International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Twitter Sentiment Analysis with Diabetic Drugs Using Machine Learning Techniques with Glowworm Swarm Optimization Algorithm

S. Radha Priya
Department of Computer Science
Government Arts College(A)
Coimbatore, India

Dr. M. Devapriya
Department of Computer Science
Government Arts College(A)
Coimbatore, India

Abstract— **Twitter is a Social Media that distributes opinions and sentiments of the people in day today life to friends and general public. Many people use twitter to communicate their side effects/benefits of diabetic medicines. Other people in turn seek these posts to gain feedback regarding their own Adverse Drug Reactions(ADR). Opinion mining of twitter data is an area that has experienced enormous growth in the last decade. For this purpose various Machine Learning(ML) Techniques and tools have been created. In this paper ML techniques were used in opinion analysis for processing information about ADR on taking diabetic drugs-Metformin(generic and brand name). The aim of this paper is to identify the optimal ML algorithm. Glowworm Swarm Optimization(GSO) is used to derive the optimal feature selection and is combined with various classification algorithms namely Naïve Bayes(NB), K-Nearest Neighbor(KNN) and Support Vector Machine(SVM). The Experimental result shows GSO+SVM combination proved maximum accuracy of 94%.**

***Keywords**— Twitter data, metformin, GSO, Naïve Bayes, KNN, SVM,ADR*

I. INTRODUCTION

Diabetes Mellitus is a metabolic disease in which the person's blood sugar levels are increased. There are two types of Diabetes Mellitus(DM) type I and II. Type I DM afflicts children and adolescents and requires insulin therapy. Type II DM occurs in older age population and requires oral anti diabetic drugs or insulin and Oral Anti Diabetic drugs(OAD) in combination. Each OAD has its benefits as well as adverse drug reactions. Hence it is essential for anti diabetic drugs to undergo pharmacovigilance in detecting ADR and there by helping physicians in preventing avoidable harm to diabetic patients[20]. Metformin(Generic and Branded) is a commonly used drug for type II DM. This study focuses on opinion of patients taking metformin regarding the ADR attained from twitter messages.

Twitter is a social media which is used by people to communicate about their health concerns and share their experiences. The number of messages shared in twitter is massive in view of the drug benefits and ADR making it an ideal resource for pharmocovigilance and early intervention[1,2]. Intelligent systems need to be created which will be accessible to patients to become aware of ADR of DM drugs from the Patients review. Fig 1 depicts the Architecture of the work flow.

In this article Section II deals with Related Works in Classification and Feature Selection Methods. Section III deals with Data Collection, Preprocessing, Feature Extraction, Feature Selection methods and ML classification algorithms. Section IV deals with data set description, Performance Evaluation and Comparative analysis of ML algorithms. Section V deals with the Conclusion.

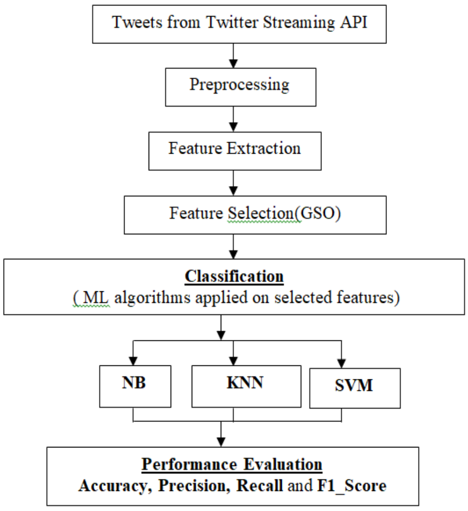
Tweets from Twitter Streaming API
Preprocessing
Feature Extraction
Feature Selection(GSO)
Classification
(ML algorithms applied on selected features)
NB
KNN
SVM
Performance Evaluation
Accuracy, Precision, Recall and F1_Score

Fig1. System Architecture

II. RELATED WORKS

While a lot of research has been completed to analyse tweet sentiments by developing techniques of ML, some of the works are described in this section. Movie reviews are taken from twitter message and are analysed using different machine learning algorithms. They are differentiated into Positive, neutral and negative. Emojis and symbols which express sentiments are excluded in this study. Like wise words with repetition of letters to express sentiments are also ignored in this study[Tripathy et al][2]. Sentiment Analysis of Arabic micro blogs has been done using deep learning systems-LSTM and GRU. Forward and backward direction has been tested with LSTM and GRU[Moslmi et al][3]. Various feature selection methods and classification algorithms have been used for Malay sentiment classification. In this SVM classifier with information gain base feature selection method proved best performance with accuracy of 85.33%[Azani et al][4]. Hager et al[21] study deals with prediction of heart disease from real time medical data regarding the people's present health status. Various machine learning algorithms are used to identify the optimal predictor. Random forest classifier proved to have the best performance in predicting heart disease with an accuracy of 94.9%, thus helping the patients in preventing cardiac catastrophe.

TABLE I . EXPLOIT OF EXISTING SENTIMENT ANALYSIS

Author & Year	Preprocessing Steps	Dataset	ML method	Accuracy in %
Haddi, E et al (2013) [7]	Expanding Acronym, On line text cleaning, stemming, removal of white space, stop words and negation handling	Movie review 1-1400 documents Movie review 2-2000 documents	TF-IDF FF FP	93.5 90.5 93
Bao, Y .et al (2014) [8]	IRL features reservation, negation transformation, normalization of repeated letters, stemming and lemmatization	Standford twitter sentiment	LibLinear	85.5
Dos Santos, F.L et al (2014) [9]	Terms standardization, stemming, spell check	Brazilian Portuguese about android apps	SVM-stage 3 Logistic Regression-Stage 3	81.08 81.29
Dwi Aji Kurniawan et al (2016) [10]	RT removal, Case converting, Website address removal, Twitter username removal, removing characters non-alphanumeric characters and Changing abbreviations to their actual phrases.	Twitter API (Real-time traffic data)	NB SVM DT	98.02 98.31 98.41
María del Pilar Salas-Zárate et al (2017) [12]	Normalization, Tokenization, Sentence formation, Assigning lexical category to each word, mapping words to their base form.	Twitter API (Diabetes)	Aspect-level sentiment classification	81.93

III. Materials And Methodology

A. Data Collection

Twitter dataset about metformin and related branded medicine were used in this research to build classification model. Twitter offers streaming Application Programming Interface (API) to permit the users to gather real time data. This is a tool which creates the communication among computer programs and web services easy. Several tools such as python, JavaScript, and R-tool services are developed to relate with twitter network also to access data in efficient way. Here, R-Tool utilized to search for tweets posted recently by users to extract real time tweets. Before collecting tweets, the drug names essential to be defined as the 'keywords' in tracking and gathering data is stored into the .csv file. The below figure illustrate the process of data collection.

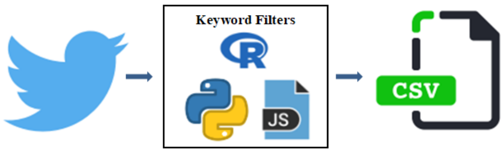

Fig2. Dataset Retrieval process

Getting Twitter API keys (API key, API secret, Access token and Access token secret) are important steps to access the Twitter Streaming API. Library files called 'Tweepy' is utilized to assess Twitter Streaming API and retrieving the twitter data. Primarily, drugs should have been on the market to treat substantial diseases, so that adequate tweets would occur for calculating their impacts. The tweets data set was retrieved with drug keyword also the tweets gathered from across the world [13].

B. Pre-Processing

Applying text preprocessing steps before analyzing the tweets is very important for achieving good results. There are several steps involved in the preprocessing stage such as URL Removal, Punctuation Removal, User name removal, Letter casing, Tokenizing, Stop word removal, Stemming and Lemmatization, to make a standard dataset [14]. Once the steps are completed, this research moves to the next main method called feature extraction. Extraction of valuable words from the tweet is called as feature extraction.

TABLE II. EXAMPLES OF PREPROCESSED DATA

Original	Pre-processed
Severe hypoglycemia(Low blood sugar) can even cause seizures, comas and hypothermia!	['severe', 'hypoglycemia' 'low', 'blood', 'sugar', 'even', 'cause', 'seizure', 'coma', 'hypothermia']
RT @Endometriosis11: Other symptoms include low-grade fevers, heavy and/or irregular periods, and hypoglycaemia	['symptom', 'include', 'lowgrade', 'fever', 'heavy', 'irregular', 'period', 'hypoglycemia']
@jonessurgery Lactic acidosis/shortness of breath/hip pain/etc - Etiology unclear at this point, differential incluâ€¦ https://t.co/4JI2FaTQAk	['lactic', 'acidosis' 'short', 'breath' , 'hip', 'pain', 'etiology', 'unclear', 'point', 'differential', 'inclu']

The above table shows the few examples of preprocessed twitter data.

C. Feature Extraction

In the Feature Extraction model, the only few selected words are identified as features which have opinion (side effects) about the metformin by their presence in the dataset. Common side effects associated with metformin (declared by WHO) are taken in this work for feature extraction process. They are illustrated in the below figure,

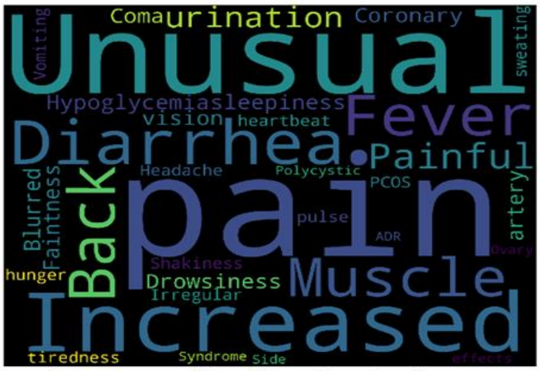

Fig3. Common Side effects of Metformin

For the twitter data analysis, twitter platform has been taken as the main source in which attributes such as the Favourite, Favourite count, Truncated, Re-tweet, isRetweet, Count of retweet were taken as the inputs. These attributes are the major constraints of the twitter platform which is normally taken for the study and it reflects the nature of message sharing logic. To identify the effects on the individuals, these attributes normally reflects the usage of medicine through its messages. Pseudo code used to represent the class label is as follows:

Pseudo code:

Consider the twitter messages such as

```
'X' medicine(Generic and Branded)  will have 'Y' side effects

If  message is favorite && favorite count =count+1&& Retweets= retweets +1
        (If Retweet count+1 && isRetweet ==true )
         Then number of retweets represents the affected people
                and they will be    experiencing same side effects

    Else if
       Number of retweets represents the affected people is less  and they will
                         Be experiencing same  side effects
    Else
        Not affected
```

Favourite represents the person who has more side effects or less side effects on the particular medicine will be liked by other persons. It shows the resemblances of like-minded people. Retweet is the repetition of another user's tweet. It Confirms the same side effects experienced by the people or nil side effects for the medicine.

The label is numbered as '1' if Favourite and Retweet count and Favourite count are equal to or greater than one else it is numbered as '0'. In this way, python codes has been programmed to extract the values based on the medicinal messages which has been posted in the twitter. These attributes represents the impact of side effects on the particular individuals through their sharing and marking as the favorites. Hence the feature engineering adopted deals with the count values, retweets mechanism which are marked as the different identifier so that it can be used for successful identification and classification of side effects.

D. Feature Selection

Feature selection technique is used in sentiment analysis that has a significant role for identifying relevant features and increasing classification (machine learning) accuracy. Glowworm Swarm Optimization is a Feature Selection technique which changes the biology system of glowworm luminescence into Arithmetic representation. The theory of GSO is: 'n' number of glowworms will be available in the search space. Initially each glowworm will have a specific Luciferin intensity value. During the movement in the search space the Luciferin intensity value changes. The glowworm with high intensity attract the glowworm with lower intensity. Glowworm will be grouped together when they fall within the circular range. Other glowworms which falls outside the perception range will be omitted. The rule of glowworm search method appears in Figure 1, where the three glowworm entities are 'a', 'b', and 'c'. Their inquiry radii are 'r_a', 'r_b'and 'r_c'. The search radius of 'a' is bigger than of 'b', and 'b' situates inside the pursuit scope of 'a'. In the event that the luminosity of 'a' is more grounded than that of 'b', the last will move towards the previous. Since 'c' is not inside the pursuit scope of 'a', neither of them will move towards one another paying little mind to whose brightness level is stronger [15].

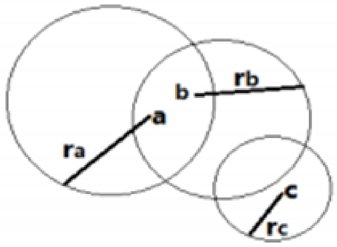

Fig4. The searching rule plan of 'GSO'

As stated by bionic principle, the two major characteristics of GSO are luminance and attractive degree. The below formulas describe the glowworm luminescence properties. Relative fluorescence luminance is denoted as

$$I = I_0 \times e^{-\lambda\gamma_{ij}} \quad (1)$$

Where, 'I_0' is the luminosity of glowworm at the position of γ =0. Higher luminance can produce best value of target function. 'λ' is a constant that denotes 'light intensity absorption' coefficient, which depicts the reduction degree of fluorescence vigour. 'γ_{ij}' is the distance among glowworm 'i' and 'j'. The attractive degree is meant as

$$\beta = \beta_0 \times e^{-\lambda \gamma^2{}_{ij}} \tag{2}$$

Where, ‘β_0‘ is the attractive degree at the point of ‘r=0’, specifically the extreme attractive degree. The below formula is used to calculate the location update (X_i).

$$X_i = X_i + \beta \times (X_j - X_i) + s \times (rand - 1/2) \tag{3}$$

From the above equation, ’s’ denotes step size factor that set as a constant. The value interval is [0, 1]. ‘x_i‘ and ‘x_j‘ are spatial positions of glowworm ‘i’ and ‘j’. ‘rand’ denotes random element and its interval value is [0, 1].

E. Machine Learning Algorithms for Classification

ML play a vital role in opinion classification. This section will discuss the some ML classification algorithms such as NB, KNN and SVM and demonstrates all classification algorithm’s characteristics and working methodology.

Naïve Bayes (NB) Algorithm

NB is a robust Machine Learning classifier, which is utilized for classification process. This method is sustained from Bayes theorem where foundational theory of ‘NB’ classifier is constructed on the independence theory. Naïve Bayes classifier presumes that the outcome of a particular attribute in a class is independent of other attributes[16]. The predictions in the bayes theorem is evaluated using the following formula.

$$P\left({}^{h}/_{D}\right) = \frac{P({}^{D}/_{h})P(h)}{P(D)} \tag{4}$$

P(h) is the probability of hypothesis h(Prior Probability) being true. P(D) is the probability of Data(Prior Probability) regardless of the hypothesis. P(h/D) is the Probability of hypothesis h(Posterior Probability) given the data D and P(D/h) is the probability of data D(Posterior Probability) given that hypothesis h was true. The classes of dataset can be easily predicted in given dataset by applying the NB classifier. Multi-class prediction is also possible. When the presumption of independence is well founded, NB is most capable than the other algorithms. Additionally, it will need less training data. If the absolute variable fits to a class that was not supervised in the training set, then the model will provide it a probability of ‘0’ that will prevent it from creating predictions. It utilises independence among its features. In actuality, it is hard to collect data that are entirely independent features.

K-Nearest Neighbor (KNN)

KNN algorithm takes a crucial part in machine learning process. It belongs to the supervised learning area and have numerous applications in intrusion detection, pattern recognition, and so on. These KNNs are applied in real life situations where non-parametric methods are needed. This technique do not create any presumptions about data distribution. In the dataset, the KNN method classifies the coordinates into clusters which are recognized by a specific character. The single idea for this method is that it is similar output for similar training samples. For the input population nearest value is identified that is ready to assign classes to all or any of the samples. Consider $X_i = \{x_1, x_2, \ldots, x_{iN}\}$ and $X_j = \{x_1, x_2, \ldots, x_{jN}\}$ the sample population, thus to measure the similarity between them and the distance is calculated as given below.

$$\text{Dist}(X_i, X_j) = \sqrt{\sum_{m=1}^{N} (x_{im} - x_{jm})^2} \tag{5}$$

In the above equation, Euclidean distance is described that evaluates similarity among two pixel points. Hence, the pixels obtain the category to which a number of them commonly resemble [17]. KNN is an Instance based learner that means it does not learn anything in the training stage. This technique does not get any discriminational function from the training data. Especially, there is no training time for this technique. It learns the characteristics of features from training dataset at the time of predictions. The training time is reduced in the algorithm so it is more rapid than other algorithms e.g. SVM, Linear Regression and so on. The classifier does not require training samples prior to making predictions. So, new data can be incorporated easily that will not influence the accuracy of the system. It requires only two parameters to execute i.e. the value of ‘K’ and the distance function (e.g. Euclidean Distance, Hamming Distance, Manhattan Distance, Minkowski Distance). This Classifier provides good results for lesser number of input features but as the number of features increase to high levels it scuffles to predict the output of new data point. One of the main problem is to select the ideal number of neighbors to be considered at classifying the new data entry. Computation expense is quite high because it is essential to calculate distance of every point occurrence to all training samples.

Support Vector Machines (SVM)

SVM is a supervised learning model that is generated for binary classification in both linear and nonlinear forms. Usually, datasets are nonlinearly indivisible, thus the main goal of the SVM method is to capture the finest available surface to make a disconnection among positive and negative training feature samples depending on peril (training and test set error) reduction principle. This method can try to describe a decision boundary with the hyper-planes in a high dimensional feature space. This hyper plane delineates the vectorized data into two classes also finds an outcome to take a decision depending on this support vector. The working method of SVM can be described as follows. Given ‘N’ linearly separable training set with feature vector ‘x’ of ‘d’ dimension. For dual optimization, Where, $\alpha \in R^N$ and $y \in \{1, -1\}$. Then the outcome of SVMs can be described as follows:

$$\vec{\alpha}^* = argmin\left\{-\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j y_i\, y_j \langle \overrightarrow{x_i}\ , \overrightarrow{x_j} \rangle\right\} \quad (6)$$

Where,

$$\sum_{i=1}^{n} \alpha_i\, y_i = 0; \qquad 0 \le \alpha_i \le C \quad (7)$$

In SVM classification, it segregates the linear dataset with a single hyper plane that can divide two classes of given feature subset. For nonlinear dataset where more than two classes to be managed, kernel functions are utilized in that state to set the data to a higher dimensional expanse that is linearly separable [18]. The technique has a normalization parameter. Consequently, it has greater simplification abilities that prevent features from over fitting. It uses the kernel trick. So, it can effectively manage the nonlinear data. The small changes in the datasets do not significantly disturb the hyper plane. So this model is stable. Selecting a suitable Kernel function (to manage the nonlinear data) is not a simple process. For high dimension Kernel, it gives rise to many support vectors that minimize the training speed considerably. For large datasets, the classifier takes much time for training. It requires feature scaling for input variables before the classification process. Algorithmic difficulty as well as memory necessity are very high. It needs lot of memory for multi class SVM to save all support vectors also this number raises sharply with the training dataset size.

IV. EXPERIMENTS, RESULTS AND DISCUSSION

In the proposed system, twitter platform has been taken as the main source in which attributes such as the retweet, favorite, isfavorite, twitter messages where taken as the inputs. These attributes are the major constraints of the twitter platform which is normally taken for the study and it reflects the nature of message sharing logic. To identify the side effects of specified drugs (metformin generic and branded) on the individuals, these attributes normally reflects the usage of medicine through its messages.

Meaning of some twitter messages which contain Generic or branded anti diabetic drugs were actually different from that of the perspective of the research. Some twitter messages showed no useful relationship between the drug name, their benefits and side effects. For example *"I need metformin"* and *"@yungliu i need to buy a white metformin but the site won't let me ship to my place?"* .Hence unrelated twitter messages were filtered out from the dataset.

Feature extraction process encounter the occurrences of all the specified words (Side effects of metformin) which is used for extraction process. The words and their occurrences are used for labelling the dataset to train the classifier. The proposed technique has been examined with the datasets after pre-processing which are retrieved from the twitter websites. Totally it has 14 attributes, but 6 attributes are more useful for the determination of side effect level of the diabetic (metformin generic and branded) drugs.

In this research, Glowworm Swarm Optimization is used for feature selection process to select significant feature from the extracted feature dataset. This optimizer is utilized to improving the classification process. These selected features can be used in machine learning algorithms for classification process.

Evaluation parameters

In this research, the execution of ML algorithms can be assessed with the components of confusion matrix on a set of testing data. The confusion matrix contains of four elements are True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FP). The evaluation metrics precision, recall, F1_score and G_mean [19] are calculated to estimate the performance level of any classifier.

Accuracy value is the proportion of the accurate number of predictions. It can be determined using the below equation:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (8)$$

Precision is the ratio of predicted positive examples which really are positive

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (9)$$

Recall also called hit rate or sensitivity; it measures how much a classifier can recognize positive examples

$$Recall = \frac{TP}{(TP + FN)} \quad (10)$$

F1_Score is the 'Harmonic Mean' of recall with precision

$$\text{F1\_Score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (11)$$

True Positive rate is a measure of sensitivity of true positive prediction

$$TPR = \frac{TP}{TP + FN} \quad (12)$$

Specificity is a measure of accuracy of True Negative Prediction

$$Specificity = \frac{TN}{TN + FP} \quad (13)$$

False Positive Rate is calculated using

$$FPR = 1 - Specificity \quad (14)$$

To describe the performance of the classification algorithms, confusion matrix has been used in this research. It permits envisioning of the execution of an algorithm. It also permits easy recognition of uncertainty between classes. The

Fig.5 depicts the Confusion Matrix. True Positives(TP) are Side Effects which are correctly identified as Side Effects. True Negatives(TN) are No Side Effects which are correctly identified as No Side Effects. False Positives(FP) are No Side Effects which are incorrectly identified as Side Effects. False Negatives (FN) are Side Effects which are incorrectly identified as No Side Effects

True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

Fig5. General form of Confusion Matrix

Evaluation of accuracy:

The performance of classification algorithms has been evaluated with Accuracy, Precision, Recall and F1_Score performance evaluation parameters. Following table shows the accuracy obtained at various ML methods for classification process.

TABLE III. COMPARISON OF ML ALGORITHMS USING ACCURACY

Algorithms	Accuracy %
GSO+NB	91%
GSO+KNN	92%
GSO+SVM	94%

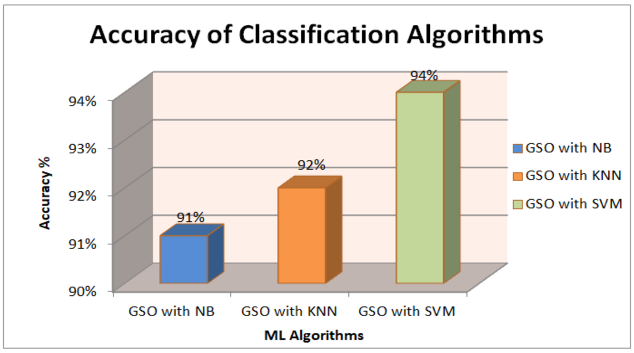

Fig 6. Accuracy of Machine Learning Algorithms

From the above figure, the SVM algorithm has showed highest accuracy rate.

TABLE IV . PERFORMANCE EVALUATION

	GSO+NB	GSO+KNN	GSO+SVM
Precision	0.89	0.96	1.00
Recall	0.85	0.90	1.00
F1-Score	0.87	0. 93	0. 96
TPR	0.94	0.96	1.00
Specificity	0.85	0.90	0.85
FPR	0.15	0.10	0.15

The above results illustrates that the SVM technique has best performance when compared to other ML algorithms.

Actual class	**Predicted class**		
		Yes (1)	No (0)
	Yes (1)	**33**	**0**
	No (0)	**3**	**17**

Fig 7 . Confusion Matrix for GSO+SVM

Among the 176 Twitter messages, 30% (53) has been taken as testing data and the 70% (123) has been taken as Training data. With the proposed algorithm for SVM with GSO while the classification has been applied on the Testing data, among the 53 Twitter messages, 33 has been classified under Side Effects. Here, misclassification (i.e.) false negative value is 0. 17 cases has been classified under No Side Effects correctly. The remaining 3 (misclassification) cases may be 'No Side Effects', but could not be predicted accurately as No Side Effects. Overall, 94% of the predictions are correct remaining 6% could not be predicted accurately (Misclassified), is shown in Fig.7.

Actual class	**Predicted class**		
		Yes (1)	No (0)
	Yes (1)	**23**	**1**
	No (0)	**3**	**26**

Fig8. Confusion Matrix GSO+KNN

In the KNN with GSO algorithm while the classification has been applied on the Testing data, among the 53 Twitter messages, 23 has been classified under Side Effects. The remaining 1 case may be side effect but could not be predicted accurately as side effect. Here, misclassification (i.e.) false negative value is 1. 26 cases has been classified under No Side Effects correctly. The remaining 3 (misclassification) cases may be No Side Effects, but could not be predicted accurately as No Side Effects. Overall, 92% of the predictions are correct remaining 8% could not be predicted accurately (Misclassified), is shown in Fig.8.

Actual class	**Predicted class**		
		Yes (1)	No (0)
	Yes (1)	**31**	**2**
	No (0)	**3**	**17**

Fig9. Confusion Matrix GSO+NB

In the NB with GSO algorithm while the classification has been applied on the Testing data, among the 53 Twitter messages, 31 has been classified under Side Effects. The remaining 2 cases may be side effects but could not be predicted accurately as side effect. Here, misclassification (i.e.) false negative value is 2. 17 cases has been classified under No Side Effects correctly. The remaining 3 (misclassification) cases may be 'No Side Effects', but could not be predicted accurately as No Side Effects is shown in Fig.9. Overall, 91% of the predictions are correct remaining 9% could not be predicted accurately (Misclassified).

V. CONCLUSION

The purpose of this research paper is to survey the potential of three ML algorithms such as NB, SVM and KNN to classify the side effects level of the antidiabetic drug 'Metformin' Generic and branded through twitter messages. In this research, GSO method is combined with the ML algorithm to select the optimal features for classification process. The efficiency of classification method is assessed in terms of Accuracy, F1_score, precision, and recall. These exploratory results show that SVM classifier is highly effectual and encouraging. The purpose of this research is

Pharmocovigilance. SVM with GSO shows 94% accuracy in prediction. TP cases are 33 and TN is 17. These 17 patients can safely continue metformin. For 33 TP patients severity of the side effect is to be accessed at the physicians clinic. If side effect is mild same medication can be continued with regular follow up. For patients with FP, reassurance can be given. Patients with FN need to attend physicians clinic for assessment. This work still has some drawbacks like SVM takes a prolonged training time for our twitter dataset, memory specification of SVM are lofty and it requires feature scaling for input variables before the classification process. All these obstacles need to be regarded for the future upcoming task to improve the twitter opinion classification.

REFERENCES

[1] Statista. (2019) Number of social media users worldwide 2010-2021. Available from: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users.

[2] Ikoro, Victoria, Maria Sharmina, Khaleel Malik, and Riza Batista-Navarro. (2018) “Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers”, in Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 95-98): IEEE.

[3] Tripathy, A. Agrawal and K. Rath, “ Classification of Sentiment reviews using n-gram machine learning approach” , Expert systems with Applications, Vol 57, PP. 117-126, 2016.

[4] T. Moslmi, S. Gaber, A. Al-Shabi, M. Albared, and N. Omar, “Feature selection methods effects on machine learning approaches in malay sentiment analysis,” in Proceedings of the 1st ICRIL International Conference on Innovation in Science and Technology (lICIST '15), pp. 444–447, 2015.

[5] S AI-Azani and ES EI-Alfy, “Emojis-based sentiment classification of Arabic microblogs using deep recurrent neural networks,” in Proceedings of the 2018 International Conference on Computing Sciences and Engineering (ICCSE), pp. 1–6, IEEE, Kuwait City, Kuwait, March 2018.

[6] K. M. Alomari, H. M. ElSherif, and K. Shaalan, “Arabic tweets sentimental analysis using machine learning,” in Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 602–610, Springer, Arras, France, October 2017,Springer.

[7] Haddi, E., Liu, X., Shi, Y.: The role of text pre-processing in sentiment analysis. Procedia Comput. Sci. 17, 26–32 (2013),Elsevier.

[8] Bao, Y., Quan, C., Wang, L., Ren, F.: The role of pre-processing in twitter sentiment analysis. International Conference on Intelligent Computing,pp 615-624,2014, Springer.

[9] Dos Santos, F.L., Ladeira, M.: “The role of text pre-processing in opinion mining on a social media language dataset.” In: Proceedings—2014 Brazilian Conference on Intelligent System, BRACIS 2014, pp. 50–54 (2014)

[10] Aji Kurniawan, Sunu Wibirama, Noor Akhmad Setiawan "Real-time Traffic Classification with Twitter Data Mining" 8th International Conference on Information Technology and Electrical Engineering (ICITEE), IEEE, 2016

[11] Jianqiang, Z., Xiaolin, G.: “Comparison research on text pre-processing methods on twitter sentiment analysis”. IEEE Access. 5, 2870–2879 (2017)

[12] María del Pilar Salas-Zárate, José Medina-Moreira, Katty Lagos-Ortiz, Harry Luna-Aveiga, Miguel Ángel Rodríguez-García, Rafael Valencia-García "Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach" Hindawi Computational and Mathematical Methods in Medicine Volume 2017.

[13] Rahman, S. A. El, F. A. AlOtaibi, and W. A. AlShehri. (2019, 3-4 April 2019). “Sentiment Analysis of Twitter Data”, in the 2019 International Conference on Computer and Information Sciences (ICCIS).

[14] Neetu Anand, Dhruvi Goyal and Tapas Kumar "Analyzing and Preprocessing the Twitter Data for Opinion Mining" Springer Nature Singapore Pte Ltd. 2018 B. Tiwari et al. (eds.), Proceedings of International Conference on Recent Advancement on Computer and Communication, Lecture Notes in Networks and Systems 34.

[15] Zhucheng Li and Xianglin Huang "Glowworm Swarm Optimization and It’s application to Blind Signal Separation" Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2016.

[16] V. Narayanan, I. Arora, and A. Bhatia, “Fast and accurate sentiment classification using an enhanced na¨ıve Bayes model,” in Intelligent Data Engineering and Automated Learning –IDEAL 2013, vol. 8206 of Lecture Notes in Computer Science, pp.194–201, Springer, Berlin, Heidelberg, Germany, 2013

[17] Najat Ali, Daniel Neagu, Paul Trundle "Evaluation of K-Nearest Neighbor Classifier performance for heterogeneous data sets" Springer, SN Applied Sciences, 2019.

[18] Huosong Xia, Yitai Yang, Xiaoting Pan, Zuopeng Zhang & Wuyue An "Sentiment analysis for online reviews using conditional random fields and support vector machines" Electronic Commerce Research, 789, Springer Science, Business Media, LLC, part of Springer Nature 2019.

[19] AlaaTharwat "Classification assessment methods" Applied Computing and Informatics, 2210-8327, Elsevier 2018.

[20] Tirthankar Deb, Abhik Chakrabarthy, Abhishek Ghosh “Adverse drug reactions in Type 2 diabetes mellitus patients on oral antidiabetic drugs in a diabetes outpatient department of a tertiary care teaching hospital in the Eastern India”, International Journal of Medical Science and Public Health Online 2016.

[21] Hager ahmed, Eman M.G. Younis, Abdeltawab Hendawi, Abdelmgid A. Ali ,’Heart disease identification from patients social posts, machine learning solution on spark’, Future Generation computer Systems, 2019.