Detecting Types of News Using Hierarchical Machine Learning Model with Text Classification

Subhadeep Chakraborty
Developer and Consultant in Artificial Intelligence
TechLearning

Abstract - **News is one of the important aspects of human life from which they can gather the required information. In the early time, the news have been gathered by readers from the newspaper or from the news channels. With the advancement of technology, Social Media comes into the scenario from where readers can get their required news and many more things. In all cases, News can be of different types which are somehow difficult for the reader to identify. Machine Learning plays an important role here to detect the type of news with the implication of Natural Language Processing to analyze the text and to identify the types. In this research, the type of news has been detected by collecting the Insorts news database from Kaggle. The news texts have been prepared by cleaning and vectorizing with the implication of Term Frequency Inverse Document Frequency and Count Vectorization and the models of machine learning have been applied. In this context, the Hierarchical Machine Learning model has been proposed that combines the selected state-of-the-art models through Stacking Classifiers and Voting Classifiers. With the application of all state-of-the-art models and the proposed model, it has been observed that the proposed model has detected the type of news with the highest accuracy (94.22% using TFIDF and Unigram) which is also seen to be higher compared to the existing approaches**

keywords - **Artificial Intelligence, Machine Learning, Text Analytics, Classification, Feature Extraction, Hierarchical Model, Model Overfitting**

I. INTRODUCTION

News and Importance

News has evolved into a form of communication. People are kept up to date on current events through the news. The evolving troubles at the national or international level, happenings, and global situations are all enclosed in the news (Harcup & O'Neill, 2017). It is critical for staying updated on the world's dynamic environment. News is indeed entertaining and fascinating. The most essential element of news is its capacity to help people become more informed and updated. News can help people get aware of the surroundings, government announcements or schemes. Thus, the importance of news in the everyday lives of people is huge. Accurate news can help people get updated about various things and at the same time, inaccurate news can spread hostility, tension, stress and conflicts among people.

Different Types of News

News on Automation Industry

There are various types of news. Industrial automation news is information about changes in companies, their manufacturing, additional features innovation, or advancement. This type of news also includes information about the usage of new technology adopted by the autonomous industry.

News on Entertainment

Entertainment news appears to be about films and shows, famous people, music, sports, and other topics that can pique people's interest (Zhang, et al., 2016) . This type of news can enhance the mood of people by giving them the content of entertainment, pleasure or enjoyment.

News on Politics

Political news is influenced by various incidents in the world and events in politics. It includes the appointment of any person to a political organisation, their concepts, and a variety of other democratic facts.

News on Science

Science news is information about new discoveries or developments or various discoveries in everyday life evolving science.

News on Sports

Sports news is associated with sports events, the location of the sports events, the winning or losing group or sportsman, rankings of the sports people, and a variety of news that provide details about sports-related events occurring.

News on Technology

Technology news is concerned with the advancement of new technology. This type of news would provide information on existing or recently developed technology (Tian & Wu, 2018).

World News

World news is information about events occurring in various countries around the world. World news includes democratic, institutional, cultural, environmental, social, and technological news from countries all over the world.

Impacts of News on Human Life

News has a huge impact on the lives, thoughts, and opinions of people. The news can make lives better by giving people knowledge about various things. News can boost confidence, respect or self-efficacy. Reading the news can have a significant impact on a person's beliefs, thoughts, and opinions. News can have a big impact on the opinions and beliefs of people. It can shape the thought process of people. People can think innovatively when they can new information through news content (Mamun & Akhter, 2018) . Positive news increases optimism and faith generating opinions among people. This can help a person become more solution-oriented and help them form positive opinions. Bad news can have a big impact on emotions. It has the potential to increase the risk of people developing depression and stress. Bad news can also generate wrath and hostility against people, communities or organisations (Suleymanov, et al., 2018).

II. PROBLEM STATEMENT

News provides people with daily updates around the world. With the advancement of digital technology, most people use to access social media or blogs to access news articles rather than prefer the newspaper (Güven, et al., 2019) . In those blogs, different types of news may arrive such as sports news, news relating to entertainment, political news etc. Most people use to navigate the news on social media or blogs and read the news articles based on the news types by heading (Mahmud, et al., 2022) . However, the heading of the news may not be enough to identify the news type. The content of the news in the news body is essential to identify the type of news. The general readers may not identify the types of news from blogs or social media and randomly read the articles by scrolling through the window (Bhattacharjee, et al., 2017) . In this project, Natural Language Processing and Machine learning will be applied to detect the types of news that have been observed in news blogs or social media. In this context, Natural Language Processing will be applied to process the text of the news and analyze those. On the other hand, machine learning will be applied to detect the types of news (Kanagavalli, et al., 2022).

III. EXISTING WORKS

Text Feature Extraction Process

The main focus of this research paper Dalaorao et al. (2019) was to identify the vacancies related to Term frequency-inverse document frequency (TFIDF) to deal with single terms. The main reason behind this is that the single terms sometimes become vague. A single term can be utilized for indexing. A single term is also general and does not have the potential to differentiate different terms. The focus of this research paper is to use Time frequency-inverse document frequency using collocation as the term features. Depending on the determination of parts of speech the collocated terms can be extracted. Three document classifiers are related to traditional and modified Term frequency-inverse document frequency. Multinomial NB, support vector machines, and random forests are the three classifiers. The experiment has demonstrated that the integrated collocation to enhance the term frequency-inverse document frequency process has surpassed the traditional TFIDF process by increasing the performance rate by 10% leading to 85.9% accuracy.

According to Rahmah et al. (2019), Technology-enhanced learning or TEL can cover a huge spectrum of discussion. In this article 40 technology enhances the learning articles by having processes that have been taken from different research databases. Luhn's significant words have been implemented as a qualitative approach in previous research works. Term frequency-inverse document frequency weight has been applied in this research paper to dressing different issues. 23 key terms have been produced in this research paper from 685 term frequency-inverse document frequency important words. The research paper has demonstrated that some words appear in the highest term frequency-inverse document frequency weight cluster.

According to Lei (2020), the research paper has focused on studying the optimization of the short text classification approach of word2vec algorithms for improving calculation accuracy and text mining degree of a short text in classification methods. To address the problems of inadequate features extracted by conventional text sign extraction approaches, the research paper has constructed a feature vector space of text using word vector models of distributed semantic expressions. Also, it has used word2vec algorithms for training word vector databases and extracting the feature vectors with the help of statistical language models. The paper has verified the practicality and reliability of the algorithms recommended in this research paper with the help of empirical analysis. The result has shown that the word vector semantic approach using distributed semantic expression has the capability to develop feature vector space of short text minimizing the probability of gradient explosion. A statistical language model can also calculate the feature vectors of a word in the short text. The result has also identified that the word2vec algorithm can improve classification performance, calculation accuracy, and text mining degree. The accuracy level of the method is more than 80%.

According to Tian & Wu (2018), TF-IWF weighted word2vec model has been proposed in this research paper as the feature extraction method for resolving the issues of ignoring the significance of worlds and missing semantic relations between multiple words in the emotional analysis of microblogs. The research paper has also been used to support vector machines for getting positive and accurate outcomes. In order to calculate the weightage of the words, TFIWF has been used and in order to calculate the words vector, Word2vec has been used. The collected information has been classified and trained with the help of support vector machines. The result has found that the performance, recall, and precision of the proposed method are improved. TF-IWF weighted word2vec has achieved an 89.4% accuracy level.

As per Yue & Chen (2015), an automatic visual bag of words for online robot mapping and navigation is the main focus area of this research paper. Various machine learning algorithms have been proposed in this research paper to implement a robot mapping and navigation process. Some errors have also been collected in this research paper using this algorithm.

As stated by Anwar et al. (2014), the image-based ancient coin classification system has been proposed in this research paper. Classification of the coins will be done on the basis of identifying the symbols minted on the opposite sides of the coins. Dense sampling depending on the bag of visual words model has also been implemented in this research paper for recognizing the symbols on the coins. Lack of spatial information in the bag of visual words model can minimize symbol recognition rates because the symbols possess some particular geometric structures. The coins are imaged under multiple rotations that result in severely rotated symbols. A novel bag of visual words model has been proposed for the classification of symbol-based coins using the spatial arrangement of visual words in a rotating manner. The model has outperformed the traditional bag of visual words model and the recommended angles histograms of pair-wise identical visual words model.

News Classification

According to Gürcan (2018), this research paper has used a classification model to classify Turkish text using supervised machine learning algorithms. Using this model classification of different news attacks including economy, politics, technology, health, and sports can be performed. Support vector machine decision tree, k nearest neighbour, Multinomial Naive Bayes, and Bernoulli Naive Bayes is used as classification models to classify Turkish news text. Multinomial Naive Bayes has achieved 90% accuracy to classify the Turkish language on different social media communication channel platforms.

According to Alsukhni (2021), Multi-label text classification refers to a natural extension process of classifying texts. The documents are assigned as per the set of levels for the classification of the text. A natural language processing system has been implemented for this purpose to modify, manipulate and comprehend natural languages with the help of computer systems. Due to the under-resourced structure of the Arabic language, it becomes difficult to classify Arabic text. The purpose of this research paper is to develop a model for classifying Arabic text in Arabic news to help the users get the news as per their interests. This paper has demonstrated the effectiveness of deep learning models to resolve the problems of the Arabic multilevel text classification process. Recurrent Neural Networks or RNN, and Multilayer perceptron or MLP employing Long Short Term Memory are used in this research paper. The result has demonstrated that long short-term memory has achieved 82% accuracy and Multilayer perceptron has achieved 80% accuracy.

As per Noppakaow & Uchida (2019), the main purpose of this research paper is to analyze automatic models for the classification of Thailand-based online news articles. 6000 news articles have been used as the data set in this research paper. These news articles have been classified into different categories including sports news, crime news, entertainment news, and political news. The main classification algorithms used for classifying news articles include support vector machines, decision trees, and deep learning methods used in this research paper. The result has highlighted that the decision tree has become able to achieve 86% accuracy. 94% accuracy has been obtained by support vector machines. On the other hand, the highest accuracy of 95% is achieved by deep learning models.

According to Rao & Sachdev (2017), it has become urgent and important to classify different news as per the requirements of people because of a huge influx of multiple news associated with different lifestyles of people. People have become more conscious of getting news about every happening, surroundings, and event. This research paper has used a model for this purpose to classify the news articles based on the needs and requirements of the people depending on the different cities. A web crawler for content extraction using HTML pages of news articles has been constructed in this research paper. The various classifiers used for this purpose include Naive Bayes, random forest, and support vector machine. The result has demonstrated that these machine learning techniques can help in achieving the goals of the research paper to improve the accuracy of resolving the issues of news classification. The random forest performed best with the highest F1 score of 85.20%.

According to Ilyas et al. (2021), numerous sources of online news articles have people get news updates about various facts and happenings. Classification of innumerable data that are produced on a regular basis has become a difficult task. Natural Language Processing as well as machine learning models have been used in this research paper to address this challenge. The main purpose of this research paper is to classify Pakistani news that is collected through open data Pakistan database. Random forest, Logistic regression, Naive Bayes, and support factor machines are used as classifiers in this research paper. The support vector machine has achieved 97.8% accuracy as a result in the case of single-level classification. On the other hand, Logistic regression has acquired 83% accuracy for classifying multi-level text.

According to Londo et al. (2019), the main purpose of this paper is to analyze and classify the texts of Indonesian news articles. The research has used support vector machine decision trees and multinomial Naive Bayes as machine learning algorithms to classify news articles in the Indonesian language. The support vector machine has achieved 93% accuracy for this purpose.

According to Miao et al. (2018), Text classification is a very important and significant mining technology used for text mining. Depending on mission learning I'll go to them. The text classification process involves multiple processes that include text pretreatment, text representation classification, and training. This research paper has designed a Chinese news text classification model. Support vector machines, K-nearest neighbour, and Naive Bayes are used. SVM has obtained the best performance rate with F1 score of 95.7%.

IV. PROPOSED METHODOLOGY

Methodology

The proposed method for the detection of different types of news is shown below:

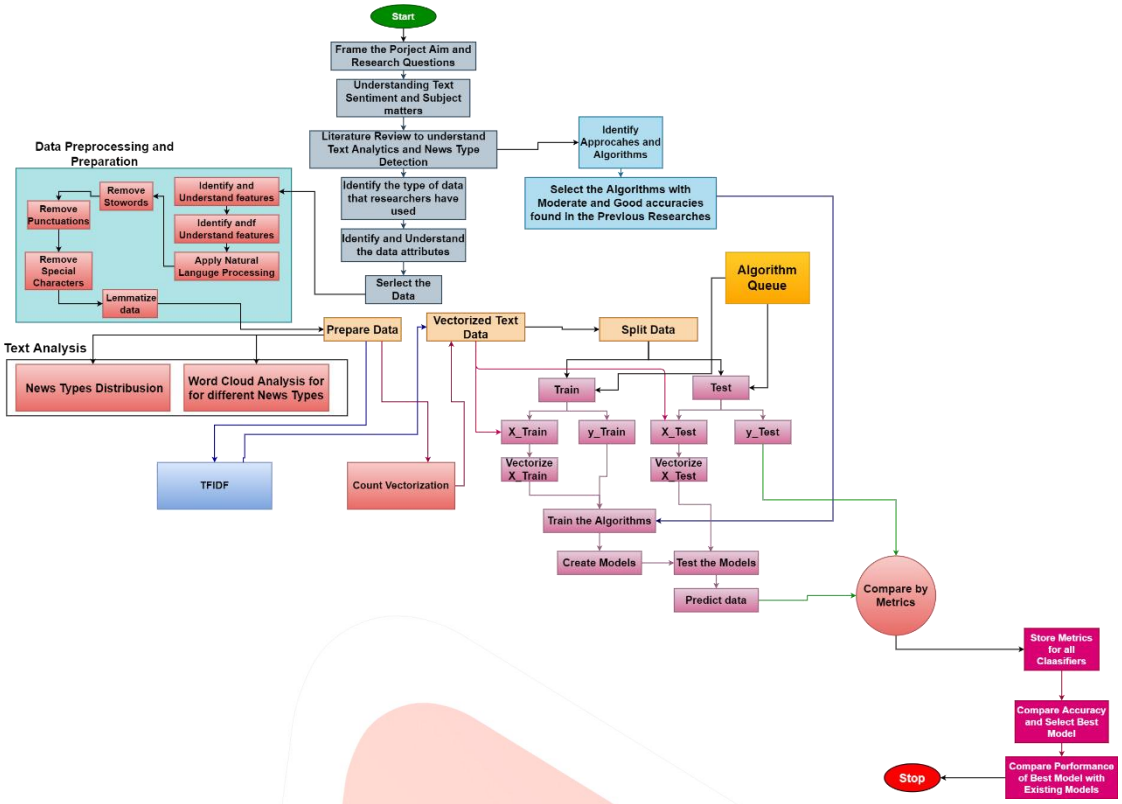

Fig-1: Proposed Methodology

In this process, the text analytics based on the news body will be done using natural language processing. For the purpose of vectorization of the news text, TFIDF and Count Vectorization process will be used. After vectorizing the news text, machine learning classifiers and the proposed model will be applied to detect the types of news. By comparing the result of classification metrics, the most effective model will be chosen for the detection of the type of news.

Data Selection

The dataset for the types of news has been collected from Kaggle (Yadav, 2021). This database contains the records of seven types of news. The data snapshot is shown below:

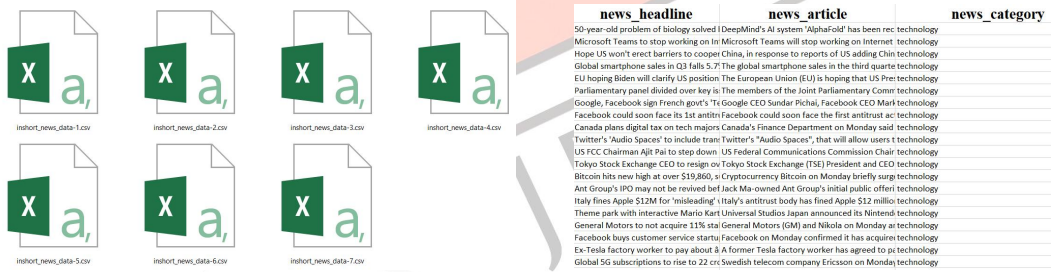

Fig-2: Inshorts News

In the news database, the distribution of the news texts is shown below:

- ✓ World News have a total of 2067 news texts
- ✓ Entertainment News have a total of 2036 news texts
- ✓ Sports News have a total of 1900 news texts
- ✓ Technology News have a total of 1791 news texts
- ✓ Politics News have a total of 1596 news texts
- ✓ Science News have a total of 1437 news texts
- ✓ Automobile News have a total of 1293 news texts

So, in the database, there are a total of 12120 news texts available in the database. These news texts will be used in this research to detect the type of news.

Text Vectorization Process

Term Frequency Inverse Document Frequency

It is one of the widely used vectorization processes for text data. It is a measure of the frequency of the word that appears in the text and the overall document. It is measured by comparing both terms and finding the weights. In this context, the TFIDF has been applied to detect the news types. While applying the TFIDF, the below-mentioned process has been applied to vectorise the news texts. Additionally, at the time of vectorizing the news text, Unigram and Bigram word sequences have been used.

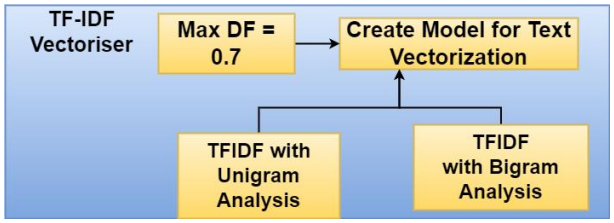

Fig-3: TFIDF Vectorization Process

Count Vectorization Process

This is another widely used process to vectorise text data. Here, it has been applied with Unigram and Bigram to vectorise the news text data. The main difference between TFIDF and Count Vectorization is that TFIDF finds the overall weight of the text in the document, and Count Vectorization finds the count of the tokens in the text document. So, eventually, the weights that have been calculated by TFIDFA may be integers or fractions but the frequency that has been found by Count Vectorizer is always an integer. The process of applying Count Vectorizer is shown below:

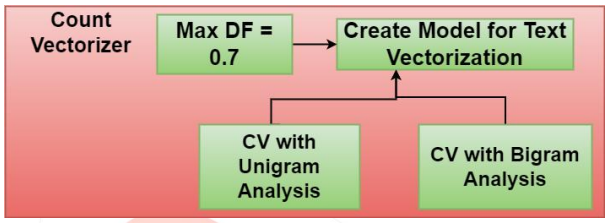

Fig-4: Count Vectorization Process

Selection of Algorithms

The algorithms that have been selected from Machine Learning are generally termed, State-of-the-Art models. By combining all the selected algorithms, the hierarchical model will be prepared.

Logistic Regression

This algorithm belongs to the linear model of machine learning that employs the sigmoid function to classify the data by determining the class labels. It can be used for binary or multi-class detection.

Decision Tree

This model is used to solve the problems of classification and regression. In this research, this model has been used for the purpose of classification of the news types. This model has the root node from where the tree starts splitting. The nodes of the trees are split through branches until the Gini impurity will become zero. The final decision can be obtained from the leaf node where the Gini impurity value is zero.

Adaptive Boosting

This model belongs to the ensemble model and is known as the meta-estimator. The meta-estimator of the model fits the classifier with the extra copies of the model to handle the weights of the incorrectly classified classes. In this context, the single base estimator can be taken as any classifier which suits best to detect the classes most precisely.

Random Forest

This model also belongs to the ensemble family but it differs from the Adaptive Boosting. The base estimator is defaulted by a decision tree. This invokes a number of heterogeneous decision trees (according to the value of n_estimators) which together makes the decision of the data labels. To classify the data, this model applies the majority voting rule.

Design of Hierarchical Model

Now, by combining the four selected models, the proposed hierarchical model has been prepared. To prepare the model, two stacking classifiers have been used. Each stacking classifier will contain two State-of-the-Art models. The outcome of both of the stacking models will be finally inputted into the Voting model to apply the majority voting system. The overall architecture of the proposed model is shown below:

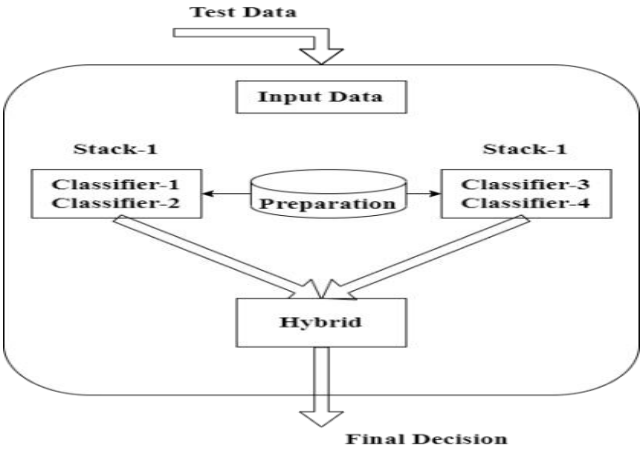

Fig-5: Structure of Hierarchical Model

In the first stacking classifier logistic regression & decision tree and in the second stacking classifier, Adaptive Boosting & Random Forest have been assigned. So, the primary classifi8cation will be done by all four models and the outcomes will be reflected through the stacking classifier. Both two outcomes have been finally to the voting model (with a hard voting process that redirects to applying the majority voting rule) from where the final decisions are obtained. The proposed model is shown below:

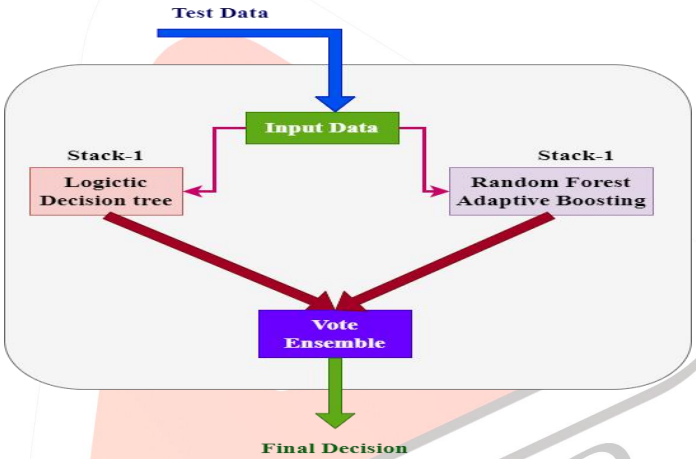

Fig-6: Proposed Model

Evaluation Metrics

The evaluation of the model performances will be done with the application of classification metrics. Hence, four metrics have been chosen for the evaluations which are stated below:

Accuracy

The accuracy can be measured using the following formula:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{Correct\ Prediction}}{\boldsymbol{Total\ Test\ Instances}}$$

-- (1)

Precision

The precision can be measured using the following formula:

$$\boldsymbol{Precision} = \frac{\boldsymbol{Correct\ Prediction}}{\boldsymbol{Total\ Test\ Instances\ Predicted\ in\ a\ Class}}$$

-- (2)

Recall

The recall can be measured using the following formula:

$$\boldsymbol{Recall} = \frac{\boldsymbol{Correct\ Prediction}}{\boldsymbol{Total\ Test\ Instances\ Present\ in\ a\ Class}}$$

-- (3)

F1-Score

The f1-score can be measured using the following formula:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

-- (4)

V. NEWS TYPE DETECTION

News Text Cleaning

The news text has been cleaned initially with the application of natural language processing. In this context, three steps have been taken where the special characters, stopwords and news tags have been removed to clean the text. The flow of news text cleaning is shown below:

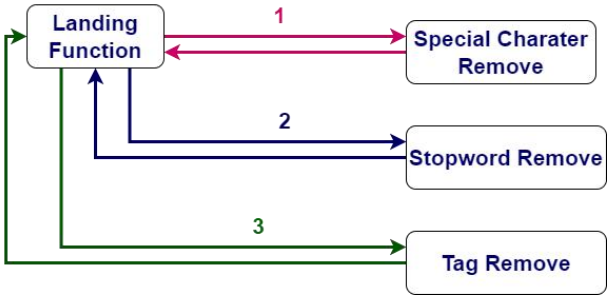

Fig-7: News Text Cleaning

With the application of the above-mentioned process, the news test has been cleaned and the cleaned text is shown below:

news_text	news_category
Bad sign: Mexican Prez after Trump's social me...	technology
Ranbir, Tripti to team up for 'Kabir Singh' ma...	entertainment
SL records 2nd-highest human deaths due to con...	science
IIT-B researchers create AI model to identify ...	science
Found Rahane's captaincy to be fabulous: Ian C...	sports
BMW to build quarter of million more EVs than ...	automobile
Mustang EV's roof doesn't come off: Ford EV he...	automobile
EY plans 9,000 new hires in India in 2021 Glob...	technology
Tesla to route India investment through Tesla ...	automobile
India played good cricket, but in patches: Har...	sports

news_text	news_category
bad sign mexican prez trumps social media acco...	technology
ranbir tripti team kabir singh makers animal r...	entertainment
sl records ndhighest human deaths due conflict...	science
iitb researchers create ai model identify mala...	science
found rahanes captaincy fabulous ian chappell ...	sports
bmw build quarter million evs planned bmw sign...	automobile
mustang evs roof doesnt come ford ev head appa...	automobile
ey plans new hires india global professional s...	technology
tesla route india investment tesla motors amst...	automobile
india played good cricket patches harbhajan in...	sports

Fig-8: Cleaned News Text

News Word Cloud

After cleaning the news text, the word clouds have been visualised to understand the most frequent words and patterns from each of the news categories. The word cloud visualization of the news types is shown below:

Fig-9: News Word Cloud

News Count and Distribution

The distribution of the news text is shown below:

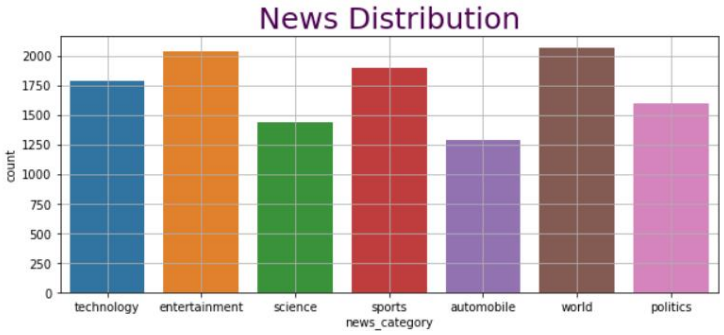

Fig-10: News Distribution

News Vectorization

To prepare the ness text data, the vectorizations have been done using TFIDF and Count Vectorization process. After applying the vectorization process, the total number of features that have been generated is shown below:

Table 1: Features after Vectorization

	Unigram	Bigram
TFIDF	22992	22992
Count Vectorization	137842	137842

News Type Detection Using Logistic Regression

The detection of news types has been done with the application of Logistic Regression and the result of the detection are shown below. It can be seen that model has been outfitted less in the Unigram but higher in the Bigram. Additionally, the accuracy is higher in Count Vectorization compared to TFIDF.

Table 2: Result of Logistic Regression

	TFIDF		Count Vectorization	
	Unigram	**Bigram**	**Unigram**	**Bigram**
Training Accuracy	90.56	95.09	96.17	95.39
Test Accuracy	87.87	88.66	93.61	87.83
Difference (Train-Test)	2.69	6.43	2.56	7.56
Precision (Test)	88.33	89.59	93.59	89.6
Recall (Test)	87.87	88.66	93.61	87.83
F1-Score (Test)	87.92	88.95	93.59	88.33

News Type Detection Using Adaptive Boosting

The detection of news types has been done with the application of Adaptive Boosting and the result of the detection are shown below. It can be seen that model has been outfitted less in TFIDF and the accuracy is higher for the text data in Count Vectorization. The accuracy of the model is seen to be very less. Additionally, the accuracy is higher for TFIDF compared to Count Vectorization.

Table 3: Result of Adaptive Boosting

	TFIDF		Count Vectorization	
	Unigram	**Bigram**	**Unigram**	**Bigram**
Training Accuracy	55.54	42.23	55.2	42.17
Test Accuracy	54.83	42.12	57.67	42.29
Difference (Train-Test)	0.71	0.11	-2.47	-0.12
Precision (Test)	60.11	57.16	63.11	68.3
Recall (Test)	54.83	42.12	57.67	42.29
F1-Score (Test)	53.91	39.79	56.59	40.15

News Type Detection Using Decision Tree

The detection of news types has been done with the application of a Decision Tree and the result of the detection are shown below. It can be seen that model is highly overfitted for both types of vectorization in Unigram and Bigram. Additionally, the accuracy is higher for TFIDF compared to Count Vectorization.

Table 4: Result of Decision Tree

	TFIDF		Count Vectorization	
	Unigram	**Bigram**	**Unigram**	**Bigram**
Training Accuracy	97	97	96.98	96.98
Test Accuracy	91.63	88.94	91.38	89.11
Difference (Train-Test)	5.37	8.06	5.6	7.87
Precision (Test)	91.51	89.39	91.28	89.87
Recall (Test)	91.63	88.94	91.38	89.11
F1-Score (Test)	91.52	89.03	91.29	89.3

News Type Detection Using Random Forest

The detection of news types has been done with the application of a Random Forest and the result of the detection are shown below. It can be seen that model is moderately overfitted in Bigram for both types of vectorization but less in Unigram. Additionally, the accuracy is higher for Count Vectorization compared to TFIDF.

Table 5: Result of Random Forest

	TFIDF		Count Vectorization	
	Unigram	**Bigram**	**Unigram**	**Bigram**
Training Accuracy	97	97	96.95	96.98
Test Accuracy	93.32	89.36	92.95	88.82
Difference (Train-Test)	3.68	7.64	4	8.16
Precision (Test)	93.32	90.18	92.93	90.26
Recall (Test)	93.32	89.36	92.95	88.82
F1-Score (Test)	93.27	89.54	92.9	89.19

News Type Detection Using Proposed Model

The detection of news types has been done with the application of a Proposed Model and the result of the detection are shown below. It can be seen that model is moderately overfitted for both types of vectorization compared to other models. Additionally, the accuracy is higher for TFIDF compared to Count Vectorization.

Table 6: Result of Proposed Model

	TFIDF		Count Vectorization	
	Unigram	**Bigram**	**Unigram**	**Bigram**
Training Accuracy	96.92	96.94	96.97	96.92
Test Accuracy	94.22	91.3	93.69	91.3
Difference (Train-Test)	2.7	5.64	3.28	5.62
Precision (Test)	94.17	91.23	93.68	91.41
Recall (Test)	94.22	91.3	93.69	91.3
F1-Score (Test)	94.16	91.24	93.67	91.3

VI. PERFORMANCE EVALUATION

Comparison of Applied Models

The classification metrics have been compared for TFIDF concerning Unigram and Bigram and it can be seen that the accuracy is highest for the Proposed model which is 94.22% for Unigram and 91.3% for Bigram as shown below:

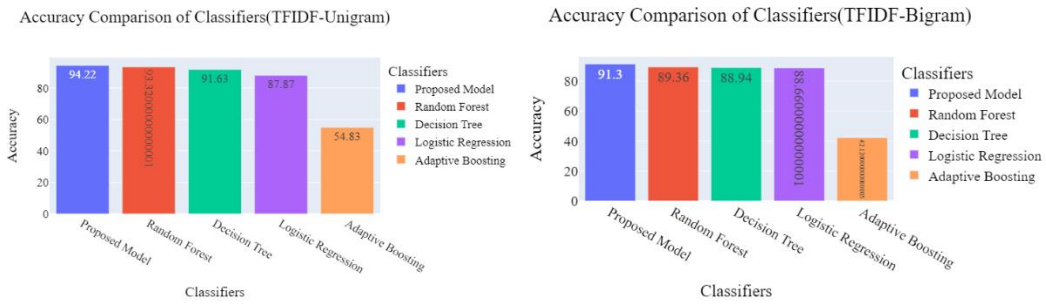

Fig-11: Comparison of Applied Models with TFIDF

The accuracies of TFIDF are shown below for all models which is showing that the Proposed Model has detected the type of news with the highest accuracy for Unigram.

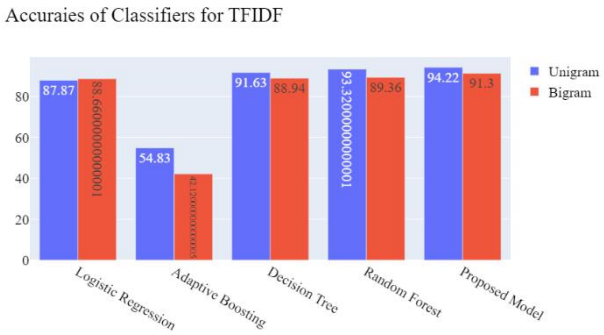

Fig-12: Accuracy Comparison for TFIDF

The classification metrics have been compared for Count Vectorization concerning Unigram and Bigram and it can be seen that the accuracy is highest for the Proposed model which is 93.67% for Unigram and 91.3% for Bigram as shown below:

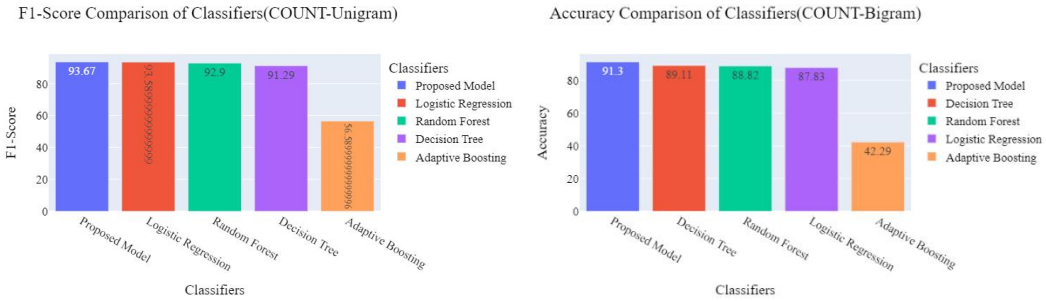

Fig-13: Comparison of Applied Models with Count Vectorization

The accuracies of Count Vectorization are shown below for all models which is showing that the Proposed Model has detected the type of news with the highest accuracy for Unigram.

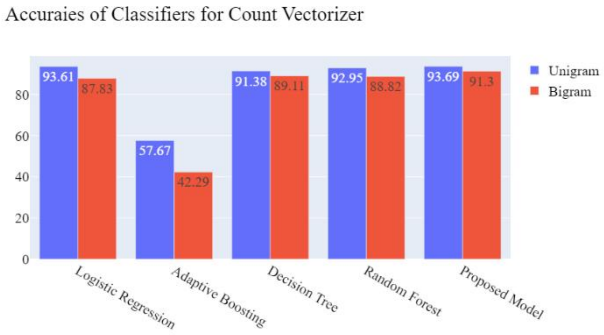

Fig-14: Accuracy Comparison for Count Vectorization

By Comparing the accuracies for TFIDF and Count Vectorization, it can be said that the proposed model has performed better for Unigram using TFIDF with an accuracy rate of 94.22%.

Achievement of Research

The performances of the proposed model have been compared with the existing approaches as discussed in the reviews of existing research. It can be seen that the proposed model has performed better compared to the previous models for the detection of news types. The comparison of the accuracy of the proposed model and the previous models is shown below:

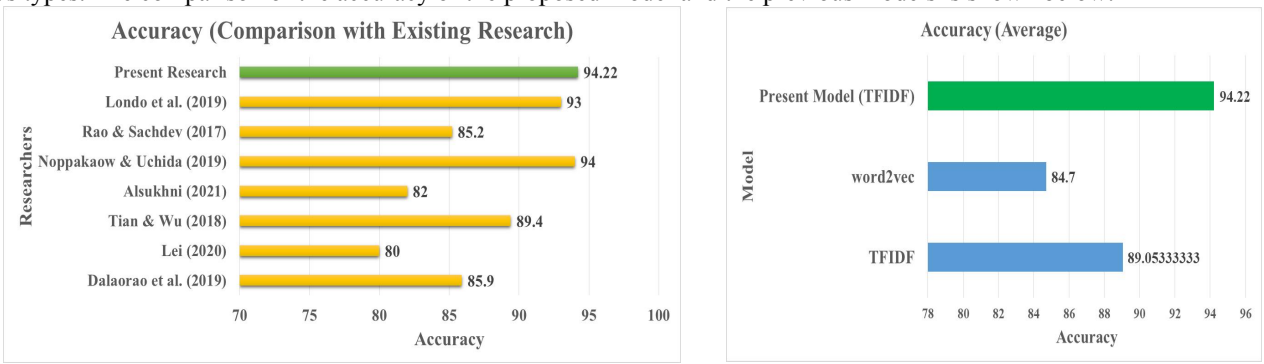

Fig-15: Accuracy (Comparison with Existing Research)

VII. Discussion and Conclusion

News is an important component of society where the reader can gain knowledge by gathering information from it. Besides the knowledge gathering, the detection of the news types is also important which the general readers may not distinguish because of mixed news texts. This research emphasised the detection of news types with the application of natural language processing & machine learning and by proposing the Hierarchical Machine Learning model. The main motivation behind the design of the Hierarchical Machine Learning model is to reduce the model overfitting so that the detection of news type will become more accuracy. With the application of TFIDF & Count Vectorization along with Unigram and Bigram, it has been observed that the Hierarchical Machine Learning model has detected different types of news by TFIDF & Unigram with the highest accuracy (94.22%) which is also higher compared to the existing models.

References

[1] Abainia, K., Ouamour, . S. & Sayoud, H., 2015 . Topic Identification of Noisy Arabic Texts Using Graph Approaches. 26th International Workshop on Database and Expert Systems Applications (DEXA), pp. 254-258.

[2] Alsukhni, B., 2021. Multi-Label Arabic Text Classification Based On Deep Learning. 12th International Conference on Information and Communication Systems (ICICS), pp. 475-477.

[3] Anwar, H., Zambanini, S. & Kampel, M., 2014. A rotation-invariant bag of visual words model for symbols based ancient coin classification. EEE International Conference on Image Processing (ICIP), pp. 5257-5261.

[4] Bhattacharjee, S. D., Talukder, A. & Balantrapu, B. V., 2017. Active learning based news veracity detection with feature weighting and deep-shallow fusion. IEEE International Conference on Big Data (Big Data), pp. 1-6.

[5] Dalaorao, G. A., Sison, A. M. & Medina, R. P., 2019. Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy. IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA), pp. 282-285.

[6] Gürcan, F., 2018. Multi-Class Classification of Turkish Texts with Machine Learning Algorithms. 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1-5.

[7] Güven, Z. A., Diri, B. & Çakaloğlu, T., 2019. Comparison of Topic Modeling Methods for Type Detection of Turkish News. 4th International Conference on Computer Science and Engineering (UBMK), pp. 1-5.

[8] Ilyas, A., Obaid, S. & Bawany, N. Z., 2021. Multilevel Classification of Pakistani News using Machine Learning. 22nd International Arab Conference on Information Technology (ACIT), pp. 1-5.

[9] Kanagavalli, N., Priya, S. B. & D, J., 2022. Design of Hyperparameter Tuned Deep Learning based Automated Fake News Detection in Social Networking Data. 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1-5.

[10] Lei, S., 2020. Research on the Improved Word2Vec Optimization Strategy Based on Statistical Language Model. International Conference on Information Science, Parallel and Distributed Systems (ISPDS), pp. 356-359.

[11] Londo, G. L. Y. et al., 2019. A Study of Text Classification for Indonesian News Article. International Conference of Artificial Intelligence and Information Technology (ICAIIT), pp. 205-208.

[12] Mahmud, F. B. et al., 2022. A comparative analysis of Graph Neural Networks and commonly used machine learning algorithms on fake news detection. 7th International Conference on Data Science and Machine Learning Applications (CDMA), pp. 1-4.

[13] Mamun, A.-A. & Akhter, S., 2018 . Social media bullying detection using machine learning on Bangla text. 10th International Conference on Electrical and Computer Engineering (ICECE), pp. 385-388.

[14] Miao, F., Zhang, P., Jin, L. & Wu, H., 2018. Chinese News Text Classification Based on Machine Learning Algorithm. 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), pp. 48-51.

[15] Noppakaow, A. & Uchida, O., 2019. Examinations on the Performance of Classification Models for Thai News Articles. 11th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1-4.

[16] Rahmah, A., Santoso, H. B. & Hasibuan, Z. A., 2019. Exploring Technology-Enhanced Learning Key Terms using TF-IDF Weighting. Fourth International Conference on Informatics and Computing (ICIC), pp. 1-4.

[17] Rao, V. & Sachdev, J., 2017. A machine learning approach to classify news articles based on location. International Conference on Intelligent Sustainable Systems (ICISS), pp. 863-867.

[18] Suleymanov, U. et al., 2018 . Empirical Study of Online News Classification Using Machine Learning Approaches. IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), pp. 1-6.

[19] Tian, H. & Wu, L., 2018. Microblog Emotional Analysis Based on TF-IWF Weighted Word2vec Model. IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), pp. 893-896.

[20] Yadav, K., 2021. News Classification. [Online]

[21] Available at: https://www.kaggle.com/datasets/kishanyadav/inshort-news

[22] Yue, H. & Chen, W., 2015. Comments on Automatic Visual Bag-of-Words for Online Robot Navigation and Mapping. IEEE Transactions on Robotics, pp. 223-224.

[23] Zhang, D. & Liu, . X., 2021. Text classification of DGCNN model based on deep learning. International Conference on Electronic Information Engineering and Computer Science (EIECS), pp. 622-625.

[24] Zhang, X., Xie, N. & Nakajima, M., 2016. Cleaning Textual and Non-textual Mixed Color Document Image with Uneven Shading. Nicograph International (NicoInt), pp. 136-136.

[25] Zhang, Y. et al., 2020. SocialCCF: Graph-text Collaborative Cleaning Framework Based on Social Networks. IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), pp. 742-747.

[26] Zheng, Y., 2019. An Exploration on Text Classification with Classical Machine Learning Algorithm. International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), pp. 81-85.