ABSTRACT

This research paper was entitled “Pattern Recognition of Suicidal Ideation”. The sole objective of this study is to explore and classify the patterns of suicidal ideation in Twitter here in Philippines. The machine learning algorithm used was K-Nearest Neighbor (KNN) ,Support Vector Machine (SVM) and Naive Bayes. The models were compared by their accuracy, class recall, precision and F-Measure as a basis for choosing the best model.

The data were tweets and collected through RapidMiner 8.0 using Search Twitter operator. Data gathered were Filipino and majority was Tagalog. Data cleaning and data preprocessing were employed after the collection of data. Classification Rule Mining was used in order to understand and classify the suicidal ideation.

The most frequent itemsets occurred are *ayoko* and *mabuhay* with a support count of 0.09 and 0.043 respectively.The resulting top trend in Filipino suicidal ideation after generating the data from repositories were (*ayoko→ mabuhay*)with a confidence of0.381 and (*sawa→ mabuhay*) with a confidence of 0.273.

Results showed that Support Vector Machine (SVM) is the best model that gave the best results in classifying the possible suicidal ideation since it has the highest sensitivity.

Keywords*: suicidal ideation, classification, best model.*

INTRODUCTION

Background of the Study

Suicidal ideation, also known as suicidal thoughts, is the thoughts of people wanting to harm themselves. Suicidal ideation is a feeling that may occur when people are not able to cope up huge situations; it could be the death of somebody they love, chronic illness, a break up, or financial problems. There are two forms of suicidal ideation, the active and passive suicidal ideation. Passive suicidal ideation involves a desire to die, but not having a specific plan to carry out death. Active suicidal ideation involves an existing wish to die including the plan on how to do it. The main thought of suicidal ideation is the act of suicide.

In the data of WHO (World Health Organization) 2015, there are close to 800,000 people dying due to suicide every year. Also, it was stated that suicide is the second leading cause of death among 15-29 years old next to road injury. Approximately, one million people commit suicide each year worldwide, that is about one death every 40 seconds or 3,000 per day.

Nowadays, social media has been widely used by many people in sharing information about their lives, interacting with one another, updating photos and up-to-minute thoughts. In 2017 there are 2.46 billion social media users, it is estimated that there will be 2.77 billion users in 2019. An article written by Conor Gaffey of Newsweek says that “two hours of social media a day linked suicidal thoughts in teens”. Also, according to the Digital Global Overview by We Are Social and Hootsuite, Filipinos lead the world in social media use with an average of 4 hours and 17 minutes a day. Fiftyeight percent of the Philippine population is active social media users on a monthly basis even though Philippines is one of the countries that have the slowest internet connection.

Facebook as the top leading social networking site with over one billion users has been investing in artificial intelligence fields like machine learning and deep neural nets to predict and prevent suicide. The social network already had a system that allows users to report posts that suggest a user is at risk of self-harm. Using those reports, Facebook trained an algorithm to recognize similar posts, which they are testing now in the US. Once the algorithm flags a post, Facebook will make the option to report the post for "suicide or self-injury" more prominent on display.

A study on suicidality in Twitter (one of the most popular social networking site) examined the level of concern for a suicide-related post; the bases were solely the content of the post. It was judged by human coders and then replicated by machine learning. The study shows that 14 percent of suicide-related tweets were 'strongly concerning'. However it was unclear for the researchers if those tweets are genuine suicide or not (O'dea et al., 2015).

Suicidal ideation risk factors in the general population may be related with the presence of family history for psychiatric illness, depressive mood, high anger and short or long sleep duration (Seung, Lee et al., 2013). Moreover, as social media became a place where everyone expresses their feelings and thoughts, users should be aware of other users that have suicidal thoughts.

With the alarming usage of Filipino social media users and from the read and gathered data, the researcher thought of conducting a study that will identify the pattern of suicidal ideation tweets in the Philippines. It will raise awareness among Filipino users and will also probably classify suicidal risk among co-users.

Statement of the Problem

The main focus of the study is to recognize and classify the incidence of suicidal ideation in Twitter which is one of the most known networking sites of this millennium.

This study shall answer the following questions:

1. What is the nature or characteristics of the data gathered?
2. What are the most frequent terms occurring in the generated database and its pattern?
3. What model gave the best results in determining the possible suicidal ideation in twitter?

Significance of the Study

First, this research will be an edge to the researcher for she will gain more knowledge and wisdom regarding her study – pattern recognition of suicidal ideation. It will also widen her knowledge in exploring big datasets.

At the same time, this study will benefit the Pangasinan State University for this study could be added to the research papers in the field of data mining.

Furthermore, it will let social media users be aware of suicidal ideation. It will also help users in recognizing and acting upon any warning signs that a family member or a friend was considering of harming themselves.

Lastly, the provided general knowledge and information on this research regarding pattern recognition can be used as a guide by future researchers' and mental health agencies.

Scope and Limitations

This research is entitled "Pattern Recognition among Suicidal Ideation". The data on this research only focused on tweet data regarding Filipino suicidal ideation which was gathered from Twitter. These tweets can be categorized as yes (with suicidal ideation) or no (without suicidal ideation).

Classification Analysis and Association Rules was used. The task of the researcher was to discover the best model that would accurately categorize Filipino tweets.

Definition of Terms

The following terms were described conceptually and/or operationally:

Suicidal Ideation – is a thought to kill oneself.

Suicide- is the act of intentionally killing himself.

Social Media- are computer-mediated technologies that leta user create and share information via virtual communities.

Users- are someone who uses social media.

Classify- is using classification rules and algorithm to further analyze Filipino tweet data.

Pattern Recognition- uses training and test set to identify the accuracy of a classification model.

REVIEW OF RELATED LITERATURE AND STUDIES

Related Literature

Suicidal Ideation is an overwhelming desire to carry out death; it usually occurs during depressive episodes. Usually, thoughts are not followed by an actual suicide

(Nugentet al.2013). It was very depressing that people having suicide thoughts cannot find their way back to positive thinking.

According to Valley Behavioral Health System, suicidal ideation has 3 causes that were clearly identified.

Genetic. A family history of suicidal behavior is associated with suicidal behavior of family members. (Brent DA, et al. 2005)

Physical. Abnormal low level of neurotransmitters such as dopamine and serotonin to the body changes the structure and function of the brain. It is highly a risk for mental illness including suicidal behavior.

Environmental. To those people that are inbad situations such as being insulted in school, or be attacked is highly at risked of suicidal ideation. (Sharma et al., 2015)

Related Studies

For the past years, there are lots of researchers that focus on the pattern recognition on suicidal ideation using different models, methods and media. The purpose of this section is to review some studies that are related to this study. The following studies are cited below:

Cheng et al. (2017), assessed the suicide risk and emotional distress in Chinese social media. A web-based survey was conducted to measure the risk factors of the user including their suicidal probability. Support vector machine and logistic regression are the classifiers that the researchers used.

O'Dea et al. (2016) and Birjali, et al. (2016) both used machine learning algorithms to detect and predict suicidal ideation posts in Twitter.

Mandge (2013) employed data mining tools in WEKA to gather data. Internal and external data of students and parents were collected. The success of the project turned data mining into an available technology in identifying students at high risk of suicide.

Seunget al. (2013) used survey method in Korea to collect their data. They conducted their decision tree analysis from Answer Tree 3.0 program. The results of the study showed that depressed group of female students with a score of delinquency had the highest rate of suicide attempts.

Price et al. (2004), this study used data mining techniques such as genetic algorithms (GA), artificial neural networks (ANN), and tree-based regression (TBR). They proposed to use the three data-mining techniques to: a) select the most predictive measures of suicide and suicidal behavior using the GA;b) examine the patternsof interaction among the most predictive measures chosen by GA; c) maximize the predictive power of theselected measures using ANN and compare the results with those by other methods; and d) to examine thestructure of associations among the most predictive measures using the information stored in the trained ANNs.

Reece et al. (NYI), predicted photos on Instagram (one of the top social networking site) with suicidal features. Face detection algorithm was used to identify photos and they used Bayesian logistic regression to determine t he strength of individual predictors.

Burnap et al. (NYI) used the four derived features sets with the most popular classifiers from the special issue on classification of suicidal topics. These were Support Vector Machine (SVM), Rule-Based (they used Decision Trees), and Naive Bayes (NB).

De Choudhury et al. (NYI) classified their data gathered in twitter using Support Vector Machine (SVM). Results showed that social media contains useful signals for characterizing the onset of depression in individuals.

Roberts et al. (NYI) annotated tweets in twitter at the tweet level given the seven emotions: anger, disgust, fear, joy, love, sadness and surprise using sentiment analysis and corpus analysis.

METHODOLOGY

Data Collection

The first process that was accomplished in this paper was to gather data. The data was gathered using RapidMiner 8.0 a software platform for data science teams that unites data prep, machine learning, and predictive model deployment. An operator in RapidMiner 8.0 “Search Twitter” will produce tweets with a maximum number of 10,000 in a single query.

The researcher collected and connected the Twitter API and access token to allow the researchers Twitter Account. Changing the value of the parameters “query” and “limit” took many runs. Query is the term that should be searched and limit determines the number of tweets. Filipino language was used in gathering tweets since it is the focus of the study. Only the tweets that are most popular or recent were gathered.

Table 1

Sample Tweets from Actual Data Set with Number of Language Used

Number of Language	**Tweets**
One (Tagalog)	Bwisit na buhay walang kwenta sana hindi na ako pinanganak!
Two (Tagalog- English)	Fuck this life!! Please kunin niyo na ako. Ayoko na mabuhay
Three (Two Filipino Dialects-English)	Everyone calls me crazy kapag sinasabi kong gusto ko namamatay. Haan ak nga bagtit!!!

Data Cleaning

Data cleaning was employed after data collection. Data gathered was pasted in Microsoft Excel. Tweets that appeared more than once were removed especially the huge amount of retweets. Words and symbols that are irrelevant were also deleted. After data cleaning, the exact total number of tweets kept was 18,258.

Data Annotation

After data collection and cleaning, the data were annotated manually. The data was labeled by “Yes” (with suicidal ideation) or “No” (without suicidal ideation). There are three annotators including the researcher. Since the researcher cannot afford to hire professionals (psychologist) to annotate her data, she chose to have an annotator that experiences issues in life specifically in school, friends and family. The tweets were validated and were stored in MS Excel.

Tweets that appeared to be not actually a suicidal thought obtained were 16,919.To avoid bias, these tweets were randomized using MS Excel and the first 1,339 was used for the whole process. A total of 2,678 tweet data was used in this data.

Table 2
Example of Annotated Tweets

Yes	No
Yung katawan ko buhay na buhay pero yung pagiisip ko patay na.	Ako puyat, ako tulog kanina ,ako kakagising lang, ako di pa kain sarap buhay.
Kapagod maging walang kwenta na tao. Magpakamatay na kaya ako?	Ang haba nung message ko nuxx. I hope you'll appreciate it heart heart
Tama na!! ayoko na etong buhay na eto	Hindi lang pala sa basketball magaling, pati sa pagiging walang kwenta, aba napakagaling!
Ayoko na mabuhay.	Hina ng wifi ey....Grabe na ito!
Siguro nga talaga ayaw kong mabuhay kasi i'm an overdue baby, meaning 10months akong asa tiyan ng nanay ko at sabing doctor, ayaw ko daw lumabas kaya ganun kaya nag cs nalang si mommy. Sana piñata nyo nalang ako. Sana hinayaan nyo nalang ako mamatay sa tiyan ng mommy ko.	Grabe ngayon nalang ulit ako nakatawang sobrang lakas dahil sa nanay ko. Namiss ko lang naman yung tumawa na abot sa kapitbahay dahil sa lakas hakhak

Data Preprocessing

This is one of the main process done by the researchers after cleaning and annotating the reviews gathered, it is the time for data preparation.

*Process Document from Data*operator was used. This operator has set of parameters. In the *vector creation*, the researcher used TF-IDF (Term Frequency-Inverse Document Frequency). This schema is a numerical statistic which shows how important a word is to a document in collection.

Data preprocessing includes case transformation, tokenization, filtering stop words, and stemming.

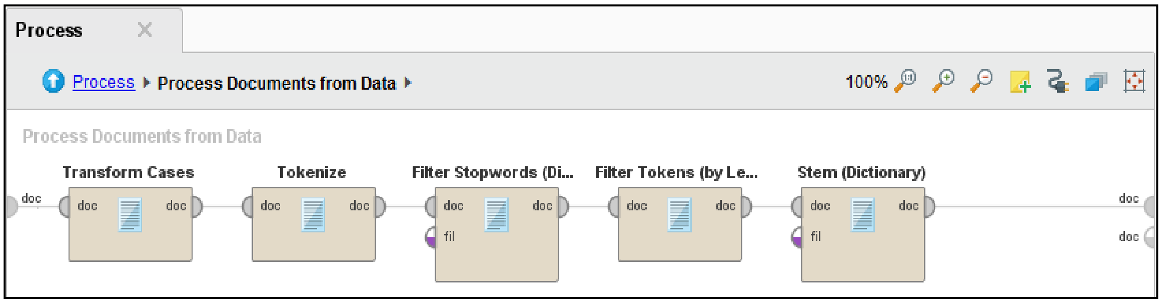


Figure 1. ***Inside Preprocessing Document Operator***

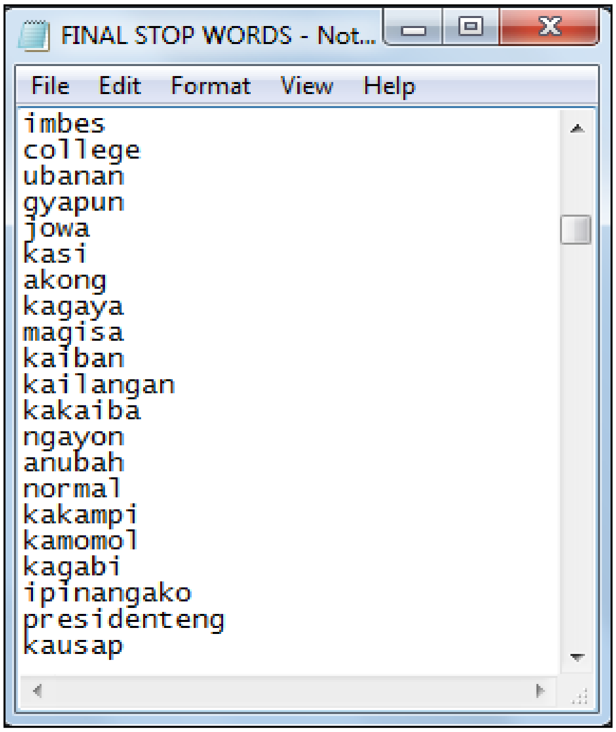
Under the preprocessing, the first step was to transform cases. This operator basically transforms the cases of letters (*eg. lower case or upper case*). The researcher chose to transform all letters to lower case for the purpose of convenience.

After case transformation, tokenization was employed. This operator splits the text into series of tokens. In the parameter section, the researcher chose non-letters.

After tokenizing the data, the researcher used Filter Stopwords (Dictionary) operator to remove all the tokens equal to the text file. The researcher created a list that is not necessary. Only some suicidal thoughts and pronoun were not included on the list. The researcher typed it in Notepad and saved it in text format.

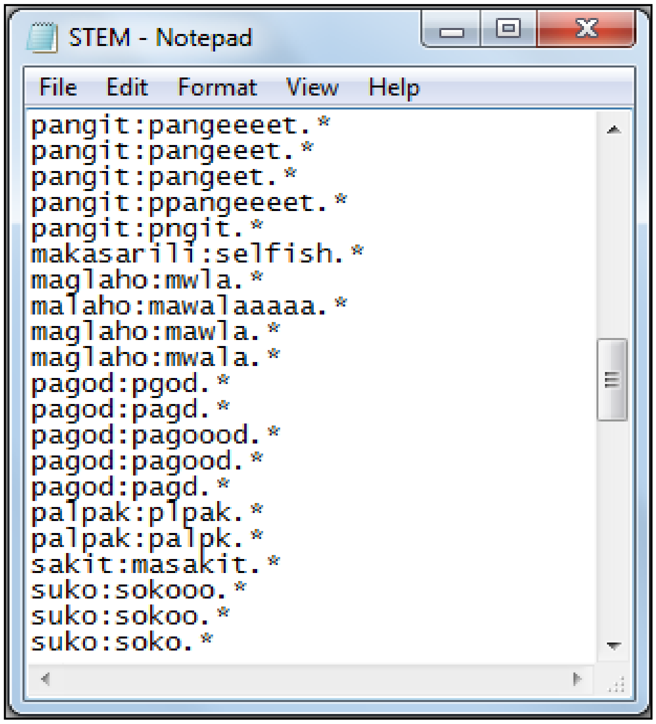
The last step that was employed was the stemming. This process needs the use of Stem (Dictionary) operator which allows the reduction of *Pattern Recognition of Suicidal Ideation*of the terms to a base form using an external file with replacement rules.

Figure 2.*Saved Text File of Stopwords Filipino Dictionary*



```
imbes
college
ubanan
gyapun
jowa
kasi
akong
kagaya
magisa
kaiban
kailangan
kakaiba
ngayon
anubah
normal
kakampi
kamomol
kagabi
ipinangako
presidenteng
kausap
```

Figure 3.*Saved Text File of Stem Filipino Dictionary*



```
pangit:pangeeeet.*
pangit:pangeeet.*
pangit:pangeet.*
pangit:ppangeeeet.*
pangit:pngit.*
makasarili:selfish.*
maglaho:mwla.*
malaho:mawalaaaaa.*
maglaho:mawla.*
maglaho:mwala.*
pagod:pgod.*
pagod:pagd.*
pagod:pagoood.*
pagod:pagood.*
pagod:pagd.*
palpak:plpak.*
palpak:palpk.*
sakit:masakit.*
suko:sokooo.*
suko:sokoo.*
suko:soko.*
```

Pattern Recognition Process

The method that the researcher applied in determining the pattern recognition of suicidal ideation in Twitter was the Association Rule Mining. In Association Rule Mining, all itemsets must meet the set value for minimum threshold for support and confidence to arrive at strong relationship between or among items. The formula for computing the support and confidence are given below:

$$\textbf{SUPPORT} = \frac{\text{occurrences of } [a,b]}{\text{total number of transactions}}$$

$$\textbf{CONFIDENCE } [a,b] = \frac{\text{occurrences of } [a,b]}{\text{total number of } [a]}$$

Frequent item sets are quantified by ***support*** which is the ratio of the number of instances where [A, B] appeared together in a single transaction to the total number of transactions while the ***confidence*** is defined as the probability of finding [item A, item B] together.

The first step includes conducting tweet analysis using FP – Growth in terms of frequent patterns of item sets. The second step includes the analyzation of strong relationship between pair of words using the Create Association Rules. To arrive at a strong relationship between items, the researcher set the value for minimum support threshold and confidence into 0.01.

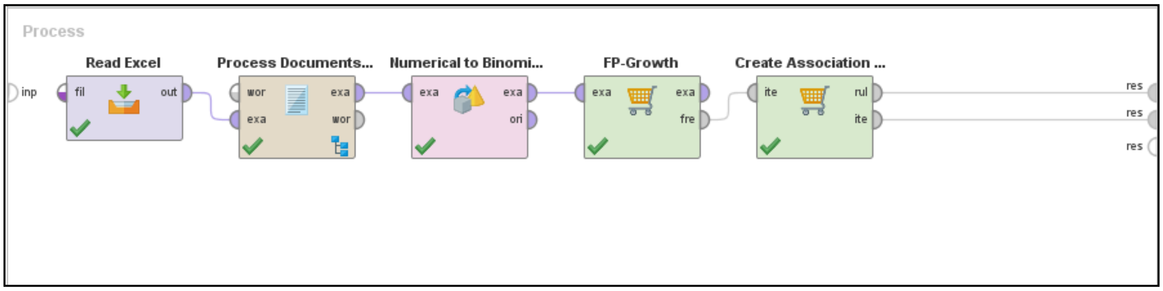


Figure 4. ***Overall Pattern Recognition Process***

Pattern Recognition Using FP – Growth

The researcher applied FP – Growth to determine the frequent pattern or patterns. By the description provided by RapidMiner, this operator efficiently calculates all frequent itemsets from the given ExampleSet using the FP-tree data structure. It is compulsory that all attributes of the input ExampleSet should be binominal. Basically, frequent itemsets are groups of items that often appear together in the data.

FP – Growth has two basic working modes: first, finding at least the specified number of itemsets with highest support without taking the 'minimum support' into account and the second one, finding all itemsets with a support larger than the specified minimum support. This approach uses FP-Tree algorithm which encodes the data set into a tree and then extracts the frequent itemsets from this tree.

Pattern Recognition by Creating Association Rules

The researcher also applied Create Association Rules that is based on Market basket Analysis. This operator generates a set of association rules from the given set of frequent itemsets. This method is generally used in supermarket and other shopping businesses because it helps to identify the trending items which are purchased together. This helps the owners of these kinds of businesses to optimize these items for more profit. When this method is applied to text mining specifically in this research study, the suicidal words or terms will become the items and the text field will become the transactions.

By using this operator, the researcher can clearly see the words that have strong relationship with each other and are commonly used together in tweets with suicidal ideation.

Determining the Best Model

After data annotation, the data were transported to RapidMiner using Read Excel, the data has been divided into two sets such as training and testing. Then undergo with the other stages which are data pre-processing where tokenization, stemming transforming and filtering stop words were done by the RapidMiner. The researchers selected the three most used algorithm in classification which is the K-Nearest Neighbor, Support Vector Machine, and Naive Bayes.

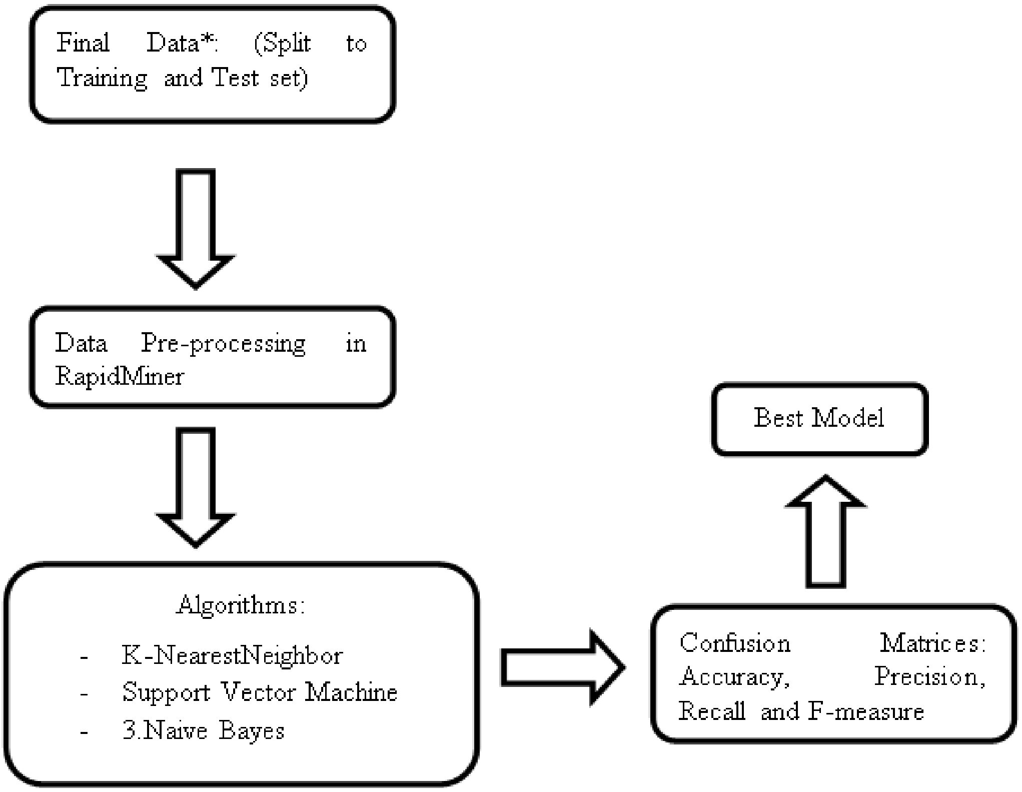
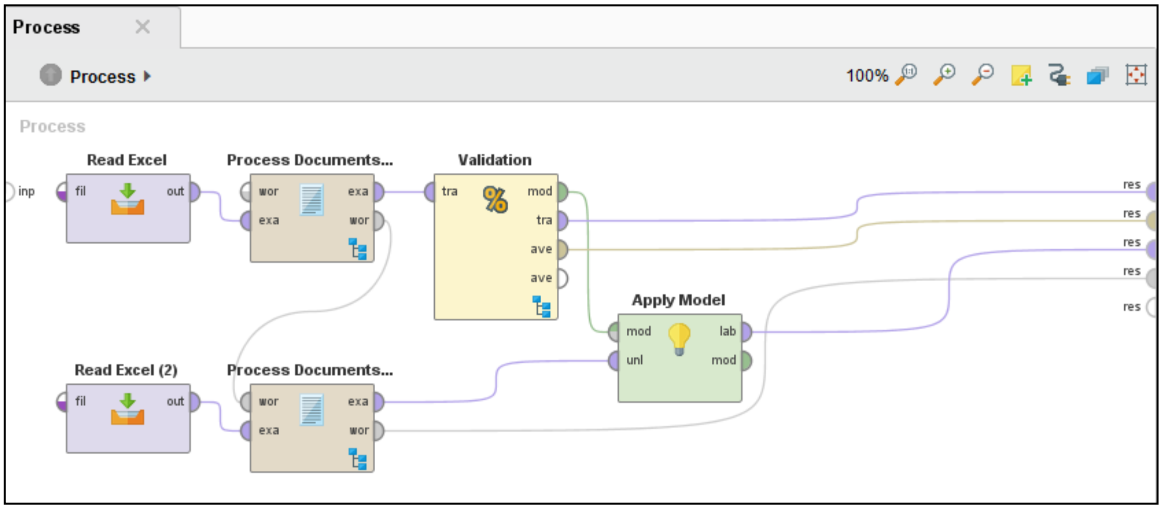

Final Data*: (Split to Training and Test set)
Data Pre-processing in RapidMiner
Algorithms:
- K-NearestNeighbor
- Support Vector Machine
- 3.Naive Bayes
Confusion Matrices: Accuracy, Precision, Recall and F-measure
Best Model

Figure 5. ***Process for Determining the Best Model***

Figure 5 shows the flow of determining the best model. The models will individually produce its confusion matrixes in the training set. The confusion matrix for test set was manually computed by the researcher.

Figure 6. ***Overall Process in Determining the Best Model***



Validation Operator

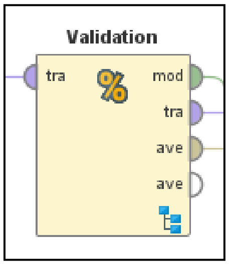
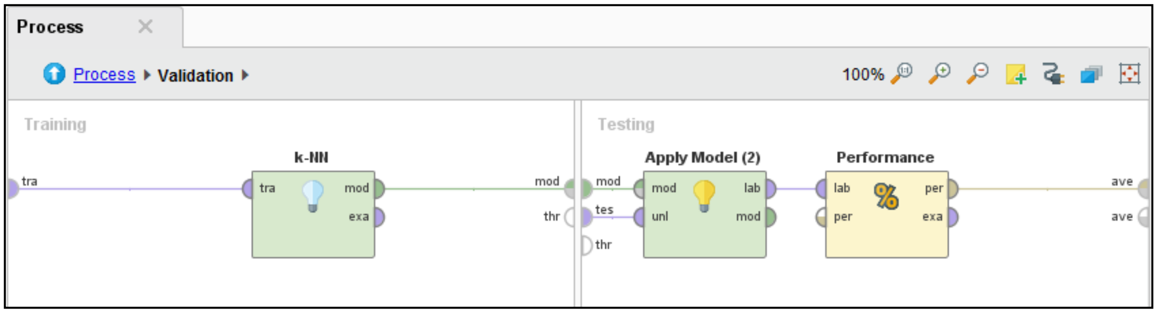


Figure 7. ***Split Validation Operator***

The validation operator used in this study was Split Validation, this operator randomly splits up the data into a training set and test set and evaluates the model. It is used to know the accuracy of the model in practice. The split ratio parameter was set into 0.7 which means the data were split as 70% data for training and the rest is for the test set. Split Validation has a sub process as shown in Figure 8 using K-Nearest Neighbor model.

Figure 8. ***Inside Split Validation Operator***



Performance operator was used to show the performance of the classification model. Apply Model (2) has same function as Apply Model in *Figure 6*, these operator applies the model on the set.

Classification Models

The algorithms selected by the researcher to use are KNN (K-Nearest Neighbor), Naive Bayes and Support Vector Machine (SVM).

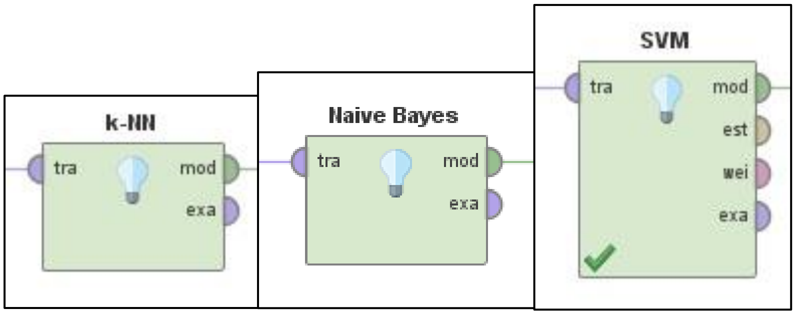


Figure 9. ***The Algorithms***

K- Nearest Neighbor is the simpliest clasification algorithm in machine learning. In KNN classification, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k-nearest neighbors. [8]

Naive Bayes is a simple technique in constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Support Vector Machine moreover is effective in high dimensional spaces. This is also effective in cases where number of dimensions is greater than the number of samples. In addition, it uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Table 3.
Confusion Matrix

	True Positive	**True Negative**	Class Precision	F-measure
Pred. Positive	A	B	$\frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}}$	$\frac{2\mathrm{A}}{2\mathrm{A}+\mathrm{B}+\mathrm{C}}$
Pred. Negative	C	D	$\frac{\mathrm{D}}{\mathrm{C}+\mathrm{D}}$	$\frac{2\mathrm{D}}{2\mathrm{D}+\mathrm{B}+\mathrm{C}}$
Class Recall	$\frac{\mathrm{A}}{\mathrm{A}+\mathrm{C}}$	$\frac{\mathrm{D}}{\mathrm{B}+\mathrm{D}}$		
Accuracy	$\frac{\mathrm{A}+\mathrm{D}}{\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{D}}$			

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset.. It can also be calculated by 1 – ERR.The **Pecision** (PREC) tells us how often the pediction in each category is correct given the actual value. It is also called positive predictive value (PPV). **Sensitivity** or **Recall** (SN) tells us how often the predicted value will give the same result. It is also called recall (REC) or true positive rate (TPR). The best sensitivity, recall and accuracy is 1.0, whereas the worst is 0.0.**F-measure** is interpreted as the weighted average of the precision and recall.

The predicted annotation for the testing set is produced by the rapid miner and was manually computed by the researcher. The basis of the best model was the accuracy.

RESULTS AND DISCUSSION

In this section, results and discussion of the study are discussed.

The frequent terms generated by the database is shown on *Table 4*, these terms were used by the filipino twitter users with or without suicidal ideation. The table below shows the top 10 most frequent word occurences in the gathered tweet data:

Table 4.
Frequent terms in the Database

Yes		**No**	
Wala	112	wala	81
Ayoko	72	mahal	21
Mabuhay	54	mali	18
Sawa	40	sarap	17
pakamatay	34	minsan	15
Sarili	34	sarili	12
Suko	28	oras	11
nahihirapan	22	gawin	10
Pangit	18	isip	10
Matapos	17	tanga	10

As shown above, the word ***wala*** is the most dominant word in both sentiment, thus this word will be disregarded. So, the researcher considered *ayoko* as the most frequent term in Yes (with suicidal ideation) with 72 occurences. In No (without suicidal ideation), the researcher considered *mahal* as the most frequent term with 21 occurences.

Table 5.
The Results after using FP-Growth

Support	Item 1	Item 2
0.09	ayoko	
0.043	mabuhay	
0.027	sawa	
0.026	sarili	
0.021	pakamatay	
0.019	suko	
0.017	mali	
0.015	minsan	
0.015	panget	
0.013	hirap	
0.012	mahal	
0.01	ayoko	Mabuhay
0.01	sawa	Mabuhay

Table 5 tells us about the produced support count using FP-Growth operator in RapidMiner 8.0. The topmost frequent itemsets occured are *ayoko* with 0.09 support count and *mabuhay* with 0.043 support count. This means that these two word are the most used Filipino suicidal term.

Table 6.
The Results after Creating Association Rule

Premise	Conclusion	Support	Confidence
ayoko	mabuhay	0.01	0.381
sawa	mabuhay	0.01	0.273
suko	mabuhay	0.01	0.239
hirap	mabuhay	0.009	0.239
ayoko,sarili	mabuhay	0.009	0.239
suko, hirap	mabuhay	0.005	0.233
pangit	mabuhay	0.005	0.216
suko, sawa	mabuhay	0.004	0.216
suko	sukong	0.004	0.183
suko	sarili	0.004	0.13
hirap	sarili	0.003	0.127
hirap, ayoko	sarili	0.003	0.112
ayoko, hirap	mabuhay	0.003	0.112

Table 6 tells us that *ayoko* together with *mabuhay* are the dominating itemsets among the others with a support of 0.01 and a confidence of 0.381. This means that the suicidal term *ayoko* implying to *mabuhay* strong relationship within each other. Also, it also implies that *sawa* has a strong relationship with *mabuhay* with a support of 0.01 and confidence of 0.273.

By thoroughly examining Table 6, it also shows that if *ayoko* (premise) were used in a tweet having a suicidal thought, then there is a high chance that *mabuhay* (conclusion) was used too in that tweet. Always keep in mind that Association Rule Mining is not consequential but co-existential. This means that if the premise occurred, then the conclusion also occurred at the same time.

The models used to classify the suicidal ideation namely: K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Naive Bayes are presented below. The confusion matrix for every model is presented that includes Precision, Recall, Accuracy and F-Measure.

Table 7.
Confusion Matrix for KNN*

		True Yes	True No	Class Precision	F-Measure
	Pred. Yes	159	18	**68.31%**	**69.43%**
Training	Pred. No	122	263	**89.83%**	**78.98%**
	Class Recall	**56.58%**	**93.59%**		
	Pred. Yes	248	53	**82.39%**	**70.96%**
Testing	Pred. No	150	352	**70.12%**	**77.62%**
	Class Recall	**62.31%**	**86.91%**		

*K-Nearest Neighbor

As reflected above, for the training set, the class precision shows that out of true labels the correctly predicted“Yes” is 68.31 % where in testing set it was 82.39 %. In training set, the correctly predicted “No” is 89.83 % and 70.12 % in testing set. This shows that out of true label, predicted “No” has higher precision in training set and it was “Yes” in testing set.

The class recall, for the training set, out of the predictions the correctly predicted “Yes” is 56.58 % and 62.31 % in the test set. On the other hand the correctly predicted “No” in the training set is 93. 59 % and 86.91 % in the test set. In both traing and testing set the correctly predicted “No” has higher percentage. Which means it has high specificity.

The F-measure shows that in both training and testing set, the higher correctly predicted is “No” having 78. 98 % and 77.62 % respectively.

Table 8.
Confusion Matrix for SVM*

		True Yes	True No	Class Precision	F-Measur
	Pred. Yes	265	194	**57.73%**	**71.62**
Training	Pred. No	16	87	**84.47%**	**45.31**
	Class Recall	**94.31%**	**30.96%**		
	Pred. Yes	370	215	**63.25%**	**74.90**
Testing	Pred. No	33	185	**84.86%**	**59.87**
	Class Recall	**91.81%**	**43.02%**		

* Support Vector Machine

As shown on the table, for the training set, the class precision shows that out of true labels the correctly predicted “Yes” is 57.73 % where as in testing set it was 63.25%. In training set, the correctly predicted “No” is 84.47 % and 84.86% in testing set. This shows that out of true labels, predicted “No” has higher precisionin training set and testing set.

The class recall, for the training set, out of the predictions the correctly predicted “Yes” is 94.31% and 91.81 % in the test set. On the other hand the correctly predicted “No” in the training set is 30.96 % and 43.02 % in the test set. This means that SVM has high sensitivity.

The F-measure shows that in both training and testing set, the higher correctly predicted is “Yes” having 71. 62% and 74.90 % respectively.

Table 9.
Confusion Matrix for Naive Bayes

		True Yes	True No	Class Precision	F-Measure
Training	Pred. Yes	258	185	**58.24%**	**71.27%**
	Pred. No	23	96	**80.67%**	**48.00%**
	Class Recall	**91.81%**	**34.16%**		
Testing	Pred. Yes	356	205	**63.46%**	**73. 94 %**
	Pred. No	46	196	**80.99%**	**60.94%**
	Class Recall	**88.56%**	**48.88%**		

The class recall, for the training set, out of the predictions the correctly predicted "Yes" is 91.81 % and 88.56 % in the test set. On the other hand the correctly predicted "No" in the training set is 34.16 % and 48.88 % in the test set. The correctly predicted "Yes" in bothh sets is higher than the "No".

As also shown on the table, for the training set, the class precision shows that out of true labels the correctly predicted "Yes" is 58.24 % where as in testing set it was 63.46%. In training set, the correctly predicted "No" is 80.67 % and 80.99 % in testing set. This shows that out of true labels, predicted "No" has higher precision in training set and testing set or simply it has high specificity.

The F-measure shows that in both training and testing set, the higher correctly predicted is "Yes" having 71. 27% and 73.94 % respectively.

Table 10.
Algorithms' Accuracy

	Accuracy	
	Training Set	Test Set
K-Nearest Neighbor	75.09%	74.71%
Support Vector Machine	62.63%	69.12%
Naïve Bayes	62.99%	68.74%

In comparing the Accuracy for training set and test set, If the training error is apparently lower than the testing error this means that the model easily overfits to the training data yet poorly generalizes. This happens usually but not always. In statistics, overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably". This is what happened in the KNN model.

KNN algorithm gives the highest accuracy with 75. 09 % and 74.71% respectively. However, as shown on Table 7 (see *pg. 24*) , KNN has high specificity which means it predicts the "No" (without suicical ideation) accurately than the "Yes" (with suicidal ideation). On the other hand, as shown on Table 8 (see *pg. 25*), SVM algorithm accurately predicts the "Yes" (with suicidal ideation) than the "No" (without suicidal ideation) with 62.63% and 69.12 % accuracy in training set and test set.

Eventhough KNN has the highest accuracy it clearly shows that Support Vector Machine (SVM) is the best model that gave the best results in determining the possible suicidal ideation.

CONCLUSION

The researcher focused on Filipino suicidal ideation in the tweet posts on Twitter, one of the most popular social networking sites of this generation. The researcher created a Filipino (specifically tagalog) dictionary to further predict the suicidal thoughts and to determine the best model that would determine this ideation.

After analyzation and interpretation of data, the following conclusions are derived:

1. The researcher arrived at a practical generalization of Filipino suicidal ideation.
2. The tweets can be classified whether yes (with suicidal ideation) or no (without suicidal ideation). The most dominant word occured in both "yes" and "no" are *wala* which was disregarded by the researcher. In yes, the terms that occured was: *ayoko, mabuhay, sawa, pakamatay, etc*. In no, the words are *mahal, mali, sarap, minsan, etc.*
3. The most frequent itemsets occurred are *ayoko* and *mabuhay* with a support count of 0.09 and 0.043 respectively. This means that these two words are the most common terms used together to express suicidal ideation in Twitter. The resulting top trend in Filipino suicidal ideation after generating the data from repositories were (*ayoko*→ *mabuhay*) with a confidence of 0.381 and (*sawa* → *mabuhay*) with a confidence of 0.273.
4. Support Vector Machine (SVM) is the best model that gave the best results in determining the possible suicidal ideation since it has the highest sentivity among the models.

RECOMMENDATIONS

After analyzing the data the following recommendations are made:

1. The researcher highly encourage the future researchers to deeply explore the Filipino suicidal ideation in Twitter or in other social networking sites like Facebook and Instagram.
2. The researcher also encourage the use of clustering techniques.
3. The researcher hopes that this study will be developed further to classify the suicidal ideation of Filipino social media user on Twitter or in other social networking sites.

BIBLIOGRAPHY

A. Theses and Dissertations

Qijin Cheng,Tim MH Li, Chi-Leung Kwok, Tingshao Zhu, and Paul SF Yip (2017). **ASSESSING SUICIDE RISK AND EMOTIONAL DISTRESS IN CHINESE SOCIAL MEDIA: A TEXT MINING AND MACHINE LEARNING STUDY.** JMIR Publication

MarouaneBirjali, AbderrahimBeni-Hssane,and Mohammed Erritali (2016). **PREDICTION OF SUICIDAL IDEATION IN TWITTER DATA USING MACHINE LEARNING ALGORITHMS.**International Arab Conference on Information Technology

Bimala Sharma, Eun Woo Nam, Ha Yun Kim,and Jong Koo Kim(November 2015). **FACTORS ASSOCIATED WITH SUICIDAL IDEATION AND SUICIDE ATTEMPT AMONG SCHOOL-GOING URBAN ADOLESCENTS IN PERU**. Int J Environ Res Public Health.

Pete Burnap, Gualtiero Colombo, and Jonathan Scourfield (2015), **MACHINE CLASSIFICATION AND ANALYSIS OF SUICIDE-RELATED COMMUNICATION ON TWITTER.**Published Online

BridianneO'Dea , Stephen Wan , Philip J. Batterham, Alison L. Calear, Cecile Paris, Helen Christensen (2015). **DETECTING SUICIDALITY ON TWITTER**. Elsevier B.V.

Omprakash L. Mandge (2013). **A DATA MINING TOOL FOR PREDICTION OF SUICIDES AMONG STUDENTS.**National Conference on New Horizons in IT - NCNHIT 2013

Seung-Min Bae, Yu Jin Lee, In Hee Cho, SeogJu Kim, JeongSooIm, and Seong-Jin Cho (2013). **RISK FACTORS FOR SUICIDAL IDEATION OF THE GENERAL POPULATION**. Journal of Korean Medical Science, 602-607

Brent DA, Mann JJ (2005). **FAMILY GENETIC STUDIES, SUICIDE, AND SUICIDAL BEHAVIOR.**US National Library of MedicineNational Institutes of Health

Rumi Kato Price, Nathan K. Risk, and Edward L. Spitznagel (2004).**DATA-MINING APPROACHES TO SUICIDE AND SUICIDAL BEHAVIOR.**Published Online

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu (NYI). **EMPATWEET: ANNOTATING AND DETECTING EMOTIONS ON TWITTER.** Published Online

Munmun De Choudhury , Michael Gamon, Scott Counts, Eric Horvitz NYI. **PREDICTING DEPRESSION VIA SOCIAL MEDIA.**Published online

Nugent, Pam M.S., NYI. **SUICIDAL IDEATION** .*psychologydictionary.org.*

B. Internet Sources:

[1] http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/ , **RETRIEVED NOVEMBER 8, 2017**

[2] https://www.statista.com/statistics/278414/number-of-worldwide-social-network- users/ , **RETRIEVED NOVEMBER 8, 2017**

[3] http://www.newsweek.com/social-media-mental-healthfacebookdepressionsocial- mediatwittermental-601955 , **RETRIEVED NOVEMBER 8, 2017**

[4]https://wearesocial.com/sg/blog/2017/01/digital-in-2017-global-overview , **RETRIEVED NOVEMBER 8, 2017**

[5] https://www.wired.com/2017/03/artificial-intelligence-learning-predict-prevent-suicide/ , **RETRIEVED NOVEMBER 8, 2017**

[6] http://www.valleybehavioral.com/suicidal-ideation/signs-symptoms-causes , **RETRIEVED NOVEMBER 8, 2017**

[7] https://www.suicideline.org.au/worried-about-someone/recognising-suicide-warning-signs/ , **RETRIEVED NOVEMBER 8, 2017**

[8] https://en.wikipedia.org/wiki/Knearest_neighbors_algoritm , **RETRIEVED JANUARY 9, 2018**

[9] https://en.wikipedia.org/wiki/Naive_Bayes_classifier , **RETRIEVED JANUARY 9, 2018**

[10] http://scikit-learn.org/stable/modules/svm.html , **RETRIEVED JANUARY 9, 2018**