Journal of Education for Pure Science- University of Thi-Qar
Vol.12, No2 (December, 2022)
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

DOI: http://doi.org/10.32792/utq.jceps.12.02.31

# Comparative Study for News Categorization by Multinomial Naïve Bayes and Support Vector Machine

**Melad Farawn Ogaib[1] , Prof. Dr. Kadhim Mehdi Hashim[2]**

[1,2] Department of Computer Science, College of Education for pure Sciences, University of Th-Qar[1], Iraq.
Imam Ja'afar Al-Sadiq University[2], Iraq.

**Received 9/5/2022 Accepted 19/5/2022 Published 01/12/2022**

## Abstract:

Online news is published in very large numbers, searches take a long time, and because of this huge number of news articles that include different genres (such as politics, sports, business, entertainment, technology, health, economics, real estate, art, etc) it will be difficult for users to access the important news that It suits their inclinations and desires. In this paper, the news group of BBC News is categorized into five categories, including (sports, business, politics, entertainment and technology). The objective of is to classify news into its own category to help users quickly and easily access relevant news without wasting any time through classification methods that use machine learning algorithms. The classification algorithms Multinomial Naive Bayes and Support Vector Machine were applied to the news data set after extracting the features from it using count vectorizer method and TF-IDF vectorizer. SVM algorithm has proven superiority over MNB in count vectorizer with an accuracy 99.1% and TF-IDF vectorizer with an accuracy 98.2%.

**Keywords:** Multinomial Naive Bayes , Support Vector Machine, TF-IDF vectorizer

## Introduction:

Text classification (TC) is a supervised learning task that entails classifying natural language text objects into one (or more) categories[1]. People communicate in a variety of ways including speaking , listening and gesturing , using specialized hand signals (such as when driving or directing traffic) and utilizing sign languages for the deaf ,text refers to words written or printed on a flat surface (paper, card, street signs, etc) or displayed on a screen or electronic device with the purpose of being read by its intended audience or by anybody passing [2]. Text categorization problems have been extensively researched and handled in a variety of real-world applications. Many researchers are currently interested in building applications that use text categorization algorithms [3]. It's also challenging to manually process and classify all of the data plainly, this resulted in the development of intelligent text processing tools to examine lexical and linguistic trends in the field of natural language processing. Text analytics employs techniques such as clustering and classification [4].

Categorization is the division of the objects into groups of entities whose members are similar in some sense[5]. An object is assigned to a predetermined group or class based on a number of observed features, classification is one of the most common decision-making processes in human activity. And many problems have been classified as classification problems ,some examples are stock market forecast, weather forecasting, bankruptcy forecasting, medical diagnosis, speech recognition and personality recognition[6].

Online news stories are one area where we come across a lot of content infrequently. People are concerned about their busy lifestyles as a result of the advancement of information technology, and they only want to read news articles related to their interests. It is a difficult process to find news relevant to a person's interests because many news articles are useful but may be irrelevant. News articles are categorized by news type/category for easy selection of people according to their likes, news and the type of news may affect the attractiveness of an individual [7].

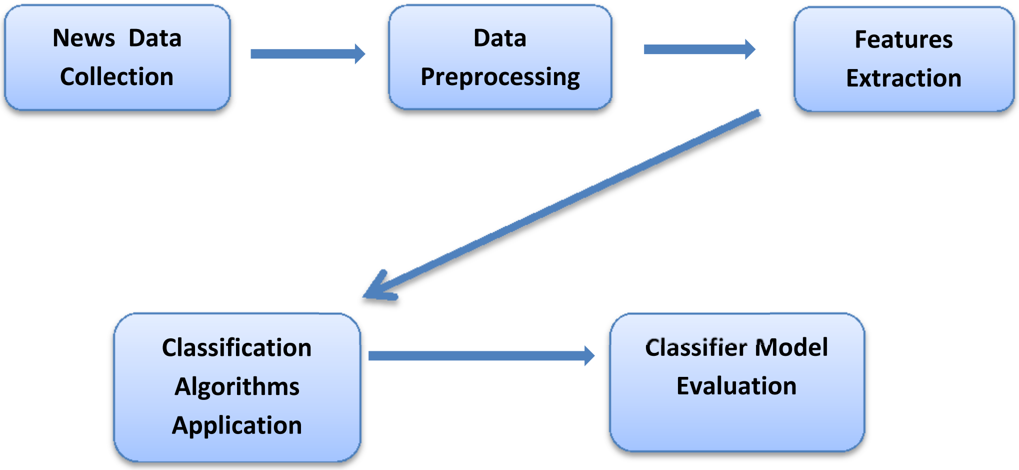
News Data Collection
Data Preprocessing
Features Extraction
Classification Algorithms Application
Classifier Model Evaluation

**Figure (1): News classification process**[7].

Machine learning algorithms cannot work with the input text directly, so the numeric attributes of the document must be extracted in order to learn the parameters and extract predictions. Artificial neural networks, support vector machines, decision trees, and other machine learning algorithms can be used to address the text classification problem. Text classification issues differ from other machine learning tasks in that they have a large feature space, as a result, text classification algorithms that can cope with thousands of features efficiently perform well[8].

## 1. RELATED WORKS

Classification is the most popular data mining method, which employs a set of pre-classified samples to develop a model that can categorize the full population of data. This type of study is very well suited to fraud detection and credit risk applications. In the data categorization process both learning and classification are involved[9].

News classification is an important field as a result of the increase in the volume of news published on various sites in general and news sites in particular. In recent years text and visual descriptions (photos and video) are included in these articles as well as multimedia resources [10]. Articles are published on the Internet in very large numbers, searches take a long time, therefore many previous studies focused on the classification of news, including: K. Thandar Nwet et al. In [11], the data set is made up of 12,000 papers divided into four categories. For Myanmar Language News categorization, a comparative analysis of machine learning techniques such as Nave Bayes (NB), k-nearest neighbours (KNN), and support vector machine (SVM) algorithms is conducted, the news corpus is used for the classifier's training and testing, the chi square algorithm, a feature selection method, achieves comparable performance across a number of classifiers. The experimental results in this paper also reveal that support vector machine has a higher accuracy of (85.5%), NB (80.5%) and KNN (81.2%).

D. Liparas Y. Hacohen-Kerner et al. In [10] , the data is classified using N-gram text properties extracted from the text and visual features produced from a single representative image. The application domain contains English-language news stories in five categories: business finance, lifestyle, entertainment, science and technology, and sports from three well-known news sites (BBC, Reuters and The Guardian).The random forest machine learning method was employed to perform multiple categorization trials, using N-gram text features gave accuracy (84.4%), while using visual features gave accuracy (53%), while using N-gram text features and visual features together resulted in somewhat higher accuracy (86.2%).

M. Ramdhani, D. Maylawati et al. In [12], a convolutional neural network (CNN) was used to classify text in the Indonesian language. The CNN algorithm has been modified to fit the features of the Indonesian language ,as a result, the stop word was removed and the derivation was found for each word during the text preprocessing phase. The experiment used 472 Indonesian news text data from multiple sources in four categories: (entertainment), (sports), (most important news), and (technology), the experiment's findings demonstrate that CNN classified Indonesian news with a high degree of accuracy (90.74%).

M.Khan et al. In [13] , using online Urdu news data to train algorithms and automatically categorize the information presented, the results showed that the Bernoulli Nave Bayes (Bernoulli NB) and Multinomial Nave Bayes (Multinomial NB) algorithms beat the others in terms of all performance criteria. The results indicate the superiority of the Multinomial NB classifier with an accuracy of 94.278 %, followed by the Bernoulli NB classifier with an accuracy of 94.274 %, a short text of news headlines can be used as input to classify the text.

Mulahuwaish , K. Gyorick et al. In[14], A useful online model has been proposed for extracting news data and organizing newspapers into five categories: business, technology, science, health, and entertainment. Four different machine learning classifiers were compared (SVM, nearest neighbors (KNN), decision tree (DT), and long-term memory (LSTM)), the results obtained show that KNN has the lowest accuracy (88.72%) and SVM has the highest accuracy (95.04%).

## 2. PROPOSED METHOD

News is categorized into different categories using machine learning algorithms. In this system news is classified into its different categories based on the features extracted from the text, and this process is considered one of the most important steps used to classify news because the classification process depends mainly on extracting features to reach the best classification. Figure (2) shows the basic stages used in this system.

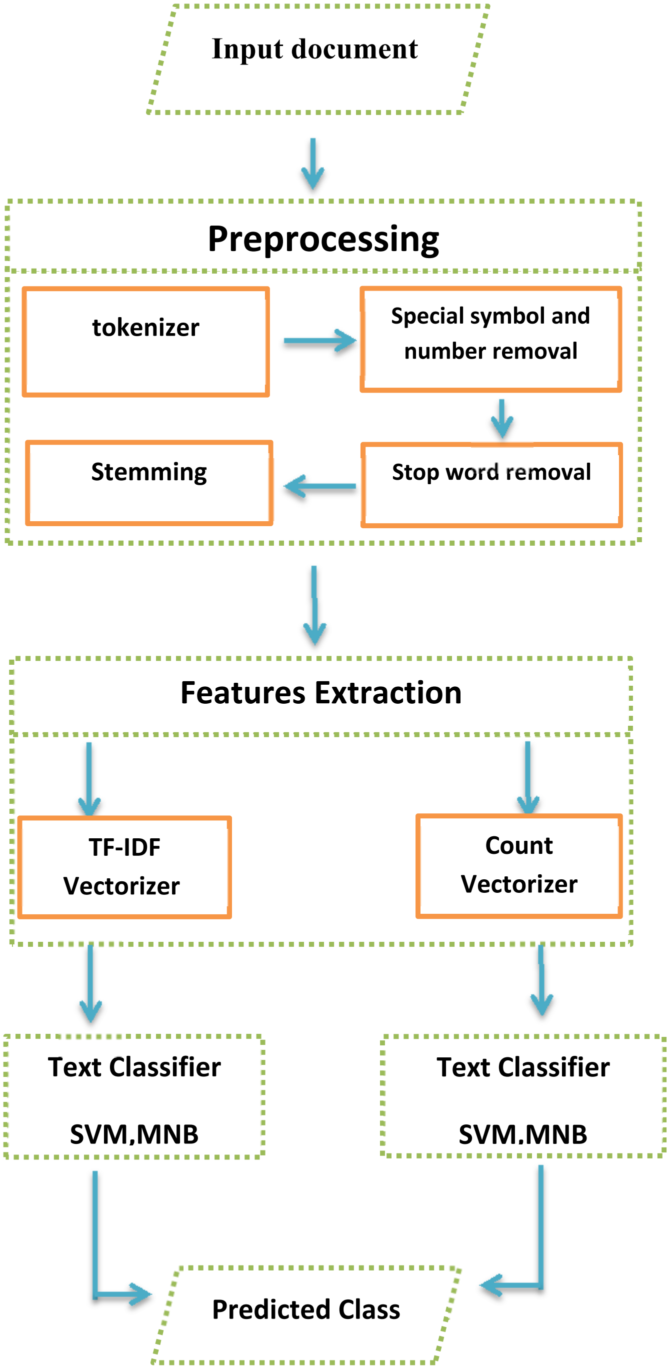

**Figure (2): The proposed System for news categorization**

## 2.1 Data Set

The data we wish to categorize is collected from BBC News in 2,225 texts divided into five categories: business, sports, technology, political and entertainment. The news data is distributed as follows:(511 sport ,510 business, 417 politics , 401 technology , 386 entertainment).

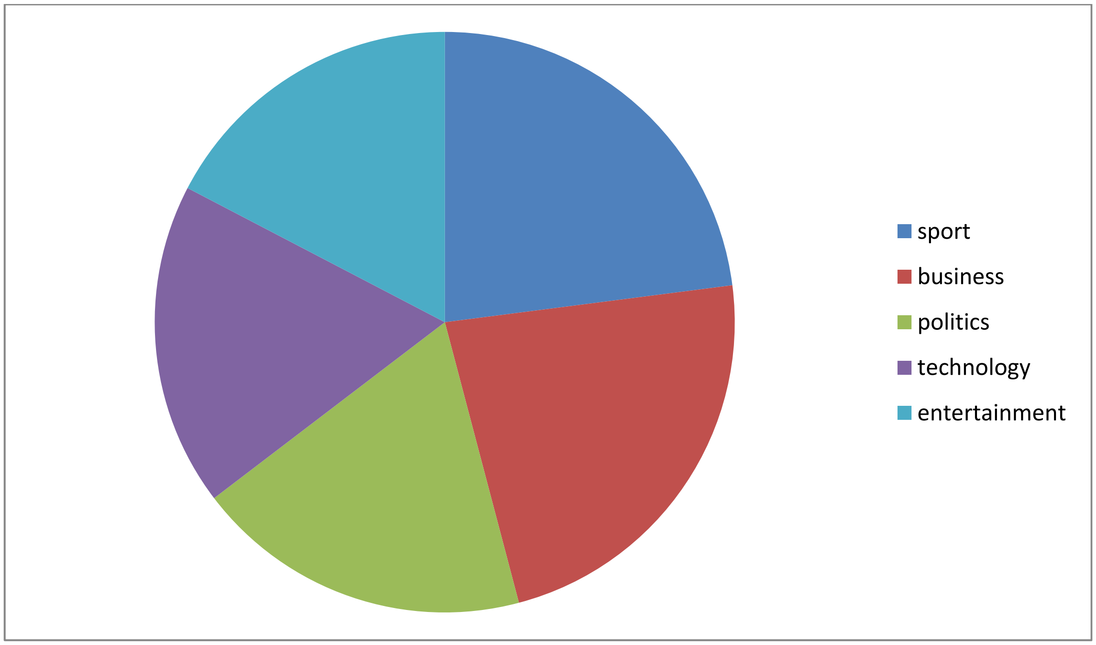

Figure (3): Categories of news dataset

## 2.2 Pre-Processing Steps

At this stage, several operations are performed on the texts to format them and make them ready for use and to reduce noise in the text for the purpose of obtaining the best performance for the classifier.

**A. Tokenization:** This process includes dividing the text into words or sentences, in word tokenization, each word is separated from the other by a separator, and each word is placed in quotation marks in order to facilitate the handling of the text and perform other operations better**.**

**B. Special symbol and number removal:** Remove any symbol from the set of special symbols (!, :, ÷, ×, º, >, <, \, /, @, #, $,%, ^, &, *, ), (, _, -, +, =, ~, ø, [, ], ‘, ”, etc) in the text to keep only related words, resulting in a much better feature extraction procedure and higher classification accuracy.

**C. Stop word removal**: Commonly used many words in the English language which are never without text such as (conjunctions, pronouns and prepositions) these words are not related to a specific category but are present in all categories, stop words do not carry any important information in the text and have a low value, therefore, they are removed and the most relevant and high valuable words are used in the classification process.

**D. Stemming**: Derivation is a very important step in the preprocessing where the number of words is reduced by returning each word to its root, since each word can come in several images, these images return to one root. The process of derivation is important because it reduces the time to extract features.

## 2.3 Feature Extraction

The main objective of feature extraction is to define a set of inputs containing relevant information that is used in the text classification process. In our proposed system, the following algorithms have been used to extract features from text:

**A. Count Vectorizer:** It is a very important tool provided by the scikit-learn library in Python whose function is to convert the text into a vector of values that represent the relevant features that will be entered into the classification algorithm in the training and testing phases, this algorithm depends on the repetition of each word in the text, which is very useful in text analysis operations. A matrix is created where we put the words in the columns, the rows represent each text sample in the document and each cell in the matrix contains the number of repeats of the word in the sample.

**B. TF-IDF Vectorizer:** This algorithm is used very widely to extract the features as it works to clarify the importance of each word in the text by calculating its weight, the greater the weight of the word, the greater its importance in the text.

## 2.4 Classification

After the process of extracting the features, the news data set is classified into its specific categories. In the proposed system multinomial naïve bayes algorithm and support vector machine are used in the classification process.

**A. Multinomial Naïve Bayes :** is one a form of naïve bayes used for multinomial distributed data . It is widely used as a baseline in text classification since it is fast and easy to construct . Furthermore, with proper preparation, it may compete with more complex methods such as support vector machines[15]. In the implementation of Multinomial Naïve Bayes (MNB). we need to add a smoothing one to the conditional probability so as to avoid zero probability of new terms in the testing set that were not available in the training set[16].

**B. Support Vector Machine:** are a supervised learning method used for classification and regression. Support Vector Machines are systems that use the hypothesis space of linear functions in a high-dimensional feature space and are taught with an optimization theory-based learning algorithm that incorporates a learning bias derived from statistical learning theory[17]**.** The time complexity of SVM is extremely high, which is one of its major drawbacks. SVMs were originally designed to deal with binary classification, however in today's world we frequently have large amounts of data to classify. Classification with more than two classes is referred to as multiclass classification[18].

### C. Count Vectorizer with MNB and SVM Classifier

The data set is divided into 80% training and 20 % testing, the classification algorithms are trained on the features extracted from count vectorizer, and then the test features are classified and the accuracy calculated for both MNB and SVM algorithms.

### D. TF-IDF Vectorizer with MNB and SVM Classifier

The data set is divided into 80 % training and 20 % testing, the classification algorithms are trained on the features extracted from TF-IDF Vectorizer, and then the test features are classified and the accuracy calculated for both MNB and SVM algorithms.

## 2.5 Evaluation

Multinomial Naïve Bayes and Support Vector Machine learning algorithms are applied to classify the selected newsgroup. The proposed model is evaluated and performance measures of each algorithm calculated using Count Vector and TFIDF Vectorizer feature extraction methods.

Some well-known classifier evaluation measures include accuracy (also known as recognition rate), sensitivity (or recall), specificity, precision, and F1 Score [7], [19], [20].

**A. Accuracy:** Is the number of correct predictions made by the classifier divided by the total data collection. The accuracy is stated mathematically as

$$\textbf{Accuracy}=\frac{(TP+TN)}{TP+FP+TN+FN} \qquad \dots\dots\dots\dots (1)$$

**B. Precision:** By dividing the number of articles that are correctly recognized (True Positive) by group (True Positive and False Positive). The precision in mathematical form is:

$$\textbf{Precision} = \frac{TP}{TP+FP} \qquad \dots\dots\dots\dots (2)$$

**C. Recall:** The recall is defined as the proportion of accurately predicted positive articles to group (True Positive and False Negative). The recall is mathematically expressed as

$$\textbf{Recall} = \frac{TP}{TP+FN} \qquad \dots\dots\dots\dots (3)$$

**D. F1 score:** The F1 score is the harmonic mean of precision and recall. When we add Precision and Recall together we get the F1 score, Because it includes both precision and recall, using one value for measurement is more efficient than using two. F1 score is presented as

$$\textbf{F1- Score}=\frac{2*Recall*Precision}{(Recall+Precision} \qquad \dots\dots\dots (4)$$

The following tables show the experimental results of the proposed model:

Table (1): MNB with count and TF-IDF

| | MNB with Count | | | MNB with TF-IDF | | | |
|---|---|---|---|---|---|---|---|
| **Category** | **Precision** | **Recall** | **F1 SCOR** | **Precision** | **Recall** | **F1 Score** | |
| **Sport** | **1.00** | **1.00** | **1.00** | **0.97** | **1.00** | **0.99** | |
| **business** | **0.97** | **0.97** | **0.97** | **0.93** | **0.99** | **0.96** | |
| **Politics** | **0.99** | **0.99** | **0.99** | **1.00** | **0.98** | **0.99** | |
| **Tech** | **0.96** | **1.00** | **0.98** | **0.96** | **0.97** | **0.97** | |
| **Entertainment** | **1.00** | **0.97** | **0.98** | **1.00** | **0.92** | **0.96** | |
| **Accuracy** | **%98.5** | | | **%97.3** | | | |

Table (2): SVM with count and TF-IDF

| | SVM with Count | | | SVM with TF-IDF | | | |
|---|---|---|---|---|---|---|---|
| **Category** | **Precision** | **Recall** | **F1 SCOR** | **Precision** | **Recall** | **F1 Score** | |
| **Sport** | **1.00** | **1.00** | **1.00** | **0.98** | **0.97** | **0.99** | |
| **business** | **0.98** | **0.97** | **0.98** | **0.97** | **0.97** | **0.97** | |
| **Politics** | **0.99** | **0.99** | **0.99** | **0.99** | **0.98** | **0.98** | |
| **Tech** | **0.98** | **1.00** | **0.99** | **0.99** | **0.97** | **0.98** | |
| **Entertainment** | **1.00** | **1.00** | **1.00** | **0.99** | **0.99** | **0.99** | |
| **Accuracy** | **%99.1** | | | **%98.2** | | | |

## 3. CONCLUSION

Machine learning algorithms do not have the ability to deal with texts only after converting them into numerical values, therefore in this paper algorithms (count vectorizer and TF-IDF vectorizer ) were used to obtain numerical values for each feature and include them in classification algorithms for training and testing. The results show the superiority of the SVM algorithm on multinomial naïve bayes algorithm with an accuracy of 98.2% with TF-IDF and 99.1% with count vectorizer. Future work may include the use of a data set that contains all news categories.

## REFERENCES

[1] **Y. HaCohen-Kerner, D. Miller, and Y. Yigal**, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, May 2020, doi: 10.1371/journal.pone.0232525.

[2] **R. Kibble,** "Introduction to natural language processing Undergraduate study in Computing and related programmes," *Roeper Rev.*, vol. 1, no. 2, p. 26, 2013.

[3] **K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown,** "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[4] **A. Wilkinson, N. Wenger, and L. R. Shugarman,** "L Iterature R Eview on," vol. 7, no. June, pp. 52–57, 2007.

[5] **E. K. Jacob,** "Classification and categorization: A difference that makes a difference," *Libr. Trends*, vol. 52, no. 3, 2004.

[6] **R. Sathya and A. Abraham**, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *Int. J. Adv. Res. Artif. Intell.*, vol. 2, no. 2, 2013, doi: 10.14569/ijarai.2013.020206.

[7] **J. Ahmed and M. Ahmed,** "Online News Classification Using Machine Learning Techniques," *IIUM Eng. J.*, vol. 22, no. 2, pp. 210–225, 2021, doi: 10.31436/iiumej.v22i2.1662.

[8] **U. Suleymanov and S. Rustamov,** "Automated News Categorization using Machine Learning methods," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 459, no. 1, doi: 10.1088/1757-899X/459/1/012006.

[9] **N. Dey, A. S. Ashour, and G. N. Nguyen,** "Deep learning for multimedia content analysis," *Min. Multimed. Doc.*, vol. 1, no. 4, pp. 193–203, 2017, doi: 10.1201/b21638.

[10] **D. Liparas, Y. Hacohen-Kerner, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris**, "LNCS 8849 - News Articles Classification Using Random Forests and Weighted Multimodal Features," 2014. [Online]. Available: http://www.bbc.com/news/business-25445906.

[11] **K. Thandar Nwet,** "Machine Learning Algorithms for Myanmar News Classification," *Int. J. Nat. Lang. Comput.*, vol. 8, no. 4, pp. 17–24, Aug. 2019, doi: 10.5121/ijnlc.2019.8402.

[12] **M. A. Ramdhani, M. A. Ramdhani, D. S. adillah Maylawati, and T. Mantoro,** "Indonesian news classification using convolutional neural network," *Indones. J. Electr. Eng. Comput. Sci.*, vol.

19, no. 2, pp. 1000–1009, Aug. 2020, doi: 10.11591/ijeecs.v19.i2.pp1000-1009.

[13] **M. B. Khan,** “Urdu News Classification using Application of Machine Learning Algorithms on News Headline,” *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 2, p. 229, 2021, doi: 10.22937/IJCSNS.2021.21.2.27.

[14] **A. Mulahuwaish, K. Gyorick, K. Z. Ghafoor, H. S. Maghdid, and D. B. Rawat**, “Efficient classification model of web news documents using machine learning algorithms for accurate information,” *Comput. Secur.*, vol. 98, Nov. 2020, doi: 10.1016/j.cose.2020.102006.

[15] **S. Xu, Y. Li, and Z. Wang,** “Bayesian multinomial naïve bayes classifier to text classification,” *Lect. Notes Electr. Eng.*, vol. 448, no. November, pp. 347–352, 2017, doi: 10.1007/978-981-10-5041-1_57.

[16] **H. M. Ismail, S. Harous, and B. Belkhouche,** “A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis,” *Res. Comput. Sci.*, vol. 110, no. 1, pp. 71–83, 2016, doi: 10.13053/rcs-110-1-6.

[17] **V. Jakkula**, “Tutorial on Support Vector Machine (SVM),” *Sch. EECS, Washingt. State Univ.*, pp. 1–13, 2011, [Online]. Available: http://www.ccs.neu.edu/course/cs5100f11/resources/jakkula.pdf.

[18] **Y. Ahuja and S. Kumar Yadav,** “Multiclass Classification and Support Vector Machine,” *Glob. J. Comput. Sci. Technol. Interdiscip.*, vol. 12, no. 11, pp. 14–19, 2012, [Online]. Available: https://globaljournals.org/GJCST_Volume12/2-Multiclass-Classification-and.pdf.

[19] **J. Han, M. Kamber, and J. Pei,** “Data Mining : Concepts and Solution Manual,” *Data Min. Concepts Tech. Solut. Man.*, p. 135, 2012, [Online]. Available: https://moam.info/data-mining-concepts-and-techniques-solution-manual_59894d1b1723ddd1695415f9.html.

[20] **S. Kumar Thapa and S. Pokhrel,** “Nepali News Document Classification using Global Vectors and Long Short Term Memory.”