Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Analysis of Indian and American poetry using topic modeling and Deep learning

K Praveen Kumar [a,*], Venkatrama Phani Kumar S [b], S.K. Lokesh Kumar [c]

[a] Department of CSE, Vignan's Foundation for Science,Technology and Research, Guntur, A.P, India
[b] Department of CSE, Vignan's Foundation for Science,Technology and Research, Guntur, A.P, India
[c] Department of Computer Science and Engineering, MLR Institute Technology, Hyderabad, India

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Indian Poetry
Stylistic features
Latent Dirichlet Allocation
Natural Language Processing
Deep Learning Methods

ABSTRACT

Text classification is a supervised machine learning technique that assigns a set of predefined categories or classes to the given text corpora based on the content of the processed text using Natural language processing techniques. Text classification is widely used in numerus applications such as categorizing the sentiment of the tweets and reviews, classification of news and web pages into multiple categories and automatic classification of emails in to spam or not spam. Under the text categories poetry is a literary text and it is special when compared with the regular prose text. A very less focus is given to the task of classification of poetry by the research community. In this context, this work aimed to classify poetry using machine learning and deep learning models and to analyze the performance of the algorithms. To perform this task, poetry corpus is categorized into multiple classes using Latent Dirichlet Allocation a topic modeling technique. The classification task is carried using Multinomial Bayesian, SGD models under machine learning methods and LSTM, Bi-LSTM and CNN models under the deep learning methods. The results are evaluated with parameter accuracy. As a result of this experiment the best classification accuracy is achieved using CNN model with 87% by outperforming other models. This shows that for literary text classification CNN can be considered as a best classifier in comparative with the LSTM and Bi-LSTM models.

Selection and peer-review under responsibility of the scientific committee of the International Conference on Advanced Materials for Innovation and Sustainability.

1. Introduction

Topic modeling is an important technique in the fields of natural language processing, machine learning and Information retrieval. Topic modeling is used to identify the latent semantic structure of documents by using Bayesian Inference[1]. Topic modeling is successfully applied in several applications that incudes topic-based clustering, classification of documents, identifying the trend of topics. Topic modelling assumes documents are formed by the mixture of topics, and topics are formed with probability distribution of words and further topic modeling represents each document by the probability distribution of topics.Fig. 1.

Poetry is a visual art, it comprises different artistic elements such as rhyme, metaphor, simile, alliterations, assonance etc. Analyzing and classifying the poetry is a challenging task. A poem can be analyzed on semantical grounds and stylistic grounds[2]. Computing stylistic features exclusively per poem consumes lot of computational effort when compared with (semantic) word features. Many researchers performed classification of poetry using word features to classify genre, mood and theme of poetry.

In this paper, we have examined the performance of LDA based feature representation in poetry classification task. In this task, Poetry is classified based on authors country(Indian or American). We have considered three data sets first one is with 128 poems; second one is with 1600 and the third one is author wise poems for this we have considered 51 authors.In all data sets we have considered notable Indian and American poets' poetry. We have performed this classification using bag of words method, here words are used basic constructs for classification.

Indian English poetry is nearly 200 years old; it can be seen in 3 phases first phase is upto1900, second phase is 1900–1947 and third phase is after 1947, these 3 phases can be categorized in to

[*] Corresponding author.
E-mail addresses: kpk_it@vignan.ac.in (K Praveen Kumar), drsvpk_cse@vignan.ac.in (V. Phani Kumar S).

https://doi.org/10.1016/j.matpr.2022.05.297
2214-7853/Copyright © 2022 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the International Conference on Advanced Materials for Innovation and Sustainability.

Please cite this article as: K Praveen Kumar, V. Phani Kumar S and S.K. Lokesh Kumar, Analysis of Indian and American poetry using topic modeling and Deep learning, Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2022.05.297

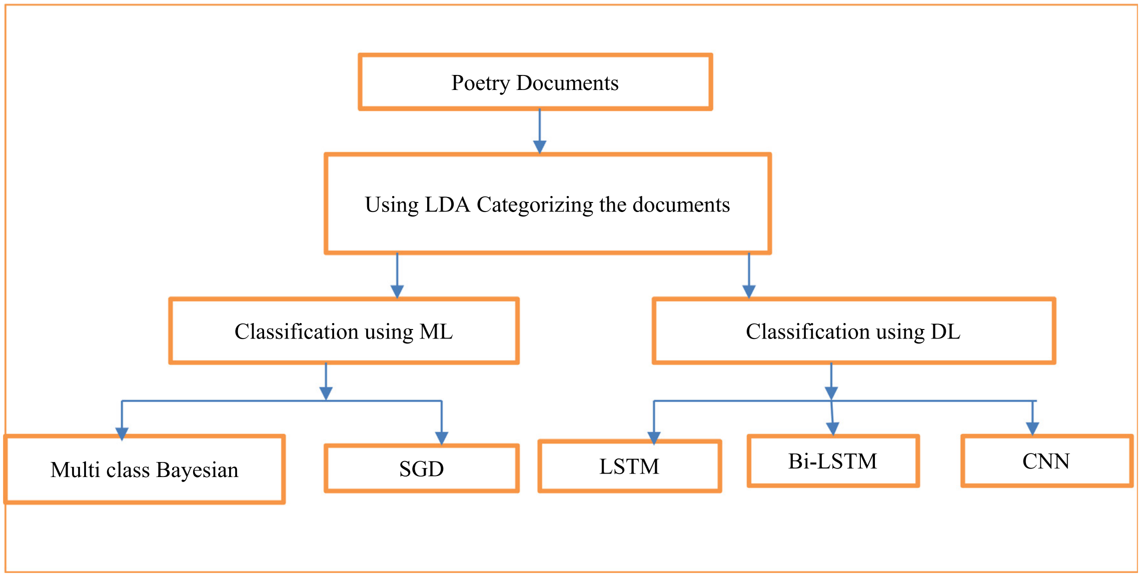


Fig. 1. Flow of the experiment.

pre independence era(up to 1947) and post independent era(after 1947)[3]. Pre independent era poets followed few foreign writers' styles and wrote on Indian history, myths and legends. Post-independence poets have witnessed a dramatic change in themes and writing styles. The post-independence authors focused on love, loneliness, social issues, feminism etc. In this work we explored the classification of Indian poetry from American poetry using topic modeling method. We evaluated the classification using logistic regression, Support Vector Machine (SVM), Random Forest, and K Nearest Neighbor(KNN) classification algorithms.

Rest of the paper is organized as follows section 2 describes about related work, section 3 presents the methodology of LDA topic modeling, section 4 describes about the methodology followed to perform the experiment further section5 presents the results and discussion and at last section 6 concludes the experimental work.

2. Related works

LDA topic modeling can be used as feature representation method in relation with text classification. Yangyang Li et.al.[4] used LDA to classify the short texts. Authors in their work web search snippets and Google news data sets are used. These data sets are classified in to 6 classes, authors used LDA model to generate the features later they have improved the feature to boost the classification. accuracy. Qiuxing chen[5] used LDA topic modeling for short text classification, authors used news data set to classify. AytugOnan et.al. [6]used LDA for sentiment classification, authors used LDA as dimensionality reduction method and reduced the sparse ness of data representation. Authors evaluated their results on 4 sentiment data sets. Ayoub Bagheri et.al. [7]used LDA for clinical sentence classification to detect patient's disease history. The authors used LDA topic modeling to enrich the representation of sentence. Initially they have obtained the TFIDF matrix and by using ETM algorithm improved the topic assignment quality to words. Authors used this model to classify the patients in to two categories 1) patients with medical history and 2) patients without medical history. They have considered 20,200 labeled sentences. Shaymaa et.al[8] used LDA for categorization of eBooks, authors used 300 books for classification purpose they have used 10 topics for topic modelling and 10 classes considered further authors compared LDA work with Latent Semantic Analysis(LSA) and the results shown that LDA performed well when compared to LSA. Xuan et.al.[9] have used LDA for classification of large text corpora, the authors named it as universal data set that is collected from Wikipedia. Authors developed a classifier that deals with short and sparse text and web documents. The classifier is built by discovering the hidden topics from large set of data collection.

This section describes about the task classification reported exclusively on poetry. Jasleen Kaur et.al [10] classified Punjabi poetry using text features. Authors reported the performance of SVM, KNN and Naive Bayes classification algorithms using Term Frequency and Inverse Document Frequency (TF-IDF) features of text. Xuemei et.al.[11] using character, rhyme, genre and overlapped words as features performed author identification on poetry named The Golden Lotus. For this, authors used Decision Tree, SVM, Naive Bayes and KNN classifiers. Noraini Jamal et.al. [12] classified Malaya poetry in to 10 themes using classifiers SVM with radial basis function(RBF) kernel and Linear kernel, for this authors used 1500 Malaya poems called pantun and TF-IDF weights are used as features. Hussein et.al. [13]classified German poetry. For this, the authors used tonality features and classified poetry in to two classes Parlando and Variable foot Poetry using AdaboostM1, IBk, Simple Logistic Regression and Random Tree classification methods. Gharbat et.al. [14]classified Arabic poetry based on the poet's era using SVM classifier. For this work authors used part of poem line as the feature, they have used 10,895 poetic lines as data set.

Abeer et.al.[15,16] performed topic modelling on Arabic poetry using cutting edge technology BERTopic that is (BERT) Bi-directional Encoder Representations from Transformers based Topic modelling which uses pretrained embeddings to find the latent topics among the documents. The authors compared these results with Latent Dirichlet Allocation (LDA) and Non-Negative Factorization (NMF) methods and the results shows that BERTopic model outperformed both the models.

3. Proposed method

Data Acquisition: Poetry is a literary art; we have downloaded the poems from Popular website Poemhunter.com. For this we have collected notable Indian and American poets' names from

Scholarly articles and Wikipedia. We have selected the poets from different time periods so that the data set contain veracity. 1600 poems were collected which includes 760 poems from 28 Indian Poets and 840 American Poems from 12 popular American authors.

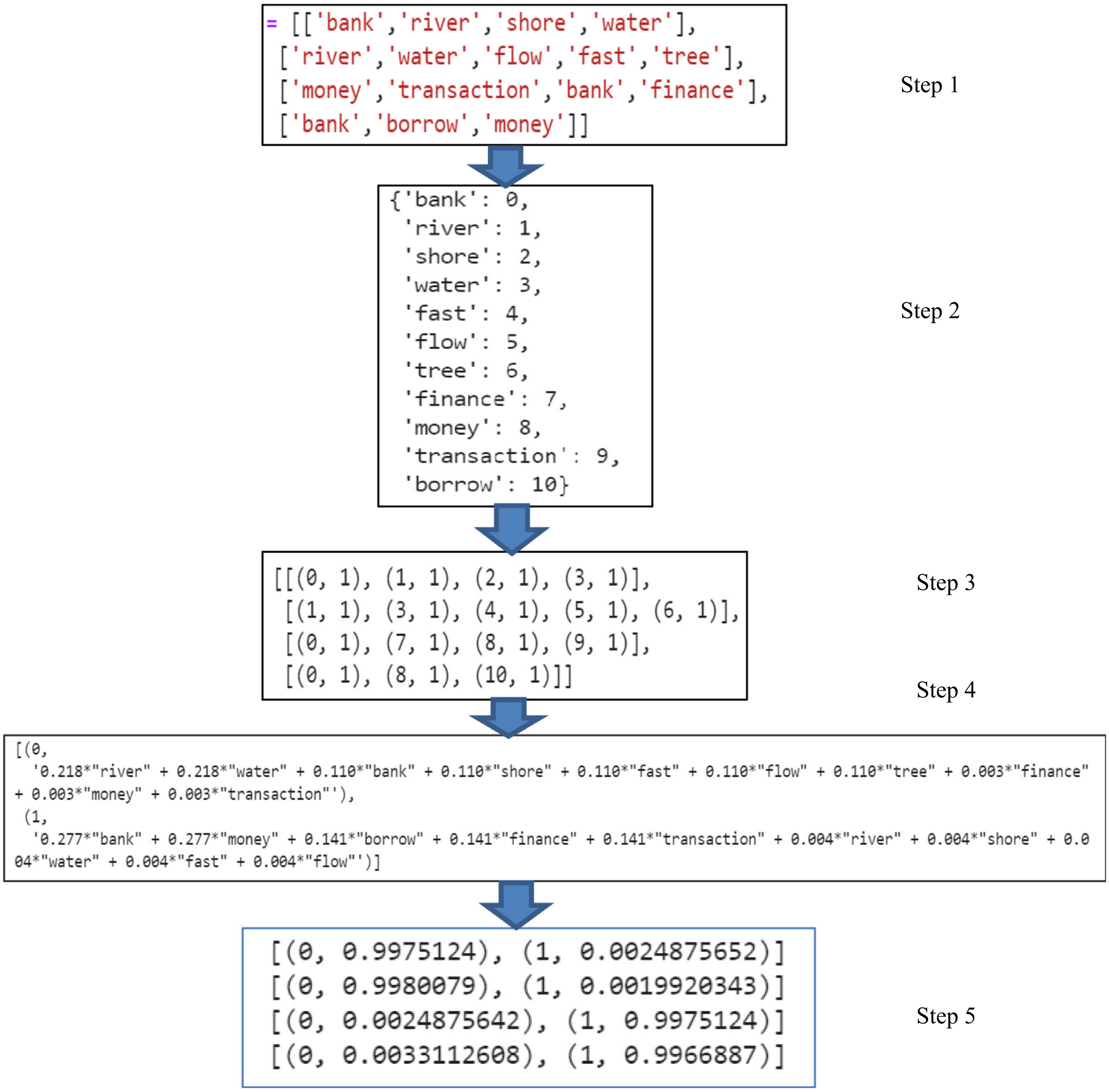
Topic Modelling: topic modelling is a technique to categories documents in to given topics. LDA is a popular topic modelling technique, that is a generative probabilistic model. Where the LDA algorithm assumes that each document is a composition of latent topics and each topic is a distribution of set of words, so by computing the probability of occurrence (distribution) of the words in each topic and the distribution of topics in each document the categorization of documents task can be achieved. To perform topic modelling genism implementation is used. Fig. 2 shows the working example of the LDA technique. In step one we provide the documents in the form of tokens for this step we tokenize the documents, in step 2 a dictionary is prepared in step 3 document to bag of words model is prepared this representation of documents is the input to the LDA model in this example we have considered 2 topics. In step 4 LDA produces topic wise word probability in step 5 document wise topic probability is computed. In Fig. 2 example 4 documents considered that contains 2 documents belongs to one topic and another 2 topics belongs to second topic, here we have shown how LDA assigns topics to these 4 documents.

Step 1 text documents are tokenized and document is represented with set of tokens, step 2 dictionary is created with the unique words, in step 3 documents are represented with dictionary values, in step 4 after completion of LDA algorithm topic wise word distribution and in step 5 topic wise document distribution is provided with the result we can conclude that the document 1 is assigned to topic 0 with a probability 0.9975124 and with probability 0.0024875652 to topic 1. from this we can conclude that document 1 is categorized under topic 0. In similar lines document 2 is categorized under topic 0 and document 3 and 4 are categorized under topic 1.

Classification methods: This section briefly discuss about the classification methods used in this paper.

Multinomial Bayesian Classifier: This classifier is a variant of Bayesian Classifier, here we consider combination variables probability of occurrence.it is used to solve the multiclass classification problems. Bayesian approach.

assumes that a document is a distribution of set of words and it can be generated using specific parametric models and the parameters can be estimated using training data.



```
= [['bank','river','shore','water'],
 ['river','water','flow','fast','tree'],
 ['money','transaction','bank','finance'],
 ['bank','borrow','money']]
```

```
{'bank': 0,
 'river': 1,
 'shore': 2,
 'water': 3,
 'fast': 4,
 'flow': 5,
 'tree': 6,
 'finance': 7,
 'money': 8,
 'transaction': 9,
 'borrow': 10}
```

```
[[(0, 1), (1, 1), (2, 1), (3, 1)],
 [(1, 1), (3, 1), (4, 1), (5, 1), (6, 1)],
 [(0, 1), (7, 1), (8, 1), (9, 1)],
 [(0, 1), (8, 1), (10, 1)]]
```

```
[(0,
  '0.218*"river" + 0.218*"water" + 0.110*"bank" + 0.110*"shore" + 0.110*"fast" + 0.110*"flow" + 0.110*"tree" + 0.003*"finance"
+ 0.003*"money" + 0.003*"transaction"'),
 (1,
  '0.277*"bank" + 0.277*"money" + 0.141*"borrow" + 0.141*"finance" + 0.141*"transaction" + 0.004*"river" + 0.004*"shore" + 0.0
04*"water" + 0.004*"fast" + 0.004*"flow"')]
```

```
[(0, 0.9975124), (1, 0.0024875652)]
[(0, 0.9980079), (1, 0.0019920343)]
[(0, 0.0024875642), (1, 0.9975124)]
[(0, 0.0033112608), (1, 0.9966887)]
```

Fig. 2. Working example of LDA.

$$P(c|d) = \frac{P(c)\Pi_{i=1}^{n} P(w_i|c)^{f_i}}{P(d)} \qquad (1)$$

Equation (1) shows the Multinomial Bayesian Model in this f_i is frequency of word occurrence w_i in document.d

$P(w_i|c)$ is conditional probability that a word wi may occur in a document d with given class c and n is the number of unique words and $P(c)$ is prior probability that a document with class c may occur in document collection. The parameters of the equation can be computed by estimating the relative frequency in data.

SGD classifier: Gradient Descent is a minimization function, when a function is given to gradient descent algorithm it will find the minimum point for that function by taking number of iterations, in this for each record coefficient value update happens where as in Stochastic Gradient Descent the coefficient updation will occur based on the pattern.

To minimize function J(0) the algorithm has to choose 0 value for this initially it will take a random value and the value is changed repeatedly to make J minimal.

LSTM: Long-Short Term Memory network is introduced to overcome the drawback of RNN. RNN prediction doesn't depends on the long sequences due to vanishing gradient problem. LSTM introduced long term memory with the help of new cell gate as shown in Fig. 3. The cell gate comprises of forget gate, input gate and output gate. Each cell information is depends on previous cell state and it is computed using three gates.

Forget gate will decide that weather the old information has to retain or erase, this gate will take the input from previous cell and current input word h_t-1,X_t respectively, the given input is multiplied with weights and a bias is added after that this result is passed through sigmoid function, this sigmoid function produces 0 to forget the data 1 to retain the data.

Input gate adds new information to the cell state by processing the given input word and the output gate is responsible to generate output based on the given input, by this way LSTM retains long sequence information with help of cell state.

Bi-LSTM: Bi-directional LSTM, is very similar architecture to LSTM. In bidirectional LSTM the sequence information is retained in both directions from left to right and right to left as shown in Fig. 4., with this the context understanding and prediction of words at every input becomes more accurate. Bi-LSTM fulfils the need of future information with the right to left direction of LSTM network.

1Dimensional Convolution Network**:** Convolution Neural Network is a Deep Learning Technique. This technique performs a Linear mathematical operation named Convolution, this operation is performed to find the features for a given input.

The CNN architecture includes input layer, a group of hidden layers and output layer, and further the hidden layers consists of several convolution, normalization, pooling and fully connected layers as shown in Fig. 5..

In CNN, the first hidden layer will be a convolution layer it is used to extract the features from input data. Next level layers are pooling layers, these layers are useful to reduce and preserve the features of the input, pooling will help to preserve the features with respect to spatial invariance. At last a fully connected layer will be there to with N dimensions where N is the number of classes to be classified. This layer is given to an activation function either Sigmoid for binary classification or SoftMax for multiclass classification.

In CNN the learning will happen with several iterations, at each iteration loss value is minimized and the accuracy of detecting the right class will increase.

Evaluation parameters: In order to evaluate the classifier results the following parameters are considered. Using these parameters, the performance of classifiers is compared.

Accuracy: this measure is calculated based on the correctly classified entities. It is the ratio between correctly classified entities to the total number of entities in dataset.

$$Accuracy = \frac{CorrectPredictions}{Totalentitiescount}$$

Precision: It is the ratio of True Positive to True and False Positive, it will tell in total positive predictions what percentage is Correctly classified as positive.

$$Precision = \frac{TotalTruePositive}{TruePositives + FalsePositives}$$

Recall: This measure tells about the correct predictions of a single class; it is the ratio of correctly identified entities of one class to total entities of that class.

$$Recall = \frac{TotalTruePositives}{TruePositives + FalseNegatives}$$

F1 score: It gives the overall performance of the model.

$$F1Score = 2 * \frac{precision * recall}{precision + recall}$$

4. Experiment and Results:

Data Pre-processing and Representation: Initially loaded the data set poetry foundation data set consists of 13,747 poems from 3128 poets, taken from Kaggle data repository. Data pre-processing performed in that special symbols, numbers are removed using regular expression package after that removed empty lines. To generate automatic class labels performed Latent Dirichlet Allocation (LDA) topic modelling, for this Sci-kit learn implementation used. Using topic coherence technique 3 topics are considered. To perform topic modelling count vectorizer method used to convert text to count vectors. Further each document is assigned a topic as class label.

After assigning the labels to the documents *tf-idf* text representation method is used as features of documents. This data is given

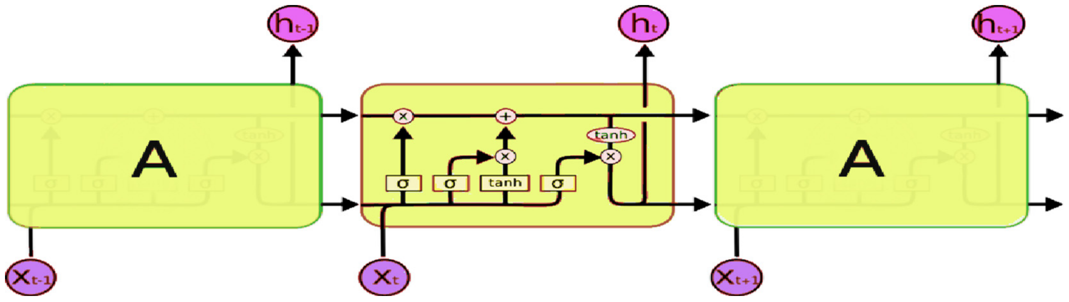


Fig. 3. Cell structure of LSTM Network.

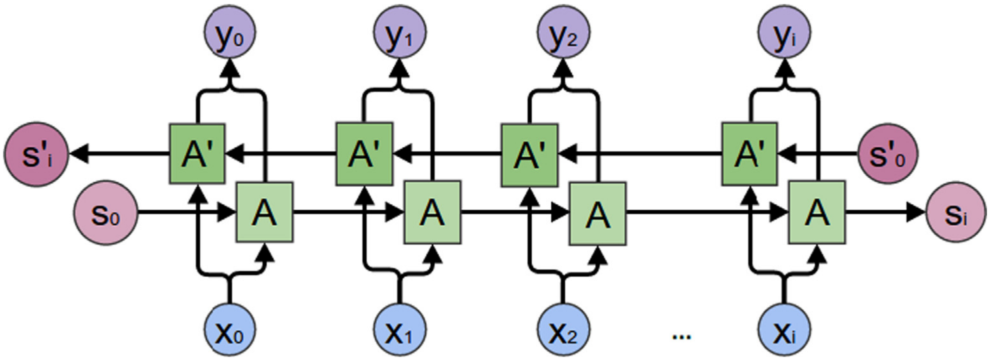


Fig. 4. Bi-LSTM network architecture.

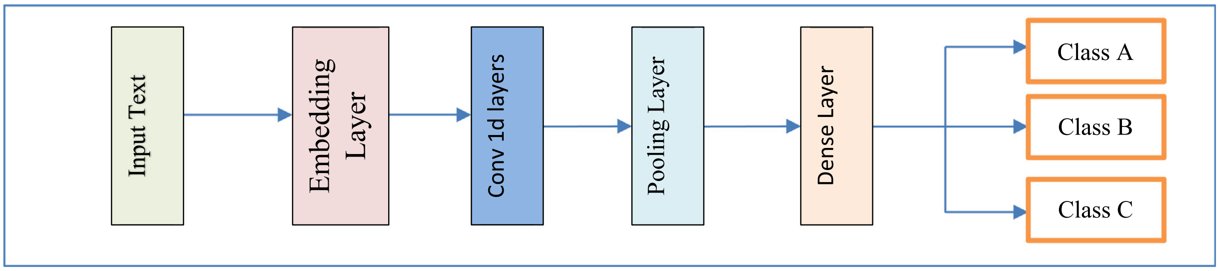


Fig. 5. Conv1d architecture.

as input to the Machine learning algorithms Multinomial Bayesian classifier and Stochastic Grading Descent algorithms. Multinomial classifier has classified the documents with 83% accuracy and Stochastic Gradient Descent classifier classified the documents with 86.5% accuracy.

The same dataset is used with deep learning techniques Bidirectional LSTM and 1 dimensional Convolution Neural Network. In Bi-LSTM maximum number of words 50,000, max sequence length is 200 and word embedding dimension considered as 150. To find the embeddings used tokenizer to find the numerical sequences, using *Pad sequences* class made all the documents lengths are equal for this padding is performed, further text is divided in to 70% training data and 30% as testing data. Bi-LSTM model is built with 64 bi-directional LSTM cells and 24 dense layers are used as hidden layers. In hidden layer Rectified Linear Unit (ReLU)is used as activation function and *rmsprop* is used as optimizer. To select this optimizer, we have tested the model with other optimizers such as *adam*, *nadam* and the best performance is achieved with RMS*prop* optimizer.

The result table Table 1 shows that the accuracy of Multinomial Naïve Bayes is 83% with this algorithm we can see that the f1 score for class 2 is very poor. This model is unable to classify the category 2 records effectively when compared with other classification methods. Stochastic Gradient Descent algorithm has shown the performance very close to the Deep learning algorithms but the F1 score is low when compared with the Deep learning methods. We can observe from the table that with subtle variance Bi-LSTM and 1Dimensional convolution methods shown their performance but the average precision values for Bi-LSTM and Conv1d are 79.3 and 81.3 respectively and the F1 score average values are almost same. With this result we can conclude that 1dimensional convolution method is performing well in classifying the poetry into multiple classes when compared with machine learning algorithms. The result further can be improved by checking various parameters of 1dimensional Convolution model.

5. Conclusion and future work:

To categorize Indian and America poetry into multiple classes we have used Machine learning and Deep learning methods. Further we have compared the results of the both methods. In this experiment identified that Conv1d deep learning method works

Table 1
Experimental results of Machine Learning and Deep Learning methods.

Method	Class label	Accuracy	Precision	Recall	F1
MNB	0	83	81	100	89
	1		96	50	10
	2		97	69	81
SGD	0	86.5	87	97	92
	1		76	29	42
	2		88	83	85
Bi-LSTM	0	86	92	89	91
	1		55	73	63
	2		91	81	86
Conv1D	0	87	90	94	92
	1		67	54	60
	2		87	85	86

not only with images but also with the complex text. In this experiment the best accuracy 87% is achieved with Conv1d deep learning method. In this experiment class labels are identified using Latent Dirichlet Allocation method later the data is given to the classification algorithms to classify the data according to the given classes.

The limitations of this work are number of poems considered for data set and to embed the text need to use a better method which will embed the stylistic features of poem that include phonemic features, syntactic features. To categories the data set LDA model is used and not tested with other advanced methods, In future to categories the data advanced methods can be used. The results are evaluated on single data set this can be tested with other data sets also in future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, ''Latent Dirichlet Allocation'' 3 (2003) 993–1022.

[2] D.M. Kaplan, D.M. Blei, A computational approach to style in American poetry, Proc. - IEEE Int. Conf. Data Mining, ICDM, 2007, pp. 553–558.

[3] dr. S. Jha, ''Status of Indian English Poetry after Independence,'' *Eur. Acad. Res.*, vol. II, no. 11, pp. 14446–14453, 2015.

[4] Y. Li, B. Liu, ''A New Vector Representation of Short Texts for, Classification'' 17 (2) (2020) 241–249.

[5] Q. Chen, L. Yao, and J. Yang, ''Short text classification based on LDA topic model,'' pp. 749–753, 2016.

[6] H.B. Aytug Onan, S. Korukoglu, ''LDA-based Topic Modelling in Text Sentiment Classification, An Empirical Analysis'' 7 (1) (2016) 101–119.

[7] F. W. Asselbergs and D. L. Oberski, ''ETM : Enrichment by topic modeling for automated clinical sentence classification to detect patients ' disease history,'' 2020.

[8] S. H. M. Al-augby, ''LSA & LDA Topic Modeling Classification : Comparison study on E-books LSA & LDA Topic Modeling Classification : Comparison study on E-books,'' *Indones. J. Electr. Eng. Comput. Sci.*, no. January, 2020.

[9] X. Phan, ''Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections,'' 2008.

[10] J. Kaur and J. R. Saini, ''Designing Punjabi Poetry Classifiers Using Machine Learning and Different Textual Features,'' no. December, 2019.

[11] X. Tang, S. Liang, Z. Liu, ''Authorship attribution of the golden lotus based on text classification methods'', *ACM Int, Conf. Proceeding Ser.*, vol. Part F1481 (2019) 69–72.

[12] M. M. and S. A. Noraini Jamal, ''Poetry Classification Using Support Vector Machines,'' vol. 8, no. 9, pp. 1441–1446, 2012.

[13] H. Hussein, B. Meyer-Sickendiek, and T. Baumann, ''Tonality in Language: The Generative Theory of Tonal Music as a Framework for Prosodic Analysis of Poetry,'' pp. 178–182, 2018.

[14] M. Gharbat, H. Saadeh, and R. Q. Al Fayez, ''Discovering the applicability of classification algorithms with Arabic poetry,'' *2019 IEEE Jordan Int. Jt. Conf. Electr. Eng. Inf. Technol. JEEIT 2019 - Proc.*, pp. 453–458, 2019.

[15] A. Abuzayed, H. Al-Khalifa, BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique, Procedia CIRP 189 (2021) 191–194.

[16] M. Ramalingam, D. Saranya, R. ShankarRam, P. Chinnasamy, K. Ramprathap, A. Kalaiarasi, An Automated Framework For Dynamic Web Information Retrieval Using Deep Learning, International Conference on Computer Communication and Informatics (ICCCI) 2022 (2022) 1–6, https://doi.org/10.1109/ICCCI54379.2022.9741044.