Article **ECORFAN Journal-Republic of El Salvador**
December 2021, Vol.7 No.13 20-33

Optimizing global processing time in the detection of depression related patterns in social networks

Optimizando el tiempo de procesamiento global en la detección de patrones relacionados con la depresión en redes sociales

MARTINEZ-DIAZ, Damian†*, LUNA-ROSAS, Francisco Javier, MEDIAN-VELOZ, Gricelda and MALO-TORRES, Miriam

Instituto Tecnológico de Aguascalientes (ITA), México.

ID 1st Author: *Damián, Martinez-Diaz* / **ORC ID:** 0000-0003-2748-305X, **arXiv Author ID:** damian.martinez.diaz, **CVU CONACYT ID:** 994749

ID 1st Co-author: *Francisco Javier, Luna-Rosas* / **ORC ID:** 0000-0001-6821-4046, **arXiv Author ID:** arXivFco19, **CVU CONACYT ID:** 87098

ID 2nd Co-author: *Gricelda, Medina-Veloz* / **ORC ID:** 0000-0002-1955-3620, **arXiv Author ID:** GrisArvix18, **CVU CONACYT ID:** 228438

ID 4th Co-author: *Miriam, Malo-Torres* / **ORC ID:** 0000-0002-0304-7897, **arXiv Author ID:** Miriammalo, **CVU CONACYT ID:** 1167058

DOI: 10.35429/EJRS.2021.13.7.20.33 Received July 25, 2021; Accepted December 30, 2021

Abstract

Depression is a common mental disorder and is on the rise worldwide according to the World Health Organization (WHO) around the world, has affected more than 322 million people, affecting mainly women than men, if this condition is not addressed in the most severe cases, it can lead people to suicide. Experts say that one of the best ways to prevent depression is to listen to the people who are close to them, social networks such as Twitter or Facebook are in a unique position to help these people to connect in real time in difficult situations, but also represents a potential risk to receive information that could later prove harmful. In this research we propose a model to optimize global time processing in detecting depression-related patterns on the social network Twitter. With the proposed methodology and with our results we demonstrate that the proposed model can be a good alternative when it comes to optimize the response time in this type of problems.

Sentiment analysis, Machine learning, Depression

Resumen

La depresión es un trastorno mental frecuente y está en aumento a nivel mundial de acuerdo con la Organización Mundial de la Salud (OMS) alrededor del mundo, ha afectado a más de 322 millones de personas, afectando principalmente a la mujer que, al hombre, si este padecimiento no se atiende en los casos más graves, puede llevar a las personas al suicidio. Los expertos afirman que una de las mejores maneras de poder prevenir la depresión, es que escuchen a las personas que está cerca de ellos, las redes sociales como Twitter o Facebook están en una posición única de poder ayudar a estas personas para conectarlas en tiempo real en situaciones difíciles, pero también representa un riesgo potencial a recibir información que posteriormente podrían resultar perjudicial. En esta investigación proponemos un modelo para optimizar el procesamiento de tiempo global en la detección de patrones relacionados con la depresión en la red social Twitter. Con la metodología propuesta y con nuestros resultados se demuestran que el modelo propuesto puede ser una buena alternativa cuando se trata de optimizar el tiempo de respuesta en este tipo de problemas.

Análisis de Sentimiento, Maquinas de Aprendizaje, Depresión

Citation: MARTINEZ-DIAZ, Damián, LUNA-ROSAS, Francisco Javier, MEDIAN-VELOZ, Gricelda and MALO-TORRES, Miriam. Optimizing global processing time in the detection of depression related patterns in social networks. ECORFAN Journal-Republic of El Salvador. 2021. 7-13:20-33.

* Correspondence to Author (Email: Damian.martinez.diaz@outlook.com)
† Researcher contributing first author.

Introduction

Depression is a mental illness that affects the emotional balance of people. Its detection is given fundamentally by different patterns of behavior of the individuals who suffer from it. Every year more and more people around the world are diagnosed with depression, including many adolescents and young adults. The impact of psychosocial factors in the adolescent and young adult population can exacerbate the intensity of the illness and exponentially increase suicidal ideation, suicidal attempts and even success.

Depression, according to the World Health Organization (WHO), is a common and treatable affective mental disorder, common in the world and characterized by changes in mood with cognitive and physical symptoms, and these can be of primary or secondary etiology when underlying diseases are found, such as cancer, cerebrovascular disease, acute myocardial infarction, diabetes, HIV, Parkinson's disease, eating disorders and substance abuse (World Health Organization, 2020).

According to WHO, depression around the world has affected more than 322 million people, if this condition is not treated, in the worst case it can lead to suicide, which each year has about 800,000 people, since suicide is the second leading cause of death in the age group 15-29 years. Although there are effective treatments for depression, more than half of those affected worldwide (and more than 90% in many countries) do not receive such treatments. Barriers to effective care include a lack of resources and trained health personnel, in addition to the stigmatization of mental disorders and inaccurate clinical assessment (WHO, 2017).

In Mexico, the National Institute of Statistics and Geography (INEGI) in the National Household Survey (ENH) 2017, shows that 31.96 million people aged 12 years and older have experienced a feeling of depression, which is equivalent to 3.17 million people of both sexes who reported feeling depressed on a daily basis, 3. 72 million reported feeling depressed weekly; 3.6 million reported feeling depressed once a month; while 21.39 million reported feeling depressed once a year, with this condition being more common among women with 60.34% of the total, equivalent to 19. 28 million, of which 2.11 million women reported feeling depressed daily, 2.34 million reported feeling depressed weekly, 2.35 million reported feeling depressed monthly and 12.46 million reported feeling depressed once a year, and men with 39. 65% of the total, equivalent to 12.67 million, of which 1.06 million men reported feeling depressed daily, 1.37 million reported feeling depressed weekly, 1.31 million reported feeling depressed monthly, and 8.92 million reported feeling depressed once a year (INEGI E. N., 2017).

Adolescence is a critical stage for human growth, since it has important physiological and psychological changes, as they make young people vulnerable, not knowing how to cope with them. These changes produce great anxiety, confusion, and despair. When going through a stress of conflicting feelings, both for family problems, school or their own personality, causing them to make unwise decisions such as alcoholism, drugs and even suicide attempts.

In addition, with the confinement suffered by the pandemic worldwide with SARS COV-2 or COVID-19 as it is colloquially known, according to the National Health and Nutrition Survey (ENSANUT), depression increased very significantly with a prevalence of 13.6% in 2018 (Cerecero-Garcia, 2020) to 27.3% in April 2020, although the reduction in depression has been significant, it still continues above the 2018 measurement affecting mainly women than men and with a lower socioeconomic level. This pandemic has detonated a growth in the number of internet users, in Mexico alone it is estimated that it had 84.1 million internet users, representing 72% of the population aged six years or more, increasing 1.9 points more than in 2019 (70.1%). Being 71.3% women and 72.7% men, within the main activities of users are communicating (93.8%), searching for information (91%) and browsing social networks (89%) (ENDUTIH, 2021).

ISSN-On line: 2414-4886

With the increase of Internet users across the country, in recent years several researches have been guided to the detection of these disorders through machine learning, analyzing it as a problem of text classification. The process of text classification consists of extracting several features from a set of previously labeled data and from these learn models that allow to distinguish between several classes. Obtaining the data is a crucial step for the correct classification of the data, however it is a process that needs a high use of computational resources in the preprocessing, especially in the classification of data from different sources that can be obtained from social networks such as Twitter, Facebook and Instagram.

This need for computational resources has led the scientific community to seek a solution to these problems by means of parallel computing, which is a technique that is being used in the fields of simulation of mathematical, statistical, climatic calculations and even with image processing that require high processing capacity, using its use in different levels of laboratories among which we can find supercomputers, distributed systems, multicore processors, graphics processors, cloud computing up to quantum parallelism, always with the same objective which is to seek an optimization of processing times.

The rest of this article has been organized as follows. Theoretical background related to social networks, analysis and classification of short texts in section 2. The materials used in the creation of the model are detailed in section 3, The methodology and implementation details have been discussed in section 4. The evaluation of the model is demonstrated in section 5. The conclusions have been presented in section 6, followed by future work in section 7.

Theoretical background

The increasing prevalence of psychological disorders such as depression and post-traumatic stress requires a serious effort to create new tools and technologies to aid in their diagnosis and treatment. In recent years, new computational approaches have been proposed to objectively analyze the patient's nonverbal behavior (Ghosh, 2014), as well as the extraction of features in video, audio and text as proposed by (Dham, 2017) or also in a more conventional way with the use of depression inventory proposed by (Beck, 1984).

As depression is a difficult situation, ethical considerations should be taken to inform the interviewees that the data will be confidential and will be used only for research purposes as done in (Granados Cosme, 2020).

Social networks such as Twitter and Facebook are increasingly associated with phenomena such as harassment, bullying, suicide or even depression (Marouane Birjali, 2016). It is therefore very important to detect potential victims as early as possible to strengthen the prevention of these phenomena on social networks. Specific linguistic features such as articles, prepositions, auxiliary verbs, adverbs, conjunctions, personal pronouns, impersonal pronouns, verbs and negations are the most important combinations that authors (Islam, 2018) have tried to find in the comments of these social networks, in search of patterns that lead them to detect depression. There are research works for the analysis of social networks specifically on the opinion of different topics such as depression, alcoholism, and drugs. These works base some of their techniques on the extraction and classification of positive, negative and neutral feelings. This classification can be achieved by statistical analysis which is divided into supervised classification (J. Pestian, 2010), (Maria Khodorchenko, 2019), (Jacques Philip, 2016), (Robert A. Fahey, 2018), and unsupervised (Matykiewicz P, 2009) with which we can explore different types of attributes or classes, to model, detect and predict [28] depression, suicide or any other keywords.

These analyses are accompanied by different techniques and tools such as the one used by (Manabu Torii, 2015) (Nguyen T. O., 2017), who made use of natural language processing (NLP) for the detection of patterns that could help to relieve the depression attitudes of relevant online users, or looking for an improved compilation of labeled dataset with the help of heuristics (Maria Khodorchenko, 2019), (Ruben Sanchez Acosta, 2019). As well as relying on tools such as nQuiry (INQUIRY, 2020), MedEx (MEDEX, 2020), Weka (Java, 2019), RIP-PER (RIP-PER, 2020), LibSVM (Lin, 2019), Stanford NER (Group, 2020), Twitter4J (API, 2020) ,WordNet (English, 2011), they managed to generate hybrid system, based on rules and machine learning (Manabu Torii, 2015).

ISSN-On line: 2414-4886

With The information generated in social networks is growing in an exponential way in content (Liang & Dai, 2013) shared by users in more than 900 social networking sites available on the Internet that are accessed today, among the most recognized we can find Facebook, Instagram, and Twitter. The latter is ranked as one of the most visited and used networks in the world, with an average of more than 58 million tweets generated per day (Li, Lei, Khadiwala, & Chang, 2012).

These "tweets" are short message posts (280 characters) which are created and shared in real time. This speed of communication is reaching the point that traditional news is becoming obsolete (Chih-Hua T., 2015), since a news story would normally take 3 hours to be reported on an incident, with a tweet shared on Twitter it would take no more than 10 minutes to be known among the users of this social network. But not everything is positive as the speed and ease of information we have on social networks causes phenomena such as harassment, bullying, depression or even suicide (Marouane Birjali, 2016), and some cases cause more unfavorable effects such as that of "copycat" called "Werther Effect" (Phillips, 1974) (Ueda M, 2017), since after the suicide of some celebrities (Marouane Birjali, 2016) makes users followers (Followers in English) of them, may come to make unwise decisions, which added to the stress and depression that one has from everyday problems due to lack of money, problems with the couple, with school or perhaps with work, school bullying etc. , can lead to a mental disorder that can be fatal.

Applying machine learning techniques to online communities is a viable method to improve our understanding of how online communication can be used to characterize people's experience of depression, as it is a very prevalent mental health problem and is a comorbidity of other mental, physical and behavioral disorders, identifying (Nguyen T. O., 2017) five subgroups of online communities: depression, bipolar disorder, self-harm, grief/grief and suicide. Psycholinguistic features and content themes were extracted from the postings and analyzed.

Even though techniques, methods and tools have been proposed, automatic detection of suicidal, depressive or stressful content in social networks is scarce, machine learning approaches are available, which have the potential to significantly impact the prediction of related events, but have not yet been able to reach short-term prediction, despite the great potential of these models, such as the low accuracy results (Bart Desmet V. H., 2018) they obtained when weighting user profiles in social networks.

Neural networks unlike some classical statistical regression models are designed to accommodate high-dimensional inputs, they can be useful for prediction. However, prospective comparisons of machine learning tools for short-term prediction have not yet been carried out, despite the tremendous potential. In part, this is because short-term risk factors derived from social networks and smartphones are not yet well characterized or validated in crucial ways, even the best computational methods for risk assessment will only be as good as the risk factor data provided to them (Torous, 2018).

Materials

Python (v3.7.x) is an interpreted programming language whose philosophy emphasizes the readability of its code. It is a multi-paradigm programming language, as it partially supports object-oriented, imperative programming and, to a lesser extent, functional programming. It is an interpreted, dynamic and cross-platform language.

Scikit-learn (v0.24.2) is a Python library that supports supervised and unsupervised learning. It also provides several tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities.

Pandas (v1.3.1) is a Python library and is intended to be the fundamental high-level building block for performing practical, real-world data analysis. It also has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in any language.

ISSN-On line: 2414-4886

TextBlob (v0.16.0) is a Python (2 and 3) library for textual data processing. It provides a simple API to dive into common natural language processing (NLP) tasks, such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and translation.

Random Forest algorithm consists of a set of individual decision trees, each trained with a slightly different sample of the training data generated by bootstrapping). The prediction of a new observation is obtained by aggregating the predictions of all the individual trees that make up the model.

Representational State Transfer (REST) completely changed software engineering starting in 2000. This new approach to the development of web projects and services was defined by (Fielding, 2000), the father of the Hypertext Transfer Protocol (HTTP) specification and one of the international references in everything related to Network Architecture. In the field of Application Programming Interfaces (APIs). Currently it is difficult to find projects or applications that do not have a REST API for the creation of professional services from that software as used by Twitter, YouTube, Facebook, etc.

Twitter REST API: Offers developers access to Twitter's core data. All operations that can be performed via the web can be performed from the API. Depending on the operation it requires authentication or not, with the same criteria as in web access. Supports formats: XML, JSON, RSS, ATOM.

The computer equipment was a Microsoft Windows 10 Enterprise HP ZBOOK 15v G5, with NVIDIA Quadro 9600 graphics card with Intel® Xeon® E-2176M CPU @ 2.70GHz 2.71GHz, 16 GB memory and 500 GB solid state hard disk.

Methodology

In this section we sought to recognize the context of the problem of our research through the approach of concepts of the most relevant terms such as the vocabulary of depression to be used, the natural language processing (NLP) tools, the selected Machine Learning algorithm, the programming language and even the perfection of the techniques that were used to calibrate the model in the executed algorithm, proposing six phases that apply the above described and will be detailed as shown in Figure 1.

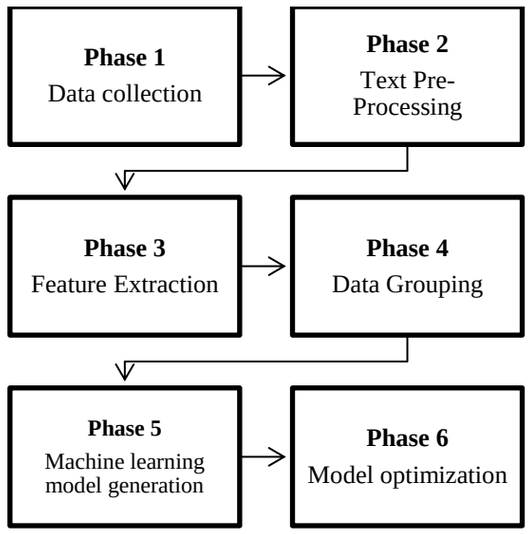

Figure 1 Phases of the proposed methodology

I. Phase 1. Data Collection.

The phase begins with the definition of the vocabulary of various words detected in the literature on depression. In Annex 1. the words that were collected in the English language vocabulary such as anxiety, anger, depression, etc., as mentioned (Ghosh, 2014), (Islam, 2018), (James G. Phillips, 2019), (Nguyen T. V., 2016), (Beltran, 2012), (Padilla-Navarro, 2016), (Marouane Birjali, 2016) can be observed. With this vocabulary related to depression topics, tweets were filtered on the social network Twitter. The collection of tweets was carried out for a period of 2 hours daily over a period of approximately 6 months, achieving a significant sample of more than 1,000,000 tweets, which were backed up in a database of files for processing in the following phases of this research.

ISSN-On line: 2414-4886

II. Phase 2. Text preprocessing

With the use of natural language processing methods and techniques, we seek to transform the data obtained into a formal representation of the tweets extracted from Twitter. In this transformation, the texts are sought to be cleaned and prepared for processing before sentiment classification, where misspellings, stop words, sentence boundaries, punctuation marks are removed and emoticons or emojis are replaced by corresponding words, among other techniques, such as those used by (Bart Desmet V. H., 2018), (Tadesse MM, 2019), (Cheeda, 2018), (Agarwal, 2011), (McDonals, 2020), which propose converting tweets to lowercase, elimination of special characters such as [:;?()!,.], blank spaces etc.; which in the cases used by the literature have worked for the proposed purpose, but for our research it is proposed to add a list of techniques which we will call from now on, techniques for eliminating noise in tweets (TER-TWS). Since, added to those used in the literature, we can obtain a tweet with less noise and with more selective text for use in the proposed model.

The tweet de-noising technique (TER-TWS) is composed of the following list of tasks.

- Replace abbreviations [don't].
- Remove old style retweets (RT).
- Change all comment text to lowercase.
- Replace website URLs with local variable.
- Replace @username with local variable.
- Replace blank spaces.
- Replace Hashtag (#).
- Remove special characters and numbers.
- Remove e-mails.
- Remove line breaks.
- Remove duplicate words.
- Apply the Stop Words technique.

III. Phase 3. Feature extraction

In this phase we seek to extract the most relevant features and that are extremely useful for the correct classification or retrieval of information to be detected in the tweet, for which tokenization techniques such as the one used by (Bart Desmet V. H., 2018) (Bart Desmet V. H., 2013) are used in the message, eliminating words that have little interest for our research, generating a feature vector for each of the tweets collected.

This vector is created from a set of N consecutive elements in a document (tweet), which includes words, numbers, symbols and punctuations, which were in some cases eliminated in the previous phase for not being relevant, this based on the N-gram modeling that is used in text mining and PLN (M. M. Tadesse, 2019). Another strategy used is the lemmatization technique which is a process by which the words of a text belonging to the same inflectional paradigm that taken to a normal form represents the whole class until the corresponding lemma is found. This lemma is the form that by convention is accepted as representing all the inflected forms of the same word. With the help of these two techniques used in this phase it is possible to visualize in the vector obtained from the tweet the following form ['seen', 'terrible', 'argument', 'weapon', 'victim', 'death', 'heartbreak', 'tragedy'], which has been run through the more than one million tweets collected and which will be used in the next phase.

IV. Phase 4. Grouping of the data.

In this phase we have challenges with the sentiment analysis, since we must determine if there is any opinion in the tweet since it could be just an objective comment or always a response to another user, as well as a topic not relevant to the depression, for this reason it is important to recognize the abbreviations and idioms of the words to be found. Unfortunately, as Twitter is an informal social network, what is expressed by users is not always the most structured, accented and popular words are used in most of the messages according to the region where it has been used and that are not necessarily in the traditional dictionary, and sometimes we can find in the same sentence positive and negative words, which makes it more complicated to determine the polarity of the option expressed by users.

ISSN-On line: 2414-4886

The polarity of each tweet is determined by assigning a score of [-1.0, 1.0] which refers to how the text can be measured in positive or negative depending on the tone of the tweet, for which -1 will indicate that they are more negative and +1 is more positive, while the value of zero will be considered as a neutral sentiment. A subjectivity score [0.0, 1.0] refers to the representation of a subjective or objective meaning, where the value close to zero represents an objective comment and close to 1 is a subjective comment (Tom De Smedt, 2020).

To detect polarity and subjectivity in the tweets collected for this research, use was made of Python's TextBlob library, which internally uses a dictionary with a total of 2920 words previously classified with polarity and subjectivity values, it can be an advantage to use a dictionary-based approach to extract tweet sentiment, since a large number of words with their orientations can be found quickly, but it can be turned into a disadvantage since such sentiment orientations of the words collected in this way are general or independent of the author's context and language.

Based on the obtained result of polarity and subjectivity we used a lexical approach since using a large number of tweets and with the bag of words created in the previous phases, we assigned an individual score to each of the words in the vector and finally calculated the sentiment by a grouping operation, with the mean of the sentiments. We classified this with 5 different numerical indicators (classes), in the more than 1,000,000 tweets used in this research, which were given a textual value for the human understanding and a numerical indicator for the machine understanding of our algorithm, using the latter for classification based on a machine learning line (Yang, 2018), as can be seen in Table 1.

Indicators	Value	Classes
P+	Very Positive	5
P	Low Positive	4
NEU	Neutral	3
N	Slightly Negative	2
N-	Very Negative	1

Table 1 Polarization-based sentiment indicators and values

Where the values are added in the Excel file used in previous phases, adding two more columns, where we can see the indicator value very positive, little positive, neutral, little negative and very negative with their respective numerical class for our algorithm. These will be used in the next phase as output data for the model to be used.

V. Phase 5. Machine Learning Model Generation

The supervised learning (Machine Learning) takes a set of data (inputs) and known responses, and looks for a way to build a predictive model that generates reasonable or adequate predictions to the new data entered. That is, given a database D={$t1$, $t2$, ...,tn} of tuples or records (individuals) and a set of classes C={$C1$, $C2$, ...,Cm}, the classification/prediction problem is to find a function f: $D \rightarrow C$ such that each ti is assigned to a class Cj. f:$D \rightarrow C$ could be a KNN Method, a Decision Tree Method, a Support Vector Machine, a Bayesian Model, a Random Forest Method and a Boosting Method (Francisco Luna Rosas, 2018).

Figure 2 shows our supervised learning model to detect depression-related patterns in social networks.

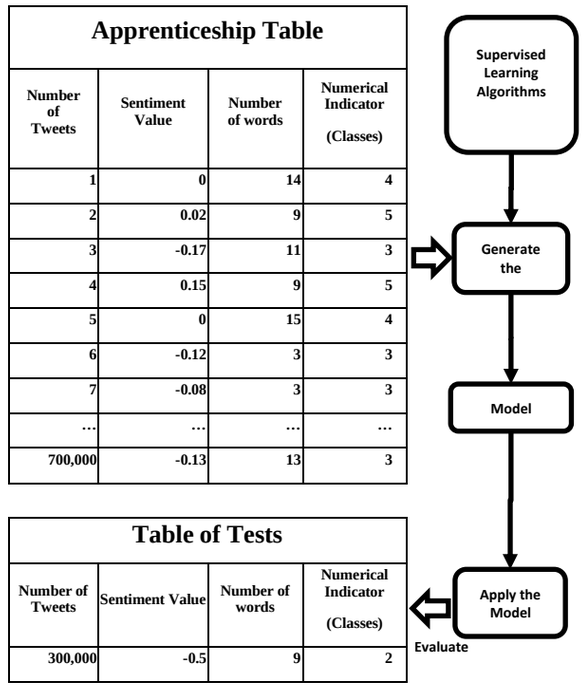

Figure 2 Supervised learning model

The Algorithm selected for this research was Random Forest, being a supervised machine learning algorithm, it has two important features, regression and classification, using the latter with a calibration of 200 decision trees in the forest (estimator). To train and test it, the Table Testing method was used, which implies that a percentage of data was used for training and another percentage was used for testing (Francisco Luna Rosas, 2018) (Maaoui, 2016), in both cases the input data were the value of the sentiment, and the number of words per tweet and the output to predict is the numerical indicator of the class, in our case of the total number of tweets collected and processed in the previous phases was 1,000,000 and of which 70% were used for training which equals 700,000 and the remaining 30% for testing which equals 300,000 as shown in Table 2.

Tweets Collected (1,000,000)	
Training (700,000)	Testing (300,000)

Table 2 Testing table method

VI. Phase 6. Model optimization

The main contribution of our research is the optimization of the overall processing time of the model, in which parallel computing is key to achieve our goal. There are different forms of parallel computing such as: bit-level parallelism, instruction-level parallelism, data parallelism and task parallelism (Rosas, 2018). The latter was the one used in our research, since it makes use of the concurrent programming paradigm that consists of assigning different tasks to each of the processors of a computing system.

Taking into consideration that social networks continue to take a wide boom worldwide and that users share millions of tweets with different contexts, the volume of data is considerable, and according to WHO estimates, depression has affected more than 322 million (INEGI, 2019) per year globally, this research can be a link between the needs of depression prevention with the interaction of technological trends of social networks with users going through this situation.

The Python programming language used in all phases of this research has support for Parallel processing, being one of the favorite languages for Big Data analysis, using libraries such as multiprocessing or threading (Python, 2020), which use "parent process" and "workers" or "helpers" architectures, formerly known as "master" and "slave" (Mariatta, 2018), as well as the management of processing threads in the system, which will help us to optimize the execution time of the preprocessing phases seen in the previous sections.

To achieve time optimization, phases I, II, III and IV were selected, because they carry a high degree of computational processing, from the extraction, cleaning, calibration and classification of the data, which causes it to take considerable time to preprocess them in a conventional sequential manner and is where parallel processing makes a big difference.

With the generation of the database that was created in phase I, 10 files with different number of tweets and file size were created. see Table 3.

Cycles	MB Size	Number of Tweets	Sequential Process (Seconds)	Parallel Process (Seconds)	Optimization %
1	49.05	100,000	733.99	151.8	79.32
2	113.66	200,000	1303.24	298.82	77.07
3	163.82	300,000	2298.57	450.68	80.39
4	214.46	400,000	2626.05	601.69	77.09
5	265.21	500,000	3516.43	752.16	78.61
6	316.73	600,000	4223.19	912.5	78.39
7	368.43	700,000	4568.22	1061.01	76.77
8	419.99	800,000	5252.52	1202.75	77.10
9	471.62	900,000	6288.38	1353.21	78.48
10	523.62	1,000,000	7151.38	1494.95	79.10

Table 3 Sequential vs. Parallel execution times

ISSN-On line: 2414-4886

Which were processed sequentially and in parallel, where in the sequential process began with the reading of file by file and tweet by tweet invoking the process of the other facets involved in the optimization, which makes the preprocessing slow and visualizing a considerable time which is doubled for each of the files, as visualized in Table IV, where we can clearly see that going through 100,000 tweets in the first file and finishing it achieves an approximate time of 733.99 seconds, and when continuing with the second one it increases to 1303.24 seconds and so on, unlike the parallel process, which reads file by file, but with the difference that the tweets are distributed by the total number of cores that the system has available, which is achieved with Python's own libraries mentioned above, to do the same tasks of the phases mentioned above, however in a parallel way, a time optimization is achieved, which in the first file we can see an improvement of 79.32% which is equivalent to 151. 8 seconds, and thus improving the next one with a time of 298.82 seconds, which is equivalent to 77.07% improvement and so on, a challenge to be considered in this optimization is the level of computation that must be considered for this type of optimizations, because as the files in the cycles grow in size and data in number of tweets, it is more complicated its processing, because it takes all the resources of the system to attend the task sequentially and the tasks in parallel.

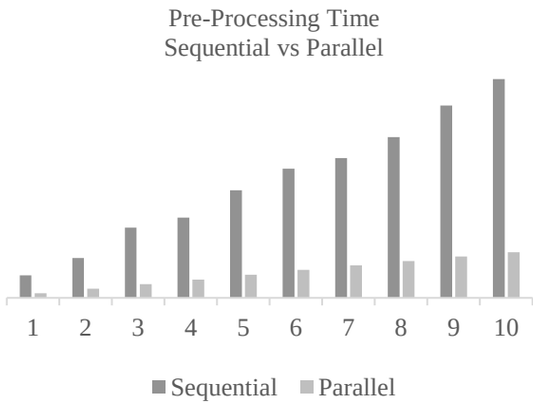

Figure 3 Preprocessing times between Sequential and Parallel

In Figure 3 we can visualize the 10 cycles with the improvements previously seen in Table IV, and in which we can clearly see how the parallel processing is separated far above the sequential processing, achieving an overall preprocessing optimization percentage of 78.23% between the comparison of the two processing techniques.

Model Evaluation

The model performance evaluation as the literature recommends, makes use of a confusion matrix (Maaoui, 2016), which contains information about the actual and estimated classifications, this matrix is NxN, where the rows are named according to the actual input classes and the columns are those predicted by the model, and are used to explicitly detail when a class is confused with another, as shown in Table 4.

Classes	1	2	3	4	5
1	**32593**	0	0	0	0
2	0	**261525**	0	0	0
3	0	0	**225622**	0	0
4	0	0	0	**174381**	0
5	0	0	0	0	**5880**

Table 4 Random Forest confusion matrix

In order to evaluate this performance, we will apply the evaluation metrics (Basu, 2012) as follows:

True positives (VP): are those samples with positive classes have been classified as positive (correctly classified).

True negatives (TN): Those samples with negative class have been classified as negative (correctly classified).

False positives (FP): Those samples with negative class that have been classified as positive (incorrectly classified).

False Negatives (FN): Those samples with positive class that have been classified as negative (incorrectly classified).

With these evaluation metrics, we are looking for:

Overall Precision (PG) which is one of the most widely used metrics for classification performance, and is defined as a ratio of correctly classified samples to the total number of samples (Tharwat, 2020) and is achieved with the following formula:

$$\boldsymbol{PG} = (\boldsymbol{VP} + \boldsymbol{VP}) \,/\, (\boldsymbol{VN} + \boldsymbol{FP} + \boldsymbol{FN} + \boldsymbol{VP}) \qquad (1)$$

ISSN-On line: 2414-4886

Accuracy (Bart Desmet V. H., 2013) is used to be able to measure the QUALITY that the model is able to identify in the classification task, and is achieved with the following formula:

$$\boldsymbol{Precision} = \boldsymbol{VP} / (\boldsymbol{VP} + \boldsymbol{FP}) \quad (2)$$

Recall (completeness or sensitivity) is used to be able to measure the QUANTITY that the model is able to identify in the classification task, and is achieved with the following formula:

$$\boldsymbol{Recall} = \boldsymbol{VP} / (\boldsymbol{VP} + \boldsymbol{FN}) \quad (3)$$

The F1-Score is used to be able to combine the measures of precision and recall into a single value, as it makes it easier to compare the combined PERFORMANCE among several solutions and is achieved with the following formula:

$$\boldsymbol{F1 - Score} = 2 * (\boldsymbol{Precision} * \boldsymbol{Recall}) / (\boldsymbol{Precision} + \boldsymbol{Recall}) \quad (4)$$

The Random Forest classifier was the algorithm we used in our model to validate its efficiency, we used an initial calibration for the estimators of 200, obtaining an accuracy of 1.00, with an error ratio of 0.0, the detail by indicator can be seen in Table 6.

Indicators	PG	ReCall	F1-Score	Total
1	1	1	1	**32593**
2	1	1	1	**261525**
3	1	1	1	**225622**
4	1	1	1	**174381**
5	1	1	1	**5880**

Table 6 Details of accuracy by indicato

Annexes

Annex 1. Depression vocabulary

Vocabulary English	
Literature	**Content**
Ghosh, S., Chatterjee, M., & Morency, L. P. (Ghosh, 2014)	sad, health, anxiety, anger, leisure, negate, hear, I, assent.
Islam, Md Rafiqul, Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (Islam, 2018)	happiness, sadness, anger, anxiety, depression, bipolar.
James G. Phillips, Leon Mann (James G. Phillips, 2019)	suicide, crowd, audience, depression, suicide attempt webcam, skype, and livestream.
Nguyen, T., Venkatesh, S., & Phung, D. (Nguyen T. V., 2016)	Sadness, health, anxiety, death, insight, negations, etc.
Beltrán, M. D. C., Freyre, M. Á., & Hernández-Guzmán, (Beltrán, 2012)	Sadness, Pessimism, Dissatisfaction, Guilt, Self-loathing, Irritability, Fatigue, etc.
Padilla-Navarro, C., Pedruelo, M. R., & Ramírez, C. L. (Padilla-Navarro, 2016)	anxiety, anger, depression, self-consciousness, immoderation, vulnerability, etc.
Marouane Birjali, Abderrahim Beni-Hssane, Mohammed Erritali. (Marouane Birjali, 2016)	fear, depression, harassment.

Conclusion

Social networks are at a very high point of popularity and this attracts the attention of people from all over the world to create social interconnections between users. Opinion mining and sentiment analysis on Twitter data are more popular with the passage of time, making users to express their sentiments with a greater ease. In this research, we have proposed a methodology to detect depression-related opinions using this type of mining and analysis. The proposed system is able to analyze a large dataset of tweets to classify them into five different classes from neutral, very negative, and little negative, as well as, little positive and very positive.

The text classification techniques used in the collected tweets have been adjusted, adapted and integrated to build a methodology to help the classification of the proposed indicators is this research and have shown that, with a good classification of sentiment polarity using tweet denoising techniques (TER-TWS), good results can be obtained, achieving in our model an acceptable accuracy of 100 with no margin of error.

Finally, it has been demonstrated that parallel processing with the use of multiprocessors and task distribution generates us better results in the preprocessing times of the more than 1,000,000 tweets classified with the methodology in its different phases, 78.23% overall time optimization has been achieved against traditional sequential processing.

ISSN-On line: 2414-4886

Future work

With the results obtained, we will seek to optimize the times with clustering techniques in distributed containerized systems, which can help us to further optimize these response times for the prediction of tweets posted by users with depression problems and some other disorder that is able to be used within our same methodology.

References

Agarwal, A. B. (2011). Sentiment analysis of twitter data. En L. 2011 (Ed.), In Proceedings of the Workshop on Language in Social Media, (págs. 30-38). New York, USA.

API, T. -A. (2020). Twitter4j.org. (twitter4j.org) Obtenido de http://twitter4j.org/en/

Bart Desmet, V. H. (2013). Emotion detection in suicide notes. Expert Systems with Applications, 40(16), 6351-6358. Obtenido de https://doi.org/10.1016/j.eswa.2013.05.050.

Bart Desmet, V. H. (May 2018 de 2018). Online suicide prevention through optimised text classification. Information Sciences, 439–440, Pages 61-78, ISSN 0020-0255, doi: https://doi.org/10.1016/j.ins.2018.02.014.

Basu, T. a. (2012). A feature selection method for improved document classification. Advanced Data Mining and Applications., 7713, 296-305.

Beck, A. T. (1984). Internal consistencies of the original and revised Beck Depression Inventory. Journal of clinical psychology, 1365-1367. doi:doi.org/10.1002/1097-4679(

Beltrán, M. D.-G. (2012). El Inventario de Depresión de Beck: Su validez en población adolescente. Terapia psicológica,, 5-13. doi:dx.doi.org/10.4067/S0718-48082012000100001

Cerecero-García, D. F.-G.-M.-A. (2020). Síntomas depresivos y cobertura de diagnóstico y tratamiento de depresión en población mexicana. Salud Pública de México, 62(6), 840-850. doi: https://doi.org/10.21149/11558

Cheeda, S. S. (2018). Automated trading of cryptocurrency using twitter sentimental analysis. J. Comput. Sci. Eng, (págs. 209-214).

Chih-Hua T., Z.-H. T.-S.-S. (2015). Mental Disorder Detection and Measurement Using Latent Dirichlet Allocation and SentiWordNet. 2015 IEEE International Conference on Systems, Man, and Cybernetics, Kowloon, 2015, 1215-1220, doi: 10.1109/SMC.2015.217.

Dham, S. S. (2017). Depression scale recognition from audio, visual and text analysis. doi:arXiv:1709.05865v1

ENDUTIH, 2. (22 de Junio de 2021). https://www.inegi.org.mx/app/saladeprensa/noticia.html?id=6606. Recuperado el 2021 de Julio de 2021, de https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/OtrTemEcon/ENDUTIH_2020.pdf

English, W. |. (1 de Junw de 2011). Wordnet.princeton.edu. (1 June 2011;) Recuperado el 2020, de https://wordnet.princeton.edu/

Fielding, R. T. (2000). Architectural Styles and the Design of Network-based Software Architectures. Recuperado el 3 de Agoust de 2021, de https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Francisco Luna Rosas, J. C. (5 de Enero-Diciembre de 2018). PCA Y SVM EN PARALELO PARA OPTIMIZAR EL DIAGNOSTICO DE CÁNCER DE MAMA BASADO EN ESPECTROSCOPÍA RAMAN. DYNA New Technologies, 14. Obtenido de http://dx.doi.org/10.6036/NT8597

Ghosh, S. C. (2014). A multimodal context-based approach for distress assessment. In Proceedings of the 16th International Conference on Multimodal Interaction, (págs. 240-246).

Granados Cosme, J. A. (2020). Depresión, ansiedad y conducta suicida en la formación médica en una Universidad en México. Investigación en educación médica,, 35. doi:doi.org/10.22201/facmed.20075057e.2020.35.20224

Group, T. S. (2020). Nlp.stanford.edu. Recuperado el 2020, de https://nlp.stanford.edu/ner/

ISSN-On line: 2414-4886

INEGI. (10 de Septiembre de 2019). ESTADÍSTICAS A PROPÓSITO DEL DÍA MUNDIAL PARA LA PREVENCIÓN DEL SUICIDIO. Obtenido de https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2019/suicidios2019_Nal.pdf

INEGI, E. N. (2017). Encuesta Nacional de los Hogares (ENH) 2017. (inegi) Obtenido de https://www.inegi.org.mx/temas/salud/: https://www.inegi.org.mx/programas/enh/2017/

INQUIRY. (2020). Merriam-webster.com. Recuperado el 2020, de https://www.merriam-webster.com/dictionary/inquiry

Islam, M. R. (2018). Depression detection from social network data using machine learning techniques. Health information science and systems, 1-12. doi:doi.org/10.1007/s13755-018-0046-0

J. Pestian, H. N. (Jan de 2010). Suicide Note Classification Using Natural Language Processing: A Content Analysis. Biomedical Informatics Insights, 3, Available: 10.4137/bii.s4706.

Jacques Philip, T. F. (2016). Relationship of Social Network to Protective Factors in Suicide and Alcohol Use Disorder Intervention for Rural Yup'ik Alaska Native Youth,. Psychosocial Intervention,, 25(1), Pages 45-54, ISSN 1132-0559, https://doi.org/10.1016/j.psi.2015.08.002.

James G. Phillips, L. M. (2019). Suicide baiting in the internet era. Computers in Human Behavior, 92, 29-36. doi:https://doi.org/10.1016/j.chb.2018.10.027.

Java, W. 3.-D. (2019). Cs.waikato.ac.nz. (University of Waikato) Obtenido de https://www.cs.waikato.ac.nz/ml/weka/

Li, K., Lei, H., Khadiwala, R., & Chang, K. C. (2012). TEDAS: A Twitter-based Event Detection and Analysis System. 2012 IEEE 28th International Conference on Data Engineering, 1273-1276, doi: 10.1109/ICDE.2012.125.

Liang, P., & Dai, B. (2013). Opinion mining on social media data. in 2013 IEEE 14th International Conference on Mobile Data Management 2013, 2, 91-96, Available: 10.1109/MDM.2013.73.

Lin, C.-C. C.-J. (10 de September de 2019). LIBSVM -- A Library for Support Vector Machines. (September 10, 2019) Obtenido de www.csie.ntu.edu.tw/~cjlin/libsvm

M. M. Tadesse, H. L. (2019). Detection of Depression-Related Posts in Reddit Social Media Forum. IEEE, 7, 44883-44893. doi:10.1109/ACCESS.2019.2909180.

Maaoui, C. F. (2016). Automatic human stress detection based on webcam photoplethysmographic signals. Journal of Mechanics in Medicine and Biology, 16(4). doi:https://doi.org/10.1142/S0219519416500391

Manabu Torii, J.-w. F.-l. (Dec de 2015). Risk factor detection for heart disease by applying text analytics in electronic medical records. Journal of biomedical informatics, Volume 58, Supplement,(ISSN 1532-0464), S164-S170, https://doi.org/10.1016/j.jbi.2015.08.011.

Maria Khodorchenko. (2019). Distant supervision and knowledge transfer for domain-oriented text classification in online social networks,. Procedia Computer Science,, 156, 166-175,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2019.08.192.

Mariatta, S. a.-i. (07 de 09 de 2018). Avoid master/slave terminology. (https://bugs.python.org/issue34605) Obtenido de https://bugs.python.org/issue34605

Marouane Birjali, A. B.-H. (2016). A Method Proposed for Estimating Depressed Feeling Tendencies of Social Media Users Utilizing Their Data. Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016), 552, págs. 413-420.

Matykiewicz P, D. W. (2009). Clustering semantic spaces of suicide notes and newsgroups articles. In Proceedings of the BioNLP 2009 Workshop 2009 , (págs. 179-184).

McDonals, B. (10 de April de 2020). University of Notre Dame. (University of Notre Dame) Obtenido de https://www3.nd.edu/~mcdonald/

MEDEX. (2020). Medex.com.bd. Recuperado el 2020, de https://medex.com.bd/

ISSN-On line: 2414-4886

Nguyen, T. O. (21 de December de 2017). Using linguistic and topic analysis to classify sub-groups of online depression communities. Science+Business Media, 76(April 2017), 1573-7721, https://doi.org/10.1007/s11042-015-3128-x.

Nguyen, T. O. (21 de December de 2017). Using linguistic and topic analysis to classify sub-groups of online depression communities. Science+Business Media, 76(April 2017), 1573-7721,. doi:https://doi.org/10.1007/s11042-015-3128-x

Nguyen, T. V. (December de 2016). Textual cues for online depression in community and personal settings. (págs. 19-34). Springer, Cham.

Organizacion Mundial de la Salud. (30 de Enero de 2020). https://www.who.int/es/news-room/fact-sheets/detail/depression. Recuperado el 31 de Julio de 2021, de https://www.who.int/es/news-room/fact-sheets/detail/depression

Padilla-Navarro, C. P. (2016). Modelado de un sistema multi-agente aplicado a la predicción de la personalidad en Twitter. Research in Computing Science, (págs. 147-156.).

Phillips, D. P. (Jun de 1974). The Influence of Suggestion on Suicide: Substantive and Theoretical Implications of the Werther Effect. American Sociological Review, Vol. 39,, 340-354,DOI: 10.2307/2094294.

Python. (2020). https://docs.python.org/3/library/multiprocessing.html. (Python) Obtenido de https://docs.python.org/3/library/multiprocessing.html

RIP-PER. (2020). TheFreeDictionary.com. Recuperado el 2020, de https://www.thefreedictionary.com/ripper

Robert A. Fahey, T. M. (December de 2018). Tracking the Werther Effect on social media: Emotional responses to prominent suicide deaths on twitter and subsequent increases in suicide,. Social Science & Medicine,, 219, 19-29, ISSN 0277-9536, https://doi.org/10.1016/j.socscimed.2018.10.004

Rosas, F. J. (2018). Optimal Design in the Removal of Fluorescence and Shot Noise in Raman Spectra from Biological Samples. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Rubén Sánchez Acosta, C. M. (Dic de 2019). Heurísticas para Data Augmentation en NLP: Aplicación a Revisiones de Artículos Científicos. RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação, 34, 44-53. doi:10.17013/risti.34.44–53

Tadesse MM, L. H. (4 de Apr de 2019). Detection of Depression-Related Posts in Reddit Social Media Forum. IEEE Access., 7, 83-93. DOI: 10.1109/ACCESS.2019.2909180 .

Tharwat, A. (2020). Classification assessment methods. Applied Computing and Informatics. doi:doi.org/10.1016/j.aci.2018.08.003

Tom De Smedt, W. D. (2020). textblob. Obtenido de https://github.com/sloria/TextBlob/blob/dev/textblob/en/en-sentiment.xml.

Torous, J. L. (28 de 06 de 2018). Smartphones, Sensors, and Machine Learning to Advance Real-Time Prediction and Interventions for Suicide Prevention: a Review of Current Progress and Next Steps. Current Psychiatry Reports, 51, 1535-1645, DOI: https://doi.org/10.1007/s11920-018-0914-y.

Ueda M, M. K. (2017). Tweeting celebrity suicides: Users' reaction to prominent suicide deaths on Twitter and subsequent increases in actual suicides. Social Science & Medicine,, 189, Pages 158-166, ISSN 0277-9536, https://doi.org/10.1016/j.socscimed.2017.06.032

WHO, W. H. (7 de april de 2017). Depression and Other Common Mental Disorders. (Global Health Estimates) Obtenido de Organización Mundial de la Salud (World Health Organization): https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/

ISSN-On line: 2414-4886

Yang, Y. a. (2018). Large scale and parallel sentiment analysis based on Label Propagation in Twitter Data. In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), (págs. 1791-1798). New York, NY, USA. doi:10.1109/TrustCom/BigDataSE.2018.00270

ISSN-On line: 2414-4886