Advances and Applications in Mathematical Sciences
Volume 21, Issue 1, November 2021, Pages 241-250
© 2021 Mili Publications, India

DIABETES TWITTER ANALYSIS USING IMPROVED ENSEMBLE MACHINE LEARNING TECHNIQUES

V. DIVIYA PRABHA[1] and R. RATHIPRIYA[2]

[1]Research Scholar, [2]Assistant Professor
Periyar University
Salem, India
E-mail: diviyaprabha7@gmail.com
rathipriyar@gmail.com

Abstract

Nowadays, sentimental analysis plays an important role in healthcare domain. Diabetes is one of the healthcare problem should be analyzed to understand people's sentiments. This paper represents a novel improved ensemble classifier (IEC) to analyze tweets in twitter. Nearly, 145713 tweets are collected to analyze the performance of the algorithm. Numerous machine learning techniques are applied to analyze the classification performance. This approach detects to classify tweets using proposed ensemble method. The effectiveness of the proposed algorithm is compared with state-of-the art approaches such as bagging, boosting and stacking.

I. Introduction

The diabetes diseases are growing from new born baby to old men and women. It is given important significance when comparing with other diseases such as cancer, asthma and chronic disease. The survey of recent reports results that adults of age over 18 improves the percentage of diabetes from 4 percentages to 8 percentages [3]. There exists different kinds of diabetes can lead to morality so it is needed to educate people about diabetic [4] awareness and initiate measures to improve human life. Diabetes drugs related post helps people to learn positive opinion and negative opinion about drugs. This suggest people best diabetic drug and reduce the risk of mortality. To support, diabetic patient and develop a positive thought which reduce the complex factors.

2020 Mathematics Subject Classification: 34Dxx, 93Dxx.
Keywords: Twitter, Machine Learning Techniques, Diabetes, Sentimental analysis.
Received June 2, 2020 Accepted January 18, 2021

All age group people share their opinion in Internet. People using social media [7] is improving during last few decades. Twitter [15] allows people to share their posts and make a social talk with their group and public. People can share their opinions in any topics such as politician for next elections, about sports, education and health care sectors. Twitter also allows extracting the tweets posted by users. Mining in health related data effective to support health care peoples [17] and improve the patient self-support. Furthermore, [5] studies conclude social media act as mediator to reduce the risk factors among patients. Social media offers an emotional support to share their opinions and lead a stress less life [6]. Due to this, health care issues are initiated aimed at opinion of people about diabetes using social media. This updated a real- time issues for the people about diabetes and its types to people. Traditional sentiment analysis confined on binary classification problem only. This paper aims to highlight the effectiveness of ensemble in machine learning and performance improvement than traditional approaches. The effectiveness of the algorithm is compared with other traditional ensemble approaches [18]. This paper is described as follow: Section 2 describes the review of related works, section 3 explains the methodology and finally section 4 illustrates the experimental results.

II. Related Work

Usage of social media is improving people's interest to share disease-related messages online. Survey suggest that small quantity of tweets during the short period of time is analyzed [1, 2] in twitter by health sectors to collect information about diabetes. The studies are unaware about the evolution of large-scale [16] conversation on twitter. Twitter [12] have analyzed content and profile about users with huge amount of diabetes related has tags. The supervised classifier in machine learning techniques such as Support Vector Classifier (SVC) and entropy proposes an approach and identify the factors about user. The practice of posting message [7] in twitter has great impact of all types of people. The survey [8] also states the usage of smart phone and internet is increased exponentially people sharing their opinion also increases. Mining twitter data [9] to analyze health issues about food. Sentiment analyses of twitter data suggest the physical and mental condition of people [10]. So, classification of this text is very important [11]. Machine learning techniques have been applied for sentimental analysis. Authors used [14] SVM, NB and RF for classification sentiments mobile review. In recent years, studies focused in ensemble approaches to improve the accuracy. This paper focused on improved ensemble approach to increase the effectiveness of classifier.

III. Methods

This section represents the different methods for the proposed IEC model. It consists of three phases such as data collection, data pre-processing and description about the proposed work.

3.1 Data Collection: Data is collected from twitter API using python coding. Searching keywords such as Diabetes as DB, Type 1 diabetes (TP1D), Type 2 diabetes (TP2D), Gestational Diabetes (GD), Young Diabetes (YD) are given as tags to search the queries. The data received in unstructured format is converted to structured format as user understandable format. The structured data are analyzed to determine the sentiments of users about diabetes.

Table 1. Dataset Description.

Dataset Name	Class	No of Tweets
DB	5	100000
TP1D	5	14157
TP2D	5	10259
YD	5	7886
GD	5	7600

3.2 Data Pre-Processing: The process of breaking tweets to tokens is a process of tokenization. Breaking the sentence into words is significantly necessary. White space is removed and each individual word is considered as tokens. Punctuations are removed in this process. Certain sentences also contain digits, comma, brackets and other symbols those symbols are removed. Remove stop-words, punctuations and other extra symbols from text. The following represents an example to represents the preprocessed tweets.

3.3 Classification Model: This section focuses to categorize the tweets using proposed ensemble model. It proposes by integrating numerous models based on the weight and threshold. This approach allows the creation of improved ensemble classifier (IEC) to improve the accuracy performance.

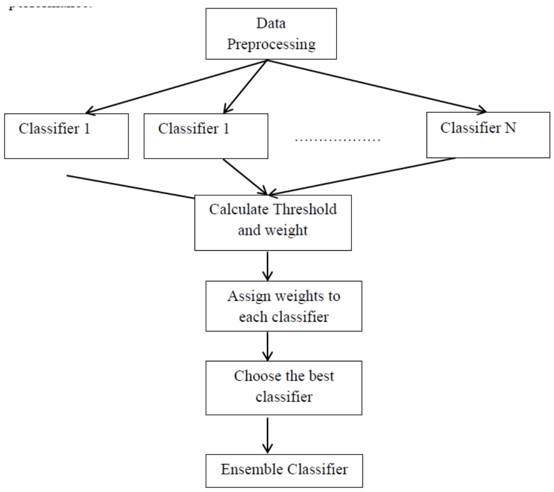


Figure 1. Proposed Ensemble Classification Model.

```
Algorithm: IEC

Input: Pre-processed tweets

Output: Tweets Sentiments

For each tweets=1 to M do
If p>=0.5 then
        Label=SP
Else if p<=0.5
         Label=P
Else if P<=-0.5
         Label =SN
Else if P>-0.5
         Label =N
Else
         Label =Neutral

Use $L$ to generate for different classifier Calculate weight
```

$$W_i = \sum_{i=1}^{n} Weight\,(C_i) * T$$

$$L = W$$

Use i to generate the classifier

$$B_i = \max(W_i)$$

Append (B_i) to ensemble

Evaluate the B_i results

Figure 1 describes the flow of the proposed diabetes tweets analyzing method using IEC. SP represents strong positive of tweets based on the polarity of tweets p. Similarly, label P represents positive sentiment of tweets, SN represents strong negative of tweets, N represents negative sentiment, SN represents strong negative sentiments. Weight W value of the classifier is calculated based on the threshold. Choosing the best classifier is identified by Bi. If the accuracy value is greater than 0.90 percentages the threshold (T) value is set to maximum of 0.5 similarly the T value is minimized based on the accuracy. W represents the weight multiplication of classifier threshold CA_T and probability of weight PB. Classifier which has the high weighted values is taken to ensemble classifier. These classifiers are appended to validate the test results.

IV. Experimental Analysis

This section represents the results of the extracted tweets using multiclass machine learning techniques such as Random Forest (RF), Logistic Regression (LR), K-Neighbor classifier (KNN), Support Vector Classifier (SVC). Classification of tweets represents the sentiments of tweets 0 for neutral, 1 for positive and 2 for strong positive, 3 for negative and 4 for strong negative. The overall weight is the multiplication of probability of weight and applying the weight based on threshold.

Table 2. Classification Results for DB tweets.

Multiclass Model	Accuracy	Probability weight	Weight based T	Overall weight	IEC
RFC	0.94	0.5	0.5	0.25	0.98
LRN	0.96	0.5	0.5	0.25	
KN	0.89	0.5	0.4	0.25	
SVM	0.92	0.5	0.5	0.25	
NB	0.90	0.4	0.5	0.20(E)	

Table 3. Classification Results for TP1D tweets.

Multiclass Model	Accuracy	Probability weight	Weight based T	Overall weight	IEC
RFC	0.94	0.2	0.5	0.1	0.96
LRN	0.92	0.2	0.5	0.1	
KN	0.83	0.1	0.5	0.05(E)	
SVM	0.95	0.2	0.5	0.1	
NB	0.88	0.2	0.4	0.08(E)	

Table 4. Classification Results for TP2D tweets.

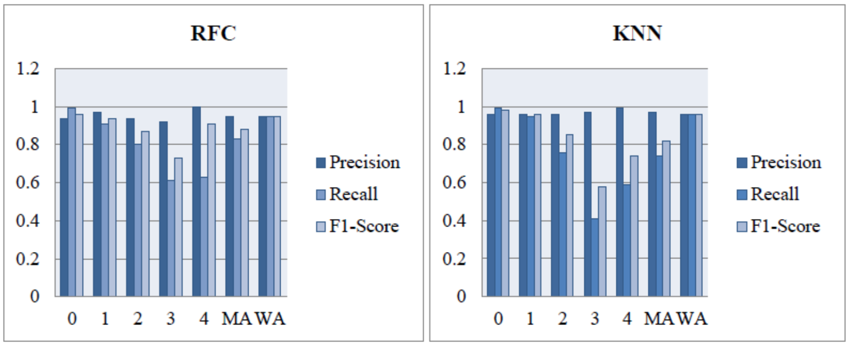
Multiclass Model	Accuracy	Probability weight	Weight based T	Overall weight	IEC
RFC	0.92	0.2	0.5	0.1	0.94
LR	0.90	0.2	0.5	0.1	
KNN	0.82	0.2	0.4	0.08(E)	
SVM	0.93	0.2	0.5	0.1	
NB	0.85	0.2	0.4	0.08(E)	

Table 5. Classification Results for YD tweets.

Multiclass Model	Accuracy	Probability weight	Weight based T	Overall weight	IEC
RFC	0.92	0.2	0.5	0.1	0.99
LR	0.97	0.2	0.5	0.1	
KNN	0.95	0.2	0.5	0.1	
SVM	0.95	0.2	0.5	0.1	
NB	0.96	0.2	0.5	0.1	

Table 6. Classification Results for GD tweets.

Multiclass Model	Accuracy	Probability weight	Weight based T	Overall weight	IEC
RFC	0.90	0.2	0.5	0.1	0.93
LR	0.89	0.2	0.4	0.0.8(E)	
KNN	0.80	0.2	0.4	0.0.8(E)	
SVM	0.90	0.2	0.5	0.1	
NB	0.83	0.2	0.4	0.0.8(E)	



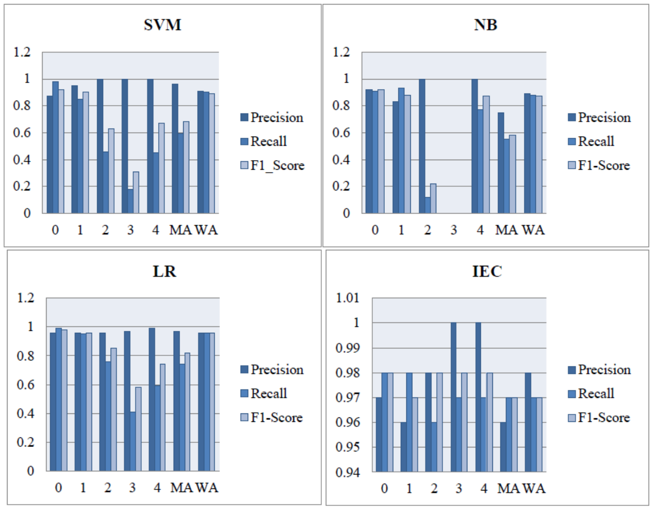


Figure 2. Comparison of accuracy metrics of existing and proposed approach.

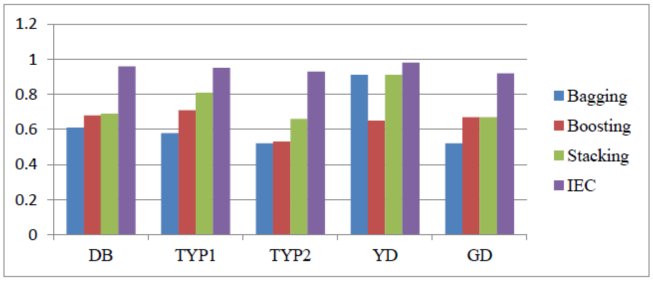


Figure 3. Accuracy of different algorithms.

Table [2-6] describes the classification accuracy of the proposed model and existing model. The traditional machine learning for multiclass classification techniques are applied. The *E* represents the elimination of the classifier due to low weight. In table 2, NB classifier is removed and other classifier such as RFC, LR, KN and SVM are selected for ensemble classification. Similarly, table 3 explains KN and NVB classifier model are removed and other models are taken for ensemble prediction. In table 4, KN and NB are removed and other classifiers are added to ensemble. Table 5 all the model is chosen for YD tweets since all classifier models have equal weight. In table 6, LR, KN and NB are removed classifier for ensemble. The overall model results suggest that NB model does not perform well due to obtaining low weight. Figure 2, illustrates the accuracy metrics for multiclass classification approaches IEC classifier performs better compared with other classifier approaches. Figure 3 explains the state-of-the art approaches compared with IEC classifier it performs better.

V. Conclusion

Sentimental analysis is a challenging task for analyzing real-time data. However, in this paper an IEC is proposed to analyze online tweets. This algorithm is capable of handling multiple classification approaches and improves the performance accuracy. From empirical results it is concluded that proposed ensemble model outperformed well in all classifiers in classification performance. It is also compared with other ensemble approaches like bagging, boosting and stacking IEC model performs superior in classification accuracy.

References

[1] Chia-Chen Chang, Gwyneth Jia and Yi Cheng, Social media, nature, and life satisfaction: global evidence of the biophilia hypothesis, Nature, Scientific Reports, 2020.

[2] E. Gabarron, M. Bradway and L. Fernandez-Luque et al., Social media for health promotion in diabetes: study protocol for a participatory public health intervention design. BMC Health Serv Res. 2018 Jun 05;18(1):414. doi: 10.1186/s12913-018-3178-7.

[3] https://www.who.int/news-room/fact-sheets/detail/diabetes

[4] Elia Gabarron and Eirik Årsand, Social Media Use in Interventions for Diabetes: Rapid Evidence-Based Review, JMIR, 2018.

[5] G. Mita, Ni MC and A. Jull, Effectiveness of social media in reducing risk factors for non communicable diseases: a systematic review and meta-analysis of randomized controlled trials. Nutr Rev. Apr. 74(4) (2016), 237-47.

[6] L. E. Johns, et al. Neighborhood social cohesion and posttraumatic stress disorder in a community-based sample: findings from the Detroit Neighborhood Health Study. Soc Psychiat Epidemiol 47 (2012), 1899-1906.

[7] J. K. Harris, A. Mart, S. Moreland-Russell and C. A. Caburnay, Diabetes topics associated with engagement on Twitter. Prev Chronic Dis. 2015; 12 doi: 10.5888/pcd12.140402.

[8] S. Greenwood, A. Perrin and M. Duggan, Social Media Update 2016 (Pew Research Center, Washington, DC, 2016).

[9] M. J. Widener and W. Li, Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. Applied Geography 54(2014), 189-197.

[10] Q. C. Nguyen, et al. Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity, JMIR Public Health Surveillance 2, e158, https://doi.org/10.2196/publichealth.5869 (2016).

[11] J, Deriu and A. De Luca, Leveraging large amounts of weakly supervised data for multi-language sentiment classification, Proceeding of the 26th International Conference on world wide web, Perth, Australia. 3-7 (2012), 1201-1211.

[12] St. Louis, Twitter hash tags associated with diabetes analyzed, Science Daily, Retrieved (2015, June 12).

[13] V. Diviya Prabha and R. Rathipriya, Readmission Prediction using Hybrid Logistic Regression, Lecture Notes on Data Engineering and Communication Technologies 2020.

[14] N. M. Danish, S. M. Tanzeel and N. Usama, A. Muhammad e, A Muhammud, A. Martinez-Enriquez and A., Intelligent interface for fake product review monitoring and removal, in Proceedings of the 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control(CCE), Mexico City, Mexico, September (2019), 1-6.

[15] Sunir Gohil, Sabine Vuik, Ara DarziSentiment Analysis of Health Care Tweets: Review of the Methods Used, JMIR, 2018.

[16] Oluwakemi Ola and Kamran Sedig, Understanding Discussions of Health Issues on Twitter: A Visual Analytic Study, Online Journal of Public Health Informatics, 2020.

[17] V. Diviya Prabha and R. Rathipriya, Sentimental Analysis using Capsule Network with Gravitational Search Algorithm, Journal of Web Engineering, 2020.

[18] Dimple Tiwari and Bharti Nagpal, Ensemble Methods of Sentiment Analysis: A Survey, International Conference on Computing for Sustainable Global Development, IEEE, 2020.