Multilingual Sentiment Analysis Based on Convolutional Neural Network and Bert

Aniket K. Shahade[1*], K. H. Walse[2], V. M. Thakare[3]

Abstract

The large amount of text type data conveyed among people across various mediums has increased as a result of the development in technology. Sentimental analysis is made possible by the accessibility of such various reviews or sentiments of people's emotions. Moreover, because there is a lack of standardised labelled data in the English, Hindi, and Marathi language area in natural language processing for sentiment classification and analysis is made significantly difficult. The most of the previous study in English, Hindi, and Marathi has focused on deep learning approach that heavily emphasise context-independent embeddings. Examples of these approaches include Word2Vec that give word for fixed appearance regardless of such contexts. In this paper demonstrated a cutting-edge classification accuracy for English, Hindi and Marathi sentiment analysis as a consequence, which greatly exceeds all models and techniques.

Key Words: Sentiment Analysis, Machine Learning, Deep Learning, Emotion, Sentiment Classification

DOI Number: 10.14704/nq.2022.20.8.NQ44712 **NeuroQuantology 2022; 20(8):6860-6867**

Introduction

The act of analysing a text or review to predict whether a person might feel about an event or perspective is known as sentiment categorization. Typically, text polarities are used to analyse the sentiments. A sentiments classification usually divides data into positive, negative, or neutral categories [1]. The core of sentiment classification is sentiment retrieval, and extensive research has been accepted in this feild. Sentiment analysis is a key subsequent stage, and it has grown significantly in previous years along with the global expansion of text data. Today, users communicate their opinions digitally on a variety of subjects, such as digital books or movie or product reviews, political comments, and other customer reviews. Determining everyone's intents consequently requires weighing many perspectives of viewpoint. In generally, sentiments describe two main idea types positive or negative across many venues in which the weight of popular opinions is relevant.

For instance, in responding to user input, food providers and online retailers regularly improve their services. For instance, ola and uber is most well-known ride sharing business model in India, uses customer opinions to enhance its offerings. The challenge there, though, is individually navigating among the comments, that requires a great deal of research and energy. Through classifying the sentiments polarity connected to a person 's viewpoint, Automated Sentiments Detection can tackle this problem. In light of one's feedback, this makes it possible to make decisions that are better aware. It can also be used in a number of NLP scenarios, including chatbots [2].

The field of Language processing has emerged as an outcome of multiple ground-breaking discoveries and scholars' steadfast work. As computational performance and the amount of publicly available information on the Internet have grown, deep learning methodologies have grown in popularity in previous decades. The word embedding has been utilised as a basic level in a number of machine and deep learning techniques because it enhances the effectiveness of neural models and the effectiveness of deep learning techniques. Early efforts to use sentiment classification in English, Hindi and Marathi relied on non-contextualized word

Corresponding author: Aniket K. Shahade
Address: [1]Research Scholar, PG Dept. of CSE, SGBAU, Amravati, [2]Professor, SGBAU, Amravati, [3]Professor & Head, PG Dept. of CSE, SGBAU, Amravati

embeddings, that show a list of fixed word representations without taking into account various different situations in that words might appear. Moreover, the concept mapping method is greatly amplified by the Bidirectional Encoder Representations from Transformers is a new emergence phenomenon [3]. BERT has emerged as the more amazing Natural language paradigm, that can accomplish wonderfully in every Natural language operation with appropriate fine tuning for particular downstream applications when the pattern shifted towards convertible designs made up of attentiveness components. BERT is a fully multimodal language architecture that has to be built on a sizable English, Hindi and Marathi [4]. There is a general mBERT framework that covers total 106 different languages [5]-[6]. The investigators created a particular language-based BERT framework which works quite comparable to the conventional BERT framework because it works poorly on some other languages activities. Like a result, utilize the improved BERT approach for sentiment classification in English, Hindi and Marathi Tweets. The main aim of this paper is presented as given below:

This paper presents the hybrid approach to overcome the problem in sentiment analysis for multilingual English, Hindi and Marathi.

To present the comparative study of proposed procedure with the existing methods.

To do this, design the English, Hindi and Marathi pre-trained BERT model.

To address the issue in sentiment classification is determining the polarization of a person's perspective. A textual corpus's elements are extracted, a classification architecture is developed, as well as its effectiveness is evaluated as part of the Sentiment analysis approach [7]. Such form of standard operating method has been used for a number of sentiments classification, such as the categorization of movies reviews [9], product opinion [10], and political reviews [8]. etc.

In [11] Suggested the model with performance score of 94 percent for identifying positive and negative tweets for restaurant comments. Whenever the investigators used SVM model to investigate the sentiment analysis about the smartphone tweets with the accuracy score 81percent [12].

Sentiment analysis on Twitter data for the Portuguese language has been defined [13]. Ombabi et al. demonstrated a sentiment classifier for the Arabic language with an accuracy of 90.75%, outperforming the state of the art [14]. The blending of two languages to create a new one is a regular occurrence in NLP. Such work has been conducted on vernacular Singaporean English, a product of the coalescence of Chinese and Malay languages [15]. However, the majority of efforts on sentiment categorization focus on English and other widely spoken languages. The biggest constraint on Bengali sentiment analysis research is a lack of appropriate resources and datasets. According to unrivalled bidirectional and attentiveness strategy, the BERT concept garners the greatest interest amongst all [16]. Therefore, investigators were monitoring how it affects subsequent Nlp techniques. Because language specific BERT systems performed better than standard mBERT approaches [17], investigator applied its own BERT approach for particular language to increase sentiment accuracy because BERT is only taught in English. Several studies have demonstrated exceptional performance outcomes in sentiment analysis [18]-[19], which aims to find components and its associated viewpoints. To assess the sentiment classification problem, several investigators was developed the novel approach based on BERT in the field of sentiment analysis [20].

Proposed Methodology

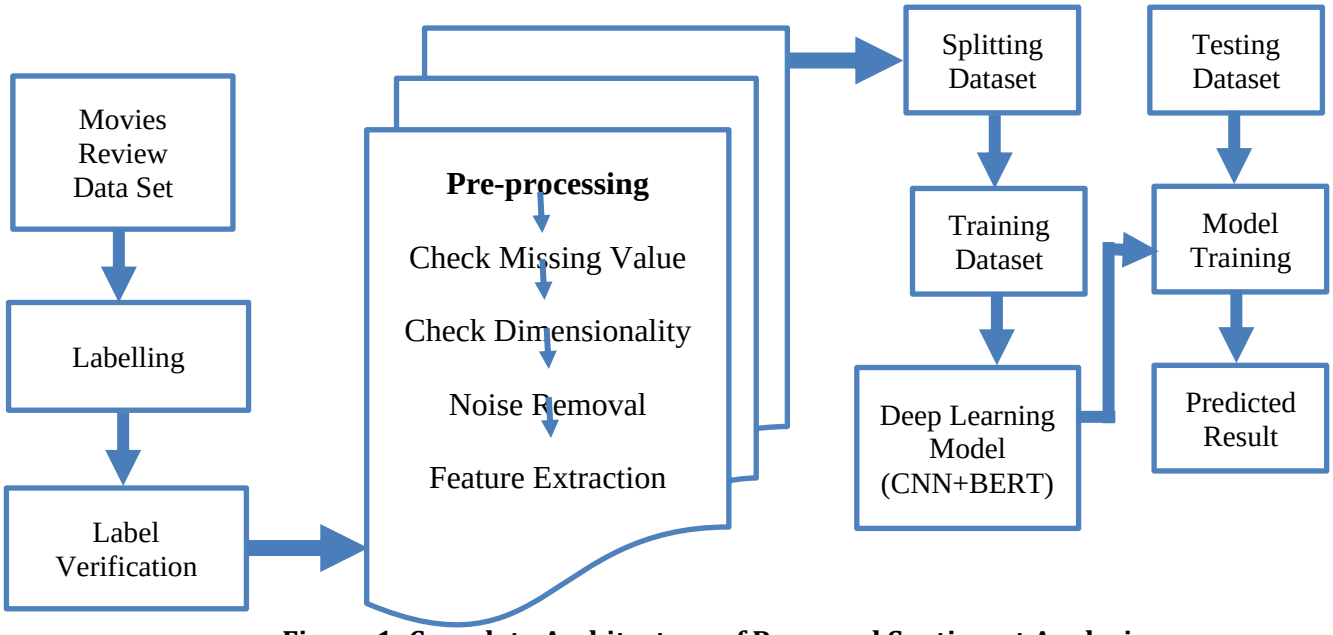

Figure 1: Complete Architecture of Proposed Sentiment Analysis

Figure 1 shows the complete Architecture of Proposed Sentiment Analysis for the movie's reviews. It shows the various steps such as gathering dataset, data labelling, pre-processing splitting the dataset, designing the model, train the model, and finally get predicted results.

Data Gathering

In this study obtained information for the collection from a variety of domains, such online media sites and blogs wherein people post their highly appreciated reviews. The majority of the information were acquired from posts on social media sites like Facebook, Twitters etc. Online movie reviews have moreover developed into a sizable component of internet advertising. Like a result, acquired information on audience reviews of movies.

Table 1: Example of collected tweets in English

Tweet	Polarity
Great Movie	Positive
Awesome Movie	Positive
Waste of Time	Negative
Fine	Neutral

Table 2: Example of collected tweets in Hindi

Tweet	Polarity
फिल्म बहुत अच्छी है	Positive
फिल्म मुझे अच्छी लगी	Positive
फिल्म अच्छी नहीं लगी	Negative
फिल्म ठीक है	Neutral

Table 3: Example of collected tweets in Marathi

Tweet	Polarity
कौतुकास्पद आहे	Positive
योग्य आहे	Positive
कौतुकास्पद नाही	Negative
कौतुक करावे की नाही	Neutral

The movie contents are 273 positive movie reviews, 240 negative movie reviews, and 205 neutral movie reviews for the Hindi language. The Twitter API then extracts the 1,600,000 tweets for the English language. The tweets can be annotated as 0 negatives and 4 positives and using 6 fields as a target, ids, date, flag, user and text can be utilized for sentiment detection. This target contains the tweet polarity as 0 negative, 2 neutrals, and 4 positives. The ids are a polarity of a tweet which consist of 2087 tweets. Data is described as the data of the tweet, flag as a query. The user describes the user that tweeted, and the text is denoted as the tweet text. Then the Marathi language consists of an 18,378 tweets dataset. The most comprehensive Marathi SA dataset accessible to date. The manually labelled Marathi tweets dataset was presented. Which contains 1 positive, -1 negative, and 0 neutral, is depicted in the following table 4.

Table 4: Dataset Description

Language	Dataset	Type of Classification	Content
Hindi	Kaggle	Movie Review Text Classification	273 Positive Movie Reviews 240 Negative Movie Reviews 205 Neutral Movie Reviews
English	Kaggle	Tweets from Twitter	The Twitter API extracted 1,600,000 tweets. The tweets have been explained (0 negatives and 4 positives) and can be employed to discern sentiment with the fields such as target, ids, date, flag, user and text
Marathi	L3CubeMahaSent	Marathi tweets	There are a total of 18,378 tweets in this dataset, which are divided into three categories - Positive (1), Negative (-1) and Neutral (0).

Data Preprocessing

In Deep learning classification, data preparation is crucial step because the neural network accuracy is highly reliant on the calibre of the dataset input. Such process is used to get data ready for automation. The numerous steps required for data preprocessing are described in the following sections.

Check Missing Values

Handling the unknown data point in dataset was where started the data processing step. There are run into two different categories of incomplete data. Many of the lack important data, whereas others offer fewer details than is necessary. If all data was missing, the complete data was discarded by wiping out the complete rows. When there were not enough data, some data was manually changed based on a comparable observational data.

Noise Removal

Update all the dataset using reducing sampling noise in dataset after accounting for incomplete data. Noise includes English, Hindi and Marathi words or symbols, special characters symbols, and emojis etc. Despite the fact that emojis are capable of conveying a wide range of expressions, observed which only a minor portion of data comprises emojis. Therefore, removing emojis is part of the cleaning procedure.

Feature Extraction

Word embedding, another name for extracting features, defines words so that associated keywords are interpreted correctly [21]. In order to determine that phrase extracting method works the best on sentiment in the English, Hindi, and Marathi language, we used four different methodologies in this investigation.

Encoding Technique

To train the suggested model using the word embedding technique after preprocessing the data [22]. In this case separately evaluated the usefulness of suggested CNN+ BERT model based on the various parameters such as window size vector size and total number of iterations perform. The Word2Vec design formulation in Gensim, an effective toolset for carrying out a range of common NLP tasks, was utilised to construct the designs, whilst the Huggingface tool was utilized for training CNN+BERT model.

Word2Vec

In this paper used word2Vec approach for embedding the text. This approach used machine learning model to evaluate how semantically comparable the word contexts [23]. There are two different model are implemented by word2Vec approach. first is CBOW and second is Skip-Gram [24]. The equation 1 shows the maximum average logarithmic probability of word are as follows:

$$ES = -\frac{1}{v}\sum_{v=1}^{V}\sum_{-c\leq m\leq c, m\neq 0} log\ log[p(w_{v+m}|w_v] \quad 1$$

Input fed to neural model using the $w1, w2, w3 \ldots wN$ layer. C represent the context or window size of word. E represent the embed size. To measure the probability of input word using the equation 2.

$$p(0) = \frac{exp(u_i^T.u_0')}{\sum_{v\in V} expexp(u_z^T.u_0')} \quad 2$$

where, V shows the amount of vocabulary used to train the model and u and v shows the input/output vector respectively.

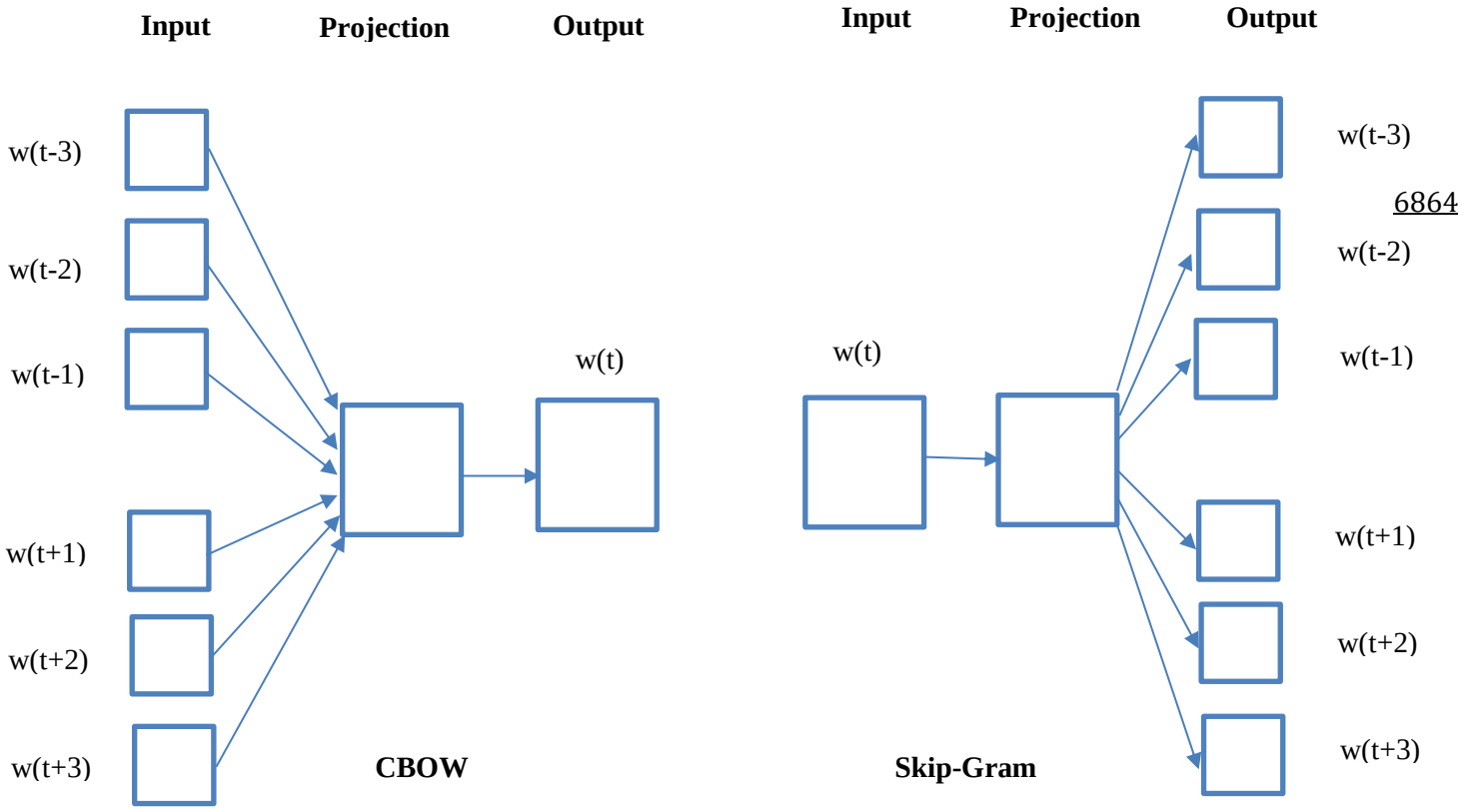

Figure 2: Conceptual framework of CBOW and Skip-Gram model

Bert

The suggested CNN+BERT framework that has been trained on a large on English, Hindi, and Marathi dataset for computing the sentiment analysis study. Since suggested CNN+BERT framework performance were compare with Devlin's approach [4]. This CNN+BERT model shows how cutting-edge outcomes of all earlier findings.

BERT encoder is the main part of neural network that operate on both directions to process the text and get meaning of every word in sentence or phrases [25]. It contains three crucial steps for

eISSN 1303-5150

preprocessing the data. When a token is numericalized (also called as tokenized) is linked to a specific value in the textual vocab. In padding takes as input a phrase of a specific Size. The size of a phrase is typically determined by the facts that dealing with. Phrases which are short than upper limit will need to have paddings added to them to cover up the difference in size. The output is largely interconnected from the entire number of blocks of the encoder. Establishing connections between input data and encoding data in the outcome are the responsibilities of a given blocks in model.

CNN Classifiers

For the classification sentiment polarity assessment used convolutional Neural Network that consist of various convolutional layers, weights and pooling layers. In proposed CNN+BERT model used local correlation approach to extract the relevant features.

Convolutional Layer

It is the main part in CNN that perform multiplication operation of every segment of input words, then adding the bias and transfer using activation function for mapping the features for next layer. If the sample input data are $t_i^0 = (t_0, t_2 \dots \dots, t_n)$ where, n is the number input samples. The final result is measured using following equation

$$C_i^{lj} = f\left(d_j + \sum_{p=1}^{P} w_m^j \, x_{x+m-1}^{0j}\right)$$

Where, f represents activation function of convolutional layer, d represent the bias, j represents mapping the feature, p represent filter, l represents index of layer.

Batch Normalization

The CNN+BERT model is train using the number of batches that contain the data input. Like a reason, every training process requires fitting the sample data distributions based on various network parameters configure, which significantly slows down models' development. A dynamically adjust all the parameters technique called batch normalisation is used to address such problem after a convolutional layer. For the training the model batch normalization measures the mean and variance of every data batch. following are the mean and variance formulas

$$u_d = \frac{1}{p}\sum_{i=1}^{p} x_i$$

$$\sigma_d^2 = \frac{1}{p}\sum_{i=1}^{p} (x_i + u_d)^2$$

Where, u_d and σ_d^2 represents the mean and variance of input data batch.

Max Pooling Layer

The suggested approach used 1-dimensional max-pooling layers for batch normalization and also used down-samples the feature to make them smaller [26]. It gathers discrete, small input pieces and generates a unique result for each one. There are numerous ways to go about doing this. In this investigation, the greatest value among a group of nearby data inputs is found using the Max-pooling method. A mapping the features pooling in layers is determined by

Performance Evaluation

In this study use python to build the deep learning model based on CNN and BERT. Since this dataset is significantly bigger than typical and necessitated the construction of suggested frameworks, employed Google Collaboratory as GPU assist. Deep learning framework, meanwhile, were implemented using Sci-kit-learn and Keras accordingly. In this model added Impact Intelligence and deep learning architecture. After the model training was finished, adjust major hyper - parameters to provide a fine-tuned network with comparable evaluation parameters. The suggested model is estimated using test dataset. The parameters for the evaluation such as accuracy Score, Precision, Recall, and f1-score are used to look over the performance of suggested model.

Result Analysis

Here, Word2vec with CNN+BERT classification performed nearly indistinguishably from one another. Word embedding, utilising the most effective machine and deep learning methods, increases performance by 3% and achieves an 97% accuracy score. Therefore, fine-tuned multilingual English, Hindi and Marathi using the various machine and proposed CNN+ BERT classification models outperform better. According to these experimental results, fine-tuning multilingual English, Hindi and Marathi using CNN+BERT led to a significant improved performance over the previous

embedding strategy. The improved performance is considerably better, which opens doors for further improvements.

Table 5: Performance of Various Machine Learning Classifiers and Proposed CNN+BERT Model

Classifiers	Accuracy	Precision	Recall	F1-Score
SVM	92%	93%	92%	92%
Adaboost	90%	91%	92%	91%
MLP	89%	90%	90%	91%
Random Forest	94%	95%	94%	93%
Decision Tree	93%	94%	93%	92%
Proposed CNN+BERT	97%	98%	97%	96%

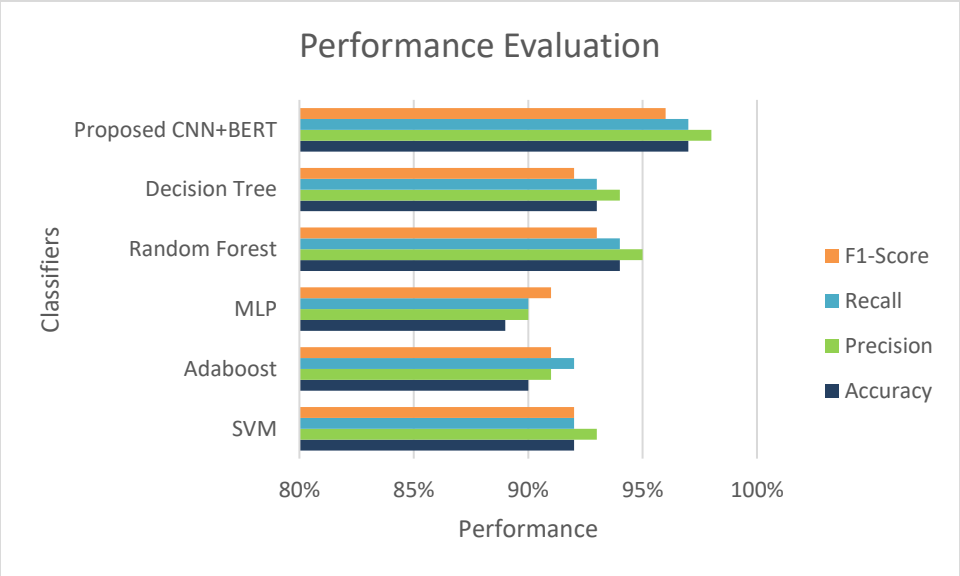

Figure 3: Bar represent the classification performance of various machine learning models with proposed CNN+BERT model.

Conclusion

In order to classification of texts as per respective themes, this research analyzes and examines several machine learning and proposed deep learning techniques. It shows that Deep learning has significant development in NLP, can outperform most prior systems. In this study, demonstrated the critical contribution of Deep CNN+BERT model to sentiment analysis for multilingual classification. For categorization challenge, a CNN+BERT deep model was developed for sentiment analysis. The proposed CNN+BERT model was implemented using open-source Python with the help of Google colab packaged. Since deep learning requires a considerable quantity of three dataset, we'll keep working to increase the database. Finally, it is found that CNN+BERT results in a remarkable accuracy score of 97%. Therefore, among the other machine learning methods proposed model has better outcome. In this study used three unbalanced datasets of multilingual English, Hindi and Marathi. In future development, plan to apply the advanced deep learning technique on a database that is better balanced and detailed and also provide a method for evaluating how well the suggested deep model performs in actual application.

References

Prottasha, N.J.; Sam, A. A; Kowsher, M.; Murad, S.A.; Bairagi, A.K.; Masud, M.; Baz, M. Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning. Sensors 2022, 22, 4157.https://doi.org/10.3390/s22114157

Kowsher, M.; Afrin, F.; Sanjid, Z.I. Machine Learning and Deep Learning-Based Computing Pipelines for Bangla Sentiment Analysis. In Proceedings of the International Joint Conference on Advances in Computational Intelligence, Online, 23–24 October 2021; pp. 343–354.

Kowsher, M.; Tahabilder, A.; Sanjid, M.Z.I.; Prottasha, N.J.; Sarker, M.M.H. Knowledge-base optimization to reduce

the response time of bangla chatbot. In Proceedings of the 2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan, 26–29 August 2020; pp. 1–6.

Rogers, A.; Kovaleva, O.; Rumshisky, A. A primer in bertology: What we know about how bert works. Trans. Assoc. Comput.Linguist. 2020, 8, 842–866.

Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018, arXiv:1810.04805.

Libovickỳ, J.; Rosa, R.; Fraser, A. How language-neutral is multilingual BERT? arXiv 2019, arXiv:1911.03310.

Kowsher, M.; Uddin, M.J.; Tahabilder, A.; Amin, M.R.; Shahriar, M.F.; Sobuj, M.S.I. BanglaLM: Data Mining based Bangla Corpus for Language Model Research. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2-4 September 2021; pp. 1435-1438.

Dashtipour, K.; Gogate, M.; Li, J.; Jiang, F.; Kong, B.; Hussain, A. A hybrid Persian sentiment analysis framework: Integrating dependency grammar-based rules and deep neural networks. Neurocomputing 2020, 380, 1-10.

Kennedy, A.; Inkpen, D. Sentiment classification of movie reviews using contextual valence shifters. Comput. Intell. 2006, 22, 110-125.

Cui, H.; Mittal, V.; Datar, M. Comparative Experiments on Sentiment Classification for Online Product Reviews; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2006; Volume 6, p. 30.

Kouloumpis, E.; Wilson, T.; Moore, J. Twitter sentiment analysis: The good the bad and the omg! In Proceedings of the 5th International AAAI Conference on Weblogs and social media, Barcelona, Spain, 17-21 July 2011; Volume 5, pp. 538-541.

Krishna, A.; Akhilesh, V.; Aich, A.; Hegde, C. Sentiment analysis of restaurant reviews using machine learning techniques. In Emerging Research in Electronics, Computer Science and Technology; Springer: Berlin/Heidelberg, Germany, 2019; pp. 687–696

Singla, Z.; Randhawa, S.; Jain, S. Sentiment analysis of customer product reviews using machine learning. In Proceedings of the 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, 23-24 June 2017; pp. 1-5.

Souza, M.; Vieira, R. Sentiment analysis on twitter data for portuguese language. In Proceedings of the International Conference on Computational Processing of the Portuguese Language, Coimbra, Portugal, 17-20 April 2012; pp. 241–247.

Ombabi, A.H.; Ouarda, W.; Alimi, A.M. Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. Soc. Netw. Anal. Min. 2020, 10, 53.

Mathews, D.M.; Abraham, S. Social data sentiment analysis of a multilingual dataset: A case study with malayalam and english. In Proceedings of the International Conference on Advanced Informatics for Computing Research, Shimla, India, 15-16 June 2019; pp. 70-78.

Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. arXiv 2020, arXiv:2003.08271.

Dai, X.; Karimi, S.; Hachey, B.; Paris, C. Cost-effective selection of pretraining data: A case study of pretraining BERT on social media. arXiv 2020, arXiv:2010.01150.

Antoun, W.; Baly, F.; Hajj, H. Arabert: Transformer-based model for arabic language understanding. arXiv 2020, arXiv:2003.00104.

Farahani, M.; Gharachorloo, M.; Farahani, M.; Manthouri, M. Parsbert: Transformer-based model for persian language understanding. Neural Process. Lett. 2021, 53, 3831-3847

Misra, J. AutoNLP: NLP feature recommendations for text analytics applications. arXiv 2020, arXiv:2002.03056.

Gupta, S.; Kanchinadam, T.; Conathan, D.; Fung, G. Task-optimized word embeddings for text classification representations. Front. Appl. Math. Stat. 2020, 5, 67.

Grohe, M. word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Portland, OR, USA, 14-19 June 2020; pp. 1-16.

Shobana, J.; Murali, M. Improving feature engineering by fine tuning the parameters of Skip gram model. Mater. Today Proc. 2021, in press.

Rush, A.M. The annotated transformer. In Proceedings of the Workshop for NLP Open Source Software (NLP-OSS), Melbourne, Australia, July 2018; pp. 52–60.

Dmitrievich, I.A. Deep Learning in Information Analysis of Electrocardiogram Signals for Disease Diagnostics. Bachelor's Thesis, Moscow Institute of Physics and Technology (State University), Moscow, Russia, 2015; p. 20.