How to Cite:
Ledalla, S., Rao, G. A., & Sesetti, . A. (2022). Sentiment analysis of Hinglish reviews using hybrid approaches. *International Journal of Health Sciences*, *6*(S2), 5432–5445. https://doi.org/10.53730/ijhs.v6nS2.6363

Sentiment analysis of Hinglish reviews using hybrid approaches

Sukanya Ledalla
Assistant Professor, Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, JNTUH, Telangana, India
Corresponding author email: ledalla.sukanya@gmail.com

Dr. G. Appa Rao
Professor, Department of Computer Science and Engineering, GITAM Institute of Technology, Visakhapatnam, Andhra Pradesh, India

Anuradha Sesetti
Assistant Professor, Department of Computer Science and Engineering, GITAM Institute of Technology, Visakhapatnam, Andhra Pradesh, India

***Abstract*---**People are using social media to communicate their views and opinions on several topics, and it is becoming more popular as a medium of communication. Nowadays people are curious to hear about a film review before they watch it for a few days. A summary of all reviews for a film can help users make their decision by not wasting time reading all reviews. Several websites can have summery because of the rise of social media.As a result, the words sentiment analysis and text analysis developed their paths to becoming important computational linguistics and text analysis components. Opinions are articulated in India in recent years, using multi-lingual terms. In the field of sentiment analysis, this has become a new challenge. Machine learning methods, such as neural networks, have proven effective in this task; however, there is space for advancing to networks with greater accuracy.In this investigation,U. S. English, Hindi dialects, and datasets like Twitter sentiment corpus, Amazon Product Reviews, IMDB Movie reviews, and Several public opinion datasets have been used in this study. The proposed network may achieve the greatest known accuracy of 92.25 percent. As an outcome, the success of the suggested network can be duplicated in various fields.

***Keywords*---**multilingual sentiment analysis, text based convolutional, neural network, IMDB movie reviews, sentiment analysis.

Introduction

Sentiment analysis is the technique of extracting and identifying subjective information in source materials using natural linguistic analysis, text analytics, and computational linguistics [1]. Its goal is to determine a speaker's or author's attitude toward a subject or a text's overall contextual polarity. The perception may be,

- His decision or his assessment.
- Emotional state (the author's emotional reactions at the time of posting).
- The emotional contact intended (that is, the author's emotional impact Wish es to get the reader on).

With the rise of the Internet combined with the popularity of endless digitally colle cted data, an analysis of numerouspublic reviews on various issues has demonstr ated its importance in the fields of marketing, politics etc. With the overwhelming number of public issues that exist, reviews and ratings have become important to analyze. The growth of online platforms, which including blogs and social networks, has sparked an attention in sentiment analysis throughout the previous decade. With the development of comments, reviews, ideas, and other forms of online communication for companies wanting to sell their products, find new possibilities, and retain their legitimacy, web opinions has become like a virtual money. Let us consider the entertainment market as an example.

Most people aren't willing to waste their time watching thousands of reviews that are created annually. Consequently, methods have arisen to help people tochoose the movies that will be worth watching. A movie review or rating can convey one person’s opinion, but thousands of reviews and ratings can givea clear understanding of the unified sentiment for a movie. This is true for movie reviews as much as it is true for music reviews and other products or services; however, movie reviews have become notoriously popular and abundant. Although ratings easily reflect numerical (e.g., 0-10) or symbolic (e.g., 1-5 stars) view sentiment, feedback cannot be analytically expressed easily.
To attain a sentiment rating, reviews are textual and thus involve a more sophisticated view; this method is called sentiment analysis or, more precisely, classification of sentiment. The purpose of the classification of sentiment is to translate textual data on a defined scale into a meaning.
The sentiment is usually graded as either positive or negative, but as an intermedi ate, some techniques use neutral emotion. A version of learning for the framework also involves the process of sentiment analysis to mirror our understanding of te xtual content.

Sentiment Analysis has a variety of goals, the most prominent of which are:

- examine an entity's social attitude as stated in a multilingual statement.
- provide accurate public review analysis.
- increase prediction performance and avoid prediction error.

According to Pew Research's Statistical Portrait of Hispanics in the United States [2], the Hispanic population in the United States made up 17.3 percent of the total population in 2014.

The researchers encountered a significant challenge in analyzing multilingual data for sentiment analysis.Many languages are a part of Indian culture. Their family speaks Hindi at home, but other social engagements demand English, thus they frequently switch between the two languages.

For example @fawadchaudhry
'अच्छा app very खुश'
model.predict(vec.transform(['अच्छा app very खुश']))
array([0], dtype=int64)
model.predict(vec.transform(['बुरा app very नाराज']))
array([0], dtype=int64)
0 indicates negative sentiment, 1 indicates positive sentiment.

The result of analysing the given sentences is a sentiment of 0. Even if the statement is positive, the algorithm predicts a negative sentiment because Hindi terms are not taken into account. When both Hindi and Englishterms are taken into account and sentiment analysis is performed, the following is the accurate result:

translations = translator.translate('अच्छा app very खुश')
model.predict(vec.transform([translations]))
array([1], dtype=int64)

To address the above challengewe created the proposed model based on this concept a novel classification technique Hinglish Multilingual Sentiment Analysis using Machine Learning and Legion Kernal Convolutional Neural Network is proposed. There is approximately 15 to 20 percent difference in the results when we consider only English and Hinglish sentences.

Machine learning is a growing field of computer science; this is a simple term that refers to any case in which a computer is needed to gain an understanding of the data that is fed. Moreover, machine learning uses data to generate an unknown purpose, as opposed to traditional computer programs that obey sequential commands to produce a known or desired output. Usually, machine learning requires unnecessary data to train on problem-solving. After training, the model can be reused and probably further trained for future data [3] .

From Naïve Bayes to Convolutional Neural Networks, there are various machine learning techniques. Many of these approaches have proven immensely successful in classifying text and sentiment.When it is applied more profoundly, machinelearning may also be defined as deep learning. Convolutional Neural Networks (CNNs) are a type of Deep Learning algorithm that takes an input image and assigns significance (learning connection weights) to different attributes in the picture, enabling it to differentiate between them. Convolutional neural networks [8] are one of the most common neural network models used during

computer vision to object recognition and patterns in images. But In thispaper we have proved that CNN can also be used on text with high accuracy.

Methodology

We have worked on different datasets as Twitter data set, IMDB dataset, Amazon Product revies data set, reviews of android apps on Google Play-Store.

model 1: Multilingual sentiment analysis of hinglish tweets

This model will provide novel sentiment analysis tactics and approaches, as well as further information on Hinglish as a sentiment analysis language. The datasets, dictionaries, and sentiment analysis packages utilized in the three trials are then described. The trials will provide statistical analysis, allowing one to draw conclusions and analyze the findings of the current work, which claims that when preprocessing and resources relevant to the Hinglish language are used, Hinglish reviews will be more accurately classified.

The basic structure proposed is illustrated in Figure 1.

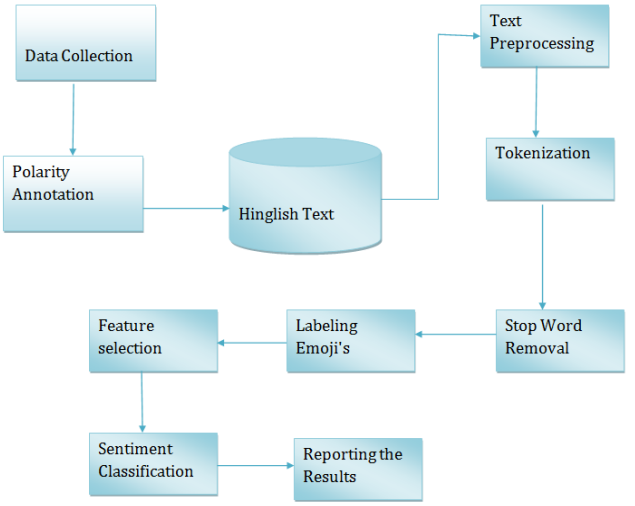


Figure-1.The basic structure of the proposed Model

The steps in the suggested model are as follows:

Step 1: Data Set Creation

We prepared a CSV file containing the entire training data set, which included 34,000 reviews. We are utilizing 80% reviews to train the modeland 20% reviews to test the model.

Step 2: Pre-processing

Text Preprocessing: Many unnecessary words, such as stop words, misspellings, and slang, can be found with typical textual and document data sets.

Text Cleaning and Pre-processing contains the following operations

Tokenization
Stop Words
Stemming

Tokenization is the process of breaking down a sentence into smaller bits, known as tokens, while also removing specific characters, such as punctuation. Tokenization is demonstrated in the following example:

The DaVinciCode is poorly written and boring						
The	DaVinciCode	is	poorly	written	and	boring

Stemming

The purpose of stemming is to reduce a word's various forms, including derivationally related forms, toa single base form.
For example: am, are, is be
car, cars, car's, cars' is car

Stop words: removing common keywords
Some exceedingly common words that appear to be of little value in assisting in the selection of texts that meet a user's needs are sometimes completely removed from the vocabulary. These are called Stop.
An example of a stop list is {A, an,is, has, he, is, it, in, be, by, etc}

Step 3: Extract Feature Vectors

We can partition each review into words in the training data and add each word to the feature vector. Some of the terms may not accurately reflect a tweet's sentiment, so we can filter them out.
Then merge each feature vector into a big list that includes all of the features and delete any duplicates.

Step 4: Create a classifier and use it to classify test tweets

The classification method seeks to discovera hyperplane that best splits the data into two classes. However, there are two types of sentiment classes (positive, negative)

Step 5: Sentiment Analysis Metrics and Evaluation

Performance metrics can be used to evaluate a classifier and determine how accurate a sentiment analysis model is. Label each emoticon in the complete emoji list for positive, negative sentiment. Obtain a positive sentiment score and a negative sentiment score for each word. Calculate the sentiment score for each word by taking the average of the score among all words in the tweet.

```
neg=(100*len(ntweets)/len(tweets))
pos=(100-(100*len(ntweets)/len(tweets)))
```

If an average value is greater than zero, consider the tweet as a positive tweet score, and if it is less than zero then consider it as a negative tweet score.
Return the positive and negative scores
The following is the procedure to translate the hindi dialect to English.

Define transliterations_English_Hindi (token,pos)

Begin:

1. If the token is in Hindi word # file contains key/value pair of Hindi & English words
   a) Open Crowd_transliteration file
   b) if token present in the file
   c) Read value of the token
2. return token,pos
3. Else ignore

End
Hinglish Sentiment Analysis algorithm achieved the accuracy of 77.43 percent on the considered sentiment dataset.

Model 2:Legion Kernel Convolutional Neural Network with LSTM ModelConvolution

People's thoughts are also expressed in a multilingual language in India. In the realm of sentiment analysis, it has become a new difficulty. In this attempt, machine learning techniques have shown to be effective. Furthermore, there is room to progress to higher-accuracy networks. As a consequence, we developed a new sentiment analysis system using a Convolutional Neural Network with Long Short-Term Memory (LSTM). Convolutional neural networks (CNNs) are indeed a sort of deep learning model that is frequently used for document classification. Despite their beginnings in image processing, CNNs have been effectively used for text summarization. Data is transferred through a deep learning model's input layer, which consists of two neurons, input, and output. The data is transferred from the input layer to the hidden layers, where it is analyzed on different levels and features.

Convolution Neural Network

(ConvNets or CNNs) Convolution layers neural networks are a type of neural network that has several layers of Convolutional processes.

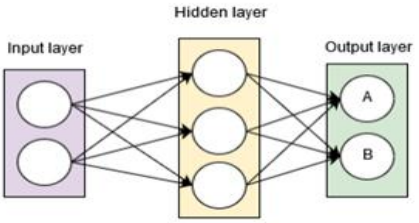


Figure-2. 2-layer ConvNets

CNN, as a person, is unable to grasp the statements. CNN's can recognize numerous words at once and comprehend the various combinations in text classification. Different combinations will trigger different emotions; a word's sentimental value may be lost when removed from its context. As a result, convolutional layers can be used in the same way as n-grams can. If the sequence is long enough, they will have difficulties transporting knowledge from earlier time steps to later ones. If you're attempting to predict something from a paragraph of text, NNs may omit crucial data at the beginning. We deployed Long Shot Term Memory to solve this problem (LSTM) [6]. The CNN layers were fine-tuned to recognize word pairings in a variety of branches. The information was used by the LSTM layers to extract sequential information from the CNN's knowledge.

Long Short-Term Memory

Contextual information can also be extracted from text using LSTM layers, but only in a sequential fashion. The convolutional output of each branch is finally processed by the branch's LSTM layer [7]. Even though the LSTM's inputs are convolved and processed, the information is still sequential text deformation. The LSTM looks at the overall meaning of reviews, sentences, and phrases, as well as the relative meaning of words. It should be noted that if the convolution layer of a branch has a kernel size of 1, the LSTM layer will train on non-convolved words.

Experimentation

Model 2.1

A kernel is used in typical convolutional networks to scan over the width and height of the data; however, when words concatenated are split apart, their meanings are lost. This problem is solved by matching the kernel to the embedding width in single-dimensional convolutional layers. A 1-dimensional convolutional layer with a kernel size of 3 will, for example, evaluate inputs having a 32x3 window shape. As a result, the kernel size represents the number of words that will be convolved at a given moment.

Best accuracy coming from the combined usage of unigrams, bigrams, and trigrams [4] motivated the use of many branches. The size of the initial convolutional layer kernel defines each branch. Each branch will analyze a different combination of word sizes. The network will work together to detect patterns in various-sized phrases and understand contextual information. For example, while the word "bad" is inherently negative, the double negative "not bad" communicates a positive feeling. “Not bad is a lie,” on the other hand, will turn the sentiment negative.CNN is unable to understand the statements like a person would. Word2vec handles it by converting text into a vector that CNN can interpret. In an n-dimensional vector space, each phrase, which is an array of words, is converted to vectors and compared to other vectors' lists of words [5].

Example

Another one of the users stated that after about one episode of Oz, you'll be hooked. They are correct, as this is precisely what occurred in my case. The first factor that influenced me regarding Oz was its ugliness and unabashed depictions of violence, which began immediately away. It was not a presentation for the fainthearted or the timid. When it comes to drugs, sex, and violence, this programme doesn't hold back. It's hardcore in the traditional sense of the term.

Vector Representation

array([[0, 0, 0, ..., 6, 6473, 1115], [0, 0, 0, ..., 3073, 2, 1496], [0, 0, 0, ..., 2289, 7, 7], ..., [0, 0, 0, ..., 18, 1, 378], [0, 0, 0, ..., 149, 120, 307], [0, 0, 0, ..., 267, 141, 824]])
Convolutional Neural Networks have the following Layers:

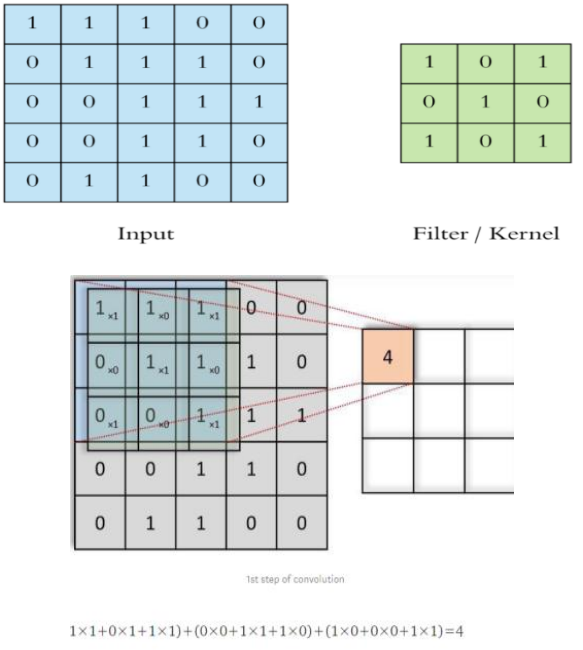
- ✓ Convolution
- ✓ ReLu Layer
- ✓ Pooling
- ✓ Flattering
- ✓ Fully Connected

Convolution Layer

This layer conducts a dot product among two matrices, the first is the source matrix another is the kernel matrix. The kernel glides from across length and breadth of the information during its forward pass, providing the information depiction of that focused area. The convolution operator in neural networks is as follows

$$O(i,j) = (I \star K)(i,j) = \sum_{k} \sum_{l} I(i+k, j+l) K(k,l)$$

I is the input and K is called the kernels



The Resultant Matrix is :

4	3	3
2	4	3
2	3	4

Rectified Linear Unit (RLU)

ReLU is a type of linear unit that has been performed an element-by-element procedure, setting all negative pixels to zero.
The function f(k)=max (0,k) is computed.ReLu enhances neural networks by accelerating training.

Pooling Layer

This layer turns each input kernel size into a single maximum observed number output. The function for pooling is*p (input height ÷ p).*By reducing the spatial size of the representation, we can reduce the amount of computation and weights required in this stage. After Performing Max Pooling Operation the Result will be:

4	3	3
2	4	3
2	3	4

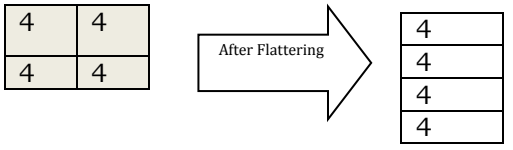
Max Pooling with 2X Stride

Max(4,3,2,4)

4	4
4	4

Flattering

You'll have to flatten the consolidated featured map once we've gotten it. The full pooling layer matrix is then flattened into a single column and fed to a neural network for computation.



Concatenation

Concatenation is used to bring the branches together at the end. The outputs of the LSTM layers are merged in an array. The total of the outputs of all the branches (l x b) in the form of this layer's output.

Fully connected layer

The previous tensor is one of the parameters of the fully connected function, as is the number of classes to predict (negative and positiveIt "demolishes" the output of each neuron into an even range between 0 and 1. The likelihood that the information corresponds to a certain class is contained in the new vector, suggesting that it includes the probabilities for all categories to forecast. So, if the output is 1, the predicted class will be positive; otherwise, it'll be negative.

Experimentation

All of the model variations were created with the Keras 2.0.4 learning library with TensorFlow 1.1.0* as the backend for the GPU. Through multiple testing phases, two high-performing models were constructedusing Keras. The networks vary in specific structure and parameters, but both produce a state of the art accuracies on IMDB translation pairings, an English word-frequency list, a Hindi word-frequency list, and a number of public opinion datasets. The parameters considered for experimentation are branch dropout, merged dropout, learning rate, and learning decay.

Model 2.1

Convolutional layers with 32 filters and kernels of three, four, and five sizes are used in the initial high-accuracy model. Despite their small size and number, these combinations are good at extracting sentiment from words and phrases.
The proposed model 2.1 is depicted in figure 3

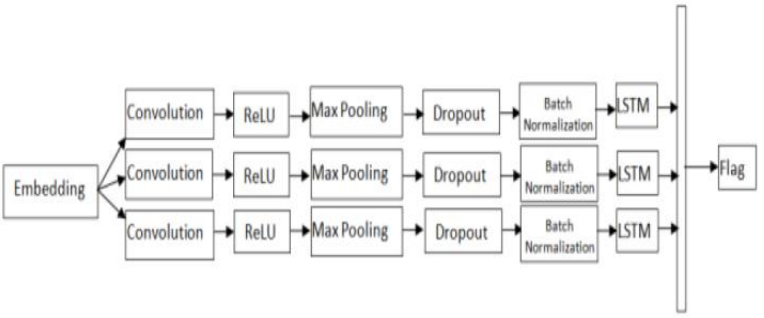


Figure-3 Proposed Model 2.1

To generalize the learning, each branch in Model 2.1 comprises regularization layers following the CNN layer. ReLU activation, Max-Pooling of size 2, and Batch Normalization are the layers in order. Before being merged and fully connected to the label output, the branches are processed by 100-unit LSTMs [8]. The Adam optimizer is employed, with a learning rate of 0.001.Model 2.1 components and their initial values are tabulated in table1.

*https://faroit.github.io/keras-docs/2.0.4/

Model Component	Model 1
Convolutional Filter Sizes	3 + 4 + 5
Convolutional Filters	32
Branch Dropout	0
LSTM	100
Merged Dropout	0
Learning Rate	0.001
Learning Decay	0

Table -1. Model 2.1 components and initial values

Model 2.2

The second model increases the number of convolutional kernels from two to seven. The proposed model 2.2 is depicted in figure 4.

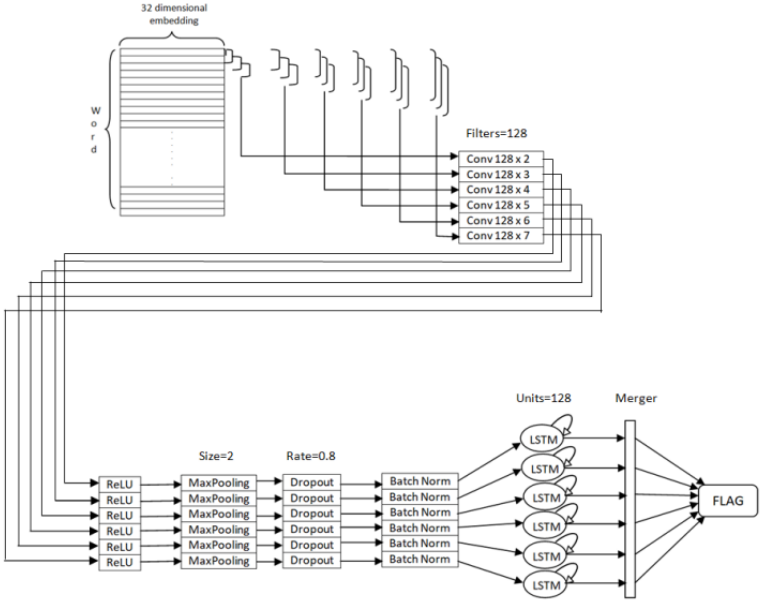


Figure-4. Proposed Model2.2

In addition to having additional branches, the network uses 128 filters and 128 LSTM units for each branch. To deal with the increased complexity and number of weights, a single dropout layer with a rate of 0.8 is added after all the branches are merged. An Adam optimizer with a rate of 0.001 is also used by the network. Table 2 lists the components of Model 2.2 as well as their beginning values.

Model Component	Model 2
ConvolutionalFilter Sizes	2+3+4+5+6+7
Convolutional Filters	128
Branch Dropout	0.5
LSTM	128
Merged Dropout	0.8
Learning Rate	0.01
Learning Decay	0.1

Table-2. Model2.2 components and initial values

Analysis of Results

All of the suggested models perform well on the IMDB, translation pairings, an English word-frequency list, a Hindi word-frequency list, and a number of public opinion datasets with accuracy variations of less than half a percent.

	English Data Set Results when considering only English Terms)		Hinglish Data Set (Results when considering both the Hindi & English Terms)	
	Positive reviews	Negative	Positive reviews	Negative reviews

	percentage	reviews percentage	percentage	percentage
Twitter Data Set	52.45	47.5	35.7	64.28
Google-Play-store-apps-reviews	62	38	75	25
Amazon Product Reviews	93.39	6.62	94.37	5.62

Table-3. Analysis of Results

Despite various changes, the models produced show very minor improvements. On the considered sentiment dataset, Model 1 had an accuracy of 77.43. Model 2.2 had a maximum accuracy of 92.25 percent. Model 2.1 came with a score of 90.75 percent accuracy. Basic structures were developed during the experiments and then tweaked to produce high-performing models. While the primary goal was to improve performance, the only method to do so was to avoid deterioration.

Specific parameters that generate the best results were determined through a process of trial, error, and tuning. The convolutional kernel sizes were the most influential factor. Models 2.2 shows that a larger kernel size is more beneficial while also demonstrating that having too many kernels reduces accuracy and increases overfitting. Although the number of filters was less important, 128 filters were sufficient for networks with a larger range of kernel sizes. Higher maximum pooling sizes, on the other hand, lost potentially relevant information. If the number of units was too high, the LSTM layers would swamp the network, hence the value was adjusted at 100 or 128.

Conclusion

On the Hinglish review sentiment dataset, the Convolutional LSTM with pooled Kernels network performed well, indicating the network's capacity to learn textual information. Despite the limitations of the Hinglish dataset, the networks that include numerous, well-established approaches provide the highest accuracy. Despite the study of several qualified machine learning algorithms, the combination of convolution and memory is capable of approaching a problem with several analyses. The CNN layers have been fine-tuned to recognize word combinations in various branches.The information was used by the LSTM layers to extract sequential information from the CNN's knowledge. The only difficulty came in preventing the start of degeneration as training progressed. The ultimate network (Model 2.2) can pursue a comprehensive comprehension of the data by incorporating numerous regression algorithms.

References

1. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, Inception- ResNet and the Impact of Residual Connections on Learning. In AAAI (pp. 4278- 4284).
2. Pew Research. “Statistical Portrait of Hispanics in the United States.” Internet. 2016.
3. J Sirisha Devi, Siva Prasad Nandyala, P Vijaya Bhaskar Reddy (2019). A Novel Approach for Sentiment Analysis of Public Posts. In: Saini H., Sayal R.,

Govardhan A., Buyya R. (eds) Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 32. Springer, Singapore.

4. Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications, 57, 117-126.
5. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.
6. Zhang, K., Chao, W. L., Sha, F., & Grauman, K. (2016, October). Video summarization with long short-term memory. In European Conference on Computer Vision (pp. 766-782). Springer International Publishing.
7. Medel, J. R., & Savakis, A. (2016). Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv preprint arXiv:1612.00390.
8. Rahman, L., Mohammed, N., & Al Azad, A. K. (2016, September). A new LSTM model by introducing a biological cell state. In Electrical Engineering and Information Communication Technology (ICEEICT), 2016 3rd International Conference on (pp. 1-6). IEEE.