Revista Gestão e Conhecimento
150-166, 2022
ISSN: 1677-9762

KNOWLEDGE DISCOVERY IN PORTUGUESE LANGUAGE SUICIDAL TWEETS USING FILTERED- EXTENDED ASSOCIATION RULES NETWORK

DESCOBERTA DE CONHECIMENTO EM TWEETS SUICIDAS EM LÍNGUA PORTUGUESA USANDO REDE DE REGRAS DE ASSOCIAÇÃO FILTRADAS ESTENDIDAS

DOI: 10.55908/RGCV16N1-010
Recebimento do original: 23/05/2022
Aceitação para publicação: 30/06/2022

Vítor Azevedo Silva
Graduando em Ciências da Computação
Instituição: Universidade Estadual do Piauí (UESPI)
Endereço: Rua Nossa Senhora de Fátima, s/n, de Fátima, Parnaíba – PI, Brasil, CEP: 64200-000
E-mail: vitorsilva@aluno.uespi.br

Matheus Henrique Silva Miranda
Graduando em Ciências da Computação
Instituição: Universidade Estadual do Piauí (UESPI)
Endereço: Rua Nossa Senhora de Fátima, s/n, de Fátima, Parnaíba – PI, Brasil, CEP: 64200-000
E-mail: matheusmiranda@aluno.uespi.br

Phillippy Cardelly Albuquerque Dos Santos
Graduando em Ciências da Computação
Instituição: Universidade Estadual do Piauí (UESPI)
Endereço: Rua Nossa Senhora de Fátima, s/n, de Fátima, Parnaíba – PI, Brasil, CEP: 64200-000
E-mail: phillippysantos@aluno.uespi.br

REVISTA gestão e conhecimento

Alinne Grazielle Mesquita De Farias
Graduanda em Ciências da Computação
Instituição: Universidade Estadual do Piauí (UESPI)
Endereço: Rua Nossa Senhora de Fátima, s/n, de Fátima, Parnaíba – PI, Brasil,
CEP: 64200-000
E-mail: alinnefarias@aluno.uespi.br

Renan de Padua
Doutor em Computação e Matemática Computacional
Instituição: Universidade de São Paulo (USP)
Endereço: Av. Trab. São Carlense, 400, Centro, São Carlos – SP, Brasil, CEP: 13566-590
E-mail: renan.padua88@gmail.com

Verônica Oliveira de Carvalho
Doutora em Computação e Matemática Computacional
Universidade Estadual Paulista (UNESP)
Endereço: Avenida 24 A, 1515, Rio Claro – SP, Brasil, CEP: 13506-900
E-mail: veronica.carvalho@unesp.br

Solange Oliveira Rezende
Professora Associada da Universidade de São Paulo
Universidade de São Paulo (USP)
Endereço: Av. Trab. São Carlense, 400, Centro, São Carlos – SP, Brasil, CEP: 13566-590
E-mail: solange@icmc.usp.br

Dario Brito Calçada
Doutor em Computação e Matemática Computacional
Instituição: Universidade Estadual do Piauí (UESPI)
Endereço: Rua Nossa Senhora de Fátima, s/n, de Fátima, Parnaíba – PI, Brasil,
CEP: 64200-000
E-mail: dariobcalcada@frn.uespi.br

ABSTRACT: The World Health Organization (WHO) and the *Global Burden of Disease study* estimate that nearly 800,000 people die by suicide each year. Social media are emerging surveillance tools that can help researchers track suicide risk factors in real time. Text Mining naturally becomes an area with greater affinity to promote studies in media such as Twitter. The discovery of terms that imply suicidality becomes crucial for its classification. The present article proposes the use of Filtered-Extended Association Rules Networks, for the selection of terms that indicate or not a suicidal tendency. The results provided the discovery of sets of terms that can be used to help classify *tweets* as being suicidal or not.

KEYWORDS: Association Rules Networks. NLP. tweets. suicide.

RESUMO: A Organização Mundial da Saúde (OMS) e o estudo *Global Burden of Disease* estimam que quase 800.000 pessoas morrem por suicídio a cada ano. Mídias sociais são ferramentas emergentes de vigilância que podem ajudar pesquisadores a rastrear fatores de risco de suicídio em tempo real. A Mineração de texto se torna naturalmente área com maior afinidade para promover estudos na mídia como o Twitter. A descoberta de termos que implicam na tendência suicida se torna crucial para sua classificação. O presente artigo propõe o uso de redes de regras de associação filtradas estendidas, para a seleção de termos que indicam ou não tendência suicida. Os resultados proporcionaram a descoberta de conjuntos de termos que podem ser usados para auxiliar na classificação de *tweets* como sendo de ideação suicida ou não.

PALAVRAS-CHAVE: Redes de Regras de Associação. PLN. tweets. suicídio.

1. INTRODUCTION

Death by suicide is an extremely complex issue that causes pain to hundreds of thousands of people every year around the world. The World Health Organization (WHO) and the *Global Burden of Disease study* estimate that nearly 800,000 people die by suicide each year, being one of the leading causes of death in young people. Thus, 1 suicide occurs every 40 seconds that could be avoided with timely and evidence-based interventions (RITCHIE; ORTIZ-OSPINA, 2015).

With an estimated 3.96 billion people actively using the internet (ABDULSALAM; ALHOTHALI, 2022), social media have become emerging surveillance tools that can help researchers track suicide risk factors in real time (JASHINSKY et al.., 2014). Many users prefer to use social media to share their thoughts and emotions, their day-to-day experiences and problems. Thus, suicidal ideation, death and self-harmful thoughts are among the most widely discussed topics on social media (ABDULSALAM; ALHOTHALI, 2022).

The works related to the construction of databases of suicidal tweets are scarce in the Portuguese language. However, in order to fill this gap, a study was developed that seeks to classify tweets by experts as "positive" and "negative" for suicidality. Preliminary results demonstrated an accuracy in the quality of the database (CARDOSO et al., 2020). The material provided by the authors went through pre-processing steps, such as removing *stopwords*, removing accents, and applying *stemming.* The representation chosen so that machine learning algorithms, classification, explainable artificial intelligence, among others, could be applied, was the *bag of words* in a *tokenization process* (CARDOSO et al., 2020).

The present study sought to develop an automatic knowledge discovery method, the Filtered-Extended Association Rules Network (*Filtered*-ExARN), in order to identify the main terms related to *tweets* with suicidal ideation written in Portuguese. A database of *tweets* classified as suicidal or not, built by Cardoso and his collaborators, was used. (CARDOSO et al., 2020).

The study results can contribute to a better accuracy in *Machine Learning* (ML) algorithms applied in the literature (ABDULSALAM; ALHOTHALI, 2022), as well as advances in the study of Explainable Artificial Intelligence (XAI) in applications that mine information in text. In order to achieve the proposed objective, a series of contributions were concluded, of which the following stand out: a new method of visualizing relevant association rules, the applicability of XAI in text mining, in addition to the listing of relevant terms in suicidal *tweets.*

2. RELATED WORKS

The works related to text mining in suicidal tweets mostly use Machine Learning algorithms. The study (HAQUE et al., 2022) offers a comparative analysis of various machine learning and deep learning models to identify suicidal thoughts from the social media platform Twitter. The experiments were conducted on a dataset of 49,178 instances retrieved from live tweets, for 18 suicidal and non-suicidal keywords, using the Python Tweepy API. Experimental findings reveal that the *Random Forest model* can achieve the highest-ranking score among machine learning algorithms, with an accuracy of 93% and an F1 score of 0.92. However, training the *deep learning classifiers* with word embedding increases the performance of ML models, where the *BiLSTM model* achieves an accuracy of 93.6% and an F1 score of 0.93 (HAQUE et al., 2022).

The study developed by (KANCHARAPU; SRINAGESH; BHANUSRIDHAR, 2022) used, in addition to the repetitive application of deep learning and machine learning algorithms, the involvement of coders. Suicide-related tweets are collected using suicide-related keywords and saved in a CSV file, which serves as a dataset. A recursive neural network was then used to classify these *tweets* to determine whether or not a suicide might occur, thus identifying individuals with a suicidal tendency (KANCHARAPU; SRINAGESH; BHANUSRIDHAR, 2022).

The work of (ROY et al., 2020) and collaborators aimed to generate an algorithm called " *Suicide Artificial Intelligence Prediction Heuristic* (SAIPH*)*" capable of predicting the future risk of suicidal thinking, analyzing data publicly available on Twitter. They trained a series of neural networks on queried Twitter data on psychological constructs associated with suicide, including overload, stress, loneliness, hopelessness, insomnia, depression, and anxiety. Using 512,526 tweets from N = 283 cases of suicidal ideation (SI) and 3,518,494 tweets from 2,655 controls, we trained a random forest model using neural network outputs to predict the binary status of suicidal ideation. The model predicted N = 830 SI events derived from an independent pool of 277 suicidal ideators relative to N = 3,159 control events, in all non-SI subjects with an accuracy of 0.88 (95% CI 0.86–0, 90).

Using an alternative approach, the model generated a temporal prediction of risk, such that peak occurrences above an individual specific threshold denote a ~7-fold increased risk for SI over the next 10 days (OU = 6.7 ± 1.1, P = 9 × 10-71). They validated the model using Twitter data obtained from the study regions, and observed significant associations of SI algorithm scores, with countywide 16-day suicide mortality rates in August and October 2019, most significantly in younger subjects. Algorithmic approaches such as SAIPH have the potential to identify individual future SI risk and can be easily adapted as clinical decision tools aiding in suicide screening and risk monitoring using available technologies (ROY et al., 2020).

Using a different approach from previous works, (HASSINE; ABDELLATIF; YAHIA, 2022) was based on classifiers and not deep learning or machine learning. However, as the number of tweets with suicide intent is small compared to the number of all tweets, this has led to an imbalanced classification problem, where the minority class (suicide intent) is more important than the majority class (no suicide intent). In such a situation, classical classifiers usually produce very imprecise results with respect to smaller classes, as they can easily discover rules that predict the majority class and ignore those related to the smaller one.

The work aimed to contribute to this line of research by introducing a new measure of interest to improve the classification process. This measure singles out the two classes regardless of their skewed distribution. Experiments performed proved that the adapted classifier outperforms the original and other pioneering baseline classification approaches in terms of prediction accuracy (HASSINE; ABDELLATIF; YAHIA, 2022).

Thus, it is observed that works that seek to predict suicidal tweets through machine learning and deep learning are widely used. However, these techniques are not XAI, losing the ability to explain the cause and effect in the data, without the "why". Only the study by (HASSINE; ABDELLATIF; YAHIA, 2022) sought a different approach from the previous ones, using classifier methods. However, despite a relevant study, the measures of interest created focus on the problem of data imbalance, without focusing on the specific content of the texts.

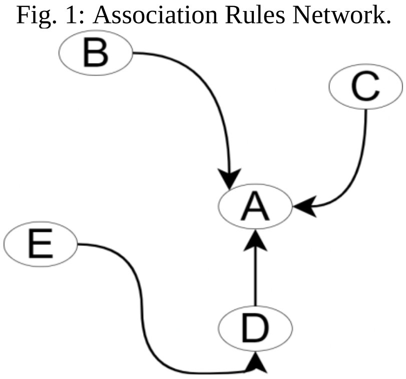
3. FILTERED- EXTENDED ASSOCIATION RULES NETWORKS (*Filtered-ExARN)*

The Apriori algorithm seeks to discover patterns of association rules by applying probabilistic measures, the main ones being support and trust. Commonly explained by the example of the supermarket, in which one wants to increase sales, the relationship between the products and each combination between them called *itemsets is* analyzed (AGRAWAL; IMIELINSKI; SWAMI, 1993).

The steps are done given a set of items *I*, a set of transactions *T*, consisting of subsets of *I*, an association rule is the relation A→B, where the subsets A and B are of *I* and A ∩ B = θ. That is, given that A happens, B also happens in *c%* of cases", where *c%* is the trust association rule. Another important measure in the association rule is support (*s%)*, which describes the percentage of transactions in which all items of the rule appear. A is called the antecedent (LHS - *Left Hand Side)* of the rule and B the consequent (RHS - *Right Hand Side)* (PADUA et al., 2018).

However, even filtering the association rules by the minimum trust and support, there are some limitations that can make the analysis process difficult, among them, the selection of relevant relations, redundancy, and the excessive number of rules. With that, some solutions were studied to solve these problems, such as the development of algorithms that build, through graphs, the links between *itemsets*, resulting in a better visualization and allowing the filtering of rules by means of edge pruning, the Association Rule Networks (ARNs) (PANDEY et al., 2009). Figure 1 shows an example of an association rule network in which B → A, C → A, E → D, D → A.

Fig. 1: Association Rules Network.



Source: authors.

Another strategy is to filter interesting rules through objective measures of interest, which can be classified as symmetric or asymmetric, that is, a measure M is symmetric if M (A → B) = M (B → A). Noting that support is a symmetrical measure and even though trust is asymmetrical, a solution study was carried out with other asymmetric measures to measure the ability that item A has to influence item B (TAN; STEINBACH; KUMAR, 2006). The studies of asymmetric measures led to the choice of gain and added value because they have the best results (CALÇADA; REZENDE, 2019) The two selected measures can be described as:

- *Added Value* [-1..0..1]: The Added Value measure indicates how much the frequency of the consequent item increases in the presence of the antecedent item, that is, it measures the gain of RHS in the presence of LHS (Equation 1. If $VA > 0$, the frequency of RHS increases in the presence of LHS.

$$AV = P(RHS|LHS) - P(RHS) = conf\ (LHS \rightarrow RHS) - P(RHS) \qquad (1)$$

- *Gain* [0..1]: Forms an exchange between support and trust, helping to select rules according to their frequency and minimum trust value (Equation 2) (FUKUDA et al., 1996).

$$Gain = [conf(LHS \rightarrow RHS) - minconf].P(LHS) \text{(two)}$$

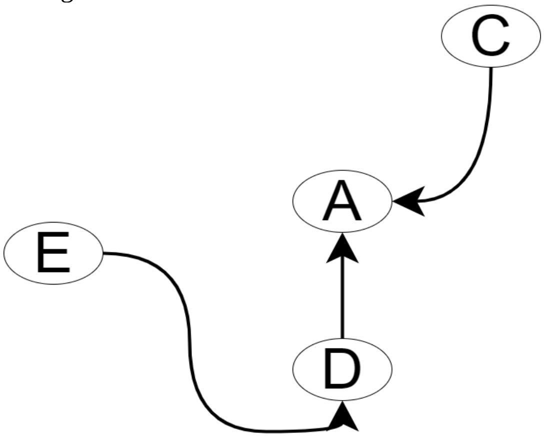
There are cases where you want to analyze a predetermined set of items. ARNs present an exploration focused on a single objective item. This exploration removes all rules that are not interesting in the context of the target item according to the minimum metrics of support and trust, showing the user only the relevant rules, but without certainty of the statistical dependence between the elements of the rules.

In order to propose a particular analysis of an objective item with probabilistic support of the influence of LHS on HRH, the Filtered-Association Rules Network (*Filtered-ARN)* was proposed, a strategy that improves the focus on visualization and filters, through gain and value aggregate, independent rules that are not perceived using only minimal trust and support (CALÇADA; REZENDE, 2019).

Following the same example as in Figure 1, if we consider that rule B →A has *Gain values* that do not satisfy chosen minimum values or null *Added Value*, this rule has no relevance, and its filtering occurs, thus modifying the network, as is shown in Figure 2.

As a strategy to improve the ARN, the Extended Association Rule Networks (ExARNs) algorithm was developed. The main objective of the ARN is to provide the analysis of rules with a focus on an itemset. ExARN proposes an improvement in knowledge extraction through the choice of N-items as the target of the network, adding value to situations that seek to find the direct and indirect correlation between two elements, especially when they are categorized as opposites, in our case, "positive" and "negative" for suicidality (PADUA et al., 2018).

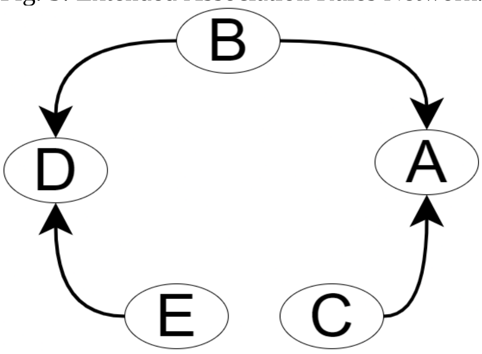
Fig. 2: Filtered-Association Rules Network.



Source: authors.

The construction of ExARN is done recursively. First, all items selected as objective item are modeled on the graph at level 0. Then, all rules that LHS items are not on the graph and have RHS items at level 0 are modeled in the network. The same process is done for all level 1 items, then for level 2 items. Until there are no more rules to be modeled (PADUA et al., 2018). An example would be the extended association rule network seen in Figure 3 with target nodes *D and A*.

Fig. 3: Extended Association Rules Network.



Source: authors.

In this research, it is proposed to combine *Filtered-ARN* with the application of ExARN to n-targets for the discovery of knowledge in *datasets*. *Filtered*-ExARN aims to combine the functionalities of *ExARN* (PADUA et al., 2018) and *Filtered*-ARN (CALÇADA; REZENDE, 2019). The use of asymmetric measures allows proportional directional analysis of the rules.

Its extension to n-target items facilitates the analysis of directly and indirectly proportional rules in a same network. For example, if we consider that rules E → D and C → A have null *Added Value or Gain* lower than an adopted minimum, they will be filtered in the execution of the algorithm, showing the direct influence of node *B* on the target nodes. The steps used for the construction are:

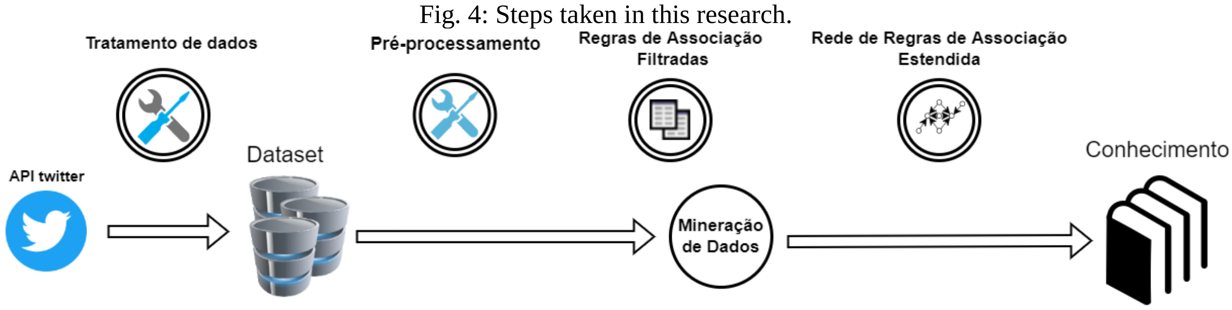
- **Construction of rules with asymmetric measures:** In this phase, rules with size 1 RHS are built using algorithms such as *Apriori*, *Apriori-TID* and *FP-Growth.* With the addition of gain and added value in the measurement calculation phase (CALÇADA; REZENDE, 2019).
- **Filtering the rules by asymmetric measures:** In this phase, the entire database is covered, excluding uninteresting rules such as: gain below the minimum and irrelevant aggregate value interval (CALÇADA; REZENDE, 2019).
- **Extended network construction:** Apply the extended rule network construction algorithm by selecting the target items. (PADUA et al., 2018)

4. METHODOLOGY

A study previously developed by (CARDOSO et al., 2020) produced the *dataset* used in this work. The study looked for relevant terms that are part of the vernacular of suicidal ideation indicated by previous related works (O'DEA et al., 2015) The *Application Programming Interface* (*API)* provided by *Twitter was used* to collect *tweets,* which were classified manually into two categories by professionals in the field, the categories chosen were "positive"' or "negative"' for suicidal ideation. The classified *tweets* totaled 699 instances, of which 211 were positive for suicidality and 488 were negative. In the next phase, the data went through common processes in text mining, which seek to transform unstructured data into structured data, seeking to represent the data in an understandable way for knowledge discovery algorithms. The techniques used were *tokenization,* punctuation removal, *stopword removal* and *stemming application*, with *bag-of-words* being the representation chosen with the simple frequency of each *token* present (CARDOSO et al., 2020).

As seen in Figure 4, the *dataset* produced by (CARDOSO et al., 2020) still did not produce the acceptable data representation for the input of the proposed *Filtered*-ExARN algorithm. The ".arff" file extensions have two possibilities of organization, one consisting of a matrix " *tweets* X terms" where only true values appear, in which each line has two values separated by ",", respectively the numeral 1 indicating that the term exists in the *tweet* followed by the numeral that indicates the index of the column in which the term is, the last part being the class to which the *tweet* belongs "positive" or "negative". The ".arff" format accepted by the the module used to extract the rules is formed by a matrix *tweets* X terms in which all the positions of the matrix appear, being 0 if the term is not in the *tweet* and 1 if it is in the *tweet,* the last column being the classification "positive" or "negative"".

Fig. 4: Steps taken in this research.



Source: authors.

Transformed the *dataset* into a valid format that the association rules module can operate on. It was followed as seen in Figure 4 for the Association Rules Extraction phase, it was developed in Java, using the *Apriori-TID algorithm* as a standard for calculating the measures. The module produces all types of Association Rule Networks, including the one proposed in this article, *Filtered*-ExARN.

For the extraction of rules, it was established that minimum support and trust were null, so that all possibilities and all terms were considered in the knowledge extraction process. The minimum gain in the filtering phase was 0.001, in addition to being automatically excluded all rules with null value added. Initially, 6,841,883 rules were extracted. After applying the filters, the number of rules was reduced to 4,107,190.

The last stage of the data mining phase involves building the networks. "Positive" and

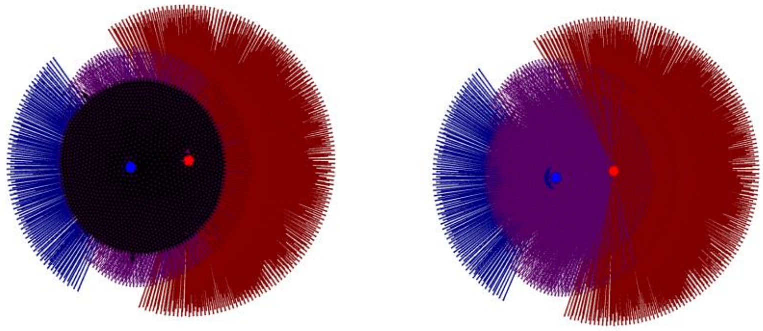
"Negative" target nodes were assigned and *Filtered*-ExARN generated. After formatting the networks, the *GEPHI* software was used for the graphic construction in order to have a visualization of the results. The most important items are those that are directly connected to the target items; therefore, three categories of items were listed: items not directly linked to the target nodes (above level 1), exclusively linked to one of the two (determining items), and linked to both simultaneously (dominant items).

As a comparison algorithm to validate the technique, the selection of attributes was used. Subsequently, the results found were verified in which category they belong. In this way, the CFS (*Correlation Based Feature Selection)* attribute selection algorithm (KAREGOWDA; MANJUNATH; MA, 2010) was selected, provided by the *WEKA package.* The output of the algorithm is a list of selected terms, which have relevance in the *dataset.* Verification of terms and analysis of the adopted categories was performed to verify that the terms of the CFS match the terms obtained by *Filtered*-ExARN.

5. RESULTS AND DISCUSSIONS

Observing the generated network (Figure 5a), several attributes are connected to both situations "[class=positive]" and "[class=negative]", totaling 2,103 terms that, for this reason, cannot be used to differentiate suicidal ideation in *tweets.* It can be inferred that these terms should be disregarded in processes that seek to classify this type of text. The use of filters for objective measures of added value and gain generated hypotheses with a higher probability of veracity (CALÇADA; REZENDE, 2019).

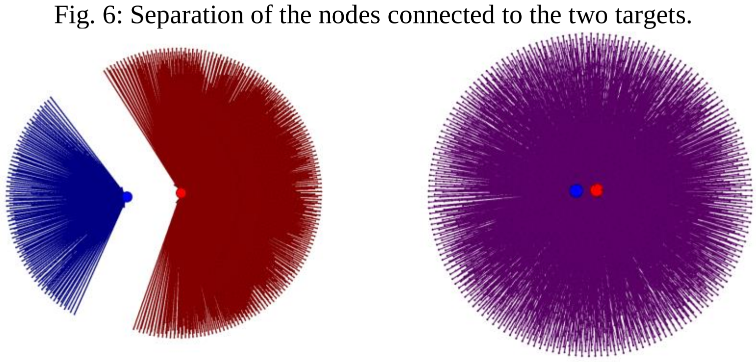
Fig. 5: Removal of nodes that are not directly linked to the target nodes.



(a) All nodes. (b) Only nodes linked to the target nodes.
Source: authors.

For a better analysis, terms that are not directly linked to the target nodes were removed (Figure 5b). The *Filtered*-ExARN generated shows only the terms represented by nodes (LHS) that directly target the classes (RHS) "[class=positive]" and "[class=negative]". However, terms that imply both target nodes such as **existenc, univers, famil** and **louc** demonstrate ambiguity when pointing to the positive class and the negative class, despite the filtering technique demonstrating the relevance of these items. The list of nodes visualized in Figure 6b are the dominant terms, in Figure 6a we can observe the separation of the terms that tend exclusively to the positive class, which indicates suicidal tendency, and the terms that are exclusively linked to the negative class, thus, the that indicate that the *tweet* does not have a suicidal tendency, form the set of determinant items.

Fig. 6: Separation of the nodes connected to the two targets.

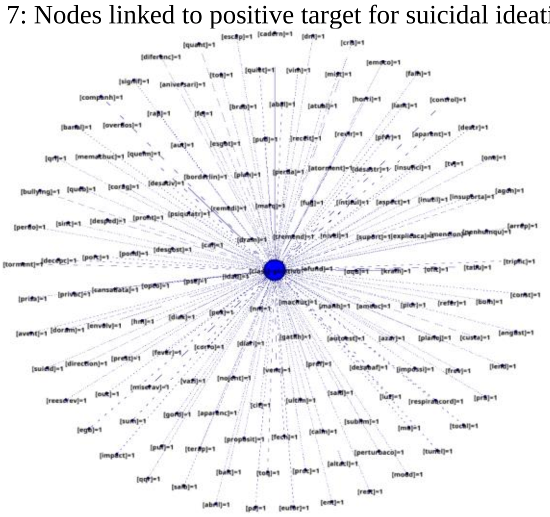


(a) Nodes linked to only one of the targets. (b) Nodes attached to both targets.
Source: authors.

Analyzing Figure 7 of all terms linked exclusively to "[class=positive]", one can see probabilistically interesting terms mined using the *Filtered*-ExARN technique. It is worth highlighting some of these terms such as: psiquiatr, dram, queim, remedi, overdos, fug, miserav, suport, corro, torment, insufficient, gord, disgusting, machuc, perd, terap, disaster. These terms stand out because they have domains that represent sentimental and emotional aspects, such as anger, anxiety, sadness, grief and hate, for example. These terms can be listed due to the particular emotive nature of the detection of suicidal ideation. In fact, emotions such as fear, anger and aggression are prominent in suicidal communication (HASSINE; ABDELLATIF; YAHIA, 2022).

Comparing the results obtained from the *Filtered*-ExARN with the selection of attributes of the CFS algorithm, it was observed the gain of analysis due to the inherent directionality of the *Filtered*-ExARN, more precisely which terms tend towards suicidal ideation and which terms tend towards non-suicidal ideation in the Cardoso *dataset* (CARDOSO et al., 2020). The terms selected by CFS were: *a, vai, q, fot, kk, com, dorm, will, dream, tip, sei, suicide, era, so, got, continu, pul, bost, suic, cart, cris, sinc.* Of these terms, only **suicide** and **sinc** are exclusively linked to the "positive" class, with the terms: "fot", "kk", "com", "tip", "era", "got", "pul" exclusively linked to the class "negative" for suicidal ideation. The rest of the terms appear in texts of both classes.

Fig. 7: Nodes linked to positive target for suicidal ideation.



Source: authors.

The results found show that of the terms selected by the CFS algorithm, 36.36% of the selected attributes belong to the negative class for suicidal ideation, only 0.9% belong to the positive class for suicidal ideation, and 54.54% of the terms are irrelevant because they demonstrate ambiguity linked to the two targets., or not be linked directly to any of the targets. Therefore, *Filtered*-ExARNs have a capacity to characterize items related to a certain class that is much superior to the selected attribute selection algorithm.

6. CONCLUSION

The proposed algorithm generated a more consistent and effective knowledge extraction than the attribute selection algorithm. It is possible to make several cuts of irrelevant rules both in the extraction phase and in the preview phase. The vectorization of the results allows the direct analysis of the impact of the rules on the targets, having the statistical support acquired by the filtering of the asymmetric rules.

The results, available online[1], can be used to improve the results of other classification algorithms, as well as serve as a basis for psychological and behavioral studies. The *Filtered*-ExARN *technique* proved to be extremely efficient, since it manages to filter the rules and assist in the analysis of n-target items simultaneously.

It is worth mentioning the requirement of high computational resources for text mining processes, so the discovery of terms with greater impact makes it possible to optimize the application of other techniques. Having scarcity in studies of this type of technique in text mining, the results were acceptable. With this, promising future works are presented, such as: construction of the weights of the network using asymmetric measures, comparison of optimization with test cases that have more severe measures, application in *datasets* of other types of applications, expansion of the *dataset.*

[1] https://drive.google.com/drive/folders/1sipwiQTSdn1JDGOWbfvW0zswaCOgcoFd?usp=sharing

References

ABDULSALAM, Asma; ALHOTHALI, Areej. **Suicidal Ideation Detection on Social Media: A Review of Machine Learning Methods**, jan. 2022. Disponível em: http://arxiv.org/abs/2201.10515.

AGRAWAL, Rakesh; IMIELINSKI, Tomasz; SWAMI, Arun. **Mining Association Rules between Sets of Items in Large Databases**. v. 2. Issue 1, 1993.

CALÇADA, Dario Brito; REZENDE, Solange Oliveira. **Filtered-ARN: Asymmetric objective measures applied to filter Association Rules Networks**. v. 22. 2019.

CARDOSO, Vinícius et al. **Detecting Suicidal Ideation on Tweets**. In: p. 178–189. doi: 10.5753/eniac.2019.9282.

FUKUDA, Takeshi et al. **Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization.** IN: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada. 1996.

HAQUE, Rezaul et al. **A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies**, MDPI AG, v. 10, p. 57, 3 abr. 2022. doi: 10.3390/technologies10030057.

HASSINE, Mohamed Ali Ben; ABDELLATIF, Safa; YAHIA, Sadok Ben. **A novel imbalanced data classification approach for suicidal ideation detection on social media**. Computing, Springer, v. 104, p. 741–765, 4 abr. 2022. issn 14365057. doi: 10.1007/s00607-021- 00984-0.

JASHINSKY, Jared et al. **Tracking suicide risk factors through Twitter in the US. Crisis**, v. 35, p. 51–59, 1 2014. issn 02275910. doi: 10.1027/0227-5910/a000234.

KANCHARAPU, Rohini; SRINAGESH, A; BHANUSRIDHAR, M. **Prediction of Human Suicidal Tendency based on Social Media using Recurrent Neural Networks through LSTM**. In: 2022 International Conference on Computing, Communication and Power Technology (IC3P). 2022. P. 123–128. doi: 10.1109/IC3P52835.2022.00033.

KAREGOWDA, Asha; MANJUNATH, A.; M.A, Jayaram. **Comparative study of attribute selection using gain ratio and correlation based feature selection**. Int J Inf Technol Knowl Manage, v. 2, jan. 2010.

O'DEA, Bridianne et al. **Detecting suicidality on Twitter. Internet Interventions**, v. 2, n. 2, p. 183–188, 2015. issn 2214-7829. doi: https://doi.org/10.1016/j.invent.2015. 03.005. Disponível em: https://www.sciencedirect.com/science/article/pii/ S2214782915000160.

PADUA, Renan De et al. **Exploring the data using extended association rule network**. In: p. 330–335. isbn 9781538680230. doi: 10.1109/BRACIS.2018.00064.

PANDEY, Gaurav et al. **Association rules network: Definition and applications. Statistical Analysis and Data Mining**, v. 1, p. 260–279. 2009. doi: 10.1002/sam.10027.

RITCHIE, Max Roser Hannah; ORTIZ-OSPINA, Esteban. Suicide. **Our World in Data**, 2015. https://ourworldindata.org/suicide. Disponível em: https://ourworldindata.org/sui cide#licence.

ROY, Arunima et al. **A machine learning approach predicts future risk to suicidal ideation from social media data. npj Digital Medicine, Nature Research**, v. 3, 1 dez. 2020. issn 23986352. doi: 10.1038/s41746-020-0287-6.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining Instructor's Solution Manual**. Pearson Addison-Wesley. 2006