Intelligent Environments 2021
E. Bashir and M. Luštrek (Eds.)

doi:10.3233/AISE210094

Self Learning of News Category Using AI Techniques

Zara HAYAT [a], Aqsa RAHIM [a], Sajid BASHIR [b] and Muddasar NAEEM [c]
[a] *National University of Science And Technology*
[b] *National University of Technology*
[c] *ICAR-CNR*

Abstract. Numerous e-news channels publish the daily happenings in the world from different sources. These huge amounts of news articles have lamentably conceived the information overload issue among the users. Hence text mining, which aims in extracting previously unknown information from unstructured text, has been widely used by several researchers to segregate full news articles however, the news headlines categorization is still specifically limited. Therefore, considering this limitation, the current research aims to propose a framework that will self-learn and automatically classify any given news headline into its corresponding news category using artificial intelligence methods i.e. text mining and machine learning algorithms. The proposed framework consists of three stages: Exploratory Data Analysis, Text Pre-processing, and Text Classification. For exploratory data analysis, the top 10 most frequent balanced news categories are chosen so that further processing of data can be done on a more balanced version of the dataset. After exploring the data, text pre-processing techniques are applied to make the data transformed, normalized, and structured. Finally, text classification is carried out with two approaches: unsupervised classification using Mean Shift and K-means algorithms and supervised classification using Logistic Regression with Bag of Words and TF-IDF algorithm. To depict the working of the proposed framework, a case study is presented on a news headlines dataset which accurately performed news headlines classification.

Keywords. Self-learning, Artificial Intelligence, Text Mining, Machine Learning, Exploratory Data Analysis, Text Pre-processing, TF-IDF, Bag of Words, Unsupervised Clustering, Supervised Classification, Logistic Regression

1. Introduction

With the escalating popularity of using the internet and smartphones, a substantial number of internet users have been amplified remarkably. By taking advantage of an enormous cluster of online users and attaining more viewers and readers, a lot of newspaper companies and news providers have started publishing and updating the news editorial articles on their websites and weblogs. Many users regularly get news from these sources [24]. These news sites include news articles aligned with a corresponding news headline. The headlines of the news exemplify the central concept of the accompanying news article in audited textual information and a much-summarized format.
Research on self-learning of news category [11] indicates that computer-aided categorization of news is more accurate and efficient than humans because once the machines

are fed with some task, they will perform it faster than humans therefore, a computer-based classification is a better solution to go for. Browsing through categories benefits in boosting the search results by allowing the users to search and filter the outcomes based on the pre-defined news categories i.e. business, weather, social, technology, politics, sports, entertainment, and many more.

Therefore, the news categorization of headlines can save the efforts and time of the readers by lessening the need to search heterogeneous full-text news articles [34]. It is observed from the several works that as much as categorization of the textual data is essential in terms of time and effort, it is also regarded as one of the toughest classification methods in machine learning. During the last decade, numerous manual newspapers and magazine companies shifted to the digital world by developing their websites to update news to online users. Reading important news is quite valuable to users, but on the other hand side, it is also cumbersome as readers have to go through the entire article to find the useful news out of the articles. Therefore, the classification of news into its various categories seems to be essential to acquire the most relevant and useful information out of the long-length articles quickly.

Recently, the development in computing technology and the introduction of new machine learning algorithms e.g. reinforcement learning [22], text mining the goal of Artificial Intelligence (AI) has become a step closer. AI has important application in diverse fields including:healthcare [5]and [21], robotics and autonomous control [26] and [2],dynamic normative environments [28], ambient assisted living techniques for improvement in the quality of life of elder persons [8], drug identification [20], intelligent environments [3], games and self-organized system [4], vision enhancing method for low vision impairments [19] scheduling and management and configuration of resources, distributed fuzzy system for inferring in real-time critical situations [27], risk management [23] and computer vision.

Text mining often referred to as text data mining [31], is the analysis of textual semi-structured or unstructured data. Since the unstructured and fuzzy text is involved in text mining, therefore, it is regarded as more complex than the data mining process [16]. Therefore, the core aim of text mining is to transform the textual data into numerical data to apply data mining algorithms to it. As text mining is a multidisciplinary area of research, the current research will follow the application of text categorization [30] from news headlines.

The rest of the paper is as follows: a literature review is described in section 2, section 3 represents the proposed framework, section 4 represents the case study and section 5 concludes the paper.

2. Literature Review

Over the recent years, plenty of methodologies have been formulated in the domain of news exploration systems and web news mining. There are numerous studies found in the literature on the automatic categorization of textual data [32]. Text Categorization is the automated allocation of textual documents into their pre-defined categories. It follows machine learning classification algorithms to build models. Several works proposed and compared various algorithms for text categorizing. A lot of techniques and algorithms are there to classify the textual data including Naïve Bayes, Support Vector Machines,

Decision Trees, Neural Network, Random forest, and many more. The notion of these classifiers is to automatically predict the incoming news article to some pre-defined class using a trained classifier.

Over the last few years, classification of news headlines has been an area of research including, classification of emotions, classification of financial news [10], headlines classification with N-gram model [18], automated categorization of news headlines [25], classification of news headlines with SVM [7], mining of emotions from headlines of news [14], and Twitter classification of news from short news headlines [9]. It is very important to have a proper news headlines categorization in our lives. Text classification is a method of allocating predefined categories to a text in conformity with its contents [12]. A well-categorized dataset of news has to be utilized for exploration such as prediction of the stock market, news categorization and trading system, and news-oriented stock trend prediction. [29].

[1] concluded that enhancement for lessening the manual effort is attained by providing the name of the category as the only input keyword for text classification. The multiplication of Word-Net and similarity-based Latent Semantic Analysis (LSA) is carried to compute the final similarity score of documents with category names. Reuters-10 corpus revealed improvement in the precision, with additional amendments and variations in lexical references and context model indicated in [1]. A category name as an initial input-based classification scheme is shown in [17]. The work in [32] carried out relative literature on different algorithms comprising of Support Vector Machines (SVM), Linear Least-Squares Fit (LLSF), K-Nearest Neighbor (kNN), Neural Network (Net), Naive Bayes (NB). He concluded that SVM generated the best performance [15].

A comparative literature review on feature selection in text classification is presented in [33]. He stated that one of the major issues in the categorization of text is the feature space's high dimensionality. The feature set for textual documents consists of unique words that appear in all the documents. To build an efficient and effective model, feature selection is a technique applied. In [13] the author worked on the accuracy of full news article classification using neural networks. Authors in [6] order the news to various classes using SVM, then applied to preprocess techniques and feature selection based on TF-IDF. He used two datasets namely BBC news and 20newsgroup respectively.

It is witnessed from the literature that major work has been presented on long-length news categorization in text mining whereas; the research work on news headlines is still specifically limited. In the current research, the core purpose is to conduct news categorization on news headlines instead of complete text news because long-length news classification is noticeably tough, tiresome, and is computationally expensive. In comparison to the news headlines, the chances of misclassification in large descriptive news articles are more conspicuous.

Therefore, the core objective of the current research is to provide a framework for the E-News channels and news portals to automatically categorize the online news headlines into their pre-defined class by applying effective text preprocessing and classification techniques on a rich news categories dataset that will ultimately improve the overall classification process and yield better accuracy and lessen the computational complexity.

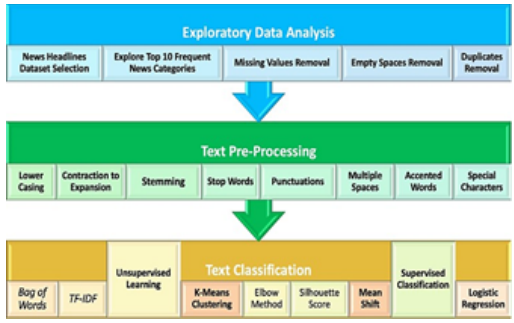


Figure 1. The proposed model.

3. System Model

The current paper provides a framework that can automatically classify news headlines into their pre-defined news categories using machine learning and text mining techniques. Figure 1 represents the proposed framework that follows three steps; exploratory data analysis, text pre-processing, and text classification. As a pre-requisite to news headlines classification, firstly a news headlines dataset is chosen so that the whole proposed process can be applied. After the selection of an appropriate news headlines dataset, the dataset is explored to see if it is an unbalanced dataset or not. Furthermore, all the duplication, empty slots, missing values are removed to make the textual data structured.

The next step is text pre-processing that will make the data clean by applying text pre-processing techniques; lower casing, contraction to expansion, stemming, removal of URL, stop words, multiple spaces, punctuations, accented words, and special characters. The last step is the text classification step which uses two approaches: unsupervised classification with K-means Clustering and Means shift algorithm. The optimal numbers of clusters are automatically predicted in K-means clustering using the Elbow method. Whereas the quality of clusters in both the algorithms i.e., K-means clustering algorithm and Mean shift algorithms, is evaluated using Silhouette score that predicts the Silhouette coefficient which tells if the clusters are defined well or not. The other approach of news headlines classification is a supervised classification which uses Logistic Regression Classifier to yield the performance of the news headlines classification efficiently.

4. Case Study

This section depicts a case study that uses the proposed framework to classify news headlines into its news categories. Each step of the proposed framework is discussed in detail below.

4.1. Dataset

For the current work, News Category Dataset is taken from the Kaggle dataset repository, which contained 200,000 news headlines from 2012 to 2018. Each news headlines have a corresponding news category. The dataset contained six columns; category, headline, author, link, short description, and date.

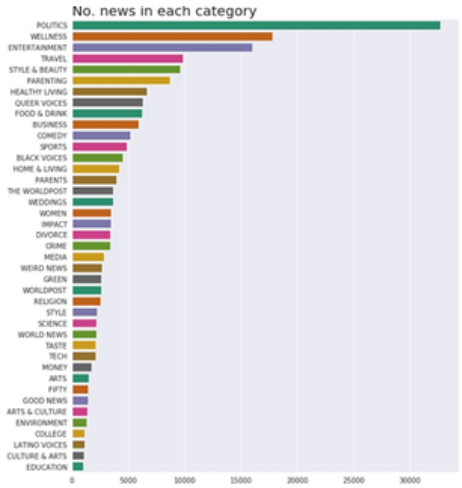


Figure 2. Total amount of news headlines in each category.

4.2. Exploratory Data Analysis

Exploratory data analysis is applied to the News Category Dataset. There were 200,583 news article values in the dataset. Figure 2 represents the 41 categories in the dataset. The exploration of data stated that the dataset was unbalanced as the three categories i.e., politics, wellness, and entertainment were more frequent than the other categories. An unbalanced dataset can highly affect the overall accuracy of the model, therefore; only those news categories were focused on that had enough news articles so that the model can be trained.

Consequently, for data exploratory only the top 10 most frequent balanced categories of news articles were explored. Subsequently, the textual redundancies, missing values, and faulty points from the sample were filtered. From a total of 200,583 news articles, only 120,008 rows were left after exploratory data analysis.

4.3. Text Pre-Processing

Next text pre-processing is applied which normalized, segmented and cleaned the data to make the data predictable by the machine learning algorithm. Firstly, a random sample of 7 rows was taken from the dataset consequently, the unique values, duplicates, empty values were checked. Authors, link and date columns from the dataset were removed as they were not giving any information. Following are the different techniques applied to pre-process, structure and clean the data.

4.3.1. Lower Casing

Firstly lower casing is applied on the text of three columns i.e., category, headline, short description to resolve the sparsity issue and get better outcomes. Making them in lower case helps in the consistency of the output.

	category	headline	short_description	headline_lemmatized_tokenized
77200	wellness	gps guide sandy c newbiggings simple steps cla...	stress strain constantly connected life wellbe...	[gps, guide, sandy, c, newbiggings, simple, st...
79743	food & drink	secrets skinny chef	combat getting weight control thinking balance...	[secret, skinny, chef]
9540	entertainment	jon stewarts war bs continues jordan klepper	opposition trying accomplish jon stewarts fina...	[jon, stewart, war, b, continues, jordan, klep...
33909	healthy living	common myths sleep aids debunked	myth 2 sleep aids improve quality sleep	[common, myth, sleep, aid, debunked]
40945	politics	conservative group urges republicans embrace e...	gop direct money medical research education in...	[conservative, group, urge, republican, embrac...

Figure 3. Data sample after Stemming.

4.3.2. Contraction to Expansion

Contraction to expansion method expanded all the abbreviated words in the data sample. The removal of contraction contributed to text standardization.

4.3.3. Uniform Resource Locators (URL)

Next step was to find the Uniform Resource Locators (URL), which are the textual references to a location on web. As they were not adding any information therefore, they were removed.

4.3.4. Stop Words

Furthermore, stop words were removed which are the most commonly used words in English , removing them means removing the meaningless words from the dataset and only concentrate on the highly meaningful words which can aid in the classification.

4.3.5. Stemming

Next stemming is applied on the data that uses a crude heuristic procedure to cut off the ends of words in the hope of properly transforming them into their root words. It helps in improving classification accuracy which saves required time and space. Figure 3 represents the data obtained after stemming. Likewise, the punctuations, multiple spaces, special characters and accented words were removed from the data as they were not adding any value to the data. Removing them made the data more standardized and usable for extracting meaningful insights.Figure 4 represents the unbalanced dataset with a huge difference between the first categories; politics as 32736 or 26.7% news headlines and the tenth category as 5937 or 4.63% news headlines. To train the machine learning model the news category was normalized with 5937 news headlines occurrences as this was the lowest value of the tenth row.

4.4. Feature Extraction

Feature extraction is the prerequisite for training a model, which transforms the textual data into numerical data. For the current research two approaches of feature extraction are used i.e., Bag of Words model and TF-IDF algorithm.

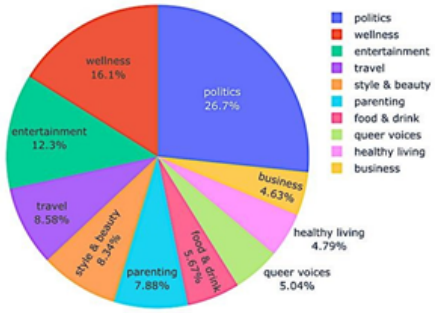


Figure 4. Top 10 categories represents it is an unbalanced dataset.

4.4.1. Bag of Words (BoW

As machine learning algorithms cannot work with raw text directly so the textual data was converted into numerical vectors of numbers using BoW. A list of unique words from the textual data obtained from the preprocessed top 10 categories were identified to design the vocabulary and then on the basis of their presence, these words were scored in each document to form a vector which can be used as an input to the machine learning model. It disregarded the order of words and all the grammatical details in the document and thus easily the textual data is represented into its equivalent vector of numbers.

4.4.2.

The next step was to apply Term Frequency-Inverse Document Frequency which spotted all the important words from the 10 categories. Words with less tfidf scores were considered as they had high importance in the document matrix. Following is the calculation of TF-IDF score for the word t in document d from the document set D:

$$\text{tf idf (t, d, D)} = \text{if (t, d). idf(t, D)} \tag{1}$$

$$\text{tf (t, d)} = \log(\text{I} + \text{freq(t, d)}) \tag{2}$$

N is the total number of documents in our text corpus, therefore, IDF Inverse Document Frequency is:

$$\text{idf (t, d)} = \log\left(\frac{N}{countd \in D; t \in d}\right) \tag{3}$$

After the text was converted into numerical data, the next step was to apply text classification using two approaches of machine learning classification, supervised classification and unsupervised classification.

4.5. Text Classification

The aim of text classification is to label the textual data into its relevant categories. In the current case study, the news headlines are divided into two data sets; train data set

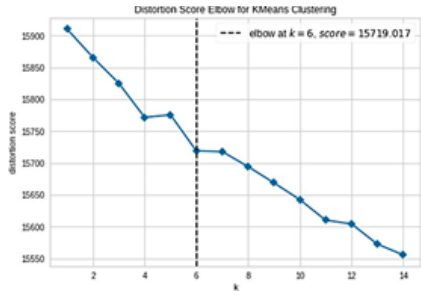


Figure 5. Distortion Score Elbow for K-Means Clustering

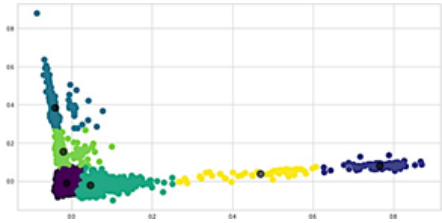


Figure 6. Visual representation of 6 clusters obtained from K-Means

which comprises of 80% of the news articles and the remaining 20% of news articles for testing. 10-fold cross validation is also applied to let the model learn from the test data. For the text classification only 20,000 news headlines samples are chosen because of memory issues.

4.5.1. Unsupervised Learning

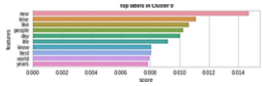
Unsupervised learning is used to train a machine learning model using unlabeled data and allows the model to discover the unknown information, hidden structure, patterns in data and find features that can be useful for categorization without any prior training. Two algorithms are used to classify news headlines categories.

-**K-Means clustering**

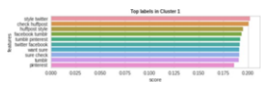
K-Means clustering also known as distance-based algorithm, is the most efficient method to cluster data. It is applied on the preprocessed news text sample that was first converted into numerical data using TF-IDF and Bag of Words model. Each datapoint in the data sample was assigned to its closest centroid to form a cluster which minimized the distance of the data points within the cluster. To choose the right number of clusters, elbow method is used.

Elbow method automatically selected the optimal number of clusters with KElbowVisualizer method for K-Means algorithm by fitting the model with a range of values for K. Figure 5 shows the distortion score Elbow for K-means clustering, where distortion is the sum of squared distances from each point to its assigned center point. Elbow method predicted 6 clusters for news categories. Figure 6 shows the visual representation of the 6 clusters obtained from Elbow method.

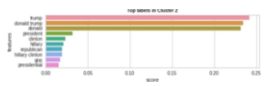
The next step was to check the quality of the clusters obtained from K-means algorithm and elbow method using **Silhouette** score with higher Silhouette coefficient which proved that the model contained well defined clusters score. After K-means clustering



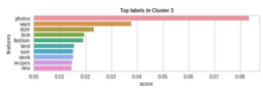
(a) Top 10 most frequent labels in Cluster 0



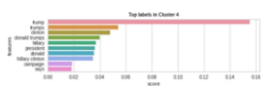
(b) Top 10 most frequent labels in Cluster 1



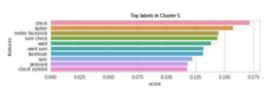
(c) Top 10 most frequent labels in Cluster 2



(d) Top 10 most frequent labels in Cluster 3



(e) Top 10 most frequent labels in Cluster 4



(f) Top 10 most frequent labels in Cluster 5

Figure 7. Top 10 most frequent labels in Cluster

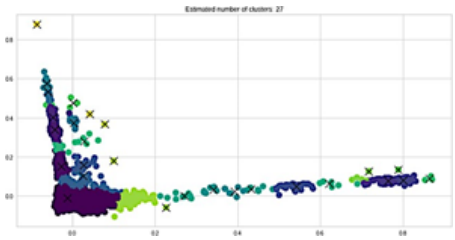


Figure 8. 27 Clusters obtained from Mean Shift Algorithm

is done, top 10 most frequent words in each 6 predicted clusters are analyzed. Figure 7 represents the results of most frequent words and their scores in each predicted cluster.

-**Mean Shift**

Next step was to use mean shift algorithm to find clusters from the data sample. Mean Shift is also referred to as centroid based algorithm or mode seeking algorithm. In a given region, it shifts the candidate data points to the closest centroids of the clusters iteratively to be the mean of the points. Then these data points are filtered in a later stage to remove the redundant points to form final set of centroids. Unlike K-means clustering, it does not requires specifying the number of clusters as the number of clusters is automatically determined by the algorithm with respect to the data. Figure 8 represents the 27 clusters obtained by Mean Shift algorithm. The quality of clusters obtained from Mean shift algorithm using Silhouette score is checked. Similarly, top 10 words from the predicted 27 clusters were obtained.

4.5.2. Supervised Classification

The last step is the supervised learning which is training the machine learning model with sorted and labeled data. As the data sample was text so it was first converted into numerical data with TF-IDF algorithm and then Logistic regression algorithm on the sample data was applied. Logistic regression algorithm is most used machine learning predictive analysis algorithm which predicts the categorical dependent variable using a set of given independent variables to assess the probability of an event's success or failure. It is very fast and efficient to train and discover the classification of the unknown records.

The experimental results obtained by logistic regression proved that prediction of news headlines into their pre-defined categories can be best gained by the use of logistic re-

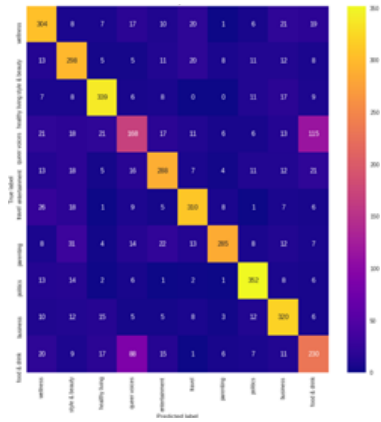


Figure 9. Heatmap representing prediction by Logistic Regression.

gression as it gave 74% accuracy in classifying the news categories. Figure 9 represents the heatmap obtained by Logistic Regression having higher values on diagonal verifies that the model is predicting the headlines to their pre-defined categories accurately. Therefore, the case study proved that with the use of the proposed framework, online news channels and news portals can easily increase their speed and efficiency of news headlines classification and reduce their computational complexity.

5. Conclusion and Future Work

We have presented a framework that self learn the category of a news. The system in the first step, exploratory data analysis on the news dataset is applied, and then text pre-processing techniques are applied to remove the unwanted and useless information from the data to get meaningful insights. Lastly, text classification is carried out using two approaches of classification, unsupervised clustering with K- means clustering and mean shift algorithm, and supervised classification with logistic regression. The proposed framework is persistent and cost-effective which will help the e-news channels and news portals to automatically classify the huge set of scattered news headlines into their pre-defined news categories efficiently and effectively. Thus, the automatic news headlines categorization will reduce the computational complexity.

In the future, the proposed framework can be extended to analyze the sentiments of users on news articles using sentiment analysis. Currently, this framework is applied to news headlines classification however, in the future; this framework methodology can be applied to other datasets such as movies or talk shows datasets to automatically categorize the movies and talk shows into their pre-defined categories efficiently and hence the accuracy can be improved. Similarly, developing a recommender system that recommends news categories to the users based on their user profiles remains a milestone to be achieved in the future.

References

[1] Libby Barak, Ido Dagan, and Eyal Shnarch. Text categorization from category name via lexical reference. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 33–36, 2009.

[2] Marina Bonomolo, Patrizia Ribino, and Gianpaolo Vitale. Explainable post-occupancy evaluation using a humanoid robot. *Applied Sciences*, 10(21):7906, 2020.

[3] A. Coronato and G. De Pietro. Tools for the rapid prototyping of provably correct ambient intelligence applications. *IEEE Transactions on Software Engineering*, 38(4):975–991, 2012.

[4] A. Coronato and G. D. Pietro. Formal design of ambient intelligence applications. *Computer*, 43(12):60–68, 2010.

[5] Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020.

[6] Seyyed Mohammad Hossein Dadgar, Mohammad Shirzad Araghi, and Morteza Mastery Farahani. A novel text mining approach based on tf-idf and support vector machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 112–116. IEEE, 2016.

[7] RR Deshmukh and DK Kirange. Classifying news headlines for providing user centered e-newspaper using svm. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2(3):157–160, 2013.

[8] Claudia Di Napoli, Patrizia Ribino, and Luca Serino. Customisable assistive plans as dynamic composition of services with normed-qos. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–26, 2021.

[9] Inoshika Dilrukshi, Kasun De Zoysa, and Amitha Caldera. Twitter news classification using svm. In *2013 8th International Conference on Computer Science & Education*, pages 287–291. IEEE, 2013.

[10] Brett Drury, Luis Torgo, and JJ Almeida. Classifying news stories to estimate the direction of a stock market index. In *6th Iberian Conference on Information Systems and Technologies (CISTI 2011)*, pages 1–4. IEEE, 2011.

[11] Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium, and Wahyu Muliady. 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)*, pages 1–4. IEEE, 2014.

[12] Rajni Jindal, Ruchika Malhotra, and Abha Jain. Techniques for text classification: Literature review and current trends. *webology*, 12(2), 2015.

[13] Sandeep Kaur and Navdeep Kaur Khiva. Online news classification using deep learning technique. *International Research Journal of Engineering and Technology (IRJET)*, 3(10):558–563, 2016.

[14] DK Kirange and RR Deshmukh. Emotion classification of news headlines using svm. *Asian Journal of Computer Science and Information Technology*, 5(2):104–106, 2012.

[15] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[16] Shu-Hsien Liao, Pei-Hui Chu, and Pei-Yuan Hsiao. Data mining techniques and applications–a decade review from 2000 to 2011. *Expert systems with applications*, 39(12):11303–11311, 2012.

[17] Chaya Liebeskind, Lili Kotlerman, and Ido Dagan. Text categorization from category name in an industry-motivated scenario. *Language resources and evaluation*, 49(2):227–261, 2015.

[18] Xin Liu, Gao Rujia, and Song Liufu. Internet news headlines classification method based on the n-gram language model. In *2012 International Conference on Computer Science and Information Processing (CSIP)*, pages 826–828. IEEE, 2012.

[19] Carmelo Lodato and Patrizia Ribino. A novel vision-enhancing technology for low-vision impairments. *Journal of medical systems*, 42(12):1–13, 2018.

[20] Muddasar Naeem, Giovanni Paragiola, Antonio Coronato, and Giuseppe De Pietro. A cnn based monitoring system to minimize medication errors during treatment process at home. In *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, pages 1–5, 2020.

[21] Muddasar Naeem, Giovanni Paragliola, and Antonio Coronato. A reinforcement learning and deep learning based intelligent system for the support of impaired patients in home treatment. *Expert Systems with Applications*, page 114285, 2020.

[22] Muddasar Naeem, S Tahir H Rizvi, and Antonio Coronato. A gentle introduction to reinforcement learning and its application in different fields. *IEEE Access*, 2020.

[23] Giovanni Paragliola and Muddasar Naeem. Risk management for nuclear medical department using reinforcement learning algorithms. *Journal of Reliable Intelligent Environments*, 5(2):105–113, 2019.

[24] [1] Pew Research Center U.S. Politics Policy. Americans spending more time following the news. Available at: ¡https://www.pewresearch.org/politics/2010/09/12/americans-spending-more-time-following-the-news, 2021.

[25] Mark W Pope. Automatic classification of online news headlines. 2007.

[26] Patrizia Ribino, Marina Bonomolo, Carmelo Lodato, and Gianpaolo Vitale. A humanoid social robot based approach for indoor environment quality monitoring and well-being improvement. *International Journal of Social Robotics*, pages 1–20, 2020.

[27] Patrizia Ribino and Carmelo Lodato. A distributed fuzzy system for dangerous events real-time alerting. *Journal of Ambient Intelligence and Humanized Computing*, 10(11):4263–4282, 2019.

[28] Patrizia Ribino and Carmelo Lodato. A norm compliance approach for open and goal-directed intelligent systems. *Complexity*, 2019, 2019.

[29] Satoru Takahashi, Masakazu Takahashi, Hiroshi Takahashi, and Kazuhiko Tsuda. Analysis of the relation between stock price returns and headline news using text categorization. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 1339–1345. Springer, 2007.

[30] Ah-Hwee Tan et al. Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases*, volume 8, pages 65–70. Citeseer, 1999.

[31] Sholom M Weiss, Nitin Indurkhya, Tong Zhang, and Fred Damerau. *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media, 2010.

[32] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, 1999.

[33] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, page 35. Nashville, TN, USA, 1997.

[34] Yingwu Zhu. Measurement and analysis of an online content voting network: a case study of digg. In *Proceedings of the 19th international conference on World wide web*, pages 1039–1048, 2010.