Contents lists available at ScienceDirect

Internet Interventions

Bootstrapping semi-supervised annotation method for potential suicidal messages

Roberto Wellington Acuña Caicedo [a,c,*], José Manuel Gómez Soriano [b], Héctor Andrés Melgar Sasieta [c]

[a] Information Technology Undergraduate Program, Universidad Estatal del Sur de Manabí, Jipijapa, Manabí, SENESCYT Scholarship Holder, Ecuador
[b] iaLife CEO and consultant at IT4IP, service provided to EUIPO, Spain
[c] Department of Engineering, Computer Engineering Section, Graduate School, Pontificia Universidad Católica del Perú, Lima, Peru

ARTICLE INFO

ABSTRACT

The suicide of a person is a tragedy that deeply affects families, communities, and countries. According to the standardized rate of suicides per number of inhabitants worldwide, in 2022 there will be approximately about 903,450 suicides and 18,069,000 unconsummated suicides, affecting people of all ages, countries, races, beliefs, social status, economic status, sex, etc. The publication of suicidal intentions by users of social networks has led to the initiation of research processes in this field, to detect them and encourage them not to commit suicide. This study focused on determining a semi-supervised method to populate the Life Corpus, using a bootstrapping technique, to automatically detect and classify texts extracted from social networks and forums related to suicide and depression based on initial supervised samples. To carry out the experiments we used two different classifiers: Support Vector Machine (SVM) (with Bag of Words (BoW) features with and without Term-Frequency/Inverse Document Frequency (Tf/Idf), as a weighted term, and with or without stopwords) and Rasa (with the default feature extraction system). In addition, we performed the experiments using five data collections: Life, Reddit, Life+Reddit, Life_en, and Life_en + Reddit. Using the semi-supervised method, we managed to increase the size of the Life Corpus from 102 to 273 samples with texts from the social network Reddit, in a combination Life+Reddit+BoW_Embeddings, with the SVM classifier, with which a macro f1 value of 0.80 was achieved. These texts were in turn evaluated by annotators manually with a Cohen's Kappa level of agreement of 0.86.

1. Introduction

The world had 7925 billion people (Worldometer, n.d.) in February 2022. From that population, 5.31 billion are unique mobile phone users (67.1%), 4.95 billion are internet users (62.5%), and 4.62 billion are active social media users (58.4%) (We Are Social, 2022). Up to 2020, humanity generated 44 zettabytes of digital data and will generate 175 zettabytes until the year 2025, most of it with information created and consumed by users through digital television, interaction with social networks, sending images and videos from camera phones between devices through the internet (IDC Corporate USA, 2012).

From this entire digital universe, 33% of the information (over 13,000 exabytes) is tagged data and 77% is untagged data (IDC Corporate USA, 2012). This data offers great opportunities for data analytics, through different types of machine learning: supervised

learning (Akpınar et al., 2019), unsupervised learning, or semi-supervised learning (Raschka and Mirjalili, 2019; Reagan et al., 2017).

Today, the explosive growth of online social networking services has changed the way people work and share their opinions, ideas, and views (Liang and Dai, 2013; Weng et al., 2010; Liao et al., 2013), regardless of their geographical location or physical limitations (Al-Garadi et al., 2016). Therefore, social networks can be extremely useful for various real-life applications, such as marketing (Leung and Chung, 2014), applications in an e-learning environment (Choudhury and Pattnaik, 2020), or discovering opinions about a particular product (Mircoli et al., 2017), knowing the geographical location, food preferences (Peschel et al., 2019), hobbies, favorite store, political tendencies (Stefanova and Kiryantsev, 2019; Rodriguez et al., 2012) and even identifying socially dangerous people (Stefanova and Kiryantsev, 2019), or those with suicidal intentions (Desmet and Hoste, 2013; O'Dea et al., 2015; Luxton

et al., 2012; Velupillai and Hadlaczky, 2019; Zhang et al., 2014; Braithwaite et al., 2016; Egmond and R. D.-C. T. J. of Crisis, 1990).

In the framework of suicide, the World Health Organization (WHO) in their report, "Suicide prevention: A global imperative", of the Mental Health Action Plan 2013–2020, published in 2014, estimated 804,000 suicide deaths occurred worldwide in 2012, representing an annual global age-standardized suicide rate of 11.4 per 100,000 people (903,450 suicides approximately in 2022). For each person who commits suicide, 20 more try to do so. In consequence, the WHO recognizes suicide as a global public health priority, which affects not only rich countries (3 men to each woman) also poor and middle-income countries (1.5 men to each woman) as well (World Health Organization, 2014). Suicide affects all age groups and is highest in persons aged 70 years or over (Jeong et al., 2020; Chang et al., 2018; Santini et al., 2015). Globally suicide is the second leading cause of violent death in the 15–29 years age group (World Health Organization, 2014; Sweeney et al., 2015). Suicide not only involves the individual in a personal way but deeply affects family members and close friends (Cerel et al., 2008; Silenzio et al., 2009).

According to this reality in 2013, in the framework of the 66th World Health Assembly, the member states of the WHO pledged, among other issues, to reduce national rates of suicide by 10% before the year 2020. To accomplish this goal, they agreed to develop and put into practice comprehensive national suicide prevention strategies, strengthening their information systems, scientific data, research, and university collaboration on mental health, and paying particular attention to groups with the highest suicide risk, such as homosexuals, lesbians, bisexuals, transsexuals, young people, refugees, migrants, and any other vulnerable group (World Health Organization, 2014; Silenzio et al., 2009). However, despite the good intentions of these organizations, their objectives have not been reached.

In general, a suicide victim goes through a period of deep personal suffering, often in silence, before making the fateful decision to end his or her own life, so predicting that someone will commit suicide has been an impossible task (Goldstein et al., 1991; Hughes, 1995; Large and Ryan, 2014; Large and Nielssen, 2012; Paris, 2006). But it is possible to detect factors that contribute to suicide risk, using standard clinical tools operated by well-trained clinicians (Beck et al., 1975; Beck et al., 1979). In addition, computer science, specifically Natural Language Processing (NLP), offers the opportunity to understand indicators of suicidal thoughts through the interaction between computers and human language (Larsen et al., 2015; Pestian and Grupp-Phelan, 2016) when these are expressed in written and spoken forms.

In this framework, social networks have provided researchers with new ways to use automated methods to analyze human language (Lacson and Khorasani, 2011), by analyzing the sentiments expressed therein, in written or spoken form by users (Abbasi et al., 2014; Girju and Moldovan, 2002; Asghar, 2016; Cole et al., 2006). Thus, by using automated methods, is possible to better understand an individual's thoughts, feelings, beliefs, behavior, and personality (Schwartz and Ungar, 2015), and to be able to successfully identify suicide notes in newsgroups and social media (Hernandez and Pontes, 2014; Huang et al., 2007; Matykiewicz et al., 2009). Machine learning algorithms have been shown to distinguish between notes written by people who died by suicide and simulated suicide notes better than mental health professionals (71% vs 70%) (Pestian and Matykiewicz, 2008). In addition to suicide notes, microblogging data have been used to build machine learning models to identify users with suicidal sentiments with 90% accuracy (Zhang et al., 2014).

The success of sentiment analysis (Pang and L. L.-F. and T. in Information, 2008; Pang and L. L.-P. of the 42nd annual meeting on, 2004; Lai et al., n.d.; Birjali et al., 2017a) relies heavily on the quality of implicit and unspecified information in the data (Chen and De Tseng, 2011; Fielding et al., 2008) that can be extracted from the large stream of information currently available (Birjali et al., 2017b; Barnes, 2007; Lieberman, 2014), which provides a better understanding of thoughts,

feelings, beliefs, behaviors, and personality of individuals (Schwartz and Ungar, 2015). With the information collected, compiled, and correctly annotated from oral or written texts, a corpus can be formed (Llisterri, 1999), which can be supervised with labeled data, which is more accurate, but generally consumes a lot of human and computer resources (Bentivogli and Pianta, 2005; Akpınar et al., 2019), or semi-supervised with unlabeled or unknown structure data, which reduces manual work and processing time (Akpınar et al., 2019; Raschka and Mirjalili, 2019), even though, the lower quality of the collected data and the accuracy of the annotations must be taken into account (Ren and Matsumoto, 2016).

Through a process of collecting information from social networks or other sources, evaluating it through agreements (Ben-David, 2008; Vieira et al., 2010; Vioules et al., 2018; Tapia et al., 2018; O'Dea et al., 2015; Hallgren, 2012; Fu et al., 2013; Canales and Strapparava, 2016), performed by a small group of experts (Bontcheva et al., 2013; Alameda-Pineda et al., 2013; Karimzadeh and MacEachren, 2019) or by multiple groups of experts through crowdsourcing (Ling et al., 2016; Karimzadeh and MacEachren, 2019), a gold standard corpus can be generated (Gundlapalli et al., 2013; Scheible et al., 2011; José, 2017; Karimzadeh and MacEachren, 2019) and used as training to add new data (Silveira et al., n.d.). Different research groups have created gold standard corpora in other knowledge areas, such as suicidality (Cremades et al., 2017a), medication abuse (O'Connor et al., 2020), depressive symptom and acquired psychosocial stressors (Mowery, 2017), Philippine species (Nguyen et al., 2019), cyberbullying (Van Hee et al., 2018), agriculture (Amorim et al., 2019) to gain a proper understanding of the relevant NLP technology to take full advantage of its capabilities (Lu, 2014).

The annotations that are made for the creation of the different corpora are generally supervised, and later in the data expansion process, these annotations are made semi-supervised, or unsupervised (Mowery, 2017; O'Connor et al., 2020; Van Hee et al., 2018; Halike et al., 2020; Amorim et al., 2019; Du et al., 2017; O'Dea et al., 2015), depending on the type of experiments carried out. The data from the experiments carried out are usually focused on a single data source: from Twitter (Jashinsky et al., 2014; Mowery, 2017; O'Dea et al., 2015; Purver et al., 2012; Wu et al., 2019), Weibo (Huang et al., 2014; Zhang et al., 2015), Netlog (Desmet and Hoste, 2018), other microblogs (Guan et al., 2015), and more than one sources (Ling et al., 2016; Cremades et al., 2017a).

This article proposes, a semi-supervised learning approach to classify potential suicide messages on social networks automatically. Our goal is to assign texts to the suicide and non-suicide category in the Life Corpus. It would be a starting point towards the semi-automatic annotation of Corpus data to detect suicide messages. Semi-automatic annotations will ease the annotation process and reduce the workload of the annotation team in terms of time and resources invested.

This document is organized as follows: Section 2 reviews the research on the use of automatic corpus annotation, Section 3 explains the methodology and resources used to develop this work. Subsequently, in Section 4 the results are presented, in Section 5 the results are discussed; finally, in Section 6 conclusions are made and opportunities for future works are proposed.

2. Related research

The research carried out on "semi-supervised corpus annotation" shows some developed works, among them the one created by Gupta et al. (Gupta et al., 2018) who studied the problem of mentions of Adverse-Drug-Reaction (ADR) in social networks. They used deep neural networks in their research process, specifically a class of Recurrent Neural Network (RNN) that is long-term memory, based on which they proposed a new RNN model based on semi-supervised learning, which can take advantage of data without tag present on social media. With the semi-supervised ADR extraction method, they obtained an f-measure of 0.75.

Brum et al. (Brum and Nunes, 2018) worked in a framework based on semi-supervised learning to extend CasSUL Corpus with unlabeled data. In the experiments, six characteristics were used: a bag of words, negation words, emoticons, emojis, the lexicon of feelings, and the label of part of the voice. As well as the classification algorithms Support Vector Machines (SVM), Naïve Bayes, Logistic Regression, Random Forest, Decision Trees, and Multilayer Perceptron, with which the best results were obtained with the combination of BoW + negation words + emoticons + emojis and feature selection using 200 estimators, entropy as a criterion and without maximum depth, with an f-measure of 0.62.

O'Dea et al. (O'Dea et al., 2015) worked on detecting whether the level of concern for a suicide-related post on Twitter could generate a training corpus for automatic learning models and implement an automated computer classifier that could replicate the accuracy of the human coders. The data for the experiments were collected from the social network Twitter, and the overall agreement rate among the human coders was 0.76. The classifiers used were the Support Vector Machine and the Logistic Regression methods. The algorithm with the best performance was Support Vector Machine with Tf/Idf without filter, obtaining an f-measure of 0.67.

Gómez (Gómez, 2014) worked on the creation of Life Corpus which is a bilingual text corpus (English and Spanish) oriented to detecting suicide ideation. This corpus was constructed retrieving texts from several social networks. Its quality was measured using mutual annotation agreement, obtaining a moderate agreement Cohen's Kappa of 0.52 in four categories, three risk classes (Possible, Urgent, and Immediate) and one the not risk. Given the imbalance of that corpus by the four categories and the small number of samples for each category, these were grouped into only two: Risk and No risk, to achieve better results in the development of different experiments carried out. In these experiments, it was decided to determine which default classification algorithm of the Weka machine learning and data mining software (Hall et al., 2009) achieved better performance when training the Life Corpus texts, using features such as Part of Speech, Wordnet Synset, and reading all numbers by a keyword, achieving the best results the KStar algorithm with the ROC area metric 0.81 and the f-measure of 0.70.

It should be noted that the quality of the corpus developed by O'Dea (O'Dea et al., 2015) and Gómez (Gómez, 2014) was determined by evaluating their texts through an agreement between annotators using Cohen's Kappa method (Cohen, 1960a), which is given by Equation one. While in their articles Gupta (Gupta et al., 2018) and Brum (Brum and Nunes, 2018) do not mention if they used any measure to evaluate the quality of the texts.

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \qquad (1)$$

In previous works with the *Life Corpus* (Life Corpus, n.d.; José, 2017), different experiments have been carried out (Caicedo et al., 2020; Parraga-Alava et al., 2019). Access to the Life Corpus is free under a Creative Commons license (Life Corpus, n.d.). Therefore, the experiments carried out with this corpus can be replicated or improved.

3. Methodology

The objective of this research was to determine a semi-supervised method to populate the Life Corpus, using a bootstrapping technique. With this, we have tried to improve the automatic detection and classification of texts extracted from social networks and forums related to suicide and depression based on the Life Corpus.

In previous works (Caicedo et al., 2020; Parraga-Alava et al., 2019) with Life Corpus, authors used machine learning techniques to systematically analyze all possible combinations of textual characteristics. They tested 28 supervised classifier algorithms using different corpus features. The study concluded that it would be interesting to increase the corpus to improve the performance.

The Life Corpus originally consisted of 102 suicidal messages, (71 texts in English and 31 texts in Spanish) 70 samples (No risk), and 32 texts (Risk), divided into four classes: No Risk, Urgent, Possible, and Immediate, all of them unbalanced (Table 1).

As the corpus was very small and there were too many categories to obtain statistically significant data (Caicedo et al., 2020), it was decided to merge the three risk classes (Possible, Urgent, and Immediate) into one, keeping the No-Risk class intact, to reduce the imbalance and therefore improve the quality of the experiments. We make the same decision in this work using the two categories (Table 2).

To increase the number of samples in the corpus, we decided to collect texts from the social platform Reddit (Gilbert, 2013), within which is the subreddit "SuicideWatch", which had 984 texts in English in the extraction date. These texts were extracted using the PRAW library (an acronym for "Python Reddit API Wrapper") in Python, which allows access to Reddit through a developer account (Reddit, n.d.). After the 984 texts had been extracted, they were preprocessed, eliminating HTML tags.

With the original supervised Life Corpus and the new unsupervised Reddit Corpus, we developed the system shown in Fig. 1 to increase the Life Corpus more samples, especially those tagged as Risk.

After this, using the samples of the Life Corpus (English + Spanish translated to English), a classifier was created based on the SVM algorithm and using the characteristics of BoW (Cao et al., 2014) with Tf/Idf terms weighting. This initial classifier was used to increase the corpus with a Bootstrapping Uncertainty Sampling technique. As the Life Corpus was very unbalanced towards samples without risk (Table 2), we were only interested in choosing the samples that the classifier categorized as Risk. In this way, considering that the classifier would have a significant error rate, samples without risk would also be included. In each iteration, the cutoff threshold was increased exponentially to limit the acceptance of new samples comparing that threshold with the confidence score given by the sklearn SVM classifier for each sample. That means if I have the sample X and Y, and the SVM score for the sample X is higher than the threshold, and the score of the sample Y is lower, sample X was added, and the sample Y was rejected in that iteration. Specifically, the threshold grew in each iteration $(1 - 0.2^n)$, where n is the iteration number. We repeated this iteration until any of the SVM confidence scores for the evaluated samples overcame the threshold. In each iteration of the bootstrapping, the unsupervised accepted samples were added to the supervised corpus, training the model again with the new samples and repeating the iteration until, as mentioned, the classifier stopped classifying no previously tagged samples as at risk. Table 3, shows the number of corpus samples of each iteration and the new calculated threshold.

The bootstrapping process above left us with a corpus of 302 Reddit samples, 200 of these samples were evaluated by six annotators divided into four groups of 50 samples, such that one of the annotators evaluated the samples of the four groups, while five independent annotators annotated only their group of 50 samples. The mutual agreement of 0.86 was reached, representing 171 agreements in the evaluation of the texts (Cohen, 1960b). The annotation results are shown in Table 4.

To ensure the quality of the data, of these 302 samples we kept the 171 in which all the annotators agreed. That is, there was mutual agreement in classifying such samples. These samples were joined to the Life Corpus samples to build five different corpora of different sizes (Table 5).

The experiments of the present work were carried out with this data.

Table 1
Number of samples for each "Alert Level" type.

Alert level	Quantity	EN	ES
No risk	70 (68.6%)	45 (63.4%)	25 (80.6%)
Urgent	19 (18.6%)	15 (21.1%)	4 (12.9%)
Possible	8 (7.8%)	6 (8.5%)	2 (6.5%)
Immediate	5 (4.9%)	5 (7%)	0 (0%)

Table 2
Number of samples for each "Alert Level" type.

Alert Level	Quantity	EN	ES
No risk	70 (68.63%)	45	25
Risk	32 (31.37%)	26	6

The reasons why we used five corpora were the following: i) the data sources were very heterogeneous and we had to verify that the results did not deteriorate using the separate corpora; and ii) as we can see in Table 2, the Life Corpus contains a mixture of messages in English and Spanish and we wanted to test how automatic translation affected the performance of the system with a corpus where all messages were in the same language with automatic translation, or keeping the original messages. There is no option for the Reddit Corpus with and without translation because all the messages were already in English in the Reddit Corpus. Therefore, we repeat all the experiments with:

Life: Only the translated version of Life.

Reddit: Only the Reddit Corpus samples with a mutual agreement between the annotators.

Life + Reddit: The combination of samples from the Life Corpus translated into English and the samples from Reddit by mutual agreement.

Life_es_en: The original and untranslated Life Corpus, with mixed samples in English and Spanish.

Life_es_en + Reddit: The original untranslated Life Corpus plus samples with mutual agreement from Reddit.

With this combination, we assessed the performance improvement when we added the Reddit Corpus, or how the automated translation of the texts affected this performance. The automatic translation of the texts in Spanish from the Life Corpus to the English language was carried out through the free and unlimited Python GoogleTrans library (Google, n.d.).

Once the different corpora were obtained, we used two different classifiers: Support Vector Machine (SVM) and Rasa intent classifier. The SVM has been widely used to classify texts, giving good results in different research processes (Suthaharan, 2016). For this learning machine, its implementation by sklearn was used (Siglidis et al., 2020). Moreover, we also wanted to conduct experiments with deep learning algorithms. Still, due to the small size of the Corpus, we decided to use the Rasa algorithm of Natural Language Understanding (NLU) (Goyal et al., 2008) that enables classifications using language models together with deep learning techniques.

With the Rasa algorithm, we use the characteristics that are defined by default in the Lexical Syntactic Featurizer algorithm: *low, title, upper, BOS, EOS, digit, pos. low* indicates if the term is lowercase or not, *title* if the word starts with a capital letter, *upper* if the word is all capitalized, *digit* if it is a number, *pos* is Part of Speech, *BOS* beginning of a sentence, and *EOS* end of the sentence.

On the other hand, the following characteristics were used for the SVM algorithms: Bag of Words (BoW) with and without Term-Frequency/Inverse Document Frequency (Tf/Idf), as term weighter, and with or without stopwords. To improve the coverage of the results, we use word embeddings to expand the terms of each message. For the use of word embeddings, we use the Polyglot library (Al-Rfou, n.d.). When given a term, this library suggests a number n of terms close to such term whose embedding vector is less than a distance d from the vector of the searched term. After several preliminary tests, we decided to set n to a value of 10 and d to 0.85.

We evaluate the results with the following metrics: simple accuracy, balanced accuracy, micro f1, macro f1, weighted f1, micro-precision, macro precision, weighted precision, micro recall, macro recall, weighted recall, micro Jaccard, macro Jaccard, weighted Jaccard. However, as the corpus were unbalanced, either towards no risk samples

in the case of the Life Corpus or risk samples for the Reddit or Life+Reddit Corpus, we decided to use macro f1 as our primary metric, since it is the one that best responded to the unbalanced corpus, which calculates the f1 statistic, separated by classes and does not use weights for grouping (Overflow, n.d.).

The experiments were performed using the 10-fold cross-validation technique and repeated 30 times, with different cross-validation random divisions, to obtain statistical significance using a *t*-test for mean differences of the values of macro f1.

4. Results

As we had two variants of the Life Corpus, one with all the samples translated into English and the other keeping the samples in Spanish, we decided to divide the experiments with all the corpora that had a variant and with those that did not. Therefore, for the first part of this section, we conducted the experiments with the Life, Reddit, and Life+Reddit collections with all the texts translated into English. However, in the second part, we used the Life_es_en and Life_es_en + Reddit Corpus, with the Life Corpus untranslated. We did not use only the Reddit Corpus for the second part because it is entirely in English.

4.1. Experiments with Corpus in English

In Fig. 2, we can observe the results using the SVM classifier (using the features BoW, BoW+Embeddings, Tf/Idf, and Tf/Idf + Embeddings) and Rasa (with the default features extraction system) for the three different English data collections: Life, Reddit, and Life+Reddit.

The results of macro f1 were better with the SVM classifier and BoW as features (without Tf/Idf weighter). There were no statistically significant differences whether we used BoW with or without word embeddings expansion. Therefore, both systems appear to have similar performance. Nevertheless, for smaller corpus, the Rasa system was the best, confirming that this classifier can improve other processes when the number of samples is small including learned language models.

In Table 6 we can see the results with more detail and observe the macro precision and macro recall, achieving a value of 0.79 of macro f1. As we can see in the table, the best system has better performance in the three measures (macro f1, macro precision, and macro recall).

4.2. Experiments with Corpus in Spanish and English

As mentioned above, in the second group of experiments, the original Life Corpus was used alone without translating any samples or in combination with Reddit, entirely in English. The objective of these experiments was to observe whether or not the translation affected the results in the classification of messages with suicidal ideation. As in previous experiments, the SVM and Rasa classifiers were used. For the first, the characteristics of BoW were extracted from each corpus, BoW expanded with word embeddings, and its weight variants with Tf/Idf. The result of these experiments can be seen in Fig. 3.

Once again, the best results were obtained using the most extensive corpus (Life_es_en + Reddit) with the classifier SVM and extracting the BoW features, with or without word embedding expansion. This system gives a macro f1 of 0.80 ($p < .01$). The best result for the Life Corpus without translation was using Rasa (0.48). In Table 7, we present the best system for each corpus with better performance in the three measures (macro f1, macro precision, and macro recall).

Although the biggest corpus precision and recall match the best results, this does not happen with the smaller corpus. The best recall system differs from the best macro f1 system ($p < .01$). Moreover, if we compare these results with the translated corpus (Table 5), we can observe a slight non-significant difference (0.80 ± 0.01 vs 0.79 ± 0.01 $p < .01$). This means that the process of automatic translation with the GoogleTrans library of the samples in Spanish from the original corpus to the English language does not significantly worsen the performance of
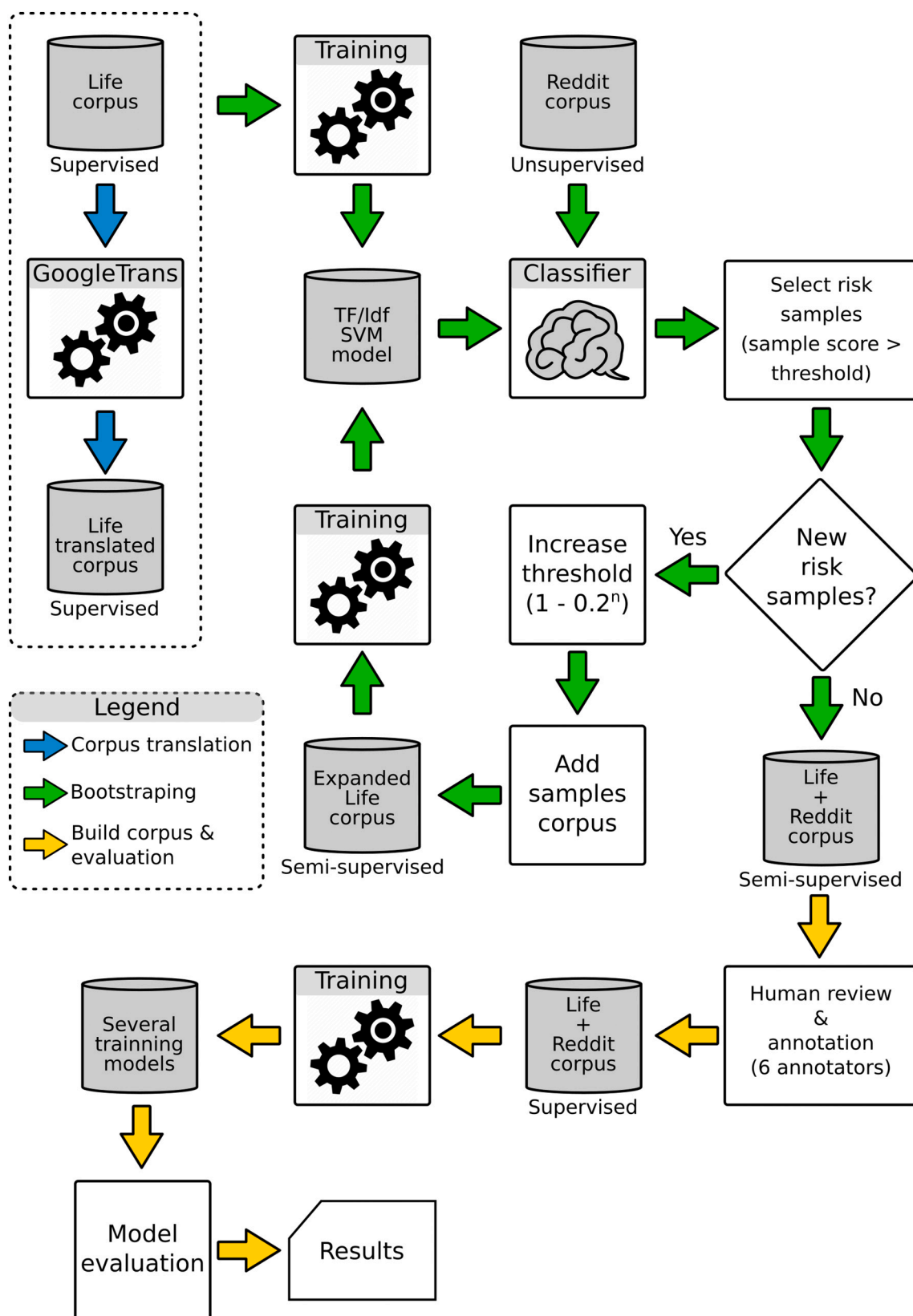
Fig. 1. System workflow scheme. The system was evaluated using, the original Life Corpus and translated Life Corpus. The system is composed of three processes: i) translation process, ii) bootstrapping Corpus expansion, iii) reviewing, building, and evaluating the final supervised corpus.

Table 3

Number of samples and threshold by iteration.

Iteration	Num samples	Threshold
1	102	0.8
2	225	0.96
3	302	0.992

Table 4

Agreements between reviewers.

Group	Reviewer	Suicide text Risk/no risk	Mutual-agreement TP/TN	Kappa Cohen
1	Reviewer GC	38/12	38/3	0.82
	Reviewer RA	47/3		
2	Reviewer CC	46/4	46/4	1.00
	Reviewer RA	46/4		
3	Reviewer KM	45/5	41/3	0.88
	Reviewer RA	47/3		
4	Reviewer AR	32/18	26/10	0.72
	Reviewer JG	31/19		
	Reviewer RA	35/15		
Totales		364/87	151/20	0.86

Table 5

Corpus used in experiments.

Corpus	Risk	Not risk	Total
Corpora only in English languages			
Life	30	72	102
Reddit	153	18	171
Life+Reddit	183	90	273
Corpus in Spanish and English languages			
Life_es_en	30	72	102
Life_es_en + Reddit	183	90	273
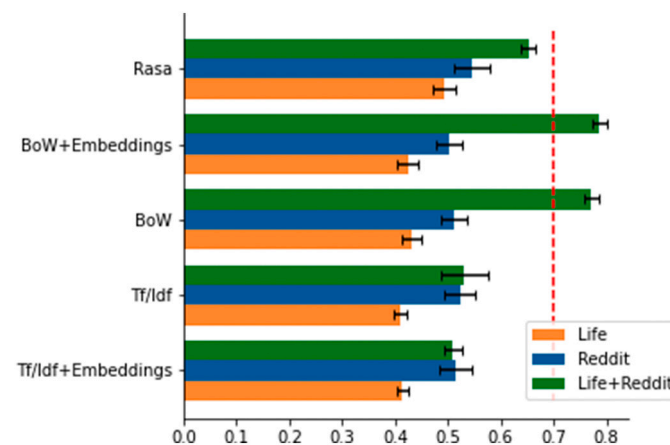
Fig. 2. The vertical pointed line is the original f-measure result with Life Corpus.

the detection of suicidal messages.

5. Discussion

Although O'Dea and Gómez developed corpus whose quality was evaluated through agreements, their development methodology was different: the first, one developed a corpus from data downloaded from Twitter in the same period, of which the 14% selected were randomly divided into two data sets to be evaluated by human coders who classified them into three categories: "Very concerning" (14%), "Possibly

Table 6

Macro f1, macro precision, and macro recall. Corpus in the English language combined with the training features. The confidence interval was calculated with $p < .01$.

Features	Macro f1	Macro precision	Macro recall
Life			
Rasa	0.49 ± 0.02	0.52 ± 0.03	0.53 ± 0.02
BoW	0.43 ± 0.02	0.40 ± 0.03	0.50 ± 0.01
BoW + Embedding	0.42 ± 0.02	0.39 ± 0.03	0.51 ± 0.02
Tf/Idf	0.41 ± 0.01	0.35 ± 0.01	0.51 ± 0.01
Tf/Idf + Embeddings	0.41 ± 0.01	0.35 ± 0.01	0.51 ± 0.01
Reddit			
Rasa	0.55 ± 0.03	0.54 ± 0.04	0.57 ± 0.03
BoW	0.51 ± 0.02	0.48 ± 0.03	0.55 ± 0.02
BoW + Embedding	0.50 ± 0.02	0.47 ± 0.02	0.54 ± 0.02
Tf/Idf	0.52 ± 0.03	0.50 ± 0.03	0.55 ± 0.03
Tf/Idf + Embeddings	0.52 ± 0.03	0.49 ± 0.03	0.55 ± 0.03
Life + Reddit			
Rasa	0.65 ± 0.01	0.76 ± 0.02	0.66 ± 0.01
BoW	0.77 ± 0.01	0.77 ± 0.01	0.79 ± 0.01
BoW + Embeddings	0.79 ± 0.01	0.80 ± 0.01	0.79 ± 0.01
Tf/Idf	0.53 ± 0.04	0.72 ± 0.06	0.58 ± 0.05
Tf/Idf + Embeddings	0.51 ± 0.02	0.68 ± 0.04	0.56 ± 0.01
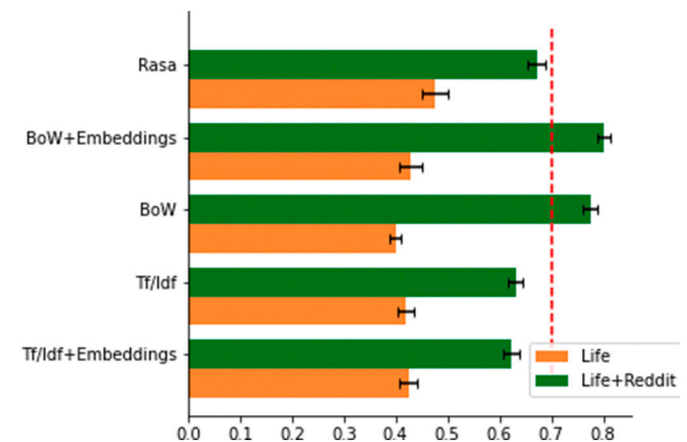
Fig. 3. The vertical pointed line is the original f-measure result with Life Corpus.

Table 7

Macro f1, macro precision, and macro recall. Corpus in English and Spanish language combined with training features. The confidence interval was calculated with $p < .01$.

Features	Macro f1	Macro precision	Macro recall
Life_es_en			
Rasa	0.48 ± 0.03	0.51 ± 0.03	0.50 ± 0.02
BoW	0.40 ± 0.01	0.35 ± 0.02	0.49 ± 0.01
BoW + Embeddings	0.43 ± 0.02	0.40 ± 0.04	0.51 ± 0.01
Tf/Idf	0.42 ± 0.02	0.37 ± 0.02	0.51 ± 0.01
Tf/Idf + Embeddings	0.43 ± 0.02	0.37 ± 0.02	0.52 ± 0.01
Life_es_en + Reddit			
Rasa	0.67 ± 0.02	0.78 ± 0.02	0.67 ± 0.01
BoW	0.78 ± 0.01	0.78 ± 0.01	0.79 ± 0.01
BoW + Embeddings	0.80 ± 0.01	0.81 ± 0.01	0.81 ± 0.01
Tf/Idf	0.63 ± 0.01	0.78 ± 0.02	0.64 ± 0.01
Tf/Idf + Embeddings	0.62 ± 0.01	0.76 ± 0.02	0.63 ± 0.01

concerning" (56%) and "Safe to ignore" (29%), with a Cohen's Kappa agreement of 0.76, while the classifier correctly identified 80% of the tweets in the category "Very concerning". The second, one initially developed a supervised corpus with annotations from different sources that, because it was evaluated with four categories, had a moderate Cohen's Kappa agreement of 0.52 (average k = 0.55), in the process of increasing the size of the corpus samples, a semi-supervised methodology was used with texts from the "SuicideWatch" subreddit, which allowed expanding the number of samples from 102 to 273 (183 risk and 90 No risk) with an agreement between Cohen's Kappa annotators of 0.86.

As we have seen in Section 4, the results are promising, being a semi-supervised learning system capable of achieving a macro f1 of 0.78–0.81, close to the mutual agreement reached by human reviewers (Cohen's Kappa of 0.86). These results came when the Life Corpus increased in size by adding the 171 samples from the Reddit Corpus, in which the annotators reached a mutual agreement. These results also show us that the semi-supervised Bootstrapping Uncertainty Sampling methodology chosen to expand the Life Corpus with new samples is valid and useful to improve the results of the automatic system for detecting messages of depression or suicidal ideation.

The Rasa NLU classifier works better than SVM for smaller corpus, maybe because it uses some pre-trained language models. However, BoW and SVM work better than the Rasa deep learning approach. Expanding the text using word embeddings does not affect significative the results ($p < .01$) or the automatic translation of the texts in Spanish from the Life Corpus to the English language using the GoogleTrans library.

On the other hand, before using this methodology we assumed that there would not be so many suicidal messages in the Reddit subgroup and that the initial classifier could not find so many messages of this type. Therefore, the final corpus will be more balanced. However, the initial classifier for the bootstrapping techniques using a preliminary SVM classifier worked better than we expected. This has led us to move from a Life Corpus where most posts had no suicide risk to another Life_es_en + Reddit Corpus, where most posts, according to reviewers, had clues of suicidal ideation or depression (Table 5).

6. Conclusions and future work

Because they have been annotated through a supervised methodology (Egmond and R. D.-C. T. J. of Crisis, 1990; Barraclough and Hughes, 1987; Huang et al., 2014), the suicide corpus population has been costly (Mircoli et al., 2017; Akpınar et al., 2019; Cremades et al., 2017b; Priyanthan et al., 2012). The objective of this research was to increase the number of samples of the Life Corpus (Liu et al., n.d.) using a semi-supervised method (Komiya et al., 2018; Braithwaite et al., 2016), which allowed to maintain the quality of the added texts, reducing the human effort.

In this work we have demonstrated two things: i) Bootstrapping Uncertainty Sampling technique used in the present work can be helpful to increase a corpus suitable for suicide prevention using supervised machine learning approaches, and ii) the expanded Life Corpus can be supportive to build a classifier in which messages of depressive or suicide ideation can be detected almost as well as the human mutual agreement.

In future work, we plan to experiment with other classification algorithms and new features, to optimize the semi-supervised annotation methodology of new samples, from microblogging networks, blogs, forums, or other sources for the Life Corpus. To constantly increase the number of samples of this corpus, to carry out more and more exact searches of potential suicidal users in social networks.

Moreover, we want to explore more the deep learning algorithms using other Rasa features or parameters, and other NLU classifiers or technologies such as BERT embeddings. On the other hand, we want to change the use of word embeddings to use their vectors directly instead

of expanding text terms with neighbor terms or to try the use of sentence embeddings.

Likewise, we intend to validate the semi-supervised annotation methodology that in this research, it was used samples from micro-blogging networks, blogs, forums, or other sources, with samples from actual clinical diagnoses, based on psychiatric epidemiological studies to determine its effectiveness in the real world and to be able to more accurately predict the probability of future thoughts of death, suicidal ideation, suicide plan or suicide attempts.

In addition, we plan to test different machine translation methodologies, such as the one provided by Google's translator through the GoogleTrans library, to generate parallel corpus expanded into different languages, from the original texts written primarily in the English language to do experiments for the detection of suicidal users in social networks in other languages, carried out by our research group or by other groups, which will be able to access this parallel corpus of the Life Corpus for free under the Creative Commons license in https://github.com/PlataformaLifeUA.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

References

Abbasi, A., Hassan, A., Dhar, M., 2014. Benchmarking Twitter sentiment analysis tools sentiment analysis on social media view project signal fusion and semantic similarity evaluation for social media based adverse drug event detection view project benchmarking twitter sentiment analysis tools.

Akpınar, M., Oral, B., Engin, D., 2019. A semi-aautomatic annotation interface for named entity and relation annotation on document images. … E. E.-2019 4th, and undefined. ieeexplore.ieee.org. … E. E.-2019 4th, and undefined.

Alameda-Pineda, X., et al., 2013. RAVEL: an annotated corpus for training robots with audiovisual abilities. J.Multimodal User Interfaces 7 (1–2), 79–91. https://doi.org/10.1007/s12193-012-0111-y. Sep.

Al-Garadi, M.A., Varathan, K.D., Ravana, S.D., 2016. Cybercrime detection in online communications: the experimental case of cyberbullying detection in the twitter network. Comput. Hum. Behav. 63, 433–443. https://doi.org/10.1016/j.chb.2016.05.051. Oct.

Al-Rfou, Rami. Welcome to polyglot's documentation! — polyglot 16.07.04 documentation. https://polyglot.readthedocs.io/en/latest/. (Accessed 14 June 2021).

Amorim, W.P., Tetila, E.C., Pistori, H., Papa, J.P., 2019. Semi-supervised learning with convolutional neural networks for UAV images automatic recognition. Comput. Electron. Agric. 164, 104932 https://doi.org/10.1016/j.compag.2019.104932. Sep.

Asghar, N., 2016. Automatic Extraction of Causal Relations From Natural Language Texts: A Comprehensive Survey. May.

Barnes, M., 2007. User-generated Metadata in Social Software: An Analysis of Findability in Content Tagging And Recommender Systems. … E. B.-G.-F. in, and undefined.

Barraclough, B., Hughes, J., 1987. Suicide: Clinical And Epidemiological Studies. Croom Helm.

Beck, A., Beck, R., Kovacs, M., 1975. Classification of suicidal behaviors: I. Quantifying intent and medical lethality. Am. J. Psychiatry 132 (3), 285–287. https://doi.org/10.1176/appi.books.9780890425596.

Beck, A., Kovacs, M., Weissman, A., 1979. Assessment of suicidal intention: the scale for suicide ideation. J.Consult.Clin.Psychol. 47 (2), 343–352. https://doi.org/10.1037/0022-006X.47.2.343.

Ben-David, A., 2008. Comparison of classification accuracy using Cohen's Weighted Kappa. Expert Syst. Appl. 34 (2), 825–832. https://doi.org/10.1016/j.eswa.2006.10.022. Feb.

Bentivogli, L., Pianta, E., 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. Nat. Lang. Eng. 11 (3), 247–261. https://doi.org/10.1017/S1351324905003839. Sep.

Birjali, M., Beni-Hssane, A., Erritali, M., 2017. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. Procedia Comput.Sci. 113, 65–72. https://doi.org/10.1016/j.procs.2017.08.290.

Birjali, M., Beni-Hssane, A., Erritali, M., 2017. In: A Method Proposed for Estimating Depressed Feeling Tendencies of Social Media Users Utilizing Their Data. Springer, Cham, pp. 413–420. https://doi.org/10.1007/978-3-319-52941-7_41.

Bontcheva, K., et al., 2013. GATE Teamware: a web-based, collaborative text annotation framework. Lang. Resour. Eval. 47 (4), 1007–1029. https://doi.org/10.1007/s10579-013-9215-6. Feb.

Braithwaite, S.R., Giraud-Carrier, C., West, J., Barnes, M.D., Hanson, C.L., 2016. Validating machine learning algorithms for Twitter data against established measures of suicidality. JMIR Ment. Health 3 (2), e21. https://doi.org/10.2196/mental.4822. May.

Brum, H.B., Nunes, M. das G.V., 2018. Semi-supervised sentiment annotation of large corpora. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11122 LNAI, pp. 385–395. https://doi.org/10.1007/978-3-319-99722-3_39. Feb.

Caicedo, R.W.Acuña, Soriano, J.M.Gómez, Sasieta, H.A.Melgar, 2020. Assessment of supervised classifiers for the task of detecting messages with suicidal ideation. Heliyon 6 (8), e04412. https://doi.org/10.1016/j.heliyon.2020.e04412. Aug.

Canales, L., Strapparava, C., 2016. Exploiting a bootstrapping approach for automatic annotation of emotions in texts. … E. B.-… C. on D., and undefined. ieeexplore.ieee.org.

Cao, J., Chen, T., Fan, J., 2014. Fast online learning algorithm for landmark recognition based on BoW framework. In: Proceedings of the 2014 9th IEEE Conference on Industrial Electronics and Applications, ICIEA 2014, pp. 1163–1168. https://doi.org/10.1109/ICIEA.2014.6931341. Oct.

Cerel, J., Jordan, J.R., Duberstein, P.R., 2008. The impact of suicide on the family. Crisis 29 (1), 38–44. https://doi.org/10.1027/0227-5910.29.1.38.

Chang, Q., Sha, F., Chan, C.H., Yip, P.S.F., 2018. Validation of an abbreviated version of the Lubben Social Network Scale ('LSNS-6') and its associations with suicidality among older adults in China. PLoS ONE 13 (8), e0201612. https://doi.org/10.1371/journal.pone.0201612. Aug.

Chen, C.C., De Tseng, Y., 2011. Quality evaluation of product reviews using an information quality framework. Decis. Support. Syst. 50 (4), 755–768. https://doi.org/10.1016/j.dss.2010.08.023. Mar.

Choudhury, S., Pattnaik, S., 2020. Emerging themes in e-learning: a review from the stakeholders' perspective. Comput.Educ. 144, 103657 https://doi.org/10.1016/j.compedu.2019.103657. Jan.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. https://doi.org/10.1177/001316446002000104. Apr.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. https://doi.org/10.1177/001316446002000104. Apr.

Cole, S.V., Royal, M.D., Valtorta, M.G., Huhns, M.N., Bowles, J.B., 2006. A lightweight tool for automatically extracting causal relationships from text. In: Conference Proceedings - IEEE SOUTHEASTCON, 2006, pp. 125–129. https://doi.org/10.1109/second.2006.1629336.

Cremades, S.Z., Gomez Soriano, J.M., Navarro-Colorado, B., 2017. Design, compilation and annotation of a corpus for the detection of suicide messages in social networks. In: Procesamiento de Lenguaje Natural, 59, pp. 65–72.

Cremades, S.Z., Gomez Soriano, J.M., Navarro-Colorado, B., 2017. Design, compilation and annotation of a corpus for the detection of suicide messages in social networks. Proces.Leng.Nat. 59, 65–72.

Desmet, B., Hoste, V., 2013. Emotion detection in suicide notes. Expert Syst. Appl. 40 (16), 6351–6358. https://doi.org/10.1016/J.ESWA.2013.05.050. Nov.

Desmet, B., Hoste, V., 2018. Online suicide prevention through optimised text classification. Inf. Sci. 439–440, 61–78. https://doi.org/10.1016/J.INS.2018.02.014. May.

Du, J., Xu, J., Song, H.Y., Tao, C., 2017. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with twitter data. BMC Med.Inform.Decis.Mak. 17 (2), 63–70. https://doi.org/10.1186/s12911-017-0469-6. Jul.

Egmond, M.Van, R. D.-C. T. J. of Crisis, 1990. The predictability of suicidal behavior: the results of a meta-analysis of published studies. Crisis 57–84.

Fielding, N., Lee, R., Blank, G., 2008. The SAGE Handbook of Online Research Methods. Sage.

Fu, K., Cheng, Q., Wong, P.W.C., Yip, P.S.F., 2013. Responses to a Self-presented Suicide Attempt in Social Media. https://doi.org/10.1027/0227-5910/a000221. Jan.

Gilbert, E., 2013. Widespread underprovision on reddit. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work. CSCW, pp. 803–807. https://doi.org/10.1145/2441776.2441866.

Girju, R., Moldovan, D., 2002. Text Mining for Causal Relations.

Goldstein, R.B., Black, D.W., Nasrallah, A., Winokur, G., 1991. The prediction of suicide. Arch. Gen. Psychiatry 48 (5), 418. https://doi.org/10.1001/archpsyc.1991.01810290030004. May.

Gómez, J.M., 2014. Language technologies for suicide prevention in social media. In: Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC), pp. 21–29. https://doi.org/10.3115/v1/W14-6903.

Google. Googletrans: Free and Unlimited Google translate API for Python — Googletrans 3.0.0 documentation. https://py-googletrans.readthedocs.io/en/latest/. (Accessed 21 November 2020). Oct.

Goyal, P., Behera, L., McGinnity, T.M., 2008. Application of Bayesian framework in natural language understanding. In: IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India), vol. 25, pp. 251–269. https://doi.org/10.4103/0256-4602.44656.

Guan, L., Hao, B., Cheng, Q., Yip, P., Zhu, T., 2015. Identifying Chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model. JMIR Ment.Health.

Gundlapalli, A.V., et al., 2013. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. J. Am. Med. Inform. Assoc. 20 (E2), e355–e364. https://doi.org/10.1136/amiajnl-2013-001946. Dec.

Gupta, S., Pawar, S., Ramrakhiyani, N., Palshikar, G.K., Varma, V., 2018. Semi-supervised recurrent neural network for adverse drug reaction mention extraction. BMC Bioinform. 19 (8), 1–7. https://doi.org/10.1186/S12859-018-2192-4. Jun.

Halike, A., Abiderexiti, K., Yibulayin, T., 2020. Semi-automatic corpus expansion and extraction of Uyghur-named entities and relations based on a hybrid method. Information 11 (1), 31. https://doi.org/10.3390/info11010031. Jan.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software. ACM SIGKDD Explor.Newslett. 11 (1), 10. https://doi.org/10.1145/1656274.1656278. Nov.

Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor.Quant.MethodsPsychol. 8 (1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023. Feb.

Hernandez, M., Pontes, J., 2014. Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)," the 5th Information Systems Research ….

Huang, Y., Goh, T., Liew, C., 2007. Hunting suicide notes in web 2.0-preliminary findings. In: Multimedia Workshops, 2007.

Huang, X., Zhang, L., Chiu, D., Liu, T., Li, X., Zhu, T., 2014. Detecting suicidal ideation in Chinese microblogs with psychological lexicons. In: 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, pp. 844–849. https://doi.org/10.1109/UIC-ATC-ScalCom.2014.48.

Hughes, D.H., 1995. Can the clinician predict suicide? Psychiatr. Serv. 46 (5), 449–451. https://doi.org/10.1176/ps.46.5.449. May.

IDC Corporate USA, 2012. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digi tal Shadows, and Biggest Growth in the Far East. https://www.idc.com/. (Accessed 23 March 2020).

Jashinsky, J., Burton, S., Hanson, C., West, J., 2014. Tracking suicide risk factors through Twitter in the US. Crisis.

Jeong, H., Nam, B., Jo, S.J., Lee, W.C., Yim, H.W., 2020. Clinical usefulness of the geriatric depression scale to identify the elderly at risk of suicide. Psychiatry Investig. 17 (5), 481–486. https://doi.org/10.30773/pi.2019.0299. May.

José, Gómez, 2017. Life! Prevención del Suicidio en las Redes Sociales | Grupo de Procesamiento del Lenguaje y Sistemas de Información. https://www.fbbva.es/equipo/analisis-sentimientos-redes-sociales-prevencion-suicidio/. (Accessed 14 November 2017).

Karimzadeh, M., MacEachren, A., 2019. GeoAnnotator: a collaborative semi-automatic platform for constructing geo-annotated text corpora. ISPRS Int. J. Geo Inf. 8 (4), 161. https://doi.org/10.3390/ijgi8040161. Mar.

Komiya, K., Suzuki, M., Iwakura, T., Sasaki, M., Shinnou, H., 2018. Comparison of methods to annotate named entity corpora. ACM Trans. Asian Low Resour. Lang. Inform. Process. 17 (4) https://doi.org/10.1145/3218820. Oct.

Lacson, R., Khorasani, R., 2011. Natural language processing: the basics (part 1). J. Am. Coll. Radiol. 8 (6), 436–437. https://doi.org/10.1016/j.jacr.2011.04.020. Jun.

S. Lai L. Xu K. Liu J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification".

Large, M., Nielssen, O., 2012. Suicidal ideation and later suicide. Am. J. Psychiatr.

Large, M., Ryan, C., 2014. Suicide risk assessment: myth and reality. Int. J. Clin. Pract. 68 (6), 679–681. https://doi.org/10.1111/ijcp.12378. Jun.

Larsen, M., Cummins, N., Boonstra, T., 2015. The use of technology in suicide prevention. In: Medicine And ….

Leung, T., Chung, F.L., 2014. Persuasion driven influence propagation in social networks. In: ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 548–554. https://doi.org/10.1109/ASONAM.2014.6921640. Oct.

Liang, P.-W., Dai, B.-R., 2013. Opinion mining on social media data. In: 2013 IEEE 14th International Conference on Mobile Data Management, pp. 91–96. https://doi.org/10.1109/MDM.2013.73. Jun.

Liao, Q., Wang, W., Han, Y., Zhang, Q., 2013. Analyzing the influential people in sina weibo dataset. In: GLOBECOM -IEEE Global Telecommunications Conference, pp. 3066–3071. https://doi.org/10.1109/GLOCOM.2013.6831542.

Lieberman, M., 2014. Visualizing big data: social network analysis. In: Digital Research Conference, pp. 1–23.

Life Corpus. GitHub. https://github.com/PlataformaLifeUA. (Accessed 29 June 2021).

Ling, W., Marujo, L., Dyer, C., Black, A.W., Trancoso, I., 2016. Mining parallel corpora from Sina Weibo and Twitter. Comput. Linguist. 42 (2), 307–343. https://doi.org/10.1162/COLI_a_00249. Jun.

X. Liu et al., "Proactive Suicide Prevention Online (PSPO): machine identification and crisis management for Chinese social media users with suicidal thoughts and," jmir.org.

Llisterri, J.Torruella Joaquim, 1999. Diseño de corpus textuales y orales.

Lu, X., 2014. Computational Methods for Corpus Annotation And Analysis. https://doi.org/10.1007/978-94-017-8645-4.

Luxton, D.D., June, J.D., Fairall, J.M., 2012. Social media and suicide: a public health perspective. Am. J. Public Health 102 (Suppl. 2), S195–S200. https://doi.org/10.2105/AJPH.2011.300608.

Matykiewicz, P., Duch, W., Pestian, J., 2009. Clustering semantic spaces of suicide notes and newsgroups articles. In: of the Workshop on Current Trends ….

Mircoli, A., Diamantini, C., Potena, D., 2017. Automatic Emotional Text Annotation Using Facial Expression Analysis.

Mowery, D., 2017. Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. J. Med. Internet Res. 19 (2), e48. https://doi.org/10.2196/jmir.6895.

Nguyen, N., Gabud, R., Ananiadou, S., 2019. COPIOUS: a gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. Biodivers. Data J. 7 (7) https://doi.org/10.3897/bdj.7.e29626. Jan.

O'Dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., Christensen, H., 2015. Detecting suicidality on twitter. Internet Interv. 2 (2), 183–188. https://doi.org/10.1016/j.invent.2015.03.005. May.

O'Connor, K., Sarker, A., Perrone, J., Hernandez, G.Gonzalez, 2020. Promoting reproducible research for characterizing nonmedical use of medications through data annotation: description of a Twitter corpus and guidelines. J. Med. Internet Res. 22 (2), e15861. https://doi.org/10.2196/15861. Feb.

Overflow, Stack. "scikit learn - What's the difference between Sklearn F1 score 'micro' and 'weighted' for a multi class classification problem? - Data Science Stack Exchange. ... E. B.-G.-F. in, and undefined. https://datascience.stackexchange.com/questions/40900/whats-the-difference-between-sklearn-f1-score-micro-and-weighted-for-a-mult. (Accessed 30 May 2021).

Pang, B., L. L.-F. and T. in Information, 2008. Opinion mining and sentiment analysis. and undefined 2008. In: Foundations and Trends{\textregistered} in Information Retrieval, 2, pp. 1–135.

Pang, B., L. L.-P. of the 42nd annual meeting on, 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. dl.acm.org. May.

Paris, J., 2006. Predicting and preventing suicide: do we know enough to do either? Harv.Rev.Psychiatry 14 (5), 233–240.

Parraga-Alava, J., Acuna-Caicedo, R., Gomez, J.M., Inostroza-Ponta, M., 2019. An unsupervised learning approach for automatically to categorize potential suicide messages in social media. In: 2019 38th International Conference of the Chilean Computer Science Society (SCCC), pp. 1–8. https://doi.org/10.1109/sccc49216.2019.8966443. Jan.

Peschel, A.O., Kazemi, S., Liebichová, M., Sarraf, S.C.M., Aschemann-Witzel, J., 2019. Consumers' associative networks of plant-based food product communications. Food Qual. Prefer. 75, 145–156. https://doi.org/10.1016/j.foodqual.2019.02.015. Jul.

Pestian, J., Grupp-Phelan, J., 2016. A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. Suicide Life Threat. Behav.

Pestian, J., Matykiewicz, P., 2008. Using natural language processing to classify suicide notes. In: Proceedings of the.

Priyanthan, P., Prasath, N., Perera, As, 2012. Opinion Mining And Sentiment Analysis on a Twitter Data Stream.

Purver, M., Battersby, S., S. Association for Computational Linguistics, 2012. Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2012). ACL.

Raschka, S., Mirjalili, V., 2019. Python machine learning: aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow, Segunda. Marcombo, S. A.

Reagan, A.J., Danforth, C.M., Tivnan, B., Williams, J.R., Dodds, P.S., 2017. Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. EPJ Data Sci. 6 (1), 28. https://doi.org/10.1140/epjds/s13688-017-0121-9.

Reddit. "Deep Web," reddit. https://www.reddit.com/r/deepweb/. (Accessed 17 April 2019).

Ren, F., Matsumoto, K., 2016. Semi-automatic creation of youth slang corpus and its application to affective computing. IEEE Trans. Affect. Comput. 7 (2), 176–189. https://doi.org/10.1109/TAFFC.2015.2457915. Apr.

Rodriguez, P., Ortigosa, A., Carro, R.M., 2012. Extracting emotions from texts in E-learning environments. In: Proceedings - 2012 6th International Conference on Complex, Intelligent, And Software Intensive Systems,CISIS2012, pp. 887–892. https://doi.org/10.1109/CISIS.2012.192.

Santini, Z.Ivan, Koyanagi, A., Tyrovolas, S., Haro, J.M., 2015. The association of relationship quality and social networks with depression, anxiety, and suicidal ideation among older married adults: findings from a cross-sectional analysis of the Irish Longitudinal Study on Ageing (TILDA). J. Affect. Disord. 179, 134–141. https://doi.org/10.1016/j.jad.2015.03.015. Jul.

Scheible, S., Whitt, R., Durrell, M., 2011. A gold standard corpus of Early Modern German. P. B.-P. of the 5th, and undefined. dl.acm.org.

Schwartz, H., Ungar, L., 2015. Data-driven content analysis of social media: a systematic overview of automated methods. Ann.Am.

Siglidis, G., Giatsidis, C., Skianis, K., 2020. GraKeL: A Graph Kernel Library in Python.

Silenzio, V.M.B., Duberstein, P.R., Tang, W., Lu, N., Tu, X., Homan, C.M., 2009. Connecting the invisible dots: reaching lesbian, gay, and bisexual adolescents and young adults at risk for suicide through online social networks. Soc. Sci. Med. 69 (3), 469–474. https://doi.org/10.1016/j.socscimed.2009.05.029. Aug.

N. Silveira et al., "A Gold Standard Dependency Corpus for English".

Stefanova, I., Kiryantsev, A., 2019. Analysis of user groups in social networks to detect socially dangerous people. In: 2018 International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T 2018 - Proceedings, pp. 789–792. https://doi.org/10.1109/INFOCOMMST.2018.8632086. Jan.

Suthaharan, S., 2016. In: Support Vector Machine. Springer, Boston,MA, pp. 207–235. https://doi.org/10.1007/978-1-4899-7641-3_9.

Sweeney, L., Owens, C., Malone, K., 2015. Communication and interpretation of emotional distress within the friendships of young Irish men prior to suicide: a qualitative study. HealthSoc.CareCommun. 23 (2), 150–158. https://doi.org/10.1111/hsc.12124. Mar.

Tapia, F., Aguinaga, C., Luje, R., 2018. Detection of behavior patterns through social networks like Twitter, using data mining techniques as a method to detect cyberbullying. In: 2018 7th International Conference On Software Process Improvement (CIMPS), pp. 111–118. https://doi.org/10.1109/CIMPS.2018.8625625. Oct.

Van Hee, C., et al., 2018. Automatic detection of cyberbullying in social media text. PLoS ONE 13 (10). https://doi.org/10.1371/journal.pone.0203794. Oct.

Velupillai, S., Hadlaczky, G., 2019. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. ... E. B.-G.-F. in, and undefined. ncbi.nlm.nih.gov.

Vieira, S.M., Kaymak, U., Sousa, J.M.C., 2010. Cohen's kappa coefficient as a performance measure for feature selection. https://doi.org/10.1109/FUZZY.2010.5584447.

Vioules, M.J., Moulahi, B., Aze, J., Bringay, S., 2018. Detection of suicide-related posts in Twitter data streams. IBM J. Res. Dev. 62 (1) https://doi.org/10.1147/JRD.2017.2768678. Jan.

We Are Social, 2022. "DIGITAL 2022: ANOTHER YEAR OF BUMPER GROWTH - We Are Social UK," Digital 2022 global overview report. https://wearesocial.com/uk/blog/2022/01/digital-2022-another-year-of-bumper-growth-2/. (Accessed 4 February 2022).

Weng, J., Lim, E.P., Jiang, J., He, Q., 2010. TwitterRank: Finding topic-sensitive influential twitterers. In: WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pp. 261–270. https://doi.org/10.1145/1718487.1718520.

World Health Organization, 2014. Preventing Suicide: A Global Imperative. World Health Organization.

Worldometer. Worldometer - World statistics in real time. https://www.worldometers.info/es/. (Accessed 4 February 2022).

Wu, T., Weld, D.S., Heer, J., 2019. Local decision pitfalls in interactive machine learning. ACM Trans.Comput.Human Interact. 26 (4), 1–27. https://doi.org/10.1145/3319616. Jun.

Zhang, L., Huang, X., Liu, T., Li, A., Chen, Z., Zhu, T., 2014. Using linguistic features to estimate suicide probability of Chinese microblog users. In: International Conference on Human Centered Computing, pp. 549–559. https://doi.org/10.1007/978-3-319-15554-8.

Zhang, A., Li, A., Zhu, T., 2015. Recognizing Temporal Linguistic Expression Pattern of Individual with Suicide Risk on Social Media," arXiv preprint arXiv:1511.09173.