Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

Cyber Bulling Detection Using Various Machine Learning Models Over Twitter

Garima Jain[1], Ms. Anjali[2]
[1]Research Scholar, Department of Computer Science and Engineering, JC Bose University at Faridabad, Haryana, India.
[2]Assitant Professor, Department of Computer Science and Engineering, JC Bose University at Faridabad, Haryana, India.

Abstract: **Twitter is one of the social media that is often used by Indians. Not a few users use Twitter to carry out negative actions such as fraud, spreading fake news, writing things that tend to contain hate speech, to online bullying (cyber bullying). Twitter has developed features such as to reply restriction, rethinking methods, blocking, muting, and reporting to prevent cyber bullying. But to identify each and every tweets that is in context to cyber bulling or not is very cumbersome. Therefore, this study will classify tweets that contain cyber bullying from specified words using the Logistic Regression, Naive Bayes, Decision Tree, and Support Vector Machine (SVM) algorithm with the inculcation of NLP models like CBOW and N-Gram models mark them with the specified label. In addition, evaluations were conducted to find the most optimal accuracy using precision and recall. The results showed that Naïve Bayes Algorithm has succeeded in classifying Cyber-bullying analysis on tweets with optimal accuracy of 94.4%. Secondly using Decision Tree has succeeded in classifying Cyber-bullying analysis on tweets with optimal accuracy of 71.1%. Thirdly, using Logistic Regression has succeeded in classifying Cyber-bullying analysis on tweets with optimal accuracy of 92.0%. Lastly using SVM has succeeded in classifying Cyber-bullying analysis on tweets with optimal accuracy of 94.9%. Consequently, cyber bullying detection system was implemented that demonstrates the validity of the proposed methodology.**

Keywords: Machine Learning, Naïve Bayes, Support Vector Machine, Logistic Regression, Decision Tree, N-Gram.

1. INTRODUCTION

The great advances in information and communication technologies have given place for applications such as social networks to be easily introduced as everyday tools for work, studies, or entertainment. Thus, social networks have taken an important turn in the way of communicating and sharing information [1]. Social media has allowed the development of techniques to analyze millions of data that are generated day by day. The processing of these data has become a fundamental piece in the definition of strategies like political [3], economic [4], or marketing [5].

The use of social networks to establish cyberbullying attacks was a practice little known in our environment, which worries the control authorities in all parts of the world. This is due to the lack of specialists in the forensic investigation in cyberspace, who can understand the problems of attacks and aggressions that occur in social networks, how the network works, how the system works, and what the vulnerabilities. These correspond to the factors where efforts should be directed to combat this type of harassment, which is increasingly they become more multifarious and complicated. However, this category of crime came under the purview of Section 354D of the IPC[6].

In addition, [7] states that the worst punishment that an adolescent can suffer is being isolated from their peer group, for example, their schoolmates. This is why there is a close relationship between bullying and suicide, and even this relationship has amplified in current years due to the appearance of cyber1bullying, which, unlike situations of face-to-face violence, these do not end when he or the adolescent returns home and they correspond to stressful situations in which sexual orientation, gender identity, religion, race, ethnicity, social condition, etc. are questioned. They are high-risk factors that result in suicidal behavior.

To prevent this problem, it is important to be able to establish mechanisms that allow for the detection and prevention of unusual events and early identification of bullying before the conditions produced, whether physical or emotional, can have repercussions. If no research work is carried out that allows for the ability to identify proactively the various types of aggression that are established in social networks, the rates of cyberbullying could increase considerably, and with this, millions of people in their family environments could be affected, for example, children and adolescents do not want to return to their schools or colleges, low self-esteem, an increase in the suicide rate, confusion in their sexual orientation, learning and concentration problems, homicides based on hate or revenge, etc. Even to prevent this type of aggression from occurring, detecting cyberbullying and providing preventive control measures are the main focuses of action to combat cyberbullying [8].

The use of data mining for the analysis and processing of information expressed in natural language allows for maximizing the identification of patterns that, at first sight of users, parents, or school guardians would go unnoticed [9]. Data analytics through the use of classification algorithms allows establishing certain characteristics of the expressions with some type of harassment and with an aggressive character towards users of social networks [10].

2. LITERATURE REFERENCES

As defined by [11], it is also known as bullying. cybernetic, and is understood, as the constant and malicious

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

damage done to a person considered as a victim, who is not able to defend himself by his means and is made or executed using electronic means such as the internet, mobile phones, or computers. Victims of cyberbullying are generally selected for meeting certain characteristics that identify them as weak both physically and emotionally. Also, I now feel different from other people, so they become an easy target for aggressors. The most frequent attacks are usually rumors, offenses, insults, threats, extortion, and intimidation, this through the use of electronic devices or means [12].

Researchers in [13] elaborated out that, even though the use of internet world has unremarkable reward for civilization, the intermittent use of these resources, also has considerable undesirable costs. This engage sexual disclosure sought after, cyber1crime, and cyber1bullying; thus sexual disclosure is whilst delinquent posing as sufferers in online commercial and incorrectly propose that their1victims are fascinated in sex. Revelation to these categories of occurrence has been associated to hopelessness, declined in self assurance, isolation, nervousness, and suicidal feelings.

Further reference uses a 2 classification approach with the same data. The first classification is based on the word abusive with the labels abusive and none, and the second approach is the classification of types of abusive, namely sexism and racism. Many algorithms are used as comparisons but the Hybrid CNN algorithm is also suggested in this study. The advantage of this research is that it is very active in classifying data into 2 steps and conducting research on various algorithm models. But in another case, the researcher said that there was still a lack of classification data, because the type of abusive was not only sexism and racism but there were others such as homophobic and others, according to the researcher, it would be difficult to find [14]. The advantage with the use of data critics that divide in 2 steps finally decided to use this reference as a reference in proving that in large numbers, data is very influential in the accuracy of the model. In addition, logistic regression algorithm can also be used on continuous data. The results of the logistic regression model have quite high precision and recall, reaching 954 and 953.

The second reference classifies the movie review using an algorithm that can use supervised (structured) data types. The data used is very large and the classification is clear step by step, so that it is used as a reference for reference. Many steps are learned from this paper because of the similar model cases. In addition, the results of the logistic regression algorithm for each model show convincing results. The deficiency in this reference is the absence of further research using data other than those conducted by researchers [15].

The third reference is a classification that customer reviews get from the Amazon API. This classification compares the Logistic Regression, Naive Bayes and SentiWordNet algorithms using the Matrix quality evaluation method and performance measures. The weakness of this reference lies in the results of the highest Naive Bayes algorithm, but in the first performance the results of the logistic regression algorithm using the Apple Iphone 5 are not so bad. The weakness of this reference lies in the research when using different datasets [16].

The fourth reference calculates the user to add validation to the information. The data taken came from Twitter with a total of 46,895 tweets. This fourth reference has a clear discussion of sentiment processes and algorithms from the ground up and goes deeper into evaluation. So that it is used as a reference in using evaluation. The evaluation results also show that the algorithm shows better results than the Naive Bayes Classifier algorithm [17]. Therefore, the four references are needed to combine the steps carried out in this study. As well as knowing how much the logistic regression algorithm can achieve optimal accuracy with Instagram post caption data.

Further research entitled " Classification and feature selection techniques in data mining", data classification is performed using the Bayes classifier method. The final result shows that the Bayes classifier method can be used for the process of classifying the resulting data from data mining [18]. The last research entitled "On the Rates of Convergence from Substitute Risk Reduction to the Bayes Optimal Classifier". States that, using the Bayes Optimal Classifier method in level comparisons convergence can minimize posteriori data from various algorithms classification [19].

3. PROPOSED ALGORITHM

3.1 ARCHITECTURE DESCRIPTION

This Cyberbullying Analysis is a model used for checking the content vide tweets that contain Cyberbullying and NonCyberbullying content. This model is created using Logistic Regression, Naïve Bayes, Support Vector Machine, and Decision Tree and will be evaluated using Split Data vide N-Gram and Confusion Matrix validation.

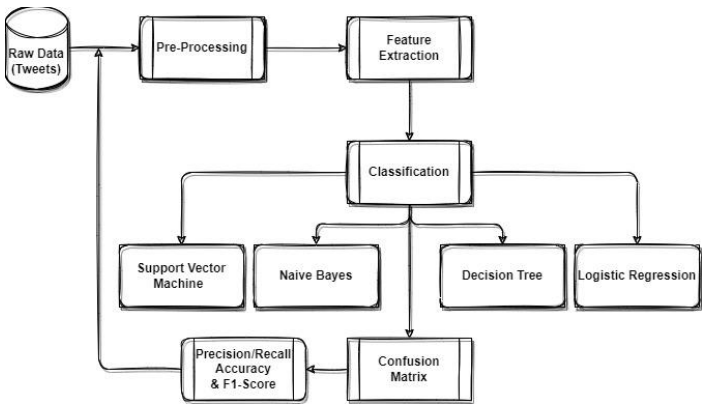

Figure 1: Research Procedure.

3.2 WORKFLOW PROCESS.

Cyberbullying analysis research was developed through the Machine Learning Models Development method which we evaluate the performance among the various techniques like

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

Logistic Regression, Naïve Bayes, Support Vector Machine and Decision Tree with the following stages:-

3.2.1 Pre-Processing

Pre-processing data, namely the process of making data labels, as well as transforming and cleaning data from noise. This stage is the stage that takes the most time because it has several processes in it, namely:

1. Cleansing, namely cleaning data from characters that can not be detected by the system. Examples such as !:, #, $
2. Case Folding, namely the process of changing words from uppercase to lowercase letters.
3. Tokenizing, namely the process of separating sentences into words.
4. Stemming, namely the process of eliminating words that contain affixes and changing words into basic words.
5. Stopwords, that is the process of removing conjunctions.

3.2.2 Logistic Regression

Logistic regression is one of the most frequent classification methods used. Binary logistic regression is used when the dependent variable is dichotomous variables. Multinomial logistic regression was used at variable times the dependent variable is a categorical variable with more than two categories. Generally logistic regression models are [20-22]:-

$$nx = \frac{e^{\beta 0+\beta 1x1 + \beta 2x2 + \cdots + \beta kxk}}{1 + e^{\beta 0+\beta 1x1 + \beta 2x2 + \cdots + \beta kxk}} \ (eq.1)$$

Where $\pi\ (x)$ is the probability value of $0 \le \pi\ (x) \le 1$, which means that logistic regression describes a probability. By transforming $\pi\ (x)$ in the above equation with logit transformation $g\ (x)$, where:

$$\boldsymbol{g(x) = In\left[\frac{\phi_k(x)}{\phi_o(x)}\right] = In\left[\frac{P(Y = K \mid x)}{P(Y = 0 \mid x)}\right]} \ \boldsymbol{(eq.2)}$$

then we get the logistic form:

$$g(x) = \beta k0 + x'\beta k,\ k = 0,1,2, \ldots, K,\ (eq.3)$$

To obtain estimates from logistic regression parameters, it can be done by using the Maximum Likelihood Estimation (MLE) as follows: Estimation of parameters in the logit model uses Maximum Likelihood with the following steps.

$$Logistic \quad (Y \le k|x) = ck(x)$$
$$= \boldsymbol{In\left[\frac{P(Y \le k \mid x)}{P(Y > k \mid x)}\right]}$$
$$= \boldsymbol{In\left[\frac{\phi_0(x) + \phi_1(x) + \ldots + \phi_k(x)}{\phi_{k+1}(x) + \phi_{k+2}(x) + \ldots + \phi_k(x)}\right]}$$
$$= \boldsymbol{\tau k - x'\beta},\ for = 0,1,2,\ldots K - 1\ (eq.4)$$

3.2.3 Naïve Bayes

Naive Bayes is a classification algorithm, used for machine learning and/or machine learning. According to the author [23], a Naïve Bayesian classifier provides a simple approach, with clear semantics, to represent, use and learn probabilistic knowledge. This method is used for supervised tasks since its main objective is to predict according to a defined class with certain characteristics, in which data is entered to know how many contain characteristics of the previously defined class.. Graphically, a Bayesian classifier is represented as shown in equation 5, where the C in the first node represents the class, and the nodes below it represent the instances that have characteristics of the class. According to [24, 25] the following example shows how a Naive Bayes algorithm works expressed in a formula.

$$P\ (C = c/\ X = x) = \frac{P(A)P(B|A)}{P(B)} \quad (eq.5)$$

However, In probability theory and data mining, a classifier Naïve Bayes is a probabilistic classifier based on Bayes' theorem and some additional simplifying assumptions. It is because of these simplifications, which are usually summarized in the hypothesis of independence between the predictor variables, that it is called naive.

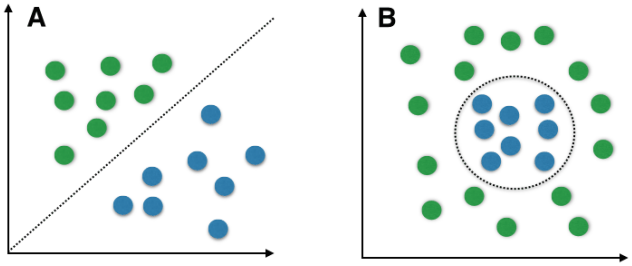

Figure 2: Naïve Bayes Model

In simple terms, a classifier of Naïve Bayes assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the variable class. For example, fruit can be considered an apple if it is red, round and about 7 cm in diameter. a classifier Naïve Bayes considers that each of these characteristics contributes independently to the probability that this fruit is an apple, regardless of the presence or absence of the other characteristics. The formula to calculate the highest probability that given certain values for the characteristics of an object belongs to a certain class is shown in Equation 6. In this formula, it can be seen how, based on the Thomas Bayes theorem, there are different probabilities for each class to which a set of characteristics of a single example could belong, according to this formula the highest probability is taken. Libraries will be used NLTK to implement the Bayesian classifier even as.

$$^{v}\ MAP \ = \ argmax_{v_j} \in V\left(P\left(V_j \middle| a_{1,} \ldots., a_n\right)\right) (eq.6)$$

Naïve Bayes is based on the simplifying assumption that the value of the grain. The conditionals are mutually independent if an output value is given or it can be given as :-

$$\hat{P}\ (\mathrm{t} \mid \mathrm{c}) \ = \frac{T_{ct}+1}{\sum t' \epsilon V T_{ct}+B'} \ (\mathrm{eq.7})$$

3.2.4 Support Vector Machine

SVM is an amalgamation of existing classification theories such as lagrange, kernel and margin hyperplane. SVM has a dividing function in the classification of the two classes linearly but due to the increasing need for classification SVM was developed not only for linear classification. However, SVM can already do non-linear classification with a combination of kernels in the feature space (high dimensional space). SVM can also perform regression whose output is real numbers or this algorithm is also called

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

Support Vector Regression (SVR). By using SVR, classification prediction between several classes can be made. The advantage of SVM is that it is effective in high dimensional spaces and different function kernels can be defined for decision functions. SVM is a machine learning system whose working principle uses Structural Risk Minimization (SRM). SRM aims to get the best dividing line (hyperplane) on the input space in the two class classification. By measuring the margin on the dividing line and finding the vertex, it will be possible to find the best dividing line between the two classes. The distance in each of the two classes is called the margin and the pattern closest to the classification between the two classes is called the support vector. The learning stages in SVM are [26-28]:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i.j}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i \vec{x}_j \quad (eq.8)$$

Where

$$\alpha_i \geq 0 (i = 1,2, \dots l) \sum_{i=1}^{l} \alpha_i y_i = 0 \quad (eq.9)$$

Information:
y_i= training data class (+ 1 / -1).
y_j= training data class (+ 1 / -1).
x_i= weight vector for the comment sentence.
x_j= weight vector for the comment sentence.

3.2.5 Decision Tree (ID3)

A decision tree consists of nodes that form a directed tree with a root node that does not have a predecessor node. All other nodes have exactly one predecessor node. A node with output arcs is known as an internal node or test node, while nodes without successor nodes are called leaf nodes or decision nodes. So, mathematically, a decision tree is a directed acyclic graph with a root node A = (G(V, E), v1), where V is a set of nodes, E is the set of links that join pairs of nodes of V, and v1 is the root node [29-32].

The learning or induction of decision trees is performed from a set of data using an algorithm that allows obtaining a model that tries to discover the relationship between the input attributes and the target class. When applied to new examples, it can predict the value of the target class and thus classify them correctly. This set of data that is used to train or obtain the model is called a training set. Attributes can be of two types: nominal or numeric. When an attribute is nominal it is useful to denote it as N = {v_i,1, v_i,2, v_i,3, ..., vi |dom(ai)|} 'this is the domain of all possible values it can have. Whereas dom(c) = {c_i,1, c_i,2, c_i,3, ..., c_i |dom(c_i)|} represents the domain of the target classes. The set of all possible instances is called the instance space and is defined as the Cartesian product of all input attributes X = {dom(a1) x dom(a1) x_1 . . . x_n dom(an)}. The training set consists of a set of m tuples, each one described by a vector of attribute values. Formally, the**1**training**1**set is represented by S(R) = (<x_1, y_1>,..., <x_m, y_m>) where $x_i \in X$ and $y_i \in$ dom (c).
Compared to other black-box techniques, decision trees are very transparent, by simple inspection, it can be argued why the model classifies in one class or another.
These models can be used to solve complex problems. Depending on the number of attributes that are evaluated in each test condition performed in a node, two types of decision trees can be induced: univariate trees and multivariate trees [29-32].
Univariable decision trees are the best known and the methods to induce them are simpler. At each node, only one attribute is used, if it is continuous then the decision is of the form

$$a_i > t_j \text{ (eq.10)}$$

doing a binary division where $\boldsymbol{a_i}$ is the i[th] input attribute and $\boldsymbol{t_j}$ is a suitably chosen constant or threshold and represents the cutoff point of the ai attribute. If $\boldsymbol{a_i}$ has a discrete value within a set of nominal values with m values, then an m-array division is performed at the node. On the other hand, in multivariable decision trees, at each node, the decision must be made by taking into account more than one attribute. For this type of tree, the algorithms used to induce them do not select the best attribute as in the case of univariates, but instead use the linear combination of the attributes. In this case, the decisions made at each node take the form:

$$\sum_{i=1}^{n} x_i a_i > t_0 \ (eq.12)$$

where $x_i$are the weights, and t_0 is a threshold. This type of tree is used when there is a complex data distribution in the training set. Univariate trees are also known as axis-parallel trees, this is because at each internal node the test conditions represent hyperplanes parallel to the axes, as can where each hyperplane is parallel to either the y-axis or the x-axis, and thus each divides or separates the space of the examples into two regions. On the other hand, multivariable trees are also called oblique, since the tests they perform at each node are equivalent to hyperplanes in an oblique orientation to the axes of the attribute space [29-32].

- **Algorithm C4.5**

This algorithm is an advanced version of the ID3 algorithm that includes the following capabilities or advantages.

a. Handling of continuous and discrete values: To handle continuous attributes, what the algorithm does is create a threshold and then divide the attribute between those that are above and below the threshold.
b. This characteristic is fundamental for this study where most of the values are continuous and their thresholds can be of high relevance.
c. It can handle missing attribute values: In the case of a missing attribute, the algorithm uses a weighting of values and probabilities instead of close or common values. This probability is obtained directly from the observed frequencies for that instance, so it can be said that algorithm C4.5 uses the most probable classification calculated as the sum of the weights of the frequencies of the attributes.
d. It is capable of generating a set of rules that are much easier to interpret for any type of tree.
e. This algorithm builds a large tree and concludes it with a "pruning" of the branches to simplify it to generate results that are easier to understand and make it less dependent on the test data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

Limitations of algorithm C4.5: Although this algorithm is one of the most popular, it has some deficiencies, among which we can name:

a. The presence of empty branches: Sometimes trees have branches that have nodes with almost zero values or are very close to them. These nodes do not help with building rules or with the classification of attributes and only make the tree larger and more complex.
b. Insignificant branches: This happens when the number of discrete attributes creates the same number of branches to build the tree, generating branches that are not relevant to the classification task, decreasing its predictive power, and generating the problem of overfitting.
c. Overfitting: This problem occurs when the algorithm selects information with unusual characteristics, causing fragmentation in the distribution process.

Fragmentation is defined as the creation of statistically insignificant nodes through which very few samples pass. Creating a tree that classifies all the training data perfectly may not lead to better generalization of data with unusual features. This problem usually manifests itself with noisy data. To solve this problem, this algorithm has the techniques of pre and post-pruning of the tree. Pre-pruning consists of stopping the growth of the tree in its construction when there is not enough data to obtain reliable results. In the case of post pruning, once the entire tree has been generated, all those subtrees that do not have enough evidence are eliminated, replacing it with the class of the majority of the remaining elements or with the probability distribution of the class. To select which subtrees should be trimmed, the following methods are used:

1. Cross-validation: It consists of separating some training data from the tree to evaluate the usefulness of the subtrees.
2. Statistical test: Use a statistical test on the training data to identify information with random origin 3. Minimum description length (MDL): It consists of determining if the additional complexity of the tree is better than simply remembering exceptions resulting from trimming.

3.2.6 N-Gram

An n-gram is a sequence of nwords of a natural language expression as they appear in the expression, and are typically used in Natural Language Processing applications such as autocorrect, speech recognition, machine translation, part of speech tagging, natural language generation, the similarity between words, identification of authorship among others. This case indicates how many elements should be taken, that is, the length of the sequence of n-gram. For example, there are bigrams, trigrams, quadrigrams (2- grams, 3-grams, 4-grams), etc [33-35].

$$Unigram = P(W_1^n) \approx \prod_i P(w_i) \quad (eq.13)$$

$$Bigram = P(W_i|W_{i-1}) = \frac{N\ (w_{i-1},w_i)}{N\ (w_{i-1})} \quad (\text{eq.14})$$

$$Trigram = P(W_1^n) \approx \prod_{i=1}^{n} P(w_i|w_{i-1}) \ (eq.15)$$

$$N-Gram = P(W_i|W_{i-2},W_{i-1}) = \frac{N(w_{i-2},w_{i-1},w_i)}{N(w_{i-2},w_{i-1})} \ (eq.16)$$

Table 1 shows an example of the sequences of 2 and 3 words (bigram and trigram) for the sentence: “The tree in my house is very big”.

GRAM	Sequences
2 Words – Bigram	The tree, tree of, of my, my house, the house is, is very, very big.
3 Words – Trigram	The tree of, tree of my, of my house, my house is, the house is very, very big.

Table 1: Examples of Ngrams obtained from a sentence.

4. RESULTS AND SIMULATION

At this stage, the system begins to be built following the results of the proposed methodology proposed in section 3. This stage also carries out the Data Preprocessing and Modeling process in the python framework. Before entering the Data Preprocessing stage, the data is annotated first. In this case, collaboration with an annotator expert was carried out to label the data into 2, namely Cyberbullying (negative) and Non-cyberbullying (positive). The annotator is embossed using polarity using the python program. After the data is annotated, the Data Preprocessing process is carried out in which it has 3 steps, namely Labeling, Pre-Processing, Classification using N-Gram (Uni/Bi/Tri) with Logistic Regression, Naïve Bayes, SVM, and Decision Tree. The following is an example of the steps in conducting data pre-processing and classification.

```
#Import libraries
import pandas as pd
import numpy as np
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import string
from nltk.stem import WordNetLemmatizer
# In[2]:
#Load dataset
import pandas as pd
df1 = pd.read_csv("dataset.csv")
```

Figure 2: Loading the Raw Data in Memory Stream

The data is taken from Twitter is in the form of English-language tweets. The tweets collected to build the model in this study were 1065 tweets as shown in Figure 4.1 with a count of 638 non-cyberbullying tweets and 427 cyberbullying tweets, respectively.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

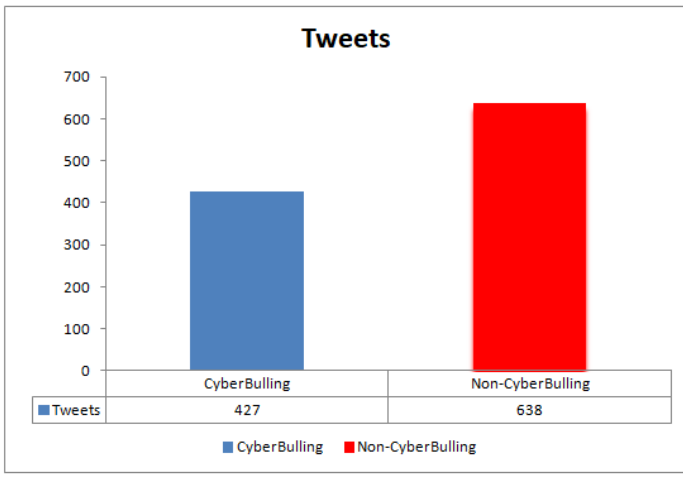

Figure 3: Tweets Count Data

4.1 PREPROCESSING

The next step the system takes in finding the tweets of an input document file. Based on Figure 3.1, it can be explained that the stages of the process or how the application system works for basic word searches for tweets with the Confix-Stripping approach in general are as follows:

1. Process 1 (file / document) The word to be stemmed is searched first in the document . If found, it means that the word is the root word, otherwise process 2 is carried out.

2. Process 2 (Parsing / tokenizing) This stage will check from the first character to the last character. If the to (i) character is not a word break then it will be added with the next character. For example, this word break character is like a punctuation mark or a space.

3. Process 3 (Stopword) This stage takes important words from the token results. Can use a stop list algorithm (remove words that are less important) or a word list (store important words). This system uses the stop list method, which is the elimination of insignificant words (stop words) in the description by checking the words resulting from the description token, whether they are included in the stop list or not. If included in the stop list, these words will be removed from the description so that the remaining words in the description are deemed important words or keywords (pattern). The stages of the stopword process are as follows:

a. The resulting word stemming is compared to the stopword table.

b. Check whether the token is the same as the stopword table or not.

c. If the token is the same as the stopword table it will be removed.

d. If the token does not match the stopword table will be displayed. That is to produce a stopword token that includes important words (keywords)

4. Process 4 (Stemming) At this stage, the Stemming process is carried out based on the input of the filter term list, the Stemming process uses the Porter process. The first step in the stemmer algorithm is to check the Precedence rule, namely the prohibition of combination of prefix and suffix, then match the terms on certain indexed elements with the list of "root words" in the dictionary database. If they match, the terms are stored in the variable stem immediately. If the Precedence rule returns the correct value, the suffix decapitation process is carried out, if not, then continue with the prefix decapitation process. Then the recording process is the process of adjusting the basic word with the rules of changing the first letter of the word, whether the recording result is the same as the dictionary. If it is true, then the root word is the same as the result of the recording otherwise the process is repeated. Therefore the below table depicts the model along with code structure.

```
for row in df1["Tweet"]:
    #tokenize words
    words = word_tokenize(row)
    print(words)
    #remove punctuations
    clean_words = [word.lower() for word in words if word not in set(string.punctuation)]
    #remove stop words
    english_stops = set(stopwords.words('english'))
    characters_to_remove = ["'",'`',"rt","https","/","\","\u200b","--","n't","'s",
    "...","//t.c" ]
    clean_words = [word for word in clean_words if word not in english_stops]
    clean_words = [word for word in clean_words if word not in set(characters_to_remove)]
    #Lematise words
    wordnet_lemmatizer = WordNetLemmatizer()
    lemma_list = [wordnet_lemmatizer.lemmatize(word) for word in clean_words]
    Tweet.append(lemma_list)
```

Figure 4: Code Block of Pre-Processing Module

4.2 FEATURE EXTRACTION AND SELECTION

- **Bag-of-words model**

The bag-of-words (BoW) model is a simple representation used for natural language processing (NLP) and information retrieval (IR), also known as vector models space. In this model, a text is in the form of a sentence or the tweets which is represented as a multi-set bag of the words contained in in it, regardless of word order and grammar yet retains diversity. Another definition for BoW is a model that studies a vocabulary from all documents, then modeled each document by counting the numbers the appearance of every word. The below is the code block represented as BoW.

From the results of the evaluation conducted, obtained an accuracy of 69.10% using above formulation and test models under the proposed scheme.

```
#Create bag of words and dictionary object
def bag_of_words(words):
    return dict([(word, True) for word in words])
Final_Data = []
for r, v in combined:
    bag_of_words(r)
    Final_Data.append((bag_of_words(r),v))
```

Figure 5: Code Block of Bag of Words

- **Generate N-gram**

After the twitter text data is normalized and the next transform case is with tokenization using the same type of token unigram, bigram and trigram as in which conducted the distribution N-grams into three types. N-gram is a combination of adjectives often appears to show a sentiment. In research using a type of unigram token, namely a twitter text data token consisting of only one word, then bigram, namely twitter text data token which consists

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

of two words and trigram is a twitter text data token consisting of three words.

```
#Bag of words for bigrams
def bag_of_bigrams_words(words, score_fn=BigramAssocMeasures.chi_sq, n=200):
    bigram_finder = BigramCollocationFinder.from_words(words)
    bigrams = bigram_finder.nbest(score_fn, n)
    return bag_of_words(bigrams)
#Bag of words for Trigrams
def bag_of_trigrams_words(words, score_fn=TrigramAssocMeasures.chi_sq, n=200):
    trigram_finder = TrigramCollocationFinder.from_words(words)
    trigrams = trigram_finder.nbest(score_fn, n)
    return bag_of_words(trigrams)
def bigrams_words(words, score_fn=BigramAssocMeasures.chi_sq,
n=200):
    bigram_finder = BigramCollocationFinder.from_words(words)
    bigrams = bigram_finder.nbest(score_fn, n)
    return bigrams
# In[35]:
```

Figure 6: Code Block of N-Gram Generation

4.3 RESULTS USING NAÏVE BAYES

The step taken after the pre-processing is classification using the Naïve Bayes algorithm the purpose of machine learning is to impart knowledge in a machine or computer so that computers can do work like humans. The task of the Naïve Bayes algorithm in sentiment analysis is to classify data sets into positive, negative, or neutral classes according to the knowledge implanted using training data. The data set and training data used in this study are text-type data, so pre-processing is a crucial part before classification. This research applies bulling class boundaries in classification using machine learning algorithms. This limitation is that the Naïve Bayes algorithm can only classify test data or original data into positive and negative classes. This limitation is applied because the training data used only provides data with positive and negative sentiment classes, so it is very unlikely to produce neutral sentiment classes. The following is an explanation for each stage in the pre-process and the application of the Naïve Bayes algorithms. Making Unigram and Bigram Unigram and bigram are part of n-gram, which is n-word chunks based on the sequential sequence of the text string. Unigram is Ingram where n is one (n-gram is one size), while bigram is n-gram where n is two (n-gram is two). For example, there is a text "YMCA University", then the unigram of the text is "University", "YMCA", while the bigram of the text is "YMCA University". The purpose of using n-gram is to increase the effectiveness of the Naïve Bayes sentiment classification model. The following is the program code for creating unigram and bigram, tokens that have been processed have been loaded in unigram form, so the next step is to form bigram. A bigram is formed by connecting a token with the next token by adding the character "_" (underscore) as a separator character.

```
0.944
bullying precision: 0.9053117782909931
bullying recall: 0.9631449631449631
bullying F-measure: 0.9333333333333333
not-bullying precision: 0.9735449735449735
not-bullying recall: 0.9308600337268128
not-bullying F-measure: 0.9517241379310345
```

Figure 7: Results of Naïve Bayes with N-Gram Generation

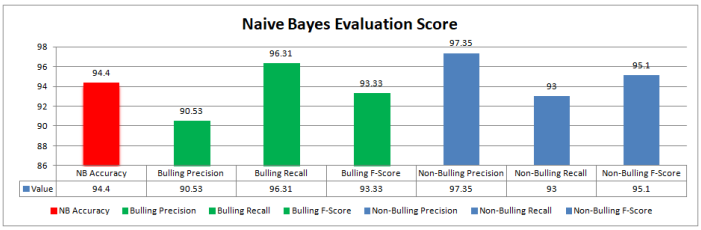

Figure 8: Results based on Naïve Bayes with Evaluation Models

For the bulling class, of the 364 copies of the validation set, 10 were misclassified. For this class, the precision of 90.53 indicates that you are very unlikely to label a bulling shows non-bullling, while the recall of 96.31 indicates that there are almost zero probabilities of labeling a bulling instance as non-bulling. Regarding the positive class, of the 157 really positive samples, 170 were erroneously classified as non-bulling. Based on precision and there call obtained for the non-bulling class, it is concluded that the classifier is unlikely to label a sample as positive negative, but at the same time unlikely to identify all the positive instances as such, that is, it could label as non-bulling a positive sample. Subsequently, the accuracy of 94.40% is achieved using Naïve Bayes.

4.4 RESULTS USING LOGISTIC REGRESSION

This step is the analysis using the Logistic Regression method is to define the Y (response variable) and X (predictor variable) data used. Following this, a simultaneous and partial test will be carried out on twitter data and its predictor variables with the data used as an example of testing which is the data that has the highest classification performance value using unigram, bigram, trigram and N-Gram. Based on the data obtained from grams, it can be seen that the value of bulling tweets are abusive eywords based on sentiwords dictionary which means it can be decided can be concluded that there is a significant influence between the keyword variables on the classification variable for bulling and non-bulling sentiments. Furthermore, based on the results of the parameter significance test, the decision was made that the auxiliary variable based on accuracy precision and recall.

```
The accuracy of Logistic Regression is  0.92
Bulling precision: 0.1141552511415525
Bulling recall: 0.18427518427518427
Bulling F-measure: 0.14097744360902256
Non-Bulling precision: 0.8858447488584474
Non-Bulling recall: 0.9814502529510961
Non-Bulling F-measure: 0.9312
```

Figure 9: Results using Logistic Regression

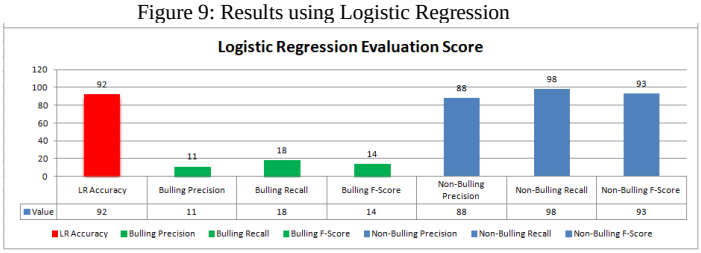

Figure 10: Results based on Logistic Regression

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

Using the above approach of the total of 364 copies of the non-bulling class, 34 were classified erroneously by Logistic Regression. For this class, the model obtained the values of 88 in precision and 98 in recall. these are very good results, very close to 1.00 (perfect performance), which indicates that for the class managed to find most of the copies and that what is labeled is of very good quality (level of certainty when labelling). Regarding its performance for the bulling class, of the total of 157 were incorrectly labeled. This allows reaching the assertion that the bulling class performed worse than the non-bulling for all the models used, for which 2 possible causes are postulated: the first, which is the class with the least representation in the training data, so the model has less data to learn from; the second, that at be two classes represented in one (bulling and non-bulling), there may be difficult the task of associating the features to the class, because it could be given the relationship of more features in fewer samples. Furthermore, consider that some features could apply to both the bulling and non-bulling class, which that usually generates difficulties in the field of machine learning.

4.5 RESULTS USING DECISION TREE

Decision Tree technique steps in the form of sentence with a limited number of words, but structured logically and systematic using UniGram, BiGram, TriGram and N-Gram. Apart from that the algorithm is a clear procedure to solve a problem using steps and is limited in number.

1. Algorithms have a beginning and an end, an algorithm must quit after working on a series of tasks. With words have finite steps using N-Gram.
2. Each step must be precisely defined so that it is not has a double meaning, not confusing with the help of UniGram, BiGram, TriGram and N-Gram.
3. Having input (input) or initial conditions.
4. Has an output (output) or final condition. The algorithm must be effective, if it is followed absolutely it will be solve the problem of Bulling and Non-Bulling Data.

```
0.711
bullying precision: 0.8010204081632653
bullying recall: 0.3857493857493858
bullying F-measure: 0.5207296849087895
non-bullying precision: 0.6890547263681592
non-bullying recall: 0.9342327150084317
non-bullying F-measure: 0.7931281317108089
```

Figure 11: Results using Decision Tree

The objective of this task is to identify aspects that generate results within the same context, selecting those characteristics that are frequently mentioned and the associated bulling or non-bulling scenario. To identify these characteristics or aspects, the simplest solution is to use a classification with one independent variable with allows us to obtain the nominal phrases of the tweets using decision tree and taking into account that most of the aspects are substantive as mentioned in result below.

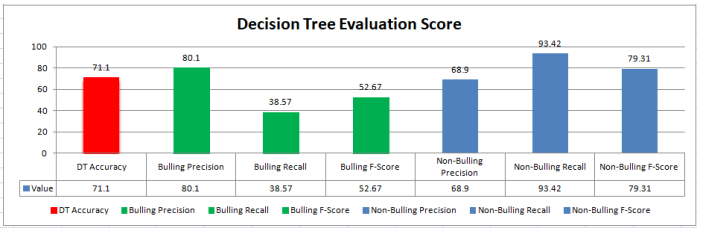

Figure 12: Results based on Decision Tree with Evaluation Models

Regarding the bulling class, of the 353 truly negative specimens, 41 were misclassified. The precision worth 80.1 is a good sign, considering that it is closer to 1.00 than to 0.5, indicating that it is unlikely to identify a positive negative sample. A very similar case is presented for the recall who got the same value of 38.57, indicating that there is the same low probability of identify a negative sample as positive as to identify a positive as negative. According to what was obtained for the positive class, of the 168 samples, 73 were misclassified. As expected, according to analyzed for the negative class, this model also delivered very similar to recall and precision for the positive class, taking the values 68.9 and 93.42 respectively, so the conclusions between the ratios of both classes is maintained as expressed in the analysis on the class negative of this model. Consequently, 71.1% of accuracy is achieved using decision tree.

4.6 RESULTS USING SUPPORT VECTOR MACHINE

This approach proposes the Support Vector Machine method for the classification process on the cyber bulling sentiment review. The result dataset from initial data processing that has been applied to three types of N-Gram is then classified using the proposed method with processed and labeled tweets and tests Model validation allows you to cross validate the different amounts of data using N-Gram to determine accuracy with precision and recall, below is results achieved using Support Vector Machine.

```
Accuracy using SVM  0.949
Bulling precision: 0.9013761467889908
Bulling recall: 0.9656019656019657
Bulling F-measure: 0.9323843416370107
Non-Bulling precision: 0.9429037520391517
Non-Bulling recall: 0.9747048903878583
Non-Bulling F-measure: 0.9585406301824211
```

Figure 13: Results using Support Vector Machine

The condition imposed on our function will be replaced by the conditions of the Lagrange multipliers, which will be easier to handle. With which, the saddle point of the following Lagrangian function must be found. Where $(wx_1 + b = +1) - (wx_2 + b = -1) \Rightarrow w(H_1 - H_2)$ is the class associated with the point x_i. By imposing this restriction, the SVM will place all the points that have the class associated $y_i = 1$ above hyperplane H_{11} and to the points that have the class associated $y_i = -1$ will place them below hyperplane H_{12}. Consequently, the results are elaborated as under.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

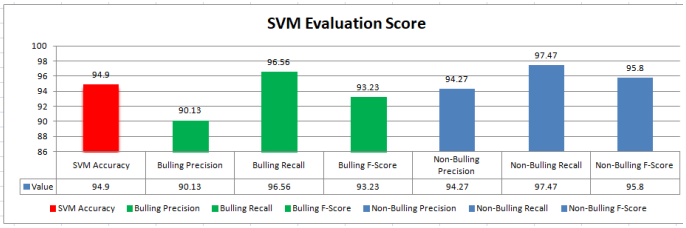

Figure 14: Results based on SVM with Evaluation Models

Regarding the performance over the bulling class, SVM mislabeled 38 samples of the total of 317 truly negative specimens. With a precision of 90.13 and a recall of 96.56 this model indicates that, for the class bulling, has greater certainty than uncertainty at the time of labeling. On the other hand, for the positive class, of the 166 truly bulling samples, 26 were wrongly labeled. Although the model had worse performance for the bulling class than the non-bulling one, it is possible to appreciate that there is a correlation between the performances of the different classes, since, being dichotomous classes, it makes sense that the strengths and weaknesses to identify one class are reflected in the strengths and weaknesses of its opposite class. Applied to this specific case (model and class), it is verified that if in the class negative the precision> recall, in its opposite class the opposite occurs (recall> precision) under the SVM. Consequently, 94.9% of accuracy is achieved using SVM.

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

Based on the results of research conducted that:-
When classical pre-processing is done and using a NLP techniques like CBOW and N-Gram, results are obtained in accordance with the literature, sufficing only with taking unigrams despite the fact that they are few messages coupled with the vague writing style, In addition, the baseline has been exceeded. Initially it was thought that by including stop words to help bigram and trigrams would have a better performance in the classifiers, but it has been verified that the n-gram is more effective approach with cbow. Based on experimentation, it was found that by not including the two characteristics pre-processing (eliminate stop words and do stemming) the results may come out with a lower value.

The bulling and non-bulling messages or tweets captured by the tweet extractor correspond to a common language and colloquial, so these messages contain quite a few misspellings, terms coined on the fly and erroneous grammar. This makes it difficult to model language and have common characteristics in different messages that serve as help the classifier. There were also small annotated corpora compared to others given in tweets and this influenced the learning of the classifiers. Other messages required more contexts because the extractor might return a message from a group that belonged therefore this lack of information further complicated the fact of being able to classify a message. Subsequently, the different models of representation of the words, it was possible to appreciate that the best is that of grams, and the best classifier are that of Naïve Bayes, Decision Tree and Support Vector Machine. When the labeled tweets for bulling and non-bulling were analyzed and based on the results it is possible to say that if the fact of having these tweets in a messages will improve or worsen the results without the cbow and gram model. So these results indicate that if it is preferable to use grams and cbow and add them as features to the training vectors of a classifier.

The justification for this is that using only symbols we have a value higher than the baseline. It can also be seen that the top line for these symbols is around 90% using unigrams/bigrams/trigrams and n-grams and as value of the cbow features. Finally in this particular analysis of tweets we have that the best values are generated when the model is of trigrams, this reinforces the idea that the more bulling tweets together, the more intensifies the gravity of the message (bulling or non-bulling) by lemmatization, stemming and applying the same models, as well as the same sorter.

Finally, when the combination of classifiers was analyzed, it was found that the Naive Bayes with the accuracy of 94.4%, Logistic Regression with the accuracy of 92.0%, Support Vector Machine with accuracy of 94.9% classifiers presented a excellent results when it was used with combining the cbow and n-gram collectively, whereas the decision tree is with 71.1% accuracy acting as the average classifier used. Therefore, the combination of cbow and grams with classifiers helps when choosing a method of weighting on the best classifiers for detecting bulling and non-bulling under the cyberbulling model.

At the end of the development of this thesis it was possible to see that it is possible to detect cyber-bulling automatically with few messages using robust techniques. Despite some limitations, based on the results obtained, we consider that the objectives of studying different processing techniques of the natural language, as well as apply machine learning techniques to meet the task of detecting cyber-bulling in the messages of the social network Twitter.

5.2 Future Work

The following are some suggestions that can be used in developing better future research.
1. Attempt to balance the proportion of data used between each class so that it is balanced and can affect the increase in the value of accuracy.
2. Voluminous Dataset can be integrated with Big-Data
3. Perform other pre-processing steps such as Map-Reduce algorithm
4. Experiments in changing the proportion of use of training data and test data along with Deep Learning Models.

REFERENCES

[1] Kapoor, Kawal & Tamilmani, Kuttimani & Rana, Nripendra & Patil, Pushp & Dwivedi, Yogesh & Nerur, Sridhar. (2018). Advances in Social Media Research: Past, Present and Future. Information Systems Frontiers. 20. 10.1007/s10796-017-9810-y.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 09, September 2022

[2] Myddleton, Johanna & Fullwood, Chris. (2016). Social Media Impact on Organisations. 10.1057/9781137517036_13.
[3] Anstead, Nick. (2015). Social Media in Politics. 10.1002/9781118767771.wbiedcs050.
[4] Dell'Anno, Roberto & Rayna, Thierry & Solomon, Offiong. (2015). Impact of social media on economic growth – evidence from social media. Applied Economics Letters. 23. 1-4. 10.1080/13504851.2015.1095992.
[5] Zhao, Lanlin. (2022). Effect of Social Media on Marketing. 10.2991/assehr.k.220110.118.
[6] https://www.statista.com/aboutus/our-research-commitment/2341/tanushree-basuroy
[7] Nixon C. L. (2014). Current perspectives: the impact of cyberbullying on adolescent health. Adolescent health, medicine and therapeutics, 5, 143–158. https://doi.org/10.2147/AHMT.S36456
[8] Nandhini, B. & Immanuvelrajkumar, Sheeba. (2015). Online Social Network Bullying Detection Using Intelligence Techniques. Procedia Computer Science. 45. 485-492. 10.1016/j.procs.2015.03.085.
[9] Sarka, Dejan. (2021). Data Mining. 10.1007/978-1-4842-7173-5_7.
[10] Thun, Lee Jia & Teh, Phoey & Cheng, Chi-Bin. (2021). CyberAid: Are your children safe from cyberbullying?. Journal of King Saud University - Computer and Information Sciences. 10.1016/j.jksuci.2021.03.001.
[11] Hinduja, Sameer & Patchin, Justin. (2010). Bullying, Cyberbullying, and Suicide. Archives of suicide research : official journal of the International Academy for Suicide Research. 14. 206-21. 10.1080/13811118.2010.494133.
[12] Marzano, Gilberto. (2022). Cyberbullying and Social Networking Sites. 10.4018/978-1-6684-5594-4.ch054.
[13] Sathyanarayana Rao, T S et al. "Cyberbullying: A virtual offense with real consequences." Indian journal of psychiatry vol. 60,1 (2018): 3-5. doi:10.4103/psychiatry.IndianJPsychiatry_147_18.
[14] Park, Ji & Fung, Pascale. (2017). One-step and Two-step Classification for Abusive Language Detection on Twitter. 41-45. 10.18653/v1/W17-3006.
[15] Samal, Biswaranjan & Panda, Mrutyunjaya & Behera, Anil. (2017). Performance Analysis of Supervised Machine Learning Techniques for Sentiment Analysis. 10.1109/SSPS.2017.8071579.
[16] Prabhat, Anjuman & Khullar, Vikas. (2017). Sentiment classification on big data using Naïve bayes and logistic regression. 1-5. 10.1109/ICCCI.2017.8117734.
[17] O. Aborisade and M. Anwar, "Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers," 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 269-276, doi: 10.1109/IRI.2018.00049.
[18] Beniwal, Sunita & Arora, Jitender. (2012). Classification and feature selection techniques in data mining. International Journal of Engineering Research and Technology. 1.
[19] Zhang, Jingwei & Liu, Tongliang & Tao, Dacheng. (2018). On the Rates of Convergence From Surrogate Risk Minimizers to the Bayes Optimal Classifier. IEEE transactions on neural networks and learning systems. PP. 10.1109/TNNLS.2021.3071370.
[20] Garcia, Guilherme. (2021). Logistic Regression. 10.4324/9781003032243-10.
[21] Roback, Paul & Legler, Julie. (2021). Logistic Regression. 10.1201/9780429066665-6.
[22] Zhou, Hong. (2020). Logistic Regression. 10.1007/978-1-4842-5982-5_6.
[23] Caraffini, Fabio. (2019). The Naive Bayes learning algorithm. 10.13140/RG.2.2.18248.37120.
[24] Berrar, Daniel. (2018). Bayes' Theorem and Naive Bayes Classifier. 10.1016/B978-0-12-809633-8.20473-1.
[25] Yang, Feng-Jen. (2018). An Implementation of Naive Bayes Classifier. 301-306. 10.1109/CSCI46756.2018.00065.
[26] Hasija, Yasha & Chakraborty, Rajkumar. (2021). Support Vector Machines. 10.1201/9781003090113-12-12.
[27] Jo, Taeho. (2021). Support Vector Machine. 10.1007/978-3-030-65900-4_8.
[28] Coqueret, Guillaume & Guida, Tony. (2020). Support vector machines. 10.1201/9781003034858-10.
[29] Jo, Taeho. (2021). Decision Tree. 10.1007/978-3-030-65900-4_7.
[30] Zhou, Hong. (2020). Decision Trees. 10.1007/978-1-4842-5982-5_9.
[31] Boehmke, Brad & Greenwell, Brandon. (2019). Decision Trees. 10.1201/9780367816377-9.
[32] Shikhman, Vladimir & Müller, David. (2020). Decision Trees. 10.1007/978-3-662-62521-7_9.
[33] Adewumi, Sunday & Abdu, Haruna. (2020). Analysis of n-gram (0. 1-4. 10.1109/ICCIS49240.2020.9257642.
[34] Sabra, Susan & Sabeeh, Vian. (2020). A Comparative Study on N-gram and Skip-gram for Clinical Concepts Extraction.
[35] Sidorov, Grigori. (2019). Generalized n-grams. 10.1007/978-3-030-14771-6_15.