Detection of Suicidal Intent in Spanish Language Social Networks using Machine Learning

Kid Valeriano[1], Alexia Condori-Larico[2], and José Sulla-Torres[3]

Escuela Profesional de Ingeniería de Sistemas
Universidad Nacional de San Agustín de Arequipa, Perú

***Abstract*—Suicide is a considerable problem in our population, early intervention for its prevention has a very important role, in order to counteract the number of deaths from suicide. Today, just over half of the world's population uses social networks, where they express ideas, feelings, desires, including suicide intentions. Motivated by these factors, the main objective is the automatic detection of suicidal ideations in social networks in the Spanish language, in order to serve as a base component to alert and achieve early and specialized interventions. For this, a Spanish suicide phrase classification model has been implemented, since currently no related works in this language with a machine learning approach were found. However, there were some challenges in performing this task, such as understanding natural language, generating training data, and obtaining reliable accuracy in classifying these phrases. To construct our classification model, two opposite and popular types of phrase embeddings were chosen, and the most widely used classification algorithms in the literature were compared. Obtaining, as a result, the confirmation that it is possible to classify phrases with suicidal ideation in the Spanish language with good accuracy using semantic representations.**

Keywords*—Spanish; suicide ideation; embeddings; machine learning; phrases classification

I. Introduction

Suicide is related to severe depression, stress, and other psychological disorders that millions of people suffer from it annually. But only a fraction receives adequate treatment, the rest of the cases mostly end in suicide [1]. According to the World Health Organization [2], more than 800,000 people die from suicide each year. It is estimated that there are approximately 20 attempts for each death and also that in the last 45 years, the suicide rate has increased by 60%. In 2012, around 804,000 suicides occurred, which represents 1.4% of the total deaths in the world, and made it the 15th leading cause of death that year, with a rate of 11.4% suicides per 100,000 inhabitants, of this 14.5% are men and 8.2% are women. This significantly affects the young population between 15 and 29 years old. Furthermore, only a quarter of those who die by suicide have been in contact with health professionals before death. For this reason, specific measures of early identification and effective interventions are needed for all those cases with a tendency to suicide.

Over the past decade, the suicide risk of Hispanics in the US has steadily increased, stress is one of the main causes of suicidal ideation, despite the growth of the suicide rate, it is relatively little studied [3]. If we refer to Spanish-speaking countries such as Uruguay with 18.4 deaths from suicide per 100,000 inhabitants, it occupies the fifteenth place in the world list of suicide rates, being the first within Spanish-speaking countries. It is also followed in order based on high suicide rates by Cuba, El Salvador, Nicaragua, Bolivia, and Spain, according to World Bank statistics [4].

Today, thanks to technological advances, it is possible to cover a large number of people with suicidal tendencies. According to a study carried out by "We are Social" until 2017, of a total of 7,530 million people in the world, of these, 4,540 million have access to the Internet. Until January 2020, the number of people using social networks is of 3.8 billion, having a growth of more than 9% since 2019 [5], from the ubiquity of social networks, people continuously express their emotions through them. These statistics prompted us to better understand and learn about the behavior of people with suicidal tendencies [6].

Although there are few works related to the study of suicide as [7] and [8], where they focus on studying the factors and behaviors that led Spanish students to commit suicide. At the time of writing this work, no works were found that use machine learning algorithms for the automatic detection of suicidal thoughts in Spanish. But there are related works in the English language, some of these are [9] and [10], most of them analyze data from the social network Twitter® as it is a public source and has an API for data extraction. In most cases, the traditional Bag-of-Words based vectorization is used, which could sometimes mean the loss of the semantic relationships of the words that make up the suicide sentences. Although the training for semantic representations requires a large amount of data considering that in real problems, generally, a small amount of data is available initially. This difficulty has already been addressed with the existence pre-trained model of words semantic representation. In our case, a pre-trained model was used in the Spanish language with data from the same source, with a similar number of characters that make up the sentences and containing the different dialects that exist in some region or country. These characteristics of the trained model helped to obtain a better semantic representation of the sentences and consequently, better classification.

The main objective of this work is to detect suicidal intentions denoted by Spanish-speaking people. To achieve this objective, a model was implemented, and a human-annotated dataset was generated and is being made available for further study. Within the set of procedures, tests are carried out with different types of text vectorization, including those with

semantic relationships, and classifiers recommended in state-of-the-art are analyzed.

After this introduction, this article is organized as follows: The related works to this article are explained in Section II, the origin of the data in Section III, the methodology in Section IV, and the configuration of experimentation in Section V, and finally the conclusions and future work in Section VI.

II. Related Work

In recent years, the work about the detection of people with suicidal ideation reflected in social networks has increased considerably. Considering that the majority of the works are oriented to the English language. The first phrase classification works were related to the feeling classification, such as the work [11] that carried out supervised feelings classification experiments divided into three classes (negative, neutral, and positive). In this case, they make use of a set of characteristics based on the subjective lexicon, specific Twitter® characteristics, and more relevant words. They managed to achieve a superior result to several unsupervised approaches that use subjectivity lexicons.

Munmun, Michael, Scott, and Eric [1] detect and diagnose the depressive disorder in people. They use crowdsourcing to collect Twitter® users diagnosed with clinical depression, according to a standard psychometric instrument. Based on the publications made during the previous year of depression, the behavioral signals reported by these users were used to create a statistical classifier that provides estimates of the risk of depression, before the known onset. Simple classification methods were used in [9], where the automatic collection of tweets with suicidal ideation is performed, according to a glossary of terms used by people expressing themselves on the social network Twitter®.

On the other hand, the work Quoc and Tomas [12] were presented, where *Paragraph Vector* was proposed, an unsupervised algorithm to learns representations of characteristics of fixed length from variable length fragments of texts (sentences, paragraphs, and documents). As a result of this study outperformed traditional Bag-of-Words models as well as other techniques for text representations. It is where this representation is used in work [13], obtaining promising results in the detection of suicidal phrases, for the training of these representations, a large amount of data was used.

Linguistic Inquiry and Word Count (LIWC) and Machine Learning-based classification approaches were applied in [14], [10]. In work [14], suicide risk levels and emotional distress are evaluated through an online survey of the Weibo social network to measure suicide risk factors (suicidal ideations, depression, anxiety, and stress levels), to feed the data with publications from these users. They were subsequently analyzed and classified, depending on the characteristics of the language, Support Vector Machine (SVM) was used for classification. They conclude that Machine Learning together with Simplified Chinese-Linguistic Inquiry and Word Count (SC-LIWC) is good, but as a whole, they need to be optimized. Similarly, Jonathan, Pete, and Gualtiero [10] created a set of classifiers of 7 classes using different types of characterization considering sentiment analysis, LIWC, software textanalysis, and regular expression. It was obtained better results using the three types of characterization as one together with the Rotation Forest algorithm.

Long short-term memory (LSTM) is applied in [15], and in [16] together with Convolutional Neural Network (CNN), and both works aim to detect quantifiable signals according to suicide attempts and ideas. In [15] the outline of an automated system is described, applying Deep Learning and LSTM. Obtaining high precision by machine learning algorithms, recommended using in a planned detection system. In [16] they worked with the social network Reddit, where they use a combined LSTM-CNN model to evaluate and compare with other classification models. Regarding [17], studies are carried out on deep learning architectures such as LSTM, CNN, and RNN, based on data from annotated tweets (suicidal intention present or absent), and comparatively obtaining better results with models based on Contextual LSTM (C-LSTM). The CNN architecture can spatially encode tweets in a one-dimensional structure, and LSTMs are more robust to noise and more capable of acquiring long-term dependencies.

Regarding the most recent works [18] and [19], data from social networks were similarly combined with Machine Learning. In work with a more practical application [18] mental health states are predicted, with reasons for health professionals to use this model as a support to identify and make a diagnosis and treatment. The main objective for [19] was to analyze the relationships between cyber victimization and suicidal ideation in adolescent victims of cyberbullying. They voluntarily and anonymously filled out study instruments (scales), which evaluated and measured the level or frequency of bullying victimization in a defined context. Thanks to the structural equation model, cyber victimization was shown to be related to suicidal ideation. They obtained as result that indirect relationships have a more significant impact on suicidal ideation compared to the direct effects of cyber victimization.

As described, most works vary in precision according to text feature extraction and the use of some classifiers. Besides, the approach used depends on the available amount of data. From the works studied, we saw that the vast majority of works obtain acceptable precision when classifying suicide phrases. Considering that the majority of these works are implemented and tested with dataset in English, that is why in this work, we intend to use the semantic word representations little discussed in the literature. The aim is to validate whether these techniques are also obtaining good accuracy results with Spanish language data.

III. Dataset

A. Data Collection

For the implementation of this model of detection of suicidal tendencies, one of the main challenges was finding a publicly available data set, due to problems with privacy and anonymity. Another of the difficulties presented was the non-existence of works related to the Spanish language similar to ours. As a result of these needs, the motivation to create a new data set in the Spanish language is born to generate a model capable of predicting people with tendencies to suicide.

In the first place, we looked for related works that have investigated the terms most used by people with a tendency

to commit suicide. The work [6] provides a list of keywords and phrases in English used by people to express their suicidal wishes (a total of 62). For the adaptation of these keywords and phrases for our work, they were translated from English to Spanish with the help of bilingual people. Table I shows some, with their respective translation in Spanish. We can also see that in rows 5 and 19, the phrases and keywords (respectively) in English, can have the same meaning in the Spanish language.

TABLE I. KEYWORDS AND PHRASES IN ENGLISH (20/62), WITH THEIR TRANSLATION INTO THE SPANISH LANGUAGE, EXPRESSING SUICIDAL IDEATION FOR TWITTER® DATA COLLECTION.

N°	English	Spanish
1	Asleep and never wake	Dormir y nunca despertar
2	Just want to sleep forever	Solo quiero dormir para siempre
3	Kill myself	Matarme/Suicidarme
4	Life is so meaningless	La vida no tiene sentido
5	Tired of being lonely / Tired of being alone	Cansado de estar solo
6	Don't want to exist	No quiero existir
7	Life is worthless	La vida no vale nada
8	Don't want to live	No quiero vivir
9	My life is pointless	Mi vida no tiene sentido
10	My life is this miserable	Mi vida es así de miserable
11	My life isn't worth	Mi vida no vale
12	Want to be dead	Quiero estar muerto
13	Not want to be alive	No quiero estar vivo
14	Hate my life	Odio mi vida
15	Want to disappear	Quiero desaparecer
16	Hate myself	Me odio a mí mismo
17	Ready to die	Listo para morir
18	Really need to die	Realmente necesito morir
19	Suicidal / Suicide	Suicida
20	Isn't worth living	No vale la pena vivir

For the extraction and study of suicidal phrases, tweets were extracted using the official Twitter® API[1]. The keywords were used as input to obtaining a total of 100,000 tweets from October to December 2019. A pre-processing was carried out only to extract the textual phrases. It was noted that not only were there phrases of people who had actually expressed their suicidal wishes and thoughts but that they also existed expressions using any or all of the suicidal keywords. For example, in phrases expressing sarcasm, prevention campaigns, and parts of song lyrics, for this reason, in order to differentiate between the sentences that truly have a suicidal tendency and those that do not, it was necessary to make a manual annotation of each phrase with suspected suicidal notation.

B. Anotation Data

A data set of 2068 text sentences was generated and annotated by humans, considering separating the tweets based on the binary criterion (1 for tweets with a tendency to suicide and 0 for tweets without a tendency to suicide). That is, assign one of the two categories, and they are selected as suicidal in case of ambiguous texts. It should be noted that suicide-prone tweets are a clear indication of the user's suicidal intent. On the other hand, non-suicide criteria are the default category for all texts that show no evidence of suicide, such as sarcasm, news, song parts, or sentences using the suicide phrases/keywords.

This manual classification can be explained more clearly in the Table II, where some examples of tweets with a tendency to suicide and those that do not are shown. As a result of the annotation 498 tweets were annotated as Suicidal (24% of the dataset), while the rest were classified as Non-suicidal.

TABLE II. COMPARISON TWEETS WITH AND WITHOUT SUICIDAL IDEATION. PER ROW SHOWS THE USE OF A KEYWORD OR PHRASE IN DIFFERENT CONTEXTS (COLUMNS).

Tweets with suicidal ideation	Tweets without suicidal ideation
ya no quiero vivir *I do not want to live anymore*	un día sin tomar coca-cola ya no quiero vivir *one day without drinking coca-cola I don't want to live*
no me soporto me quiero morir odio ser yo *can't stand myself I want to die I hate being me*	queriéndome morir ni siquiera empecé y ya falté *wanting to die I didn't even start and I'm already missing*
La vida no tiene sentido sólo vivo para odiar a mi reflejo *life has no meaning I just live to hate my reflection*	mi vida sin audífonos no tiene sentido *my life without headphones is pointless*
me estoy ahogando en un mar de lágrimas tienes que ser fuerte me dicen pero yo ya no aguanto más *I'm drowning in a sea of tears you have to be strong they tell me but I can't take it anymore*	pero no puedo siento que muero me estoy ahogando sin tu amor como quisiera (canción) *But I can't feel I'm dying I'm drowning without your love as I wish (song)*

IV. METHODOLOGY

The present work mainly consists of two main modules, in the Fig. 1, we can see the training module that is only performed once, and the prediction module is used in each consultation made by a phrase with suspected suicide intention. As can be seen in each module, almost the same three components are executed: A) data pre-processing, B) data vectorization, and C) training or inference using classification model, which is detailed below.

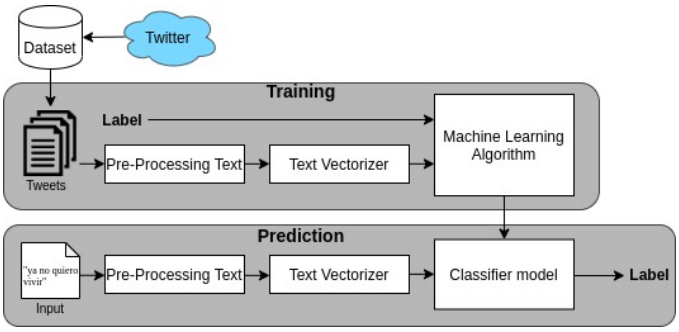

Fig. 1. Architecture of the Proposed Model of Automatic Detection of Suicidal Ideation in the Spanish Language.

A. Pre-Processing

This step is mainly focused on treating the text of the extracted sentences, how to eliminate redundant phrases, remove some noise to improve the accuracy of the model. Different procedures are applied, among which stand out:

1) Removal of URLs, special characters, and numbers: Some tweets, for the most part, contain special characters. For example, these may denote admiration, question, or some reference to a website. These characters and numbers are extracted to remove noise for both vector representation and the classification model.

[1]https://developer.twitter.com/

2) Tokenization: This process is in charge of separating the phrases into words(tokens). This separation of tokens return a words list, this procedure is the most important and vital for subsequent processes.

3) Anonymization of tweets that contain names: The identity of the person who made the post and the names mentioned in the collected tweets are anonymized. This procedure was done in order to avoid harming or exposing someone without their permission.

4) Remove of StopWords: In natural language, there are words that by themselves do not add any meaning in a sentence. These empty words generally have a high frequency of use. In the Spanish language, empty words tend to be articles, conjunctions, and pronouns.

B. Text Vector Representation

The method developed in this work is mainly based on natural language processing techniques and Machine Learning. Machine Learning algorithms operate in an attribute value configuration, where, in most cases, these attributes are associated with a numeric value. For this reason, it is necessary to establish a way of representing the elements of natural language as attributes understood by some Machine learning techniques. Today, this is commonly done with a vectorization technique, where words or sentences are represented in a vector space. The traditional Bag-of-Words approach [20] consists of transforming the text into a set of tokens. Such values can be simply Boolean variables, indicating the word presence or the word absence in the text. Also, they can be numeric, computed from a frequency measure of the words.

1) TF-IDF: The vectorization of documents using the Term Frequency-Inverse Document Frequency (TF-IDF) [21] measures the importance of the words in a document. Computing the frequency that a word appears in it (Term-Frequency *TF*), but taking into account the existence of very frequent words in the documents (*e.g.*, words such as 'and', 'so', *etc*) to reduce the weight of them. Thus, the relevance of a word increases proportionally to the frequency, but it is offset by the frequency of the word in the entire corpus (IDF – the Inverse Document Frequency). The resulting TF-IDF value is computed from the product of these two measures, as showed in the Equation 1, where the final value is normalized between 0 and 1, t is the term, and d is the document.

$$\mathrm{tfidf}_{t,d} = \mathrm{tf}_{t,d} \times \log\left(\frac{N}{df_t}\right) \tag{1}$$

where

$$\mathrm{tf}_{t,d} = \frac{\text{Number of times t appears in a document d}}{\text{Total number of terms} \in d}$$

and

$$\mathrm{df}_t = \text{Number of documents with term t } \in \text{it.}$$

2) Word Embeddings: The semantic aspects of words in a text can vary depending on the context considered. Therefore, assuming that the values associated with an attribute can range from 0 to 1, the words "king" and "queen" in a royalty context must have values close to each other (close to 1). On the other hand, in a gender context, the values associated with these same two words must be distant from each other (close to 0), since they deal with different genres. This type of semantic vector representation makes it easy to represent a word in different contexts and assign appropriate values to those attributes.

To work around this problem, it has become a standard practice to use a numeric vector to represent the tokens extracted from texts. Such representations are known as *embeddings* [22], and are usually defined as d-dimensional vectors, learned automatically from several texts. Thus, each dimension of the vector may reflect a distinct context, and the value associated with the dimension is learned accordingly. At the end of the learning process, it is expected that the words with the closest semantics will be mapped to close positions in the vector space.

The most commonly used implementations of such techniques are *Word2vec* [23] (which implements the Skip-gram [24] and CBOW [25] algorithms) and GloVe [26], all of them using neural networks with a hidden layer to obtain the learned representations. The vectors referring to the attributes of words are extracted from the weights of the hidden layer, making this form of learning to receive the name of neural language models [27]. The purpose of the neural model learning is to maximize the value of:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p\left(w_t | w_{t-k}, \ldots, w_{t+k}\right) \tag{2}$$

Where w_i represents a word in a sequence of words $w_1, w_2, \ldots, w_T$, and $w_{t-k}, \ldots, w_{t+k}$ represents a window of words of size t, where $w_{t-k}, w_k, w_{t+k} \subset w_1, w_2, \ldots, w_T$. Each prediction task is usually defined as a *softmax* classifier, as follows:

$$p\left(w_t | w_{t-k}, \ldots, w_{t+k}\right) = \frac{e^{y_{wt}}}{\sum_i e^{y_i}} \tag{3}$$

Where y_i is the non-normalized logarithm of the word probability i be the output of the model, calculated as:

$$y = b + Uh\left(w_{t-k}, \ldots, wt + k; W\right) \tag{4}$$

Where U and b are the weights of the classifier and h is either the concatenation or the mean of the word vectors in W. The neural language models are trained with the gradient descent optimization method, where the gradient is obtained from the Backpropagation algorithm [28].

C. Word2vec-Mean for Sentence Embeddings

Similar to the Word2vec representation, there is an evolved representation (Paragraph Embedding[29]), which manages to represent phrases and documents of different sizes in vectors with the same dimension. This type of presentation is not addressed in this work due to the amount of data necessary for training that is not currently available for this work.

Another way to approach this principle is to map each word that makes up the phrase in the pre-entered word2vec model. Subsequently, the average of all the mapped words is

calculated, as in Equation 5. The resulting vector is the vector representation of the phrase.

$$Phrase\_vec(t_i) = \frac{1}{n}\sum_{j=0}^{n} Enc\_word(w_{ij}, m\_w2v) \quad (5)$$

Where n is the number of words in the phrase, w_{ij} is a word to mapping in the word2vec model matrix m_w2v, Let Enc_word: $(w_{ij}, w2v) \rightarrow \vec{w}_{ij}$, where $\vec{w}_{ij} \in$ m_w2v is an encoding function, mapping word tokens to their vector representations for i^{th} words, where i $\in [0, n]$.

D. Classification Algorithms

Machine learning algorithms are categorized into several types, such as supervised, unsupervised, semi-supervised learning, and reinforcement learning. To differentiate a tweet with a tendency to suicide from a non-suicidal tweet, it is necessary to know that it is a supervised classification problem. In conclusion, to train the model, there must be a priori output for each input with the category to which it belongs.

Detecting people with suicidal tendencies is a binary classification problem. For each tweet $t_i \in D$, the data is noted by a binary variable $y_i \in \{0, 1\}$, where $y_i = 1$ denoting that the tweet t_i has a suicide intention and $y_i = 0$ the opposite. The classifier, after training, must determine if any of your sentences t_i have any structure or any word/phrase that denotes the existence of suicidal thought.

The vectorization of the tweets is a previous step to the training of our model, and later the classification algorithms were analyzed. The following steps are performed in each tweet:

$SentenceEmbeddings$: A space vector is generated for each phrase (TF-IDF or Word2Vec-Mean model).

$Classification$: Finally, once the representation of the tweets has been obtained, they feed the model. Classification algorithms such as Support Vector Machine (SVM) [30] and Logistic Regression (LR) [31] were used and compared.

V. Experimental Setup

For the construction of the classification model of all the collected tweets, exactly 2068 tweets were selected to be annotated. In total, 498 were annotated as tweets with a suicidal tendency and 1570 as tweets without risk of suicide. As they were unbalanced categories, 500 tweets were randomly taken from the 1570 non-suicide tweets; that is, 500 *non-suicidal* tweets and 498 *suicidal* tweets were made available to balance the database. Subsequently, 20% of the balanced data were assigned for tests and 80% for the training of the algorithms.

A. Exploring Data

In Fig. 2, two word-clouds were graphed from the frequent words used in the two categories a) Non-suicidal word-cloud and b) Suicidal word-cloud. In the two word-clouds, the most used word is *QUIERO* (want), which in Spanish is a word that expresses desire. Understandably, all kinds of wishes are displayed on a social network because it is a medium where users express their emotions. Besides, we can see the words most used in the suicidal category have more meaning and relationship than the non-suicidal word-cloud.

(a) Non-suicidal word cloud (b) Suicidal word cloud

Fig. 2. Word clouds with most frequent words present in a) non-suicidal, and b) suicidal ideation.

In Fig. 3, was graphed using the Scattertext framework [32], where the X-axis and Y-axis indicate the term frequency no-suicidal and suicidal texts, respectively. For instance, the upper-left area shows the terms frequently occurring suicidal texts, while the lower-right area shows the frequent terms in non-suicidal texts. In general, the visual shows an intersection where the majority of words are used in both categories. This happens because both categories are the result of searching for phrases related to suicide. Besides, we can also observe that there are words that make a distinction between the not-suicidal and suicidal. The terms used most exclusively and frequently in suicidal include 'desaparecer', 'existir', 'despertar_nunca', and 'matarme'.

B. Words Embedding Configuration

As mentioned above, little data is available, making it impossible to train our semantic word representation model. For this reason, a pre-trained model is used to achieve a better understanding of the Spanish language. This pre-trained model manages to capture the dialects and slang used in each Spanish-speaking region or country. Thanks to the capture of these relationships, our vocabulary is not closed, that is, the words not seen in the classifier training will have some semantic relationship with any word of the existing vocabulary. These vectors have a dimensionality of 100, and a vocabulary of X words.

C. Phrases Embedding Configuration

For the representation of tweets (phrases) in this work, two types of vectorization were performed. For the TF-IDF, *inverse-document-frequency* was used as the only input parameter. Where the vocabulary number of the training data is equal to the dimensionality, in this case, our vector has 2306 dimensions. For Word2vec-Mean the configuration is by default, each vector has the dimension of the pure Word2Vec vectors, that is, in our case, each resulting vector of the sentence has 100 dimensions.

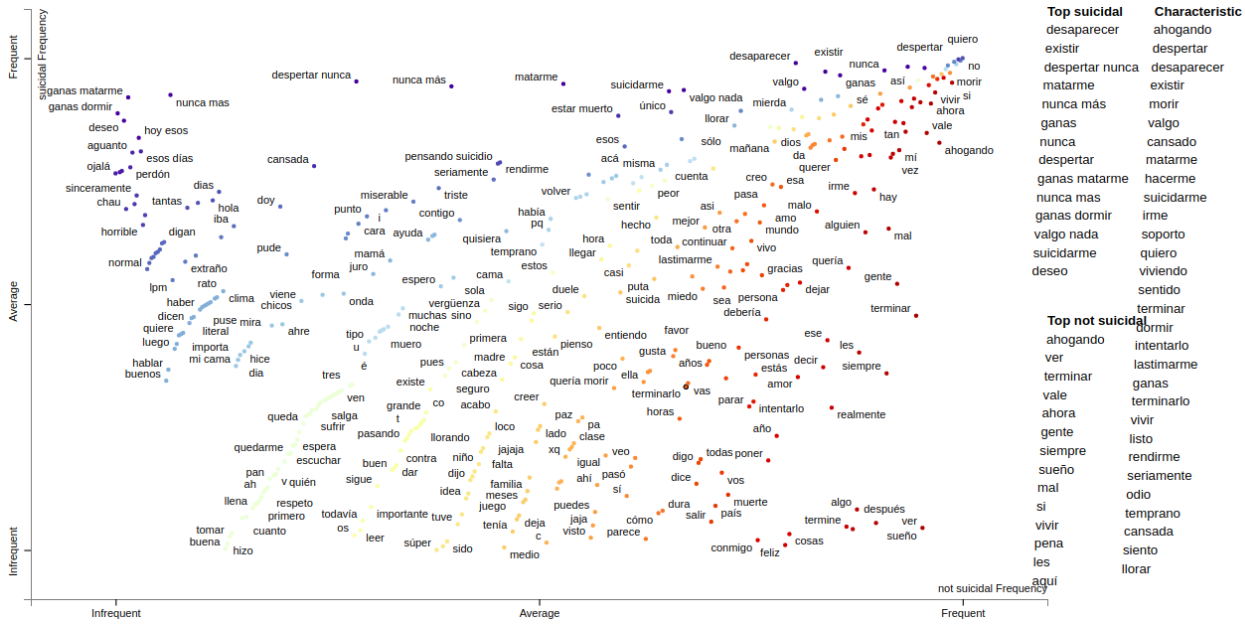

Fig. 3. Graph of unique terms and shared terms between Non-suicidal and Suicidal-ideation.

D. Classifier Details

In this work for the classification, the two most used classification algorithms in the literature were taken with excellent results in the texts classification. The classifiers were used with the default settings, both the SVM without kernel and binary Logistic Regression (LR) classifier.

E. Implementation Details

The following technologies were used for the implementation model. Also, both the dataset and the code [2] are available for further study.

- Python[3] programming language.
- The Sklearn [33] and Gensim [34] libraries are used for vectorization and classification.
- Joblib[4] library to save trained binary models.
- Flask [35] library to lift the server.
- The pre-trained Word2Vec model in Spanish[5].

F. Evaluation Metrics

The different forms of vectorization and classification algorithms are compared to each other in terms of the following metrics:

1) Precision: $\frac{t_p}{t_p+f_p}$

2) Recall: $\frac{t_p}{t_p+f_n}$

3) F1 score: $\frac{2t_p}{2t_p+f_p+f_n}$

4) Accuracy: $\frac{t_p+t_n}{t_p+t_n+f_p+f_n}$

Where t_p is the number of true positives, t_n is the number of true negatives, f_p is the number of false positives, and finally f_n is the number of false negatives.

G. Results Model Classification

Table III shows the results of the two classification algorithms based on the two types of vectorization addressed, conforming to different suicide tweet detection models in terms of the evaluation metrics. In general, the different configurations addressed in this work obtain a considerable good result in the classification of tweets in Spanish. The two main rows show the results for the two classification algorithms using TF-IDF and Word2Vec-Mean vectorization as a basis. It can be seen that obtaining greater accuracy gain depends of vector representation type. Word2Vec-Mean gets a better profit since the word count method fails in situations where the pattern to be extracted must take into account the semantics and not only the lexical aspects. Furthermore, it is observed that the use of TF-IDF representation does not differ much in its accuracy from the type of classification algorithm used. Contrary to the vectorization Word2Vec-Mean obtains a better accuracy using the Logistic Regression classifier, and it was possible to obtain the best model with a maximum accuracy of 0.79.

TABLE III. RESULTS IN TERMS OF METRICS OF THE TWO CLASSIFICATION ALGORITHMS LR AND SVM BASED ON THE TF-IDF AND WORD2VEC-MEAN VECTORIZATION TYPES.

	Model	Accuracy	Precision	Recall	F1 Score
LR	**TF-IDF**	0.72	0.74	0.72	0.73
	W2V-m	**0.79**	0.79	0.79	0.79
SVM	**TF-IDF**	0.71	0.76	0.71	0.72
	W2V-m	0.74	0.76	0.74	0.75

[2]https://github.com/kvvaldez/suicidio
[3]https://www.python.org/
[4]https://joblib.readthedocs.io
[5]https://www.datos.gov.co/Ciencia-Tecnolog-a-e-Innovaci-n/WORD2VEC-Twitter-Espa-ol-para-Latinoam-rica-Espa-a/79c6-2d7z

H. Implementation Experiment

The implementation of the model is a service that, in this case, is consumed by a web application, where the two types of vectorization are shown for their performance comparison using the Regression logistic classification algorithm that obtained better accuracy. Three examples of operation are described below, where the entries are phrases used in the model validation phase, in order to understand the model's behavior better.

To put a degree of difficulty to the model, we start from the following phrase "I don't want to live in this heat anymore", this phrase is more likely to be classified as suicide-prone for having words related to that intention, for example, in Fig. 4, is classifying a part of the phrase as "I do not want to live anymore" both classifiers classify well as not suicidal, with confidence for TF-IDF 70.8% and Word2Vec 75.3%.

Contrary, if we classify the complete phrase, "I don't want to live in this heat anymore" as shown in Fig. 5. For TF-IDF, it is a phrase with a tendency to suicide with the confidence of 65% for having words related to suicide. Opposite to Word2Vec-Mean, it is a phrase without a tendency for suicide with the confidence of 64.7%, considering that this phrase does not express any suicide intent. These peculiarities may explain why the classification using the Word2Vec-mean vectorization gave better accuracy.

Finally, in Fig. 6, the case where the phrase "I just want to sleep hugging you" is classified as non-suicidal with the confidence of of 71.4% for TF-IDF and 78.2% for Word2Vec-mean.

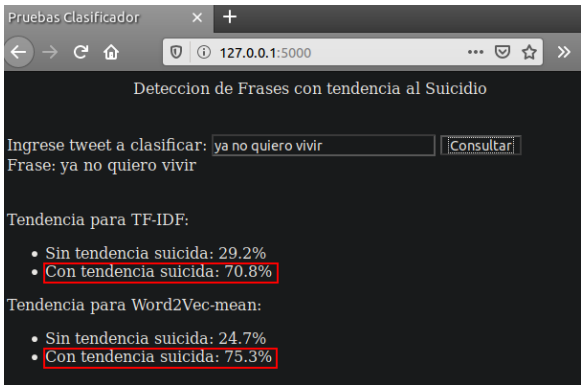

Fig. 4. Suicidal-ideation classification example

VI. Conclusion and Future Works

This work presents the detection of suicidal tweets in the Spanish language and makes a comparison of the different settings to obtain an optimal model. Discuss the importance of using semantic representations to improve the classification of suicide phrases. It also explains the challenges in building a suicide classifier other than the English language and generating training data. This work concludes that a Spanish-language tweet classification model can be constructed with relatively good accuracy. Considering the use of trained semantic representations, it has a better performance together with the logistic regression classifier. Furthermore, it is concluded that the use

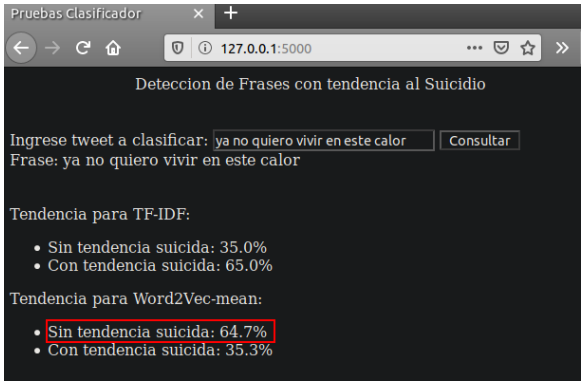

Fig. 5. Word2Vec-Mean and TF-IDF classification model comparison

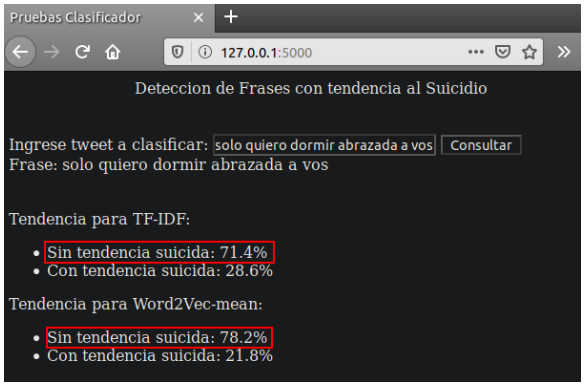

Fig. 6. Non-suicidal classification example

of the procedures and algorithms have some differences, but like the English language, good results can be obtained for the Spanish language.

For future work, this work can be extended in the first place to improve the accuracy of the model because it is an incremental model, mainly in the generation more considerable amount of data for better training. With a large amount of data, it possible to try using architectures based on deep learning algorithms to improve the model. Furthermore, this work can be extended to different environments, not only to the use of data from Twitter® or the Spanish language.

References

[1] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[2] Organización Mundial de la Salud. Suicidio. www.who.int/es/news-room/fact-sheets/detail/suicide, 2019.

[3] Caroline Silva and Kimberly A Van Orden. Suicide among hispanics in the united states. *Current opinion in psychology*, 22:44–49, 2018.

[4] Distintas Latitudes. Suicidio en américa latina: esta es la situación en siete países de la región. distintaslatitudes.net/explicadores/suicidio-jovenes-en-america-latina, 2018.

[5] We are social. Digital 2020: 3.8 billion people use social media. wearesocial.com/blog/2020/01/(digital-2020-3-8-billion-people-use-social-media, 2020.

[6] Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications*, 73:291–300, 2016.

[7] Maria Jesús Blasco, Pere Castellví, José Almenara, Carolina Lagares, Miquel Roca, Albert Sesé, José Antonio Piqueras, Victoria Soto-Sanz, Jesús Rodríguez-Marín, Enrique Echeburúa, et al. Predictive models for suicidal thoughts and behaviors among spanish university students: rationale and methods of the universal (university & mental health) project. *BMC psychiatry*, 16(1):122, 2016.

[8] Carolina Lagares-Franco, José Almenara-Barrios, Cristina O'Ferrall-González, Pere Castellvi-Obiols, Andrea Gabilondo, María Jesús Blasco-Cubedo, Andrea Miranda-Mendizábal, Oleguer Parés-Badell, José Antonio Piqueras, Miquel Roca, et al. Medidas de frecuencia utilizadas en estudios de cohortes para evaluar el comportamiento suicida en jóvenes (12-26 años): Una revisión sistemática. *Revista de Psiquiatría y Salud Mental*, 12(4):213–231, 2019.

[9] Amayas Abboute, Yasser Boudjeriou, Gilles Entringer, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Mining twitter for suicide prevention. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 250–253. Springer, 2014.

[10] Pete Burnap, Walter Colombo, and Jonathan Scourfield. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 75–84, 2015.

[11] Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics, 2013.

[12] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[13] Victor Ruiz, Lingyun Shi, Wei Quan, Neal Ryan, Candice Biernesser, David Brent, and Rich Tsui. Clpsych2019 shared task: Predicting suicide risk level from reddit posts on multiple forums. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 162–166, 2019.

[14] Qijin Cheng, Tim MH Li, Chi-Leung Kwok, Tingshao Zhu, and Paul SF Yip. Assessing suicide risk and emotional distress in chinese social media: A text mining and machine learning study. *Journal of medical internet research*, 19(7):e243, 2017.

[15] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860, 2018.

[16] Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7, 2020.

[17] Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 167–175, 2018.

[18] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2019.

[19] Begoña Iranzo, Sofía Buelga, María-Jesús Cava, and Jessica Ortega-Barón. Cyberbullying, psychosocial adjustment, and suicidal ideation in adolescence. *Psychosocial Intervention*, 28(2):75–81, 2019.

[20] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[21] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. 18(11):613–620, 1975.

[22] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[24] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4, 2006.

[25] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[27] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[28] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

[29] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[30] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[31] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[32] Jason S. Kessler. Scattertext: a browser-based tool for visualizing how corpora differ. 2017.

[33] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[34] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[35] Gareth Dwyer, Shalabh Aggarwal, and Jack Stouffer. *Flask: Building Python Web Services*. Packt Publishing Ltd, 2017.