(12) INNOVATION PATENT (11) Application No. **AU 2021104534 A4**
(19) AUSTRALIAN PATENT OFFICE

(54) Title
A SYSTEM AND METHOD FOR ANALYZING TAMIL TWEETS INTO POSITIVE AND NEGATIVE SENTIMENT

(51) International Patent Classification(s)
G06F 40/30 (2020.01) *G06N 3/02* (2006.01)
G06Q 50/00 (2012.01)

(21) Application No: **2021104534** (22) Date of Filing: **2021.07.24**

(45) Publication Date: **2022.05.05**
(45) Publication Journal Date: **2022.05.05**
(45) Granted Journal Date: **2022.05.05**

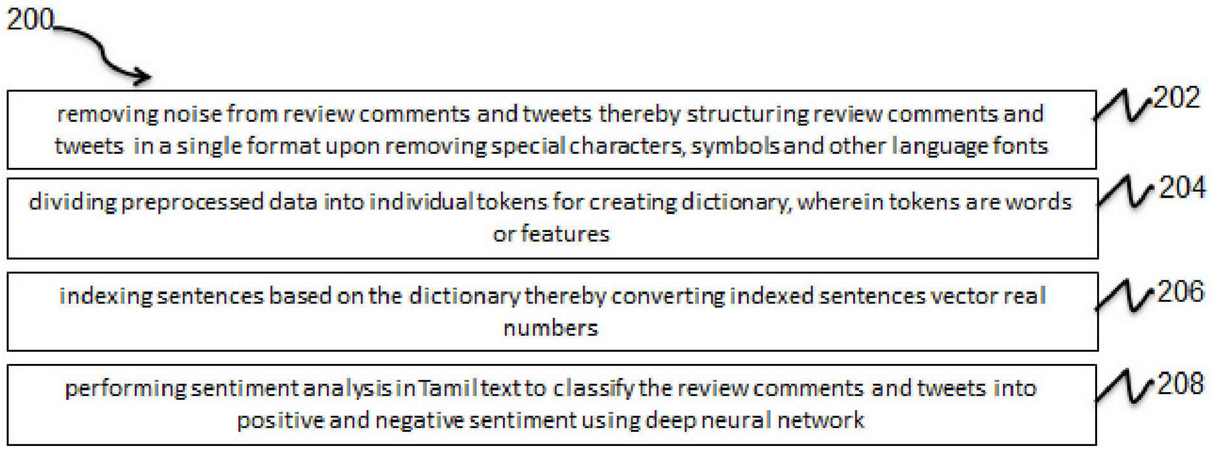
(71) Applicant(s)
Hindustan Institute of Technology and Science;S. Gokila;S. Rajeswari

(72) Inventor(s)
Gokila, S.;Rajeswari, S.

(74) Agent / Attorney
Dr. S. Gokila, 22 Glenroy road Glenroy, Victoria, VIC, 3046, AU

ABSTRACT

The present invention generally relates to a system and method for analyzing Tamil tweets into positive and negative sentiment comprises a pre-processing unit for removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts; a natural language processing unit configured for dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features; wherein sentences are indexed based on the dictionary thereby converting indexed sentences vector real numbers; and a prediction unit equipped with deep neural network for sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment.



removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts

dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features

indexing sentences based on the dictionary thereby converting indexed sentences vector real numbers

performing sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment using deep neural network

Figure 2

A SYSTEM AND METHOD FOR ANALYZING TAMIL TWEETS INTO POSITIVE AND NEGATIVE SENTIMENT

FIELD OF THE INVENTION

The present disclosure relates to a system and method for analyzing Tamil tweets into positive and negative sentiment.

BACKGROUND OF THE INVENTION

Tamil is one of the Indian languages which still requires the state-of-the-art model for opinion mining or sentiment analysis (SA). Particularly in nations like India, where a greater number of regional languages are spoken. In everyday life, such extravasation takes place. The public's viewpoint is being accepted by the E-Commerce industry, entertainment channels, and even delivery systems. Those technology-oriented activities are ready for a regional language assessment.

However, a common feedback system is used to keep track of a variety of critical decision-making processes. This type of polling allows for responses in regional languages, which will undoubtedly generate a wide range of opinions. In such cases, systemized auto analysis is necessary to forecast the customer's and consumer's viewpoints. The commercial/entertainment platform is one such sector that receives feedback.

In the view of the forgoing discussion, it is clearly portrayed that there is a need to have a system and method for analyzing Tamil tweets into positive and negative sentiment.

SUMMARY OF THE INVENTION

The present disclosure seeks to provide a system and method for analyzing Tamil tweets into positive and negative sentiment using a deep neural network.

In an embodiment, a system for analyzing Tamil tweets into positive and negative sentiment is disclosed. The system includes a pre-processing unit for removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts. The system further includes a natural language processing unit configured for dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features. Sentences are indexed based on the dictionary thereby converting indexed sentences vector real numbers. The system further includes a prediction unit equipped with deep neural network for sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment.

In an embodiment, the pre-processing include removal of English, removal of space and removal of special characters and the like for structuring.

In an embodiment, word dictionary is included in the implementation process, due to which missing index does not occur in any point of vectorization.

In an embodiment, prediction are processed from neural network to machine learning and deep learning, wherein prediction techniques applied for natural language processing with narrow down on regional language still requires an enhancement because style, slang and words used are invariably unique to person to person.

In an embodiment, the prediction unit is examined in terms of dictionary size, vector dimension, number of hidden layer and batch size.

In an embodiment, the vector size of the word is equalized to the size of the dictionary and the dictionary includes all possible words used in Tweet, wherein in Tamil some of the statements is having peculiar or rare but strong supporting words that are based on the expertization in language.

In an embodiment, the deep neural network is trained with the tokens segregated from Tamil review comments and tweets with five layers of network with a defined dense number.

In another embodiment, a method for analyzing Tamil tweets into positive and negative sentiment is disclosed. The method includes removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts. The method further includes dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features. The method further includes indexing sentences based on the dictionary thereby converting indexed sentences vector real numbers. The method further includes performing sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment using deep neural network.

In an embodiment, the dictionary is formed with individual tokens including emoji's and stop words.

In an embodiment, natural language process is configured with a dense set of vocabulary and a well-equipped learning architecture.

An object of the present disclosure is to extract the opinions from tweets, reviews and comments for development of natural language processing in Tamil language.

Another object of the present disclosure is to analyze the Tamil tweets into two polarities as positive and negative.

Yet another object of the present invention is to deliver an expeditious and cost-effective method for analyzing Tamil tweets into positive and negative sentiment.

To further clarify advantages and features of the present disclosure, a more particular description of the invention will be rendered by reference to specific embodiments thereof, which is illustrated in the appended drawings. It is appreciated that these drawings depict only typical embodiments of the invention and are therefore not to be considered limiting of its scope. The invention will be described and explained with additional specificity and detail with the accompanying drawings.

BRIEF DESCRIPTION OF FIGURES

These and other features, aspects, and advantages of the present disclosure will become better understood when the following detailed description is read with reference to the accompanying drawings in which like characters represent like parts throughout the drawings, wherein:

Figure 1 illustrates a block diagram of a system for analyzing Tamil tweets into positive and negative sentiment in accordance with an embodiment of the present disclosure;

Figure 2 illustrates a flow chart of a method for analyzing Tamil tweets into positive and negative sentiment in accordance with an embodiment of the present disclosure;

Figure 3 illustrates a process flow to classify Tamil tweet with emoji in accordance with an embodiment of the present disclosure;

Figure 4 illustrates a five layer deep neural network architecture to classify Tamil tweet review with an emoji in accordance with an embodiment of the present disclosure; and

Figure 5 illustrates sample tweet before and after pre-processing in accordance with an embodiment of the present disclosure.

Further, skilled artisans will appreciate that elements in the drawings are illustrated for simplicity and may not have necessarily been drawn to scale. For example, the flow charts illustrate the method in terms of the most prominent steps involved to help to improve understanding of aspects of the present disclosure. Furthermore, in terms of the construction of the device, one or more components of the device may have been represented in the drawings by conventional symbols, and the drawings may show only those specific details that are pertinent to understanding the embodiments of the present disclosure so as not to obscure the drawings with details that will be readily apparent to those of ordinary skill in the art having benefit of the description herein.

DETAILED DESCRIPTION

For the purpose of promoting an understanding of the principles of the invention, reference will now be made to the embodiment illustrated in the drawings and specific language will be used to describe the same. It will nevertheless be understood that no limitation of the scope of the invention is thereby intended, such alterations and further modifications in the illustrated system, and such further applications of the principles of the invention as illustrated therein being contemplated as would normally occur to one skilled in the art to which the invention relates.

It will be understood by those skilled in the art that the foregoing general description and the following detailed description are exemplary and explanatory of the invention and are not intended to be restrictive thereof.

Reference throughout this specification to "an aspect", "another aspect" or similar language means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the present disclosure. Thus, appearances of the phrase "in an embodiment", "in another embodiment" and similar language throughout this specification may, but do not necessarily, all refer to the same embodiment.

The terms "comprises", "comprising", or any other variations thereof, are intended to cover a non-exclusive inclusion, such that a process or method that comprises a list of steps does not include only those steps but may include other steps not expressly listed or inherent to such process or method. Similarly, one or more devices or sub-systems or elements or structures or components proceeded by "comprises...a" does not, without more constraints, preclude the existence of other devices or other sub-systems or other elements or other structures or other components or additional devices or additional sub-systems or additional elements or additional structures or additional components.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. The system, methods, and examples provided herein are illustrative only and not intended to be limiting.

Embodiments of the present disclosure will be described below in detail with reference to the accompanying drawings.

Referring to **Figure 1**, illustrates a block diagram of a system for analyzing Tamil tweets into positive and negative sentiment is illustrated in accordance with an embodiment of the present disclosure. The system 100 includes a pre-processing unit 102 configured for removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts.

In an embodiment, a natural language processing unit 104 is in connection with the pre-processing unit 102 for dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features. Sentences are indexed based on the dictionary thereby converting indexed sentences vector real numbers.

In an embodiment, a prediction unit 106 is equipped with deep neural network is associated with the natural language processing unit 104 for sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment. The Tamil tweets about the movies and other movie components are taken for prediction.

In an embodiment, the pre-processing include removal of English, removal of space and removal of special characters and the like for structuring.

In an embodiment, word dictionary is included in the implementation process, due to which missing index does not occur in any point of vectorization.

In an embodiment, prediction are processed from neural network to machine learning and deep learning, wherein prediction techniques applied for natural language processing with narrow down on regional language still requires an enhancement because style, slang and words used are invariably unique to person to person.

In an embodiment, the prediction unit 106 is examined in terms of dictionary size, vector dimension, number of hidden layer and batch size.

In an embodiment, the vector size of the word is equalized to the size of the dictionary and the dictionary includes all possible words used in Tweet, wherein in Tamil some of the statements is having peculiar or rare but strong supporting words that are based on the expertization in language.

In an embodiment, the deep neural network is trained with the tokens segregated from Tamil review comments and tweets with five layers of network with a defined dense number.

Figure 2 illustrates a flow chart of a method for analyzing Tamil tweets into positive and negative sentiment in accordance with an embodiment of the present disclosure. At step 202, the method 200 includes removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts.

At step 204, the method 200 includes dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features.

At step 206, the method 200 includes indexing sentences based on the dictionary thereby converting indexed sentences vector real numbers.

At step 208, the method 200 includes performing sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment using deep neural network.

In an embodiment, the dictionary is formed with individual tokens including emoji's and stop words.

In an embodiment, natural language process is configured with a dense set of vocabulary and a well-equipped learning architecture.

Figure 3 illustrates a process flow to classify Tamil tweet with emoji in accordance with an embodiment of the present disclosure. The dataset contains various social media contents required for a preprocessing step to remove the noise from the review comments because the data are in an unstructured format due to the presence of special characters, symbols and other language fonts. Therefore to structure it, preprocessing steps include removal of English, removal of space and removal of other special characters ((), [] / @ # ? *) and the like. The preprocessed data in the text document will be divided into individual tokens for the purpose of dictionary creation. Here tokens are referred to as words or features.

In corpus, it contains 1015 Tamil tweets with various special characters, English fonts and spaces are removed and a dictionary will be created with the 7933 individual tokens including emoji's and stop words. The sample sentence of the Tweet before preprocessing and after preprocessing have been given in Table 1.

The sentences are indexed based on the dictionary. Indexed sentences are converted into vector real numbers. The whole word dictionary is included in the implementation process, due to which the missing index does not occur in any point of vectorization.

The prediction methodologies are processed from the neural network to machine learning and deep learning. All the methods are enhancing and improving the accuracy of prediction. The prediction techniques applied for natural language processing with the narrow down on regional language still requires an enhancement because the style, slang and words used are invariably unique to person to person. Due to this nature some research is happening to predict the author of a book by analysis the style of expression, words used and content handled in it. Tamil words are morphologically rich and also have similar synonyms. The NLP methodology needs a dense set of vocabulary and also needs a well-equipped learning architecture. The proposed work merges the above two challenges by developing a dictionary with more than 7000 words and experimenting with the Tamil Tweet prediction using Deep Neural Network.

Figure 4 illustrates a five layer deep neural network architecture to classify Tamil tweet review with an emoji in accordance with an embodiment of the present disclosure. Opinion mining refers to the process of determining whether the given text is positive or negative. The aim of this article is to determine the best approach to do the sentiment analysis in Tamil text. The corpus is created with the preprocessed Tamil tweets gathered from the social media. The features or word tokens are converted into vectors or numerical values to train the DNN model to classify the comments into positive and negative.

The layers are fully connected (dense) by the neurons in the network layer. Each neuron in a layer receives input from all the neurons present in the previous layer referred to as densely connected. Therefore the learning takes place from the combination of all features(words) from the previous networks.

All the internal layers are built with Rectified Linear Unit (ReLU) activation function to handle optimal number of neurons and it also balances with input shape. The internal structure of proposed DNN model is shown in Figure 4. The model is built with four hidden layers. The output layer is structured using Sigmoid activation function which handles the binary class sentiment comment given against the Tweets. The variation and depth usage of vocabulary knowledge could be identified in Tamil Tweets, due to the high synonyms feature. The multilayer could handle the convex region accurately, have been concentrated more and the model has been built with four hidden layers. The dense of the first two hidden layers are kept as higher than the last two hidden layers. This structure optimally sends the output to the next layer. The loss rate is also fixed with the gradual reduction from the first layer to fourth layer.

The model is examined in terms of Dictionary size, Vector Dimension, Number of hidden layer and batch size. The dictionary has been formed from words used in the Tweets taken for analysis. The vector size of the word is equalized to the size of the dictionary. And the dictionary also included all possible words used in Tweet. In Tamil some of the statements are peculiar or rare but strong supporting words. It is based on the expertization in language. In such case the dictionary with all possible words will yield good accuracy.

The stop words also have a direct relation with sentiment. The gradual increase of dictionary size has been tested which shows the improvement in accuracy of prediction.

For evaluation of the system, 1015 Tamil tweets are segregated into 7933 tokens. The DNN model is trained with these tokens with the five layers of network with dense number as 50. The accuracy of the DNN model is having the direct proposition of the dense and number of hidden layers defined. Here the model is trained with the various batch sizes (32,64,128,256,512) and it is noted that the higher batch sizes lead to lower training accuracy with recovering lost accuracy from the larger

2021104534 24 Jul 2021

batch size by increasing the learning rate. The level of accuracy increases when the size of the word dictionary is kept with the maximum possible Tamil words. The implemented method of keeping the emoji which has been used to express sentiment is also considered as a tokenized vector, it shows the fraction of accuracy changes from 2 to 3%. Batch Size - (32,64,128,256,512), Training Accuracy(%) - (86, 84.5, 82.9, 79.3, 76.3). The Accuracy of the model increases gradually in each epoch and saturates in the 10th epoch. At 10th epoch the DNN model attains the training accuracy as 99% and testing accuracy as 77%.

The drawings and the forgoing description give examples of embodiments. Those skilled in the art will appreciate that one or more of the described elements may well be combined into a single functional element. Alternatively, certain elements may be split into multiple functional elements. Elements from one embodiment may be added to another embodiment. For example, orders of processes described herein may be changed and are not limited to the manner described herein. Moreover, the actions of any flow diagram need not be implemented in the order shown; nor do all of the acts necessarily need to be performed. Also, those acts that are not dependent on other acts may be performed in parallel with the other acts. The scope of embodiments is by no means limited by these specific examples. Numerous variations, whether explicitly given in the specification or not, such as differences in structure, dimension, and use of material, are possible. The scope of embodiments is at least as broad as given by the following claims.

Benefits, other advantages, and solutions to problems have been described above with regard to specific embodiments. However, the benefits, advantages, solutions to problems, and any component(s) that may cause any benefit, advantage, or solution to occur or become more pronounced are not to be construed as a critical, required, or essential feature or component of any or all the claims.

WE CLAIM

1. A system for analyzing Tamil tweets into positive and negative sentiment, the system comprises:

a pre-processing unit for removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts;

a natural language processing unit configured for dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features;

wherein sentences are indexed based on the dictionary thereby converting indexed sentences vector real numbers; and

a prediction unit equipped with deep neural network for sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment.

2. The system as claimed in claim 1, wherein the pre-processing include removal of English, removal of space and removal of special characters and the like for structuring.

3. The system as claimed in claim 1, wherein word dictionary is included in the implementation process, due to which missing index does not occur in any point of vectorization.

4. The system as claimed in claim 1, wherein prediction are processed from neural network to machine learning and deep learning, wherein prediction techniques applied for natural language processing with narrow down on regional language still requires an enhancement because style, slang and words used are invariably unique to person to person.

5. The system as claimed in claim 1, wherein the prediction unit is examined in terms of dictionary size, vector dimension, number of hidden layer and batch size.

6. The system as claimed in claim 1, wherein the vector size of the word is equalized to the size of the dictionary and the dictionary includes all possible words used in Tweet, wherein in Tamil some of the statements is having peculiar or rare but strong supporting words that are based on the expertization in language.

7. The system as claimed in claim 1, wherein the deep neural network is trained with the tokens segregated from Tamil review comments and tweets with five layers of network with a defined dense number.

8. A method for analyzing Tamil tweets into positive and negative sentiment, the method comprises:

removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts;

dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features;

indexing sentences based on the dictionary thereby converting indexed sentences vector real numbers; and

performing sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment using deep neural network.

9. The method as claimed in claim 8, wherein the dictionary is formed with individual tokens including emoji's and stop words.

10. The method as claimed in claim 8, wherein natural language process is configured with a dense set of vocabulary and a well-equipped learning architecture.

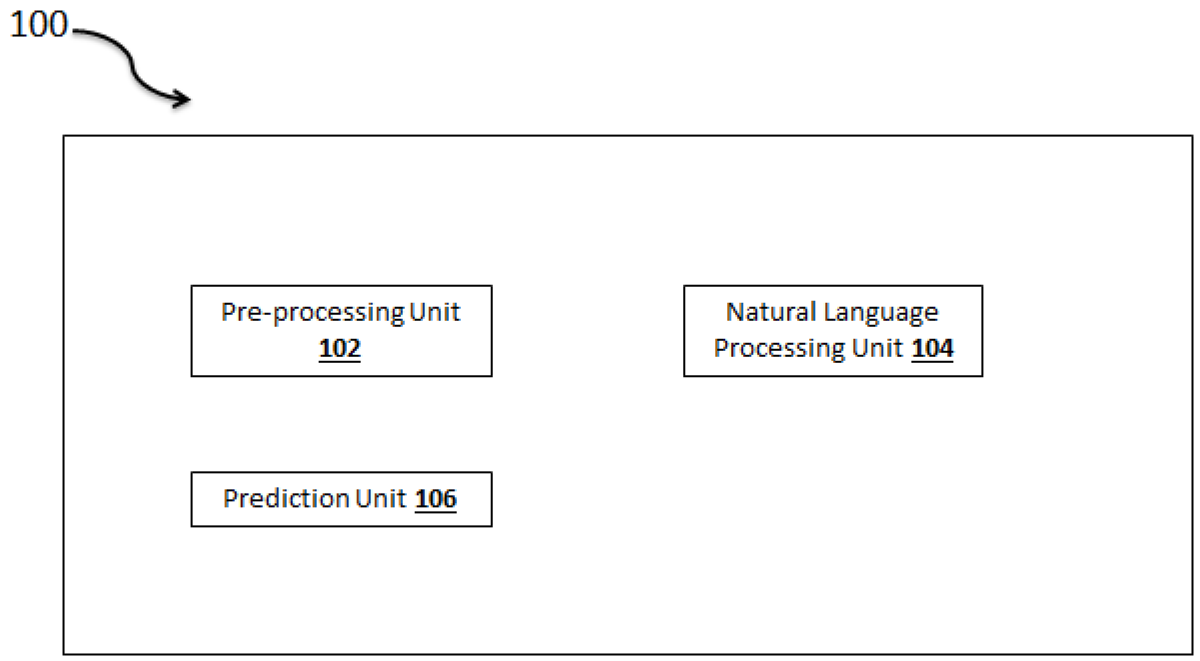
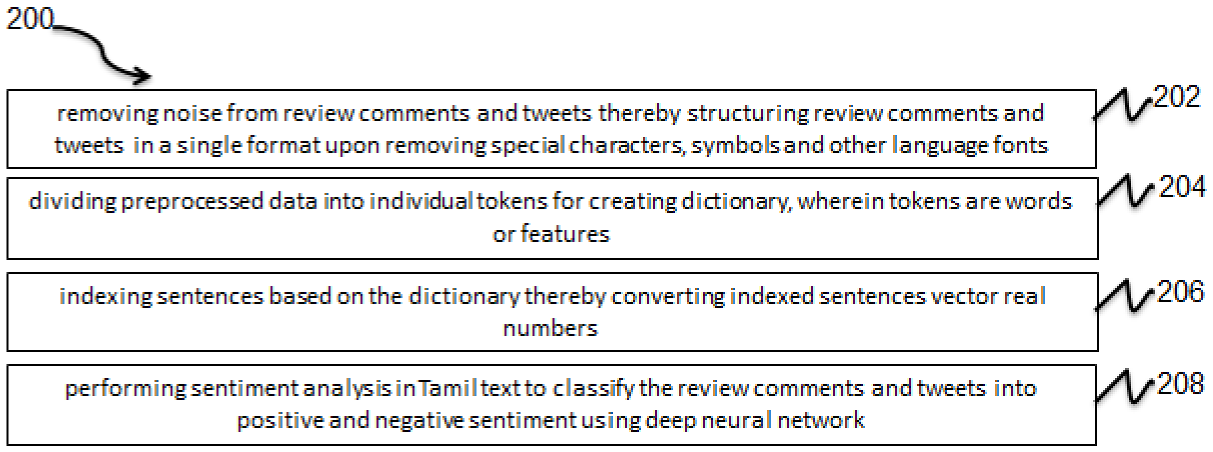

100
Pre-processing Unit 102
Natural Language Processing Unit 104
Prediction Unit 106

Figure 1



removing noise from review comments and tweets thereby structuring review comments and tweets in a single format upon removing special characters, symbols and other language fonts

dividing preprocessed data into individual tokens for creating dictionary, wherein tokens are words or features

indexing sentences based on the dictionary thereby converting indexed sentences vector real numbers

performing sentiment analysis in Tamil text to classify the review comments and tweets into positive and negative sentiment using deep neural network

Figure 2

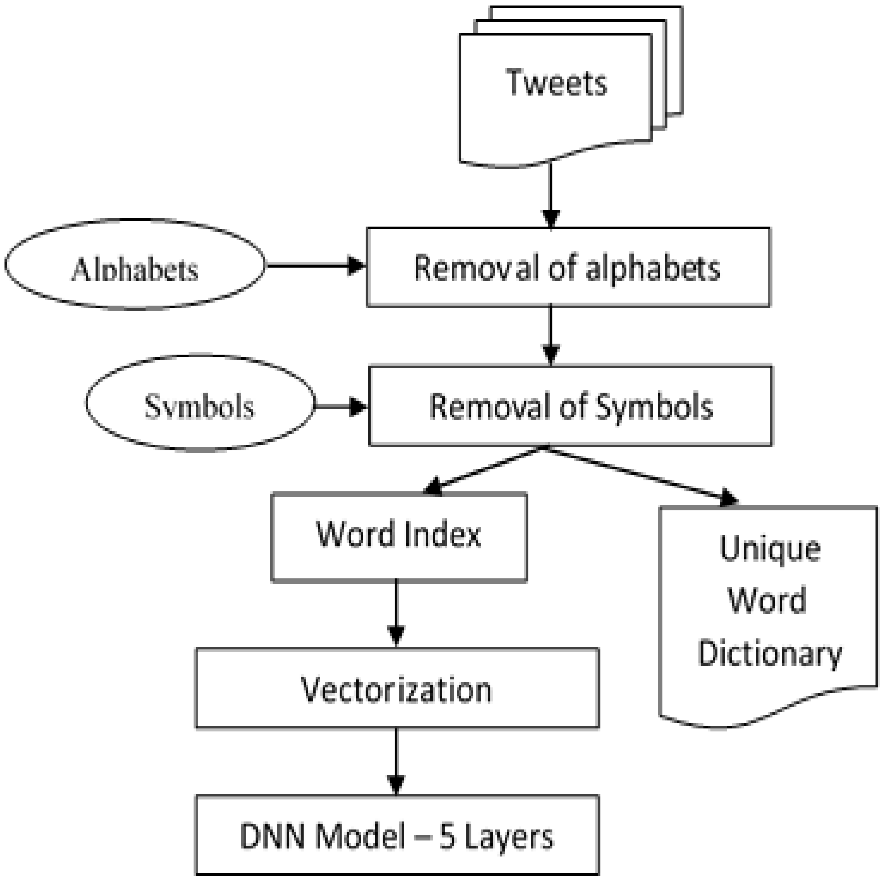

Tweets
Alphabets
Removal of alphabets
Symbols
Removal of Symbols
Word Index
Unique
Word
Dictionary
Vectorization
DNN Model – 5 Layers

Figure 3

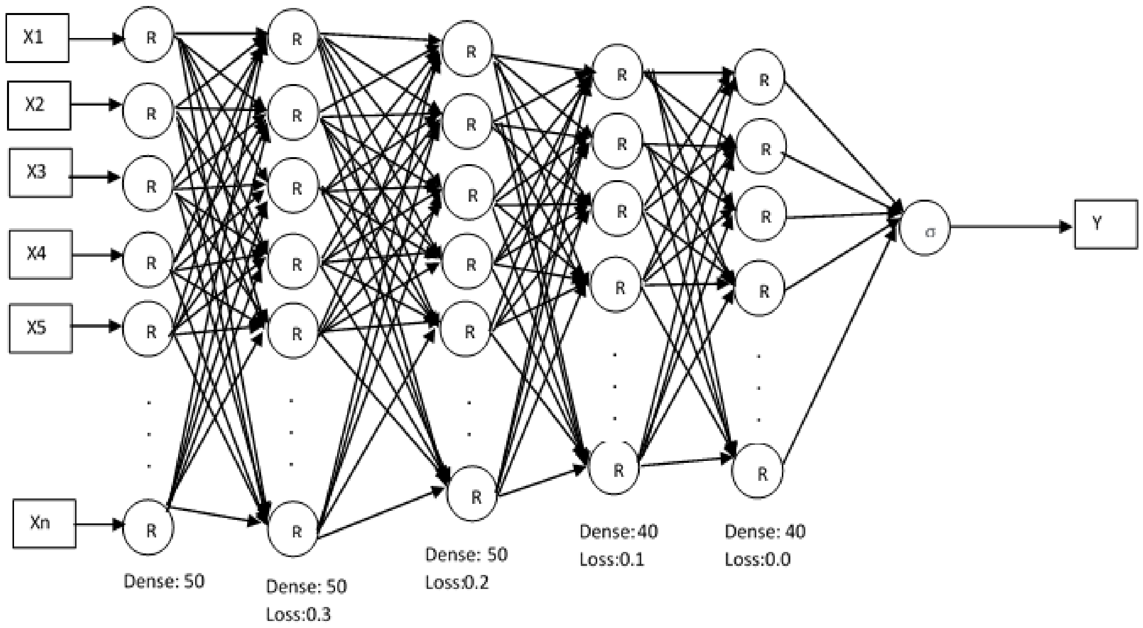

X1
X2
X3
X4
X5
Xn
R
Dense: 50
Dense: 50
Loss:0.3
Dense: 50
Loss:0.2
Dense:40
Loss:0.1
Dense: 40
Loss:0.0
σ
Y

Figure 4

Table 1: Sample Tweet Before and After pre-processing

S. No	Before Pre-processing	After Pre-processing
1	"RT LavanPath: எதை நாம் இங்கு கொண்டு வந்தோம், எதை நாம்	எதை நாம் இங்கு கொண்டு வந்தோம் எதை நாம்
2	நிலை மாறுமா.? நிஜம் சேருமா.?	நிலை மாறுமா நிஜம் சேருமா

Figure 5