Nano World Journal

https://doi.org/10.17756/nwj.2023-s4-070

Research Article

Open Access

Nanoscale Text Classification with Bi-LSTM: Enhancing Precision

Shaik Mohammed Basha[1*], Anusha Balemla[1], Saniya Anjum[1] and A. Ramesh[2]

[1]*Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, Telangana, India*

[2]*Department of Mechanical Engineering, Vardhaman College of Engineering, Hyderabad, Telangana, India*

[*]Correspondence to:
Shaik Mohammed Basha
Department of Computer Science and Engineering,
Vardhaman College of Engineering,
Hyderabad, Telangana, India.
E-mail: bashakphb786@gmail.com

Received: September 19, 2023
Accepted: November 28, 2023
Published: December 01, 2023

Citation: Basha SM, Balemla A, Anjum S, Ramesh A. 2023. Nanoscale Text Classification with Bi-LSTM: Enhancing Precision. *Nano World J* 9(S4): S417-S420.

Published by United Scientific Group

Abstract

Vast volumes of data are produced from numerous sources, including websites and social media. It is necessary to extract significant information from text data, categorize it, and forecast end-user behavior or emotions. Natural-language processing research on text categorization is an example of a topic that organizes unstructured text input into useful category classifications. The Bi-Directional Long Short-Term Memory (Bi-LSTM) model, which has recently been used to a number of natural language processing (NLP) applications, is being employed in this study to classify sentences. LSTM models are more suited for text classification because they can capture long-term dependencies between word sequences. Our suggested approach performs better than the current Machine Learning models.

Keywords

Natural language processing, Bidirectional long short-term memory, Nanotechnology

Introduction

More people are starting to freely share their opinions online thanks to the internet's and social networks' recent rapid growth. As a result, a large amount of public comments data is produced online. For instance, reviews of merchandise are published for e-commerce websites like Jindong and Taobao, while reviews of accommodations are written for travel websites like C trip and E Long. Manually reviewing the comments is difficult due to their rap id increase. The application of artificial intelligence technology to mine the emotional tendencies of comment messages may be helpful in the age of big data to swiftly understand network public opinion. Finding the sentiment tendency in the remarks requires conducting sentiment analysis study [1-3].

Text categorization in the form of sentiment analysis incorporates NLP, machine learning, data mining, information retrieval, and other study areas. The sentiment orientation study of the comments corpus reveals how individuals have either favorable, negative, or neutral attitudes towards events or items, is the primary focus of sentiment classification of opinions [4, 5]. There are several ways to analyze sentiment, such as through examination of the comments made on blog posts, brands, entertainment, along with other methods. These comments represent the opinions of online users on a variety of products, hot topics, etc. Retailers can manage customer satisfaction by making relevant product remarks. By reading these product reviews, prospective customers can assess the products [6-9].

Text categorization serves as sentiment evaluation's main function, to which various words contribute in diverse ways. It is possible to acquire a phrase's

low-dimensional, non-sparse word vector form and essential for sentiment evaluation tasks. The word representation that is used the most is the Word2vec technology's distributed word vector. The phrase's meaning content is stored in the low-dimensional word vector. Word emotion information is absent from distributed word vectors, though. The traditional TF-IDF technique is used in this work to produce the word vector with strengths by integrating word's sentimental information [10].

Accurate short text classification in the dynamic field of nanotechnology is achieved through the powerful utilization of Bi-LSTM models. This approach involves meticulous data preparation, assembling labeled datasets of succinct nano-related texts, and employing advanced preprocessing techniques. The model culminates in a dense layer with softmax activation for efficient multi-class classification. Rigorous compilation with appropriate loss functions and optimizers precedes training, where datasets are intelligently split for validation. Hyperparameter fine-tuning refines the model's performance. This approach finds application across diverse domains, including document classification, research paper recommendation, and quality control in nanofabrication, illustrating its versatility and impact in advancing nano technology research and applications. Regular updates and adaptations ensure the model remains attuned to the ever-evolving landscape of nanotechnology.

A difficult issue during machine learning, information retrieval, and NLP is linguistic material categorization. A perfect illustration of this type of material is news articles. Because they frequently have sharp, consistent information. Daily news stories are produced in large quantities by the majority of news organizations; these articles should be sorted and categorized for different audiences. It takes a lot of time and effort to manually label news items. It becomes challenging for an application to manually classify and provide viewers with the most recent news stories in real time. To enable users to access the most recent news items that have been labelled, it is necessary to implement technology that can automatically categorize news items in real time [11-13].

Using technology to analyze news data in order to identify trends in news output and content is the fundamental purpose of news categorization. The value of topicalizing news stories has been demonstrated in numerous research and applications. Automated labelling of media articles on the internet, topic accumulation, and the development of news recommendation systems are all made possible by this textual interpretation and modelling, which is important in considering how the news and media affect society and politics. It also helps in the discovery of unidentified biases and justifications behind news stories.

In general, there are three types of text sentiment analysis methodologies now available: deep learning-based, machine learning-based, and a sentiment dictionary-based sentiment classification. In the sentiment dictionary-based method, the dictionary is used in finding words that express sentiment in the text and derive sentiment scores. The words sentiment possibility is then determined using the guidelines for sentiment calculation. The literatures introduced representative sentiment dictionary-based study. Words sentimental evaluation sentiment dictionary-based method is simple to implement and does not involve manual labelling of samples. However, the sentiment vocabulary has a significant impact on the analysis's quality. The majority of sentiment dictionaries suffer from issues including a dearth of specific topic terms and adequate sentiment word covering.

Pang's research was the first to use machine learning to analyze text sentiment. To analyze the sentiment of movie reviews, they employed the Support Vector Machine (SVM) algorithm, maximum entropy method, and naive Bayesian algorithm.

Experimentation

Proposed method

Overview of Bi-LSTM

Recurrent neural networks (RNNs) that are bi-directional are essentially just two distinct RNNs connected. This system allows the networks to retrieve facts regarding the order in each time step, forward as well as backward. The information you provide while using bi-directional will just be handled in two separate ways: one goes from the present to the future, whereas the second goes backwards from the future to the current. In contrast to unidirectional methods, these preserves prediction performance using LSTMs that run downstream. By incorporating the two hidden layers, you may conserve information from the past and the future at any given time.

Short text classification using Bi-LSTM

Finally, the experimental findings demonstrated that the SVM [9] method was more effective in handling the sentiment categorization of movie reviews. A semi-supervised classification technique based on graphs was proposed by Goldberg and Zhu [12], and it received ratings of 0 to 4.0 levels for both favorable and non-favorable remarks. It was suggested to use a sentiment analysis model with high-dimensional mixed feature built on SVM that is based on emoji, unfavorable traits, and sentiment features. Yet the caliber of the corpus with polarity labels determines the precision of the sentiment analysis approach employing machine learning.

Many academics have applied deep learning's sentiment analysis methodology with success in recent years. By synthesizing semantics on the syntax tree of binary sentiment polarity and producing positive sentiment analysis results in the data set of movie reviews, Socher's Recursive Neural Tensor Network model created a sentiment graph collection. In a bid to improve the sentiment classification of brief messages like twitter posts, the CharSCNN (Character to Sentence Convolutional Neural Network) model uses two convolutional layers to extract information from related phrases and keywords. A model that makes use of tree-structured long short-term memory networks that has produced successful results in sentiment analysis and semantic association. When added the attention mechanism to the LSTM, excellent outcomes were obtained from the SemEval-2017 Task4 Tweets sentiment classification.

We endorse this application since it helps to alleviate the limitations imposed by traditional and other existing approaches, making it a worthwhile system. In the suggested approach, a Bi-LSTM is used to draw knowledge from the intricate patterns in the data and classify the text. Because of the robustness of this method, accuracy is improved.

This proposed system has the following benefits: (i) Saving time, (ii) Low computational cost, (iii) Minimal complexity, and (iv) High precision.

Algorithm

Data: Input data sentences

1. Data loaded into the colab.
2. Reset the index of each sentence present in the data set.
3. Split the data set into train and test data.
4. Cleanup the train and test data by removing the stop words using NLTK.
5. Run Bi-LSTM on the train data and train the dataset.
6. Validate the test data set and calculate the accuracy.

Architecture diagram

Figure 1 presents the architecture diagram used.

Dataset description

The Bi-LSTM algorithm is trained and tested using the tweets data. The data is taken from the tweets and comments. The data is taken as the short text and that text is split into three attributes. The attributes are negative, neutral, and positive. Each short text in the dataset is divided based on the words used in the short text. The dataset consists of 179114 tweets and with three columns (Username, short text, sentiment). The dataset is divided into test and train dataset in the 7:3 ratio respectively.

Results and Discussion

Short text classification using Bi-LSTM

The short text classification is based on the feedback or tweets given. A data set is defined with three main attributes (negative, neutral, and positive). The final result analysis will be the classification of the sentence. Firstly, the sentence is given with the index starting from 1. The sentences are then cleaned up. For cleaning the sentences, we use the stop words removing method. After removing the stop words, sort the data as test and training data.

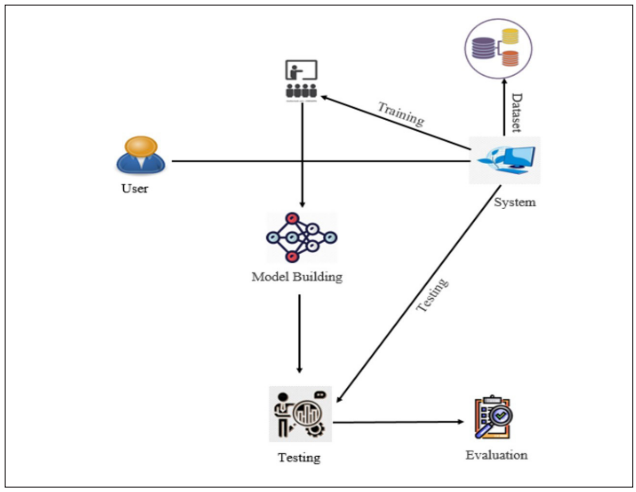

Figure 1: Architecture diagram.

Training the dataset

Divide the data set into the train data set by defining a split ratio. Use the following method to split the data.

Train, test = train_test_split(df1,test_size = 0.3,Stratify = df1.result.values)

Check the shape of the train data. After that clean up the data and redefine the data into two attributes (Clean text and Result).

Data validation

The testing information is utilized to confirm that the evaluation of the results is accurate. The test data is cleaned up and then redefined into clean text and result attributes. The test data is divided into 5 epochs and each epoch consists of 4478 data elements. The training accuracy and test accuracy will be analyzed with the help of the graph (Figure 2). Table 1 presents accuracy comparison.

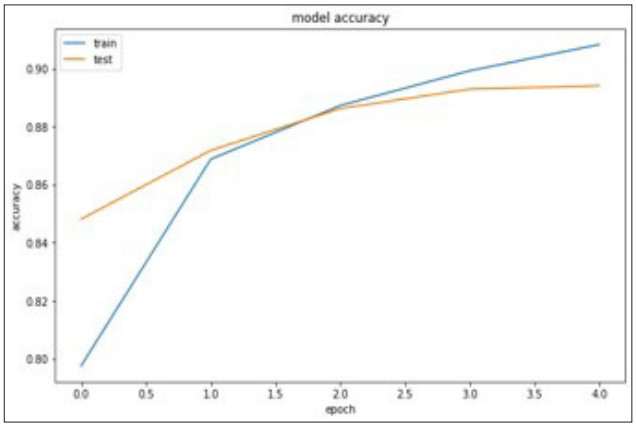

Figure 2: Test and trained data accuracies.

Table 1: Accuracy comparison table.

S. No.	Algorithm name	Accuracy
1	SVM	0.726
2	LSTM	0.897
3	Bi-LSTM	0.916

Conclusion

The research objectives of this project are to compare the performance of the models (accuracies) like Bi-LSTM, LSTM, RNN, BERT and SVM and proposes that our Bi-LSTM models outperforms all the other models in terms of accuracies. We have successfully developed a system to compare the accuracies of different models. This is created in a user-friendly environment with Python programming on Jupyter Notebook and colab.

Future Scope

We intend to evaluate the prognosis approach and label the concise text applying the most accurate and applicable machine learning algorithms using the latest data set.

Acknowledgements

None.

Conflict of Interest

None.

References

1. Vijayan VK, Bindu KR, Parameswaran L. 2017. A comprehensive study of text classification algorithms. In International Conference on Advances in Computing, Communications and Informatics, Udupi, Karnataka, India.
2. Shi K, Li L, Liu H, He J, Zhang N, et al. 2011. An improved KNN text classification algorithm based on density. In IEEE International Conference on Cloud Computing and Intelligence Systems, Beijing, China.
3. Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes L, et al. 2019. Text classification algorithms: a survey. *Information* 10(4): 150. https://doi.org/10.3390/info10040150
4. Learning multi-label topic classification of news articles.
5. Li S, Wang Y, Xue J, Zhao N, Zhu T. 2020. The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. *Int J Environ Res Public Health* 17(6): 2032. https://doi.org/10.3390/ijerph17062032
6. Carreira R, Crato JM, Gonçalves D, Jorge JA. 2004. Evaluating adaptive user profiles for news classification. In Proceedings of the 9th International Conference on Intelligent User Interfaces, Madeira, Funchal, Portugal.
7. Hao Z, Cai R, Yang Y, Wen W, Liang L. 2017. A dynamic conditional random field based framework for sentence-level sentiment analysis of Chinese microblog. In IEEE International Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing, Guangzhou, China.
8. Rehman ZU, Bajwa IS. 2016. Lexicon-based sentiment analysis for Urdu language. In Sixth International Conference on Innovative Computing Technology, Dublin, Ireland.
9. Manek AS, Shenoy PD, Mohan MC. 2017. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World Wide Web* 20: 135-154. https://doi.org/10.1007/s11280-015-0381-x
10. Mubarok MS, Adiwijaya A, Aldhi MD. 2017. Aspect-based sentiment analysis to review products using Naïve Bayes. *AIP Conf Proc* 1867(1): 020060. https://doi.org/10.1063/1.4994463
11. Bouazizi M, Ohtsuki T. 2017. A pattern-based approach for multi-class sentiment analysis in Twitter. *IEEE Access* 5: 20617-20639. https://doi.org/10.1109/ACCESS.2017.2740982
12. Goldberg AB, Zhu X. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In Proceedings of TextGraphs: The First Workshop on Graph Based Methods for Natural Language Processing, New York, USA.
13. Qiu X, Sun T, Xu Y, Shao Y, Dai N, et al. 2020. Pre-trained models for natural language processing: a survey. *Sci China Technol Sci* 63(10): 1872-1897. https://doi.org/10.1007/s11431-020-1647-3