International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

COMMENT TOXICATION USING DEEP LEARNING NETWORKS

Dhaneshwar Mardi[1], Rethvik Sai[2], Nimilitha[3], Aishwarya[4], Ganesh[5]

[12345]*Computer Science And Engineering, Gitam University*

Abstract - Online communication has been plagued by the issue of toxic comments more and more. The goal of this research is to identify how toxic a specific comment is. Comments using explicit language can be categorized as Toxic, Severe Toxic, Obscene, Threatening, Insulting, and Identity Hating. To identify and categorise hazardous remarks, the effort uses natural language processing techniques and machine learning algorithms. The project's objective is to create a comprehensive model that can be used to recognise and censor harmful remarks in real time across a range of online venues. To preprocess the text data, extract features, and construct the classification model, the project employs a mix of natural language processing (NLP) techniques and machine learning (ML) algorithms. Tokenization is used in the preprocessing steps, and word embeddings are used in the feature extraction phase.We pass this preprocessed data through the deep learning model.
Next, we'll assess the models' performance before finishing by building a Gradio app. The project resulted in a model capable of identifying various kinds of toxic comments . The model can be integrated into online platforms to allow automatic comment moderation and to promote healthy online discussions. The project has far-reaching consequences for promoting online safety, reducing cyberbullying, and cultivating a positive online environment.

1.INTRODUCTION

Toxic remarks can be made against individuals or groups based on their ethnicity, gender, sexual orientation, religion, nationality, physical appearance, or other traits. These remarks might vary from minor insults to threats and harassment. These can have major psychological and emotional consequences on the recipient, including the following:

Physical health problems: Toxic comments can also lead to physical health problems such as headaches, insomnia, and digestive problems. Prolonged exposure to negative comments can also weaken the immune system, making individuals more susceptible to illness.

Social isolation: People who are exposed to toxic comments may become isolated and avoid social interactions, leading to a loss of social support and a decline in mental and emotional health.

Cyberbullying: Toxic comments can also escalate into cyberbullying, which can have a devastating impact on the individual's mental health, leading to suicidal thoughts and attempts.

Impact on career and personal life: Toxic comments can also affect an individual's personal and professional life, leading to negative consequences such as job loss, damaged relationships, and social stigma.

It is a form of cyberbullying, and they can occur in various forms, including direct attacks, trolling, and passive-aggressive comments. They are often used to intimidate, humiliate, or silence others, and can have a negative impact on the online community as a whole.

People have the ability to openly voice their opinions on a variety of issues and events through online forums and social media.

These online remarks occasionally contain explicit language that readers may find offensive.

Many people stop speaking up and stop looking for helpful advice when they are subjected to harassment and abuse.

2. LITERATURE REVIEW

A)Title: Machine learning methods for toxic comment classification: a systematic review. (Reference: 2)
Authors: Darko Androcec
Summary: The research gives a thorough assessment of machine learning algorithms utilised in online platforms for harmful comment categorization. The authors point out that the issue of identifying and categorising toxic language on online platforms has grown in importance owing to the harmful impact it may have on individuals and communities. They also mention that machine learning methods have showed potential in tackling this issue, but that further research is needed to analyse and compare the various methodologies deployed. The authors looked at 35 research that used different machine learning techniques to categorise harmful remarks. They discovered that supervised learning approaches were utilised in the majority of investigations, with logistic regression, decision trees, and support vector machines being the most widely used classifiers. The authors also discovered other research that employed deep learning approaches to obtain improved accuracy, such as recurrent neural networks and convolutional neural networks.
B)Title: A Neuro-NLP Induced Deep Learning Model Developed Towards Comment Based Toxicity Prediction. (Reference: 6)
Authors: Kulaye Shreyal Ashok, Shaikh Mohammad Bilal Naseem, Kulaye Aishwarya Ashok.
Summary: The research proposes a deep learning model constructed using a neuro-linguistic programming (NLP)

technique for anticipating harmful remarks. According to the authors, existing NLP algorithms are restricted in their capacity to detect and categorise poisonous language effectively. As a consequence, they offer a model that blends natural language processing techniques with a neural network design to get more accurate results. The model's performance is evaluated using two datasets: one from the Kaggle Toxic Comment Classification Competition and another from the Civil Comments platform. The suggested model beats numerous baseline models, with an F1 score of 0.79 on the Kaggle dataset and 0.91 on the Civil Remarks dataset.

3. OBJECTIVE

The major goal of this toxic comments ML research is to create a machine learning model capable of identifying harmful remarks in real time. Additional goals are as follows:

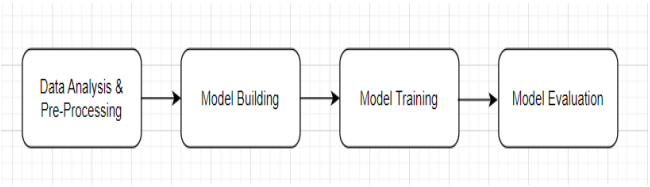
Gather and preprocess a dataset of harmful and non-toxic remarks. Each of the remarks is assigned a binary label that determines how hazardous it is. After tokenization, which occurs during the preprocessing stage, all of the individual words will be represented by integers.

To construct the most accurate model, experiment with various machine learning algorithms and approaches, such as natural language processing and deep learning.

Use numerous metrics to assess the model's efficacy, such as accuracy, precision, and so on. The approach is integrated into a platform to quickly identify and eradicate negative statements. The Gradio application will make use of the model.

Reduce the occurrence of toxic remarks to help promote a healthier and more polite online community.

4. SYSTEM METHODOLOGY



5. DATA ANALYSIS AND PREPROCESSING

Data Analysis and Pre-Processing is the process of taking raw data from a dataset and transforming it into a machine-readable format. This procedure entails employing statistical and/or tokenization tools systematically to describe and demonstrate, compress and recapitulate, and assess data.

6. MODEL BUILDING

Now that we have put the comments into the proper format for the machine to process, we now use that data to build a model which can predict the comment from the input by tokenization. We trained on our training set using the sequential model. We included Dropout and a dense layer in our sequential model. Dropout is a very simple method to prevent overfitting in neural networks.

7. MODEL TRAINING

After understanding which model to use, we put the dataset to use by training the model with the dataset. We do this to get the accuracy of a correct prediction of the result. A training set using 10 epochs was used to train the Sequential model. The validation loss and loss curves were calculated to examine the model. We took 10 epochs, which consisted of 1100 iterations. This number of epochs was taken to show the accuracy of the model, which changes with every epoch, and we needed the highest accuracy. The loss value was also computed along with the accuracy to show how much loss occurred.

8. OVERVIEW OF TECHNOLOGIES

8.1. MACHINE LEARNING

Machine learning (ML) is a subfield of artificial intelligence (AI) that allows machines to automatically learn from data and past experiences in order to spot trends and make predictions with minimum human intervention. Machine learning is concerned with utilising data and algorithms to replicate how humans learn, with the goal of constantly increasing its accuracy. Machine learning may complete tasks without being explicitly instructed. During training and testing, machine learning evaluates algorithms and structures inside a dataset, and qualities from that dataset are utilised as inputs to the algorithms. The correctness of the input-data representation is often heavily dependent on the success of a Machine learning algorithm. It has been demonstrated that an appropriate data

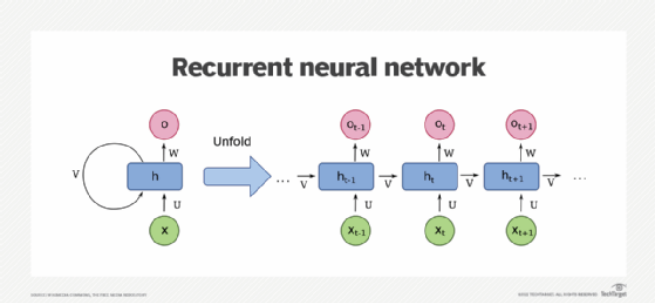
representation improves performance over a poor data representation.

8.2. DEEP LEARNING

Deep learning (DL) is a subset of Machine Learning, which is a subset of Artificial Intelligence. Deep learning is a subject in which computer algorithms learn to evolve on their own. Deep learning involves Artificial Neural Networks (ANN), which are supposed to emulate how humans see and learn, whereas machine learning employs more basic concepts. Deep Neural Networks (DNNs) are networks in which each layer can perform complex operations such as representation and abstraction to make sense of images, sounds, and text. Deep learning algorithms are a complex and mathematically demanding evolution of machine learning algorithms. Deep Learning builds multi-layer learning models by combining transformations and graph technologies. Deep learning techniques extract features automatically.

8.3. RECURRENT NEURAL NETWORKS

Artificial neural networks that employ time series or sequential input are known as recurrent neural networks (RNNs). Deep learning algorithms like this are utilised in well-known apps like Siri, voice search, and Google Translate. For ordinal or temporal issues, they are often utilised in language translation, natural language processing (NLP), speech recognition, and picture captioning. Like feedforward and convolutional neural networks (CNNs), recurrent neural networks (RNNs) learn from training data. They are distinguished by the fact that they have "memory," which allows them to alter current input and output by using data from previous inputs. Recurrent neural networks' output is governed by the basic components of the sequence, whereas traditional deep neural networks presume that inputs and outputs are independent of one.



Since they are the only algorithm with internal memory, RNNs are a strong and stable type of neural network, and one of the most promising ones currently in use. Recurrent neural networks are a popular deep-learning technology that has been around for a long time. Despite the fact that they were developed in the 1980s, we have only just begun to realise their full potential. RNNs have received a lot of attention as a result of advances in computer power, vast quantities of data currently available, and the introduction of long short-term memory (LSTM) in the 1990s.

RNNs can forecast what will happen next because they have internal memory that allows them to recall essential data about the input they received. Time series, speech, text, financial data, audio, video, weather, and many other types of sequential data are the algorithms of choice. In comparison to other algorithms, recurrent neural networks may develop a far deeper grasp of a sequence and its surroundings.

Types of Recurrent Neutral Network

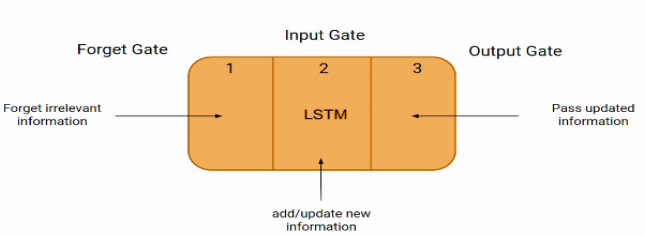
They are three types of Recurrent Neutral Networks: Long Short-Term Memory(LSTM), Gated Recurrent Units(GRU), and Bidirectional RNNs.

Long Short-Term Memory (LSTM)

Recurrent neural networks have long short-term memory. The output of the previous phase is sent into the current step of an RNN as input. Hochreiter and Schmidhuber devised the LSTM. It addressed the issue

of long-term RNN dependency, which happens when the RNN can predict words from current input but not words from long-term memory. RNN performance diminishes as the gap length increases. By default, LSTM may keep data for an unlimited amount of time. It is utilised in time-series data processing, forecasting, and classification. Long short-term memory (LSTM) is a kind of recurrent neural network (RNN) that can handle sequential data such as time series, audio, and text. In order to perform tasks like language translation, speech recognition, and time series forecasting, LSTM networks must understand long-term connections in sequential data. Because just one hidden state is conveyed throughout time, a typical RNN may struggle to grasp long-term dependencies. LSTMs address this issue by including a memory cell—a unit that can store data for a long length of time.

The input gate determines what data is stored in the memory cell. The forget gate determines what data is deleted from the memory cell. Moreover, the output gate controls the memory cell's data output. LSTM networks may learn long-term dependencies by determining which information to keep and which to reject as it comes through the network. Deep LSTM networks may be formed by stacking LSTMs and can recognise increasingly complex patterns in sequential data.



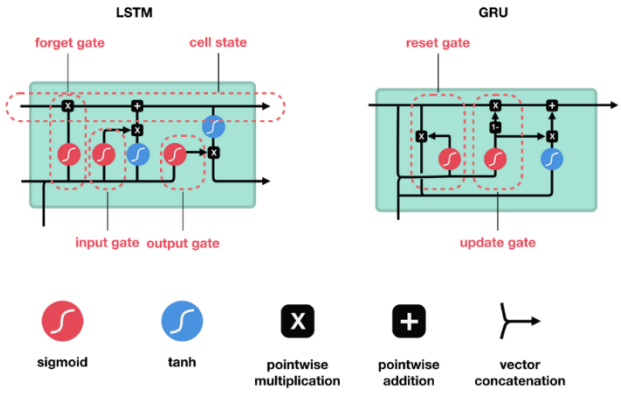
Gated Reconnect Units (GRU)

In 2014, Cho et al. presented the Gated Recurrent Unit (GRU) kind of recurrent neural network (RNN) as a more straightforward alternative to Long Short-Term Memory (LSTM) networks. GRU, like LSTM, is capable of analysing sequential data such as text, audio, and time series.

The main assumption of GRU is to employ gating mechanisms to update the network's hidden state selectively at each time step. The gating mechanisms regulate the flow of data into and out of the network. There are two gating systems on the GRU: the reset gate and the update gate.

The update gate controls how much of the input is used to update the hidden state, whereas the reset gate controls how much of the prior hidden state is wiped. The updated hidden state is used to calculate the GRU output.

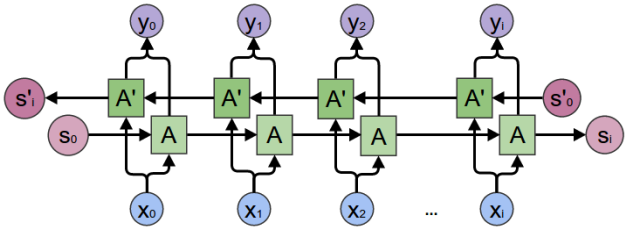
Lastly, GRU networks are a subclass of RNNs that successfully represent sequential data by selectively updating the hidden state at each time step through gating techniques. They have demonstrated efficacy in a wide range of natural language processing applications, including speech recognition, machine translation, and language modelling.



Bidirectional RNNs

Bidirectional RNNs, or BRNNs, are used to allow input to be traversed straight through the past and backward through the future. A BRNN is composed of two RNNs, one of which moves forward from the beginning of the data sequence and the other which moves backward from the beginning of the data sequence. The network

blocks of a BRNN can be LSTMs, GRUs, or simple.



Sequential Model and its Layers

Machine learning models that input or output data sequences are known as sequence models. Text streams, audio and video clips, time series data, and other types of sequential data are all examples of sequential data. Recurrent neural networks (RNNs) are a popular approach in sequence modelling.

The discovery of Sequence Models was motivated by the study of discrete sequential data, such as text sentences, time series, and other sequential data. Convolutional Neural Networks are more suited to dealing with spatial data, whereas these models are better suited to dealing with sequential data.

The most crucial thing to know about sequence models is that we are no longer working with samples that are independently and identically distributed (i.i.d.) since the data we are working with is ordered sequentially.

Bidirectional layer

Bidirectional LSTM is a kind of recurrent neural network that is commonly used in natural language processing (BiLSTM). It can accept input from both sides, and unlike traditional LSTM, the input flows both directions. It is an effective tool for simulating the sequential exchanges of words and phrases in both directions. Lastly, BiLSTM reverses the direction of information flow by adding another LSTM layer. It simply means that the input sequence flows backward in the additional LSTM layer. The results of the two LSTM layers are then combined using a variety of methods such as average, sum, multiplication, and concatenation.

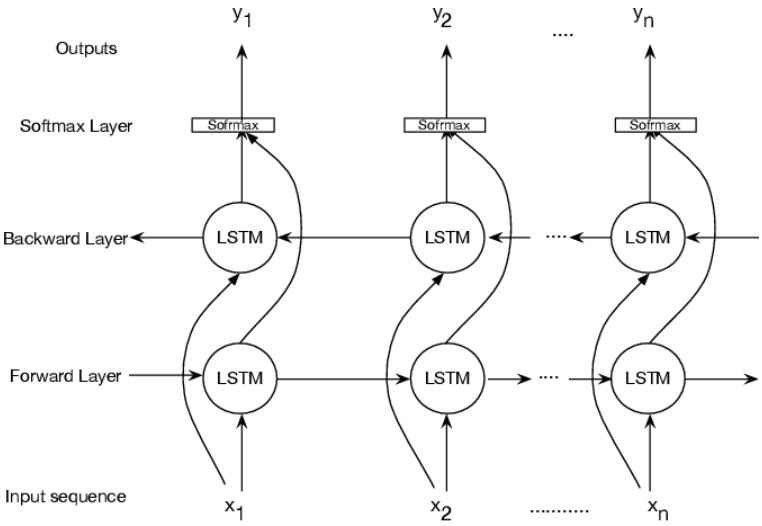
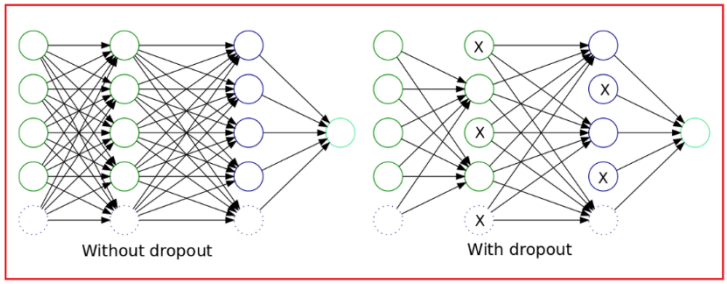


Fig. 1: Bidirectional LSTM architecture.

Dropout Layer

The process of removing nodes from the input and hidden layers of a neural network is known as "dropout." The parent network is used to create a new network architecture by temporarily eliminating all forward and backward links with lost nodes. With a probability of p, the nodes are deleted.

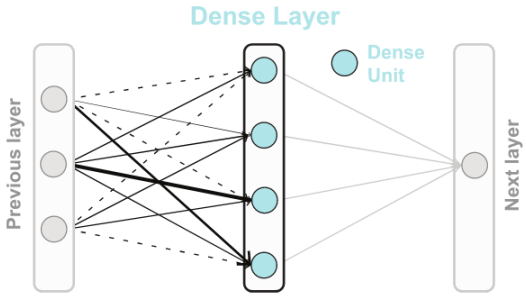


Dense Layer

A dense layer in a neural network is one whose preceding layers are intimately coupled, meaning that each layer's neurons are connected to every other layer's neurons. The most frequently utilised layer in artificial neural network networks is this one.

A model's dense layer neurons do matrix-vector multiplication and receive output from all of the neurons in the layer above them. When performing matrix-vector

multiplication, make sure that the row vector of the dense layer's output from the previous layers is equal to its column vector.

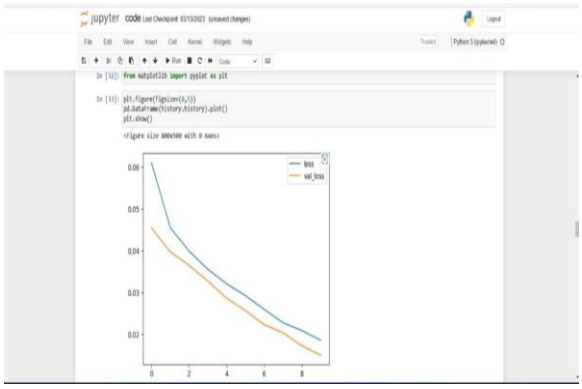


9. TESTING

After completing the coding part, we now compare the output of the techniques that we have used in the coding part. The Sequential model was trained for 10 epochs using the training set. The validation loss vs loss curves were computed to investigate the model.

EPOCH	LOSS	VALIDATION LOSS
1	0.061	0.0454
2	0.0455	0.0397
3	0.0399	0.0365
4	0.0355	0.0327
5	0.0321	0.0286
6	0.0259	0.0257
7	0.0288	0.0223
8	0.0292	0.0204
9	0.0209	0.0173
10	0.0186	0.0151

Table : Comparison of loss and validation loss



In these curves, the loss curve is increasing while the val_loss curve is decreasing. After validating the model, the loss curve decreases from its initial point.

10. RESULTS AND DISCUSSIONS

The below screenshots show the gradio app, in which we have given random comments as input. It gives an output that tells if the comment is Toxic, Severe Toxic, Obscene, Threatening, Insulting, and Identity Hating.

11. CONCLUSION

The goal of the ML toxic comments project was to create a machine learning model capable of reliably classifying hazardous comments from internet discussions.This model proved successful in identifying hazardous remarks and distinguishing them from non-toxic comments.

This project gave us the opportunity to work with two different deep learning models and put them on a Natural Language Processing use-case. The project's multiple data pre-processing and feature engineering phases made us aware of efficient approaches for cleaning textual data. We comprehended the operation of different deep-learning models, including CNN, LSTM, and the LSTM-CNN hybrid model. We learned about word embedding and the advantages of using pre-trained word embedding.

The project's findings show that machine learning has the capacity to detect and prevent hazardous behaviour

in online groups. It should be noted, however, that machine learning models are not a perfect solution to the problem of online toxicity because they are restricted by the quality and representativeness of the data used to train them. To achieve additional success in this field, it is vital to keep improving the quality of the training data and developing more sophisticated models capable of accounting for the various intricacies of human language and behavior.

12. REFERENCES

A. Journals/Article

1. Daniel Jurafsky, James H Martin, "Speech and Language Processing: An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition", 2/e, Prentice Hall, 2008.

2. Machine learning methods for toxic comment classification: a systematic review, by Darko Androcec

3. C. Manning, H. Schutze, "Foundations of Statistical Natural Language Processing", MIT Press. Cambridge, MA, 1999.

4. Jacob Eisenstein, Introduction to Natural Language Processing, MIT Press, 2019.

5. Jalaj Thanaki, Python Natural Language Processing: Explore NLP with macMachinehine Learning and deep learning techniques, Packt, 2017.

6. A Neuro-NLP Induced Deep Learning Model Developed Towards Comment Based Toxicity Prediction By Kulaye Shreyal Ashok, Shaikh Mohammad Bilal Naseem, Kulaye Aishwarya Ashok

B. Website

1. Sequential Model— https://analyticsindiamag.com/a-tutorial-on-sequential-machine-learning/

2. Architectures of Rnn, https://pub.towardsai.net/introduction-to-the-architecture-of recurrent-neural-networks-rnns-a277007984b7#:~:text=Recurrent%20neural%20networks%20(RNNs)%20are,information%20later%20in%20the%20sequence.

3. Classification of LSTM Layers, https://towardsdatascience.com/toxic-comment-classification-using-lstm-and-lstm-cnn-db945d6b7986

4. Illustrated: 10 CNN Architectures, https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d