Diabetes Treatment Pattern Identification Through Layered Tri-Skip-Gram Approach

P.Vasudha Rani, K.Sandhya Rani

Abstract— It's very much crucial to suggest a compatible treatment pattern for a disease, based on its various symptoms at different stages of it. Twitter is a powerful form of social media for sharing information about various issues and can be used to raise awareness and collect pointers about associated risk factors and preventive measures. Twitter tweets retrieved with the Hashtag '#Diabetes' is the origin resource of my paperwork. Here I proposed a recommendation model for suggesting treatments for Diabetes considering social media's Twitter data set which has undergone Bigram &Trigram Analysis to provide the analysis report of people suffering from diabetes at different stages and the treatments suggested for different stages differently. The technique proposed is a layered Tri-skip-gram approach which finds widespread use in the analysis of Textual data.

Here the process starts with relevant diabetes tweets retrieval and then the task of normalization to get required tweet text for the textual analysis. This retrieval process ensures the presence of the unigram 'diabetes' in all the tweets. At the second level, Bigram classification was implemented for the status tracking of Diabetes for which the outcome is different clusters groups of diabetes tweets for different stages of it. Now the main task of this paper is classifying these cluster groups. The proposed model uses feature extraction to find treatment groups and then uses term-tweet classification method to derive the patterns of diabetes treatments suggested for different stages of diabetes as Prediabetes, Type2, and Type1.

Index Terms—Twitter, Diabetes, Bigram Classification, Trigram Classification.

I. INTRODUCTION

The amount of data getting exchanged was witnessing an increase of useful data and no of users in the social media. It is observed that at an instant time there are 638,849,096 tweets shared in a day and there are 338,522,956 Twitter active users on the Internet. So we can observe more involvement and interactive behavior of people in sharing their opinions, feelings, experiences, and emotions, etc that create big data [2]. One of the more common online platforms getting used is Twitter. The Twitter application is designed in such a way that every Twitter user is mapped with many followers and people to follow in the different area of specialisms. Several studies investigated the use of Twitter to analyze ongoing health-related events.

The visible presence of social media is on track in the domain of health care is from 2008. Patients have started searching and enquiring on the internet for chronic health Issues, such as cancer, diabetes, or heart disease for asking and sharing and interacting with patients similar to them. This was a study from 2011. People familiarized to share their health information and their loved once through social media. Nowadays this is a common behavior people expose to seek help from others, in their difficulties. The use of Twitter to analyze ongoing health-related events is investigated in several ongoing studies.

Social media data has got its more application into health domain [1] nowadays. The growing pervasiveness of health studies in social media leads to support the collection and analysis of health-related data in the real world in real time. Twitter is one of the social media with exponential growth. Twitter [3], a short-message micro-blogging service allows people to share their feelings, opinions, emotions, daily life activities, health status, treatments taken, what are the causes, foods suggested etc information with the people in the world around us, as Tweets.

So, we recommend that Twitter data can provide an opportunity to detect and manage public health 'events'. The health event chosen in the proposed methodology is about Diabetes status tracking and treatments suggested at its different stages, for its control. Diabetes is one of the prominent diseases Nowadays, can be effectively managed when caught early. However, when left untreated, it can lead to potential complications that include heart disease, stroke, kidney damage, and nerve damage, etc. Among many diseases, Diabetes disease related data has got more scope for research nowadays.

So, the proposed algorithm experiments on Diabetes Tweets data retrieved through the Twitter API by executing a search query with the hash term as "Diabetes". The algorithm uses a Skip-gram Classification [5] method for the detection of treatment patterns for diabetes through a layered trigram approach. Wherein N-grams/Skip-grams involves Unigrams, Bigrams, and Trigrams to be used in the analysis. The layered process is significant in its nature, and more effective by the method of implementation through Skip-grams. Skip-grams outline the concept of searching for a multi-lexicon in the given tweet in spite of the position of each lexicon in the tweet.

Revised Manuscript Received on April 12, 2019.

P.Vasudha Rani, Research Scholar, CS Dept, SPMVV, Tirupati, Sr.Asst Professor, IT Dept, GMRIT

Dr.K.Sandhya Rani, Professor, CS Dept, SPMVV, Tirupati

Table 1. Summarizes different Classification Methods, Approaches, and Domains of Interest of the Work with Limitations/Outcomes

Author	Classification Method	Approach	Domain	Limitation or Outcome
M. Ghiassi et al. [22]	Sentiment Analysis	N-gram Analysis-ANN	Twitter-Brand related	More accurate estimation of sentiment in experimentation on the Justin Bieber
Yaakov et al. [7]	Topic-based Classification	Unigram Unmasking	Multi-Domain-books	Results based on the most frequent unigrams hurts performance
Abinash et al. [23]	Sentiment Analysis	N-gram M/C learning	Reviews and Blogs	Symbol Analysis not happening
John Violos et al. [21]	Text Classification	N-gram graph Rep Model	Reuters and Newsgroups	characteristics that exploit the assets of graph representations
Ron Bekkerman et al. [6]	Topic-based classification	Unigram+Bigram	Newsgroup Dataset	Attempt to incorporate bigrams to the document representation and report an improvement in the result
Vasudha et al [9]	Tweet based Classification	Layered Bi-skip-gram Classification	Diabetes Tweets	Combination of Unigrams + Bi-skip-grams

The first classification of tweets is to retrieve health-related tweets from non-relevant tweets to Diabetes from the Twitter web site. For this Naive Bayes model classification algorithm is used to classify tweets into two categories: Tweets with the unigram "diabetes" and without unigram "diabetes". The unigram [2] "diabetes " is already present in all the resultant health tweets. These unigrams are used at the next level.

The second level classification requires the combinations of unigrams which are called bi-grams. For which, feature extraction needs to be done from the Diabetes tweets for retrieving features related to stages of Diabetes. Highest Ranking Feature Extraction Algorithm is used for this purpose. And the Proposed model uses Term-tweet classification with these features to classify the tweets into clusters. These Unigrams & Bigrams are used in the next level classification.

The third level classification requires the combinations of unigrams & bi-grams [6] which are called tri-grams. At this level, feature extraction needs to be done from the Diabetes tweets for retrieving features related to treatment suggestions of Diabetes. Highest Ranking Feature Extraction Algorithm is used for this purpose. And the Proposed model uses Term-tweet classification with these features to classify the tweets into clusters of clusters.

To justify the layered nature of the method, the process refers to my earlier paperwork [9] where it experimented with Bi-skip-grams generated from Diabetes Tweets regarding the tracking of the different stages of diabetes such as Type2, Type1, Insulin, Death, Prediabetes, Suffering, Symptoms. The process of Analysis considered the Bi-Skip-grams as Diabetes-Type2, Diabetes-Type1, etc. Bigram corpus is a subset of Bi-skip-grams. The skip-gram method can be explained with the following example. A typical tweet is considered as

Tweet: RT @ABCcatalyst: Exercise packs a 4-way punch against diabetes:- It helps you lose weight- Shrinks abdominal fat.

Unigrams: 'Exercise' ,'packs', '4-way', 'punch', 'against', 'diabetes', 'Helps', 'lose' ,'weight', 'Shrinks', 'abdominal', 'fat' where all the single words are considered.

Bi-grams: 'Exercise packs', 'packs 4-way', '4-way punch', 'punch against', 'against diabetes', 'Diabetes helps', 'helps lose',' lose weight', 'weight Shrinks', 'shrinks abdominal', 'abdominal fat' where all the two sequent word combinations are considered.

Skip-gram model is a method of checking for N-words not required to be in sequent from a given tweet text. Skip-grams are defined with the orders of {1,2,3,4….}. Skip-grams defined with order 1 are unigrams, with order 2 are Bi-grams, with order three are Trigrams. Higher order skip-grams refers to four-grams, five-grams, etc. At the unigram level, both Skip-gram model and N-grams work similar, but where it differs from Bi-gram level.

Bi-skip-grams: In addition to the above list, it also includes 'Exercise diabetes', 'exercise weight', 'Diabetes lose', exercise abdominal' etc where all the words including In-sequent word combinations.

A. Feature Reduction

In the second level classification, term-tweet classification helps to get the Feature vector with their contribution counts either '1' or '0'. Now applying the Highest Ranking Feature Extraction Algorithm, to get the counts of features contributing more to the Diabetes are identified by specifying a threshold value. Applying term tweet matrix Classification at the last, to get the mapping of tweets with the features.

Now applying the Highest Ranking Feature Extraction Algorithm, Features related to different treatments suggested contributing more to Diabetes are identified. These features are applied for Term-Tweet Classification technique using term-tweet matrix Classification, upon the tweets resulted from Bigram Classification and again applying Term-Tweet Classification considering the features extracted for "diabetes" treatments.

Assume some features are identified such as "Insulin", "Symptoms", "Suffering", Type1, Type2, Death, and Pre-diabetes, etc…Now each of these unigrams needs to be taken for bigrams preparation such as Diabetes- suffering", Diabetes -Type1, Diabetes-Type2, Diabetes –Death…will be considered for analysis. Actually, these features can be treated as skip-grams with length [2] because there is no condition verified that those two words should be sequent in the tweet.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

At the next level Classification, again feature selection algorithm is applied to select the Diabetes Treatment-related features such as "Diet", "Exercise", "Therapy", "Medication", "Yoga", "Vitamins" from the resultant Tweets from the Bigram Analysis to form Trigrams for Diabetes Treatment Analysis making use of the Tri-Skip-Grams {Diabetes-Type1-Diet, Diabetes-Type1-Exercise, Diabetes-Type1-Medication, Diabetes-Type1-Therapy, Diabetes-Type1-Yoga, Diabetes-Type2-Diet, Diabetes-Type2-Exercise, Diabetes-Type2-Therapy, Diabetes-Type2-medication, Diabetes-Type2-yoga, Diabetes-Insulin-Diet, Diabetes-Insulin-Exercise, Diabetes-Insulin-Yoga,Diabetes-Insulin-medication, Diabetes-Insulin-Therapy, Diabetes-Pre-diabetes-Diet, Diabetes-Pre-diabetes-Medication, Diabetes-Pre-Diabetes-Yoga, Diabetes-Pre-diabetes-Exercise, Diabetes-Pre-diabetes-Therapy.

The prediction results of this trigram analysis are the treatment suggestions for Diabetes for all the different stages i.e "People with Symptoms of Diabetes", "People suffering from Type1 level Diabetes", "People suffering from Type2 level Diabetes", "People using Insulin for Diabetes" and the corresponding treatment suggestions in the order, are Diet, Exercise, Therapy, Yoga, Medication etc.

B. Authentication of Health Tweets Data Set

Nowadays people are using Twitter as a common platform to communicate and share their feelings and experiences related to their lifestyle, opinions on general, public, political and also includes health issues with the public. So, the Twitter platform can be utilized to develop a disease surveillance system to mine health-related information to be useful for the public. The direct users of this information are the patients and the general public. This necessitates the validation of disease-related tweets by health care professionals to ensure they are evidence-based and credible information is used to make critical decisions. The main focus of this paperwork is related to Diabetes Disease Tracking of public health. For this work, Twitter is used as the main source of collecting Diabetes information in the form of tweets. Analysis is done to evaluate the Diabetes related tweets taken as input in the experimentation for their validity to show that they are evidence-based tweets.

The tweets are classified based on user's sources and locations and obtained 730 different twitter accounts wherein 60% of them are NGO's working regarding different health issues, and remaining 40% are regular user account assuming that they may be patients or accompanies to patients. Some of the lists of NGO 's operating related to Diabetes are @Nutrientology_, @MyHealthTest, @HealthyInteract, @HlProviders, @MyHealthTest, @E4Diabetes, @CitizensHealth etc..All the Health-NGO's related to Diabetes are maintaining a blog and a twitter account to provide regular updates to the public.

II. LITERATURE SURVEY

A. Related Works

In 2003, Ron Bekkerman et al. [6] have proposed a framework for Text categorization based on Bigrams in addition to Unigrams. The author demonstrated an attempt to incorporate bigrams in a document representation based on distributional clusters of unigrams, and report an improvement to our baseline results on Newsgroups dataset.

In 2018, Yaakov HaCohen-Kerner et al. [7] has presented an alternate method to feature reduction i.e Unigram-Unmasking instead focusing on the concept of "Bag-of-words" vector. The proposed model uses Topic-based Classification approach for the text classification for the domain of "Online free textbooks" categorized into Career and Study Advice, Economics and Finance, IT Programming, Natural Sciences, Statistics, and Mathematics.

In 2012, Badr Mohammed Badr et al. [5] have conducted a series of classification experiments using two machine learning algorithms i)SVM ii)Multinominal Naïve Bayes. His goal was a three-fold i) To investigate whether or not (POS) features are useful ii) To study the effectiveness of sparse phrasal features (bigrams and skip-grams) [7] to capture sentiment information iii) To investigate the impact of combining unigrams with phrasal features on the classification performance and the data domain is Internet-based movie reviews.

In 2016, Abinash Tripathy et al. [23] have discussed four different Supervised machine learning methods such as Naïve Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM). He started this work with a motivation that a higher level of N-grams is considered, the result is expected to be better. Experimentation was done for all the cases such as unigrams, Bigrams, Trigrams, Unigrams+Bigrams, Bigrams+Trigrams, Unigram+Bigram+Trigram features individually for each classification method and better results shown with SVM. Sentiments of people have been analyzed on movie reviews.

In 2013, M. Ghiassi et al. [22] have developed comparable Sentiment Classification models using SVM. It introduces an approach to supervised feature reduction using n-grams and statistical analysis to develop a Twitter-specific Lexicon for sentiment analysis which yields improved sentiment analysis accuracy. He compared and proved that the devised twitter specific lexicon gives significantly more effective recall and accuracy metrics values when compared to traditional sentiment lexicon.

In 2018, John Violos et al. [21] have applied the classification over high-frequency data streams. In his classification model, the text is represented as N-gram graphs and the classification process takes place using text pre-processing, graph similarity, and feature classification techniques following the supervised machine learning algorithms. The proposed model and various parameters are evaluated experimentally and the high-frequency stream emulated using two public data sets 20NewsGroup and Reuters-21578.

B. Review

Table 1 summarizes the Author details, Classification Method used, Approach, Data Domain and Limitations or Outcomes of various n-gram approach for text classification for both Health and other related data. Ron Bekkerman

et al. [6] has shown a positive scope of research for text classification by incorporating bigrams for the representation of the document in addition to unigrams. The experimentation used an approach of topic-based classification. Badr Mohammed Badr et al. [5] also performed a similar kind of experimentation to Ron that phrasal features are combined with unigrams to show better performance with classification either by using SVM or NB model. Yaakov HaCohen-Kerner et al. [7] proposed and implemented an alternate method to feature reduction namely Unigram-Unmasking. Abinash Tripathy et al. [23] has experimented with different machine learning algorithms such as SVM, NB, SGD, ME for the different feature vector combinations as Unigrams+Bigrams, Bigrams+Trigrams, Unigrams+Bigrams+Trigrams, etc. M. Ghiassi et al. [22] has developed a framework for sentiment analysis experimented with Traditional Lexicon and Twitter specific Lexicon. And proved that Twitter specific Lexicon model has given better performance. John Violos et al. [21] have applied a different strategy that n-gram graphs are used to represent features. To further strengthen the works in text classification, a new layered approach is experimented and proved useful results for diabetes data with the inclusion of features combinations {unigrams, bi-skip-grams, Tri-skip-grams}.

III. PROPOSED METHODOLOGY

A. Layered Tri-skip-gram Approach

In this paper, Tri-skip-gram classification model is proposed for the Diabetic treatment pattern identification related to different stages of Diabetes. The proposed model can effectively identify the suitable treatment suggestion and profusely utilize the experiences, shared as opinions of people who are facing a similar problem of diabetes. This model uses a Layered Architecture which is based on bigrams, Trigrams classification of tweets data. The detailed architecture is shown in **Fig.1**. This work is a part of Natural Language Processing (NLP). The proposed model consists of six phases namely i) Relevant Tweets retrieval ii) Standardizing the tweets iii) First level feature extraction iv) Bi-skip-gram Classification v) Second level feature extraction vi) Tri-skip-gram Classification. The Tri-skip-gram approach which is proposed in this phase is presented in the form of an algorithm by the name **Algorithm 1**. A brief description of each phase is discussed.

Phase 1: Relevant Tweets Retrieval

At first, the key job is to focus on classifying health tweets from all other non-health categories of tweets [8]. This is done by the execution of a search query through the Twitter API by specifying the search terms related to the health aspect. Here the aspect of health is identified as 'Diabetes' or 'Diabetic'. Implementation of search twitter method in R uses the Naïve Bayes algorithm for the retrieval of diabetes-related Tweets from non-relevant and result is a filename.CSV file with the information Tweet text, date&time of creation, screenName, status source, retweet count, isRetweet, location attributes as longitude and latitude. Tweet text information from filename.CSV is given as input for the next phase i.e Phase 2.

Algorithm 1 Trigram Classification Approach

Input: T_D - Tweets on Diabetes Dataset
R_{t2}- Resultant Diabetes-Type2 class Tweets
R_{t1}- Resultant Diabetes-Type1 class Tweets
R_{In}- Resultant Diabetes-Type2 class Tweets
R_{Sm}- Resultant Diabetes-Type1 class Tweets
Output: Identified Treatment Patterns database T_p
Classification Process
Step1: Preprocessed Diabetes Tweets T_d
Step2: Apply Feature Extraction Algorithm
Step3: Apply Bigram Classification Process
Step4: Apply Term-Tweet Classification using selected Features {type2, type1, insulin, symptoms}
Step5: Resultant four classes of Tweets Datasets $R_{t2,}$ R_{t1}**,** R_{in}**,** R_{sm}
Step6: For each Dataset R_{t2} as Input for Trigram Classification Method
Step7: For each class of Tweets which already Bigram classified
Step8: Apply Feature Extraction Algorithm
Step 9: **Apply Term-tweet Classification {Diet, Exercise, Medication, Yoga, Therapy}**
Step 10: Result is the Classified tweets with the one treatment skip gram Diet …
Step 11: Repeat from step 7 for the remaining treatments
Step12: End

Phase 2: Standardizing the Tweet Text

This is the basic and compulsory phase of the proposed model. Tweets are a textual description of people's opinion posted in the social networking environment. The proposed model requires only the cleaned tweet text as input to it. For which the Diabetes Data file has to undergo tweet normalization process in which tweets are preprocessed to eliminate spaces, duplicates, unrequired attributes, extra unnecessary symbols like @, # from tweet text. The duplicate tweets from the file have no impact on the accuracy and performance of the model. Removing duplicates is one of the key tasks of the normalization of data.

Phase3: First Level Feature Extraction

For the purpose of research, there are different ways of collecting features. They are i) utilizing unigrams ii) unigrams with their frequency iii)bi-grams iv) bi-grams with their frequency. In this paper, all these feature techniques are used to develop a classification model. As it is a layered process, at the first level among the unigram feature chosen is "Diabetes". At the second level, the highest ranked feature extraction algorithm is executed for the tweet text to identify all the unigrams with their frequency counts. Among which only the features with a threshold value are chosen for Bi-skip-gram classification. From these features, unigram word features describing different stages of Diabetes are taken for the Next level classification.

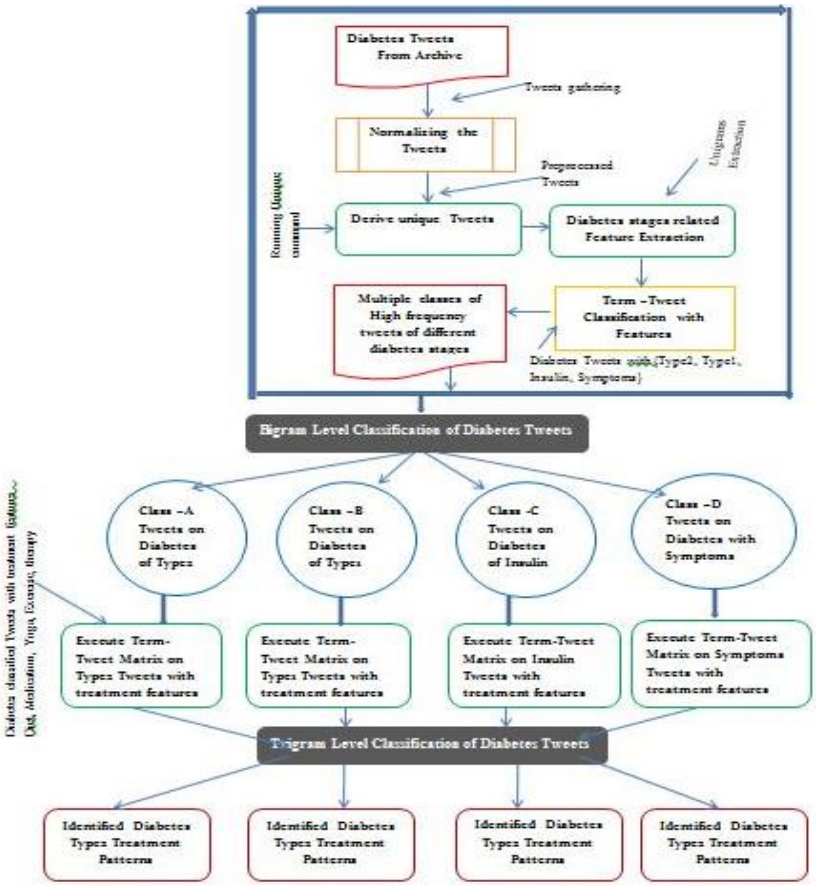

Fig.1. Detailed Layered Architecture of the Proposed Work of Diabetes Treatment Identification

Phase4: Bi-skip-gram Classification

In our earlier work [9] detailed process of Layered Bi-gram classification of Diabetes Tweets data for categorizing the tweets based on different stages of diabetes, into multiple classes of tweets has been discussed. It has been proved that more people are suffering from Diabetes and the highest percentage of people are at the stage of insulin usage,next is type1 and then type2. The different tweet clusters obtained from bigram classification are diabetes-Type2 tweets, Diabetes-Type1 Tweets, Diabetes-Insulin Tweets, Diabetes-Symptoms Tweets. These are the input for next level Tri-skip-gram Classification. The Architecture of Bigram Classification is shown in Fig.2.

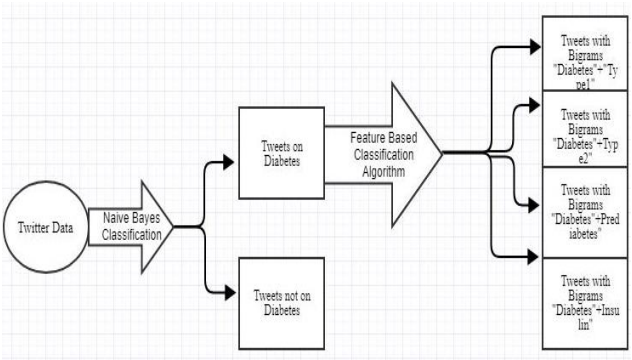

Fig.2: System Architecture for Bigrams Classification Approach on Tweets with status feature's information

Phase5: Second Level Feature Extraction

The input for this phase are the four .csv files containing the tweet clusters from the previous phase. They are Diabetes-Type2.csv, Diabetes-Type1.csv, Diabetes-Insulin.csv, Diabetes-Symptoms.csv. Now for the selection of features, each type of tweet text file is taken for processing the unigram features. The process explained in **Phase3** is repeated for all the four classes of tweets for identifying the treatment features. The list of treatment features identified are Diet, Yoga, medication, exercise, therapy.

Phase6: Tri-skip-gram Classification

Text classification is the task of automatically assigning documents to a fixed no of categories. Four classes of Tweets are the result of bigram level classification. All the four classes of tweets with the bigrams Diabetes-Type2, Diabetes-Type1, Diabetes-Insulin, Diabetes-Symptoms are taken to be categorized for the different treatment features. The Resultant tweets and treatment features "Diet, Yoga, medication, exercise, therapy" from the previous Phase are the Input for the current phase analysis. At this phase again term-tweet classification [11] is done by giving the treatment features also as an input. The outcome of this phase are the identified treatment patterns for each stage of diabetes.

By executing the commands for classification, it results in binary vector contains '1' for the presence of a term in a tweet and '0' for the absence of a term in a tweet. Finally by performing summation functions on the numerical vectors, identifies the strength of each treatment pattern for the stages of diabetes. This experimentation is repeated for all the four clusters to get all the treatment patterns with their strengths prioritized. The process flow is shown in Fig.3.

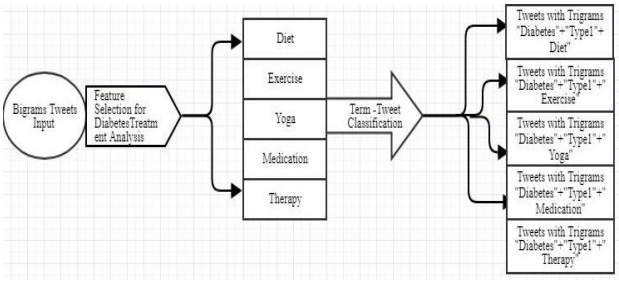

Fig. 3: System Architecture for Level-3 Trigrams Approach on Health Tweets with Treatment Feature's Information

High Ranked Feature Extraction Algorithm

The resulted feature vectors from term-tweet classification are the input to this algorithm to extract only high ranked features based on the count score for each feature. The step by step process is shown in Algorithm 2.

Algorithm 2 High Ranked Feature Extraction

Input : Tweets Dataset Td

Output: Unigram features extracted

Feature Extraction

Step1: Preprocessing Td to filter data attributes

Step2: Preprocessing Tweet text to remove symbols like @,#, etc and stop words

Step3: Result of preprocessing is a .csv file with tweet text

Step4: Tokenization Process

Step5: Unigrams corpora Tc from tokenization

Step6: For all the unigrams of the corpora Tc

i)For each unigram Ui, save wi<-rank_of_Ui(UI,Ci)

Step7: Each wi-count of unigram Ui is compared with threshold limit

Step8: Treatment features identified

Step9: End

IV. EXPERIMENTAL RESULTS

In this paper work, Diabetes Tweets retrieved from Twitter is the source of data for experimentation. To retrieve diabetes tweets from Twitter, a search query need to be executed with the search term "diabetes" or "diabetic", "lang", and "no of tweets required" through Twitter API. The lang option specifies the language in which tweets to be retieved. This parameter is set to English. The retirved tweets are preprocessed to get the clean tweet text. data for the next bigram level classification. The proposed model of Tri-skip-gram classification is experimented to derive patterns of treatments of Diabetes.

A. Preprocessing of Tweets

We need to prune Diabetes tweets data to be suitable for text classification. Example tweet is shown below

Before Preprocessing:

"RT @250HealthyFoods: Newly published meta-analysis of data from 2832 people suggests #nuts may improve #insulin sensitivity—for nutrient co"

After Preprocessing :

"HealthyFoods Newly published meta-analysis of data from 2832 people suggests nuts may improve insulin sensitivity for nutrient co"

B. Feature Extraction

In this work, experimenting the highest ranked feature extraction algorithm on Diabetes tweets data and prediabetes, type2, type1, insulin classes of tweets data results into bigrams and trigrams. The results from the first level feature extraction are shown in Table 2.

Table 2. Features generated and to be used for Bigram Classification –Diabetes – Stages

Feature1	Feature 2	Feature 3	Feature 4	Feature 5	Feature6
Symptoms	**Type1**	**Type2**	**Insulin**	**Death**	**Suffering**

These table data features are used in the Bigram classification process to get categories of tweets for all the unigram features. The resultant tweets of Bigram classification are stored in the tweet files namely Diabetes-Type2.csv, Diabetes-Type1.csv, Diabetes-Insulin.csv, Diabetes-Symptoms.csv. These tweets data is used in trigram analysis. Before the execution of trigram classification, again feature extraction was done to retrieve treatment features listed in Table 3.

Table.3. Features generated and to be used for Trigram Classification

Feature1	Feature 2	Feature3	Feature 4	Feature5	Feature6
Exercise	**Diet**	**Medication**	**Yoga**	**Therapy**	**Vitamin**

The generated bigrams and trigrams data from Table.2. & Table.3 are represented through many forms as "WORD CLOUD", mind map, etc. Word clouds of unigram, bigram and trigram are shown if Fig.4, Fig.5 and Fig.6 respectively.

1) Unigram Features collected from Diabetes Tweets

Fig.4. Word Cloud of Unigrams Generated from the Diabetes Tweets Dataset

2) Bigram Features collected from Diabetes Tweets

Fig.5.Word Cloud of Bigrams Generated from the Diabetes Tweet Dataset with Feature Dataset

3) Trigram Features collected from Diabetes Tweets

Fig.6.Word Cloud of Trigrams Generated from the Diabetes Tweets

C. Tri-skip-gram Classification

The classification takes two inputs. They are Tweets Data file and the extracted features from the previous phases. The experimentation of this classification is done through the generation of Term Tweet Classification Matrix.

Proposed Term Tweet Matrix Classification

Diabetes tweet corpus and extracted features is the input for Term-Tweet Matrix generation. Either the TermTweetMatrix or TweetTermMatrix can be generated based on whether you want terms as rows or tweets as columns, or vice versa generates sparse matrices for corpora. Taking the selected Features into consideration to find out their strengths in the Diabetes Tweets Data set TermTweetMatrix is generated with the corresponding frequencies and using Sigma ∑ function at the column level. If ∑ function is used at row level produces strength of each tweet in terms of features. And the resultant values are placed in Table 4.

Table 4.Term Frequencies for the Selected Ranked Features

Terms			
TweetNo	Type1-Diet	Type 2-Exercise	Insulin-Exercise
144	0	1	2
236	1	2	1
237	0	1	1
242	1	1	1
246	0	2	0
248	1	1	1
273	2	1	2
489	1	0	1
502	0	1	0

After classifying tweets with these selected features, each feature gets its count as its strength. The preferred treatment for each category is shown as a Priority sequences in Table 5.

Table 5. Treatment Pattern Identified and their Priorities

	Treatment Patterns and their Priorities				
Prediabetes	Diet- 1	Exercise - 2	Medication- 3		
Type2	Diet- 1	Exercise - 2	Medication- 3	Yoga- 4	Therapy-5
Type1	Yoga- 1	Medication-2	Diet- 3	Exercise -4	Therapy-5
Insulin	Diet-1	Medication-2	Therapy-3	Exercise -4	Yoga-5
Symptoms	Exercise -1	Diet-2	Medication-3	Yoga-4	Therapy-5

Note: Priorities are represented from 1 to 5

Comparison with Google Trends

The comparison among different treatments suggested for Diabr=etes based on Google Trends is shown in Fig.7.

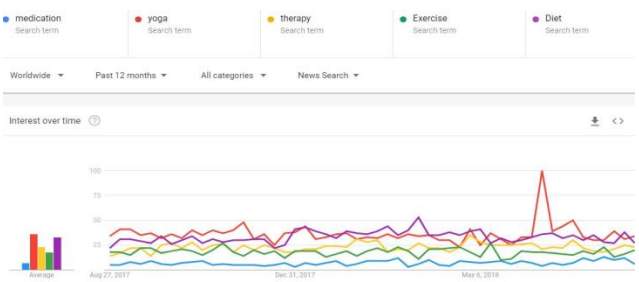

Fig.7. Google Trends Based Comparison Among Different Treatments

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

D. Analysis of strengths of all the treatment patterns for different Diabetes Stages

The numerical strengths derived from tweets data of each treatment pattern for a particular Diabetes stage is projected in the below Table 6. Individual treatments on X-axis and No of tweets support that treatment on Y-axis The identified treatment patterns for each stage of Dibetes and its strengths are shown in Fig.9. The final deriving conclusion is that for the stage of Type2, at most preferred treatment is "DIET' & for the stage of Type1, at most preferred treatment is "Medication' & for the stage of Insulin, at most preferred treatment is "DIET' & for the stage of Symptoms, at most preferred treatment is "DIET' & Therapy.

Table 6. Data Table of Treatments and the Corresponding Tweet Count

S. No	Treatment Features	Frequency strength (I)	Frequency strength (t2)	Frequency strength (t1)	Frequency strength (s)
1	**Diet**	5000	2300	100	2500
2	**Exercise**	900	300	100	2000
3	**Physical Activity**	1100	300	100	700
4	**Yoga**	400	200	500	200
5	**Medication / Drugs**	2800	400	1800	2200
6	**Therapy**	2700	100	100	2700

Comparison of Treatment patterns

Diabetes Treatments are identified for all the stages of Diabetes. When we compare treatment patterns utilized for all the stages, taking individual treatments on X-axis and No of tweets supporting that treatment on Y-axis, the comparison analysis is shown through Barcharts in Fig.8.

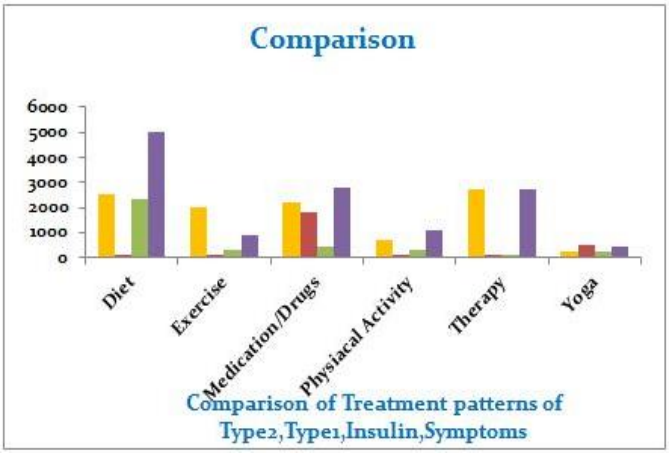

Fig.8. Comparison s among Different Stages of Diabetes -Type2, Type1, Insulin, Symptoms and their Treatments through Bar chart

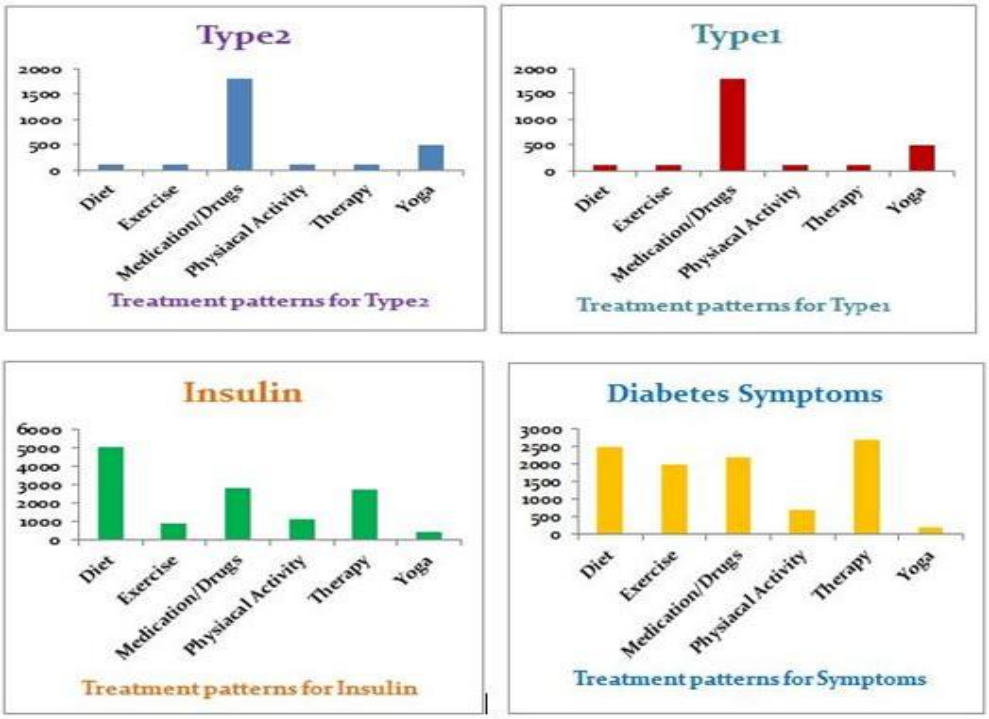

Fig.9. Treatments Patterns Identified for Diabetes at different Stages of Type2, Type1, Insulin, Symptoms

V. CONCLUSION

In this paper has a model is proposed for tweet text classification for the Diabetes data. The primary focus of the paper is to introduce an effective method for tweet text analysis related to Diabetes. The model uses a layered Tri-skip-gram approach which works at two levels, Bigram level & Triram level. The two key processes involved are feature extraction and then classification at each level. The outcome of this paper is useful for the people to understand the significance of identified treatments patterns which are

helpful to prevent and control Diabetes. The proposed model effectively utilized all levels of layered approaches such as unigram, bigram and trigrams to identify the treatment patterns for all the stages of Diabetes.

REFERENCES

1. Ahmed Ali, Walid Magdy, and Stephan Vogel, "**A Tool for Monitoring and Analyzing HealthCare Tweets",** Qatar Computing Research Institute Qatar Foundation, Doha, Qatar, 2013.
2. An Article on "Big Data is the Future of Healthcare", by *cognizant* 20-20 insights | September 2012.
3. A Avinash Kumar Barnwal, Govind Kumar, Amit Kumar Das,"**Application of Twitter in Health Care Sector for India**", 3rd Int Conf. on Recent Advances in Information Technology RAIT-2016.
4. Chris Okugami1, Ross Sparks1* and Sam Woolford2, "**Twitter Data Offers Opportunities for Public Health Professionals**", Bentley University, Australia.
5. Badr Mohammed Badr, S. Sameen Fatima, Osmania University, "Using Skipgrams, Bigrams, and Part of Speech Features for Sentiment Classification of Twitter Messages", 2015, IIIT, ICON 2015 Proceedings.
6. Ron Bekkerman, James Allan, "**Using Bigrams in Text Categorization**", 2003, Department of Computer Science, University of Massachusetts.
7. Yakoov HaCohen-Kerner, Avi Rosenfeld, Asaf Sabag, Maor Tzidkani, **"Topic-Based Classification through Unigram Unmasking",** Science Direct, 2018.
8. Sandeep Kumar, Rahul Rishi, MD University, Rohtak, India "Data Collection And Analytics Strategies of Social Networking Websites", 2015 IEEE.
9. V.Vasudha Ran, Dr.K.Sandhya Rani, **"Efficient Tool for Diabetes Tracking through Layered Bigram Approach"**, SSRN Elsevier DataBase, Volume No 04 Issue No 04 2018.
10. Demitrios E. Pournarakis, Dionisios N. Sotiropoulos, George M. Giaglis*,"**A computational model for mining consumer perceptions in social media",** 2016 Elsevier, Greece.
11. Sunmoo Yoon, RN, Ph.D., Noémie Elhadad, Ph.D., and Suzanne Bakken,"**A Practical Approach for Content Mining of Tweets**", Columbia University, New York, New York, July 2013.
12. Tauhid R. Zaman, Ralf Herbrich," **Predicting Information Spreading in Twitter**", Cambridge, MA, Cambridge, UK.
13. Cynthia Chew, Gunther Eysenbach," Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak", November 2010, Volume 5, Issue 11, PLOS ONE.
14. Introduction to the tm Package, **Text Mining in R**, Ingo Feinerer, December 6, 2017.
15. Andrei Sechelea, Tien Do Huu, Evangelos Zimos, and Nikos Deligiannis, "**Twitter Data Clustering and Visualization"** Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels, Belgium, January 2016.
16. Matthew Mayo, "**A Clustering Analysis of Tweet Length and its Relation to Sentiment**", TSYS School of Computer Science, Columbus State University, 2015
17. Kim Holmberg1,2, Kristina Eriksson-Backa2,3, and Stefan Ek2,3, "**Tweeting about Diabetes and Diets – Content and Conversational Connections**", Department of Organization Sciences, VU University Amsterdam, Amsterdam, Netherlands, Springer 2014.
18. Shubhi Choudhary, Vijay Singh, Goutam Chakraborty, "**Application of Text Mining on Tweets to Analyze Information about Type-2 Diabetes** " , Oklahoma State University, OK, US, 2015.
19. Vincenza Carchiolo, Alessandro Longheu(B), and Michele Malgeri, "**Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics**", Universit`a Degli Studi di Catania, Catania, Italy ,2015.
20. Jianguo Chen a , Kenli Li a , b , ∗, Huigui Rong a , ∗, Kashif Bilal c , Nan Yang d , Keqin Li, "**A disease diagnosis and treatment recommendation system based on big data mining and cloud computing**" , ScienceDirect 2018.
21. John Violos, Konstantinos Tserpes, Iraklis Varlamis, and Varvarigou, **"Text Classification Using the N-Gram Graph Representation Model Over High Frequency Data Streams",** Frontiers in Applied Mathematics and Statistics, 2018.
22. M. Ghiassi, J.Skinner, D.Zimbra, "Twitter brand sentiment Analysis: **"A hybrid system using n-gram analysis and dynamic artificial neural network**", Elsevier 2013.
23. Abinash Tripathy, Ankit Agarwal, Santanu Kumar Rath, **"Classification of Sentiment reviews using n-gram machine learning approach",** Elsevier, 2016.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication