https://doi.org/10.31436/iiumej.v22i2.1662

ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

JEELANI AHMED[*] AND MUQEEM AHMED

Department of Computer Science and Information Technology,
Maulana Azad National Urdu University, Hyderabad, India

[*]Corresponding author: Jeelani.jk@gmail.com

(Received: 18th October 2020; Accepted: 31st March 2021; Published on-line: 4th July 2021)

ABSTRACT: A massive rise in web-based online content today pushes businesses to implement new approaches and resources that might support better navigation, processing, and handling of high-dimensional data. Over the Internet, 90% of the data is unstructured, and there are several approaches through which this data can translate into useful, structured data—classification is one such approach. Classification of knowledge into a good collection of groups is significant and necessary. As the number of machine-readable documents proliferates, automatic text classification is badly needed to classify these documents. Unlabeled documents are categorized into predefined classes of labeled documents using text labeling, a supervised learning technique. This paper reviewed some existing approaches for classifying online news articles and discusses a framework for the automatic classification of online news articles. For achieving high accuracy, different classifiers were tried. Our experimental method achieved 93% accuracy using a Bayesian classifier and present in terms of confusion metrics.

ABSTRAK: Peningkatan tinggi pada masa kini pada maklumat dalam talian berasaskan web menyebabkan kaedah baru dalam bisnes telah diguna pakai dan sumber sokongan seperti navigasi, proses, dan pengurusan data berdimensi-tinggi adalah perlu. 90% data di internet adalah data tidak berstruktur, dan terdapat pelbagai kaedah data ini dapat diterjemahkan kepada data berguna, lebih berstruktur — iaitu melalui kaedah klasifikasi. Klasifikasi ilmu kepada koleksi kumpulan baik adalah penting dan perlu. Seperti mana mesin-boleh baca dokumen berkembang pesat, teks klasifikasi automatik juga sangat diperlukan bagi mengklasifikasi dokumen-dokumen ini. Dokumen yang tidak dilabel dikategori sebagai pengelasan pratakrif dokumen berlabel melalui teks label, iaitu teknik pembelajaran berpenyelia. Kajian ini mengkaji semula pendekatan sedia ada bagi artikel berita dalam talian dan membincangkan rangka kerja bagi pengelasan automatik artikel berita dalam talian. Bagi menghasilkan ketepatan yang tinggi, kami menggunakan pelbagai alat klasifikasi. Kaedah eksperimen ini mempunyai ketepatan 93% menggunakan pengelas Bayesian dan data dibentangkan berdasarkan matriks kekeliruan.

KEYWORDS: *text classification; naïve Bayes; support vector machine; news articles*

1. INTRODUCTION

Today, the overwhelming volume of digital content is expanding and growing constantly. Automatic text classification into existing categories is considered a primary method to process and manage this enormous amount of data.

This kind of textual data is generated from various sources and is available in conference materials, editorials, digital/electronic documents, web pages, emails, publications, and journals. Many people use these online sources for day-to-day information access instead of being limited to printed material such as newspapers, magazines, and books. The entrance to such information has become more comfortable; however, the organization of knowledge is difficult, making it challenging to manage. Organizing this kind of digital data is considered a critical method to effectively classify this digital information.

Text classification has a vital role in text retrieval, information abstraction and summarization, and question-answering. Usually, the data available on the web used for classification are diverse, from various sources like broadcast or printed news, bulletin boards, newsgroups, advertisements, and movie reviews. Because of their varied nature, like being multi-sourced and having different vocabularies, different formats, and different writing styles for documents, automatic text classification is essential [1]. Due to the enormous amount of text that is stored in the electronic format, it is necessary to understand and examine such data and extract similar details, which may be useful when making decisions [2-4].

When a document is classified under an already defined category this action is termed Text classification (Fig. 1). A document can be single labeled or multi-labeled, depending upon the classes. A document is termed as a single label when it is allocated to a single class and is named multi-label if the document assigns it to multiple classes [5].

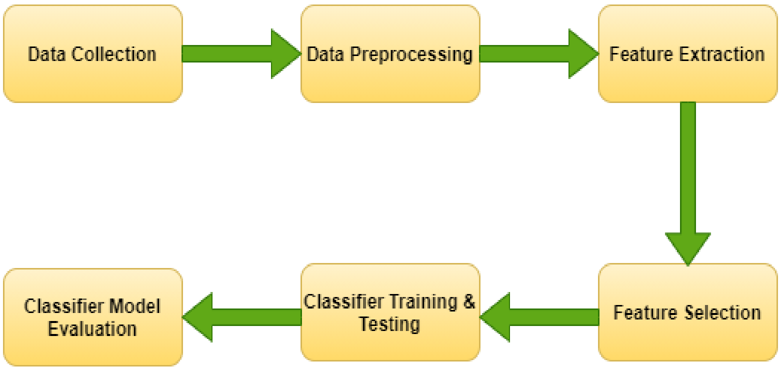


Fig. 1: Text classification process.

Text classification comprises feature selection, document representation, applying the algorithm, and performance assessment. Because of the sudden growth of online transactions and the colossal accessibility of text documents, it is essential to retrieve vital data from documents using classification. Today, organizing and managing text information is considered a crucial practice [6]. Neural Networks [8], Support Vector Machines [7], k-Nearest Neighbor [9], and Naïve Bayesian [10] classification are algorithms that may be used for constructing classification methods.

One area where we encounter a considerable amount of daily text is online news articles. Due to Information Technology developments, persons are worried about their busy lifestyle and thus they desire to only read news articles of their interests [11]. It is a big task to mine pertinent news concerning an individual's interest because much of the news articles are enlightening but could be less significant. An individual's lure can be contingent on many factors such as the news location and the type of news articles [12]. Here, we classify the news articles rendering to the type/category of the news articles. For example, imagine an individual who prefers to read news related to politics. Since the web is a sea full of news articles, it would be challenging for a particular person to read only articles related to politics.

In this work, to customize the news articles for a specific category, we have applied machine-learning techniques. The category may be politics, crime, sports, entertainment, or world news. We trained and tested the classification model using the huff post data set, which contains around 75k news articles. After training and testing, we applied this model for the news articles taken from different live news websites such as the Hindu, the Print, Indian express, and the Quint.

2. BACKGROUND

2.1 Naive Bayes (NB)

This algorithm is used as the classifier for the news articles. It is a prevalent method of classification when dealing with multi-class classification. It works based on the probabilistic technique first proposed by Lewis [13] and drives its roots from the Bayes theorem. This algorithm can run very efficiently with a large dataset. For text classification problems, Naïve Bayes is used as a standard because it has a fast run time as compared to other classifiers. Naïve Bayes is also used to solve problems like spam detection. Due to its simplicity, it outpaces many advanced classification techniques.

2.2 K-Nearest Neighbors (KNN)

The problems of regression and classification domain can be solved using this supervised machine learning algorithm [9]. KNN works by calculating the space between a query and all the instances (K) nearer to the question and then selects the majority recurrent label. It is comprehensible and applied very quickly. It slows down its performance when there is an upsurge in the data sample. It stores all the training data, due to which it is a somewhat costly algorithm. It requires high memory storage as compare to other methods. KNN is used in many areas, such as in politics; it can classify a voter into various classes like 'Will not Vote,' 'Will Vote,' or 'Will Vote to AAP Party'.

2.3 Support Vector Machine (SVM)

SVM is a supervised machine-learning algorithm [7]. SVM works when provided training data along with associated labels. Once the training is over, if supplied a data set, the model assigns a label to it. The support vector machine works well when dealing with linear classification problems. For classification, it makes a hyper-plane by choosing the maximum distance between adjacent data points. It is both a flexible and powerful technique of machine learning that is used for regression and classification. When dealing with high dimensional space, SVM works well and offers excellent accuracy, and it requires very little memory for processing.

2.4 Logistic Regression (LR)

It is the best regression analysis to be conducted where the dependent variable is binary. As for all studies of regression, logistic regression is a statistical process. It describes data and describes the relationship between one conditional dependent variable and one or more interval, ordinal, nominal, or ratio-level independent variables.

3. RELATED WORK

With the rapid increase in demand for managing substantial text databases, text classification is a popular and dynamic research field of data mining. It becomes necessary and essential to competently handle the textual data to search and access any document quickly due to the enormous growth of digital textual information [14,15].

Text classification uses unique rules to grant classes from a category of already specified classes to unlabeled text documents. Text classification typically operates manually, but because the classification rules are generated manually, such a procedure is time-consuming and costly. Therefore, Machine Learning is another methodology that uses automatic rule creation to classify documents [16].

The numerous classifiers are: Support Vector Machines [7], Neural Networks [8], Bayesian classification [10], Nearest Neighbor (KNN) [9], Association based classification [17-19], Term Graph Model [20,21] and Decision Tree (DT) [22,23], etc. Many authors use the mentioned machine learning techniques; the table below was formulated after a brief study of these techniques.

Table 1: Related work in the field of Text classification by various authors

References & year	Datasets	Algorithm / Technique used	Conclusion/Note
[24] (2014)	Reuter-21578	KNN algorithm	The accuracy of the KNN is maximum compared to Term Graph and Naïve Bayes. Its time complexity is maximum as compared to others is the only drawback.
[25] (2014)	Sohu laboratory corpus	SVM-KNN algorithm	By making recommendations and improving the classifying probability, the SVM-KNN algorithm will improve the classifier's efficiency. This algorithm increases the low amount of processing difficulty. Moreover, SVM is efficient and fast for classification.
[26] (2015)	Single-label news corpus	SVM, TF-IDF	For the classification of Indonesian news posts, the authors used algorithm adoption and problem transformation methods. The authors develop a combination technique for the automatic classifying of the news articles and achieve better results with an f-measure is 85.13%.
[27] (2015)	Reuters-21578	Hyper Rectangular Keyword Extraction	The authors proposed vital extraction based on the document categorization hyper rectangular method. This method generates excellent and accurate results when compared to traditional classification approaches.
[28] (2015)	Czech News Articles	Linear SVC, SGD, PA, NB	The authors proposed a method for enhancing the classification accuracy of multi-document classification. The baseline classifier Naïve Bayes performed better than other classifiers.
[29] (2016)	Various news websites	Naïve Bayes (NB)	Various news webpages were classified using structure attributes and URL content. The naïve Bayes algorithm showrd better results compared to existing approaches when implemented on the same dataset.
[30] (2016)	News articles of emotional topics	NB, C45, DT, SVM vectors, Winnow, Balanced Winnow, and Max Entropy	The authors experimented with hierarchical classification with the purpose of sentiment analysis of news articles. SVM and TF-IDF were used with six and four classification algorithms respectively and achieved better results.
[31] (2016)	Stock new of China	Softmax training algorithm, Weighted sort algorithm, and selection algorithm	For making investment decisions, authors classified stock news by proposing a novel algorithm. The authors achieved better accuracy and availability with this method compared with a general algorithm.

References & year	Datasets	Algorithm / Technique used	Conclusion/Note
[32] (2016)	Online news articles	Neural Network Classifier	The performance of the ANN using feature reduction performed better as compared to the essential ANN and gave better results for the classifications. The training process was prolonged.
[7] (2017)	Various news websites articles	Naïve Bayes, SVM, Random Forest	The performance of the SVM was at the bottom when compared with Naïve Bayes and Random Forest. Also, SVM took much training time.
[33] (2017)	Korean News Articles	KNN, NB, SVM, and Logistic Regression (LR)	The author performed two studies using four classifiers. Study 1, which had the data with less complexity, achieved a higher level of accuracy. Study 2, which had the data with more high complexity, may have harmed the classification results.
[34][2017]	Reuters.com news articles and Gold Standard dataset	SVM, SDG, LR	The authors used various classification algorithms to classify text articles to detect opinion, spam, and fake news. Classifiers achieved good results with an increase in the features.
[35] (2017)	Yahoo News US Edition	Keyword matching, Newsmap	The work developed a semi-supervised classifier for the classification of geographical news. The overall classification accuracy was 0.80, which was relatively low when compared to other traditional classifiers.
[36] (2018)	NLPCC2014, REV1-v2	Attention GRU Bi-RNN	Centered on the process of neural network interest, the authors proposed a bi-directional recurrent neural network algorithm that performed classification on two datasets and achieved a high accuracy of 83.9 compared to other classifiers.
[37] (2018)	BR news dataset US news dataset	ANOVA, SVM	The work performed classification on two news datasets and proved noticeable differences between reliable and unreliable news sources.
[38] (2018)	Indonesian news articles	Latent Dirichlet Allocation (LDA)	Classification results of Indonesian news articles represented using a word cloud, which allowed easy understanding of each category's results and trends; However, more explanation was needed in terms of classification.
[39] (2018)	China News	SVM, CNN, MAXENT	The authors applied three algorithms in two schemes for the classification of Chinese news articles. SVM performed better than the other two, but a detailed explanation of the experiment is needed.
[40] (2019)	News articles and tweets	Word2vec Convolutional Neural Networks	The algorithm used is classified tweets and news articles into related and unrelated words. Word2vec enhanced performance by learning the semantic relations between words.
[41] (2019)	Arabic news articles from websites	SVM, NB, DT, RF, LR	The work classified Arabic news articles collected from various news websites. SVM performed better than all other classifiers and achieved an accuracy of 87%.
[42] (2019)	Indonesian news articles	Convolutional neural networks, recurrent neural networks	The work performed a systematic classification of risk document to get the financial risk information in real-time. Authors achieved better classification results using extensive data. However, manual labeling of new data was a tedious process.

References & year	Datasets	Algorithm / Technique used	Conclusion/Note
[43] (2019)	Thai PBS, Khaosod, and Dailynews	SVM, Decision Tree, Deep Learning	The authors discussed the performance of various classifiers using Thai news as a dataset. Deep Learning outperformed the SVM and Decision Tree.
[44] (2019)	BBC datasets, five groups of 20Newsgroup	SVM, Naïve Bayes, Bi-LSTM, LSTM, CNN	This work used five classifiers for the classification of two Chinese news datasets. They also pointed out the difference between ML and DL classification that ML was a time-consuming classification process that required preprocessing and feature extraction. In contrast, DL did not require cleaning activities. However, using ML, they obtained efficient and reliable results and the best accuracy.
[45] (2020)	Society channel of Sina	C4.5 Decision Tree Algorithm	The authors proposed a classification approach for emotions in news articles. They classified emotions into fear, joy, and sadness. This method achieved a high accuracy of 87.83% and compared it with the SVM classifier.
[46] (2020)	Uzbek Daryo online news	SVM, DT, RF, LR, Multinomial Naive Bayes	The authors used six algorithms for news article classification in Uzbek languages. The approach achieved the highest accuracy of 86.88%.
[47] (2020)	Tigrigna news dataset	SVM, DT, LR, RF, KNN, etc.	The authors constructed the Tigrigna news dataset and have experimented with eight various classifiers. SVM outperformed other classification algorithms. Moreover, it achieved the highest accuracy.

We reviewed several papers; Table 1 contains some recent papers. Several scholars studied and categorized texts in various languages, such as Korean [33] and Chinese [39]. There were still several works available covering Arabic language classification [41], shedding light on academic articles based on the usage of classical classifications of supervised machine learning, like SVM [34,37], KNN [24,25], Decision Tree [45], and Naïve Bayes [7,29]. Although other authors focused on classification through neural networks [32] and deep learning [42], the overall output was higher. Finally, it is evident that the classification algorithm’s performance in every language was significantly impaired by the consistency of the data source, as the unnecessary and repetitive characteristics of the data undermined the precision and output of the classifier.

4. RESEARCH DESIGN AND IMPLEMENTATION

Based on the above literature, we proposed a framework to perform text classification of English news articles obtained from various Indian news websites. For ease of understanding, we divide the proposed framework into three main modules: Data Extraction, Data Preprocessing, and Classifier Module as illustrated in Fig. 2.

First, the dataset is cleaned using data preprocessing. Then, we divide the dataset into train and test parts. Train data hold 70%, and test data hold 30% of the dataset. The next module trains the classifier to predict the class labels for the collected news articles and assign a category label in the next step. Thereby performance evaluation of the classifier is carried out using some performance metric. The next section contains a detailed explanation of each module.

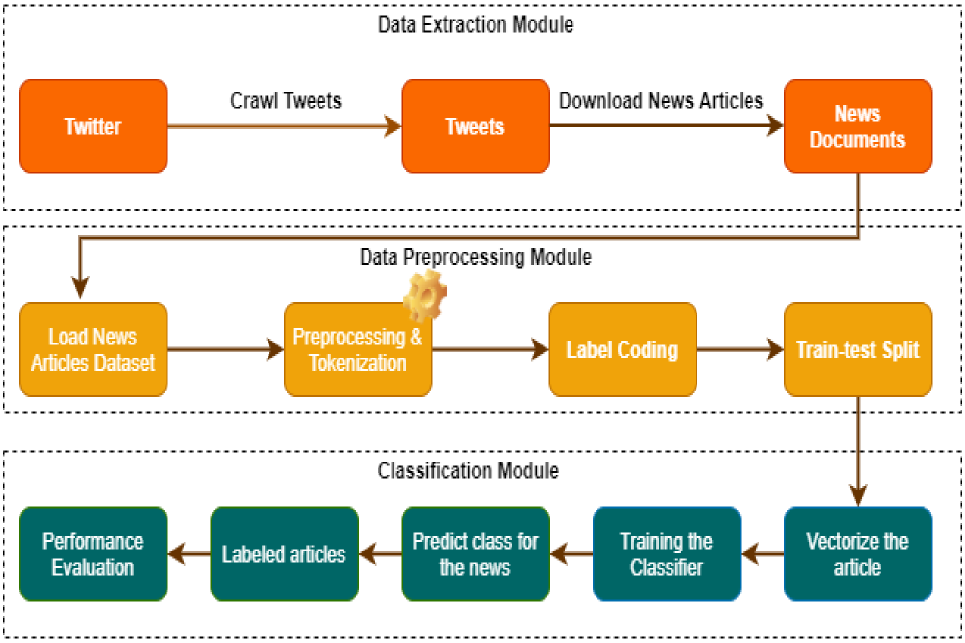


Fig. 2: News article classification process.

4.1 Data Extraction

The process starts with a data collection module, in which news articles from seven different news websites are crawled, as shown in Fig. 2. Firstly, tweets of various news Twitter handles are crawled from Twitter using tweepy. Various news articles were extracted from crawled tweet URLs. Further, news article web pages from each news website are extracted by visiting each URL. In the module, the data extraction module downloads tweets and news articles from seven news websites.

We have also used a dataset that contains around 75k news headlines with its content and category from the year 2012 to 2018 taken from Huff Post. These news articles contain eight categories: entertainment, crime, politics, business, world news, sports, media, and technology. The percentage of each class of news articles is shown in Fig. 3.

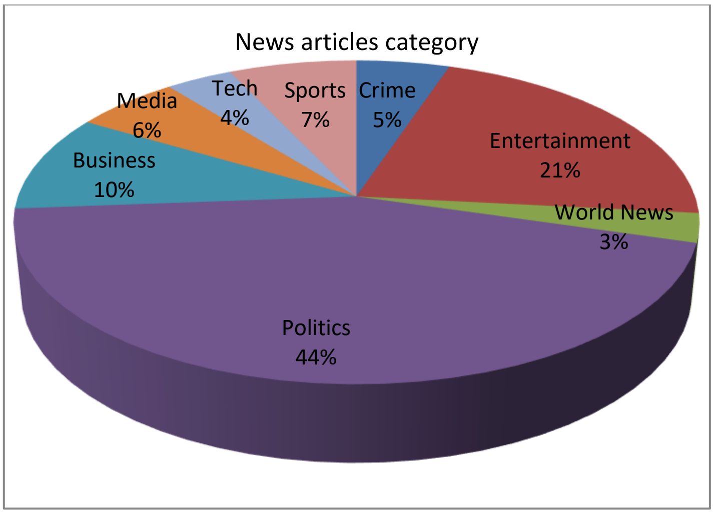
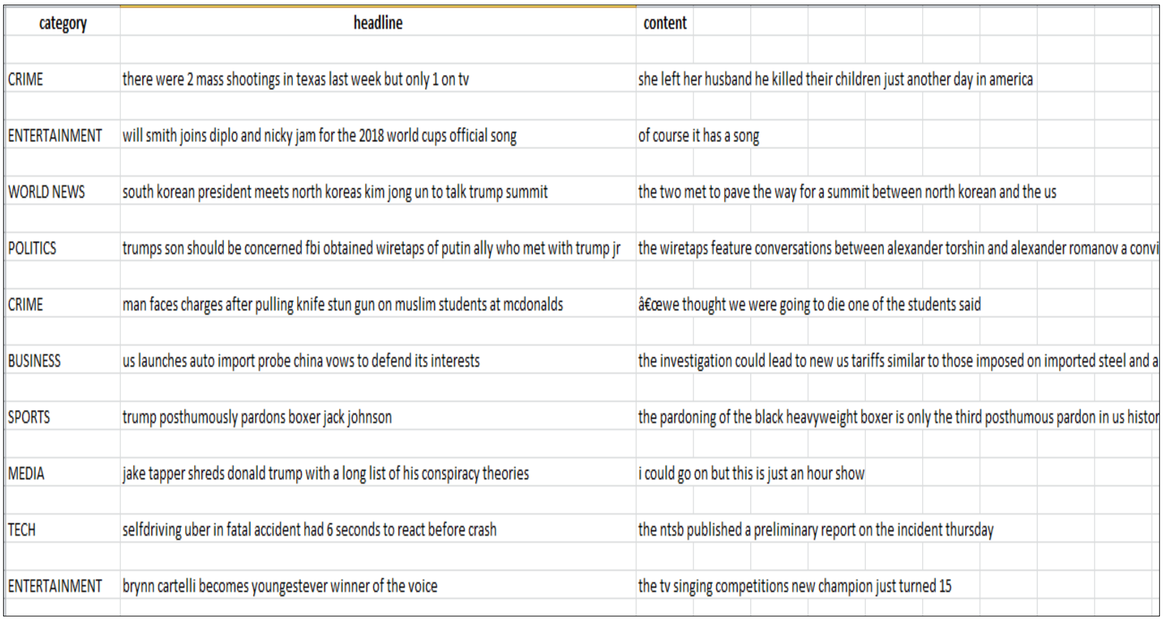


Fig. 3: Percentage of each news article category.

The dataset that was used is shown in Fig. 4. It contains the first ten lines of the dataset, which have eight categories of news articles in CSV format, as shown below.



category	headline	content
CRIME	there were 2 mass shootings in texas last week but only 1 on tv	she left her husband he killed their children just another day in america
ENTERTAINMENT	will smith joins diplo and nicky jam for the 2018 world cups official song	of course it has a song
WORLD NEWS	south korean president meets north koreas kim jong un to talk trump summit	the two met to pave the way for a summit between north korean and the us
POLITICS	trumps son should be concerned fbi obtained wiretaps of putin ally who met with trump jr	the wiretaps feature conversations between alexander torshin and alexander romanov a convi
CRIME	man faces charges after pulling knife stun gun on muslim students at mcdonalds	â€œwe thought we were going to die one of the students said
BUSINESS	us launches auto import probe china vows to defend its interests	the investigation could lead to new us tariffs similar to those imposed on imported steel and a
SPORTS	trump posthumously pardons boxer jack johnson	the pardoning of the black heavyweight boxer is only the third posthumous pardon in us histor
MEDIA	jake tapper shreds donald trump with a long list of his conspiracy theories	i could go on but this is just an hour show
TECH	selfdriving uber in fatal accident had 6 seconds to react before crash	the ntsb published a preliminary report on the incident thursday
ENTERTAINMENT	brynn cartelli becomes youngestever winner of the voice	the tv singing competitions new champion just turned 15

Fig. 4: Description of the dataset.

4.2 Data Preprocessing

The data preprocessing module performs the cleaning of the dataset, which is considered to be an important task to achieve good results. Firstly, we perform the tokenization of articles; the module changes a group of characters to groups of strings with some recognizable meaning. Next, by using the python nltk package, we remove stop words such as 'what' and 'the' as these are the words that have minimal prominence and occur commonly.

4.2.1 Label Encoding

We used the LabelEncoder class from the Scikit-Learn library in python to transform categorical text data into model comprehensible numerical data. From the Scikit-learn library, we use LabelEncoder class to encode the first column, then fit and transform the data of that column. Then, the new encoded data replaced the existing text data. After running the LabelEncoder piece of code, we get the following table.

Table 2: Label encoding for different news category

Category Name	Category Code
Crime	0
Entertainment	1
World News	2
Politics	3
Sports	4
Business	5
Media	6
Tech	7

4.2.2 Train-test Split

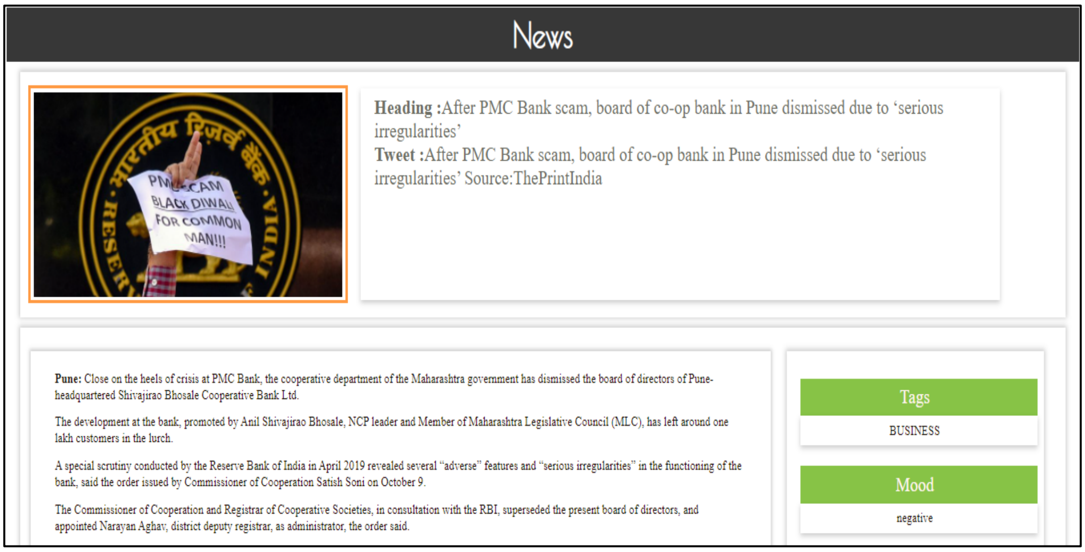
Our work includes a single labeled classification of the news articles collected from seven different news websites. For training and testing purposes, we divided our dataset into 70% for training and 30% for testing. The training data set consisted of 52635 labeled news articles and the testing set consisted 22558 of news articles. In training data, we have used the 10 fold cross-validation, fitted the final model to it, and then evaluated with the unclassified data and obtained evaluation metrics that showed as little bias as possible.

4.3 Training the Classifier

The training set and the set of labels corresponding to it are the input data for the classifier. The labels we used along with training set are: 'Crime', 'Entertainment', 'World News', 'Politics', 'Sports', 'Business', 'Media', and 'Tech'. Based on the pre-decided category tag, processed news articles were numerically labeled. Because the classifier input consists of two vectors, it was essential to vectorize news articles and the labels. They functioned as an input to the classifier after the vectorization of these two entities.

4.4 Testing the Classifier

The classifier model was tested on the testing data once the classifier was trained with the trained data. The classifier predicted the category of the corresponding news articles. An example of the news article's 'business' category from the testing set is presented in Fig. 5. Naive Bayes classifier performed better, with 93% accuracy. Furthermore, our model is sure that the above 'business' article would belong to the same category.



Heading : After PMC Bank scam, board of co-op bank in Pune dismissed due to ‘serious irregularities’
Tweet : After PMC Bank scam, board of co-op bank in Pune dismissed due to ‘serious irregularities’ Source:ThePrintIndia

Pune: Close on the heels of crisis at PMC Bank, the cooperative department of the Maharashtra government has dismissed the board of directors of Pune-headquartered Shivajirao Bhosale Cooperative Bank Ltd.

The development at the bank, promoted by Anil Shivajirao Bhosale, NCP leader and Member of Maharashtra Legislative Council (MLC), has left around one lakh customers in the lurch.

A special scrutiny conducted by the Reserve Bank of India in April 2019 revealed several “adverse” features and “serious irregularities” in the functioning of the bank, said the order issued by Commissioner of Cooperation Satish Soni on October 9.

The Commissioner of Cooperation and Registrar of Cooperative Societies, in consultation with the RBI, superseded the present board of directors, and appointed Narayan Aghav, district deputy registrar, as administrator, the order said.

Tags
BUSINESS

Mood
negative

Fig. 5: News article classified as business category.

4.5 Performance Metrics

We made use of Scikit-learn to implement all the classification algorithms. The performance measurements for classifiers were studied to represent the Act imbalances on conventional measuring instruments like accuracy, recall, and precision. These metrics use knowledge about the classes currently and projected classes from classification tasks. The following confusion matrix may reflect all possible scenarios of the findings.

Table 3: Confusion matrix

		PREDICTED	
		Positive	**Negative**
ACTUAL	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

When an instance belongs to a class, it is considered a positive case, and when an instance does not belong to a specific class, it is considered a negative case. An instance is called true positive (TP) when it belongs to the positive case and is appropriately labeled as such. The false negatives (FN) are such cases that are incorrectly classified as negative cases when they belong to the positive cases. The false positives (FP) are such cases that are incorrectly classified as positive when they belong to the negative cases. The instances belonging to the negative case are correctly classified as such. These are called the true negatives (TN). Below we highlight some of the performance metrics based on the confusion matrix.

4.5.1 Accuracy

Accuracy is the number of all right predictions the classifier has made divided by the overall data collection [48]. The accuracy mathematically expressed as

$$Accuracy = \frac{True\ Positive + True\ Negative}{(True\ Positive + False\ Positive + True\ Negative + False\ Negative)} \tag{1}$$

The error rate (ERR) is the sum of all incorrect forecasts separated by the dataset's total number. The highest error rate is 0.0, while 1.0 is the worst. The error mathematically expressed as

$$Error\ Rate = 1 - Accuracy \tag{2}$$

4.5.2. Precision

The number of articles that are properly identified (True Positive) by the number of articles that the classifier estimates corresponds to a specific group (True Positive and False Positive). The precision [48] in mathematical form is

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \tag{3}$$

4.5.3. Recall

The ratio of positive articles correctly predicted to all articles in the actual class is termed a recall [49]. The recall is mathematically expressed as

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \tag{4}$$

A good recall implies that much of the positive cases (TP+FN) are rated as positive (TP). This situation is likely to result in more FP measurements and lower average precision. A poor recall implies we have a significant FN volume (should have been positive but categorized as negative). This ensures that we have greater confidence that this will be a positive case if we find a positive scenario.

On imbalanced classification, when scholars use machine learning techniques, accuracy is not accurate. In imbalanced conditions, higher accuracy can be obtained since all data is predicted as the majority class. Hence, the F1 score and the receiver operating characteristic (ROC) curve are accepted as fair measurements by the machine learning community. The tradeoff between false positive and true positive is indicated by the ROC curve. When the false positives get overlooked, and the emphasis is distorted, it would probably reflect precision only for the true positives. On the other hand, if the real positives get overlooked, and the emphasis is distorted for false positives, the ratings would most definitely represent the recall. The classifier's efficiency is reflected by the area under the curve (AUC).

4.5.4 F1-Score

The harmonic mean of precision and recall is the F1 score. If we combine Precision and Recall, then it will become the F1 score [50]. The optimal and worst values for F1 score are 1 and 0, respectively. It is efficient to use one value for measurement instead of using two precision and recall values as it combines both. F1 score is presented as

$$F1\ score = \frac{2 * Recall * Precision}{(Recall + Precision)} \tag{5}$$

5. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

For every classification algorithm checked on our dataset, Table 4 displays the assessment of comparative results of SVM, LR, KNN, and Naive Bayes algorithms using multiple news websites in terms of all mentioned performance metrics. Accuracy and F1 score are nearly similar. Figure 6 illustrates the analysis of all four algorithms in graphical form.

Table 4: Result table

Classifier	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.93	0.93	0.93	0.9
LR	0.81	0.75	0.79	0.81
SVM	0.76	0.76	0.76	0.75
KNN	0.72	0.64	0.52	0.72

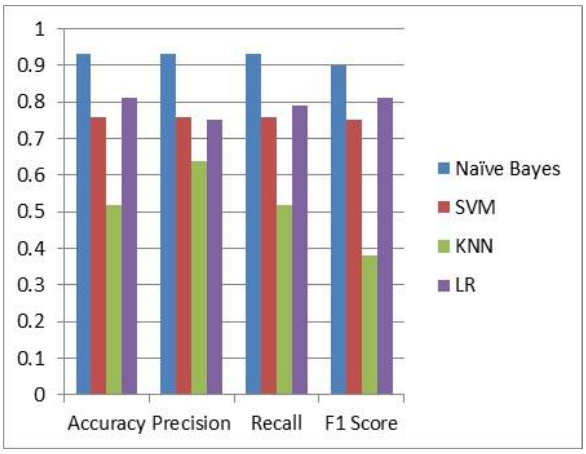


Fig. 6: Comparing all four classifiers.

In all four classifiers, the Naive Bayes classifier performed better than others with 93.0% accuracy. Moreover, the worst result was displayed by the KNN classifier at 72.0%. The best (NB) and worst (KNN) classifier's confusion matrices are shown in Fig. 7 and Fig. 8.

Figure 9 shows classifier output by imbalanced ratios. When training data for minority groups is sparse compared to other algorithms, Naive Bayes worked better than the other classifiers. It touched a 0.5 F1 score while the imbalance was 11%, whereas others reached 22% of the imbalance with the similar F1 score. Nonetheless, the efficiency of the classifier will not improve after imbalanced levels dropped to 44%, whereas some algorithms eventually demonstrated better results after more class imbalance training data were produced. Overall, the lowest result was obtained by the KNN classifier. To attain an average F1 score above 0.7, it took an imbalance level of 66 percent, while the majority of the classifiers achieved an average F1 score above 0.7 at a 44 percent disparity rate. In the most imbalanced conditions, Naive Bayes worked the best.

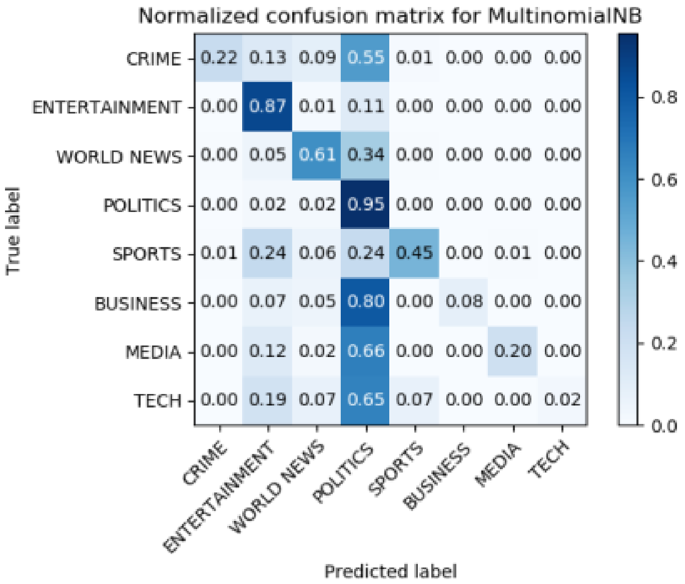


Fig. 7: Confusion matrix for the Naive Bayes classifier.

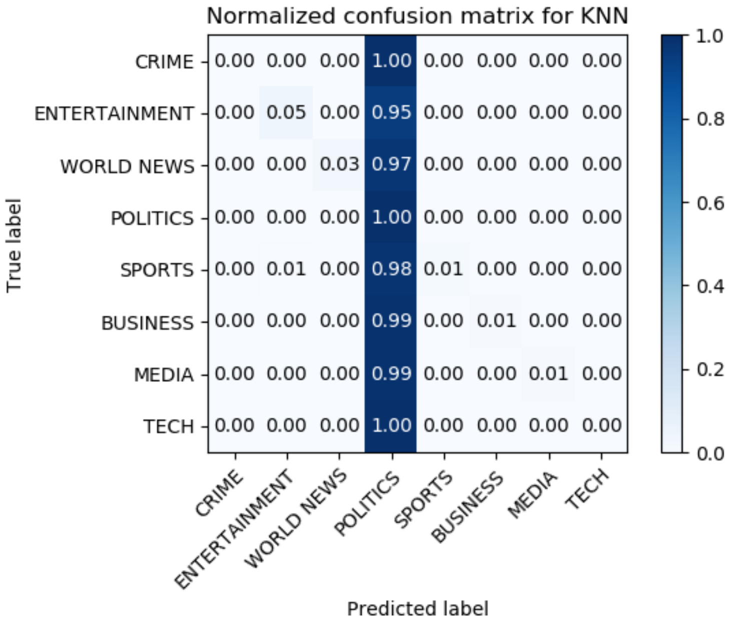


Fig. 8: Confusion matrix for the KNN classifier.

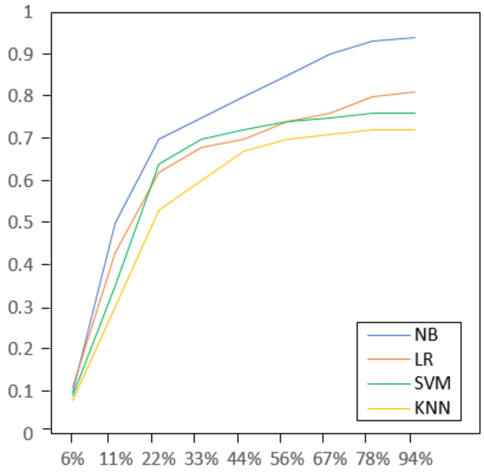


Fig. 9: Performance of classifiers - Naive Bayes performs best while KNN performs worst.

6. CONCLUSION

In this research, for online news posts, we implemented a single-class text classifier system. We presented a sample dataset containing approximately 75k news articles sliced from seven various websites with their tags. We defined the dataset's collection, cleaning, and construction phases. We analyzed our dataset by adding four distinct classifiers. We used Naïve Bayes and several other classification algorithms for classification and performed a comparison of different classifiers' outcomes from seven various sites on the same dataset, and the findings confirm that Naïve Bayes showed better performance than most classifiers; when working with various news datasets, it offers sufficient classification accuracy. This study has many possible extensions. Our future aim is to discover and implement a classification methodology in various regional languages.

ACKNOWLEDGEMENT

The authors send their humble gratitude to the university for providing all the facilities during this research.

REFERENCES

[1] Jindal R, Malhotra R, Jain A. (2015) Techniques for text classification: Literature review and current trends. Webology, 12(2): Article 139. https://www.webology.org/2015/v12n2/a139.pdf

[2] Turney P. (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Computing Research Repository, 417-424. doi:10.3115/1073083.1073153.

[3] Wilson T, Wiebe J, Hoffmann P. (2009) Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. Computational Linguistics, 35(3): 399-433. doi:10.1162/coli.08-012-r1-06-90

[4] Quan C, Ren F. (2009) Construction of a blog emotion corpus for Chinese emotional expression analysis. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 3: 1446-1454.

[5] Wang TY, Chiang HM. (2011) Solving multi-label text categorization problem using support vector machine approach with membership function. Neurocomputing, 74(17): 3682-3689. https://doi.org/10.1016/j.neucom.2011.07.001

[6] Harrag F, El-Qawasmah E, Al-Salman AMS. (2010) Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm. In Proceedings of the 2010 First International Conference on Integrated Intelligent Computing, Bangalore, India, 2010, pp 6-11. https://doi.org/10.1109/iciic.2010.23

[7] Sapankevych N, Sankar R. (2009) Time series prediction using support vector machines: A survey. IEEE Computational Intelligence Magazine, 4(2): 24-38. https://10.1109/MCI.2009.932254.

[8] Zhihang Chen, Chengwen Ni, Murphey, YL. (2006) Neural Network Approaches for Text Document Categorization. In Proceedings of the IEEE International Joint Conference on Neural Network Proceedings, pp.1054–1060. https://doi.org/10.1109/ijcnn.2006.246805

[9] Zhang X, Bicheng Li, Xianzhu Sun. (2010) A k-nearest neighbor text classification algorithm based on fuzzy integral. In Proceedings of the Sixth International Conference on Natural Computation, pp 2228–2231. https://doi.org/10.1109/icnc.2010.5584406

[10] Martinez-Arroyo M, & Sucar LE. (2006) Learning an Optimal Naive Bayes Classifier. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), pp 748–752. https://doi.org/10.1109/icpr.2006.748

[11] Pendharkar B, Ambekar P, Godbole P, Joshi S, Abhyankar S. (2007) Topic categorization of RSS news feeds. Group, 4, 1.

[12] Rao V, Sachdev J. (2017) A machine learning approach to classify news articles based on location. In Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS), pp 863-867. https://doi.org/10.1109/iss1.2017.8389300

[13] Lewis DD. (1998) Naive (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning, 4-15). Springer, Berlin, Heidelberg

[14] Chen N, Blostein D. (2006) A survey of document image classification: problem statement, classifier architecture and performance evaluation. International Journal of Document Analysis and Recognition (IJDAR), 10(1): 1-16. https://doi.org/10.1007/S10032-006-0020-2

[15] Gupta V, Lehal GS. (2009) A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence, 1(1): 60-76. https://doi.org/10.4304/jetwi.1.1.60-76

[16] Manning CD, Raghavan P, Schutze H. (2008) Introduction to information retrieval? Cambridge University Press, pp 405-416.

[17] Li W, Han J, Pei J. (2001) CMAR: Accurate and efficient classification based on multiple class-association rules. In Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 2001, pp. 369-376. doi: 10.1109/ICDM.2001.989541.

[18] Yin X, Han J. (2003) CPAR: Classification based on predictive association rules. In Proceedings of the 2003 SIAM International Conference on Data Mining, pp 331-335.

[19] Berzal F, Cubero J, Marín N, Sánchez D, Serrano J, Vila A. (2005) Association rule evaluation for classification purposes. Actas del III Taller Nacional de Mineria de Datos y Aprendizaje, pp 135-144.

[20] Jiang C, Coenen F, Sanderson R, Zito M. (2010) Text classification using graph mining-based feature extraction. In Research and Development in Intelligent Systems XXVI, Springer, London, pp 21-34.

[21] Huynh D, Tran D, Ma W, Sharma D. (2011) A new term ranking method based on relation extraction and graph model for text classification. In Proceedings of the Thirty-Fourth Australasian Computer Science Conference, 113: 145-152.

[22] Han J, Kamber M. (2001) Data mining concepts and techniques, Morgan Kaufmann Publishers. San Francisco, CA, pp 335-391.

[23] Chen J, Huang H, Tian S, Qu Y. (2009) Feature selection for text classification with Naïve Bayes. Expert Systems with Applications, 36(3):5432-5435. https://doi.org/10.1016/j.eswa.2008.06.054

[24] Bijalwan V, Kumar, V, Kumari P, Pascual J. (2014) KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application, 7(1):61-70. https://doi.org/10.14257/ijdta.2014.7.1.06

[25] Lin Y, Wang J. (2014) Research on text classification based on SVM-KNN. In Proceedings of the IEEE 5th International Conference on Software Engineering and Service Science, Beijing, China, pp 842-844. https://doi.org/10.1109/ICSESS.2014.6933697.

[26] Rahmawati D, Khodra ML. (2015) Automatic multi-label classification for Indonesian news articles. In Proceedings of the 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Chonburi, Thailand, 2015, pp 1-6. https://doi.org/10.1109/ICAICTA.2015.7335382.

[27] Hassaine A, Mecheter S, Jaoua A. (2015) Text categorization using hyper rectangular keyword extraction: Application to news articles classification. In Proceedings of the International Conference on Relational and Algebraic Methods in Computer Science, pp 312-325. https://doi.org/10.1007/978-3-319-24704-5_19

[28] Lehečka J, Švec J. (2015) Improving multi-label document classification of Czech news articles. In Proceedings of the International Conference on Text, Speech, and Dialogue, pp 307-315.

[29] Arya C, Dwivedi SK. (2016) News web page classification using URL content and structure attributes. In Proceedings of the 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 2016, pp. 317-322, https://doi.org/10.1109/NGCT.2016.787743.

[30] Li J, Fong, S, Zhuang, Y, Khoury, R. (2015) Hierarchical classification in text mining for sentiment analysis of online news. Soft Computing, 20(9), 3411-3420. https://doi.org/10.1007/s00500-015-1812-4

[31] Weng, W., Liu, Y., Wang, S., & Lei, K. (2016) A multiclass classification model for stock news based on structured data. In Proceedings of the Sixth International Conference on Information Science and Technology (ICIST), Dalian, China, pp 72-78. https://doi.org/10.1109/ICIST.2016.7483388.

[32] Kaur S, Khiva NK (2016). Online news classification using deep learning technique. International Research Journal of Engineering and Technology, 3(10): 558-563.

[33] Suh Y, Yu J, Mo J, Song L. (2017) A comparison of oversampling methods on imbalanced topic classification of Korean news articles. Journal of Cognitive Science, 18(4): 391-437.

[34] Ahmed H, Traore I, Saad S. (2017) Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9. https://doi.org/10.1002/spy2.9

[35] Watanabe K. (2017) Newsmap. Digital Journalism, 6(3): 294-309. https://doi.org/10.1080/21670811.2017.1293487

[36] Du C, Huang L. (2018) Text classification research with attention-based recurrent neural networks. International Journal of Computers Communications & Control, 13(1): 50-64. https://doi.org/10.15837/ijccc.2018.1.3142

[37] Gruppi M, Horne BD, Adali S. (2018) An exploration of unreliable news classification in Brazil and the US. arXiv preprint arXiv:1806.02875.

[38] Kusumaningrum R, Adhy S. (2018). WCLOUDVIZ: Word cloud visualization of Indonesian News Articles Classification Based on Latent Dirichlet Allocation. Telkomnika, 16(4): 1752-1759.

[39] Cecchini D, Na L. (2018) In Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, pp 681-684. https://doi.org/10.1109/BigComp.2018.00125.

[40] Jang B, Kim I, Kim JW. (2019) Word2vec convolutional neural networks for classification of news articles and tweets. PLOS ONE, 14(8): e0220976. https://doi.org/10.1371/journal.pone.0220976

[41] Qadi LA, Rifai HE, Obaid S, Elnagar A. (2019) Arabic text classification of news articles using classical supervised classifiers. In Proceedings of the 2nd International Conference on new Trends in Computing Sciences (ICTCS), Amman, Jordan, pp 1-6. https://doi.org/10.1109/ICTCS.2019.8923073.

[42] Gumilang M, Purwarianti A, Nurdinasari F. (2019) Systemic risk document classification on Indonesian news articles using deep learning and active learning. In Proceedings of the International Conference on Electrical Engineering and Informatics (ICEEI), Bandung, Indonesia, pp 46-51. https://doi.org/10.1109/iceei47359.2019.8988829

[43] Noppakaow A, Uchida O. (2019) Examinations on the Performance of Classification Models for Thai News Articles. In Proceedings of the 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), Pattaya, Thailand, 2019, pp. 1-4 https://doi.org/10.1109/iciteed.2019.8929959

[44] Huang CM, Jiang YJ. (2019) An empirical study on the classification of Chinese news articles by machine learning and deep learning techniques. In Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, pp 1-6. https://doi.org/10.1109/icmlc48188.2019.8949309

[45] Winster SG, Kumar MN. (2020) Automatic classification of emotions in news articles through ensemble decision tree classification techniques. Journal of Ambient Intelligence and Humanized Computing, 1–12. https://doi.org/10.1007/s12652-020-02373-5

[46] Rabbimov IM, Kobilov SS. (2020) Multi-class text classification of Uzbek news articles using machine learning. Journal of Physics: Conference Series, 1546(1): 012-097.

[47] Fesseha A, Xiong S, Emiru ED, Dahou A. (2020) Text classification of news articles using machine learning on low-resourced language: Tigrigna. In Proceedings of the 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, pp 34-38. https://doi.org/10.1109/ICAIBD49809.2020.9137443.

[48] Sharma A, Mishra PK. (2020) State-of-the-art in performance metric and future direction for data science algorithm. Journal of Scientific Research, 64(2): 221-238.

[49] Saura JR. (2020) Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics. Journal of Innovation & Knowledge, 6(2): 92-102. https://doi.org/10.1016/j.jik.2020.08.001

[50] Pereira L, Nunes N. (2020) An empirical exploration of performance metrics for event detection algorithms. Non-Intrusive Load Monitoring. Sustainable Cities and Society, 62: 102399. https://doi.org/10.1016/j.scs.2020.102399