Computing (2022) 104:741–765
https://doi.org/10.1007/s00607-021-00984-0

SPECIAL ISSUE ARTICLE

A novel imbalanced data classification approach for suicidal ideation detection on social media

Mohamed Ali Ben Hassine[1] · Safa Abdellatif[2] · Sadok Ben Yahia[2]

Received: 6 March 2021 / Accepted: 6 July 2021 / Published online: 17 August 2021

Abstract
Suicide has become a serious social health issue in modern society. Suicidal ideation is people's thoughts about committing or planning suicide. Many factors, such as long-term exposure to negative feelings or life events, can lead to suicidal ideation and suicide attempts. Among these approaches to suicide prevention, early detection of suicidal ideation is one of the most effective ways. Using social networking services provides a platform for people to express their sufferings and feelings in the real world, which provides a source for a deeper investigation into models and approaches for the detection of suicidal intent to enable prevention. This paper addresses the early detection of suicide ideation through the associative classification approach applied to Twitter social media. However, since the number of suicide intention tweets is tiny compared to the number of all the tweets, this leads us to an imbalanced classification problem, in which, the minority class (suicide intention) is more important than the majority class (no suicide intention). In such a situation, classical classifiers usually yield very inaccurate results regarding minor classes, since they can easily discover rules predicting the majority class and overlook those related to the minor. This paper aims to contribute to this line of research by introducing a new interestingness measure to enhance the classification process. This measure highlights the two classes regardless of their imbalanced distribution. Carried out experiments proved that the adapted CBA outweighs in terms of prediction accuracy the original one, and other pioneering baseline classification approaches.

✉ Sadok Ben Yahia
sadok.ben@taltech.ee

Mohamed Ali Ben Hassine
mohamedali.benhassine@fst.utm.tn

Safa Abdellatif
Safa.abdellatif@fst.utm.tn

[1] Faculty of Sciences of Tunis, University of Tunis El Manar, LIPAH-LR11ES14, El Manar, 2092 Tunis, Tunisia

[2] Department of Software Science, Tallinn University of Technology, Akadeemia tee 15a, 12618 Tallinn, Estonia

Keywords Suicidal ideation detection · Feature extraction and selection · Associative classification · Imbalanced datasets

Mathematics Subject Classification 06-08

1 Introduction

According to the world health organization, close to 800,000 people die because of suicide every year, which is one person every 40 seconds. Suicide is a global phenomenon and occurs throughout the lifespan. Effective and evidence-based interventions can be implemented at population, sub-population, and individual levels to prevent suicide and suicide attempts. There are indications that for each adult who died by suicide, there may have been over 20 for others attempting suicide. People with depression are highly likely to commit suicide, but many without depression can also have suicidal thoughts. Many factors, such as long-term exposure to negative feelings or life events, can lead to suicidal ideation (*aka* suicidal thinking), and suicide attempts. Generally, suicide utters were significantly more likely to leave a suicide note than attempters [9]. Thus, any written suicidal sign is viewed as a worrying sign, and an individual should be questioned on the existence of individual thoughts.

Owing to the advances in social media and online anonymity, an increasing number of individuals turn to interact with others on the Internet. Online communication channels are becoming a novel way for people to express their feelings, suffering, and suicidal tendencies. Hence, online channels have naturally acted as a surveillance tool for suicidal ideation, and mining social content can provide a source for a deeper investigation into models and approaches for the detection of suicidal intent to enable prevention. Among all the ways of suicide prevention, early detection of suicidal ideation is one of the most effective ones. It determines whether the person has suicidal ideation or thoughts by a given tabular data of a person or textual content written by a person.

Many approaches have been proposed to deal with suicidal ideation detection, ranging from traditional classification approaches, deep learning, traditional machine learning to AI technologies [19]. This paper addresses the early detection of suicide ideation through the associative classification approach applied to Twitter social media. In fact, Associative classifiers extract first association rules [5] between patterns and then choose a subset of them with a pre-specified column called *class label* to predict the class of a record [2]. However, one of the major problems of associative classifiers is their inability to manage the huge number of class association rules generated from real-world datasets and to capture the most valuable and potentially gainful ones. To overcome this problem, an evaluation and a selection of association rules have to be performed. To do so, several interestingness measures have been proposed in the literature to sort and select association rules based on various sights and concerning some pre-defined goals. Another issue related to datasets used in suicidal ideation detection is the number of suicide intention tweets, which could be considered very small compared to the number of all the tweets. This leads us to an imbalanced classification problem, in which the minority class (suicide intention) is more important than the majority class (no suicide intention).

As a rule of thumb, in the classification's case of imbalanced datasets [23], where the class of interest is relatively rare compared to the other ones, the existing measures are no more beneficial for the sorting or the selection of association rules [18]. In fact, two scenarios are plausible: (*i*) existing measures favor rules belonging to major classes and consider others as uninteresting; and (*ii*) they are focusing on the rules of minor classes without considering ones of the major classes, which will badly affect the whole global accuracy of the Associative Classifier. To overcome this shortcoming and be able to keep highly interesting rules from both classes regardless of their imbalanced distribution, we introduce five novel interestingness measures that will be applied respectively as a primary criterion for rule selection. These measures offer a high recall on the minority class while maintaining a high precision over the major class.

Experiments carried out regarding three assessment measures showed evidence backing up our claims that the newly proposed measure offers a stronger predictive power compared to the existing ones and compared to other several well-known classification approaches.

We organize the rest of this paper as follows. Section 2 reminds the basic concepts related to association rule-based classification, interestingness measures, and imbalanced datasets. Section 3 underscores the limits of the existing measures and painstakingly describes the introduced ones. Section 4 is dedicated to the application of the novel measures in the suicidal ideation detection field. We report the results of the experiments in Sect. 5. We allude to our takeaway messages and sketch issues of future works in Sect. 6.

2 Background and related work

We provide, in the following, the basic concepts related to this work.

2.1 Classification based on association rules

Let D be a training dataset containing $|D|$ instances and $\mathcal{I}$ a set of *m+1* distinct items with $\mathcal{I} = \{i_1, i_2, \ldots, i_{m+1}\}$. An Association Rule (AR) is the pattern of the form $X \rightarrow Y$ where X and Y are non-intersecting subsets of $\mathcal{I}$. Using an AR mining technique on D, frequent itemsets are mined and Class Association Rules (CARs) are extracted where X is a subset of items and Y is a class label.

Associative Classification (AC) is a rule-based approach proposed to classify by first discovering a complete set of CARs from the training set and then using it to predict class labels for objects with an unknown class [2]. An AC is composed of three major phases which are generating, filtering, and selecting rules for class prediction. The generating phase comprises extracting frequent rules based on a measure m and a threshold t. The second phase comprises ranking and pruning the complete set generated during the first phase to discard worthless rules. The last phase comprises selecting a rule among the collection of CARs kept predicting the class label of new objects. In the following subsection, we briefly sketch the interestingness measures used in association rule mining.

2.2 Interestingness measures in association rule mining

The aim of an AC is to extract knowledge from data by detecting interesting associations between patterns in a given database. However, one of the problems of AC is that it can generate a huge number of ARs including a lot of uninteresting ones. Delving through this huge number of ARs in order to identify the most interesting ones stands in the furthest from being a straightforward task. To solve this problem, it's of utmost importance to establish a "well-established" Interestingness Measure (IM) to evaluate the quality and importance of an AR.

AC aims to extract knowledge from data by detecting interesting associations between patterns in a database. However, the key drawback of AC is that it can generate an overwhelming number of ARs including a batch of pointless ones. Delving through this huge number of ARs to identify the most interesting ones stands in the furthest from being a straightforward task. To solve this problem, it's of utmost importance to establish a "well-established" Interestingness Measure (IM) to assess the quality and importance of an AR.

Several IMs have been proposed in order to unveil gainful ARs. We can classify these measures into three main categories: *objective*, *subjective* and *semantics-based* ones [13]. Objective measures are neither application-specific nor user-specific. They are quantitatively representable, relatively visual, easy to operate, and only depend on the input raw data. Whereas, the subjective ones are those that take into consideration both the data and the user who leverages the data. However, unlike objective IMs, they may not be representable by an easily interpretable mathematical formula. Semantic-based measures are special types of subjective measures that take into consideration the semantics of a pattern that is domain-specific. Without loss of generality, in the remainder of this paper, we only focus on objective measures.

There are many objective interesting measures available in the literature [13]. One of the most used interestingness measures is the *support*, which gauges how often the itemset appears in a single transaction in the database.

$$Support(A \rightarrow C) = P(A \cap C) \tag{1}$$

Another worthy of mention criterion is the *confidence*, which is the conditional probability of having C given A.

$$Confidence(A \rightarrow C) = \frac{P(A \cap C)}{P(A)}. \tag{2}$$

Conviction is another interestingness criterion that assesses the deviation from independence by considering outside negation.

$$Conviction(A \rightarrow C) = \frac{P(A)P(\overline{C})}{P(A\overline{C})} \tag{3}$$

Laplace is also a measure commonly used for classification tasks. It is a confidence estimator that takes the support to compute the Laplace value. As far as the value of the support decreases, the value of Laplace also decreases.

$$Laplace(A \rightarrow C) = \frac{N(AC) + 1}{N(A) + 2} \tag{4}$$

The lift measures the dependency between A and C. It reflects the positive/negative correlation of antecedent and consequence of the rule.

$$Lift(A \rightarrow C) = \frac{Confidence(A \rightarrow C)}{P(C)} = \frac{P(AC)}{P(A)P(C)} \tag{5}$$

To unveil some broad strokes of this large number of measures, they have proposed some properties in the literature. We consider, in the following, three sets of properties for the IM related to a rule as $A \rightarrow C$. Piatetsky-Shapiro [25] proposed three prime properties that are desirable to be fulfilled by any IM.

- P_1 : $IM = 0$ if A and C are independent, i.e., $P(AC) = P(A)P(C)$.
- P_2 : *IM monotonically increases with* $P(AC)$ *when* $P(A)$ *and* $P(C)$ *remains the same.*
- P_3 : *IM monotonically decreases with* $P(A)$ $($*or* $P(C))$ *when* $P(AC)$ *and* $P(C)$ $($*or* $P(A))$ *remain the same.*

Tan et al. [29] proposed five other properties which are, unlike Piatestky-Shapiro's ones, not desirable in every IM and used to classify IM into different categories. These properties are thoroughly described in [13].

Lenca et al. [21] have also proposed five properties for the IM evaluation, namely :

- L_1: *An IM is constant if there is no counterexample to the rule.*
- L_2:*An IM decreases with* $P(A\overline{C})$ *in a linear, concave or convex fashion around* 0^+.
- L_3: *An IM increases as far as the total number of records also increases.*
- L_4: *The threshold is easy to set.*
- L_5: *The semantics of the IM is easy to express.*

A myriad of other of utmost importance factors, influencing the interestingness of an association rule, need to be taken into consideration. These factors include the imbalance of class distribution and the cost of misclassification. In the following, we pay heed to the imbalance of class distribution factor.

2.3 Imbalanced datasets classification

Dealing with class imbalance has become a popular problem faced in data mining [22,30]. Several studies around the world have steadily treated this issue for years. According to Bao-Gang et al. [18], a dataset is called *imbalanced*, whenever the number of records of one class (called the *major* or *negative* class) is highly surpassing the number of records belonging to the other class (called the *minor* or *positive* class). Ironically, this class (the minor one) is often the sought-after one in many applications and should be rightly recognized.

One very popular approach to addressing imbalanced datasets is to over-sample the minority class, to wit ”artificially” duplicating examples in the minority class, although these examples don't add any new information to the model. Doing so was referred to in the literature as the Synthetic Minority Over-sampling Technique (SMOTE) [17]. In the remainder, we avoid such artificial over/under-sampling by adapting the interesting measures to cope with the imbalance of classes.

The class imbalance problem has become more and more marked while applying machine learning algorithms to real-world problems. These applications range from medical diagnosis, text classification, fraud detection, sentiment analysis, to name but a few. It is worthy of mention that classical classifiers cannot handle the problem of imbalanced datasets, since they mistakingly assume a balanced class distribution and do not consider the minority classes. For example, in a medical diagnostic problem where the disease cases are usually quite rare as compared with normal populations, the recognition goal is to detect people with these diseases. Hence, a genuine classifier is the one that provides a high accuracy on the disease class [1]. However, in this scenario, classical classifiers usually yield very inaccurate results regarding minor classes, since they can easily discover rules predicting the majority class and consider rules predicting minor classes as uninteresting ones. We root this problem back in using some specific IMs, during the rule selection phase, which are not suitable for the classification of imbalanced datasets.

3 Proposed interestingness measures

We underscore, in the following, the limits of the existing measures. Then we introduce the novel ones, and we scrutinize their theoretical properties.

3.1 Limits of the existing measures

The dedicated literature witness a myriad of interestingness measures (IM) to improve the associative classifiers by managing the huge number of rules generated from real-world datasets and capturing the most significant ones. These IMs are not suitable for the extraction of important association rules from imbalanced datasets. In fact, those proposed in the literature tend either to emphasize the extraction of rules belonging to major classes and consider others as uninteresting. Alternatively, they focus on the extraction of rules belonging to minor classes and ignoring rules from major classes, which badly affects the whole global accuracy of the classifier.

One of these IMs is the confidence, that has the downside of its inability to actually select interesting rules. In fact, the latter only considers the conditional probability of rules. Taking the example of a rule $A \rightarrow C$ and if $N(A) = 10$, $N(C) = 9,000$, $N(AC) = 9$ and $N = 10,000$ where N is the size of data and $N(X)$ denotes the frequency of X, then the confidence value of the rule $A \rightarrow C$ is 90%.

Based on this high value of confidence, the association rule $A \rightarrow C$ will be considered as an interesting one. However, this is the furthest from the truth according to the first property of Piatetsky-Shapiro, which shows that A and C are independent (i.e. P(AC)=P(A)P(C)) and the obtained value of confidence is exactly equal to the probability of C regardless of A. Finally, this association rule might be just drawn from noise. Furthermore, the Laplace measure is not considered as a flawless one when selecting rules in case of the classification of imbalanced datasets. Indeed, it heavily relies on support, that is as far as the value of the support decreases, the value of Laplace also decreases. That is to say, assigning a high value to Laplace may lead to only select rules having obvious knowledge and ignore other ones. Besides, choosing a low value of Laplace yields to select an overwhelming number of rules which may be noisy and redundant. The Conviction is another well-known IM proposed in the literature. This measure is also not helpful for rule selection in case of imbalanced datasets classification. In fact, as presented in Equation 3, the rule $A \rightarrow C$ is considered as interesting whenever the conviction value surpasses 1. However, taking the example of a binary imbalanced dataset having C_1 as the major class and C_2 as the minor one. We have the conviction's value for the rule $A \rightarrow C_2$ is equal to $conviction(A \rightarrow C_2) = \frac{Sup(A)Sup(\overline{C_2})}{Sup(A\overline{C_2})}$. Considering $sup(C_1) \simeq 1$ and $\overline{C_2} = C_1$, we have $sup(A\overline{C_2}) = sup(AC_1) \simeq sup(A)$ and $conviction(A \rightarrow C_2) \simeq 1$. By and large, the conviction measure cannot also select significant rules from minor classes.

Unlike the above IMs, the lift measure favors rules from minor classes. In fact, we consider an association rule as significant whenever its lift's value highly surpasses 1. However, for a major class C_1, the rule $A \rightarrow C_1$ has as a lift value $lift(A \rightarrow C_1) = \frac{sup(AC_1)}{sup(A)sup(C_1)} \simeq 1$ since $sup(C_1) \simeq 1$ and $sup(AC_1) \simeq sup(A)$.

As a result, the lift measure may hardly select any significant rules from major classes. However, it is beneficial for the selection of association rules from minor classes. The lift measure has a major drawback, that is its symmetry. Bluntly, for an association rule $A \rightarrow C$, the lift measure takes A and C in an equivalent position, which is not true in the case of imbalanced datasets.

3.2 New interestingness measure description

In this sub-section, we introduce alternative IMs intending to overcome the above cons by looking for selecting highly interesting rules from both types of classes regardless of their imbalanced distribution.

The first introduced measure is the $ModifiedLift$ and is defined as follows:

$$ModifiedLift(A \rightarrow C) = \frac{lift(\overline{A} \rightarrow \overline{C})}{lift(A \rightarrow \overline{C})} \times \frac{P(\overline{A})}{P(A)} = \frac{P(\overline{AC})}{P(A\overline{C})} \qquad (6)$$

We can see the first part of the Equation 6 as the lift measure of the rule $\overline{A} \rightarrow \overline{C}$ divided by the lift measure of $A \rightarrow \overline{C}$. In fact, as long as the value of the lift measure of $\overline{A} \rightarrow \overline{C}$ increases, both of $\overline{A}$ and $\overline{C}$ are considered more and more dependent. This means that A and C are likely to be dependent too.

Besides, the higher the value of the lift measure of $A \rightarrow \overline{C}$ is, the more independent A and C are. Therefore, we have proposed to set the lift measure of $\overline{A} \rightarrow \overline{C}$ as nominator and the lift measure of $A \rightarrow \overline{C}$ as a denominator. This ratio could indefinitely increase. To overcome this issue, we suggest making some improvements by multiplying the first ratio by a corrective ratio. This latter comprises the no occurrence probability of the antecedent $P(\overline{A})$ divided by the occurrence probability of the antecedent $P(A)$.

After the multiplication of these two ratios, we get as a result a new ratio which consists on the probability of observing $\overline{A}$ and $\overline{C}$ together, $i.e.P(\overline{AC})$ divided by the probability of observing A and $\overline{C}$ together, $i.e.P(A\overline{C}$. This measure yields higher values whenever $P(\overline{AC})$ surpasses $P(A\overline{C})$.

In the following, we provide evidences backing up our posit that the $ModifiedLift$ measure is suitable for the classification of imbalanced datasets.

Considering the example of a binary imbalanced datasets, having C_1 as the major class and C_2 as the rare one, the *ModifiedLift* of the rule $A \rightarrow \overline{C_1}$ is equal to:

$$ModifiedLift(A \rightarrow C_1) = \frac{P(\overline{AC_1})}{P(A\overline{C_1})} = \frac{P(\overline{A}C_2)}{P(AC_2)} \simeq \frac{P(C_2)}{P(C_2)} \simeq 1$$

This assessment is explained as: *(i)* $\overline{C_1} = C_2$ and *(ii)* C_2 is the rare class, then, in the major class, the number of instances including C_2 is much more rare than the number of instances including A, i.e. $P(AC_2) = P(C_2)$ and $\overline{A}$, i.e. $P(\overline{A}C_2) = P(C_2)$.

The probability of observing A and C_2 together or $\overline{A}$ and C_2 together will have as a maximum value the number of instances of C_2, that is $P(C_2)$.

In the same vein, the rule $A \rightarrow \overline{C_2}$ has as a $ModifiedLift$ value equal to:

$$ModifiedLift(A \rightarrow C_2) = \frac{P(\overline{AC_2})}{P(A\overline{C_2})} = \frac{P(\overline{A}C_1)}{P(AC_1)} \simeq \frac{P(\overline{A})}{P(A)}$$

Based on the estimation shown above, we can notice that the value of the $ModifiedLift$ turns around 1 for association rules extracted from minor classes. For rules extracted from minor classes, this measure depends on both the occurrence and no occurrence of the antecedent part. Formally, we have : for major classes,

$$ModifiedLift(A \rightarrow C_1) \simeq 1$$

and for minor classes,

$$ModifiedLift(A \rightarrow C_2) < 1 \text{ if P(A)} > P(\overline{A}) \; ModifiedLift(A \rightarrow C_2) > 1 \text{ if P(A)} < P(\overline{A})$$

Therefore, it is clearly noticed that if we have a low support of the antecedent we, then have a high value of the $ModifiedLift$.

Fortunately, this is gainful for the classification of imbalanced datasets. In fact, for this type of data, and minor classes specifically, we prefer to extract specific rules since they provide and unveil new knowledge. As a rule of thumb, these rules have a long antecedent A and consequently a low support of A.

This measure inversely depends on the value of $P(AC)$ which is suitable in case of imbalanced data sets since the value of $P(AC)$ for minor classes is generally lower than the value of $P(AC)$ for major classes.

DM_2 and DM_3 also derive from $ModifiedLift$ and inversely depend on the probability of occurrence of the conclusion $P(C)$ which is also convenient for the case of imbalanced data sets since $P(C)$ of minor classes is much less than $P(C)$ of major classes. We defined these two measures as follows:

$$DM_2(A \rightarrow C) = ModifiedLift(A \rightarrow C) * \frac{1}{\sqrt{P(C)}} = \frac{P(\overline{AC})}{P(A\overline{C})\sqrt{P(C)}} \tag{7}$$

$$DM_3(A \rightarrow C) = ModifiedLift(A \rightarrow C) * \frac{P(A)}{P(C)} = \frac{P(\overline{AC})P(A)}{P(A\overline{C})P(C)} \tag{8}$$

Last but not least, DM_4 is also a derived measure from $ModifiedLift$ but it inversly depends on the probability of occurrence of the premise $P(A)$. This could be suitable for imbalanced data sets since it favors rules having low $P(A)$, which generally correspond to long antecedents. DM_4 is defined as follows :

$$DM_4(A \rightarrow C) = ModifiedLift(A \rightarrow C) * \frac{1}{\sqrt{P(A)}} = \frac{P(\overline{AC})}{P(A\overline{C})\sqrt{P(A)}} \tag{9}$$

3.3 Scrutiny of the formal properties

Table 1 sketches seven properties mentioned in Sect. 2.2 for the new proposed measure and some well-known objective IMs that will be used thereafter for the comparison. By "Yes", we mean that the measure fulfills that property.

For mining and selecting association rules in an imbalanced dataset, P_2, P_3, and L_1 are the more sought-after properties to be fulfilled by an IM.

In fact, the measure that complies with the property P_2 is significant for the classification of imbalanced datasets since the association between rare variables A and B becomes more interesting with the increase of $P(AB)$ while keeping both $P(A)$ and $P(B)$ constant.

Moreover, an IM that fulfills P_3 is considered suitable for the classification of imbalanced datasets. Taking the example of two rare variables A and B, if $P(A)$ is increased while maintaining $P(AB)$ constant, A no longer remains as a rare variable. Thence, the association between a rare and a frequent variable is less interesting than an association between two rare variables.

L_1 is also an interesting property to be fulfilled by the interestingness measure. In fact, associations that have a *confidence* value equal to 1, should have the same

Table 1 Properties interestingness measures

Measures	P1	P2	P3	L1	L2	L3	Symmetric
ModifiedLift	No	Yes	Yes	No	Yes	Yes	No
DM_2	No	Yes	Yes	No	Yes	Yes	No
DM_3	No	Yes	Yes	No	Yes	Yes	No
DM_4	No	Yes	Yes	No	Yes	Yes	No
Confidence	No	Yes	Yes	Yes	Yes	No	No
Laplace	No	Yes	Yes	No	Yes	yes	No
Lift	Yes	Yes	Yes	No	Yes	No	Yes
Conviction	Yes	Yes	Yes	Yes	Yes	No	No

value of interestingness regardless of the *support* which is gainful for the association between rare variables.

4 Use case: suicidal ideation detection

We devote this section to the validation of the efficiency of the proposed measures in a real-world domain, which is the text mining domain and precisely the suicidal ideation detection in social media.

4.1 Why suicidal ideation detection?

According to the World Health Organization (WHO), worldwide, someone dies by suicide every 40 seconds, i.e., every year close to $800,000$ people take their own life and there are many more people who attempt suicide. Every suicide is a tragedy that badly affects families, communities, and entire countries and has long-lasting effects on the people left behind. Suicide occurs throughout the lifespan and is sadly considered as the second leading cause of death among 15 and 29 years old. [1]

The WHO recently approved several universal suicide prevention interventions, including two promising strategies targeting vulnerable groups and individuals, and facilitating their access to crisis helplines. The timely identification of these groups and vulnerable individuals and the balance of identifying high-risk cases, without too many false positives, is still a thriving challenge. This has led to increasing efforts in clinical settings, which further increases the financial and temporal costs. For these reasons, public health set as its utmost priority to explore novel approaches and identify people at risk of suicide without increasing costs or adding burdens to the existing clinical systems. This effort could benefit from the introduction and proliferation of new social media technologies.

Social media has provided researchers with alternative ways to use automated methods of natural language analysis and to better understand the thoughts, feelings, beliefs, behaviors, and personalities of individuals. Studies on data-based methodologies rely-

[1] https://www.who.int/news-room/fact-sheets/detail/suicide.

ing on language use have proven useful for monitoring psychological states and public health issues such as influenza, cardiovascular disease, alcohol problems, and smoking. Infodemiology or infotainment is an emerging field related to data-driven computing methodologies and other studies that use social media to utterly understand and monitor the taking root of this plaguing health problem [28]. Twitter is a social media application that allows users to broadcast news, information, and personal updates to other users (subscribers) in tweets or statements of 140 characters or fewer. We see these statements as important markers that provide a general impression of the psychological states of social media users. Hence, social networks, especially Twitter, offer the opportunity to take advantage of this information to explore public health issues, in particular depression and suicide risk.

Proactively detecting suicide is one of the most effective ways to drastically reduce suicidal rates. For this reason, we are paying close heed to this search field.

4.2 Data collection and pre-processing steps

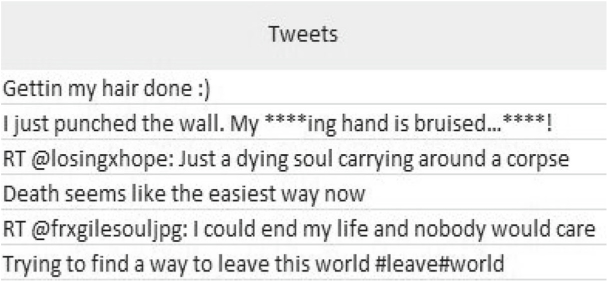
The prediction of suicidal ideation, from Twitter posts, comprises five major steps as shown in Fig. 3. These phases are: *(i)* data extraction from Twitter; *(ii)* preprocessing the text within a tweet; *(iii)* feature extraction from preprocessed tweets for the suicidal ideation identification; *(iv)* evaluation and selection of best and relevant features that can improve the performance of the classification model and *(v)* Classification and identification of tweets exhibiting suicidal ideation. We thoroughly describe each phase below.

Please note that we will present an illustrative example of real tweets and within each step, we will sketch the transformation made.

4.3 Step 1: data gathering

Traditionally, finding data on individuals having a mental illness and suicidal ideation is an arduous task, because of the social stigma placed upon mental health. However, nowadays people are turning to the anonymity of the internet to express their frustration, discuss mental health issues, and seek help [8,27].

For the sake of fringing any individual privacy disclosure, neither a direct quote from any data is shown, nor any identifying information. A dataset of distressed and no distressed Twitter users' IDs was kindly provided by [31]. In fact, a Twitter API was used to collect tweets based on key phrases related to suicide which are the risk factors defined by the American Psychiatric Association (APA) [3] and the warning signs identified by the American Association of Suicidology (AAS) [26]. The authors of these tweets were randomly investigated to extract distressed users having suicidal ideation and writing frequently about depression, suicide, and self-harm. Many everyday users were also randomly collected. Professionals with expertise in mental health research validated the selection of distressed and everyday users. Based on this dataset, we have randomly extracted 100 tweets from distressed users and 160 tweets from everyday users for 260 tweets. These tweets were subsequently reviewed and validated by three annotators. These annotators annotate tweets by attributing ”Yes”



Tweets
Gettin my hair done :)
I just punched the wall. My ****ing hand is bruised...****!
RT @losingxhope: Just a dying soul carrying around a corpse
Death seems like the easiest way now
RT @frxgilesouljpg: I could end my life and nobody would care
Trying to find a way to leave this world #leave#world

Fig. 1 A sample of the gathered Tweets

or ”No” answer to the question ”Does this text imply self-harm tendencies or suicidal intent?” [27] In all, 96 of all tweets were annotated as suicidal ones and 164 as non-suicidal ones. These tweets are subsequently preprocessed and used to train and test classifiers to pinpoint suicide ideation.

Example 1 We show a few samples of posts randomly selected from the Tweet data set in Fig. 1. In this sample, the first two posts are *Non-Suicidal* ones whereas, the four other ones are *Suicidal*. Notwithstanding, this distribution does not reflect the real distribution of the data set.

4.4 Step 2: data pre-processing

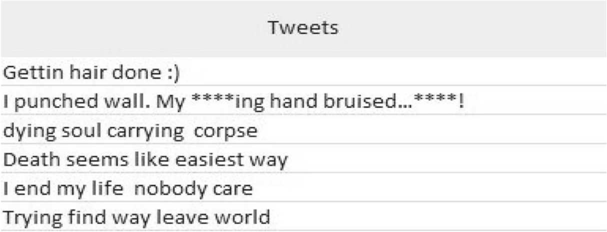
Since online texts, specifically tweets, usually contain a lot of noise and uninformative data, the cleaning step is primordial to improve the result of the classification. Cleaning involves many steps. The most worthy of mention ones are [32]:

- Identification and removal of the retweets shown by ”rt”, the shortened URLs, and the user mentions which have the format of @username.
- Removal of the hashtags ”#”. Indeed, so many of them are concatenated words, which amplifies the vocabulary size.
- Removal of all the stop words, which do not offer any further information on the general orientation of the text.

Example 2 In this step, we apply the cleaning process on the six Twitter posts sketched in the previous step. Figure 2 depicts the cleaned posts. For the comprehensibility of the tweets in the following steps, we present the posts only cleaned from the hashtags and URLs but without removing the stop-words.

4.5 Step 3: feature extraction

Two individuals, even having suicidal ideation, may not express their symptoms in the same way. In fact, texts exhibiting the same level of suicidal risk may have vastly different content. That’s why it is barely impossible to create a specific dictionary that engulfs all terms related to suicide.



Tweets
Gettin hair done :)
I punched wall. My ****ing hand bruised...****!
dying soul carrying corpse
Death seems like easiest way
I end my life nobody care
Trying find way leave world

Fig. 2 Pre-processed Tweets

To address this problem, we have used the Empath tool [11] to transform the tweets into a vector of more specified emotional and topical categories (features).

In fact, Empath is a text analysis tool that allows users to analyze text across 200 built-in, pre-validated set of emotional and topical categories drawn from existing knowledge bases and literature. It also allows the generation and validation of new lexical categories on demand from a user-generated set of seed terms. Like the popular LIWC [24], Empath analyzes text documents and returns scores for various psychological and other types of dimensions based on specific dictionaries. It attempts to capture topical and especially emotional dimensions that it strongly relates suicidal tendencies and other mental health problems to.

However, Empath is considered better than LIWC since it can analyze text through a broad range of a topic while LIWC has only 40 topical and emotional categories and many of which contain fewer than 100 words.

Using Empath, we have transformed the tweet text to a vector of categories (features). We derived these categories from the text as follows:

- **General lexical categories:** These categories represent general lexical domains such as death, health, family, home, religion, work, grief, college, etc.
- **Affective lexical categories:** we relate these categories to domains representing sentimental and emotional aspects such as anger, anxiety, happiness, sadness, grief, love, hate, etc. These categories were incorporated because of the particular emotive nature of the suicidal ideation detection. In fact, emotions such as fear, anger, and aggressiveness are prominent in suicidal communication.

Besides the above categories, we have generated a new category that is more specific to the particular domain of suicidal ideation detection. We generate this category based on a list of seed terms, such as *suicidal; suicide; kill myself; my suicide note; my suicide letter; end my life; never wake up; can’t go on; not worth living; ready to jump; sleep forever; want to die; be dead; better off without me; better off dead; suicide plan; suicide pact; tired of living;*

Moreover, researchers and clinicians have shown that individual at-risk use fewer emojis and more first pronouns such *”I”, ”myself, ”me”, ”mine”, etc.*, that is why we thought that including these features could help us predict the suicidal ideation.

Given the large number of features associated with each tweet, and the potential inclusion of some irrelevant, redundant and noisy ones, the classification model building will be subsequently a challenging and time-consuming task and can, even worse, lead to a poor predictive performance. To solve the problem of high dimensionality, a dimension reduction procedure has to be used.

Example 3 Using Empath, we have transformed the list of posts to a matrix having 6 rows which is the number of posts, and 30 columns which correspond to the number of features generated by Empath. We split these features into five categories, namely Emotions, General, Suicide, personal pronouns, and others. We sketch the obtained matrix in Fig. 4. Each value V_{ij} in the matrix is a numerical representation for a category j in a post i.

4.6 Step 4: feature selection

As mentioned above, we have associated an overwhelming number of features with tweets. Redundant, irrelevant, or partially relevant features can negatively affect the classification model performance.

Like the popular LIWC [24], Empath analyzes text documents and returns scores for various psychological and other types of dimensions based on specific dictionaries. Scores are computed based on the *OddRatio* measure. In fact, to assess how much more likely a word is to occur in a category c, the Empath tool computes the oddRatio of a word occurring in a category c based on the formula 10. These oddRatio scores are, subsequently, aggregated by category to form the final score.

$$oddRatio(w, c) = \frac{P(w|category = c)}{1 - P(w|category = c)} \tag{10}$$

The Empath tool attempts to capture topical and especially emotional dimensions that suicidal tendencies and other mental health problems are strongly related to. However, Empath is considered better than LIWC since it can analyze text through a broad range of topics while LIWC has only 40 topical and emotional categories and many of which contain fewer than 100 words.

To tackle this issue, a dimension reduction process has to be performed. It comprises selecting a subset of relevant features that contain reliable information related to the original dataset while removing noisy and redundant features that would not be of help during the classification process. Hence, feature selection improves the results' accuracy, reduces the running time of the classifier, and increases comprehensibility. The three attribute selection approaches applied in the remainder are [14]: *(i)* Correlation Attribute Eval (CA); *(ii)* GainRatio Attribute evaluation (GR); and *(iii)* InformationGain Attribute evaluation (IG). These approaches, used via the WEKA software system [16], are detailed in the following:

- ***Correlation Attribute Eval method (CA):*** It gauges the worth of an attribute regarding the target class. In fact, it uses Pearson's correlation method to measure the correlation between the attribute and the class [15]. This attribute selection method selects only attributes that have a moderately positive or negative correlation and drops attributes with a low correlation, i.e. value close to zero.
- ***Gain Ratio Attribute Eval (GR):*** This method assesses the worth of an attribute by measuring the gain ratio regarding the target class. We compute it using the

following formula where H represents the entropy [15].

$$GainR(class, Attribute) = \frac{(H(class) - H(Class|Attribute))}{H(Attribute)} \tag{11}$$

This attribute evaluation method is subsequently combined with the ranking method of search and then applied to the dataset.

- ***Info Gain Attribute Eval (IG):*** This approach assesses the significance of an attribute regarding the target class based on the value of Information Gain. It is computed by the following formula, where H represents the entropy [15].

$$InfoGain(class, Attribute) = (H(class) - H(Class|Attribute)) \tag{12}$$

Subsequently, this attribute evaluation method is combined with the ranking of search and then applied to the dataset.

Example 4 To reduce the large number of features associated with each tweet, we have chosen in this example to apply the Correlation Attribute Eval method using the Weka tool. The latter selection method maintained 13 attributes among the 30 ones of the input dataset. The removed attributes are those, which are redundant or constant with all the posts such as health, school, 1st plural pronoun, etc. Thence, as sketched in Fig. 5, we obtain a novel matrix having 6 rows and 14 columns (13 features + 1 target class).

4.7 Step 5: classification

Generally, in such domains, e.g., suicidal ideation detection, terrorism detection, harmful and racist speech detection to name but a few, datasets gathered from the web are usually imbalanced. Analyzing this type of data with classical approaches may head to inaccurate results, since they can easily discover rules predicting the majority class and overlook those predicting minor classes, since they are deemed pointless rules.

For this reason, we will use the proposed interestingness measure with the well-known CBA. We will apply the latter IM as the primary criteria to sort and select rules to improve it and make it beneficial for the classification of imbalanced data sets.

In the following section, we pay close attention to the experimental evaluation and thoroughly discuss the obtained results.

5 Experimental results

The adapted CBA with the new measure will be compared to the original CBA, the adapted CBA with three existing measures (i.e. Laplace ”Lap”, Lift and Conviction ”Conv”) and several well-known classification approaches, which are implemented in WEKA [16] including Random Forest (RF) [6], Ripper (JRip) [7], Part [12] and MultiLayerPerceptron (NN).

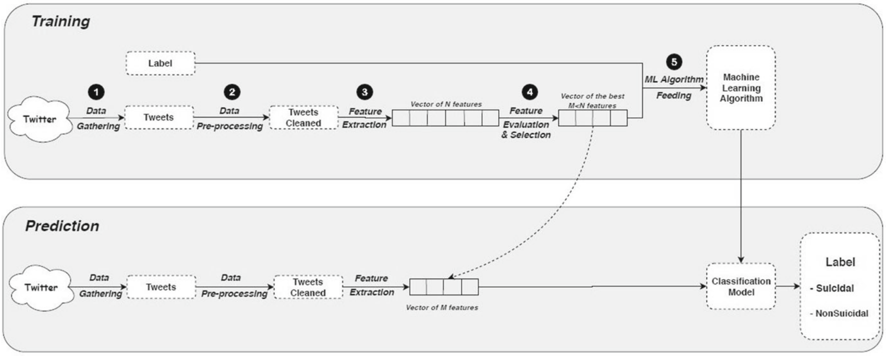
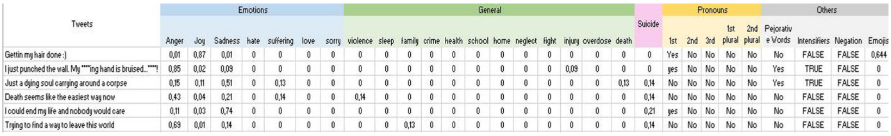
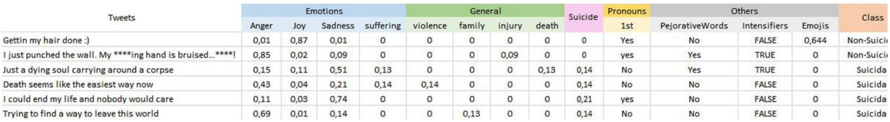


Fig. 3 Suicidal ideation detection process



Tweets	Emotions							General												Suicide	Pronouns					Others			
	Anger	Joy	Sadness	hate	suffering	love	sorry	violence	sleep	family	crime	health	school	home	neglect	fight	injury	overdose	death		1st	2nd	3rd	1st plural	2nd plural	Pejorative Words	Intensifiers	Negation	Emojis
Gettin my hair done :)	0,01	0,87	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Yes	No	No	No	No	No	FALSE	FALSE	0,644
I just punched the wall. My ****ing hand is bruised...****!	0,85	0,02	0,09	0	0	0	0	0	0	0	0	0	0	0	0	0	0,09	0	0	0	yes	No	No	No	No	Yes	TRUE	FALSE	0
Just a dying soul carrying around a corpse	0,15	0,11	0,51	0	0,13	0	0	0	0	0	0	0	0	0	0	0	0	0	0,13	0,14	No	No	No	No	No	Yes	TRUE	FALSE	0
Death seems like the easiest way now	0,43	0,04	0,21	0	0,14	0	0	0,14	0	0	0	0	0	0	0	0	0	0	0	0,14	No	No	No	No	No	No	FALSE	FALSE	0
I could end my life and nobody would care	0,11	0,03	0,74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,21	yes	No	No	No	No	No	FALSE	FALSE	0
Trying to find a way to leave this world	0,69	0,01	0,14	0	0	0	0	0	0	0,13	0	0	0	0	0	0	0	0	0	0,14	No	No	No	No	No	No	FALSE	FALSE	0

Fig. 4 Transformation of the Tweets using the Empath tool



Tweets	Emotions				General				Suicide	Pronouns	Others			Class
	Anger	Joy	Sadness	suffering	violence	family	injury	death		1st	PejorativeWords	Intensifiers	Emojis	
Gettin my hair done :)	0,01	0,87	0,01	0	0	0	0	0	0	Yes	No	FALSE	0,644	Non-Suici
I just punched the wall. My ****ing hand is bruised...****!	0,85	0,02	0,09	0	0	0	0,09	0	0	yes	Yes	TRUE	0	Non-Suici
Just a dying soul carrying around a corpse	0,15	0,11	0,51	0,13	0	0	0	0,13	0,14	No	Yes	TRUE	0	Suicida
Death seems like the easiest way now	0,43	0,04	0,21	0,14	0,14	0	0	0	0,14	No	No	FALSE	0	Suicida
I could end my life and nobody would care	0,11	0,03	0,74	0	0	0	0	0	0,21	yes	No	FALSE	0	Suicida
Trying to find a way to leave this world	0,69	0,01	0,14	0	0	0,13	0	0	0,14	No	No	FALSE	0	Suicida

Fig. 5 Features maintained after the application of the Correlation Attribute Eval

For the feature selection methods, we have used the Correlation Attribute Eval method, the Gain Ratio Attribute Eval method, and last but not least the Info Gain Attribute Eval one.

We will carry experimentation regarding three assessment measures, which are the Global Accuracy, Geometric Mean, and F-measure.

5.1 Global accuracy

Global Accuracy is the most commonly used metric [10]. It is considered as the overall accuracy rate of a classifier. This measure estimates the proportion of instances that are correctly classified. We compute the global accuracy as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

The experimental results are illustrated in Table 2.

The first column stands for the Feature Selection methods, while the second one represents the adapted CBA with the new measure. The third column represents the original CBA. Columns from 4 to 6 represent the adapted CBA with three existing measures. Columns from 7 to 10 represent the classical classifiers used. The two last

Table 2 Global accuracy results

	Proposed measures	Original CBA	Existing measures			Existing algos				Avg Rank	Avg Rate
	ML	Conf	Lap	Lift	Conv	JRip	Part	RF	NN		
Correlation Attribute Eval	0.76	0.65	0.65	0.58	0.78	0.65	0.70	0.72	0.71	1.22	0.68
Gain Ratio Attribute Eval	0.70	0.63	0.63	0.63	0.70	0.65	0.70	0.70	0.67	1.88	0.66
Information gain	0.70	0.63	0.63	0.56	0.72	0.65	0.67	0.68	0.66	2.87	0.65
All attributes	0.70	0.63	0.63	0.34	0.63	0.70	0.59	0.70	0.65	2.77	0.61
Avg Rank	**1.50**	6.25	6.25	7.00	2.00	4.75	4.50	2.00	4.50		
Avg Rate	**0.71**	0.64	0.64	0.53	0.71	0.66	0.67	0.70	0.67		

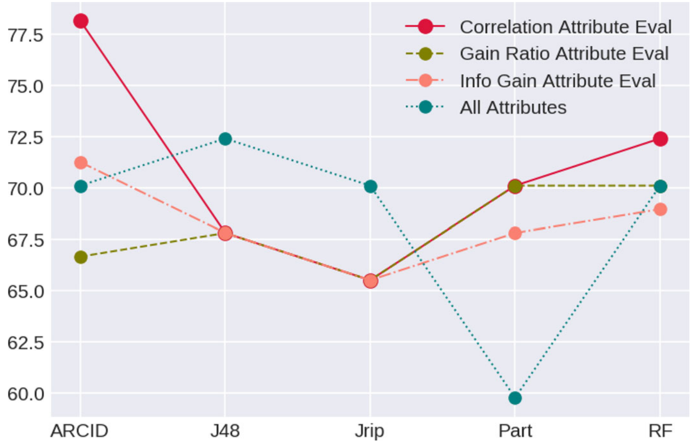


Fig. 6 Classification results based on global accuracy

columns represent, respectively, the average rate and average rank of each Feature Selection method. The first three rows represent the Feature Selection methods, while the fourth row represents the results of the classifier using no feature selection method, i.e., we use all features. The two last rows represent the average rate and the average rank of each classifier. To compute the average rank, we apply a non-parametric Friedman test.

As we can notice from Table 2, the adapted CBA with the novel measure yields better results compared to the adapted CBA with the Confidence, Laplace, and Lift measures and compared to the state of art algorithms Jrip, Part, and NN. However, we notice that the adapted CBA using the Conviction measure yields competitive results to those got with the CBA approach adapted with the novel measure. That the Conviction measure favors rules of major classes while considering those belonging to minor classes as uninteresting could explain this. We also notice that the RF classifier performs competitively to the adapted CBA with the new measure, since these types of algorithms aim to maximize the global prediction and minimize the error rate to which the minor class rarely contributes. If we take a deeper look at Table 2, we notice that the feature selection methods have highly enhanced the global accuracy, especially the Correlation Attribute Eval method, which yields an average rate of 0.68 compared to an average rate of 0.61 using all the attributes. The Correlation Attribute Eval method also enhanced all the adapted versions of CBA and three among the four baseline algorithms. By and large, we may conclude that even the proposed measure is dedicated to handle the imbalance aspect of the dataset, and to more emphasis on correctly predicting the suicidal tweets (minor class), they, nevertheless, offer a good global classification accuracy.

Moreover, if we take a deeper look at Fig. 6, we may notice that using the Correlation Attribute Eval Method for Attribute Selection has sharply enhanced the global accuracy for the ARCID classifier from 70.11 using all the features to 78.16 using the subset of features extracted based on the Correlation Attribute Eval method.

Based on these results, we may conclude that even though ARCID is dedicated to handle the imbalance aspect of the data set and to more emphasize on correctly predicting

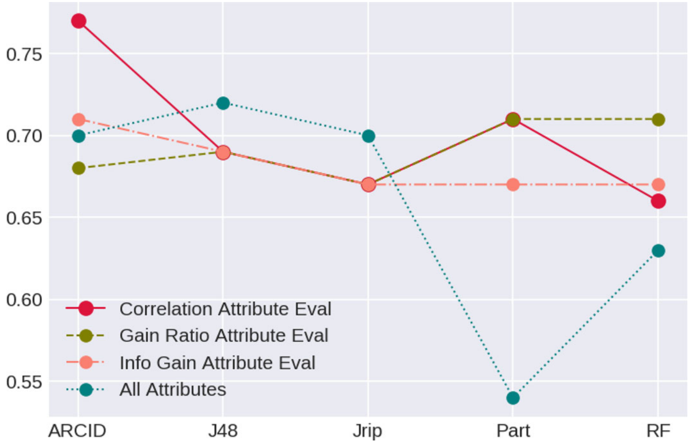


Fig. 7 Classification results based on geometric mean

the suicidal tweets (minor class), it enhances the prediction of everyday tweets (major class) and subsequently offers a very good global classification accuracy.

5.2 Geometric mean

As we mentioned above, the adapted versions of CBA using the novel measure and combined with the Correlation Attribute Method for Attribute Selection provide statistically good results compared to the other adapted versions of CBA and the state-of-art algorithms in terms of global accuracy.

However, the accuracy could not only be used to evaluate the performance of a classifier, especially with imbalanced datasets. In fact, based on Global Accuracy, we consider a classifier as good, even if it misclassifies all the instances of a minor class.

To overcome this problem, we have proposed to use an additional evaluation measure which is the Geometric Mean, suggested by Kubat et al. [IM34], as the product of the prediction accuracies of both classes. Thus, Gmean assesses the balanced performance of a classifier between the minority and majority classes. We can only achieve a high Gmean value with high prediction accuracies in both classes. The Gmean measure is defined as follows:

$$GeometricMean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{14}$$

According to the results depicted in Table 3 and Fig. 7, we clearly notice that the CBA approach, using the novel measure, outperforms by several ranks the CBA approach using the existing measures and the existing algorithms. In fact, the CBA approach, using the novel measure, yields the best average rank (1.5) and the best average rate (0.7). Whereas for CBA using the existing measures and for the baseline algorithms, the average ranks range from 2.00 to 8.25 and the average rates range from 0.06 to 0.68.

Table 3 Geometric mean results

	Proposed measures	Original CBA	Existing measures			Existing algos				Avg Rank	Avg Rate
	ML	Conf	Lap	Lift	Conv	JRip	Part	RF	NN		
Correlation Attribute Eval	0.77	0.25	0.25	0.60	0.72	0.67	0.71	0.66	0.67	1.44	0.89
Gain Ratio Attribute Eval	0.70	0.00	0.00	0.65	0.70	0.67	0.71	0.71	0.68	1.44	0.53
Information gain	0.68	0.00	0.00	0.57	0.60	0.67	0.67	0.67	0.63	2.66	0.49
All attributes	0.70	0.00	0.00	0.55	0.57	0.70	0.54	0.63	0.61	3.00	0.47
Avg Rank	**1.50**	8.25	8.00	6.75	4.00	9.5	2.00	4.00	4.50		
Avg Rate	0.70	0.06	0.06	0.59	0.65	0.60	0.66	0.67	0.65		

Table 4 F-measure results

	Proposed measures	Original CBA	Existing measures			Existing algos				Avg Rank	Avg Rate
	ML	Conf	Lap	Lift	Conv	JRip	Part	RF	NN		
Correlation Attribute Eval	0.71	0.11	0.11	0.60	0.66	0.62	0.65	0.58	0.60	1.66	0.52
Gain Ratio Attribute Eval	0.68	0.00	0.00	0.66	0.68	0.62	0.65	0.65	0.62	1.44	0.50
Information gain	0.60	0.00	0.00	0.59	0.52	0.62	0.60	0.59	0.55	2.66	0.45
All attributes	0.65	0.00	0.00	0.54	0.50	0.64	0.42	0.55	0.52	3.11	0.43
Avg Rank	**1.25**	8.00	8.00	4.00	4.00	3.25	4.00	4.50	5.50		
Avg Rate	**0.66**	0.03	0.03	0.60	0.59	0.63	0.58	0.59	0.57		

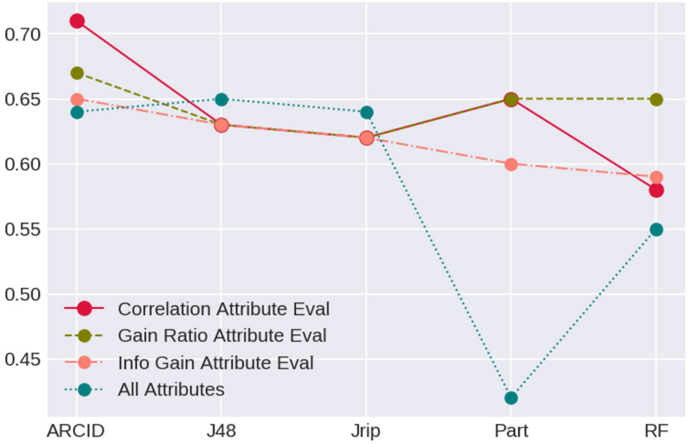


Fig. 8 Classification results based on F-measure

Based on these results, we may conclude that the novel interestingness measure is statistically backed up to be suitable for the selection of rare but high-quality rules while maintaining a good global accuracy. Regarding the attribute selection method, we notice the Correlation Attribute Eval method is the best Attribute Eval method whenever compared to all its competitors.

To better assert the efficiency of the novel measures in case of the classification of imbalanced datasets, we will use another additional assessment measure, which is the F-Measure.

5.3 F-measure

The F-measure combines Precision and Recall into a single measure. It also reflects the soundness of a classifier in the presence of rare classes. Precision is a measure of exactness. It calculates the percentage of instances that are actually correctly labeled among instances labeled as positive. The recall is a measure of completeness that assesses how many examples of the positive class were correctly labeled. We define these measures as follows:

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{17}$$

Table 4 presents F-measure results obtained from the experiments. As plainly underscored by Fig. 8, we can observe that the CBA using the proposed interestingness provides statistically better results. In fact, similar to the performance's conclusions got with the Gmean assessment measure, the difference of performances between the

CBA using the novel interestingness measure and CBA using the existing interestingness measure and the existing algorithms is undoubtedly significant. Indeed, the average rank (average rate) of CBA using the proposed measure is 0.66 (resp. 1.25), whereas, the average rank (resp. average rate) using is ranging between 4.00 to 8.00 (resp. 0.03 to 0.60) for the CBA using the existing measures and 4.00 to 5.50 (resp. 0.57 to 0.59) for the existing algorithms. Moreover, the Correlation Attribute Eval yields the best average rate among all the rest.

Based on all these findings, we may assert that the new measure successfully focuses on increasing the accuracy of the minor class while trading off the accuracy of the major class.

5.4 Discussion

Summing up the results, we can conclude that the introduced measure separately added to the CBA algorithm has improved it to be a suitable algorithm for predicting suicidal ideation through tweets. In fact, they proved their efficiency for the extraction of significant knowledge from minor classes (suicidal tweets), which is statistically proved using the Geometric Mean and the F-measure, while perfectly maintaining and even enhancing the predictive accuracy of the whole classifier which is statistically proved by the Global Accuracy.

Furthermore, it is clearly noticed that the Correlation Attribute Eval method is the most adapted and suitable feature selection method for the adapted version of CBA and most of the state-of-art algorithms.

6 Conclusion

This paper proposed a new interestingness measure in order to handle the problem of association rule selection for prediction when datasets are imbalanced. In lieu of the existing measures, such as Confidence and Laplace, we use the proposed measure as the primary criterion to filter out rules that will be of use by the well-known associative classifier CBA. This measure aims to keep highly interesting rules from both types of rules simultaneously in order to offer a high recall on the minority class while maintaining a high precision for the majority class.

Experiments in the domain of suicidal ideation detection in social networks and regarding three assessment measures proved that the newly proposed measures offered a better predictive power compared to the existing ones and compared to other several well-known classification approaches.

Even though in this paper, the classification is carried out using one association rule-based classifier which is CBA [1], the results can be straightforwardly generalized to other associative classifiers, such as ARC-BC [4], ACN [20].

The obtained results in this paper open many perspectives. In the following, we present some promising future research paths from which we cite:

- Handling the problem of multi-label: In fact, one single label classifier may be insufficient in some cases since it only associates the obvious class to the rule and omits others. This could mislead in several cases, namely in medical diagnosis. For instance, taking the example of a patient who suffers from cough and food poisoning simultaneously, a multi-label classifier is required. Therefore, the goal of the classifier turns to recognize the categories of diseases rather than only to recognize instances as diseases or not.

- Learning from imbalanced streams: Another fruitful perspective consists in learning from imbalanced streams. In fact, because of the vast amount of data collected every second, adding and modifying the training set presents a new challenge to the classifier when an imbalanced distribution is expected. In fact, in case of a repetitive change, the relation between classes is no longer permanent and the imbalance ratio is changing with the stream of progress. In this respect, close attention should be paid to the recent COVID 19 pandemic lockdown, During that period, communities have faced mental health challenges related to COVID-19-associated morbidity, mortality, as well as physical distancing and stay-at-home orders.
- A white-box model is used to generate explanations. We need to pay close attention to the quality of the provided explanations to explain suicidal ideation predictions, to wit self-explaining models, i.e., accurate predictions coming along with a narrative explanation in natural language easy to understand by different intervenes. Different models of explainability, e.g., SHAP, LIME, Contractive, Counterfactual, could be explored and assessed in terms of proximity to the user's mental model.

Acknowledgements Sadok Ben Yahia is supported by the Astra funding program Grant 2014-2020.4.01.16-032

References

1. Abdellatif S, Ben Hassine MA, Ben Yahia S, Bouzeghoub A (2018) ARCID: a new approach to deal with imbalanced datasets classification. In: SOFSEM 2018: theory and practice of computer science - 44th international conference on current trends in theory and practice of computer science, Krems, Austria, January 29–February 2, 2018, Proceedings, pp 569–580
2. Ali K, Manganaris S, Srikant R (1997) Partial classification using association rules. In: Proceedings of the third international conference on knowledge discovery and data mining (KDD-97), Newport Beach, California, USA, August 14–17, 1997, pp 115–118
3. American Psychiatric Association (2003) Practice guideline for the assessment and treatment of patients with suicidal behaviors. Am J Psychiatry 160:1–60
4. Antonie ML, Zaiane OR (2002) Text document categorization by term association. In: Proceedings of 2002 IEEE international conference on data mining. IEEE, pp 19–26
5. Ben Yahia S, Gasmi G, Nguifo EM (2009) A new generic basis of factual and implicative association rules. Intell Data Anal 13:633–656. https://doi.org/10.3233/IDA-2009-0384
6. Breiman L (2001) Random forests. Mach Learn 45:5–32
7. Cohen WW (1995) Fast effective rule induction. In: Proceedings of the twelfth international conference on machine learning, pp 115–123
8. Coppersmith G, Leary R, Whyne E, Wood T (2015) Quantifying suicidal ideation via language usage on social media. In: Joint statistics meetings proceedings, statistical computing section, JSM
9. DeJong TM, Overholser JC, Stockmeier CA (2010) Apples to oranges? A direct comparison between suicide attempters and suicide completers. J Affect Disord 124:90–97
10. Elfeky MG, Verykios VS, Elmagarmid AK (2002) Tailor: a record linkage toolbox. In: Proceedings 18th international conference on data engineering. IEEE, pp 17–28

11. Fast E, Chen B, Bernstein MS (2016) Empath: understanding topic signals in large-scale text. In: Proceedings of the 2016 CHI conference on human factors in computing systems, ACM, pp 4647–4657
12. Frank E, Witten IH (1998) Generating accurate rule sets without global optimization
13. Geng L, Hamilton HJ (2006) Interestingness measures for data mining: a survey. ACM Comput Surv (CSUR) 38:9
14. Gnanambal S, Thangaraj M, Meenatchi V, Gayathri V (2018a) Classification algorithms with attribute selection: an evaluation study using WEKA. Int J Adv Netw Appl 9:3640–3644
15. Gnanambal S, Thangaraj M, Meenatchi V, Gayathri V (2018b) Classification algorithms with attribute selection: an evaluation study using WEKA. Int J Adv Netw Appl 9:3640–3644
16. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 11:10–18
17. He H, Bai Y, Garcia EA, Li S (2008) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE World Congress on Computational Intelligence, pp 1322–1328
18. Hu B, Dong W (2014) A study on cost behaviors of binary classification measures in class-imbalanced problems. arXiv:1403.7100
19. Ji S, Pan S, Li X, Cambria E, Long G, Huang Z (2021) Suicidal ideation detection: a review of machine learning methods and applications. IEEE Trans Comput Soc Syst 8:214–226. https://doi.org/10.1109/tcss.2020.3021467
20. Kundu G, Islam MM, Munir S, Bari MF (2008) ACN: an associative classifier with negative rules. In: 2008 11th IEEE international conference on computational science and engineering. IEEE, pp 369–375
21. Lenca P, Vaillant B, Meyer P, Lallich S (2007) Association rule interestingness measures: experimental and theoretical studies. In: Quality Measures in data mining. Springer, pp 51–76
22. López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141. https://doi.org/10.1016/j.ins.2013.07.007
23. Patel H, Rajput DS, Reddy GT, Iwendi C, Bashir AK, Jo O (2020) A review on classification of imbalanced data for wireless sensor networks. Int J Distrib Sens Netw. https://doi.org/10.1177/1550147720916404
24. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates 71
25. Piatetsky-Shapiro G (1991) Discovery, analysis, and presentation of strong rules. Knowledge discovery in databases, 229–238
26. Rudd MD, Berman AL, Joiner TE Jr, Nock MK, Silverman MM, Mandrusiak M, Van Orden K, Witte T (2006) Warning signs for suicide: theory, research, and clinical applications. Suicide Life-Threaten Behav 36:255–262
27. Sawhney R, Manchanda P, Singh R, Aggarwal S (2018) A computational approach to feature extraction for identification of suicidal ideation in tweets. In: Proceedings of ACL 2018, student research workshop, pp 91–98
28. Sueki H (2015) The association of suicide-related twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan. J Affect Disord 170:155–160
29. Tan PN, Kumar V, Srivastava J (2004) Selecting the right objective measure for association analysis. Inf Syst 29:293–313
30. Thabtah FA, Hammoud S, Kamalov F, Gonsalves A (2020) Data imbalance in classification: experimental evaluation. Inf Sci 513:429–441. https://doi.org/10.1016/j.ins.2019.11.004
31. Vioulès MJ, Moulahi B, Azé J, Bringay S (2018) Detection of suicide-related posts in twitter data streams. IBM J Res Dev 62:1–7
32. Xiang G, Fan B, Wang L, Hong J, Rose C (2012) Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In: Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, pp 1980–1984