Project Report on

“Sentiment Analysis on Covid-19 Awareness Conversations Using CNN”

Project submitted

In partial fulfillment of the necessities for the degree of

MASTER OF COMPUTER APPLICATION

By

Debarati Saha

RollNo:**001910503023**

RegistrationNo:**149885** of**2019-20**

Under the supervision of

Dr. Dipankar Das

Department of Computer Science and Engineering
Faculty of Engineering and Technology

Jadavpur University
Kolkata – 700032, India

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY, KOLKATA

Certificate of Recommendation

This is to certify that **Debarati Saha**(Reg. No.: 149885 of 2019-2020, Roll No: 001910503023, Exam Roll No: MCA226022) is a student in the Master of Computer Applications course and also the project entitled "***Sentiment Analysis on Covid-19 awareness conversations using CNN***" could be a bonafide record of labor carried out by her, is accepted in partial fulfillment of the necessity for the degree of Master of Computer Application from the Department of Computer Science and Engineering, Jadavpur University during the educational year 2021-2022. She has been ready to follow all the instructions in an exceedingly calm and responsible way and successfully distributed her research work. Wish herall the best for her future endeavors.

Dr. Dipankar Das (Project Supervisor)

Assistant prof., Dept. of Comp. Science & Engineering

Jadavpur University, Kolkata-700032

Prof.Anupam Sinha

Head of the Department, Dept. of Comp. Science & Engineering

Jadavpur University, Kolkata-700032

Prof.Chandan Majumdar

Dean, Faculty Council of Engineering & Technology

Jadavpur University, Kolkata-700032

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY, KOLKATA

CERTIFICATE OF APPROVAL

This is to clarify that the project entitled ***"Sentiment Analysis on Covid19 Awareness Conversations Using CNN"*** has been completed by ***Debarati Saha.*** This work is carried out under the supervision of **Dr. Dipankar Das** in partial fulfillment of the requirements for the degree of ***Master of Computer Application*** of the department of ***Computer Science and Engineering, Jadavpur University,*** during the session *2021-2022*. The project report has been approved as it satisfies the academic requirements with respect to project work prescribed for the said degree.

(Signature of the internal examiner) (Signature of the external examiner)

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY, KOLKATA

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby certify that I am the only real author of this thesis and no part of this thesis has been published or submitted for publication. I certify that, to the most effective of my knowledge, my thesis doesn't infringe upon anyone's copyright nor violate any proprietary rights which any ideas, techniques, quotations, or the other material from the work of people included in my thesis, published or otherwise, are fully acknowledged in accordance with the quality referencing practices. I declare that this is often a real copy of my thesis, including any final revisions, which this thesis has not been submitted for a higher degree to the other University or Institution.

____

(Signature of the Candidate)

Candidate's Name: ***Debarati Saha***
Class Roll No.: *001910503023*
Project Title:Sentiment Analysis on Covid-19 Awareness Conversations Using CNN

ACKNOWLEDGEMENT

With my most sincere gratitude, I would like to thank **Dr. Dipankar Das**, Department of Computer Science & Engineering, my supervisor, for his overwhelming support throughout the duration of the project. His motivation always gave me the required inputs and momentum to continue with my work, without which the project work would not have taken its current shape. His valuable suggestions and diverse discussions have always inspired new ways of thinking. I feel deeply honored that I got this opportunity to work under him.

I would like to express my sincere thanks to all my teachers for providing a sound knowledge base and cooperation.

I would like to thank all the faculty members of the Department of Computer Science & Engineering of Jadavpur University for their continuous support.

Last, but not least, I would like to thank my batch mates for staying by my side when I need them.

Debarati Saha
Roll No.: *001910503023*

ABSTRACT

Natural Language Processing (NLP) has become one of the most significant technologies of the 4th Industrial Revolution and a well-liked field of AI with the rise of voice interfaces and chatbots. The discipline of NLP has produced a wide range of beneficial applications, which are rapidly expanding. They range from basic to intricate. One is sentiment analysis.

Two years back, the whole world was hit by a pandemic, the Covid-19. Many people lost their lives, many people lost their jobs. Schools, colleges, universities, offices all were shut. The whole world came to a standstill. Even today, the effect of Covid has not been erased completely. Daily new cases are being reported. Standing in this situation, if we could get the help of an advisor who can give the necessary solutions would have been a great help. This could also help in spreading awareness about the pandemic. The goal is to create an automated advisor using NLP. A chatbot can be trained over conversations on Covid-19.

For training the chatbot, a large number of conversations are required. Along with it, intent words and sentiments are also required to understand the intention of the user and the depth of the situation. We have developed a dataset where two users (User1 and User2) will take turns to chat among themselves regarding Covid-19. However, we have employed several machine learning and one deep learning model like CNN to explore the insights of sentiment analysis on COVID discussions.

Keywords: Sentiment Analysis, Natural Language Processing, Machine Learning, COVID dialogues.

Table of Contents

1. *INTRODUCTION*

The automatic software manipulation of natural languages, such as speech and text, is known as natural language processing, or NLP for short. NLP has become one of the most significant technologies of the 4th Industrial Revolution and a well-liked field of AI with the rise of voice interfaces and chatbots. The discipline of NLP has produced a wide range of beneficial applications, which are rapidly expanding. Sentiment analysis, a subfield of NLP is becoming popular for its emergence in adding intelligence to machine in a human like way. Thus, once it is added with chat or any communication system, its interactive power enhances in a significant way.

To ascertain the emotional meaning of communications, sentiment analysis is a type of analytical technique that combines machine learning, natural language processing, and statistics.Businesses assess content such as social media posts, online reviews, call center interactions, and consumer messaging using sentiment analysis. Sentiment analysis can track changes in attitudes towards companies, products, or services, or individual features ofthose products or services.

COVID-19 derived from SARS-CoV-2 is spreading dramatically worldwide and causing millions of infections and deaths amongst the human population. Computer technologies provide profound opportunities to fight infectious disease outbreaks and have a remarkable role, especially in sentiment analysis; this importance is due to their tremendous role in analyzing public sentiment.

Thus, in the present project work, the CNN based deep learning approach is employed to investigate the roles of sentiment in a scale of -5 to +5 for COVID related utterances and the results shows that the performance can be improved not only by enhancing the size of the dataset, but to explore with additional clues related to sentiment. The dataset has been developed based on two users 1 and 2 which can be interchangeably used to play the roles of a user and bot in training and can lead to more interesting insights in future attempts.

1.1. CHALLENGES

Firstly, developing large amount of data set which contains conversation on COVIDhas been a major challenge for this database. Also tagging sentiments for thousands of data is very time consuming. A large part of research time has been consumed for preparing dataset.

1.2. GAP IDENTIFICATION

There has not been much research going on dialogue sentiment, especially on COVID-19 data.Our idea was to prepare a model which is able to help to recognize a sentiment automatically whether it is positive or negative or neutral. On this pandemic day, it is very hard for us to keep ourselves mentally healthy. This model could be beneficial to fill that gap and can be of any help.

1.3. MOTIVATION

Two years back, the whole world was hit by a pandemic, the COVID-19. Many people lost their lives, many people lost their jobs. Schools, colleges, universities and offices all were shut. The whole world came to a standstill. Even today, the effect of COVID has not been erased completely. Daily new cases are being reported. Standing in this situation, if we could get the help of an advisor who can give the necessary solutions would have been a great help. This could also help in spreading awareness about the pandemic. The goal is to develop an automated advisor using Natural Language Processing (NLP). A chatbot can be trained over conversations on COVID-19.

The idea is to prepare a chatbot that will understand the sentiment of the user and provide useful information about COVID-19. This can help in spreading awareness among common people about the global pandemic.

For training the chatbot, a large number of conversations are required. Along with it, intent words and sentiments are also required to understand the intention of the user and the depth of the situation. We have developed a dataset where two users (User1 and User2) will take turns to conversant among themselves regarding COVID-19.

1.4. PROBLEM STATEMENT

Firstly, a dataset is to be made with two user utterances. In place of bot another human is chosen for replying. Training different machine learning models using COVID-19 dialogue dataset is to be done and make a note or supervise how the model is able to predict the test sentiments more accurately.

We proposed two types of evaluation strategies e.g., Relaxed and Strict to investigate the roles of sentiments on two user level utterances. The evaluation is also to be done separately on the whole dataset for each model.

1.5. OBJECTIVE

This paper describes about incorporating sentiment analysis of two users. Our goal is to review dialogues of users in the dataset and find out the sentiments and analyze those sentiments.

The main objective of the research is to check the accuracy with results, training the user data and to test part of user data to check other sentimentsare capable of predicting the same or not. The sentiment will be either positive or negative. And similarly, we have to train data and test user data, to check that part of user is capable of predicting the same sentiment or not.

1.6. CONTRIBUTION

The complete dataset annotated with sentiment and intents has been developed by us with above mentioned characteristics.In addition, aCNN based system is designed with highest notable accuracy of 63% on user utterances for classifying sentiments.

2. *RELATED WORK*

In the paper, “Deep learning for sentiment analysis: A survey.” by Zhang, Lei, Shuai Wang, and Bing Liu [1], they have introduced numerous deep learning architectures and their uses in sentiment analysis in this work. For a variety of sentiment analysis applications, many of these deep learning algorithms have produced cutting-edge results. They anticipated that deep learning for sentiment analysis research will continue to advance thanks to recent developments in both theory and applications. This paper initially provides an overview of deep learning before conducting a thorough investigation of the ways it is currently being used in sentiment analysis.

This article "Techniques and applications for sentiment analysis." By Feldman, Ronen [2] addressed many algorithms that attempt to address each of the major research issues within the subject of sentiment analysis. They have also listed a few significant open tasks and outlined some of the main sentiment analysis applications. To avoid these open difficulties, many commercial sentiment analysis systems continue to employ basic methodologies, which negatively impact their performance.

It is informed in the article “CASA: Conversational Aspect Sentiment Analysis for Dialogue Understanding.” by Song, L., Xin, C., Lai, S., Wang, A., Su, J., &Xu, K.[3] the task of conversational aspect sentiment analysis (CASA), which can provide beneficial fine-grained sentiment information for dialogue understanding and planning, is introduced in order to hasten the development in this domain. Overall, this work significantly adapts the usual aspect-based sentiment analysis to the conversational environment. They annotated 3,000 chit-chat dialogues (27,198 sentences) with fine-grained sentiment data, including all sentiment expressions, their polarities, and the corresponding target mentions, to help with the training and evaluation of data-driven techniques. They also annotated a 200-dialogue test set that is outside of the domain to assess robustness. Additionally, we create a variety of baselines for the preliminary investigation based on either pre-trained BERT or self-attention. According to experimental results, their BERT-based model performs well for both in-domain and out-of-domain datasets, and a thorough analysis points to a number of potential possibilities for further advancements.

Regarding our BERT model, it gets respectable results of 78.44 and 79.08 on Sentiment Extraction and Mention Extraction (aspects in common language) sub-tasks, respectively, on the in-domain DuConv test set. On the other hand, we notice considerable declines in performance on the out-of-domain test set, where SE and ME, respectively, see performance drops of about 16 and 15 points.

In the study “Sentiment classification in customer service dialogue with topic-aware multi-task learning” by Wang, J., Wang, J., Sun, C., Li, S., Liu, X., Si, L., ... & Zhou, G.[4],they concentrated on the task of sentiment classification in a crucial type of dialogue—customer service dialogue and they suggested a novel method that captures all relevant information to improve the classification performance. They specifically suggested a topic-aware multi-task learning (TML) strategy that, by recording multiple types of subject information, learns topic-enriched utterance representations in customer service dialogue. In the experiment, we suggest a sizable and high-quality annotated corpus for the sentiment classification problem in customer service discourse. Empirical analyses on the proposed corpus demonstrate that their technique significantly outperforms various strong baselines. The TML system has an F1 score of 75.9.

In “Sentiment analysis of tweets using svm” by Ahmad, M., Aftab, S., & Ali [5], it is stated that Support Vector Machine (SVM) has been employed in Weka for sentiment analysis in this work. SVM is one of the popular supervised machine learning techniques for determining the polarity of text. Two pre-classified datasets of tweets are used to analyze the performance of SVM, and three metrics—Precision, Recall, and F-Measure—are utilized to compare results. One of the most popular tools for examining how machine learning and data mining algorithms operate is Weka. The University of Waikato in New Zealand created Weka using the Java programming language. Its user-friendly GUI interface contributes to its widespread acceptance. Because of its portability and General Public License, this utility is fairly well known. Results show that the average Precision, Recall, and F-Measure is 55.8 %, 59.9 %, and 57.2% respectively.

3. *DATA PREPARATION*

3.1. DESCRIPTION

We developed a dataset from conversations between two users. Two users took turns among themselves to continue conversations about COVID-19.

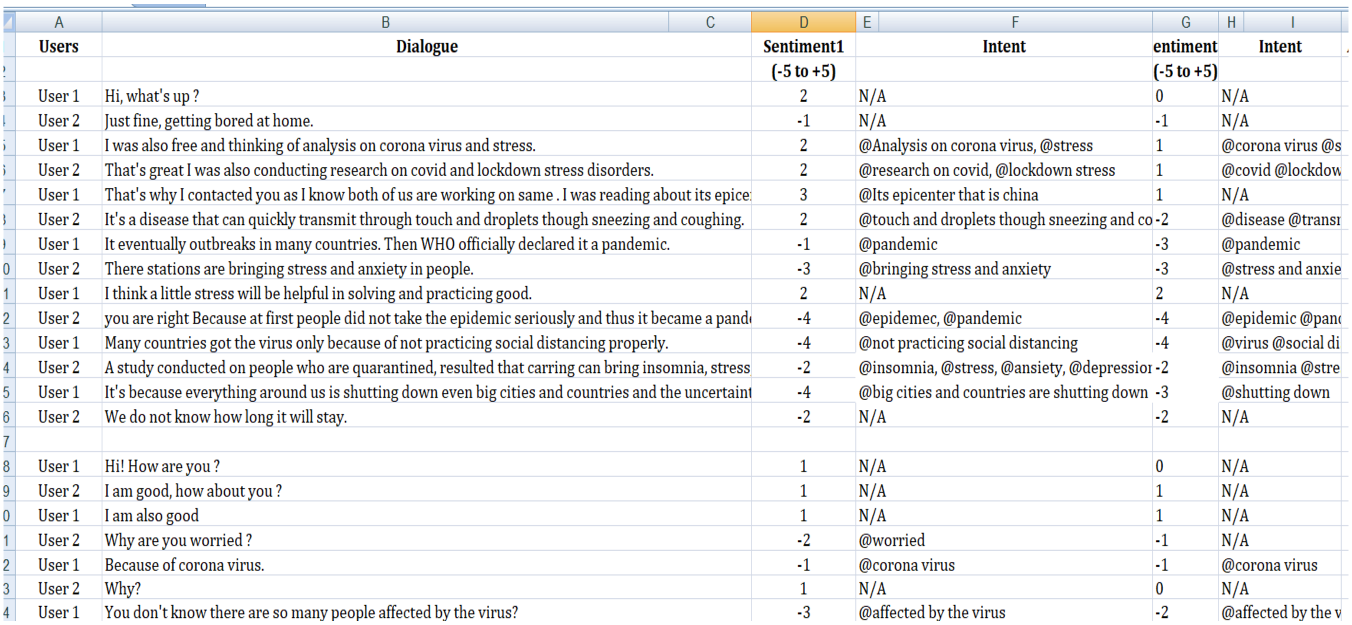
The columns are Speaker, Statements, Intents and Sentiments.

The dataset consists of 36 conversations with 713 utterances. Out of which, User1 has 357 utterances and User2 has 356 utterances. For each utterance, intent words are noted down and Sentiment scores of range -5 to +5 are given by two annotators. Different models(Logistic regression, SVM,CNN) for Sentiment classification are used.

For every utterance, two parameters have been introduced, Sentiment Score (range of -5 to +5) and Intent Word. Sentiment score will help understand the depth of the situation while Intent words will help to identify the topic of the conversation.

We have used majorly 80% of the dataset for the training purpose and 20% of the dataset for testing.

3.2. **SAMPLE DATASET**



Users	Dialogue	Sentiment1	Intent	entiment	Intent
		(-5 to +5)		(-5 to +5)	
User 1	Hi, what's up ?	2	N/A	0	N/A
User 2	Just fine, getting bored at home.	-1	N/A	-1	N/A
User 1	I was also free and thinking of analysis on corona virus and stress.	2	@Analysis on corona virus, @stress	1	@corona virus @s
User 2	That's great I was also conducting research on covid and lockdown stress disorders.	2	@research on covid, @lockdown stress	1	@covid @lockdow
User 1	That's why I contacted you as I know both of us are working on same . I was reading about its epice	3	@Its epicenter that is china	1	N/A
User 2	It's a disease that can quickly transmit through touch and droplets though sneezing and coughing.	2	@touch and droplets though sneezing and co	-2	@disease @transı
User 1	It eventually outbreaks in many countries. Then WHO officially declared it a pandemic.	-1	@pandemic	-3	@pandemic
User 2	There stations are bringing stress and anxiety in people.	-3	@bringing stress and anxiety	-3	@stress and anxie
User 1	I think a little stress will be helpful in solving and practicing good.	2	N/A	2	N/A
User 2	you are right Because at first people did not take the epidemic seriously and thus it became a pand	-4	@epidemec, @pandemic	-4	@epidemic @pan
User 1	Many countries got the virus only because of not practicing social distancing properly.	-4	@not practicing social distancing	-4	@virus @social di
User 2	A study conducted on people who are quarantined, resulted that carring can bring insomnia, stress	-2	@insomnia, @stress, @ansiety, @depressior	-2	@insomnia @stre
User 1	It's because everything around us is shutting down even big cities and countries and the uncertaint	-4	@big cities and countries are shutting down	-3	@shutting down
User 2	We do not know how long it will stay.	-2	N/A	-2	N/A
User 1	Hi! How are you ?	1	N/A	0	N/A
User 2	I am good, how about you ?	1	N/A	1	N/A
User 1	I am also good	1	N/A	1	N/A
User 2	Why are you worried ?	-2	@worried	-1	N/A
User 1	Because of corona virus.	-1	@corona virus	-1	@corona virus
User 2	Why?	1	N/A	0	N/A
User 1	You don't know there are so many people affected by the virus?	-3	@affected by the virus	-2	@affected by the v

Fig 3.2: A sample view of the dataset

4. METHODOLOGY

In this section we describe the complete pipeline of our model which includes
(1) Pre-processing and Sentence/Dialogue Processing
(2)Feature Extraction (TF-IDF for supervised models)
(3)Model Training(Supervised & Deep learning models)

4.1. PRE-PROCESSING

The pre-processing mainly involves solving the issues related to data cleaning. In the preliminary step first thing we did is to extract the dialogue and sentiment columns from the dataset into a columnar format where 1st column is the dialogue and 2nd column is sentiment given by annotator1 and the 3rd column is sentiment given by annotator2. These sentences from the 1st columns are further processed to remove and clean the data with the help of regular expressions. We have also prepared a list of contractions and substituted them with their full form. A full stop is added to the end of sentence in absence of a punctuation mark.

SENTENCE/DIALOGUE PROCESSING

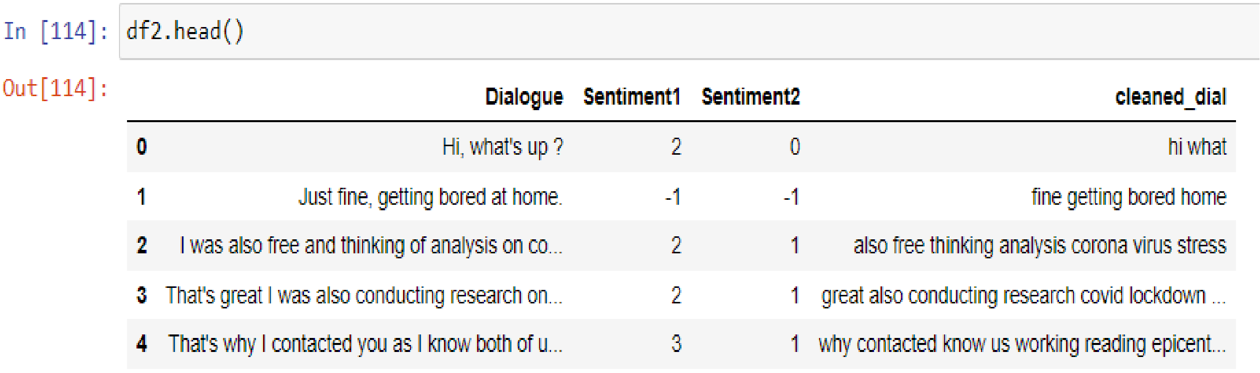
By preprocessing dialogues we get processed data which is suitable for our model. This helps to achieve more accuracy for model training.
All unnecessary elements of the sentence, such as auxiliary verbs and punctuations etc are deleted or removed in this section. Pre-procession is done for our model in the following manner-

- Deleting empty rows, whitespaces.
- **Lemmatization-** Lemmatization is the process of merging two or more words into single word. This examines the morphology of the word and gets rid of

endings like shocked to shock, caught to catch, etc. For our model lemmatization is done by **spacy** library.

- **Removing Stop-words**–Stop-words refer to most common words used in the English language which doesn’t have any contribution towards sentiment analysis. “Is”, “of”, “the,” “by,” etc. are a few examples of stop words. So these words need to be eliminated before model training.
- **Lower casing** all words.
- **Tokenization** – This step breaks the large paragraphs called chunks of text is broken into tokens which are actually sentences. These sentences can further be broken into words. For example, consider the sentence, before word tokenization –“Ram went to banabas with Sitamaiya”. And after tokenization it becomes: {‘Ram’,’ went’, ‘banabas’, ‘with’, ‘to’,’ Sitamaiya’}
- **Expanding short words**–Replacing can’t,don’t with can not, do not etc. is done here.
- **Cleaning** all the **non-letter characters, including numbers.**

After preprocessing our data looks like-



```
In [114]: df2.head()
Out[114]:
```

	Dialogue	Sentiment1	Sentiment2	cleaned_dial
0	Hi, what's up ?	2	0	hi what
1	Just fine, getting bored at home.	-1	-1	fine getting bored home
2	I was also free and thinking of analysis on co...	2	1	also free thinking analysis corona virus stress
3	That's great I was also conducting research on...	2	1	great also conducting research covid lockdown ...
4	That's why I contacted you as I know both of u...	3	1	why contacted know us working reading epicent...

Fig 4.1 : A sample view of data after pre-processing

4.2. *SENTIMENT ANALYSIS MODELS*

4.2.1. *IMPLEMENTING WITH SUPERVISED LEARNING MODEL*

Using labeled training data, supervised learning allows training a function that can later generalize to new examples. A critic is used in the training to show whether a function is accurate or not and then to change the function to obtain the desired output. The back-propagation method is one of the most well-known instances, although there are numerous additional algorithms as well.

A subspace with a hyperplane has a dimension that is 1 less than the feature space. A two-dimensional (2D) plane that crosses a space described by a feature vector of size 3 is known as a hyperplane.

SVM chooses extreme vectors that help in creating the hyperplane. These extreme vectors are called support vectors.

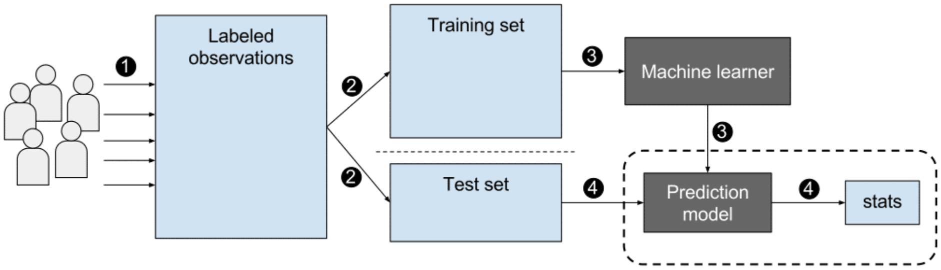


Fig 4.2.1 : Working Principle of supervised learning model

For our model we have used 80% of the data for training the model and 20% for testing. Random state of 82 for annotator1 and random state of 37 for annotator2 is used while splitting the training data.

4.2.1.1. TF-IDF

For feature extraction we have used TF-IDF.The Term Frequency-Inverse Document Frequency (also called TF-IDF), is a well-recognized method to evaluate the importance of a word in a document. Term Frequency of a particular term (t) is calculated as number of times a term occurs in a document to the total number of words in the document. IDF (Inverse Document Frequency) is used to calculate the importance of a term. There are some terms like “is”, “an”, “and” etc. which occurs frequently but don’t have importance. IDF is calculated as IDF (t) = log(N/DF), where N is the number of documents and DF is the number of document containing term t. TF-IDF is a better way to convert the textual representation of information into a Vector Space Model (VSM).

4.2.1.2. LOGISTIC REGRESSION

Logistic Regression is a classification Model. It is a simpler and more efficient model for binary and linear classification problems. The range of logistic regression is bounded between 0 and 1. Logistic regression doesn’t require the relationship between input and output variables. This is because of applying a non-linear log transformation to the odds ratio.

Logistic function is defined by:

logistic function = 1/(1+e-x)

Here, x is the input variable.

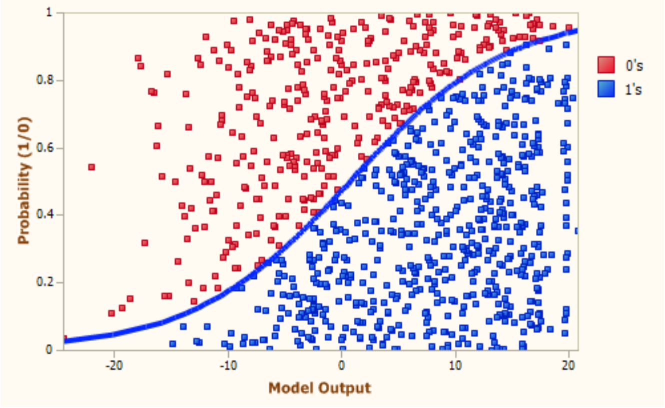


Fig 4.2.1.1 : Working Principle of Logistic Function

4.2.1.3. SUPPORT VECTOR MACHINE

SVMs are a popular supervised learning model that you can use for classification or regression. This approach works well with high-dimensional spaces (many features in the feature vector) and can be used with small data sets effectively. When the algorithm is trained on a data set, it can easily classify new observations efficiently. It does this by constructing one or more hyperplanes to segregate the data set between two classes.

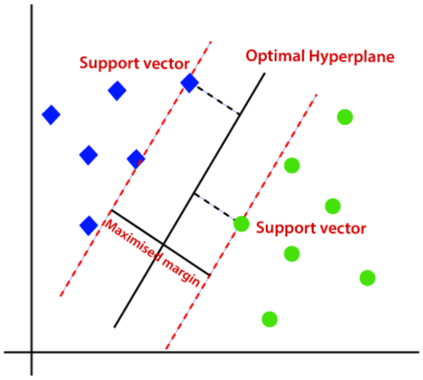


Fig 4.2.1.2 : Working Principle of SVM Function

4.2.2. DEEP LEARNING MODEL

The hidden layers of the neural network are modified via deep learning using a multilayer technique. Features are defined and extracted via feature selection techniques or manually in conventional machine learning procedures. Deep learning models, on the other hand, automatically learn and extract information, improving accuracy and performance. Typically, classifier models' hyper parameters are also measured automatically.

We have used CNN model to complete our analysis.

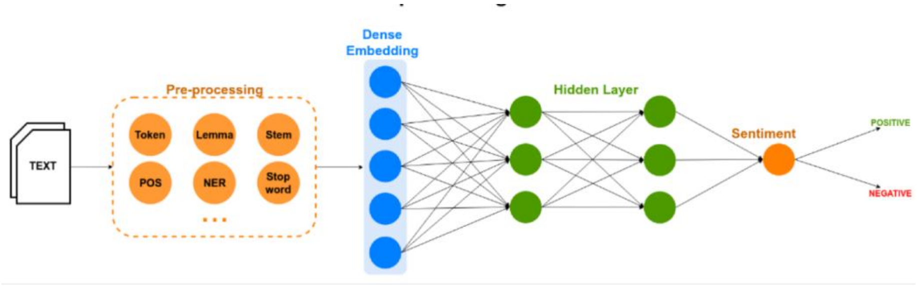


Fig 4.2.2 : Working Principle of Deep learning model

4.2.2.1. CNN MODEL

A convolutional neural network is a special type of feed-forward neural network originally employed in areas such as computer vision, recommender systems, and natural language processing. It is a deep neural network architecture, typically composed of convolutional and pooling or subsampling layers to provide inputs to a fully-connected classification layer. Convolution layers filter their inputs to extract features; the outputs of multiple filters can be combined. Pooling or subsampling layers reduce the resolution of features, which can increase the CNN's robustness to noise and distortion. Fully connected layers perform classification tasks.

In the case of our model CNN, 2D convolutional layers have been used. Each of the convolution layers is followed by a Max-pooling layer. The number of filters is set to 256 and 128 for the first and second convolutional layers, respectively whereas the size of the filter is chosen 5 for both the layers. Learning rate of 0.0001 is applied throughout the training process. Exponential Linear Unit (ELU) was used in convolutional layers as well as fully connected layers. Among the three dense layers, the first one has 128 hidden neurons with a dropout value of 0.7 and the second one has 64 hidden neurons with dropout value of 0.5. Six softmax units have been used to classify each user input for the first dataset and eleven softmax units have been used for the second dataset in the last dense layer. RELU is used for intermediate dense layers.

We have performed only relaxed evaluation in case of CNN model. Train dataset if taken as 70% of the total dataset for annotator 1 and 80% for the annotator 2.

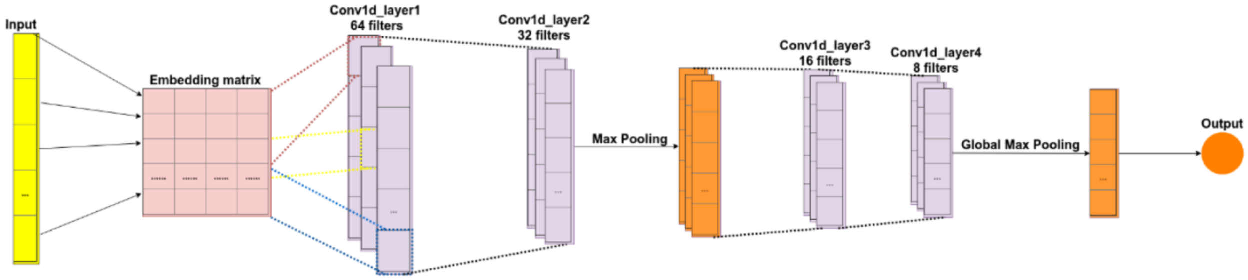


Fig 4.2.2.1: CNN Architecture

5. *EXPERIMENTS AND RESULTS*

5.1. COMPARATIVE STUDY

- In this study, word-embedding for deep learning model and features TF-IDF (word level) for supervise models were both taken into consideration separately done for **annotator1's sentiment** and **annotator2's sentiment**. Table 1 displays the results of three classification algorithms (SVM, Logistic Regression, and CNN) utilising the TF-IDF feature (four performance metrics, accuracy, precision, recall, and f-score). As observed from the tables below, SVM performs better in both situations, and our task is to determine which model perform better than others and for which annotator.

Table 1 : All over accuracy of 3 models' strict evaluations

Annotator	*Algorithms*	*Accuracy %*	*Precision %*	*Recall %*	*F1-score %*
First	*Logistics*	39	0.34	0.23	0.22
	SVM	44	0.35	0.33	0.33
	CNN	60	0.30	**0.50**	0.38
Second	*Logistics*	49	**0.59**	0.35	0.37
	SVM	50	0.46	0.37	**0.39**
	CNN	**63**	0.21	0.33	0.26

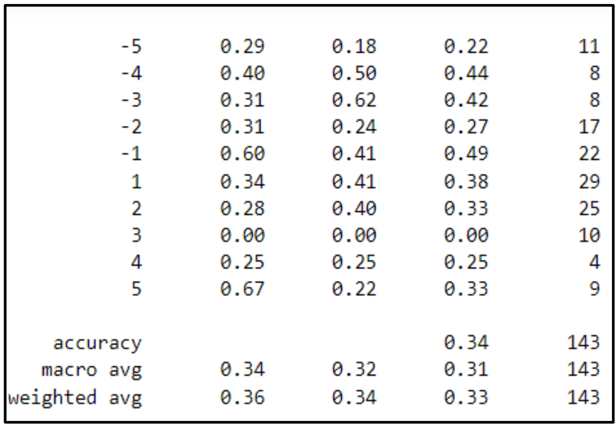
We can see **accuracy is better forannotator2's** sentiments than annotator1's sentiment. Also for **SVM model** accuracy score is higher than all other supervise models.

- Also both **Strict Evaluation** and **Relaxed Evaluation**have been done on the dataset.

In strict evaluation, we have focused on 11 sentiment classes, -5 to +5 and in relaxed evaluation, we have focused on 3 sentiment classes which are positive, neutral and negative sentiments only (in some cases test data missed neutral sentiment which is 0, hence only 10 or 2 classes are evaluated in classification report).

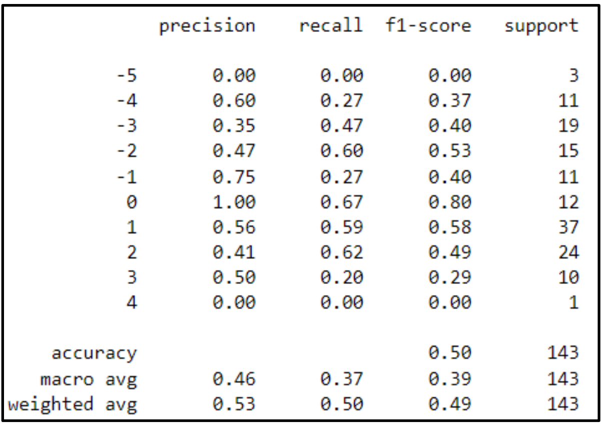
- ***Classification report on basis of strict evaluation***

Annotator1 & SVM:



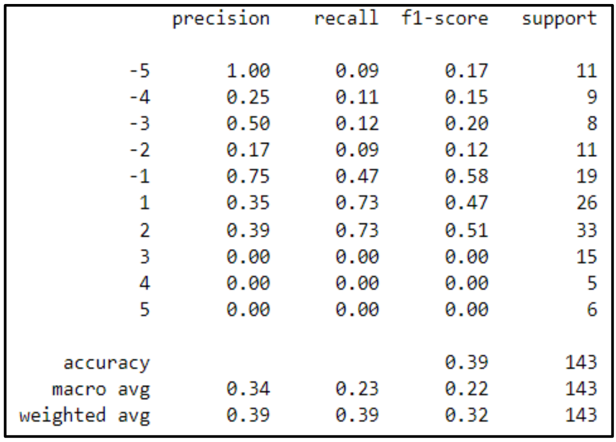
```
          -5       0.29      0.18      0.22        11
          -4       0.40      0.50      0.44         8
          -3       0.31      0.62      0.42         8
          -2       0.31      0.24      0.27        17
          -1       0.60      0.41      0.49        22
           1       0.34      0.41      0.38        29
           2       0.28      0.40      0.33        25
           3       0.00      0.00      0.00        10
           4       0.25      0.25      0.25         4
           5       0.67      0.22      0.33         9

    accuracy                           0.34       143
   macro avg       0.34      0.32      0.31       143
weighted avg       0.36      0.34      0.33       143
```

Annotator2 & SVM:



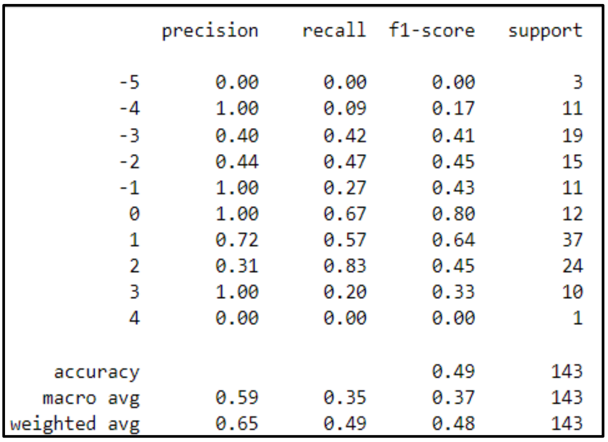
```
              precision    recall  f1-score   support

          -5       0.00      0.00      0.00         3
          -4       0.60      0.27      0.37        11
          -3       0.35      0.47      0.40        19
          -2       0.47      0.60      0.53        15
          -1       0.75      0.27      0.40        11
           0       1.00      0.67      0.80        12
           1       0.56      0.59      0.58        37
           2       0.41      0.62      0.49        24
           3       0.50      0.20      0.29        10
           4       0.00      0.00      0.00         1

    accuracy                           0.50       143
   macro avg       0.46      0.37      0.39       143
weighted avg       0.53      0.50      0.49       143
```

Annotator1 & Logistics Regression:



```
              precision    recall  f1-score   support

          -5       1.00      0.09      0.17        11
          -4       0.25      0.11      0.15         9
          -3       0.50      0.12      0.20         8
          -2       0.17      0.09      0.12        11
          -1       0.75      0.47      0.58        19
           1       0.35      0.73      0.47        26
           2       0.39      0.73      0.51        33
           3       0.00      0.00      0.00        15
           4       0.00      0.00      0.00         5
           5       0.00      0.00      0.00         6

    accuracy                           0.39       143
   macro avg       0.34      0.23      0.22       143
weighted avg       0.39      0.39      0.32       143
```

Annotator2 & Logistics Regression:

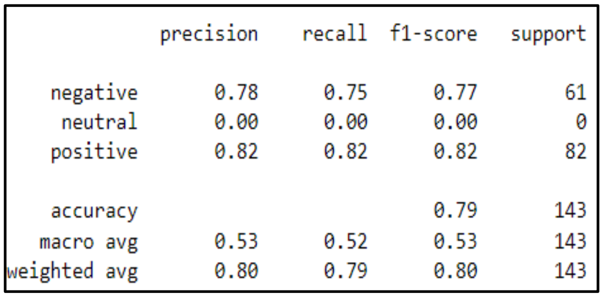


```
              precision    recall  f1-score   support

          -5       0.00      0.00      0.00         3
          -4       1.00      0.09      0.17        11
          -3       0.40      0.42      0.41        19
          -2       0.44      0.47      0.45        15
          -1       1.00      0.27      0.43        11
           0       1.00      0.67      0.80        12
           1       0.72      0.57      0.64        37
           2       0.31      0.83      0.45        24
           3       1.00      0.20      0.33        10
           4       0.00      0.00      0.00         1

    accuracy                           0.49       143
   macro avg       0.59      0.35      0.37       143
weighted avg       0.65      0.49      0.48       143
```

Hereas we can see for some classes testing data was insufficient which results 0 precision. For example classes -5,4 in **Annotator 2-Logistics.**

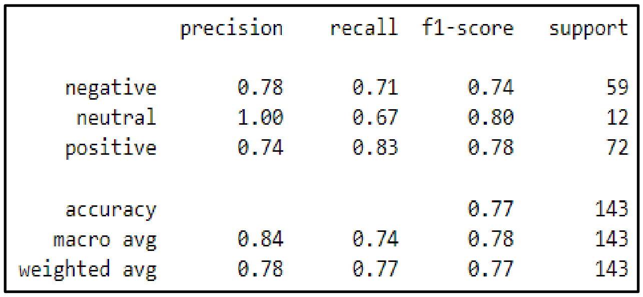
- ***Classification report on basis of relaxed evaluation***

Annotator1 & SVM:



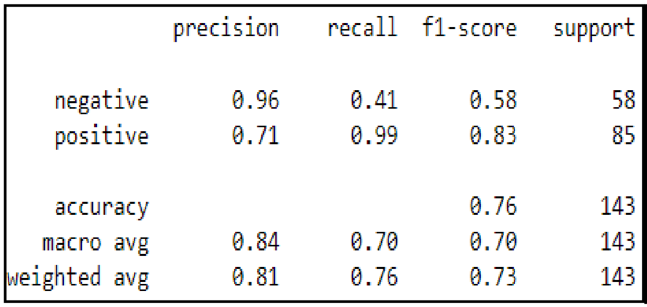
```
              precision    recall  f1-score   support

    negative       0.78      0.75      0.77        61
     neutral       0.00      0.00      0.00         0
    positive       0.82      0.82      0.82        82

    accuracy                           0.79       143
   macro avg       0.53      0.52      0.53       143
weighted avg       0.80      0.79      0.80       143
```

Annotator2 & SVM:



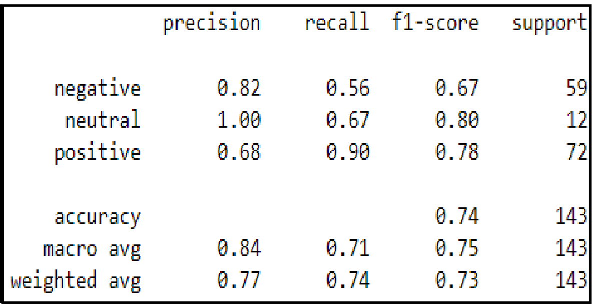
```
              precision    recall  f1-score   support

    negative       0.78      0.71      0.74        59
     neutral       1.00      0.67      0.80        12
    positive       0.74      0.83      0.78        72

    accuracy                           0.77       143
   macro avg       0.84      0.74      0.78       143
weighted avg       0.78      0.77      0.77       143
```

Annotator1 & Logistics Regression:



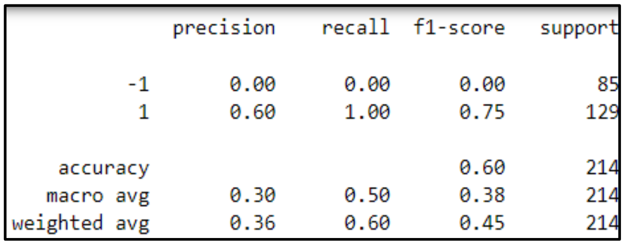
```
              precision    recall  f1-score   support

    negative       0.96      0.41      0.58        58
    positive       0.71      0.99      0.83        85

    accuracy                           0.76       143
   macro avg       0.84      0.70      0.70       143
weighted avg       0.81      0.76      0.73       143
```

Annotator2 & Logistics Regression:



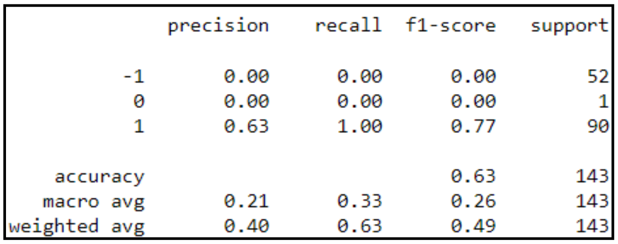
```
              precision    recall  f1-score   support

    negative       0.82      0.56      0.67        59
     neutral       1.00      0.67      0.80        12
    positive       0.68      0.90      0.78        72

    accuracy                           0.74       143
   macro avg       0.84      0.71      0.75       143
weighted avg       0.77      0.74      0.73       143
```

Annotator1 & CNN:



```
              precision    recall  f1-score   support

          -1       0.00      0.00      0.00        85
           1       0.60      1.00      0.75       129

    accuracy                           0.60       214
   macro avg       0.30      0.50      0.38       214
weighted avg       0.36      0.60      0.45       214
```

Annotator2 & CNN:



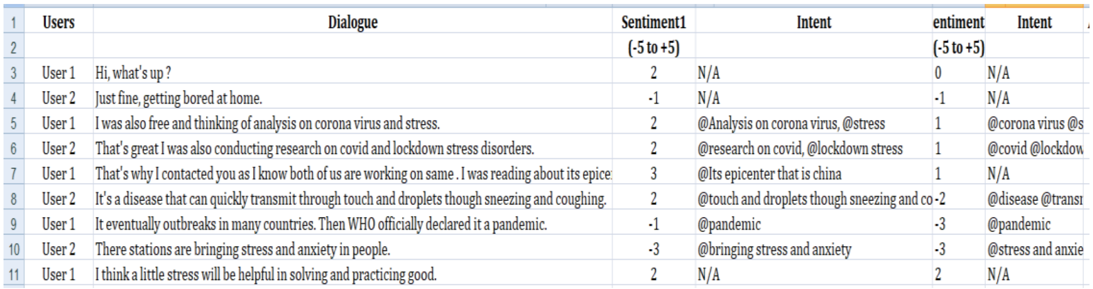
```
              precision    recall  f1-score   support

          -1       0.00      0.00      0.00        52
           0       0.00      0.00      0.00         1
           1       0.63      1.00      0.77        90

    accuracy                           0.63       143
   macro avg       0.21      0.33      0.26       143
weighted avg       0.40      0.63      0.49       143
```

Here we can see accuracy score of CNN model is the highest but for both annotator1 & annotator2 it is unable to detect all negative and neutral sentiments resulting f1-score of class -1,0 **zero**. So we can say if we consider accuracy without any data loss then SVM has best performance.

6. OBSERVATION & DISCUSSIONS

6.1. SYSTEM ANALYSIS

- **Challenges faced due to labeling sentiments manually-**In our dataset, all the sentiments have been marked by two annotators. From the following table, we can see that two sentiments are differently marked by annotators which can surely affect the accuracy of our model. Marking sentiments manually has several drawbacks, it depends on the mood and situation in which the annotator is in. Also, it is not sure that if he is asked again to label the same sentiment for the 2nd time then the annotator will mark the sentiment as similar to the previous one.

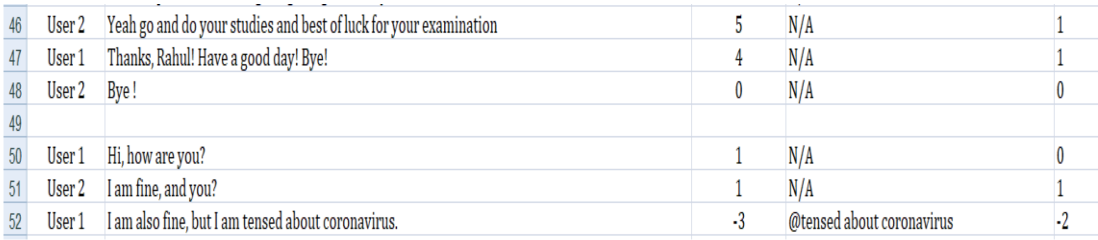


1	Users	Dialogue	Sentiment1	Intent	entiment	Intent
2			(-5 to +5)		(-5 to +5)	
3	User 1	Hi, what's up ?	2	N/A	0	N/A
4	User 2	Just fine, getting bored at home.	-1	N/A	-1	N/A
5	User 1	I was also free and thinking of analysis on corona virus and stress.	2	@Analysis on corona virus, @stress	1	@corona virus @s
6	User 2	That's great I was also conducting research on covid and lockdown stress disorders.	2	@research on covid, @lockdown stress	1	@covid @lockdow
7	User 1	That's why I contacted you as I know both of us are working on same . I was reading about its epice	3	@Its epicenter that is china	1	N/A
8	User 2	It's a disease that can quickly transmit through touch and droplets though sneezing and coughing.	2	@touch and droplets though sneezing and co	-2	@disease @transı
9	User 1	It eventually outbreaks in many countries. Then WHO officially declared it a pandemic.	-1	@pandemic	-3	@pandemic
10	User 2	There stations are bringing stress and anxiety in people.	-3	@bringing stress and anxiety	-3	@stress and anxie
11	User 1	I think a little stress will be helpful in solving and practicing good.	2	N/A	2	N/A

Fig 6.1.1 : A sample view of the dataset where different sentiment is shown for same dialogue with restpect to two 23nnotators

- **Lack of data –**In the case of the deep learning model CNN, the accuracy has been reduced due to a lack of sufficient data. Each deep learning model is data-hungry, requires a large amount of data, and needs powerful computing resources.

- **Unnecessary dialogues –** Since this dataset is based on dialogues, it consists of many unnecessary sentences like 'hi!', 'what's up?' etc. as shown in the following image. Either simple chats/words are almost removed during pre-processing (stop-words removal) or their sentiments are tagged as neutral. These uninformative words do not serve much for analysis.



46	User 2	Yeah go and do your studies and best of luck for your examination	5	N/A	1
47	User 1	Thanks, Rahul! Have a good day! Bye!	4	N/A	1
48	User 2	Bye !	0	N/A	0
49					
50	User 1	Hi, how are you?	1	N/A	0
51	User 2	I am fine, and you?	1	N/A	1
52	User 1	I am also fine, but I am tensed about coronavirus.	-3	@tensed about coronavirus	-2

Fig 6.1.2 : Uninformative utterances are visible in dataset

- **Fluctuation** - We have observed that each time there is a fluctuation in accuracy due to fewer data. It is not stuck on a point.

6.2. ERROR ANALYSIS

- **Feature extraction**- This is the most important part of supervised learning. And in this, I have used TD-IDF which can be the main factor as the feature extraction is done manually by the supervised learning model.

- **Manual sentiment tagging** - Sentiments are given by two different annotators not calculated in this dataset. Accordingly, the accuracy and scoring matrix differs.
  For example, the accuracy of the SVM model is predicted as 50% for sentiments tagged by annotator 2 whereas 34% for annotator 1. For other models also we can see fluctuation in accuracy.

- **Unbalancednumber of class data** -It is clear that neutral sentiment carries the most weight, followed by negative sentiment. It results an unbalanced classification problem that can affect the accuracy of machine learning or deep learning based models.

- **Confusion matrices of the modelsstrict vs. relaxed evaluation and annotator1 vs. annotator2.**

Table 2: comparative studies of confusion matrices on relax and strong evaluation for annotator1 & annotator2

	Model	*Annotator1*	*Annotator2*
Strict	***SVM***	[[5 3 1 0 1 1 0 0 0 0] [1 2 1 4 0 0 1 0 0 0] [2 1 0 2 2 0 1 0 0 0] [0 0 0 5 2 3 1 0 0 0] [0 2 1 2 10 2 1 0 1 0] [0 0 0 0 2 18 3 1 2 0] [0 1 3 2 0 4 18 4 0 1] [0 0 0 2 0 3 3 4 2 1] [0 0 2 0 1 1 0 0 0 1] [0 0 0 1 0 2 2 0 0 1]]	[[0 0 2 0 0 0 1 0 0 0] [0 3 7 0 0 0 0 1 0 0] [0 0 9 5 0 0 3 2 0 0] [0 0 2 9 0 0 1 3 0 0] [0 1 0 1 3 0 4 2 0 0] [0 0 0 0 0 8 2 1 0 1] [0 1 2 2 1 0 22 9 0 0] [0 0 1 2 0 0 4 15 2 0] [0 0 2 0 0 0 2 4 2 0] [0 0 1 0 0 0 0 0 0 0]]
	Logistics	[[1 2 0 0 1 4 3 0 0 0] [0 1 1 2 1 2 2 0 0 0] [0 0 1 0 0 4 3 0 0 0] [0 0 0 1 1 5 4 0 0 0] [0 1 0 2 9 3 4 0 0 0] [0 0 0 0 0 19 7 0 0 0] [0 0 0 0 0 8 24 1 0 0] [0 0 0 1 0 7 7 0 0 0] [0 0 0 0 0 1 4 0 0 0] [0 0 0 0 0 2 4 0 0 0]]	[[0 0 2 0 0 0 0 1 0 0] [0 1 3 2 0 0 0 5 0 0] [0 0 8 5 0 0 0 6 0 0] [0 0 1 7 0 0 1 6 0 0] [0 0 1 0 3 0 3 4 0 0] [0 0 0 0 0 8 1 3 0 0] [0 0 4 1 0 0 21 11 0 0] [0 0 0 1 0 0 3 20 0 0] [0 0 1 0 0 0 0 7 2 0] [0 0 0 0 0 0 0 1 0 0]]
Relaxed	***SVM***	[[46 0 15] [0 0 0] [13 2 67]]	[[42 0 17] [0 8 4] [12 0 60]]
	Logistics	[[24 34] [1 84]]	[[33 0 26] [0 8 4] [7 0 65]]
	CNN	[[0 85] [0 129]]	[[0 0 52] [0 0 1] [0 0 90]]

7. CONCLUSION

A difference in accuracy is found for manual sentiment tagging. SVM model performs better than the logistic regression model. Also CNN model performs better than other machine learning models.

For future purposes, annotators must be more careful while tagging sentiments. The dataset should consist of at least 5000 utterances to make our deep learning model more effective. Sampling could have been a solution for unbalanced classification which needs to be explored in our later experiment.

REFERENCES

[1]Zhang, Lei, Shuai Wang, and Bing Liu. "Deep learning for sentiment analysis: A survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.4 (2018): e1253.

[2]Feldman, Ronen. "Techniques and applications for sentiment analysis." Communications of the ACM 56.4 (2013): 82-89.

[3]Song, L., Xin, C., Lai, S., Wang, A., Su, J., &Xu, K.. CASA: Conversational Aspect Sentiment Analysis for Dialogue Understanding. Journal of Artificial Intelligence Research,(2022), 73, 511-533.

[4]Wang, J., Wang, J., Sun, C., Li, S., Liu, X., Si, L., ...& Zhou, G. (2020, April). Sentiment classification in customer service dialogue with topic-aware multi-task learning.In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 9177-9184).

[5]Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment analysis of tweets using svm. Int. J. Comput. Appl, 177(5), 25-29.

[6]Ahuja, Ravinder, et al. "The impact of features extraction on the sentiment analysis." *Procedia Computer Science* 152 (2019): 341-348.

[7]A.H. Alamoodi, B.B. Zaidan, A.A. Zaidan, O.S. Albahri, K.I. Mohammed, R.Q. Malik, E.M. Almahdi, M.A. Chyad, Z. Tareq, A.S. Albahri, HamsaHameed, MusaabAlaa, “Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review”,Expert Systems with Applications,Volume 167,2021,114155,ISSN 0957-4174

[8]Naqvi, Uzma, Abdul Majid, and Syed Ali Abbas. "UTSA: Urdu text sentiment analysis using deep learning methods." *IEEE Access* 9 (2021): 114085-114094.