DEVELOPMENT OF ADVANCED TEXT CLASSIFICATION SYSTEM USING TEXT AUGMENTATION TECHNIQUES

Khemika Kashyap[1], Piyush Pratap Singh[2]

M.Tech Student In Computer Science And Engineering[1]

Associate Professor, SCSS, JNU, Delhi[2]

School Of Computer And System Sciences, Jawaharlal Nehru University, Delhi, India

Abstract: *Text classification in NLP is very common in every domain to classify opinions in various sectors in this world. In the industry, it helps to better corporate with the challenges in contributing company with their classified approached set of tasks. When it comes to medical the data from collected tests of patients are classified as the number of diseases and the medicines to cure a particular disease. Also in the media, to classify the sections and headlines of the newspaper content classification, word classification is mandatorily implemented. The words with similar meaning, grammar, and text make the same interpretation when combined with any other sentence. This is generally called augmentation where not only images can be made in the form of visualization at different angles but NLP also does by beautifying the class of labels based on the intention of writing and speaking. The motivation of text augmentation came from communication as well as from computer vision in the sense that when we speak and see anything around, a different set of images is formed in the brain, so with this, the implementation of a text classification by using the techniques of augmentation has been studied. this technique represents a benchmark for the larger dataset is compared. Data analysis and algorithms is been applied to develop a TensorFlow model where it provides GPU support and prebuild libraries. The study of the paper shows the approaches of augmentation and classification comparison used in today's world.*

Keywords: NLP, Text Classification, Tensorflow, Text-Augmentation

1. INTRODUCTION

Natural language processing on a broader view is how a computer understands human language when it is spoken and written. It is broadly characterized by two that is natural language understanding and generation which involves data preprocessing and highlights followed by generation/development. Some techniques are possible to integrate into this area by using deep learning and machine learning methods, algorithms can effectively analyze the depth of data. In NLP, the main motive is to manipulate the natural language in all possible forms in the same way as it is done by humans i.e, from a generation of language to understanding and learning. creating an NLP pipeline for a problem statement goes with the following chain starting with a text document undergoing sentence segmentation, tokenization, part of speech tagging, lemmatization, stop words, dependency parsing, noun phrases, named entity-relationship, coreference resolution, and at last data structure representing parsed text. NLP provides certain stages of learning where input data can be processed and result in meaningful human-like understanding. The steps involve lexical analysis where it notices the structure of words and phrases. Syntactical analysis checks the grammar and relationship amongst the words. The semantic analysis maps the words with their matching dictionary meaning. The next step is disclosure integration brings about the meaning of the immediately succeeding sentence. The Applied area of NLP involves searching contextual contents, and machine translation techniques by statistical, rule-based, or neural machine translations. When a machine can categorize information it can sense

different perspectives of the same information in the form of augmentation where a simple sentence can be twisted and make a huge difference in meaning. Image augmentation in computer vision is highly used for collecting huge data from an individual image using techniques of rotation, filtering, resizing, and blur effects. Similarly, when the same is implied to NLP data, text can be represented in various forms with the same semantics. Data augmentation is a strategy usually used when there is less dataset for training and the chances of model overfitting dataset are very high. Augmentation lets to add diversity to a dataset and generalize well even in the unseen situation. Text augmentation is derived from the parent domain of computer vision where trivial operations is been performed on images to make understand the patterns inside neurons and learn proactively. Augmentation of a text enables certain techniques where a sentence can be written in such a way that the original sentence looks like the same meaning by just replacing and applying transformation semantically in the data. Many tools and theories is been implemented to achieve augmentation. There are lexical substitution, thesaurus-based substitution, word embedding substitution, masked language model, TF-IDF-based word replacement, back translation, text surface transformation, random noise injection, spelling error injection, and sentence shuffling.

2. LITERATURE SURVEY

Data augmentation is used as standard practice for making a dataset huge and to perform validations over a greater environment. With a single image numbers of other transformed images are generated i.e by rotating, randomness, and additional effects. Similarly for NLP in text-based datasets by doing many transformations and randomness in the text semantically many meanings of the same sentence can be generalized for a large set of neural network training. In the paper “ImageNet Classification with Deep Convolutional Neural Networks, Alex Krizhevsky [1]”, trained a deep convolution neural network with fully connected layers on a GPU environment to reduce overfitting in the model. This shows that a large convolution neural network results better in a huge dataset with a pure supervised learning approach. This has a limitation that it cannot produce a satisfactory response if the convolution layer is removed. This paper ” A survey on Image Data Augmentation for Deep Learning, Connor Shorten & Taghi M. Khoshgoftaar [2]” focuses on data augmentation on limited data. This discussed the algorithms of geometric transformations, color-space augmentation, kernel filter, mixing images, random erasing, adversarial training, and applications of GAN are covered. It shows how a model can improve performance by taking advantage of big data.” Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations, Sosuke Kobayashi [3]”: this paper focused on the word predictions which is generated by the bidirectional LSTM model at word positions. they used datasets like SST5 and TREC with more than two labels. The accuracies of the model show that contextual augmentation improves model performance on different domains majority on synonym-based augmentation.” EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, Jason Wei, Kai Zou [4]” It follows four operations synonym replacement, random insertion, random swap, and random deletion. On these tasks, EDA performs well for both convolution and recurrent neural networks. this paper set a benchmark on five datasets and concluded that it improves huge accuracies for smaller datasets. also, this is limited to the standard practice of NLP. The survey was done based on defining a structured plan for stating augmentation, its goals, trade-offs, interpretation, techniques, and methods. the paper “A Survey of Data Augmentation Approaches for NLP by Steven Y. Feng [5]” defined three different techniques i.e., rule-based, interpolation, and model-based techniques. Along with these techniques, it also studied about applications of data augmentation. In the paper “Rethinking complex neural network architecture for document classification, Ashutosh Adhikari [6]” for a document classification BiLSTM

model with appropriate regularization exceeds the performance of state of art of four standard benchmark datasets. It says that this implementation can be further explored for embeddings language models. " Data Augmentation Using Pre-trained Transformer Models, Varun Kumar [7]", studies different transformer-based pre-trained models such as the autoregressive model(GPT-2),auto-encoder model(BERT), and seq2seq model(BART) for conditional data augmentation. This paper "Do Not Have Enough Data? Deep Learning to the Rescue! Ateret Anlaby-Tavor [8]", introduced a method of LAMBADA that is experimented on small data. It says that it improves classification performance on a variety of datasets.it contributes on three fronts:-1.statistically improves classification accuracy 2. Outperforms state of art of data augmentation methods in scarce data situations 3. Suggest a computing alternative to semi-supervised techniques when unlabeled data does not exist." Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis Franco M. Luque [9]", here in this study, it took a Spanish sentimental tweet dataset to classify the tweets based on polarity. It uses two techniques of augmentation i.e., two-way translation augmentation and instance crossover augmentation. This paper can train a linear classifier and ensemble models to get higher competitive results.

3. MOTIVATION

Text augmentation is used as standard practice for making a dataset huge and to perform validations over a greater environment. With a single image numbers of other transformed images are generated i.e by rotating, randomness, and additional effects. Similarly for NLP in text-based datasets by doing many transformations and randomness in the text semantically many meanings of the same sentence can be generalized for a large set of neural network training.

4. PROBLEM STATEMENT

With a broader domain in NLP, a study of text classification techniques and algorithms to get implemented with performance and visualization to be deployed on the cloud. Further than a twist with NLP by computer vision by using the image and text augmentation. A depth implementable approach of a model by LSTM, CNN, RNN, etc . by TensorFlow and PyTorch as an implementation model.

5. APPROACHES FOR TEXT AUGMENTATION

4.1. Augmentation With Paraphrasing:

Parrot is a paraphrase-based framework used to augment text into paraphrases. parrot minimizes the gap of NLPAug, sentence transformer, and paraphrasing mining utility. A good paraphraser is validated on two factors i.e., (1) if the generated text conveys the same meaning as the original context (Adequacy) and (2) if the text is fluent / grammatically correct English (Fluency) [10]. the three key metrics that measure the quality of paraphrases are: Adequacy (Is the meaning preserved adequately?) Fluency (Is the paraphrase fluent in English?) Diversity (Lexical / Phrasal / Syntactical) (How much has the paraphrase changed the original sentence?) [19].

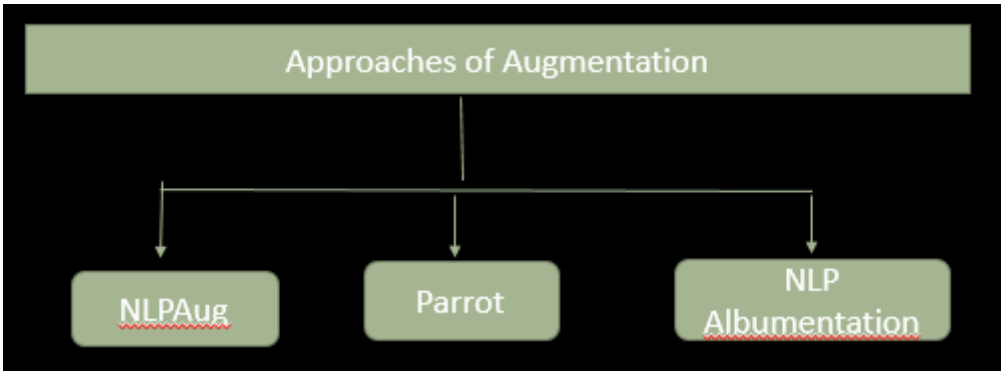

Figure 1: Text Augmentation Approach

4.2. NLPAug:

This is the python library where the augmenter is the basic element of augmentation for a machine learning project. It is a lightweight library and can easily expect fewer lines of code to get the desired result. Augmenter is a sequential pipeline that applies augmentation functions sequentially [20]. There are many steps to augment text inside these, they are:

a. Character Augmenter: Augmentation in character level includes the image to text and chatbot. While recognizing the text we need the OCR model. OCR augmenter simulates the error for augmentation using OCRAug(). For chatbot there is KeyboardAug() is introduced to simulate typing kind of errors. It also has a random augmenter, Swap characters randomly, deletes characters randomly, etc.

b. Word Augmenter: In this technique, we make use of Word2vecAug, GloVeAug, and FasttextAug. It uses word embeddings to search for the most similar relatable words to make sentence worthy.

c. TF-IDF Augmenter: Augmenter that leverage TF-IDF statistics to insert or substitute word.

d. Contextual Word Embeddings Augmenter: Insert, substitute word by contextual word embeddings (BERT, DistilBERT, Roberta, or XLNet)

e. Sentence Augmentation: Here in this method, we can use GPT2 and XLNet for contextual word embeddings.

4.3 NLP Albumentation: It is a computer vision tool that boosts the performance of deep convolutional neural networks. this is used for text classification as well as for augmentations.

6. METHODOLOGY

Data augmentation is a method for increasing the diversity of training examples using a dataset already available. Augmentation in computer vision is explored in more depth as compared to NLP domains. NLPAug library is focused on text and signal processing which has formulations for character-wise, word wise and sentence-wise augmentation. This paper will have prime concern on focusing on text augmentation techniques in NLP dataset, their challenges for training, and system-friendly space. Research design would follow the majority of the sampling method or criteria, the tools, procedure and materials,

and the measured performance. The initial phase of design would be the collection of a dataset, a systematic approach of collecting samples from a dataset and making some statistical analysis regarding data like which distribution data is heading to, which is most affected area, where the region of curve is growing and making out some data visualization reporting for the same. After gathering the dataset next step is to make what you want to implement from that, a detailed study of text classification and its handling techniques were explored to evaluate the performance, and consistency of producing outputs in a larger dataset. Text classifications using machine learning algorithms like naïve Bayes, support vector machine, and deep learning algorithms. The two main deep learning architectures for text classification are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The idea behind classification is to classify a sentence into a defined relation class. Deep learning is hierarchical machine learning, using multiple algorithms in a progressive chain of events. It's similar to how the human brain works when making decisions, using different techniques simultaneously to process huge amounts of data. The next phase of the experiment is to evaluate an image-based labeled dataset, a concept of computer vision, and its application will be useful in natural language processing. While training images with computer vision libraries it becomes easy to collect millions of extra images from a tiny set of images. Deep learning algorithms to train a set of large images will result in good accuracy and an F1 score. This phase of image dataset and text dataset will differentiate how images get easily augmented whereas for words it requires many steps of transformations and high complexity. Now diving into the study of augmentation and how text augmentation plays a good role in NLP applied areas. Image augmentation is used to increase the training data size for training a deep neural network on a dataset. Along with similar lines, text augmentation is explored in the field of text processing for improving the efficiency of models. The last phase of this proposal is to implement these augmentation methods for better model performance. Word embeddings of sentences to generate augmented data to increase data size and trained a multi-class classifier on data. Also, word similarity and synonym methods were used to generate new texts and interpolation and extrapolation for the embedding level.

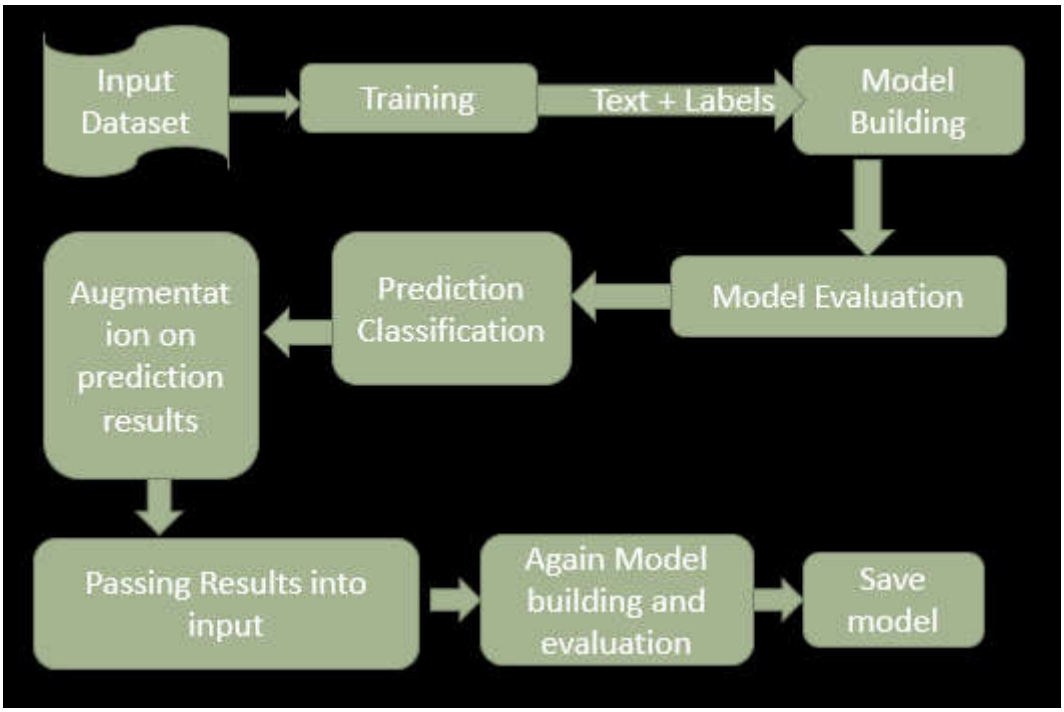

Figure 2: Dataflow Diagram Of Development Process

The objective of this implementation is to show observations for enhancing the development of classification followed by augmentation. The implementation is divided into these many stages:

a. **System Requirement:** Minimum of 8GB RAM, TensorFlow GPU runtime, google colab.

b. **Dataset Analysis:** The dataset taken here is from a wiki corpus which was rated by human raters for toxicity. The corpus contains 63M comments from discussions relating to user pages and articles. For preprocessing libraries like pandas, NumPy, matplotlib, seaborn, sklearn, TensorFlow, re, tqdm, word clouds, stopwords, etc, are required. From the analysis, it is observed that classes are imbalanced and the data is huge. Imbalance data hamper the accuracy of a model So to faster training and balance the dataset, we have downsampled the majority class. Downsampling means training on a disproportionately low subset of the majority class examples. After this, the distribution of data needs to be checked to analyze whether it is skewed or not. So now, to preprocess the data by checking for missing values.

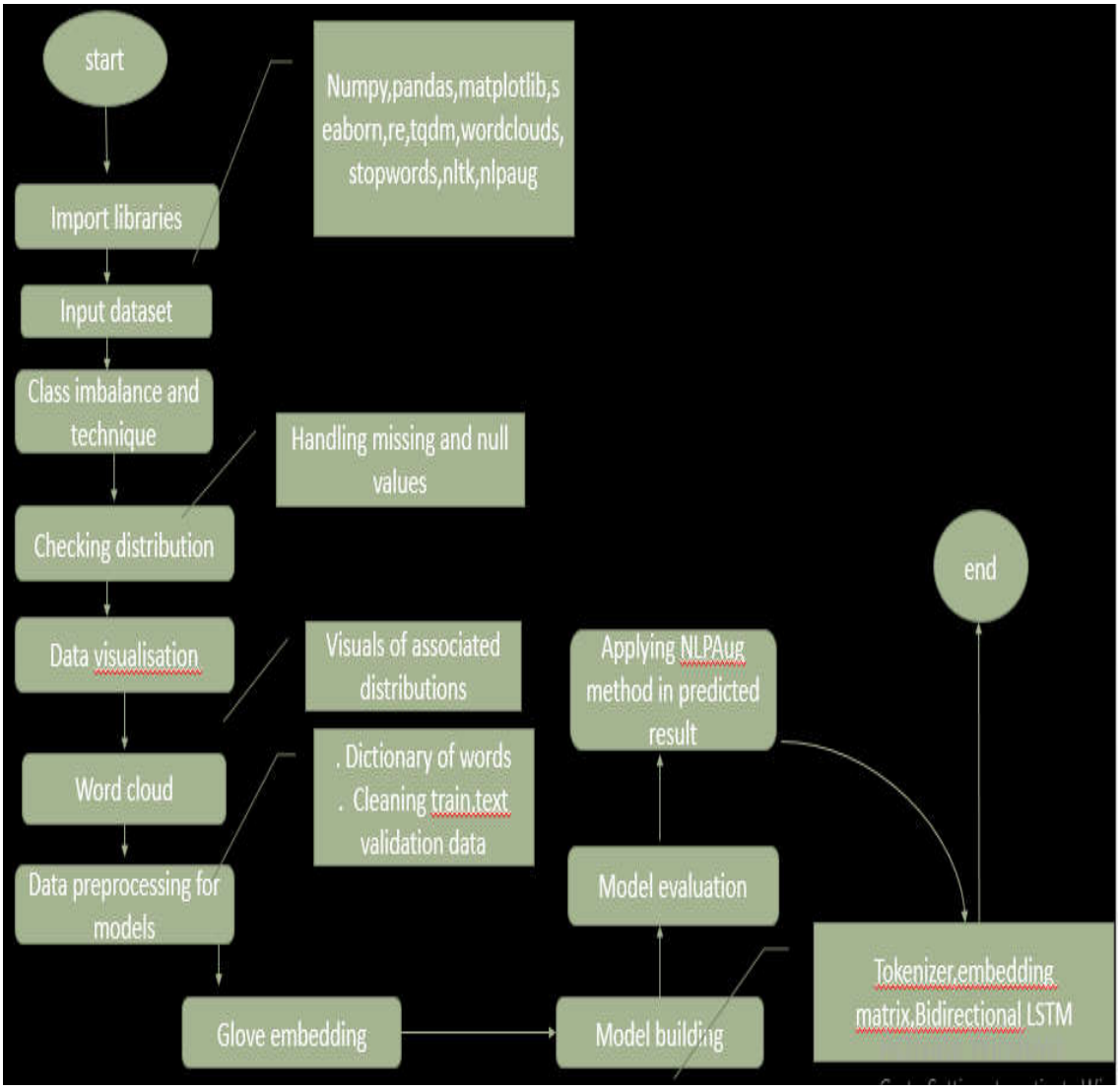

Figure 3: Flow Diagram Of Model

c. **Data Visualization:** In this section, dataset variations, plots, heatmap, and wordcloud are presented to massage the data. Wordcloud is a visualization tool for texts that are used to visualize keywords from texts and also the size of a particular word becomes huge when there is a higher frequency count for each word. So here, words are visualized based on the frequency of comments.

d. **Model Building:** For this step, firstly highlight the text properties and feature creation. So the property is like it is sequential data which is linearly separable high dimensional data. For creating a feature remove the stops words, and special characters, and create feature vectors. The model used is a Bidirectional LSTM recurrent neural network. After defining this all, cleaning up data. now dividing training and testing data into 80-20 percent. Taking training data and tokenizing into sequences and defining a dictionary of words. create an embedding matrix using words of vectors i.e., glove.6B.200d which has 4lac vectors. Now coming to the recurrent neural network model i.e., Bidirectional LSTM where dense layer activation layer and the stochastic gradient is used as optimizer. Here softmax function is RELU, max pool layer is applied.

e. **Model Compilation And Prediction:** For model compilation, binary cross-entropy loss function and adam optimizer are taken. prediction is performed on a validation set.

f. **Text Augmentation In A Predicted Comment:** so finally, here comes augmentation. the approaches discussed above are applied like using NLPAug, and parrot framework, and with these models, accuracy is far better and with speedy performance.

Methods	Accuracy
1.BiLSTM + RNN	92%
2.USING NLPAug + BiLSTM	95%
3.Naïve Bayes classification	94%
4.TF-IDF NB-Logistic classification	98%
5.NLPAug TF-IDF	98%
6.CNN + Tensorflow (Text Classification)	95%
7.NLPAug CNN	94.06%
8.EDA	98%
9.Augmentation using paraphrasing	97%

Table 1: Observations On Following Algorithms

7. RESULT AND CONCLUSION

In this paper, comprehensive and structured implementation of text augmentation for NLP is highlighted. Simple data augmentation operations can boost performance on text classification tasks. Improving model performance for classification by replacement with synonyms, random words, and antonyms. Parrot framework which is used for augmenting text data is to be achieved. Leveraging the NLP cloud APIs such as spacy, google neural machine translation, wordnet, nltk, etc. Augmentation with google tool Augly. All the techniques are experimentally performed and later on can be used with an applied area of computer vision.

8. REFERENCES

1. Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton: "ImageNet Classification with Deep Convolutional Neural Networks"
2. Connor Shorten and Taghi M. Khoshgoftaar: "A survey on Image Data Augmentation for Deep Learning"
3. Sosuke Kobayashi: Contextual Augmentation: "Data Augmentation by Words with Paradigmatic Relations"
4. Jason Wei, Kai Zou : "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks"
5. Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, Eduard Hovy: "A Survey of Data Augmentation Approaches for NLP"
6. Ashutosh Adhikari, Achyudh Ram, Raphael Tang, Jimmy Lin:"Rethinking Complex Neural Network Architectures for Document Classification"
7. Varun Kumar, Ashutosh Choudhary, Eunah Cho: "Data Augmentation using Pre-trained Transformer Models"
8. Ateret Anaby-Tavor, et al.: "Not Enough Data? Deep Learning to the Rescue!"
9. Franco M. Luque: Atalaya at TASS 2019: "Data Augmentation and Robust Embeddings for Sentiment Analysis "
10. Shuxiao chen : "A group-theoretic framework for data augmentation 2020 "
11. Vukosi Marivate, Tshephisho Sefara: "Improving short text classification through global augmentation methods 2019"
12. Zhilin Yang: XLNet: "Generalized Autoregressive Pretraining for Language Understanding 2019"
13. Jacob Devlin: BERT: "Pretraining of Deep Bidirectional Transformers for Language Understanding 2019"
14. Sam Shleifer: "Low resource text classification with ulmfit and backtranslation 2019"

15. Jeffrey Pennington, Richard Socher, Christopher Manning: GloVe: "Global Vectors for Word Representation 2014"

16. Simon Tong: "Support vector machine active learning with applications to text classification 2002"

17. Pengfei Liu: "Recurrent neural network for text classification with multi-task learning 2016"

18. Marzieh fadaee: "Data augmentation for low-resource neural machine translation 2017"

19. PrithivirajDamodaran: https://github.com/PrithivirajDamodaran/Parrot_Paraphraser

20. Edward Ma: Github repository https://github.com/makcedward/nlpaug