en el ciberespacio. A fin de probar el método propuesto, se analizaron páginas web sobre el suicidio como caso de estudio con la intención de identificar y detectar rasgos en estudiantes con tendencias suicidas. El procedimiento considera un Web Scraper para localizar y descargar información de Internet, así como técnicas de Procesamiento de Lenguaje Natural para la recuperación de los vocablos. Con el propósito de explorar la información, se construyó un conjunto de datos basado en Tablas Dinámicas y Ontologías Semánticas, especificando las variables predictivas en jóvenes con inclinación suicida. Por último, para evaluar la eficiencia del modelo se utilizaron algoritmos de Aprendizaje de Máquina y Aprendizaje Profundo. Cabe destacar que se optimizaron los procedimientos para la construcción del dataset (utilizando Algoritmos Genéticos) y obtención del conocimiento empleando Cómputo Paralelo y Aceleración con Unidades de Procesamientos de Gráfico (GPU). Los resultados revelan una precisión del 96,28% sobre la detección de las características en adolescentes con tendencia suicida, alcanzando el mejor resultado a través de una Red Neuronal Recurrente con un 98% de precisión. De donde se infiere que el modelo es viable para establecer bases sobre mecanismos de actuación y prevención de comportamientos suicidas, que pueden ser implementados en instituciones educativas o distintos actores de la sociedad.

KEYWORDS | PALABRAS CLAVE

Suicidal behavior, cybersuicide, web mining, machine learning, deep learning, recurrent neural networks.
Conducta suicida, cibersuicidio, minería web, aprendizaje de máquina, aprendizaje profundo, redes neuronales recurrentes.

Received: 2021-09-27 | Reviewed: 2021-10-24 | Accepted: 2021-12-03 | OnlineFirst: 2022-01-15 | Published: 2022-04-01
DOI https://doi.org/10.3916/C71-2022-08 | Pages: 101-112

1. Introduction

The benefits that new technologies have brought to our lives cannot be denied, however, they have also generated problems that were previously unknown, one of them is known as cybersuicide. According to López-Martínez (2020), cybersuicide is a phenomenon that refers to the influence of the information that circulates on the Internet and incites a person to commit a suicidal act. Suicide is the second leading cause of death in the world population aged 10 to 24, which represents 100,000 dead adolescents per year. Suicide is a complex behavior that is constructed over time and which depends on multiple biological, family, social, and educational factors, among others.

The internet and social media have rapidly and substantially transformed the way adolescents and people communicate and access suicide-related information. This large number of websites and volumes of information is commonly referred to as Big Data. Molina and Restrepo (2018), mention that users not only search web pages that provide information related to methods and ways to commit suicide, but also seek help, support, and guidance in the face of the suffering they experience derived from suicidal thoughts, sadness, loneliness, and anxiety. These reasons require early detection and immediate response to this public health problem, where information is essential for its analysis and a useful means to educate and prevent suicidal behavior. Educational institutions play a very important role in promoting healthy lifestyles and providing early support to at-risk youth.

From different approaches, many efforts have been made to study the large volumes of information that are generated on the Internet by applying automatic techniques to analyze and predict events that affect society, such as suicide. With this perspective, there are programs, developments, algorithms, and evolving processes that continue to be studied by researchers. According to Nalini and Sheela (2014), understanding the relationship between analytical skills and the characteristics of a criminal event can help researchers use these techniques more efficiently to identify trends and patterns, address different issues, and even predict crime. For these reasons, this study proposes a new approach to Web Mining processes, integrating Machine Learning and Deep Learning techniques using GPU-accelerated and Parallel Computing to process large volumes of data in order to identify and detect traits in adolescents with suicidal tendencies.

Research such as those of Bonami et al. (2020), Denia (2020), Kim and Chung (2019), Anggraini et al. (2018), and Roy et al. (2017), have aimed to treat large volumes of data, analyzing information with Artificial Intelligence techniques and Natural Language Processing (NLP), which are summarized in Table 1.

Table 1. Characteristics of studies related to Big Data and Artificial Intelligence

Related Jobs	Big Data	Artificial Intelligence					Research Focus
				Computation			
	Data Source	Data Mining	NLP	Sequential	Parallel	GPU Acceleration	
Bonami, Piazentini and Dala-Possa (2020)	Sensor	Yes	No	Yes	No	No	Educational
Denia (2020)	Social Network	No	Yes	Yes	No	No	Social impact of scientific speeches on Twitter
Kim and Chung (2019)	Websites	Yes	Yes	Yes	No	No	Health
Anggraini, Sucipto and Indriati (2018)	Social Network	Yes	Yes	Yes	No	No	Cyberbullying
Roy et al. (2017)	Websites	Yes	Yes	Yes	No	No	Cybercrime Intrusion detection
Current investigation	Websites	Yes	Yes	Yes	Yes	Yes	Suicide

Some of these research projects present knowledge-seeking methods aimed at data mining (Bonami et al., 2020; Kim & Chung, 2019; Anggraini et al., 2018; Roy et al., 2017), other studies have focused on information retrieval using Natural Language Processing techniques (Denia, 2020; Kim & Chung, 2019; Anggraini et al., 2018). One of the complexities of this research is that the information is unstructured and located in large amounts of data dispersed on websites with different security and communication protocols, in addition, the texts have spelling errors and abbreviations. To begin the process of detecting the characteristics of students with suicidal tendencies, it is necessary to transform the information into structured data, continuing with the analysis using Machine Learning techniques to obtain patterns that become knowledge and added value.

This proposal is based on the combination of Big Data and Artificial Intelligence techniques with a genetic strategy using Parallel Computing. Also, we adapted some processes of the Semantic Web with procedures based on Semantic Ontologies, Vocabularies, and Dynamic Tables, integrating NLP methods for the processing of data, such as cleaning, separation of information from the computational code and removing meaningless words, tokenization (word separation), synonyms, stemming (word root) and term frequency, in order to generate valuable and value-added information, obtained from the analysis of a large number of pages with suicidal content downloaded from the web. It should be noted that, with these processes, we build a Semantic Ontology that classifies information through concepts validated with Machine Learning and Deep Learning techniques, applying GPU-accelerated Parallel Computing.

The main purpose of our research is to explore the benefits that can be obtained from the information that circulates on the web, seeking to discover patterns, unknown correlations, additional information that can be very useful to support the decision-making process in problem-solving. The study objectives contemplate the detection of traits in students with suicidal tendencies, as well as, carrying out a technical practice of recovering web pages, pre-processing, analyzing and classifying data. Finally, we create a categorization of the linguistic corpus of suicide. The main contribution of the article focuses on a study model on data related to cyberbullying problems on the Internet, specifically cybersuicide. Within the framework of the study, the limit is specifically the detection of cybersuicide words on the Internet through semantic ontologies using predictive analysis on a computer. It should be noted that the descriptive analysis of the words is not considered within the scope of the research.

1.1. Cybersuicide and its impact on students

López-Martínez (2020) mentions that suicide is a multifactorial phenomenon, and that, at present, with the emergence of the new Information and Communication Technologies (ICTs), it constitutes a new scenario and with it, a new problem for the prevention of suicidal behavior. In this context, a new concept is created, cybersuicide, which refers to the action of taking one's own life, motivated by the influence of prosuicidal pages, forums, and chat rooms on the Internet, among other variables. Within the same framework, Olivares (2019) points out that cybersuicide refers to the influence of information circulating on the internet, as well as the incitement that exists in those media to exercise it. In addition to the situation, Moreno & Blanco (2012) indicate that it is a phenomenon that spreads rapidly throughout the planet, increasing suicide cases year after year.

Durkheim (2008) indicates that suicide is any case of death that results directly or indirectly, from any act, positive or negative, performed by the victim itself, knowing that the victim should produce this result. From another perspective, Berengueras (2018) mentions that suicide is when individuals violate the laws of their own nature by making the decision to end their own life, in a deeply expressive and essential act of being understood and heard in their previous manifestations. For his part, Marchiori (2015) establishes that the most frequent instruments to commit suicide are firearms, knives, ropes, wires, fabrics for suffocation, medicaments, drugs, jumping from bridges, buildings, to train tracks or passing cars, asphyxia by immersion, poisons, fuels (gas, coal, kerosene, naphtha), among others.

Research such as Arevalos (2020), SeGob (2021), Luna and Dávila (2018), Sánchez-García et al. (2018), Blanco (2019), Healy (2019), and the World Health Organization (WHO) (2019) reveal that suicide is a global problem, which can occur at any age. However, it is preventable through timely interventions, where the education sector plays an important role, since it is reported that in 2016 it was the second leading cause of death for young people between the ages of 15 and 19. In Argentina, testimonies from young people enrolled in peripheral urban junior high schools related to suicidal behavior were obtained, in which the following features stand out: belittling, humiliation, bullying, neglect, obesity, family abuse and lack of support, parental divorce, rape, abuse, complexes with their body or image, dating problems, drugs, alcohol, social indifference and visualization of few opportunities in the future (Arevalos, 2020).

In Mexico, the Secretary of the Interior reports the impact of the Covid-19 pandemic on girls and boys concerning suicide, wherein 2021 a record figure of 1,150 suicides were registered, with an increase of 37% in children between 10 and 14 and of 12% in adolescent women between 15 and 19 (SeGob,

2021), along the same lines it is identified that aggression, family violence, educational delay, alcohol or tobacco consumption are risk factors associated with the suicide attempt in adolescents, especially in younger women (Luna & Dávila, 2018). In Spain, it is revealed that 7.7% of adolescents between 12 and 19 years old showed difficulties in emotional adjustment and greater suicidal ideation derived from behaviors related to bullying, and tobacco and cannabis use (Sánchez-García et al., 2018). Along the same line, the records of the National Institute of Statistics reported 3,679 deaths by suicide in 2017 with 74% for men and 26% for women, with an average of 10 suicides per day (Blanco, 2019). In the United States, suicide claimed the lives of 5,016 men and 1,225 women between the ages of 15 and 24 in 2017, with a youth suicide rate of 14.6 per 100,000 inhabitants (Healy, 2019). For its part, the WHO (2019) reveals that in the world about 800,000 people take their own lives and many more try to do it, where 79% of suicides took place in low, and middle-income countries. Ingestion of pesticides, hanging, and firearms are the most common suicide methods.

1.2. Suicidal behavior in adolescents

Beaven-Ciapara et al. (2018), describe a study on the risk factors associated with suicidal behavior in young people aged 13 to 18 in the community of Guaymas, Sonora, Mexico. The results indicate that the psychosocial factor that is being presented is dysfunctional families, causing depression and low self-esteem, where the model used relates suicidal behavior with this finding in 37%, occurring more frequently in females. The study was carried out on 120 middle and high school students (41% males and 59% females), and its statistical analysis was done using SPSS software version 21.0. In turn, Mosquera (2016), presents a non-systematic review of the literature on child suicidal behavior, revealing that among the most prominent risk factors are: being male, having previous suicide attempts, social exclusion, emotional conflict. In addition, high comorbidity is observed with depressive disorders, bipolar disorder, and schizophrenia. Among the most effective treatments are dialectical-behavioral and cognitive-behavioral therapy. On the other hand, Carballo-Belloso and Gómez-Peñalver (2017) found a strong causal association between individual vulnerability factors and stressors, such as bullying experiences in childhood, and the subsequent development of thoughts and/or self-inflicted-injury behaviors, highlighting the importance of an adequate detection of this potentially modifiable risk factor.

1.3. Suicide analysis with Artificial Intelligence

Table 2 shows a comparison of related works for the analysis and prediction of suicide with Machine Learning techniques, in which Ramírez-López et al. (2021) seek to predict possible cases of suicide in the city of Aguascalientes, Mexico, using an SQL Server database of the 911 emergency service that records suicides. In their tests, they implement a geospatial analysis with Weka's EBK (Empirical Bayesian Kriging) method to locate the places where suicides have taken place. The results reveal a 99.22% prediction with a Bayesian classifier in MatLab identifying graphically the probability areas where suicide occurs and where it does not. Gen-Min et al. (2020) research how to predict suicidal ideas in army personnel since they have greater psychological stress and are at a higher risk of suicide attempts compared to the general population. The analysis uses Machine Learning techniques that include Logistic Regression, Decision Trees, Random Forests, Regression Trees, Vector Support Machines, and Multilayer Perceptron, considering five psychopathological domains (BSRS-5), anxiety, depression, hostility, interpersonal sensitivity, and insomnia. Their results exceed 98% accuracy in the classification using a questionnaire-based dataset of 3546 people. For their part, Pérez-Martínez et al. (2020) describe a Twitter analysis in which school harassment, rape, and suicide-related to the Netflix series "13 Reasons Why" are discussed in several countries, retrieving 154,470 tweets for exploration.

The results reveal that 51% of tweets were about suicide, 24% about bullying, and 23% about rape, where the United States and Spain were the countries with the greatest participation. The hashtags #13ReasonsWhy and #PorTreceRazones (in Spanish) were used to recover tweets, eliminate duplicate tweets, and form 3 samples with words in Spanish, English, French, Portuguese and Italian. The projections and calculations were made with statistical analysis of the SPSS software. Chiroma et al. (2018), identify suicide text on Twitter with Decision Trees, Bayes, Random Forests, and Vector Support

Machines, obtaining an accuracy between 34.6% and 77.8%, achieving better results with the first. On the other hand, Du et al. (2018) extract psychiatric stressors from suicide-related Twitter data using a Deep Learning approach and transfer learning strategy, where an accuracy of 78% is obtained with convulsion Neural Networks and 67.94% with Recurrent Neural Networks. Finally, Hermosillo-De la Torre et al. (2015) show the relationship of depressive symptoms, hopelessness, and psychological resources on suicide attempts in a sample of 96 adolescents in Aguascalientes, Mexico. Using SPSS they applied descriptive statistics for the analysis of proportions and estimation of population parameters, and nonparametric statistics for the comparisons of study groups. Subsequently, they implemented Spearman's Rho statistics to find out how variables were associated, and linear regression to observe the relationships between them. The results show that the development of the capacity to properly manage sadness is one of the factors to be considered as a suicide prevention measure to be encouraged and developed in adolescents.

Table 2. Comparison of data analysis techniques for suicide detection

Related studies	Data source	Data Analysis Techniques						
		Statistics	Machine Learning	Deep Learning	Data Pre-processing			Software Used
					Web Scraper	NLP	Semantic Ontology	
Ramírez-López et al. (2021)	SQL Server	No	Yes	No	No	Yes	No	Weka y MatLab
Gen-Min et al. (2020)	Questionnaire	Yes	Yes	No	No	No	No	Non-specific
Perez-Martínez et al. (2020)	Twitter	Yes	No	No	No	No	No	SPSS
Chiroma et al. (2018)	Twitter	No	Yes	No	No	Yes	No	Non-specific
Du et al. (2018)	Twitter	No	Yes	Yes	No	Yes	No	Non-specific
Hermosillo-De la Torre et al. (2015)	Psychological resource scales	Yes	No	No	No	No	No	SPSS
Present research	Websites	Yes	Yes	Yes	Yes	Yes	Yes	Java and Python

2. Materials and methods

2.1. Methodology to help to detect traits in students with a suicidal tendency on websites

The methodology applied in this study begins by obtaining data from cyberspace and ends with the detection of the traits of students with suicidal tendencies. The procedure consisted of three stages that integrated Big Data Analytics, Natural Language Processing, Semantic Web, and Artificial Intelligence, which are described below.

2.1.1. Location and download of suicide web pages (stage 1)

For locating and downloading websites with suicidal content, an open-source crawler was developed with the JSOUP libraries using the Java programming language, followed by scraping techniques to get the name of the website, the internet address (URL), and to make a copy of the file on the computer hard drive. The information is stored in databases to avoid duplication, along with control fields that allow the pre-processing of documents.

2.1.2. Construction of datasets for tests (stage 2)

One of the most crucial and complex parts for the detection of traits on students with suicidal tendencies, is the construction of the dataset since it is the main part where the classification and prediction tests are carried out using Artificial Intelligence techniques. To represent different sets of suicidal characteristics in adolescents, we used an approach based on semantic ontologies that allow us to associate concepts through object-oriented techniques (classes-objects-attributes), facilitating the grouping of different suicide conditions such as signs of suicide, ways to carry it out, types of suicide, risk factors, prevention, influences, and synonyms. Where the class symbolizes the main theme as suicide, the objects are the subtopics as signs of suicide, and the attributes are the characteristics, such as euphoria, distress, sleep, farewell, and isolation, among others.

The dataset is the result of the transformation of unstructured data to structured data, this part begins with the elimination of the computational code of the text, followed by the separation of words (tokenization), suppressing the words without meaning (stop word), such as prepositions, pronouns, articles, adverbs, conjunctions, and some verbs. Finally, the terms that will be part of the linguistic corpus of suicide were stored in a database. To indicate the importance of each suicide trait and increase the analytical accuracy, the techniques of Term Frequency (TF) and Document Inverse Frequency (IDF) were used, where the term was replaced with the root word, obtained with the Porter algorithm (2006) through the technique of lemmatization (stemming), where for example we considered for the search of the term suicide, the words suicide, suicides, suicidal and commit suicide, generating a more accurate result.

Subsequently, dynamic tables were constructed in MySQL using the characteristics defined in the Semantic Ontology as metadata linking them to the linguistic corpus to generate the dataset. It is important to mention that the dataset construction process was optimized with Parallel Computing through a Genetic Strategy to equally distribute the transformation of the web pages establishing a cluster with the processor cores simulating a chromosome and its genes. The process of the Genetic Algorithm evolves until it reaches the optimum distribution of web pages and meets the completion criteria with an adaptation function based on the mean. The evolution of the population is based on elitism, selection by tournament, point crossing, and mutation with random replacement. Finally, the target binary variable is established, with the name "correct" and the values "yes/no" that determine the response to be generated by the Artificial Intelligence algorithm from the predictor variables defined in the Semantic Ontology. To carry out this experiment, the variable to predict is determined with the value "yes", when the frequency of groups (signs of suicide, influences, types of suicide, synonyms, ways of carrying it out, prevention, and risk factors) is greater than zero, and "no" otherwise.

2.1.3. Artificial Intelligence applied to the detection of traits of students with a suicidal tendency (stage 3)

To evaluate the dataset, Artificial Intelligence techniques were selected: Random Forests, Neural Network, Decision Tree, and Logistic Regression, because they present greater similarities with related works (Gen-Min et al., 2020, Chiroma et al. 2018). In the construction of the algorithms, Python programming language was used, employing Machine Learning techniques with Sequential Computation based on the Sklearn API, Pandas, and NumPy.

Within the same testing framework, we considered, to optimize the above strategies using GPU-Accelerated Parallel Computing, applying Apple's scalable technology that allows the development of custom models through the Turicreate API with the object SFrame, which can mutate and scale to Big Data, considering the algorithms Random Forests, Logistic Regression, and Decision Tree. Finally, Deep Learning tests were performed using Recurrent Neural Networks (RNN), where the greater the number of layers and neurons, the greater the depth of the network and its learning capacity. The proposed RNN model integrates a 4-layer sequence (input layer, recurrent hidden layer with Long Short-Term Memory (LSTM), hidden layer, and output layer) that are densely connected, where all the neurons in one layer are connected to all the neurons in the next layer.

Among the optimization algorithms, Adam's algorithm, the SGD Stochastic Gradient Descent Method, and the RMSProp incremental learning technique were considered. Adam's algorithm combines the advantages of the AdaGrad and RMSProp algorithms, which calculates the learning rate of adaptive parameters based on the mean value of the first moment and makes full use of the mean value of the second moment of the gradient based on the non-centered variance. The SGD method maintains a single learning rate to update all weights throughout the training. Finally, the RMSProp technique considers a different training factor for each dimension, where the scaling of the training factor is performed by dividing it by the mean of the exponential decline of the square of the gradient.

The RNN algorithm proposed is based on the synchronous data parallelism of Keras and TensorFlow, where each layer is represented by a tensor with global information (called a global lot) and is divided into sublots according to the number of GPUs (called local lots), where the gradient calculations are performed

considering the loss of the model. Subsequently, the updates originated in the local gradients are merged with the rest of the replicates, thus remaining synchronized with the process.

2.2. Materials

The operating systems implemented in the different tests include MacOS Big Sur, Linux Ubuntu 20.04, and Windows 10. The software tools integrate Java programming language, 8.0 edition, and the MySQL database manager, 5.7 edition, used for the localization and download of web pages (stage 1). Also, for the construction of the dataset through the Genetic Algorithm, NLP techniques, vocabulary construction, and the design of semantic ontologies based on objects (stage 2). For the Artificial Intelligence processes, Python programming language, 3.8 edition, was used, with the Sklearn, Pandas, Numpy, Turicreate, Keras, and TensorFlow libraries, with Parallel Computation and GPU acceleration (stage 3). Finally, the computer equipment used in this study was a MacBook Pro with 2.4 GHz Intel Core i9 processor, 16 GB DDR4 memory, 500 GB Flash hard drive, Intel UHD Graphics 630 GPU, and Radeon Pro 560X GPU, connected to a 100MB internet service for locating and downloading web pages.

3. Tests and results

3.1. Testing procedure

The test procedure developed allows the analysis of small and large datasets, where thousands to millions of records can be processed according to the procedure described below:

Stage 1:

- A set of web pages related to suicide is located and a copy is made on the hard disk for analysis.

Stage 2:

- A Semantic Ontology is established on the domain related to traits of students with a suicidal tendency extracting a set of words from books and processed using the classes-objects-attributes technique.
- A set of meaningless words is selected based on a Google SEO standard (Landaeta, 2014).
- A linguistic corpus of suicide is generated using NLP and semantic web techniques, optimized with Parallel Computing through Genetic Algorithms to balance the load.
- The cybersuicide dataset is constructed for the tests linking Semantic Ontology with the linguistic corpus, and the binary target variable is established.

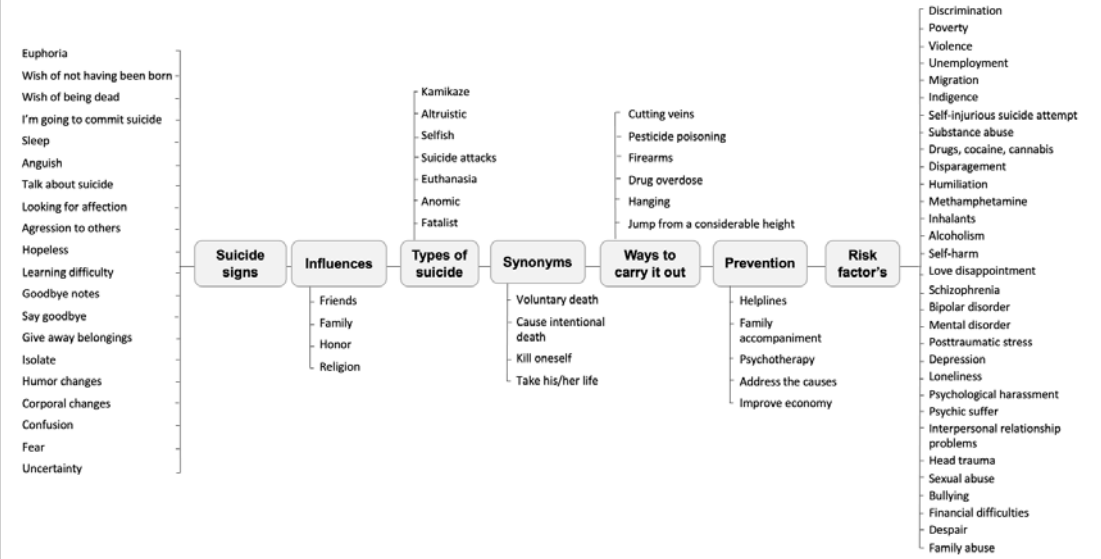
Stage 3:

- For the construction of programs with Sequential Computation, the Sklearn library was used, specifically with the algorithms: Random Forests (RandomForestClassifier), Neural Network (MLPClassifier), Decision Trees (DecisionTreeClassifier) and Logistic Regression (LogisticRegression).
- For the optimization of the programs with Parallel Computation and acceleration with GPU, the algorithms based on the Turicreate Library were used, specifically: Random Forests (tc.random_forest_classifier), Decision Trees (tc.decision_tree_classifier) and Logistic Regression (tc.logistic_classifier).
- Finally, for the construction of Recurrent Neural Networks with Deep Learning, an algorithm that connects several layers was considered. The Input Layer (inputLayer) receiving 97 suicidal traits, the recurrent occult layer with Short-Term Memory (RecurrentLayer LSTM) defined with 97 neurons, which consider possible combinations of suicidal traits, where the parameters obtained are the result of linear and non-linear transformations by calculating the current prediction considering the previous result of weights and biases, before using the activation function (using what has been learned), which allows having a short-term memory, the Hidden Layer dense (HiddenLayer) with 48 neurons, reducing the calculations of the transformations and filtering the optimal results to interpret the output of the hidden LSTM layer, finally the Output Layer (outputLayer_Sigmoid) of one neuron compatible with binary predictions, which returns the final prediction.

3.2. Analysis and results

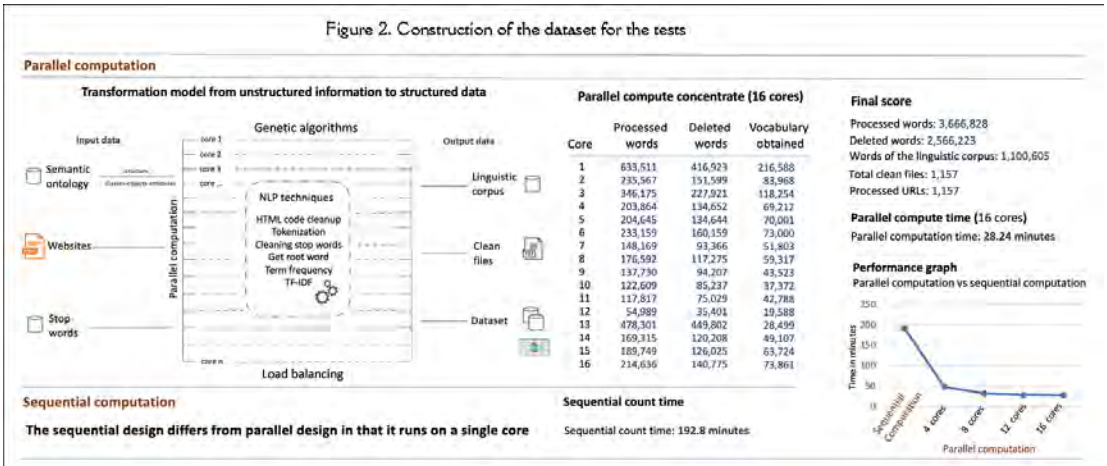
Regarding suicide-related web pages for testing purposes, 1,157 websites were located using the crawler and were downloaded to the hard drive. Concerning Semantic Ontology, the result is presented in Figure 1, where the traits of students with a suicidal tendency are described, organized by groups that reflect risk factors: ways of prevention, ways of carrying out suicide, similarities of suicide, types of suicide, factors that influence carrying suicide out, and signs indicating suicidal tendency. It should be emphasized that the information is based on the books: Suicide (Durkheim, 2008), The thought of suicide in adolescence (Villardón-Gallego, 2013), Suicide criminological approach (Marchiori, 2015), When nothing makes sense: reflections on suicide from logotherapy (Rocamora, 2017), Suicide the unbearable need to be another (Berengueras, 2018), The footprint of hopelessness: prevention strategies and coping with suicide (Urra, 2019), and Suicide: a comprehensive and integrative look (García-Peña, 2020).

Figure 1. Categorization of traits of students with suicidal tendencies (Semantic Ontology)



The results of the dataset construction for the tests are shown in Figure 2, illustrating the model of transformation from unstructured information to structured data, the Parallel Computing concentrate (taking as an example the 16 processor cores), specifying the core used, words processed, words removed, and vocabulary obtained.

Figure 2. Construction of the dataset for the tests



Likewise, the overall results of the process are presented, where 3,666,828 words were extracted, from which 2,566,223 empty words were eliminated, obtaining a linguistic corpus of 1,100,605 words of cybersuicide and 1,157 clean files with information extracted from websites. Finally, the time results of the Sequential Computation are included.

It should be noted that in the execution of the tests a better performance is obtained with the Parallel Computation (16 cores) on the Sequential Computation, obtaining an optimization of 682%, with a time of 28 minutes 24 seconds, against 192 minutes 8 seconds, obtaining the best time response when using a 16-core chromosome, as shown in the performance graph. Similarly, it can be seen, that the computation time begins to stabilize at 12 cores. On the other hand, the results of trait classification in students with suicidal tendencies are shown in Table 3. These were organized using Machine Learning techniques (Sequential Computation and Parallel Computation) and Deep Learning. Specifying the algorithm used, the values of the confusion matrix (true positives, true negatives, false positives, and false negatives), the percentage of precision and execution time.

Table 3. Results of prediction of characteristics in students with a suicidal tendency

Learning Types	Learning Techniques with Python	Learning Assessment						
		Algorithm	Confusion matrix				Precision (%)	Execution time (seconds)
			True Positive	True Negative	False Positive	False Negative		
Machine Learning	Sequential Computation (Sklearn)	Random Forests	293	44	6	5	96.84%	0.58 seg
		Neural Network	294	43	6	5	96.84%	2.09 seg
		Decision Tree	290	46	8	4	96.55%	0.03 seg
		Logistic Regression	293	46	6	3	97.41%	0.06 seg
	Parallel Computation (Turicreate) CPU + GPU AMD	Random Forests	289	52	14	4	94.98%	0.02 seg
		Logistic Regression	281	47	10	9	94.52%	0.07 seg
		Decision Tree	288	43	10	7	95.11%	0.01 seg
Deep Learning	Parallel Computation (TensorFlow y Keras) CPU	Recurrent Neural Network	297	44	6	1	98.00%	195.73 seg

For Machine Learning with Sequential Computation (Sklearn library), the algorithm that obtained the best classification result was Logistic Regression with 97.41% accuracy and a processing time of 0.06 seconds. The Random Forests, Neural Network, and Decision Tree showed stability above 96.55% very close to the first. The best time was obtained with the Decision Tree algorithm with 0.03 seconds.

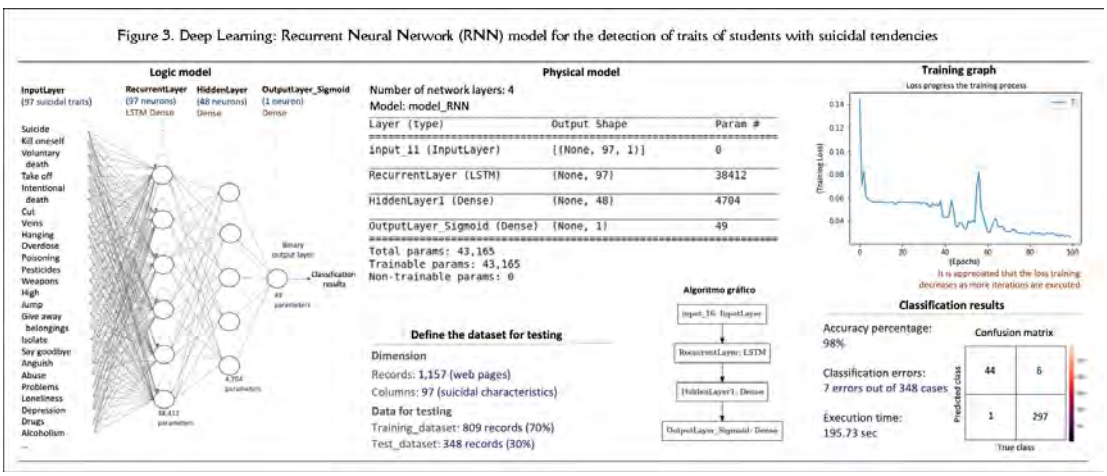


Figure 3. Deep Learning: Recurrent Neural Network (RNN) model for the detection of traits of students with suicidal tendencies

Regarding Machine Learning with GPU accelerated Parallel Computing (Apple's Turicreate Library), the algorithm that obtained the best classification result was a Decision Tree with 95.11% accuracy and a response time of 0.01 seconds. Random Forest and Logistic Regression algorithms obtained a stable

percentage close to 95%. According to the tests performed, Turicreate API algorithms do not operate on Windows Operating Systems, in addition, no support for Neural Networks was found on the official website of the library.

Figure 3 presents the solutions of the RNN model for the detection of traits of students with suicidal tendency. It shows the logical and physical model with the outputs generated (layers), the graphical algorithm, the preparation of the dataset for the tests, and the training and classification results.

The results of the RNN reveal a 98% prediction and 195.73 seconds of computation, obtaining the best classification result compared to the rest of the algorithms. The procedure examines the dataset that represents the 1,157 web pages and the 97 characteristics of Semantic Ontology, employing 70% for training and 30% for testing. It should be noted that to optimize the model, the method based on Adam's Learning Rhythm was implemented since it obtained better performance in the tests than the SGD and RMSProp methods. The accuracy of the model can be seen in the training graph; in which it is observed that the loss decreases as more iterations of the algorithm are performed. In the same line, the confusion matrix reveals the accuracy of the classification by reporting 7 errors and 341 hits. It is worth mentioning that 43,165 solutions were generated in the different layers of the RNN model to obtain the optimal solution.

4. Discussion

First, it is important to emphasize that the combination of different techniques and procedures has a greater technological and methodological scope, which can be seen in the excellent results. Putting Big Data Analytics techniques in practice into the information recovery and transformation processes, as well as Machine Learning for the discovery of knowledge, allows speeding up the analysis and classification of web pages for the user with an acceptable response time. One of the most important findings in the study is that the downloaded sample of 1,157 URLs, represents the web pages about suicide in Spanish because the web crawler searches showed repeated URLs mostly after 1,000.

As for computation time related to Artificial Intelligence processes, the algorithm that showed the best performance was Decision Tree using Parallel Accelerated Computation with GPU, and in contrast, the longest time was obtained by the RNN algorithm, associated with the same nature of recurrence of the algorithm. On the other hand, in the transformation of unstructured data to structured data, it was possible to optimize time by 682% by applying Parallel Computation compared to Sequential Computation.

Regarding the objective of the research (to detect traits or characteristics in students with a suicidal tendency in websites), the proposed methodology and architecture reached 98% accuracy in the classification, identifying patterns related to suicide signs, risk factors, ways to carry it out, people who influence the suicidal tendencies, types of suicide and forms of prevention. This allows the establishment of foundations for mechanisms of action and prevention of suicidal behaviors, which can be implemented by different social entities, including educational institutions, government agencies, and pro-suicide associations.

In relation to other authors, such as Gen-Min et al. (2020), Chiroma et al. (2018), we agree that the Machine Learning techniques employed (Neural Network, Decision Tree, Logistic Regression, and Random Forests) are highly effective in the classification and prediction of suicide. Regarding Deep Learning with RNN, the present research obtained better classification results than Du et al. (2018), with 98% and 67.94% accuracy respectively, derived from applying different techniques for the construction of the dataset.

Finally, functionality tests were carried out with the Artificial Intelligence algorithms in different operating systems, finding that the Linux Ubuntu 20.04 Operating System is ideal for working with the different hardware and software technologies.

5. Conclusions

Cyberspace becomes a shared ecosystem of information on websites, social networks, and people who comment on suicidal experiences based on anonymity, in which they interact and express a specific opinion or information without the need to expose their identity. Suicidal behavior in students highlights the

cracks in our contemporary society and confronts academic communities because they generate frustration, impotence, guilt for not having done what was needed at the time. This results in questioning the current educational system, which sometimes can be severe. It is necessary to understand education as a critical process in society since it is the starting point of socialization outside the family nucleus and in which citizenship is incubated. Therefore, it concerns a privileged place to train new generations in human values, which enable the support and cultural recreation of a specific society and can prevent problems such as instances of suicidal events.

It should be noted that the model proposed in this study may represent an interesting contribution to the analysis of data in cyberspace related to suicidal tendencies in students and adolescents. Thus, as the model is applied to new websites, it acts as an expert system, to identify patterns related to suicide signs, ways to carry it out, risk factors, ways to prevent it, and influences (terms defined in the semantic ontology). Thus, it is possible to establish grounds for the development of protocols (pro-suicide) in educational institutions, which enable the prevention of suicidal behavior through timely information and sensitization of both students and parents.

Among the results, high percentages are reported in the detection of traits in students with a suicidal tendency on websites, associated with the different dataset construction and Machine Learning techniques used. The results show an improvement in the transformation time from unstructured data to structured data using Parallel Computing Techniques with Genetic Algorithms, obtaining a 682% time saving compared to Sequential Computing. Likewise, an average accuracy of 96.28% is obtained, reaching an optimal value of 98% accuracy with the RNN algorithm. Therefore, it is inferred that a Recurrent Neural Network is a robust architecture for dealing with text analysis, where the output of the previous state is the feedback to preserve the memory of the network over time or sequence of words. Similarly, it is concluded that the proposed methodology and architecture are suitable for identifying and classifying suicidal signs in students with information from the web.

In relation with the results of the study, the exploration of descriptive analysis techniques on the suicidal traits obtained in the present study were considered to establish the degree of association between the groups and their variables. As future work, it is proposed to explore the suicidal behavior in adolescents based on short texts (tweets) from the social network Twitter to help in the detection of suicidal incidents, and to establish a hybrid data analysis model that combines sentiment analysis techniques, Semantic Ontologies, and Natural Language Processing, integrating Convolutional Neural Networks with Deep Learning for text classification and pattern search.

Authors' Contribution

Idea, I.C.Z.; Literature review (state of the art), I.C.Z.; Methodology, I.C.Z., F.J.L.R; Data analysis, I.C.Z., F.J.L.R.; Results, I.C.Z., J.I.L.V.; Discussion and conclusions, I.C.Z., J.I.L.V; Writing (original draft), I.C.Z.; Final revisions, I.C.Z., F.J.L.R, J.I.L.V; Project Design and sponsorships, I.C.Z.

Funding Agency

National Technological Institute of Mexico, Llano Aguascalientes Campus.

References

Anggraini, I.Y., Sucipto, S., & Indriati, R. (2018). Cyberbullying detection modelling at Twitter social networking. *Jurnal Informatika*, *6*(2), 113-118. https://doi.org/10.30595/juita.v6i2.3350

Arevalos, D.H. (2020). El sentido de la vida y las prácticas ligadas al suicidio. Testimonio de jóvenes escolarizados. Revista Latinoamericana de Estudios sobre cuerpos, emociones y sociedad. *RELACES*, *32*, 52-63. https://bit.ly/3pXSVC8

Beaven-Ciapara, N.I., Campa-Álvarez, R.A., Valenzuela, B.A., & Guillen-Lúgigo, M. (2018). Inclusión educativa: Factores psicosociales asociados a conducata suicida en adolescentes. *Prisma Social*, *23*, 185-207. https://bit.ly/3GRlO9X

Berengueras, M. (2018). *Suicidio la insoportable necesidad de ser otro*. Universidad Autonoma del Estado de Morelos. https://bit.ly/3F7a8iW

Blanco, C. (2019). El suicidio en España, respuesta institucional y social. *Revista de Ciencias Sociales*, *33*(46), 79-106. https://doi.org/10.26489/rvs.v33i46.5

Bonami, B., Piazentini, L., & Dala-Possa, A. (2020). Education, Big Data and Artificial Intelligence: Mixed methods in digital platforms. [Educación, Big Data e Inteligencia Artificial: Metodologías mixtas en plataformas digitales]. *Comunicar*, *65*, 43-52. https://doi.org/10.3916/C65-2020-04

Carballo-Belloso, J.J., & Gómez-Peñalver, J. (2017). Relación entre bullying, autolesiones, ideación suicida e intentos auutoliticos en niños y adolescentes. *Revista de estudios de Juventud*, *115*, 207-218. https://doi.org/10.3916/C65-2020-04

Chiroma, F., Liu, H., & Cocea, M. (2018). Text Clasiffication For Suicide Related Tweets. In *International conference on Machine Learning and Cybernetics (ICMLC)* (pp. 587-592). https://doi.org/10.1109/ICMLC.2018.8527039

Denia, E. (2020). The impact of science communication on Twitter: The case of Neil deGrasse Tyson. [El impacto del discurso científico en Twitter: El caso de Neil deGrasse Tyson]. *Comunicar*, *65*, 21-30. https://doi.org/10.3916/C65-2020-02

Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., & Xu, H. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Medical Informatics & Decision Making*, *18*(43), 77-87. https://doi.org/10.1186/s12911-018-0632-8

Durkheim, E. (2008). *El suicidio*. Grupo Editorial Éxodo. https://bit.ly/3p6C8h7

García-Peña, J.J. (2020). *El suicidio: Una mirada integral e integradora*. Universidad Católica Luis Amigó. https://doi.org/10.21501/9789588943619

Gen-Min, L., Szu-Nian, Y., & Yueh-Ming, T. (2020). Machine Learning based suicide ideation prediction for military personnel. *IEEE Journal of Biomedical and Healt Informatics*, *24*(7), 1907-1916. https://doi.org/10.1109/JBHI.2020.2988393

Healy, M. (2019). *Alcanzan máximo histórico los índices de suicidio de adolescentes y adultos jóvenes en EE.UU*. Los Angeles Times. https://lat.ms/3IYhebO

Hermosillo-De-La-Torre, A.E., Vacío-Muro, M.A., Méndez-Sánchez, C., Palacios-Salas, P., & Sahagún-Padilla, A. (2015). Sintomatología depresiva, desesperanza y recursos psicológicos: una relación con tentativa de suicidio en una muestra de adolescentes mexicanos. *Acta Universitaria*, *25*(2), 52-56. https://doi.org/10.15174/au.2015.900

Kim, J., & Chung, K. (2019). Associative feature information extraction using text mining from health big data. *Wireless Pers Commun*, *105*, 691-707. https://doi.org/10.1007/s11277-018-5722-5

Landaeta, G. (2014). *Lista de stop words o palabras vacías en español*. SEO para Google. https://bit.ly/3p2ysg3

López-Martínez, L.F. (2020). Suicidio, adolescencia, redes sociales e Internet. *Norte de SaludMental*, *17*, 25-36. https://bit.ly/3sg25g2

Luna, M., & Dávila, A. (2018). Adolescentes en riesgo: factores asociados con el intento de suicidio en México. *Revista Gerencia y Política de Salud*, *17*(34), 1-14. https://doi.org/10.11144/Javeriana.rgsp17-34.arfa

Marchiori, H. (2015). *El suicidio enfoque criminológico*. Editorial Porrúa. https://bit.ly/3sfMYDu

Molina, M.J., & Restrepo, D. (2018). Internet y comportamiento suicida en adolescentes: ¿Cuál es la conexión? *Revista Pediatría*, *51*, 30-39. https://doi.org/10.14295/pediatr.v51i2.109

Moreno, P., & Blanco, C. (2012). Suicidio e Internet. Medidas preventivas y de actuación. *Revista Psiquiatria.com*, *16*(18). https://bit.ly/3E110Lb

Mosquera, L. (2016). Conducta suicida en la infancia: Una revisión critica. *Revista de Psicología Clínica con Niños y Adolescentes*, *3*, 9-18. https://bit.ly/3p5lkG8

Nalini, K., & Sheela, L. (2014). A survey on Datamining in Cyber Bullying. *International Journal on Recent and Innovation Trends in Computing and Communication*, *2*(7). https://bit.ly/3q7GHqt

Olivares, S. (2019). Uso de Internet y conductas suicidas en adolescentes de 14 a 18 años en México. *Visión Criminológica-Criminalística*, (pp. 6-21). https://bit.ly/3p7uKSF

Organización Mundial de la Salud (Ed.) (2019). *Suicidio. Información obtenida el 6 de abril de 2021 en la dirección de Internet*. OMS. https://bit.ly/3q8TkBG

Pérez-Martínez, V.M., Aparicio-Vinacua, B., & Rodríguez-González, M.D. (2020). Acoso escolar, violación y suicidio en Twitter: Segunda temporada de «Por trece razones». *Vivat Academia*, *153*, 137-168. https://doi.org/10.15178/va.2020.153.137-168

Porter, M.F. (2006). An Algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, *40*, 211-218. https://doi.org/10.1108/00330330610681286

Ramírez-López, C.M., Montes, M., Ochoa-Zezzatti, A., Ponce-Gallegos, J.C., & Guzmán-Mendoza, J.E. (2021). Identification of possible suicide cases using a Bayesian Classifier with the database the Emergency Service 911 of Aguascalientes. *International Journal of Combinatorial Optimization Problems and Informatics*, *12*(1), 43-57. https://bit.ly/32gamWt

Rocamora, A. (2017). *Cuando nada tiene sentido: Reflexiones sobre el suicidio desde la logoterapia*. Editorial Desclée de Brouwer. https://bit.ly/3F9TxLa

Roy, S.S., Mallik, A., Gulati, R., Obaidat, M.S., & Krishna, P.V. (2017). A deep learning based artificial neural network approach for intrusion detection. In D. Giri, R. Mohapatra, H. Begehr, & M. Obaidat (Eds.), *Mathematics and Computing. ICMC 2017. Communications in Computer and Information Science* (pp. 44-53). Springer. https://doi.org/10.1007/978-981-10-4642-1_5

Sánchez-García, M.A., Pérez-De-Albéniz, A., Paíno, M., & Fonseca, P. (2018). Ajuste emocional y comportamental en una muestra de adolescentes españoles. *Actas españolas de psiquiatria*, *46*, 205-216. https://bit.ly/3yAOv8a

SeGob (Ed.) (2021). *Impacto de la pandemia en niñas y niños*. Secretaria de Gobernación de México. https://bit.ly/3J1bsGg

Urra, J. (2019). *La huella de la desesperanza: Estrategias de prevención y afrontamiento del suicidio*. Ediciones Morata. https://bit.ly/30H1QiQ

Villardón-Gallego, L. (2013). *El pensamiento de suicidio en la adolescencia*. Publicaciones de la Universidad de Deusto. https://bit.ly/33FtwFV