|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

**Comparative Study of Various NLP Algorithms**

Shivam Pandey  
Computer Science Engineering (Core)  
SRM Institute of Science and Technology Chennai, India  
[sp7287@srmist.edu.in](mailto:sp7287@srmist.edu.in)

Mayank Sinha  
Computer Science Engineering (Core)  
SRM Institute of Science and Technology Chennai, India  
[ms7651@srmist.edu.in](mailto:ms7651@srmist.edu.in)

***Keyword:***

*classification, algorithm, dataset, NLP, training, accuracy.*

**ABSTRACT**

This document is a brief information about our research that we are doing upon various NLP algorithms and classifications and our task is to identify the shortcomings in the algorithms that are being used in the current world*.*

1. **INTRODUCTION**

The field of Natural Language Processing currently has a variety of algorithms which are used for the toxic language detection, but these algorithms are not devised keeping in mind the production environment. In the production environment, when the algorithm runs in real-time then the speed of classification has a severe impact on the application usage. If the algorithm is slow in its’s classification task, then the algorithm is good only for theoretical usage. In real world we require algorithms which are highly accurate and fast in terms of operations. We aim to devise an algorithm which is viable in real world i.e., the algorithm is fast and accurate at the same time and there is no trade-off between two. We will be conducting a study between various NLP algorithms, and we will be testing the trained classifiers based on variety of parameters like accuracy, recall, precision, F1 score, time taken for prediction. Based on the parameters we will define what are shortcomings in various NLP algorithms.

1. *Literature Survey*

In artificial Intelligence Natural Language processing (NLP) is used in machines to understand and to solve and get the proper meaning of human language in an easy way and more understandable way. Therefore, we can conclude that most of the coders can perform task or organize automatic paragraphing, decoding text, or languages, having the sentiment analysis, speech etc. Therefore, it is often considered as a huge problem in the world of computers as human language is very rarely precise and very often spoken plainly and to understand this human language and not only the words but also the way it has been said is a very difficult task and how they are linked together to create a meaningful sentence and the emotion behind the sentence. Therefore, Despite the language, spoken by the people is the easiest thing for us human beings but ambiguity of NLP is what makes it a difficult problem for computers to master.



1. *Neural conversational model*

Conversational modelling can use the mapping of a sequence to another sequence.

1. *Neural approaches to conversational AI*

This paper provides Neural Network for 3 different scenarios which are Q/A task completion, social chat.

1. *Language models are unsupervised multitask learners*

A language model which learns without any explicit supervision when trained on a new dataset.

1. *XLNET: Generalized autoregressive pretraining for language understanding*

XLNET enables learning bidirectional contexts by maximising the expected likelihood permutations of the factorization order

1. *Classification of useful comments in stack overflow*

Detection of useful comments of stack overflow platform

1. *Research on text classification based on BERT-BIGRU model*

The usage of BERT model for the text classification helps to better classify the data which we are dealing with the modern world.

1. *Identification of toxic comments using machine learning*

Lemmatization of word for the generation of root word.

1. *Toxic comment classifications*

Multiple deep learning architecture to learn semantic embedding

1. *Toxic comment detection: Analyzing the combination of text and emoji*

Emojis also play a role in the classification of toxic text.

1. *Long short-term memory recurrent neural network architectures for large scale acoustic modelling*

LSTM is specific RNN which takes into the account the complete sequence of input feeded.

1. *BERT-PRE-Training of deep bidirectional transformers for language understanding*

BERT is designed to pre-train deep bidirectional representation from unlabelled text by jointly conditioning on both left and right context in all layers.

1. *Comparative study of CNN and RNN for natural language processing*

This paper explores two most famous DNN for the NLP task which are CNN and RNN.

1. *A critical review of recurrent neural networks for sequence learning*

Recurrent neural network is a superset of feedforward neural networks, augmented with the ability to pass information across time.

1. *ALBERT: a lite Bert for self-supervised learning of language representation*

This is a modified version of traditional Bert model. This model incorporates two parameters reduction techniques which are factorized embedding parameterizing and cross layer parameter.

*3. Existing Algorithms and Drawbacks*

1. *Logistic Regression*

* Fails to classify complex patterns
* Assumes linearity between variables

1. *Decision Tree*

* Highly sensitive towards small data changes
* Cannot handle large dataset
* Sensitive towards biasness

1. *Naïve Bayes*

* Requires Large Dataset
* Assumes all predictors are independent (rarely seen)

1. *Random Forest*

* Results of random forest are very tough to interpret
* Very slow for real time predictions.

1. *K nearest Neighbour*

* Selecting the value of K is a complex task
* Cannot handle large dataset
* Sensitive to outliers

1. *ANN (Artificial Neural Network)*

* Deep neural Network suffer from vanishing gradient problem
* Suffers from exploding gradient problem

1. *LSTM (Long short-term memory)*

* The model is prone to overfitting
* It is difficult to apply the dropout algorithm to curb the overfitting problem.

1. *RNN (Recurrent Neural Network)*

* Computation of this neural network is slow
* Faces vanishing gradient and exploding gradient problem.

1. *Auto Encoders*

* The algorithm becomes too much specialized to a specific use-case.
* The algorithm is very lossy.

1. *Transformers*

* The training of transformers is very complex process.
* The interpretation of result is a very tough task.

*4. Research Gap*

* The field of natural Language Processing is of high interest amongst the researchers. Researchers are working consistently towards devising new algorithms and improving the current algorithms.
* The current research does not consider the practical implementations and working of the algorithm in production environment.
* We will test the NLP models in the production environment using various metrics to judge the performance of these algorithms. This will indicate the variety of algorithms to be used to solve the real-world problems

.

*5. Requirement gathering*

* For this project we need a huge, annotated dataset of toxic comments labelled as toxic and non- toxic.
* We will need to clean this data for our usage.
* We will train multiple NLP classifications models on this dataset and then we will test our models in real world scenario to identify their shortcomings.
* We will use 2 major sources for our data gatherings:
* Kaggle-> it is an online repository which is used by data scientist for gathering data according to their requirements.
* Twitter API-> This is an API which is provided by twitter for the data collection. It is both unpaid and paid depending on the data needs of user. We will use the unpaid service; we will collect tweets which are labelled as toxic and non-toxic.

*6. References*

1. A Neural Conversational Model, BAHDANAU D, CHO K,BENGIO Y,Google Research.
2. Neural approaches to conversational AI, Jianfeng Gao, Michel Galley, Lihong Li, Microsoft Reseach.
3. Alec Radford, Jeffrey Wu,Rewon Child , Language models are unsupervised multitask learners, OpenAI SanFrancisco.
4. Systematic Exploration and Classiﬁcation of Useful Comments in Stack Overﬂow Prasadhi Ranasinghe, Nipuni Chandimali, Chaman Wijesiriwardana Faculty of Information Technology University of Moratuwa Katubedda, Sri Lank
5. Classification of Online Pernicious Comments using Machine Learning Aniket L Sulke1 Akash S Varude2 1,2Researcher 1,2MIT Academy of Engineering, Pune, India
6. Research on text classification based on BERT-BiGRU MODEL,Qing Yu, Ziyin Wang and Kaiwen Jaing,Tianjin Key laboratory of intelligence,China
7. Cyberbullying Detection: An Ensemble Based Machine Learning Approach Kazi Saeed Alam Department of Computer Science and Engineering Khulna University of Engineering & Technology Khulna, Bangladesh [saeed.alam@cse.kuet.acbd](mailto:saeed.alam@cse.kuet.acbd)
8. improving language understanding by generative pre-training.alee radford,karthik narasimhan,tim salimans,ilya sutskever.albert: a lite bert for self-supervised learning of language
9. Hasim sak, Andrew senior, francoise beaufays, long short term memort recurrent neural network architectures for large scale acoustic modeling, google brain.
10. Jacob devlin, Kenton lee, bert pre-training of deep bidirectional transformers for language understanding, google brain.
11. Wenpeng yin, katharina kann, mo yu, comparative study of cnn and rnn for nlp ,IBM research USA
12. Zachary C liption, a critical review of recurrent neural network for sequence learning,universirt of san diego.
13. mingda chen,kevin gimpel,piyush sharma,radu soricut.google research toyota technological institute at chicago.
14. emotion-cause pair extraction:a new task to emotion analysis in texts.riu xia,zixiang ding. nanjing university of science and technology,chin
15. Authoir name,title, source,vol,issue,year