I.J. Information Engineering and Electronic Business, 2021, 1, 15-24
Published Online February 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2021.01.02

Bengali News Headline Categorization Using Optimized Machine Learning Pipeline

Prashengit Dhar
Department of Computer Science and Engineering, Cox's Bazar City College, Bangladesh
Email: nixon.dhar@gmail.com

Md. Zainal Abedin
Faculty of Science Engineering and Technology, University of Science & Technology Chittagong- 1079, Bangladesh
Email: jakcse99@gmail.com

Received: 15 October 2020; Revised: 10 November 2020; Accepted: 25 December 2020; Published: 08 February 2021

Abstract: Bengali text based news portal is now very common and increasing day by day. With easy access of internet technology, reading news through online is now a regular task. Different types of news are represented in the news portal. The system presented in this paper categorizes the news headline of news portal or sites. Prediction is made by machine learning algorithm. Large number of collected data are trained and tested. As pre-processing tasks such as tokenization, digit removal, removing punctuation marks, symbols, and deletion of stop words are processed. A set of stop words is also created manually. Strong stop words leads to better performance. Stop words deletion plays a lead role in feature selection. For optimization, genetic algorithm is used which results in reduced feature size. A comparison is also explored without optimization process. Dataset is established by collecting news headline from various Bengali news portal and sites. Resultant output shows well performance in categorization**.**

Index Terms: Bengali text; tokenization; stop word deletion; genetic algorithm; categorization

1. Introduction

Use of internet has increased rapidly. We cannot think of a day without internet as well as social media. With the increasing number of users, a large portion is online newsreader. As news portals offer low data cost for their sites, readers also becomes much interested in online news. There was a time when people wait for newspaper hawker to get and read newspaper. The time has gone. Availability of smartphone, computer and internet technology, make it easier to get or read any news with a click within a second. With the advancement of Unicode module and Internet, use of bengali texts in digital domain have become standard and also increased. A large number of people spoke Bangla around the world and ranked at 7th among most spoken language. Most of the official govt. sites of Bengladesh use bangla now. It is in such a way that any user can identify related documents easily without any difficulty in understanding the site. Automatic Text categorization system can help in retrieving required document successfully within short time.

Web technologies related to internet has made a great change in our society. Categorization of text is a research in the area of text mining. Automatic categorization of text helps in saving much time, which is costly in manual system. Previously few researches have been done in Bengali text mining. In this research genetic algorithm for feature reduction is used in bengali text categorization. Sentiment analysis is one of the known researches in text mining or NLP.

2. Literature Review

Iqbal used genetic algorithm as feature selection in sentiment analysis [1]. They combined lexicon and ML based approaches to achieve higher accuracy. It was on twitter and review dataset. Machine learning usually focuses on global classification than specific aspects. Generally three parts are included in analyzing sentiment such as Machine Learning, lexicon-based and rule grounded approaches [2]. Sentiment analysis is basically opinion mining as like as characterization of sentiment.by collecting data from text and using NLP, machine learning etc [3]. Research on analyzing sentiment is going on from last few years. The rule- based method looks for opinion related words from the text and later categorizes it depending on the amount of positive words and negative words [4]. It calculates different rules for classification such as boosted words, dictionary polarity, idioms, negation words etc. The lexicon-based tactic

includes measuring the sentiment polarity of text which is done by calculating the semantic orientation of sentences or words [5]. The semantic orientation is a measure of subjectivity and opinion in text. Khan proposed a rule dependent domain-independent approach [6]. Its task is to classify subjective sentences and objective sentences from comments of reviews and blog. Score and polarity were measured using SentiWordNet.

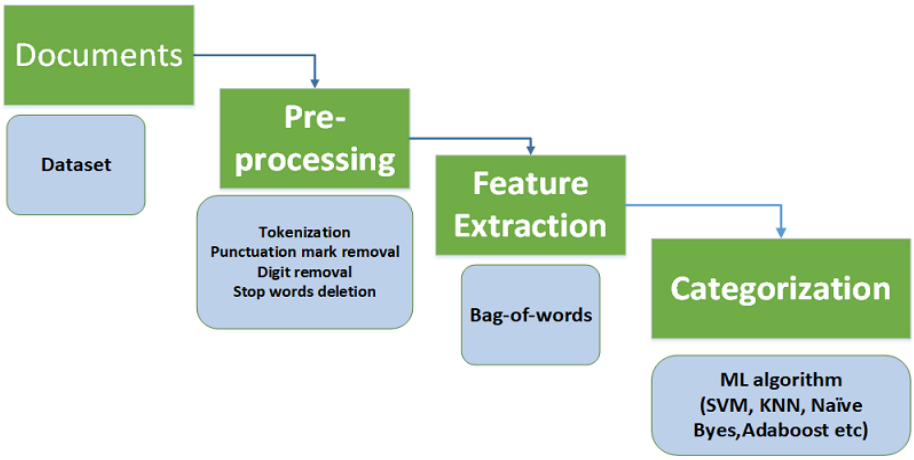

Fig.1. General approach of Bengali text classification

Along with English language, many researches on other languages like- Spanish [8], Arabic [7], Chinese [9] is occurred and still going on. Ahmad explores arabic text classification based on semantic ontology [7]. Term frequency-inverse document frequency (TF-IDF) is performed on arabic text documents. An unsupervised learning on Chinese text is applied by Zagibalov [9]. The task was to analysis sentiment. In different text data classification work, Naïve Bayes as ML algorithm is used frequently [10,12,13]. Classification of web sites from content home page is done by using Naïve bayes [10]. Suguna used improved KNN with genetic algorithm for classifying text [11]. Anurag and Debabrata presented a model to classify text based on frequencies of text parameter [18]. Wf-icf (based upon the well-known tf-idf property of information retrieval) property based parameter s used in that model.Vinay and Shishir proposed a method for election outcome prediction by collecting data from social media [19]. Topic modelling and dynamic keywords are represented as main features to predict election outcome.Shoumick and Pramod proposed awork that classifies text SVM[20]. A multithreading approach was presented to speed up the pre-processing task.

In this study, the main purpose is to categorize news headline. It is better to classify a news based on headline rather than whole article. Classification based on whole article requires large feature size also as well as time consuming also. This study shows news headline classification with reduced feature based on GA (Genetic Algorithm) and optimized ML pipeline.

3. System Overview

The methodology of the proposed system is described in this section. The system is mainly comprises of five modules. Fig.2 shows step based depiction. The first one is data acquisition. It is a process of collecting labeled news for sentiment analysis. Then the second module undergoes several steps of pre-processing. It is a process of transforming data into a suitable form that can simply used for subsequent analysis. The third module is the extraction of features for creating a classification model. Fourth module is optimization of features based on genetic algorithm approach. To reduce large feature size, optimization of features maintaining accuracy is a standard way. The goal of using GA is to reduce feature size.

At this phase population is a term that is a set of candidate solution which progresses towards a superior solution over the certain generations with the purpose of solving a problem. Few genetic operators like - mutation, crossover are employed to a population. Evolution of fitness is a core step. As various candidate solution is available, then which solution is going to be exist in next generation is decided by fitness value returned by the fitness function. The last module is train and test using machine learning.

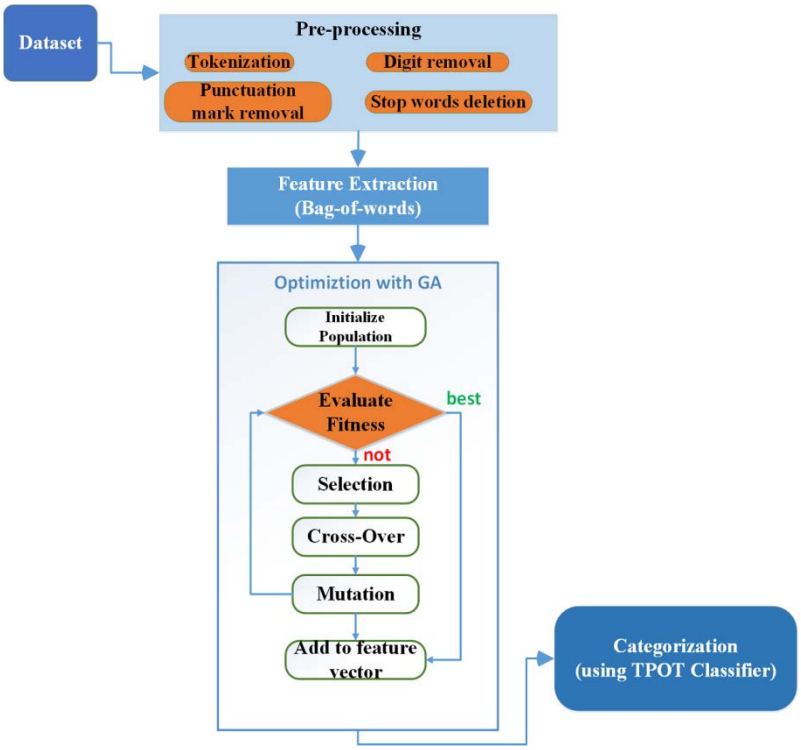

Fig.2. Proposed text categorization framework

4. Proposed Methodology

4.1. Dataset

A crucial part of any classification system is dataset. There many datasets are available online in different language. Most commonly found dataset is in English language. However, in case of Bengali language related dataset(corpus) is very rare. So custom dataset is needed to create for this research purpose. Text from various online news portal is collected. Self created text data is also added in the dataset. Data from popular news portal is mostly collected. Data are collected in three categories such as- health, sports, and technology. Size of the dataset is 1000x3 .Table 1 shows statistics of dataset.

Table 1. Proposed Dataset statistics

Category	Number of documents
Health	158
Sports	158
Technology	158

4.2. Pre-processing

Proposed system uses bag-of-words module .So suitable and appropriate form of words is necessary and also can perform good in classification stage. Pre-processing is mandatory to get associated features. Correct features of relevant information are needed to extract from the input text data, it may result in high dimensional feature size. Optimization algorithm can reduce the generated feature matrix. However, pre-processing is essential before features extraction from the documents. In pre-processing phase, text document is represented as “Bag of words”. The following operations are done on each document as pre- processing.

1. Tokenization

It refers to breaking the sentence or text documents into individual word, symbol phrases or other elements which is delimited by white space, new line or tab. Tokenization output are set of words. For example: "১০ নভেম্বর থেকে বিশ্বকাপ ক্রিকেট শুরু !", Tokenization result for the line will be- "১০" "নভেম্বর" "থেকে" "বিশ্বকাপ" " ক্রিকেট" "শুরু" "!"

2. Removing digit or numbers

Text documents may contain digits or number. It may be in Bengali or English. Digits are unwanted element in this research. A text documents or sentence any contain digits or numbers, but is unnecessary to keep those digits. Meaningful word are of preferences in the this research work. At this phase, digits are removed from text. It results like- "নভেম্বর" "থেকে" "বিশ্বকাপ" " ক্রিকেট" "শুরু" "!"

3. Removal of punctuation marks

For text classification, removal of punctuation mark is mandatory, Punctuation marks are most unnecessary in text classification task. Besides punctuation marks, special symbols like - {,},[,],(,),<,>, :, ,^,&,*,! etc. are also removed. An outcomes like- "নভেম্বর" "থেকে" "বিশ্বকাপ" " ক্রিকেট" "শুরু"

4. Stop words removal

Words that contain no relevant information are known as stop words. Examples are: আমি,তুমি,এবং etc. Stop words set is more important specially for bengali text classification. A strong stop word set leads relatively low dimensional feature size. Custom stop words set are created for bengali text documents. Day, months, country names, probable common name are also included. In addition, single letter word is also removed in this stage. So finally we got only these three words. **"বিশ্বকাপ" " ক্রিকেট" "শুরু".**

5. Feature Extraction

In this stage, features are extracted for feeding into a classification model. Extracted features are stored in a suitable format so as to directly applicable to classifier. After pre-processing, remaining words are used to represent document. The remaining words are known as term. Few techniques are available to extract features such as - count vectorizer and term frequency-inverse document frequency (TF–IDF). In this research, TF-IDF is employed for extracting the required feature matrix. It imitates the significance of terms to the corpus data in a text. CountVectorizer counts the frequency of word in a document. TF-IDF results in score that represents the relative significance of a term in the document and the entire corpus. TF refers to Term Frequency and Inverse Document Frequency for IDF.

$$TF(t,d) = TF(t,d) * log\frac{N}{DF(t)} \quad (1)$$

For equation (1), t= term, d= documents, TF(t)= term frequency, N= number of documents in the corpus, DF(t)= number of documents in the corpus containing the term t, N= number of documents in the corpus.

Proposed system is simulated on both CounterVectorizer and TFIDF feature extraction technique. The corpus contains 3903 after removal of punctuation marks, digits and stop words, the word size is 3506. In addition, features are also optimized. Genetic algorithm is used as feature optimizer through Tree-Based Pipeline Optimization Tool (TPOT) and also employed TPOT as an AutoML method. This research proposes bengali text classification using optimized feature.

6. Classification Algorithms

After extracting feature matrix, it is then send to the classifier. The classifier trains and tests the feature matrix. There are several types of classification model. In this study following classification model is considered. These are described in this section.

6.1. Support Vector Machine

Support Vector Machine is one of the popular and simplest linear classifier. SVM is a supervised and binary classifier. It works by finding a hyper plane to separate classes. The hyper plane is chosen depending on the highest maximum and margin among various classes [14]. The principle of SVM algorithm is the risk minimization. SVM also supports outlier and regression. Fig 3 shows how SVM works.

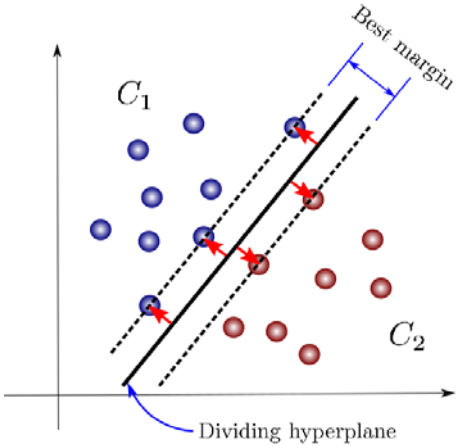

Fig.3. SVM principle visualization

The equation to predict for an input with the dot (.) product between the input as x and every support vector as xi is computed as given in (2).

$$f(x) = B0 + sum(ai * (x, xi)) \quad (2)$$

This is an equation that comprises calculating the inner products of a new input (x) with entire support vectors in training data. The coefficients B0 and ai of each input should be evaluated from training data by the ML algorithm.

6.2. Naive Byes

Naïve Byes is probabilistic classification model which formulates renowned Bayes theorem. It is used frequently used in different classification model. Naïve bayes generally provides good results. It's main idea is from bayes theorem and also uses Bag of word technique in feature extraction. As it measures the prior probability and posterior probability of a category or class. That means the distribution words in a document is observed[15]. Due to this it has an widespread use. This kind of classifier is suitable for classifying text data.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

Bayes theorem(equation (3)) finds the probability of incident A, given that B has occurred. B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent.

KNN is another finest classification algorithm. It uses statistical approach for classifying data. For a test sample, the algorithm looks for k nearest neighbor of the given test sample from all the training document. Then scores the class candidates depending on the category of k neighbor. If numerous of the KNN documents supports same category, then the total score of that category is the matching score of the category in respect to the test sample. It then sorts the scores of the candidate categories and system allocates the candidate category with the maximum score to the test sample. Generally Euclidean distance is formulated to calculate the nearest neighbor. For two points (x, y) and (a, b). The Euclidean distance (d) formula will be is in eqn(4)

$$d = sqrt((x-a)^2+(y-b)^2) \quad (4)$$

The smallest Euclidean distance is tried out to consider and on the basis of smaller distances, calculations are performed.

6.3. Adaboost

AdaBoost is short for Aadaptive Boosting. . It is utilized in conjunction with several alternative forms of learning algorithms to enhance performance. The output of the opposed learning algorithms ('weak learners') is joined into a weighted add that denotes the ultimate output of the boosted classifier. AdaBoost is sensitive to clattering information and outliers. In some issues it is less at risk of the overfitting downside than alternative learning algorithms. The individual learners is weak, however as long because the performance of every one is slightly higher than random shot, the ultimate model is tried to converge to a tough learner.

6.4. Tree-Based Pipeline Optimization Tool (TPOT)

TPOT automates the difficult process of building an auto ML pipeline by demonstrating pipelines as binary expression trees with ML operators as primitives [17]. Pipeline elements include algorithms from the extensive library of scikit-learn [16] besides other effective implementations like extreme gradient boosting. With the help of genetic programming, TPOT optimizes machine learning pipelines. TPOT systematizes the most critical part of a machine learning algorithm. TPOT explores numerous number of possible pipelines intelligently so as to grab the best output for data.

Automated ML (AutoML) is another field with the objective of making it simple to choose AI calculations, their boundary settings, and the pre-preparing strategies that improve their capacity to identify complex examples in enormous information. The Tree Based Pipeline Optimization (TPOT) strategy that utilizes double articulation trees to speak to ML pipelines with improvement gave by hereditary programming and other stochastic hunt strategies. Each AI framework has hyper boundaries, and the most fundamental assignment in mechanized AI (AutoML) is to consequently set these hyper boundaries to advance execution. Computerized diminish the human exertion essential for applying AI. This is especially significant with regards to AutoML. Hyper boundary streamlining (HPO) has a few significant use cases; it can improve the presentation of AI calculations improve the reproducibility and decency of logical examinations

TPOT is a python-based AutoML library consists of genetic programming algorithm to find out the best performing ML pipelines, which are built on top of scikit-learn. Automation of TPOT consists of feature selection, model selection and parameter optimization mainly except “wrangling” part(Data Cleaning/Data Clensing) to do transformation of data, a dimension reduction or a scale change and model validation uses genetic programming to optimize a ML pipeline that maximizes the score on the provided features and target. This pipeline optimization technique uses internal k-fold cross-validaton to evade overfitting on the provided data. At final state of the pipeline optimization process, the best pipeline is then trained on the entire set of provided samples. A pipeline is mainly a succession of AI measures (typically different preprocessors, trailed by a characterization or a regression method) that are precise applied to a bit of information, to deliver a last mode. The methodology utilized in this paper look through the space of pipelines that can tackle a given ML issue utilizing GP.

Currently, the grammar used in the design of TPOT only considers two types of primitives; data pre-processing procedures (feature transformation and selection), and classifiers.

7. Result and Analysis

For analyzing the performance quality of a learner, an easy and common way is understanding the confusion matrix. Column of confusion matrix represents predict class and row represents the actual class. Different evaluation metrics are used in text categorization. In this study, popular and common ways are considered for measuring performance. Precision, recall and F1- score are used here. Precision computes the quantity of positive class predictions that truly belong to the positive class. Equation (5) shows the formula of precision.

$$\text{Precision, } P = TP/(TP+ FP) \tag{5}$$

Recall measures the amount of positive class that predicts beyond all positive sample in the dataset. Equation shown in eqn (6).

$$\text{Recall } R = TP/(TP + FN) \tag{6}$$

F1 (F measure) is the combination of recall and precision. F1 calculates overall performance accuracy of a model. Equation (7) shows the formula.

$$F = 2 * ((\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})) \tag{7}$$

Accuracy can also be measure in following way of equation (8).

$$\text{Accuracy} = (TP+TN)/(TP+FP+FN+TN) \tag{8}$$

After extracting features, it is then trained and tested with different classifier. Classification result with different classifier is depicted in following table 2 to 7.

Table 2. Classification accuracy statistics

Classifier	**TFIDF (%)**	**CountVectorizer(%)**
Logistic Regresson	69.47	74.73
KNN	69.47	67.36
Aaboost	62.105	66.31
SVM	70.52	73.68
Naïve Byes	**73.68**	**75.78**

Table 3. Result of Logistic Regression using TFIDF

	precision (%)	recall (%)	f1-score(%)
health	89	48	63
Sports	71	88	78
Technology	58	72	65
macro avg.			69

Table 4. Result of KNN using TFIDF

	precision (%)	recall(%)	f1-score(%)
health	89	48	63
Sports	71	88	78
Technology	58	72	65
macro avg.			69

Table 5. Result of Adaboost using TFIDF

	precision(%)	recall (%)	f1-score(%)
health	76	39	52
Sports	91	61	73
Technology	46	90	61
macro avg.			62

Table 6. Result of SVM using TFIDF

	precision(%)	recall (%)	f1-score(%)
health	59	70	64
Sports	92	70	79
Technology	68	72	70
macro avg.			71

Table 7. Result of Naïve Byes using TFIDF

	precision(%)	recall (%)	f1-score(%)
health	77	61	68
Sports	93	76	83
Technology	60	86	70
macro avg.			74

From table 2 to 7, it is clear that Naïve Byes performs better than others while using TFIDF. On the basis of macro average of F1 score, NB achieved average accuracy of 74% while adaboost is 62%, SVM is 71%, KNN and Logistic regression are achieved 69%.

A simulation using only countvectorizer is also done. In this regard, classification result using only countvectorizer is shown in table 8 to 12.

Table 8. Result of Logistic Regression using countvectorizer

	precision(%)	recall (%)	f1-score(%)
health	89	48	63
Sports	71	88	78
Technology	58	72	65
macro avg.			69

Table 9. Result of KNN using countvectorizer

	precision(%)	recall (%)	f1-score(%)
health	89	48	63
Sports	71	88	78
Technology	58	72	65
macro avg.			69

Table 10. Result of Adaboost using countvectorizer

	precision(%)	recall (%)	f1-score(%)
health	76	39	52
Sports	91	61	73
Technology	46	90	61
macro avg.			62

Table 11. Result of SVM using countvectorizer

	precision (%)	recall (%)	f1-score(%)
health	59	70	64
Sports	92	70	79
Technology	68	72	70
macro avg.			71

Table 12. Result of Naïve Byes using countvectorizer

	precision (%)	recall (%)	f1-score(%)
health	77	61	68
Sports	93	76	83
Technology	60	86	70
macro avg.			74

Classification result using countervectorizer as feature extraction method, Naïve Byes works well and provides 74% macro avg. on f1 score.

A feature optimization method is applied using TPOT reduces feature size and classifies with better performance in this study. Table 13 and 14 shows classification result and confusion matrix of TPOT returns using optimized features from countervectorizer. TPOT returns Logistic regression as best fit ML while considering countervectorizer. Collecting feature matrix from TFIDF by TPOT, it returns Naïve Byes as best fit ML. Outcomes of classification and confusion matrix using TFIDF are depicted in table 15 and table 16. Both TFIDF and countervectorizer results in 81% accuracy rate on macro average of f1 score returned by TPOT.

Table 13. Result of TPOT returns (Logistic Regression) using countervectorizer

	precision(%)	recall(%)	f1-score(%)
health	82	79	81
Sports	84	82	83
Technology	76	81	79
macro avg.			**81**

Table 14. Confusion Matrix of TPOT returns (Logistic Regression) using countervectorizer

	health	Sports	Technology
health	**23**	**3**	**3**
Sports	**1**	**27**	**5**
Technology	**4**	**2**	**26**

Table 15. Result of TPOT returns (Naïve Byes) using TFIDF

	precision (%)	recall (%)	f1-score(%)
health	67	83	74
Sports	96	76	85
Technology	84	84	81
macro avg.			**81**

Table 16. Confusion Matrix of TPOT returns (Naïve Byes) using TFIDF

	health	Sports	Technology
health	**24**	**1**	**4**
Sports	**7**	**25**	**1**
Technology	**5**	**0**	**27**

8. Conclusion and Future Works

A very efficient system for Bengali text classification is a challenging task. Bengali language has numerous number of diversity and it is rich. Test is applied on both counter-vectorizer and TF-IDF method. The dataset is created in a way that covers most of the words that are relevant with the classes. The paper proposes Bengali news headline categorization with optimized ML. An average accuracy of 81% is achieved in this study. This paper made a clear comparison of performance between regular ML and optimized ML pipeline.

References

[1] Farkhund Iqbal, Jahanzeb Maqbool, Benjamin C M Fung, Rabia Batool, Asad Masood Khattak, Saiqa Aleem, and Patrick C K Hung, "A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction," in IEEE Access, vol 7, pp 14637-14652, 2019, doi: 1 1109/ACCESS 2019 2892852

[2] A Collomb, C Costea, D Joyeux, O Hasan, and L Brunie, "A study and comparison of sentiment analysis methods for reputation evaluation," Rapport de recherche RR-LIRIS-2014-002, 2014

[3] B Pang and L Lee, "Opinion mining and sentiment analysis," Foundations and trends in information retrieval, vol 2, no 1-2, pp 1–135, 2008

[4] X Ding, B Liu, and P S Yu, "A holistic lexicon-based approach to opinion mining," in Proceedings of the 2008 International Conference on Web Search and Data Mining ACM, 2008, pp 231–240

[5] M Taboada, J Brooke, M Tofiloski, K Voll, and M Stede, "Lexiconbased methods for sentiment analysis," Computational linguistics, vol 37, no 2, pp 267–307, 2011 Chen, S Y, & Hsieh, J W Boosted road sign detection and recognition In Proc of Intl Conference on Machine Learning and Cybernetics, 2008 pp 3823–3826

[6] A Khan, B Baharudin, and K Khairullah, "Sentiment classification using sentence-level lexical based semantic orientation of online reviews," Trends in Applied Sciences Research, vol 6, no 10, pp 1141–1157, 2011

[7] Hawalah, Ahmad 2019 "Semantic Ontology-Based Approach to Enhance Arabic Text Classification " Big Data Cogn Comput 3, no 4: 53

[8] F Ciravegna, L Gilardoni, A Lavelli, S Mazza, W J Black, M Ferraro, et al , "Flexible text classification for financial applications: the FACILE system," in ECAI, 2000, pp 696-700
[9] T Zagibalov and J Carroll, "Automatic seed word selection for unsupervised sentiment classification of Chinese text," in Proceedings of the 22nd International Conference on Computational LinguisticsVolume 1, 2008, pp 1073-108
[10] A S Patil and B Pawar, "Automated classification of web sites using Naive Bayesian algorithm," in Proceedings of the International MultiConference of Engineers and Computer Scientists, 2012, pp 14-16
[11] N Suguna and K Thanushkodi, "An improved K-nearest neighbor classification using Genetic Algorithm," International Journal of Computer Science Issues, vol 7, pp 18-21, 2010
[12] L Jiang, Z Cai, H Zhang, and D Wang, "Naive Bayes text classifiers: a locally weighted learning approach," Journal of Experimental & Theoretical Artificial Intelligence, vol 25, pp 273-286, 2013
[13] Q Yuan, G Cong, and N M Thalmann, "Enhancing naive bayes with various smoothing methods for short text classification," in Proceedings of the 21st international conference companion on World Wide Web, 2012, pp 645-646
[14] C Cortes and V Vapnik, "Support-vector networks," Machine learning, vol 20, pp 273-297, 1995
[15] Anuja P Jain and Padma Dandannavar 2016 "Application of machine learning techniques to sentiment analysis", International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 628- 632
[16] Pedregosa,F (2011) Scikit-learn: machine Learning in Python J Mach Learn Res , 12, 2825–283
[17] Olson R S , Moore J H (2019) TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning In: Hutter F , Kotthoff L , Vanschoren J (eds) Automated Machine Learning The Springer Series on Challenges in Machine Learning Springer, Cham https://doi org/1 1007/978-3-030-05318-5_8
[18] Anurag Sarkar, Debabrata Datta,"A Frequency Based Approach to Multi-Class Text Classification", International Journal of Information Technology and Computer Science, Vol.9, No.5, pp.15-22, 2017.
[19] Vinay K. Jain, Shishir Kumar, "Towards Prediction of Election Outcomes Using Social Media", International Journal of Intelligent Systems and Applications, Vol.9, No.12, pp.20-28, 2017.
[20] Soumick Chatterjee, Pramod George Jose, Debabrata Datta, "Text Classification Using SVM Enhanced by Multithreading and CUDA", International Journal of Modern Education and Computer Science, Vol.11, No.1, pp. 11-23, 2019.

Authors' Profiles

Prashengit Dhar received his B.Sc. degree in Computer Science and Engineering from University of Science and Technology Chittagong (USTC) and M.Sc. degree in Computer Science and Engineering from Port City International University. Currently he is working as a lecturer in a college. He has published several papers in conferences and journal. His research interests include image processing, pattern recognition and machine learning.

Zainal Abedin received the BSc and MSc degree in Computer Science & Engineering from Chittagong University of Engineering and Technology (CUET), Chattagram, Bangladesh. He is a faculty member of the Department of Computer Science and Engineering at University of Science and Technology Chittagong. He published a good number of articles in international conferences and journals. His research interest includes Machine Vision with Deep Learning, Optimization of Machine learning model, Natural Language Processing, Signal Processing for Health care application and block chain technology.

How to cite this paper: Prashengit Dhar, Md. Zainal Abedin, " Bengali News Headline Categorization Using Optimized Machine Learning Pipeline", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.13, No.1, pp. 15-24, 2021. DOI: 10.5815/ijieeb.2021.01.02