Instagram Comment Classifier

R. Sai Manognya[1], Ramsha Mehreen[2], T. Hemanth[3], Dhanush[4]

[1]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.

[2]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.

[3] Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.

[4]Department of Electrical and Electronics Communication Engineering, SR University, Warangal, Telangana, India.

ABSTRACT

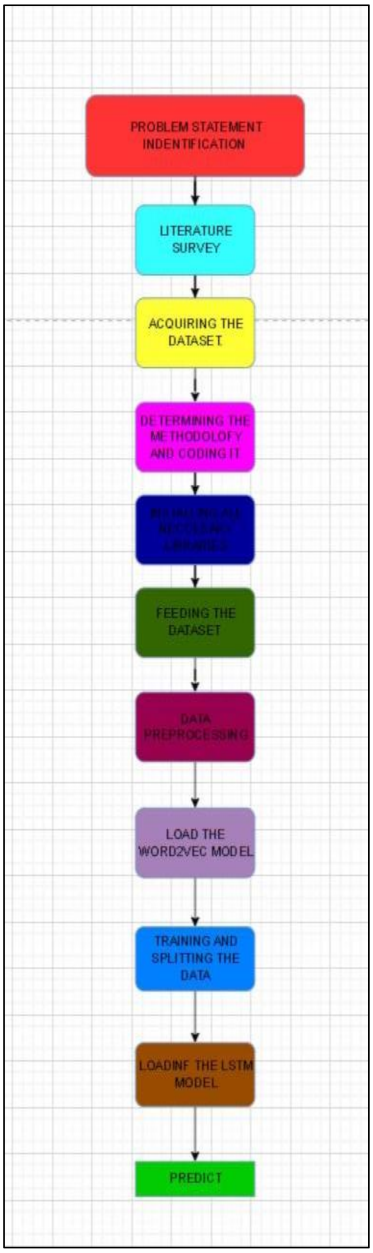
Instagram (IG) is a web-based and mobile social media application where users can share photos or videos with available features. Upload photos or videos with captions that contain an explanation of the photo or video that can reap spam comments. Comments on spam containing comments that are not relevant to the caption and photos. The problem that arises when identifying spam is non-spam comments are more dominant than spam comments so that it leads to the problem of the imbalanced dataset. A balanced dataset can influence the performance of a classification method. It is more necessary for implementing such models for letting people know the nature of the comment and make them aware of their feed. This helps them to stay aligned instead of reading and worrying about such unnecessary comments. Our model gave training accuracy of 97.1% and validation accuracy of 97.3% which can be implemented further more in future by increasing more input features.

1.INTRODUCTION

Instagram is a social media for sharing photos and videos that allow users to like or like to comment on media that has been shared. Online forums and social media platforms have provided individuals with the means to put forward their thoughts and freely express their opinion on various issues and incidents. In some cases, these online comments contain explicit language which may hurt the readers. Comments containing explicit language can be classified into myriad categories such as Toxic, Severe Toxic, Obscene, Threat, Insult, and Identity Hate. The threat of abuse and harassment means that many people stop expressing themselves and give up on seeking different opinions. To protect users from being exposed to offensive language on online forums or social media sites, companies have started flagging comments and blocking users who are found guilty of using unpleasant language. Several Machine Learning models have been developed and deployed to filter out the unruly language and protect internet users from becoming victims of online harassment and cyberbullying.

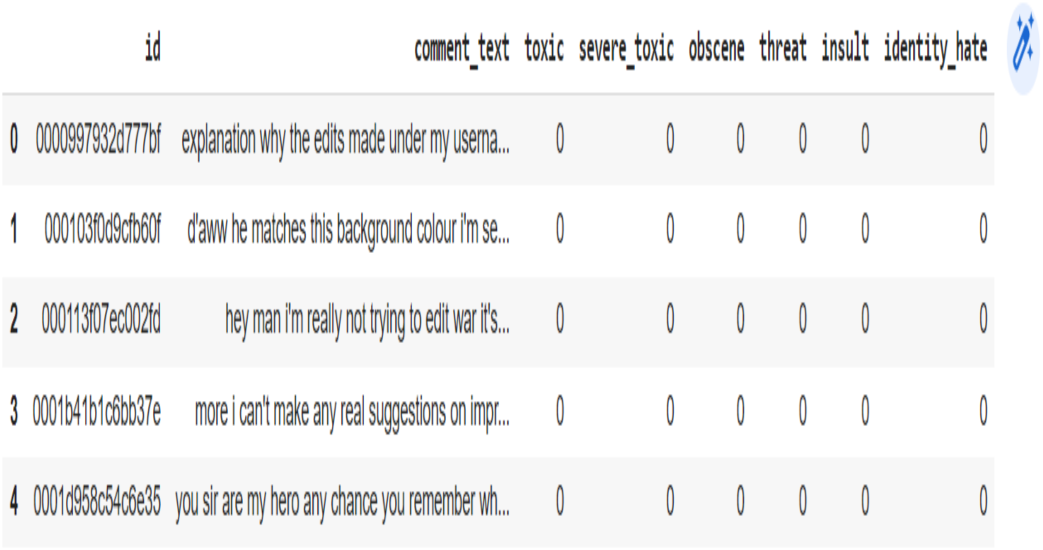
2.PROBLEM DEFINITION

- Threat to social media users and may lead them to depression.
- Broken public trust due to bad comments.
- Achieve reliability.
- Understanding the intuition behind the machine learning algorithms of Natural language processing tool kit and its packages.



3.DATASET AND ATTRIBUTES

The data set for building the classification model was acquired from the competition site and it included the training set as well as the test set. The steps elaborated in the workflow below will describe the entire process from Data Pre-Processing to Model Testing. There are 159571 samples. The input is comment_text and the output Label is a model which predicts a probability of each type of toxicity for each comment.



	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	explanation why the edits made under my userna...	0	0	0	0	0	0
1	000103f0d9cfb60f	d'aww he matches this background colour i'm se...	0	0	0	0	0	0
2	000113f07ec002fd	hey man i'm really not trying to edit war it's...	0	0	0	0	0	0
3	0001b41b1c6bb37e	more i can't make any real suggestions on impr...	0	0	0	0	0	0
4	0001d958c54c6e35	you sir are my hero any chance you remember wh...	0	0	0	0	0	0

4. DATA PRE PREOCESSING

The first step of tokenization is when the phone breaks up the words into tokens. One way this might happen is when the phone looks for spaces, punctuation marks, and capital letters, and then uses those clues to divide up the text into individual words. Tokenization is a process that occurs when there are white spaces in between letters. Tokenizers divide up text into individual tokens.

Lemmatization happens when we replace all of the different forms of a word with its root form. For example, in English we often use “read,” “reads,” and “read,” but in some languages like Spanish and French, each of these forms will be replaced with their root form: “leer.” The process is known as lemmatization. The third step is called stemming**,** where we replace a word with its root form or stem. For example, the most common way to stem words in English is by removing “-ing” and replacing it with “.” There are also more complicated ways of stemming that involve removing parts of words that aren’t actual letters, such as the suffixes “-tion,” “-sion,” and “-ment.” When we remove these suffixes and replace them with just their root forms they become -tion — “dictation”, -sion — “revision”, and -ment — “measure.”

Finally, there’s inflectional morphology, which includes certain grammatical rules that usually involve making a word plural, or changing it from a verb to a noun. For example, if you add ‘er’ to the word ‘design’, it becomes ‘designer’ but the former is a verb whereas the latter becomes a noun. One hot encoding is the most widespread approach ,and it works very well unless your categorical variable takes on a large number of values(i.e., you generally won’t it for variables taking more than 15 different values . It’d be a poor choice in some cases with fewer values ,though that varies).One hot encoding creates new(binary) columns , indicating the presence of each possible value from the original data.

5. METHODOLOGY

Models

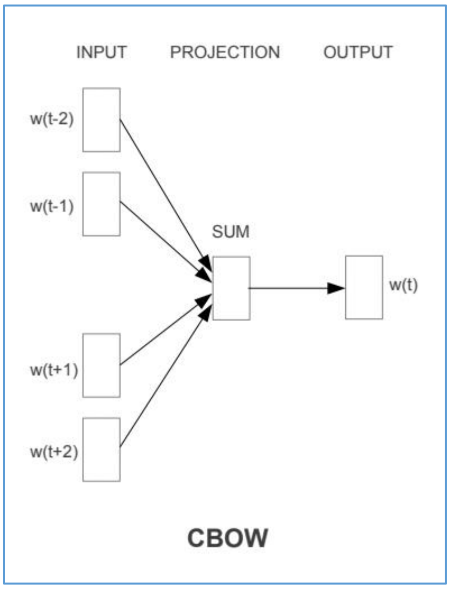
After Data pre-processing we are going to perform word embedding using the Word2Vec vectorizer.

Word2Vec

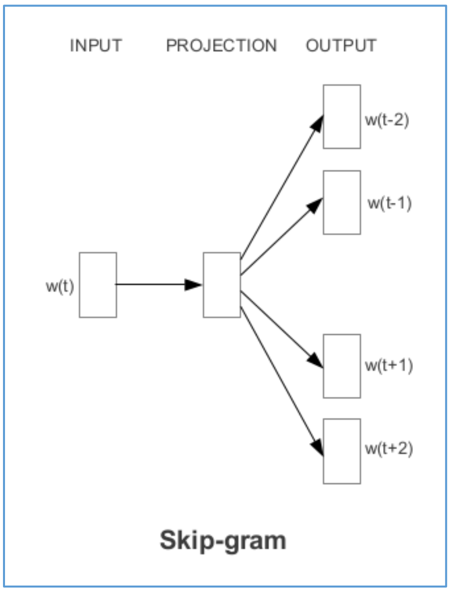
Word2vec is not a singular algorithm, rather, it is a family of model architectures and optimizations that can be used to learn word embeddings from large datasets. Embeddings learned through word2vec have proven to be successful on a variety of downstream natural language processing tasks.

These papers proposed two methods for learning representations of words:

- **Continuous bag-of-words model**: predicts the middle word based on surrounding context words. The context consists of a few words before and after the current (middle) word. This architecture is called a bag-of-words model as the order of words in the context is not important.



- **Continuous skip-gram model**: predicts words within a certain range before and after the current word. A worked example of this is given below.



After word embedding, we are going to load the LSTM model for text classification.

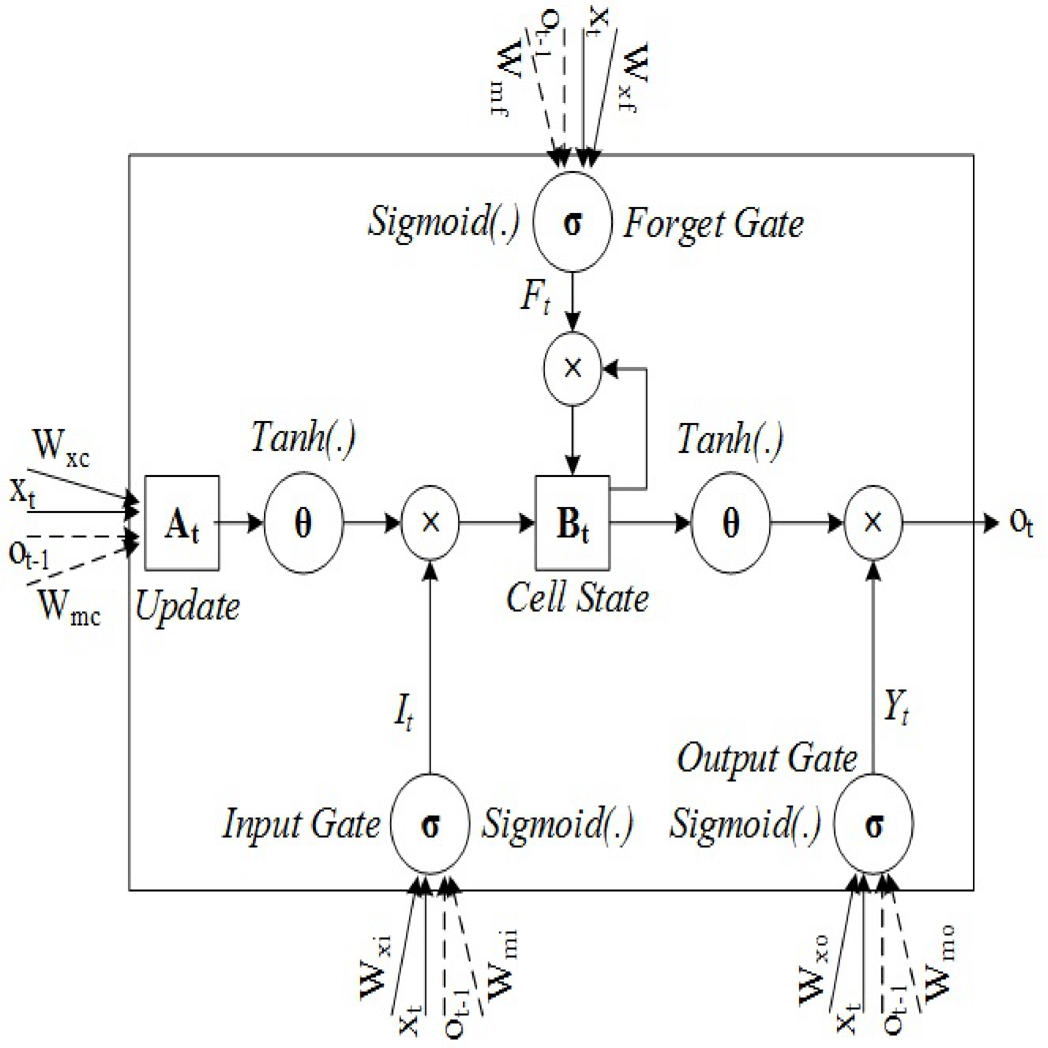
LSTM

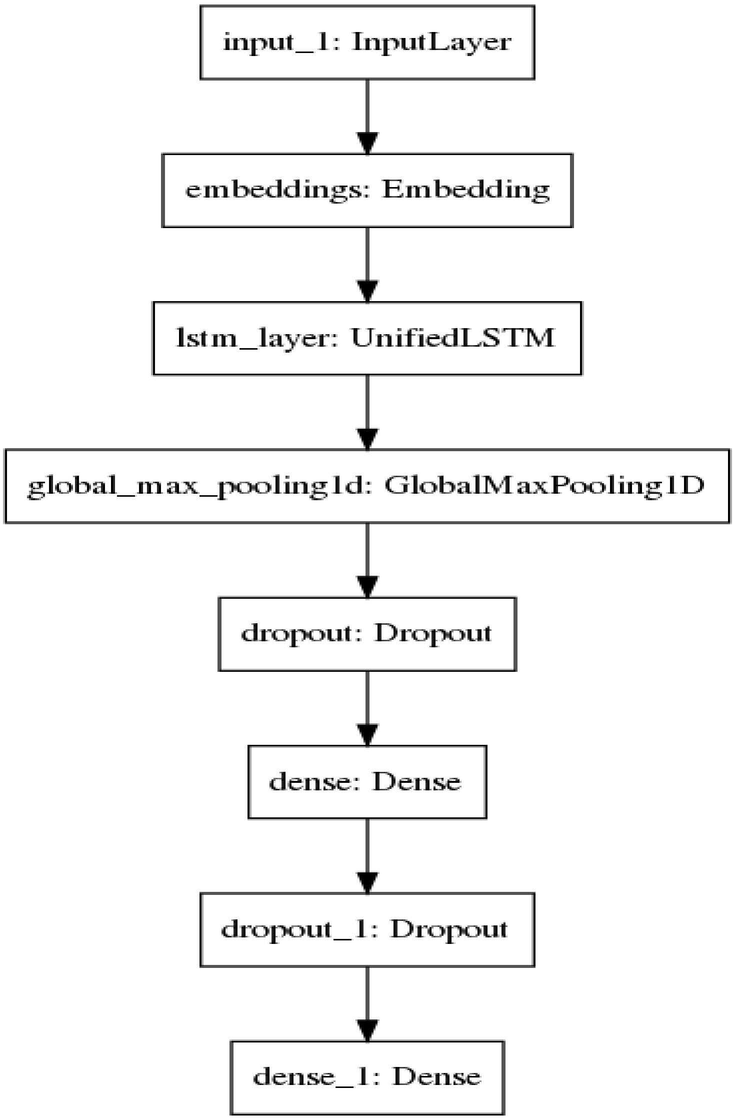
LSTM (Long Short-Term Memory) network is a type of RNN (Recurrent Neural Network) that is widely used for learning sequential data prediction problems. As every other neural network LSTM also has some layers which help it to learn and recognize the pattern for better performance. The basic operation of LSTM can be considered to hold the required information and discard the information which is not required or useful for further prediction.

The Architecture of LSTM

A simple LSTM network consists of the following components.

- Forget gate
- Input gate.
- Output gate

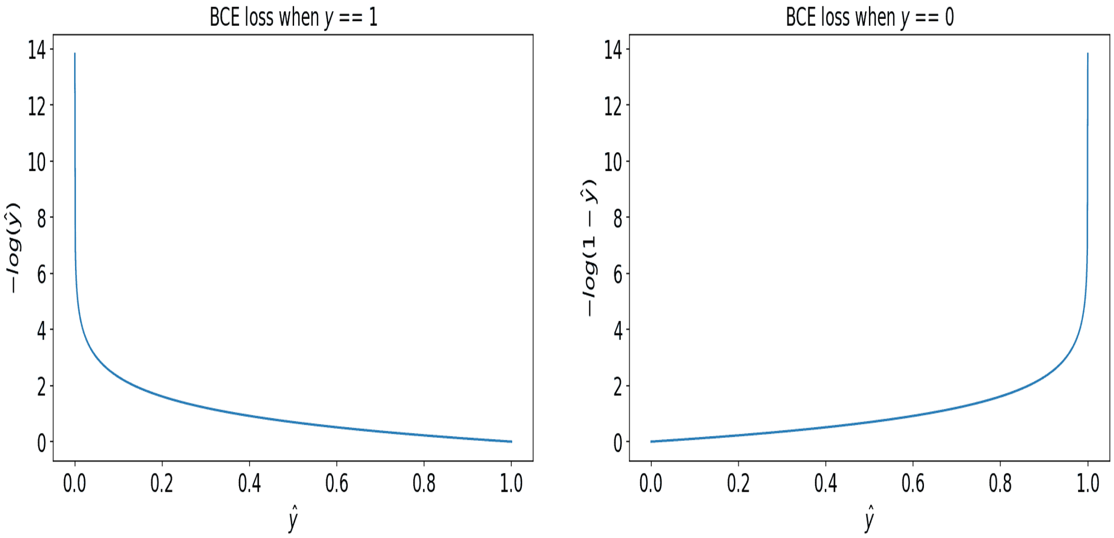




Binary Cross Entropy:

The most common loss function for training a binary classifier is binary cross entropy (sometimes called log loss). Binary classifiers, such as logistic regression, predict yes/no target variables that are typically encoded as 1 (for yes) or 0 (for no). When the model produces a floating point number between 0 and 1 (yhat in the function above), you can often interpret that as p(y == 1) or the probability that the true answer for that record is “yes”. The data you use to train the algorithm will have labels that are either 0 or 1 (y in the function above), since the answer for each record in your training data is known.

The y and (1 - y) terms act like switches so that np.log(yhat) is added when the true answer is “yes” and np.log(1 - yhat) is added when the true answer is “no”. That would move the loss in the opposite direction that we want (since, for example, np.log(yhat) is larger when yhat is closer to 1 than 0) so we take the negative of the sum instead of the sum itself.



Adam Optimizer:

Adaptive Moment Estimation (Adam) is among the top-most optimization techniques used today. In this method, the adaptive learning rate for each parameter is calculated. This method combines advantages of both RMSprop and momentum. stores decaying average of previous gradients and previously squared gradients.

Advantages:

1. Easy Implementation
2. Requires less memory
3. Computationally efficient

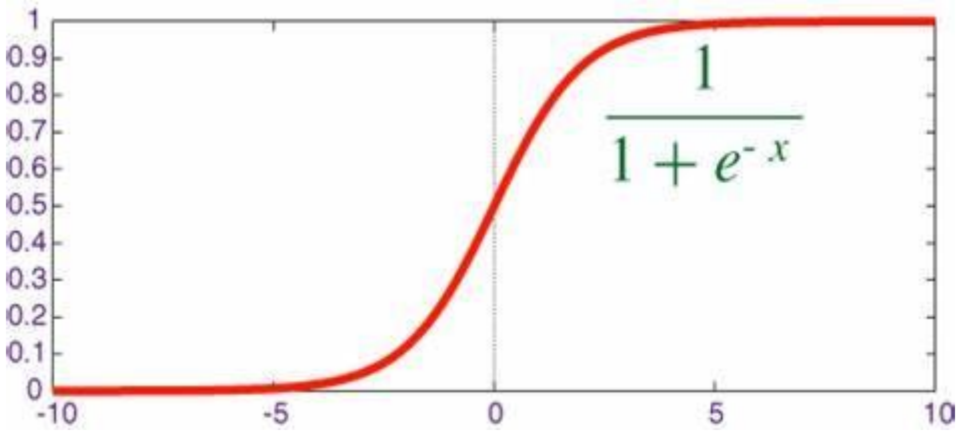
Disadvantages:

1. Can have weight decay problem
2. Sometimes may not converge to an optimal solution

Sigmoid Activation Function:

The sigmoid function also known as logistic function is considered as the primary choice as an activation function since it’s output exists between (0,1). As a result, it's especially useful in models that require the probability to be predicted as an output. Because the likelihood/probability, of anything, only occurs between 0 and 1, sigmoid turns out to be the best option.

So, to sum it up, when a neuron's activation function is a sigmoid function, the output of this unit will always be between 0 and 1. The output of this unit would also be a non-linear function of the weighted sum of inputs, as the sigmoid is a non-linear function. A sigmoid unit is a kind of neuron that uses a sigmoid function as an activation function.



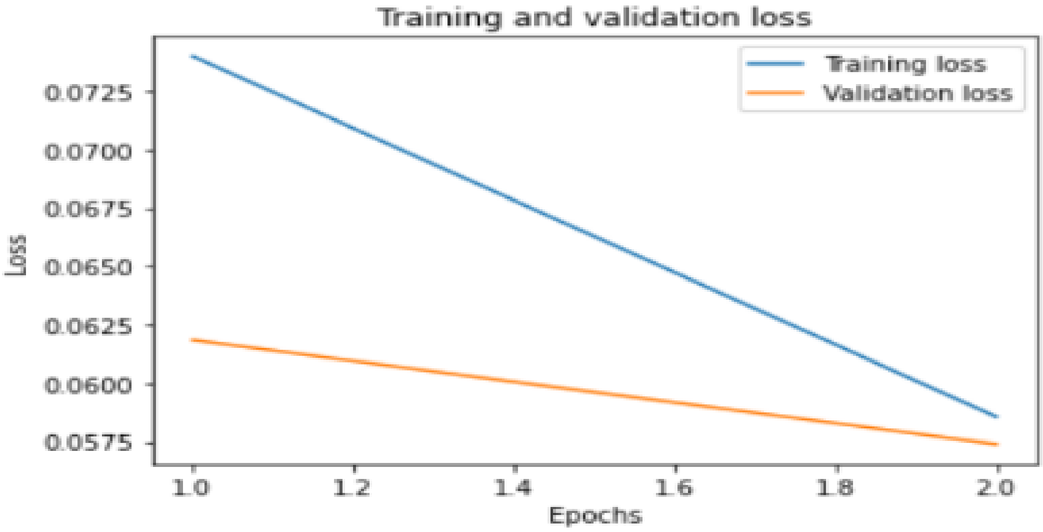
Relu Activation Function:

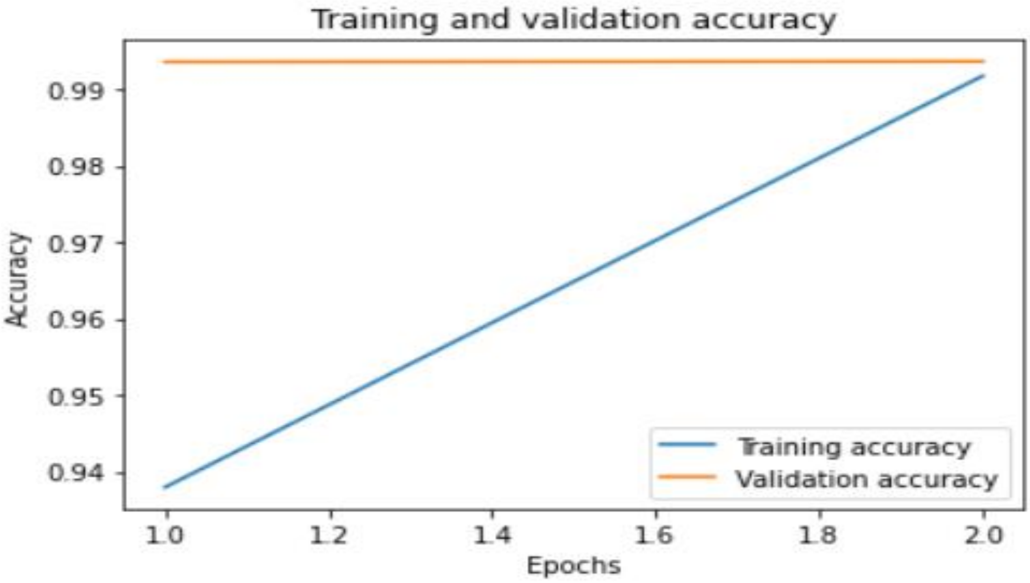
The rectified linear activation function or ReLU is a non-linear function or piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. It is the most commonly used activation function in neural networks, especially in Convolutional Neural Networks (CNNs) & Multilayer perceptrons.

$$f(x) = max(0, x)$$

6.RESULTS

Here are the results we obtained from training our model i.e. training loss is around 0.05 and validation loss is 0.0574 whereas training accuracy is 97.1% and validation accuracy is 97.3%.





7.CONCLUSION

In most of the online conversation platforms, social media users often face abuse, harassment, and insults from other users. Due to which, many users stop expressing their ideas and opinions. Platforms struggle to facilitate conversations effectively. After evaluating the results procured during the training phase of the project and the results that we received, we can claim that the LSTM Model performs better which states that the LSTM model is the right choice for the Instagram Comment Classification use-case. Hence from the above results we can say that our model is working good and can further be implemented by increasing the no. of input features.

8. REFERENCES

[1] https://elibrary.unikom.ac.id/id/eprint/1112/14/UNIKOM_DANIAR%20NUR%20AMIN_JURNAL%20DALAM%20BAHASA%20INGGRIS.pdf

[2] https://github.com/ArielZeev/classification-comments-instagram/blob/main/README.md

[3] https://ieeexplore.ieee.org/document/8938575/metrics#metrics

[4] https://www.researchgate.net/publication/334904656_Detection_Of_Spam_Comments_On_Instagram_Using_Complementary_Naive_Bayes

[5] https://www.socialmediatoday.com/social-networks/instagrams-rolling-out-new-tools-remove-toxic-comments

[6] https://github.com/IBM/MAX-Toxic-Comment-Classifier

[7] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6837517.pdf