Int J Syst Assur Eng Manag (June 2022) 13(3):1341–1355
https://doi.org/10.1007/s13198-021-01471-7

ORIGINAL ARTICLE

An intelligent model based on integrated inverse document frequency and multinomial Naive Bayes for current affairs news categorisation

Sachin Kumar[1] · Aditya Sharma[1] · B Kartheek Reddy[1] · Shreyas Sachan[1] · Vaibhav Jain[1] · Jagvinder Singh[2]

Received: 6 December 2020 / Revised: 20 September 2021 / Accepted: 19 October 2021 / Published online: 7 November 2021

Abstract Digital technologies, their product and services have empowered the masses to generate information at a faster pace. Digital technologies based information sharing platforms such as news websites and social media platforms such as Facebook, Twitter, Instagram, What's app etc have flooded the information space due to the easy generation of information and dissemination to the masses instantly. Information classification has been an important task, especially in newspapers and media organisations. In another area also, information or text classification has an important role to play so that important and vital information can be classified based on the already predefined categories. In journalism, editors and resources persons were allocated the task to recognise and classify the news stories so that they can be placed in the predefined categories of economy and business news, political news, social news, editorial section, education and career, and sports information etc. Nowadays the process of classification and segregation of textual information has become challenging due to the flow of diverse, vast information. Additionally, the pace of information and its updates, access and competition among the media House have made it more challenging. Hence automated and intelligent tools which can classify the information and text accurately and efficiently is needed to reduces human efforts, time and increase productivity. This paper presents an intelligent, efficient and robust intelligent machine learning model based on Multinomial Naive Bayes(MNB) to classify the current affairs news stories. The proposed Inverse Document Frequency(IDF) integrated MNB model achieves classification accuracy of 87.22 per cent. The experiment results are also compared with other machine learning models such as Logistics Regression(LR), Support Vector Machine(SVM), K-Nearest Neighbours(KNN) and Random forest(RF). The results demonstrate that the presented model is better in term of accuracy and may be deployed in real world information classification and media domain to improve the productivity, efficiency of the current affairs news classification process.

Keywords News Articles · Classification · Intelligent Methods · Machine Learning · Support Vector Machine · Multinomial Naive Bayes · Inverse Document Frequency(IDF)

✉ Sachin Kumar
skumar@cic.du.ac.in

Aditya Sharma
aatom098@gmail.com

B Kartheek Reddy
bkartheekreddy@gmail.com

Shreyas Sachan
shreyas.sachan@gmail.com

Vaibhav Jain
vabsweb@gmail.com

Jagvinder Singh
jagvinder.singh@gmail.com

[1] Cluster Innovation Centre, University of University, Delhi, India

[2] Dept of Management, Delhi Technological University, Delhi, India

1 Introduction

Present-day technological advancements, especially digital technologies have transformed the way information is generated, transmitted and disseminated (Van Veldhoven and Vanthienen 2021). Digital technologies and software tools developed have immense power and scale to circulate information all over the world (Ashuri 2016). At present, most news and media organisations responsible for sharing information have maintained their online web and app platform to reach the masses in a very efficient and effective manner (Bail 2017). The news and information collection, processing, publication and dissemination has become very challenging with new technological tools being used and upgraded in each of the phases of the news life cycle (Canito et al. 2018). These phases of the news life cycle require automated and intelligent methods of information processing to classify the information and news stories. In the present day competitive world, it requires fast, efficient, effective, intelligent solutions to be deployed in each phase so that the news reaches the masses in a very fast manner and the classification process is automated (Hogenboom et al. 2016). The digital news articles platforms and the media industry is growing and there is a need to create an intelligent system that automates hectic and time-consuming tasks such as manual news categorisation after reading the articles at the primary level (Thomson et al. 2020). For example, news articles come for publication, media organisations need an expert to classify them after reading so that they can be published in the appropriate section of newspapers or websites. This is a time-consuming process. Artificial Intelligence with Natural Language Processing(NLP) research and development can reduce and accurately classify the information automatically (Marconi 2020). Such practical issues come under text processing and categorisation in traditional models of classification works. For example, classification of the news related text in natural languages such as English, Hindi or regional languages into a predefined set of categories such as politics, business, entertainment, health, sports etc (Medagoda 2016). The mentioned classification task requires a lot of human effort of reading the article or asking the author itself for the classification but accuracy needs to be ensured (Daud et al. 2017). The large volume of the articles, their large quality and the complexity of information and facts necessitates the efficient system automating the classification process for the news agencies and organisations. News article classification, or in broader terms the text classification, have changed from core mathematical models to statistical and machine learning (Hu et al. 2017). Most of the research work in the fields of text classification in news articles have been short of producing benchmark level accuracy of the text classification. Some of the work produced, used only headlines and tags of the news stories leaving the body of the news, which did not give the good accuracy of the news story classification (Kanan and Fox 2016). Such models suffered from the loss of complex information ingrained in the natural language and did not result in the best accuracy for the models (Salminen et al. 2019). Additionally, they suffered from over fitting and did not make them eligible to be deployed in a real-world environment in practical situations (Singh et al. 2019). This research work creates an intelligent model to facilitate news article categorisation which will aid the users, writers, editors and news organisations and other stakeholders to accurately label the news articles with minimal efforts in terms of human and computational resources Miller and (Oswalt 2017). The proposed model reads the whole article with title and body of news and tags and extracts meaningful relationships between words of the news articles to label the article into one of the following categories: Business and Economy, Education and Career, Entertainment, Food and Health, International, Politics and Governance, Science and Technology, and Sports. The research work of the paper proposes the following three improvements and findings.

1 It proposes a model which takes news titles and news stories for classification of the news articles which is an improvement over previous work.
2 It proposed a machine learning-based model which is very efficient and effective in news categorisation.
3 It performs classification and comparison on newly developed data set of news stories.

This paper is organised in the following manner. Section 1 introduces the background of the research problem and its context. Section 2 highlights the recent work done in the domain and research gaps. Section 3 describes the mathematical models, their foundations with evaluation criteria. Section 4 explains the methodology of the experimentation and proposed model. Section 5 discusses the results and provides their analyses in the comparison to the existing work. Section 6 concludes the paper.

2 Literature survey

Text processing is one of the important and challenging topics for the researchers. Getting insights from diverse texts after their processing have several useful applications such as sentiment analysis, information categorisation, user feedback, language generation, hate speech and abusive speech detection, misinformation campaign detection etc. (Ur-Rahman and Harding 2012). With such diverse applications, several methods of text processing and

classification have been developed. The developed models have their foundations on mathematical and statistical inferences. Recent developments in intelligent models based on Artificial Intelligence/machine learning have also been used. Machine learning based methods have been applied in several diverse fields such as energy consumption predictions (Kumar et al. 2019, 2018b), pattern recognition (Kiranyaz et al. 2014), occupancy detection (Kumar et al. 2020), health informatics and governance (Ravì et al. 2016; Kumar et al. 2018c), fake news detection (Conroy et al. 2015), education and learning Lykourentzou et al. (2009), object recognition (Ramík et al. 2014), financial predictions (Ahmed et al. 2016; Lin et al. 2011), intelligent transportation Kumar et al. 2018a, environment and weather prediction (Rasouli et al. 2012). Machine learning approaches involve basic tasks such as classification, regression and clustering (Witten et al. 2005). Several studies have been conducted in text classification and processing which have used a variety of methods (Fong et al. 2013). Some of research work have developed approaches based on rigid mathematical understanding and decision support systems (Pröllochs et al. 2016). The rigid mathematical models suffer from low accuracy in a dynamic environment. Other statistical approaches have their own limitations (Paragios et al. 2006). Machine learning methods have produced some of the relevant models of some of the natural language processing work. Boumans et. al. worked on content analysis for digital journalism (Boumans and Trilling 2016). The work provided an automated content analysis for online news and print news. Alhothali et al. (2015) applied affect control theory to analyze readers'reaction towards news articles. Text classification and processing have been applied in health domain as well for developing personalised health systems. Mukwazvure et al. (2015) even provided a hybrid approach to analyze sentiments of news stories and their comments. Some have used hierarchical classification approaches for the classification of online news articles (Li et al. 2016) with heading only. Valdivia et. al. (2017) applied machine learning methods for the sentiments analysis and text analysis of trip advisor website. Several techniques have been proposed for the improvement in the accuracy of the machine learning text processing methods (Carstens and Toni 2017; Onan et al. 2016). Some researchers have also applied machine learning approaches on the news data collection and processing. Wen-Lin Hsu et. al. (1999) proposed various intelligent and conducted experiments on NETNEWS data set. The proposed were based on vector space paradigm and categorised the news articles quite efficiently. Carlos et.al. (2014) presented a study of the new article's life cycle using data from large international news networks and social media. The data set was generated had more than 3,000,000 website visits and 200,000 social media reactions. They showed that it is quite possible to accurately predict and model the overall traffic articles received in their life cycle by observing the first 10–20 minutes of social media reactions. Michela et.al. (2016) demonstrated a quantitative cross-platform analysis of public discourse and news consumption on online social media on the Italian referendum. They found that users from well-separated communities tend to restrict their attention to a specific set of articles on Facebook and Twitter. From Hakim et al. (2014), proposed a frequency-inverse document frequency (TF-IDF) framework which was weighting for classification of news articles into sports, business and science and technology categories. In their approaches, for assigning weights to the words of articles, the TF-IDF scheme was used, for which each news article is transformed into a vector of weights. For the training of the system, a total of 300 categorized news articles were collected and were tested on 60 randomly extracted articles. In their work, Yaocheng Gui et. al. (2012). used a different approach to classify news articles. They used a hierarchical text classification approach to classify text data documents into a pre-generated hierarchy of news article categories. Their work utilizes named entities as features for the classification task of news articles into a pre-generated hierarchy of articles about international relations. The experiment results analysis reported improved performance of classification for news articles using the named entities. After going through the recent work of article classification, it has been observed that the works have majorly used only the title/headline of news stories and some of the work have used just tags and body of news article alone. In the present work, we propose models based on machine learning with integrated text processing techniques for news article classification in which the input data includes three dimensions such as title of news articles, news article body and their tags/keywords. Additionally in our work, the number of news articles categories to be classified have been raised from 3 to 8 with newly generated data set.

3 Foundational approaches

3.1 Machine learning approaches

News classification problem is text processing and classification problem and can be solved using supervised machine learning approaches. Supervised learning problems deal with labeled data sets paired with an input object and the desired output value (Sen et al. 2020). It is processed by the machine learning approaches with some other feature engineering approaches to do the text categorisation. Following are the mathematical foundations of

the machine learning methods used in the research work. Logistic Regression (LR) Kurt et al. (2008) is considered one of the simplest machine learning methods, LR tries to predict, quantitatively, probability of a classification outcome having only two values. The prediction is based on the use of one or more predictors. These predictors may be of type numerical and categorical. A logistic model (P) is represented by as follows

$$P = h(x) = 1/(1 + e(-z(x))) \quad (1)$$

where $z = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_p * x_p$. is a linear model. The function $1/(1 + e(-z))$ is often called the "logistic"or"sigmoid"function which produces a logistic curve, which is limited to values from 0 and 1 to be interpreted as the P probability. To classify the news articles, our goal is to search for a value of $z(x)$ in such as way that the probability $P(y = 1|x) = h(x)$ is large when x represents news articles belonging to the"1"class and small when x belongs to the"0" class. Multinomial Naive Bayes (MNB) (Kibriya et al. 2004) is fundamentally based on naive bayes approaches and based on bayes' theorem. It has the independence assumptions between the features in order to predict the category of a given category. Naive Bayes method is referred to as the model of posterior probability. This method updates value of prior belief of an event if new information comes. The result of the event is the probability of the /categories occurring given the new data Korb and Nicholson 2010).

$$P(class/attributes) = \frac{P(class) * P(attributes/class)}{P(attributes)}$$

where P(class/attributes) is called posterior probability, P(class) is called the class prior probability, P(attributes/class) is called the likelihood and P(attributes) is called the predictor prior probability. MNB comes under probabilistic classifiers and calculates the probability of each category of news articles using Bayes theorem (McCallum et al. 1998). The news stories category with the highest probability will be output. So, the model calculates the probability that news article d belongs to a class c which is given as follows.

$$P(c/d)P(c) * k = 1ndP(tk/c) \quad (2)$$

In the above equation, $P(tk/c)$ is the called the conditional probability of the term tk of class news article c. We can interpret the probability $P(tk/c)$ as a measurement which is , how much evidence tk contributes for the class determination that c class is the correct class of d news article. $P(c)$is the prior probability of a news article occurring in class c. The term MNB simply explains to us that each $p(fi/c)$ represents the multinomial distribution, rather than some other distribution. K-Nearest Neighbours (KNN) Zahid et al. (2001) very simple and essential algorithms. It is a non-parametric method. More importantly, it does not assume something about the distribution of news article data. It is widely applied in real-life scenarios. The algorithm locates clusters of similar type in the dataset. If any unclassified news story is given, KNN assigns it to a cluster or group by observing what group its nearest neighbors belong to. KNN has an important parameter like k which defines the number of neighbors considered. For example, fitting a KNN model with parameter value $k = 2$, the two closest neighbors are taken to smooth the estimate at a given news article. Support Vector Machine (SVM) (Suykens and Vandewalle 1999) is one of the most important and widely used ML methods. SVM can be applied for many types of problems such as classification and regression (Tong and Koller 2001). In SVM, each news article item is plotted as a point in an n-dimensional space. In the space n is the number of attributes/features the data set and in this case news article. They are plotted the value of each feature being the value of a particular coordinate on the space (Noble 2006). News article classification is performed by finding the hyperplane that differentiates the two news article classes very well and in an optimal way (Joachims 1998). It is also robust to outliers type of news articles. Support vectors are simply the coordinates of individual observation. SVM is a frontier that best segregates the two news article classes using hyper-planes and maximizing the distances between the nearest news article data point. This distance is called Margin. In case a linear hyperplane is not sufficient, SVM has a technique namely kernel trick. Kernel functions take low dimensional input space and transform it into a higher-dimensional space. it converts not separable problems to separable problems, these functions are called kernels. Following is the mathematical representation of SVM. A hyperplane in an N dimensional feature space represented by the following equation as

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b = \sum_{i=1}^{n} x_i w_i + b = 0$$

We divide the above equation by $||\mathbf{w}||$ and obtain the following equation.

$$\frac{\mathbf{x}^T\mathbf{w}}{||\mathbf{w}||} = P_{\mathbf{w}}(\mathbf{x}) = -\frac{b}{||\mathbf{w}||}$$

This equation indicates that the projection of any point $\mathbf{x}$ on plane onto the vector $\mathbf{w}$ is always $-b/||\mathbf{w}||$. More simply, $\mathbf{w}$ represents the normal direction of the hyperplane, and $|b|/||\mathbf{w}||$ is the distance from the origin point to the hyperplane in space. Here the important thing to note is that the hyperplane is not unique from the equation. $cf(\mathbf{x}) = 0$ can represent N number of planes for varying

values of c. The N-dimensional space is divided into two regions by the plane by a mapping function $y = sign(f(\mathbf{x})) \in \{1, -1\}$,

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b = \begin{cases} > 0, & y = sign(f(\mathbf{x})) = 1,\ \mathbf{x} \in P \\ < 0, & y = sign(f(\mathbf{x})) = -1,\ \mathbf{x} \in N \end{cases}$$

Here any point in hyperspace $\mathbf{x} \in P$ on the positive side of the plane is mapped to value 1, while any point in hyperspace $\mathbf{x} \in N$ on the negative side is mapped to -1. A point $\mathbf{x}$ of unknown class will be classified to P if $f(\mathbf{x}) > 0$, or N if $f(\mathbf{x}) < 0$. For a decision type of hyper-plane $\mathbf{x}^T\mathbf{w} + b = 0$ to separate the two classes of points representing the news articles $P = \{(\mathbf{x}_i, 1)\}$ and $N = \{(\mathbf{x}_i, -1)\}$, it has to satisfy

$$y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq 0$$

for both $\mathbf{x}_i \in P$ and $\mathbf{x}_i \in N$. Among all such hyper planes satisfying this condition, there is need to find the optimal one H_0 that separates the two classes with the maximal margin. Hence, the optimal plane must be in the middle of the two different classes of classification problem. This makes the distance from the dividing plane to the closest point on either side of news articles or data points equal. Two additional separate hyper planes are defined and plotted to set the boundary to the maximum of the middle point with following conditions.

$$\mathbf{x}_i^T\mathbf{w} + b = y_i$$

And the following holds for all support vectors:

$$b = y_i - \mathbf{x}_i^T\mathbf{w} = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_i^T\mathbf{x}_j)\,.$$

Now the problem of finding the optimal decision plane in terms of $\mathbf{w}$ and b can be formulated as:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}||\mathbf{w}||^2 \quad \text{(objective function)} \\ &\text{subject to } y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq 1,\ \text{or}\ 1 - y_i(\mathbf{x}_i^T\mathbf{w} + b) \leq 0,\ \ (i = 1, \cdots, m) \end{aligned} \tag{3}$$

The solution of the above equation gives the maximum margin of the points so that the news articles can be determined to be categorized in one of the categories.

4 Research methodology

4.1 Data collection

The experiment is designed with the following phases. First the experiment has collected the news articles from the three sources namely The Hindu, NDTV and Indian Express website. There are 1800 articles. It is to be noted that majorly, previous research studies have mostly used the headlines/titles of the news articles for the classification. There was no big dataset of news articles, along with the categories/classes , tags/keywords and content of the news articles/body. It has been a disadvantage of using only the article's headline for classification . The reason is that the headlines are not always honest in their interpretation for the sake of catching attention. Hence, the experiment tried to remove that or reduce that loop by collecting the large amount of data of new articles tagged with their categories and body of the news articles and headlines for our categorization problem. Three diverse sources have made the data more practical and reduces the chances of overfitting. Collected articles belongs to the following categories:-

1 Business and Economy
2 Education and Career
3 Entertainment
4 Food and Health
5 International
6 Politics and Governance
7 Science and Technology
8 Sports

These 8 categories cover the majority of the articles that are read in our daily life and the articles that do not fall into any of these categories were removed. The data set can be accessed on the following URL:https://github.com/sachinblessed95/Current-Affairs-News-Categorisation-Data-Set.

4.2 Data processing

Obtained newspaper articles are rich in the content. But for better accuracy, more concise information and feature rich content is required. To get the relevant and feature rich information, data cleaning is required. Data cleaning is the first step where all the text is converted into lowercase for simplicity and uniformity followed by tokenization by breaking the sentences into words and removing the stop-words. Alphanumeric characters, numbers and punctuation are also removed followed by the redundancy of words (Chava et al. 2018). However, there are still some words that are left out after these steps of data processing which are redundant for the learning model for differentiating the news articles. For example, consider the sentence"He is playing Football"and"He plays Football", Both the words "playing"and"plays"have similar meanings and increasing the complexity of the model may even lead to decreased accuracy. Hence, lemmatization needs to be incorporated which means converting these two words into a single meaningful form Gui et al. (2009). After processing the data, it contains only the essential words required for classification. Words are mapped to numbers which are taken as features for the mathematical models. There can be many features but there is a need to have important

features selected through selection and generation of relevant features to improve the accuracy.

4.3 Feature engineering

Articles processed require the embedding approaches for prepossessing the text of the news articles. There are many widely used embedding techniques but a suitable one is the Count Vector and frequency'inverse document frequency(TF-IDF) encoding which will work as a feature preparation phase (Wang et al. 2010). Count Vector embedding helps in forming a vocabulary from the given corpus or the news articles by calculating the frequency of each word and returns a vector with these frequencies assigned to the respective words of the news articles. On the other hand, TF-IDF vectorizer embedding technique calculates the TF-IDF frequency of each word in each document. These frequencies generate sparse matrices which are further used as the feature sets for the model. The mathematical formula for Tf-IDF frequency is given as $Tf - Idf(t) = Tf(t) * Idf(t)$.

Here $Tf(t)$ represents the number of times the words t comes in a targeted document) divided by the total number of words in the targeted document. $Idf(t)$ is log_e of the total number of documents divided by the number of documents with the word t in them.

4.4 Proposed model

After preprocessing and feature engineering phase, a proposed model with the integration of the feature enhancement and categorisation module has been presented as follows. The proposed model is based on the penetration of feature enhancement and machine learning model MNB. This proposed model takes three contents of the news articles to learn and classify the news articles which includes the title of the news article, news articles body/content and tags of the news articles. Following is the proposed model 1.

where $A^{'}$ represents the article that is free from stop words and punctuation. Then Lemmatization using StanfordNLP python library for the Lemmatization process Al Omran et al. (2017) is conducted. Then the vectorization of the articles is required to assign weights to the words in the article which has been carried out using two different commonly used approaches-Count Vectorization and TF-IDF Vectorization. Finally, the encoded articles are given as input to the classifiers. The proposed model in pictorial form is depicted in the Fig. 1. In the proposed model, procedural, news articles are obtained from three sources. In the proposed model, procedurally, news articles are obtained from three sources. They are combined together to make a centralised data set. In the next phase, the news articles are moved to the data cleaning phase. After this phase, the data of news articles is processed for removal of insignificant words, removal of redundant words. This makes the processing of the articles efficient and these types of words do not contribute to disturbing the articles from other categories. After this phase, the feature engineering process starts with conversion of the words into the feature in numerical form through vectorization. There are two feature generation techniques for the news articles techniques. The proposed model uses two vectersiation techniques namely count vector and TF-IDF vectorization. They both are integrated for the feature generation and selection integrated with ML models. Data set with property feature engineering is divided into two sets of 80 percent and 20 percent for the training and testing of the proposed model. Now the proposed model TF-IDF+MNB is run and results are obtained. To ensure the robustness of the model, it uses the cross validation techniques to ensure that results don't suffer from overfitting. The model is evaluated on several criteria of classification problems such as accuracy, confusion matrix, f-score, recall, etc. Model is also compared with other machine learning models such as SVM, RF, LR and KNN on several criteria.

Algorithm 1: Proposed model(A)

Result: Category of the article
Let $A^{'}$ be a new string array;
while $token \in A$ **do**
  if $token \notin (stopwords \cup punctuation)$ **then**
    $A^{'}$ $A^{'} \cup$ token;
  end
end
A_{lemma} lemmatize($A^{'}$);
A_{vec} vectorize(A_{lemma});
category classify(A_{vec});
return category;

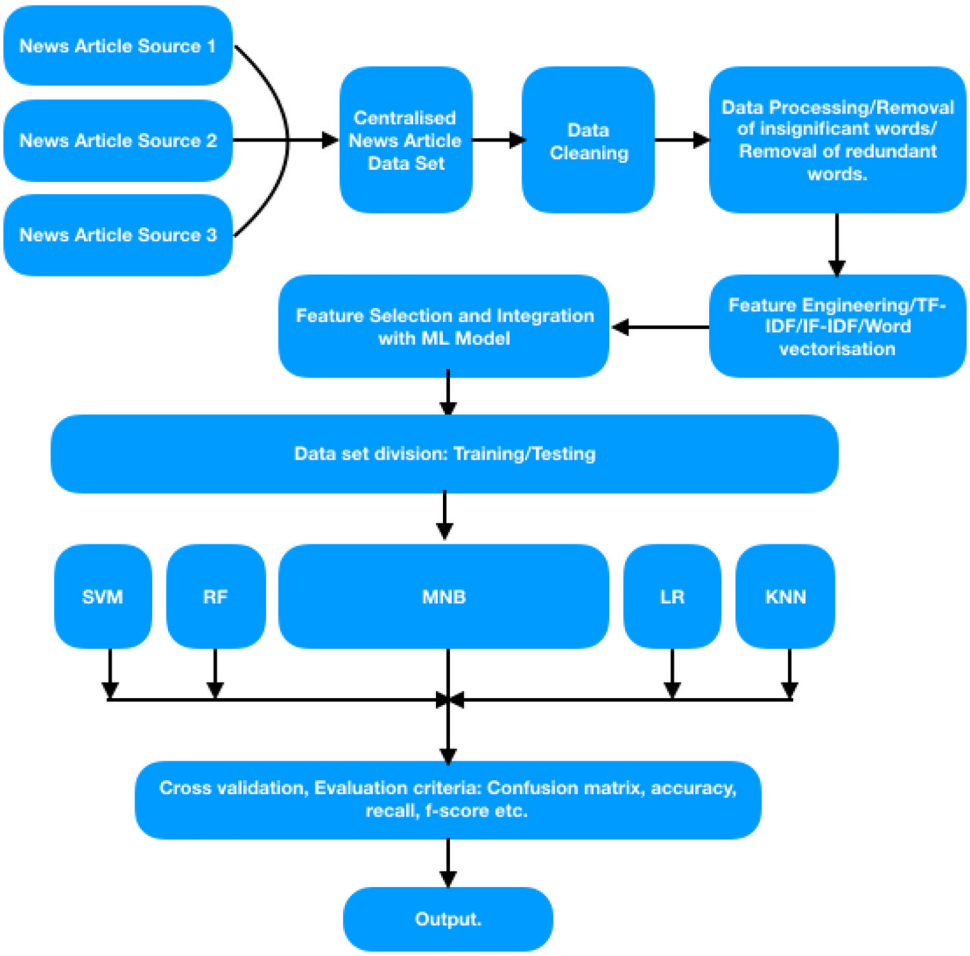

Fig. 1 Proposed model based on TF-IDF integrated MNB

4.5 Evaluation criteria

News article categorisation falls under the classification task and there are many standard evaluation matrices for the classification tasks such as confusion matrix, accuracy, precision and recall and f-measure. The proposed model solves multiclass classification problems of assigning news articles to pre-existing 8 categories of articles. The first performance metric is the confusion matrix that is described as follows. A confusion matrix provides a summary of classification results. Confusion matrix summarizes the number of correct and incorrect classifications with their count values by each class. Evaluation metric accuracy is considered as the most intuitive measure of performance (Sokolova and Lapalme 2009). It is simply a ratio of correctly predicted observation to total observations and defined as the $accuracy = (TP + TN)/(TP + TN + FP + FN)$. Other metrics are precision and recall that can be defined mathematically as follows. Let us define an experiment from P positive instances and N negative instances for some conditions. Then the precision and recall will be formulated as: $Precision(p) = TP/(TP + FP)$ and $Recall(r) = TP/(TP + FN)$. The last evaluation metric is F-measure, which is the harmonic mean of measurement precision and recall. F1 score reaches its best

Fig. 2 Experimental data with category, key words and article

	Category	Keywords	Article
0	2	\nBring on the laughs	dash stand comedy lot improvisation ton audien...
1	2	\nTamannaah Bhatia to play leading lady in Sun...	Actor Tamannaah Bhatia excited sign acclaim di...
2	2	\nSensation Rise 2018: Rise for the city and f...	say never forget first tricity music festival ...
3	2	\nThe Beatles in India\n	year 1968 Beatles spend time away limelight le...
4	2	\nNetflix renews Sacred Games for a second sea...	Sacred game Netflix first original indian show...

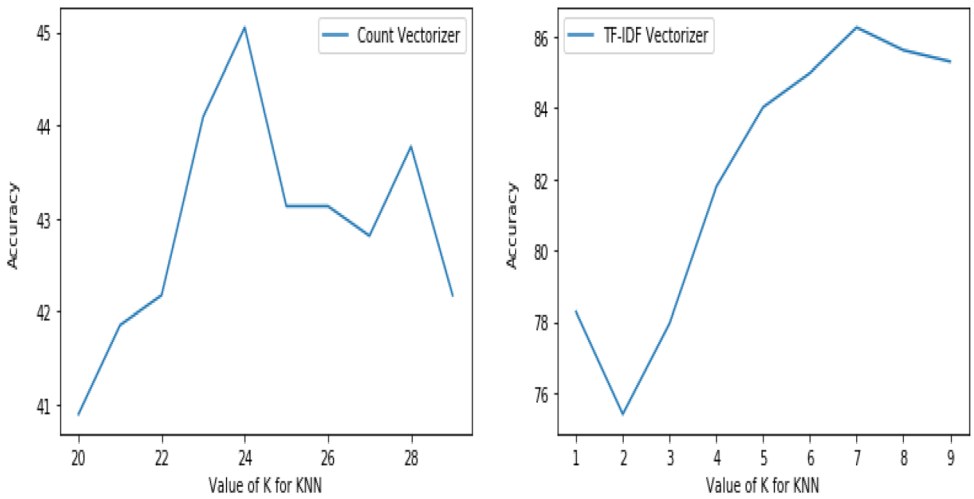

Fig. 3 Plot of accuracy versus the number of nearest neighbors using **a** Countvectorizer. **b** TF-IDF Vectorizer.

Table 1 Accuracy of all the classifiers using different vectorizers for categorization.

ML Models	Count Vectorizer	TF-IDF Vectorizer
Multinomial Naive Bayes(MNB)	**87.22**	**85.62**
Random Forest(RF)	79.55	81.15
Support Vector Machine(SVM)	82.42	85.62
K Nearest Neighbor(KNN)	45.04 (k=24)	86.26 (k=7)
Logistic Regression(LR)	84.02	85.26

value at 1 which is the perfect precision and recall and worst at 0. $F - measure = 2 * p * r/(p + r)$.

4.6 Experiment setup

The experiment has been conducted on Intel Core(TM) i5-4690 CPU@3.20 GHz with RAM 4 GB, 64 bit OS, x64 based processor. The work has been done using a Python programming language.

5 Experimental results and analysis

The articles were split into training and testing data in 80 percent and 20 percent respectively. The training dataset contained around 160 articles of each category and 40 articles of each category in the testing dataset. The sample dataset looks as given in Fig. 2. It contains the fields such as **Category**, **Keyword/Tags** and **Article**. The Category field contains a category id for each of the categories of the news articles, the keyword column is the about the headline of the article. The article column contains the article body which model would be processed to classify the news articles together with other features.

The results have been compiled in tables and have been put in pictorial representations through plots. Several models apart from the proposed model have been evaluated on evaluation criteria such as confusion matrix, accuracy,

Fig. 4 Bar plot comparing accuracy of different algorithms

recall, f-score etc. The good results have been highlighted and compared with our proposed model on MNB. Let us take each ML model and compare it with the proposed model integrated with TF-IDF and MNB. KNN faces difficulty as it needs to decide and choose the correct value of the number of neighbours for good accuracy in the article categorization. This is achieved by testing the KNN algorithms for certain values of the number of neighbours. This gives the trend of how the change in the number of neighbours is changing the accuracy of the KNN model. The below Fig. 3 shows the trend of accuracy versus the number of neighbours, parameters of the model, for different vectorizers in the news article classification. It is observed that as the value of K increases, the value of

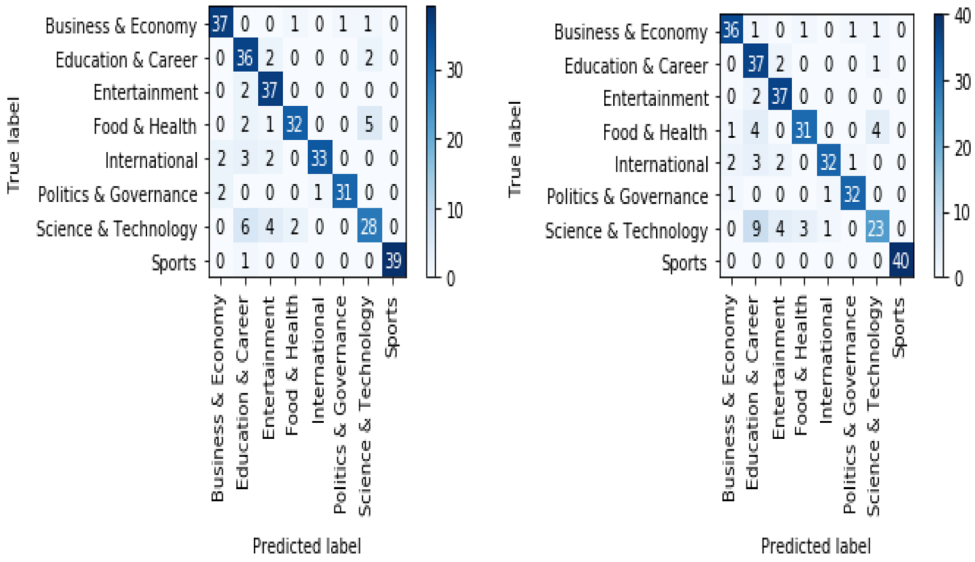

Fig. 5 Confusion matrix for Multinomial Naive Bayes **a** Using Count Vectorizer. **b** Using TF-IDF Vectorizer.

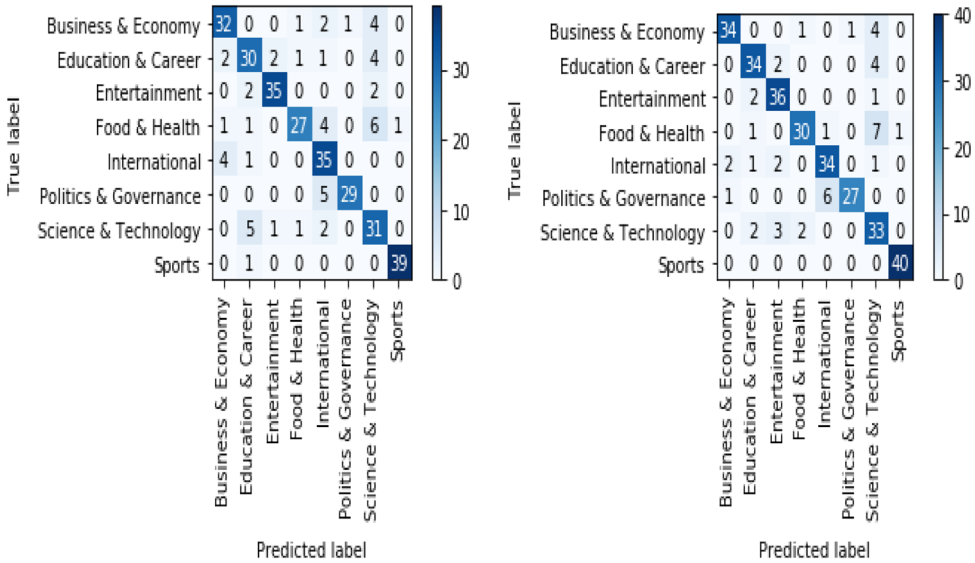

Fig. 6 Confusion matrix for Support Vector Machines **a** Using Count Vectorizer. **b** Using TF-IDF Vectorizer.

accuracy increases and on values of K equals to 4 and 8, KNN model shows the least accuracy. In the second Fig. 3 , with an increase in value of K, the accuracy of the KNN model increases.

The model got iterated in the range of 1 to 40 for the parameter value. The number of nearest neighbours of the model was quite costly in terms of efficiency. Hence the range of K has been restricted to 40 only and obtained good results with TF-IDF vectorizer. It needs a lot of storage to store all the training data. The other three algorithms were computationally less costly compared to KNN (Table1). Proposed model of TF-IDF + MNB produced 87.22 percent with Count Vectorizer as shown in Table1 and it was also computationally cheap as compared to other ML Models. Additionally, It scales relatively well to high dimensional data due to its properties. It is also depicted that the TF-IDF + KNN model produced good results with the value of K being less than 10 while other models of MNB produced good results with the value of K in range from 20 to 30. Specifically, KNN model with Count Vectorizer feature selection technique produced accuracy of 45.04 with the value of K equal to 24 while the KNN with TF-IDF produced the article classification accuracy of 86.26 with value of K equivalent to 7. Coming to the MNB based proposed model, Count Vecteriser + MNB model produced the best results of the accuracy 87.22 and TF-IDF + MNB produced 85.62. This is the only proposed model which has performed well for both feature selection techniques and closer to this accuracy comes the very simple model of LR. Count Vectoeriser + LR classified the news articles with accuracy of 84.02 and TF-IDF + LR produced the articles with 85.26 accuracy which is close to the

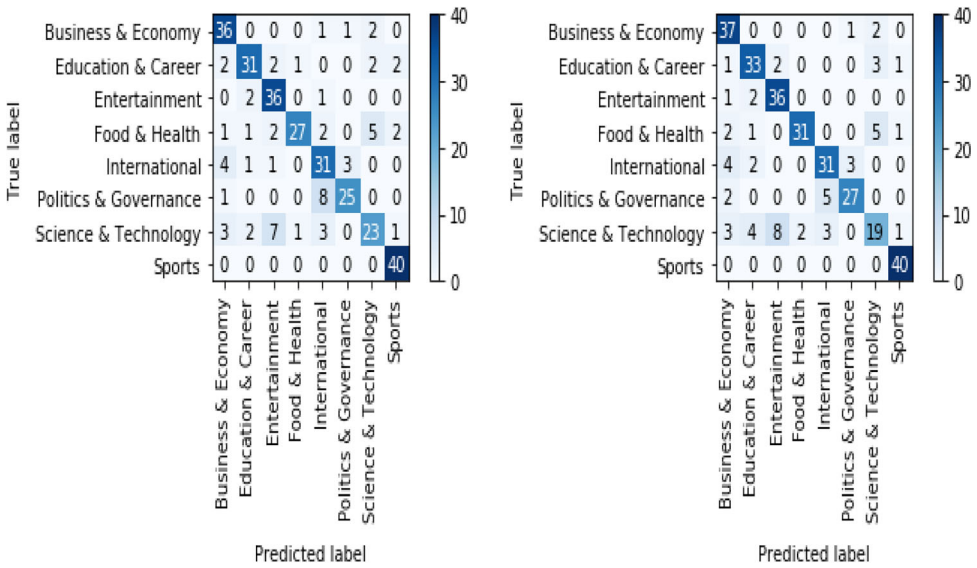

Fig. 7 Confusion matrix for Random Forest Neighbours **a** Using Count Vectorizer. **b** Using TF-IDF Vectorizer.

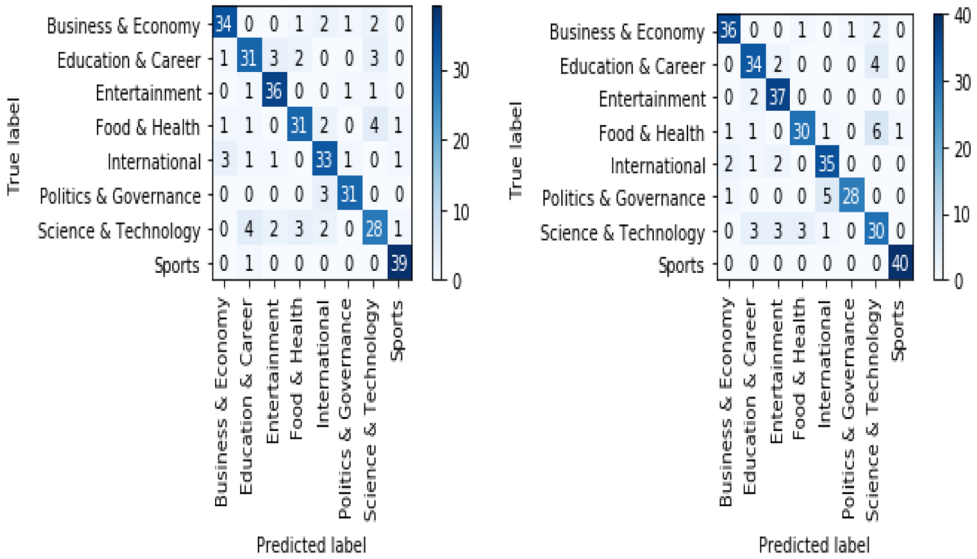

Fig. 8 Confusion matrix for Logistic Regression **a** Using Count Vectorizer. **b** Using TF-IDF Vectorizer.

proposed model. The count vectoriser + SVM and TF-IDF + SVM models produce the accuracy of 82.42 and 85.62. The parameter of the SVM used are $C = 1.0$, $kernel =' rbf'$, $degree = 3$, $gamma = 0.0$, $coef = 0.0$ The SVM produced very low accuracy compared to other ML models. It scales relatively well to high dimensional data. The more samples in the training set the more complex and slow the classification process. Table 1 lists the accuracy of the all ML models with counter vectorizer and TF-IDF. RF based integrated model with count vector and TF-IDF with 79.55 and 81.15 respectively. The results of RF have been obtained with parameter $estimator = 100$, $jobs = -1$.

Figure 4 shows the plot of the accuracies of all ML models used for the classification of news articles. The maximum accuracy is obtained with MNB followed by LR, KNN, SVM, and RF. Only MNB gives better accuracy with TF rather than TF-IDF processing of articles. The rest of the models perform better in processing with TF-IDF vectorisation. KNN accuracy is significantly low. The confusion matrix of the classification gives us a better visualization of how precisely the proposed machine learning models such as TF-IDF+MNB are able to classify each of the categories. The diagonal values of the confusion matrix represent the true positives. Figure 5 displays the confusion matrix of both count vectoriser+MNB and TF-IDF+MNB. It shows that MNB classifies articles with sports categories very accurately. TF-IDF+MNB faces difficulties in classifying the science and technology articles and performing the least accurately and confuses them with education and career and with food articles also. Figure 6 displays results of confusion matrix with SVM model. SVM faces problems in classifying articles on food

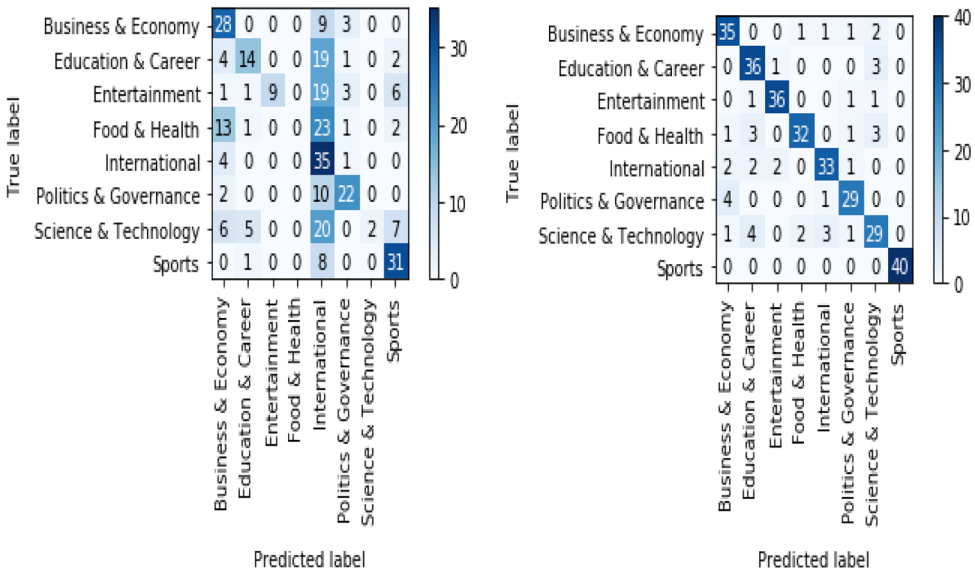

Fig. 9 Confusion matrix for KNN **a** Using Count Vectorizer. **b** Using TF-IDF Vectorizer.

and health and confuses them with science and technology and international news. Although international categories can overlap with health. SVM also performs better in the sports category. Figure 7 displays results with RF models and gives better accuracy with sports articles and less with science and technology. Similar results are obtained in the TF-IDF preprocessing phase. Figure 8 displays results of LR using a count vector and TF-IDF vectoriser. Here also similar results are obtained with highest classification accuracy with sports category and lowest in science and technology. Figure 9 displays results of KNN models with the best accuracy in the sports category but it gives a worse performance in science and technology articles and confuses them with business education. It is observed that the precision of the sports category for most of the algorithms is quite good which shows that the algorithms are good at recognising the articles belonging to the sports category. The confusion matrix of the Support Vector Machine for both TF and TF-IDF vectorizer is very disturbed which shows that the algorithm is not quite good at predicting the classes correctly which is also depicted by its accuracies.

The confusion matrix of the classification gives us a better visualization of how precisely the algorithms are able to classify each of the categories. The diagonal values of the matrix represent the true positives i.e., the classification label and the true label are the same. Figure 5 displays the confusion matrix of both CV and TF-IDF preprocessing articles methods and gives a clear idea of each category article classification and assessment. It shows that MNB classifies articles with sports category very accurately and science and technology articles least accurately and confuses them with education and career and with food articles. Figure 6 displays results of confusion matrix with SVM model. SVM faces problems in classifying articles on food and health and confuses them with science and technology and international news. Although international categories can overlap with health. SVM also performs better in the sports category. Figure 7 displays results with RF models and gives better accuracy with sports articles and less with science and technology. Similar results are obtained in the TF-IDF preprocessing phase. Figure 8 displays results of LR using a count vector and TF-IDF vectoriser. Here also similar results are obtained with highest classification accuracy with sports category and lowest in science and technology. Figure 9 displays results of KNN models with the best accuracy in the sports category but it gives a worse performance in science and technology articles and confuses them with business education. It is observed that the precision of the sports category for most of the algorithms is quite good which shows that the algorithms are good at recognising the articles belonging to the sports category. The confusion matrix of the SVM for both TF and TF-IDF vectorizer is very disturbed which shows that the algorithm is not quite good at predicting the classes correctly which is also depicted by its accuracies.

The Table 2 also lists the Precision, Recall and F-Measure. The precision metric shows how correctly the algorithms can classify into each category and recall shows out of all his categories how well can the algorithm recall a particular category. According to the results, the proposed model with TF-IDF+MNB, in the sports category, recall, f-measure and precision is 1.0 which is highest. Other machine learning models don't produces that level of accuracy and recall and f-score as depicted in the Table 2. The least results are obtained in health and food and science and technology category due to their similar nature of content which confuses the models to distinguish them.

Table 2 Precision, Recall and F-Measure of category classification of all the algorithms

Category	ML Models		Precision	Recall	F-Measure
Business and Economy	MNB	TF	0.9250	0.9024	0.9135
		TF-IDF	0.9000	0.9000	0.9000
	LR	TF	0.8500	0.8717	0.8607
		TF-IDF	0.9000	0.9000	0.9000
	SVM	TF	0.8000	0.8205	0.8101
		TF-IDF	0.8500	0.9189	0.8831
	RF	TF	0.9000	0.7659	0.8275
		TF-IDF	0.9250	0.7400	0.8222
	KNN (k=7)	TF-IDF	0.8750	0.8139	0.8433
Education and Career	MNB	TF	0.9000	0.7200	0.7999
		TF-IDF	0.9250	0.6607	0.7708
	LR	TF	0.7750	0.7948	0.7848
		TF-IDF	0.8500	0.8292	0.8395
	SVM	TF	0.7500	0.7500	0.7500
		TF-IDF	0.8500	0.8500	0.8500
	RF	TF	0.7750	0.8378	0.8051
		TF-IDF	0.8250	0.7857	0.8048
	KNN (k=7)	TF-IDF	0.9000	0.7826	0.8372
Entertainment	MNB	TF	0.9487	0.8043	0.8705
		TF-IDF	0.9487	0.8222	0.8809
	LR	TF	0.9230	0.8571	0.8889
		TF-IDF	0.9487	0.8409	0.8915
	SVM	TF	0.8974	0.9210	0.9090
		TF-IDF	0.9230	0.8372	0.8780
	RF	TF	0.9230	0.7500	0.8275
		TF-IDF	0.9230	0.7826	0.8048
	KNN (k=7)	TF-IDF	0.9230	0.9230	0.9230
Food and Health	MNB	TF	0.8000	0.9143	0.8533
		TF-IDF	0.7750	0.8857	0.8266
	LR	TF	0.7750	0.8378	0.8052
		TF-IDF	0.7500	0.8824	0.8108
	SVM	TF	0.6750	0.9000	0.7714
		TF-IDF	0.7500	0.9090	0.8219
	RF	TF	0.6750	0.7500	0.8275
		TF-IDF	0.7750	0.9393	0.8493
	KNN (k=7)	TF-IDF	0.8000	0.9142	0.8533
International	MNB	TF	0.8250	0.9705	0.89189
		TF-IDF	0.8000	0.9411	0.8648
	LR	TF	0.8250	0.7857	0.8049
		TF-IDF	0.8750	0.8333	0.8536
	SVM	TF	0.8750	0.7142	0.7865
		TF-IDF	0.8750	0.8974	0.8860
	RF	TF	0.7750	0.6739	0.7209
		TF-IDF	0.7750	0.7948	0.7848
	KNN (k=7)	TF-IDF	0.8250	0.8684	0.8461

Table 2 continued

Category	ML Models		Precision	Recall	F-Measure
PoliticsandGovernance	MNB	TF	0.9117	0.9688	0.9394
		TF-IDF	0.9411	0.9411	0.9411
	LR	TF	0.9117	0.9117	0.9117
		TF-IDF	0.8235	0.9655	0.8888
	SVM	TF	0.8529	0.9666	0.9062
		TF-IDF	0.7941	0.9642	0.8709
	RF	TF	0.7352	0.8620	0.7936
		TF-IDF	0.7941	0.8709	0.8307
	KNN (k=7)	TF-IDF	0.8529	0.8529	0.8529
Science and technology	MNB	TF	0.7000	0.7777	0.7368
		TF-IDF	0.5750	0.7931	0.6667
	LR	TF	0.7000	0.7368	0.7179
		TF-IDF	0.7500	0.7142	0.7317
	SVM	TF	0.7750	0.6595	0.7126
		TF-IDF	0.8250	0.6600	0.7333
	RF	TF	0.5750	0.71875	0.6388
		TF-IDF	0.4750	0.6551	0.5507
	KNN (k=7)	TF-IDF	0.7250	0.7631	0.7435
Sports	MNB	TF	0.9750	**1.0000**	0.9873
		TF-IDF	**1.0000**	**1.0000**	**1.0000**
	LR	TF	0.9750	0.9286	0.9512
		TF-IDF	**1.0000**	0.9756	0.9877
	SVM	TF	0.9750	0.9750	0.9750
		TF-IDF	**1.0000**	0.9756	0.9876
	RF	TF	**1.0000**	0.8888	0.9411
		TF-IDF	0.1000	0.9302	0.9638
	KNN (k=7)	TF-IDF	**1.0000**	**1.0000**	**1.0000**

Table 3 Time taken during the training period by the classifiers using different vectorizers for categorization.

ML Models	Count Vectorizer	TF-IDF Vectorizer
Multinomial Naive Bayes(MNB)	**1.82s**	**3.86s**
Random Forest(RF)	6.88s	6.74s
Support Vector Machine(SVM)	130s	147s
K Nearest Neighbor(KNN)	28.9s	29s
Logistic Regression(LR)	37.3s	16.2s

Table 3 gives a running time analysis of all machine learning models with both Counts vectoriser and TF-IDF vectoriser. The time taken is maximum in SVM and least in proposed model MNB. The MNB performs the best in both the cases of accuracy and time efficiency in the classification of news articles.

6 Conclusion and future works

Digital technologies based information sharing platforms such as news websites and social media platforms such as Facebook, Twitter, Instagram, What's app etc have flooded the information space. Information or text classification has an important role to play so that important and vital information can be classified based on the already

predefined categories. Nowadays the process of classification and segregation of textual information has become challenging due to the flow of diverse, vast information. Automated and intelligent tools which can classify the information and text accurately and efficiently are needed to reduce human efforts, time and increase productivity. The research work contributes in three ways. It proposed a model which takes news titles and news stories for classification of the news articles which is an improvement over previous work. It proposed a machine learning-based model which is very efficient and effective in news categorisation. Secondly, it performs classification and comparison on newly developed data sets of news stories. This paper presents an intelligent, efficient and robust intelligent machine learning model. The proposed IDF integrated MNB model achieves classification accuracy of 87.22 per cent and results have been compared with state of art machine learning algorithms such as LR, KNN, SVM and RF. The proposed model Count Vectorizer integrated MNB and IDF integrated MNB is also better in terms of efficiency with running time of 1.82s and 3.86s respectively. The work can also be improved by introducing more selection methods and deep neural network architects in future works.

Acknowledgements Acknowledgements, if any, should follow the conclusions, and be placed above any Appendices or the references.

Funding No funding received for this research work.

Declaration

Conflict of Interest There is no conflict of interest in this work.

References

Ahmed M, Mahmood AN, Islam MR (2016) A survey of anomaly detection techniques in financial domain. Future Gener Comput Syst 55:278–288

Al Omran FNA, Treude C (2017) Choosing an nlp library for analyzing software documentation: a systematic literature review and a series of experiments. In: 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), IEEE, pp 187–197

Alhothali A, Hoey J (2015) Good news or bad news: Using affect control theory to analyze readers'reaction towards news articles. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1548–1558

Ashuri T (2016) When online news was new: Online technology use and constitution of structures in journalism. Journal Stud 17(3):301–318

Bail CA (2017) Taming big data: Using app technology to study organizational behavior on social media. Sociol Methods Res 46(2):189–217

Boumans JW, Trilling D (2016) Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. Digit Journal 4(1):8–23

Canito J, Ramos P, Moro S, Rita P (2018) Unfolding the relations between companies and technologies under the big data umbrella. Comput Ind 99:1–8

Carstens L, Toni F (2017) Using argumentation to improve classification in natural language problems. ACM Trans Internet Technol (TOIT) 17(3):30

Castillo C, El-Haddad M, Pfeffer J, Stempeck M (2014) Characterizing the life cycle of online news stories using social media reactions. In: Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing, ACM, pp 211–223

Chava RVSP, Dhar S, Gaur Y, Rambhakta P, Shetty S (2018) Big data text summarization-hurricane irma

Conroy NJ, Rubin VL, Chen Y (2015) Automatic deception detection: Methods for finding fake news. Proc Assoc Inf Sci Technol 52(1):1–4

Daud A, Khan W, Che D (2017) Urdu language processing: a survey. Artif Intell Rev 47(3):279–311

Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. Proc Nat Acad Sci 113(3):554–559

Fong S, Zhuang Y, Li J, Khoury R (2013) Sentiment analysis of online news using mallet. In: Proceedings of the 2013 International Symposium on Computational and Business Intelligence, IEEE Computer Society, Washington, DC, USA, ISCBI '13, pp 301–304, 10.1109/ISCBI.2013.67

Gui Y, Gao Z, Li R, Yang X (2012) Hierarchical text classification for news articles based-on named entities. In: International Conference on Advanced Data Mining and Applications, Springer, pp 318–329

Habash N, Rambow O, Roth R (2009) Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In: Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt, vol 41, p 62

Hakim AA, Erwin A, Eng KI, Galinium M, Muliady W (2014) Automated document classification for news article in bahasa indonesia based on term frequency inverse document frequency (tf-idf) approach. In: 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), IEEE, pp 1–4

Hogenboom F, Frasincar F, Kaymak U, De Jong F, Caron E (2016) A survey of event extraction methods from text for decision support systems. Decis Support Syst 85:12–22

Hsu WL, Lang SD (1999) Classification algorithms for netnews articles. In: Proceedings of the Eighth International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '99, pp 114–121, 10.1145/319950.319965

Hu Y, Ye X, Shaw SL (2017) Extracting and analyzing semantic relatedness between cities using news articles. Int J Geogr Inf Sci 31(12):2427–2451

Joachims T (1998) Making large-scale svm learning practical. Technical report, Tech rep

Kanan T, Fox EA (2016) Automated Arabic text classification with p-s temmer, machine learning, and a tailored news article taxonomy. J Assoc Inf Sci Technol 67(11):2667–2683

Kibriya AM, Frank E, Pfahringer B, Holmes G (2004) Multinomial naive bayes for text categorization revisited. In: Australasian Joint Conference on Artificial Intelligence, Springer, pp 488–499

Kiranyaz S, Ince T, Gabbouj M (2014) Multidimensional particle swarm optimization for machine learning and pattern recognition. Springer, Berlin

Korb KB, Nicholson AE (2010) Bayesian artificial intelligence. CRC Press, Florida

Kumar S, Kalia A, Sharma A (2018a) Predictive analysis of alertness related features for driver drowsiness detection. Adv Intell Syst Comput 736:368–377

Kumar S, Pal SK, Singh R (2018b) A novel method based on extreme learning machine to predict heating and cooling load through design and structural attributes. Energ Build 176:275–286

Kumar S, Singh R, Pal SK (2018c) A conceptual architectural design for intelligent health information system: Case study on india. Quality, IT and Business Operations: Springer Proceedings in Business and Economics, vol 1. Springer, Singapore, pp 1–15

Kumar S, Saibal KP, Singh R (2019) A novel hybrid model based on particle swarm optimisation and extreme learning machine for short-term temperature prediction using ambient sensors. Sustain Cities Soc

Kumar S, Singh J, Singh O (2020) Ensemble-based extreme learning machine model for occupancy detection with ambient attributes. Int J Syst Assur Eng Manag

Kurt I, Ture M, Kurum AT (2008) Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl 34(1):366–374

Li J, Fong S, Zhuang Y, Khoury R (2016) Hierarchical classification in text mining for sentiment analysis of online news. Soft Comput 20(9):3411–3420

Lin WY, Hu YH, Tsai CF (2011) Machine learning in financial crisis prediction: a survey. IEEE Trans Syst, Man, Cybern, Part C (Appl Rev) 42(4):421–436

Lykourentzou I, Giannoukos I, Nikolopoulos V, Mpardis G, Loumos V (2009) Dropout prediction in e-learning courses through the combination of machine learning techniques. Comput Educ 53(3):950–965

Marconi F (2020) Newsmakers: artificial intelligence and the future of journalism. Columbia University Press, Columbia

McCallum A, Nigam K, et al. (1998) A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization, Citeseer, vol 752, pp 41–48

Medagoda N (2016) Sentiment analysis on morphologically rich languages: An artificial neural network (ann) approach. In: Artificial Neural Network Modelling, Springer, pp 377–393

Miller K, Oswalt A (2017) Fake news headline classification using neural networks with attention. Tech. Rep., California State University

Mukwazvure A, Supreethi K (2015) A hybrid approach to sentiment analysis of news comments. 2015 4th International Conference on Reliability. Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), IEEE, pp 1–6

Noble WS (2006) What is a support vector machine? Nat Biotechnol 24(12):1565–1567

Onan A, Korukoğlu S, Bulut H (2016) Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl 57:232–247

Paragios N, Chen Y, Faugeras OD (2006) Handbook of mathematical models in computer vision. Springer, Berlin

Pröllochs N, Feuerriegel S, Neumann D (2016) Negation scope detection in sentiment analysis: Decision support for news-driven trading. Decis Support Syst 88:67–75

Ramík DM, Sabourin C, Moreno R, Madani K (2014) A machine learning based intelligent vision system for autonomous object detection and recognition. Appl Intell 40(2):358–375

Rasouli K, Hsieh WW, Cannon AJ (2012) Daily streamflow forecasting by machine learning methods with weather and climate inputs. J Hydrol 414:284–293

Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ (2016) Deep learning for health informatics. IEEE J Biomed Health Inf 21(1):4–21

Salminen J, Yoganathan V, Corporan J, Jansen BJ, Jung SG (2019) Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type. J Bus Res 101:203–217

Sen PC, Hajra M, Ghosh M (2020) Supervised classification algorithms in machine learning: A survey and review. In: Emerging technology in modelling and graphics, Springer, pp 99–111

Singh G, Kumar B, Gaur L, Tyagi A (2019) Comparison between multinomial and bernoulli naïve bayes for text classification. 2019 International Conference on Automation. Computational and Technology Management (ICACTM), IEEE, pp 593–596

Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437

Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300

Thomson T, Angus D, Dootson P, Hurcombe E, Smith A (2020) Visual mis/disinformation in journalism and public communications: current verification practices, challenges, and future opportunities. Journalism Practice pp 1–25

Tong S, Koller D (2001) Support vector machine active learning with applications to text classification. J Mach Learn Res 2(Nov):45–66

Ur-Rahman N, Harding JA (2012) Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Syst Appl 39(5):4729–4739

Valdivia A, Luzón MV, Herrera F (2017) Sentiment analysis in tripadvisor. IEEE Intell Syst 32(4):72–77

Van Veldhoven Z, Vanthienen J (2021) Digital transformation as an interaction-driven perspective between business, society, and technology. Electron Mark pp 1–16

Wang N, Wang P, Zhang B (2010) An improved tf-idf weights function based on information theory. In: 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering, IEEE, vol 3, pp 439–441

Witten IH, Frank E, Hall MA, Pal C, DATA M (2005) Practical machine learning tools and techniques. In: DATA MINING, vol 2, p 4

Zahid N, Abouelala O, Limouri M, Essaid A (2001) Fuzzy clustering based on k-nearest-neighbours rule. Fuzzy Sets Syst 120(2):239–247