STM JOURNALS

Journal of
Computer Technology & Applications

ISSN: 2229-6964 (Online)
ISSN: 2347-7229 (Print)
Volume 12, Issue 3, 2021
DOI (Journal): 10.37591/JoCTA

http://computerjournals.stmjournals.in/index.php/JoCTA/index

Review JoCTA

Unravelling Modern News Classification Methods: A Systematic Review

Aarushi Gupta[1,*], Pulkit Sharma[2], K.C. Tripathi[3], M.L. Sharma[4]

Abstract

Nowadays, the news is being generated each second from every corner of the world, with millions of news articles generated every day. Some assume that at least 1.8 million articles are published yearly, in about 28,000 journals. It has become difficult to recognize what's fake and what's genuine due to the overflow of millions of articles every day. Not every person reads every news, so the classification of news according to people's interests has become significant. 80% of the information is unstructured and "text" being the most common type of unstructured data has led to the emergence of numerous news classification methods. Some of the modern news classification methods stretch from using Machine Learning models like "SVM", "Naive Bayes", "Decision Tree" to the modern complex transformer and deep learning models like "BERT" and "LSTM". So, the main aim of this review is to better understand the pipeline and shortcomings of the news classification processes aforementioned. This article will help researchers to distinguish between various models and choose the one best for their projects.

Keywords: news classification, machine learning, deep learning, data preprocessing

INTRODUCTION

The world is a very happening place and, each second somewhere some incident is taking place. This has led to the generation of numerous news articles each day. We need a way to analyze this information and extract facts and figures from it. In the last decade, various NEWS showcasing platforms, both online and traditional, have emerged. It has become easy for anyone to access news from any part of the world [1]. A significant portion of NEWS occurs in an unstructured textual form whose study can be propitious for various industries. Classification of news is a complex process as it requires cleaning and preprocessing of data, feature selection, models application, and lastly, evaluating the accuracy of these models. With millions of news articles available, it has become difficult for readers to skim through each news article and read the one of their interest. Therefore, the classification of news has become essential in the modern age. Categorization of news helps readers navigate easily through the articles [1]. It helps in sentiment analysis, which helps us define the tone of a particular news category. In recent times, when it's difficult to differentiate between fake and truth, news classification plays an important role in identifying genuine and trustworthy news sources.

In the current time, various supervised Machine Learning methods are being used for various text classification methods like SVM, Naive Bayes, Decision Trees, Random Forest, etc. These classifiers are a part of modern news classification systems because of their high speed and good accuracy they are still used till now. Modern researchers are also

***Author for Correspondence**
Aarushi Gupta

[1,2]Student, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India
[3]Associate Professor, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India
[4]HOD, Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India

Received Date: December 13, 2021
Accepted Date: December 20, 2021
Published Date: December 27, 2021

Citation: Aarushi Gupta, Pulkit Sharma, K.C. Tripathi, M.L. Sharma. Unravelling Modern News Classification Methods: A Systematic Review. Journal of Computer Technology & Applications. 2021; 12(3): 7–17p.

© STM Journals 2021. All Rights Reserved

using complex Transformer models like BERT to gain more accuracy at the stake of slow speed. These advanced models can understand the relationship between the words and thus provide better accuracy while most of the ML models ignore these relationships by treating all words independently.

With so many different classification models present, it's baffling for an individual to look for a model that fits their requirements. Not only the variety of models but the hyperparameter tuning also makes it difficult for the readers to select a good model that fits their needs. Our review aims at bringing ease to the readers by making a detailed analysis of all modern techniques used for the classification of news so far. It will also help researchers to make advancements in the existing technologies or may help them to bring up some new technologies.

Our review paper has the following structure. Section 1 is an introduction section followed by section 2 giving a brief about the different approaches used so far for news classification. Section 3 is a Literature review where a detailed analysis of machine learning and deep learning research papers is done by us. It is followed by section 4 discussing the comparisons and results of the analysis done by us. Finally, section 5 provides a conclusion of our paper.

DIFFERENT APPROACHES FOR NEWS CLASSIFICATION

News classification is mostly performed by two approaches: Machine Learning-based or Deep Learning or Advance Transformer-based models. Machine Learning shows decent accuracy while Deep Learning and advanced transformer-based models show high accuracy but are difficult to train and are slow when compared with Machine Learning models.

Machine Learning approaches can be classified into two types, based on “how the learning is carried out”. These two approaches are Supervised Learning and Unsupervised Learning. The majority of the applications mentioned in the literature review are supervised learning applications. Two sets of annotated data are necessary for both training and testing in supervised learning. Given a text, we try to predict a class based on real-world continuous data related to the text's meaning in a supervised learning framework. Thus, a sentence can lie to any of the listed categories.

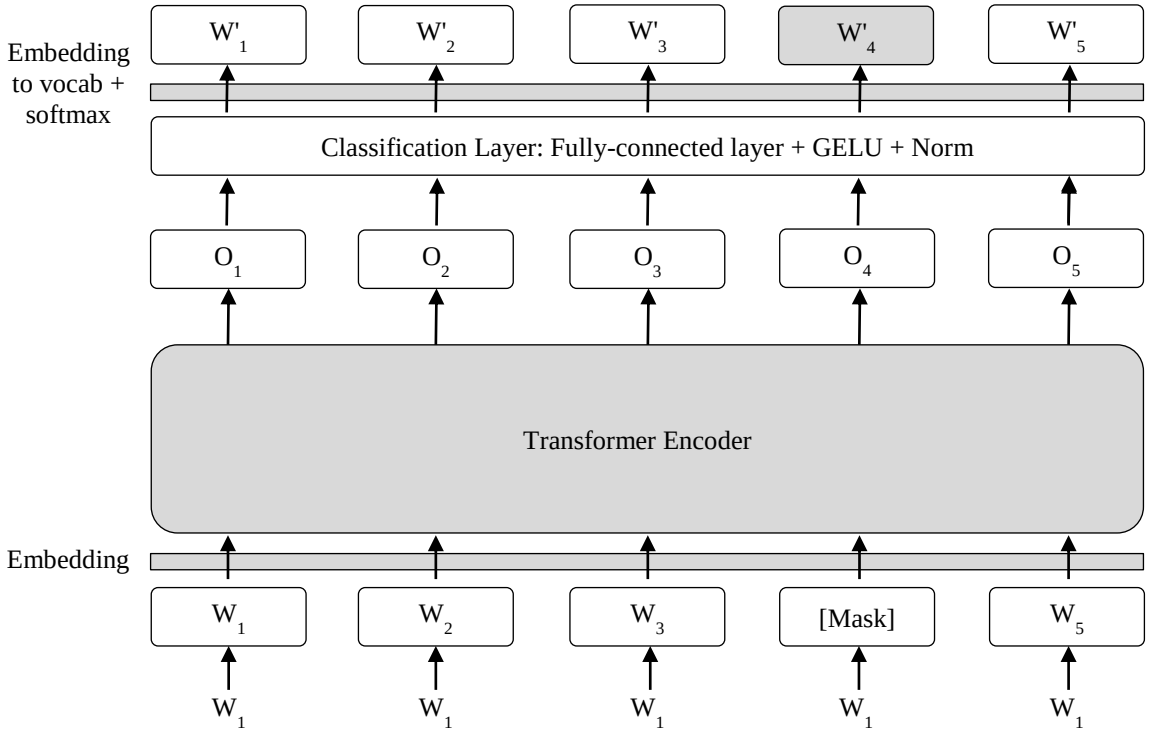

Figure 1. Block Diagram of BERT Classification Model.

© STM Journals 2021. All Rights Reserved

Deep learning is a machine learning technique that allows computers to learn by example in the same way that humans do. A computer model learns to execute categorization tasks directly from images, text, or sound in deep learning. Deep learning models can attain state-of-the-art accuracy, even surpassing human performance in some cases. In Deep Learning, the News classification is mostly performed by RNN models which includes LSTM and RNN layers for training and more complex sequence to sequence transformer models that adopt the principle of "attention" which brings much greater insights from the text and helps in better categorization. Nowadays transformer architecture models have also emerged which are proving to be more efficient and accurate than any other classification model. BERT is an example of a transformer architecture deep learning model. Figure 1 shows a block diagram of the BERT model used for classification.

In both these approaches, qualitative information available in the text needs to be converted to quantitative, so that it can be utilized by the algorithm. Thus, we need to follow a well-defined pipeline that not only converts the textual information to vectors but also reduces the load of models by some techniques like stemming and stopwords removal. Figure 2 demonstrates the text processing pipeline that must be followed while data mining.

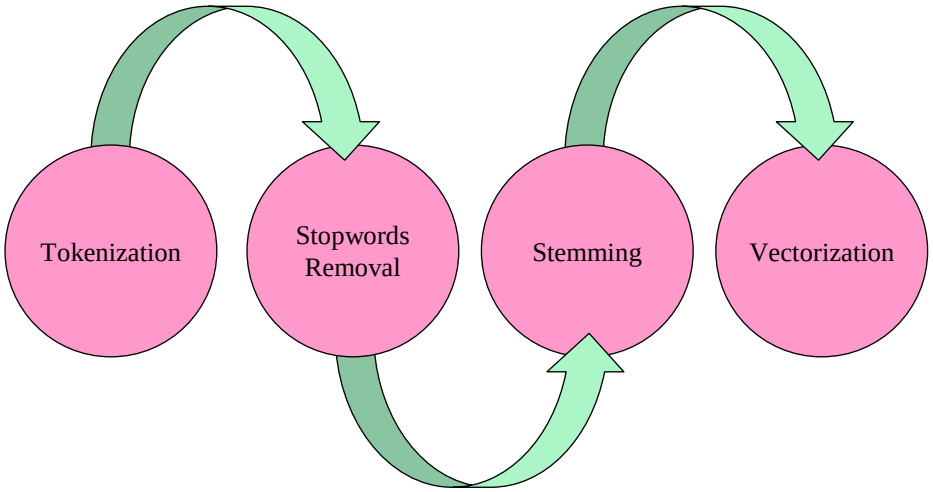

Figure 2. Text Pre-processing Pipeline.

A proper pre-processing pipeline is discussed below:

- **Tokenization:** It includes various steps, first, changing the whole text to one case (lowercase preferably), second is parsing the whole sentence and breaking down the sentence into words without taking care of any persisting relationship that exists among them.
- **Stop Word removal:** This is filtering out Stop Words from the tokenized list of words. The list generally includes words that are used on a very high frequency like a, an, the, were, they, etc. These words have very little importance in a sentence, thus can be removed. It reduces the load for model training as the length of texts gets reduced. But even so, stop word removal is not a necessary step. The reason is simple as many times the text might lose its meaning after stop word removal. E.g., "This TV show is not bad at all!" will be changed to ["TV", "show", "bad"]. Thus, a model will learn the sentence with different meanings. The elimination of stop words is greatly depending on the task at hand and the aim we are attempting to achieve.
- **Stemming:** It will change the word to its base form by removing affixes from it. Like running will be changed to run. This can significantly reduce the vocabulary of the model, thus, in turn, reducing the model's training time also.
- **Vectorization:** This is an important step. It will change the list of words to vectors. Various techniques are being followed for vectorization. One technique is TFIDF, which vectorizes the tokenized sentences to vectors based on the frequency of words. While another approach involves using pre-trained word2vec models like glove50D to get the desired vectors.

© STM Journals 2021. All Rights Reserved

Figure 3 depicts popular machine learning approaches as well as significant pre-processing processes that are widely employed for News Classification. The most commonly utilized techniques are support vector machines (SVM), Naive Bayes (NB), Random Forest (ANN), and Decision Trees (DT). For text mining, according to Khadijah [2], SVM is the most commonly used machine learning algorithm, followed by NB. When it comes to evaluating unstructured text data, other machine learning techniques have received less attention.

Figure 4 depicts popular deep learning approaches that can be used to employ news classification. The most common techniques are to use RNN/LSTM based TensorFlow models or to use advanced seq2seq transformer-based models like BERT.

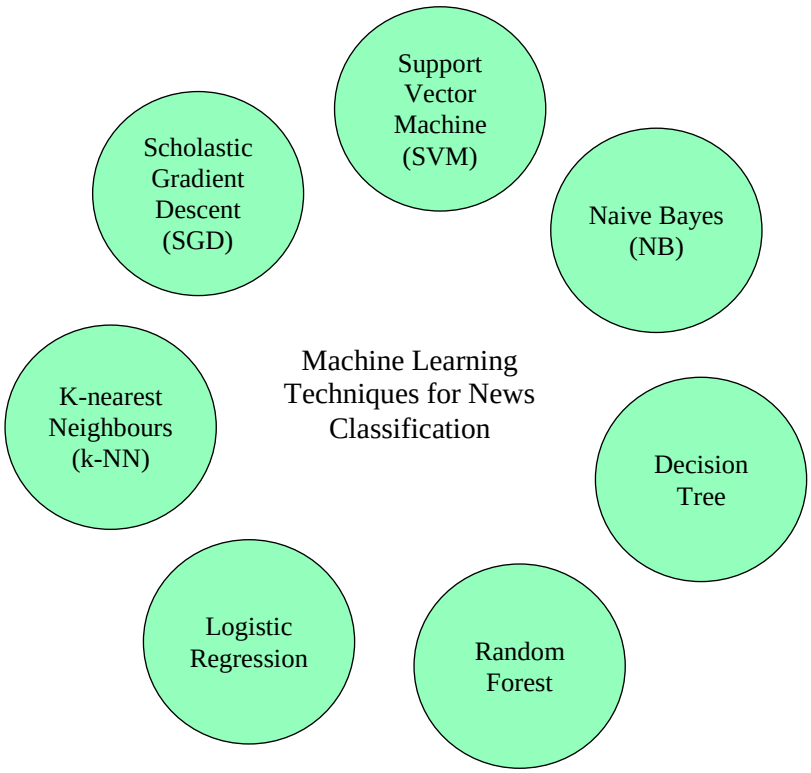

Figure 3. Machine Learning Technique Used For News Classification.

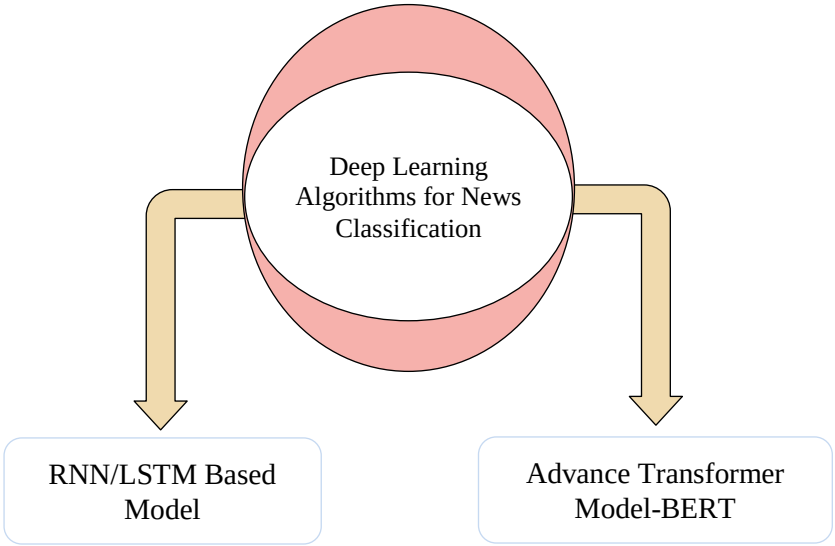

Figure 4. Deep Learning Techniques for News Classification.

© STM Journals 2021. All Rights Reserved

LITERATURE REVIEW

We have studied various research papers on "News Classification", in which the authors have come up with different models (ML-based or DL-Based), different accuracies, different languages (e.g. English, Indonesian, etc.), and different pipelines. Analyzing a variety of research papers will help us to gain new insights that are unknown to the present world and will, in turn, help future developers and researchers. To analyze these research papers, we have categorized them into Machine Learning models and Deep Learning models. We have come up with some parameters, based on which we will differentiate among various models. These parameters are:

- Pre-processing Pipeline
- Dataset used
- Type of model used
- Accuracy gained

Based on these parameters we will analyze the work of various researchers and will compare the work to get the finest ones in both categories. We will further compare the models of both categories to gather some invaluable insights.

Analysis of Machine Learning Model

Machine Learning is the most common approach among researchers for News Classification. We are analyzing different papers whose machine learning models used are one among SVM, naive Bayes, Decision Tree, and Random Forest because of their vast popularity among the researchers. We have collected some Research work from the last decade which uses Machine Learning models for News Classification. We have studied their research work in-depth and analyzed the same in the coming sections.

Rama [3] applied an SVM-based Machine learning approach to build a system that can classify "Financial News" as either "rise" or "drop". It uses a combination of two types of data: Textual past news articles and past intraday price. The system uses TF-IDF for vectorization of input sentences, but no information is provided for stop words removal and stemming. The model was trained on a sample size of 149 records. Their model achieves an accuracy of 83% on the training data. Also, the model's recall and precision for "rise" is 67% and 87% respectively and for "drop" the recall and precision are 93% and 81% respectively. To get a better insight into the model, we calculated the F1 score for the model for better insights. Model's F1 score for "rise" is 0.74 while that for "drop" is 0.86. Even though the model achieved a significantly high accuracy (83%), it is more vulnerable to underfitting due to its low training dataset (149 samples). This may have happened because of the small training dataset used.

Saigal [4] used SVM and its variants for news classification. They have used 3 SVM-based classifiers namely: Twin Support Vector Machine (T-SVM), Least Square Support Vector Machines (LS-SVM), Least Squares Twin Support Vector Machine (LS-T-SVM). All these three models are trained on two different datasets: Reuters-21578 (21578 samples 4 categories) and 20 Newsgroups (approx. 20000 samples 4 categories). The researchers have followed a proper pipeline with tokenization, stop word removal, stemming, and vectorization. They have used TF-IDF for creating the word embeddings. For the Reuters-21578 dataset, the accuracy and F1 score for T-SVM is 86.67% and 0.75, for LS-SVM it's 96.72% and 0.96 and for LS-T-SVM its 98.21% and 0.98. For the Newsgroup 20 dataset, the accuracy and F1 score for T-SVM are 83.33% and 0.81, for LS-SVM it's 90.10% and 0.98 and for LS-T-SVM it's 92.96% and 0.98. All the 3 classifiers have shown exceptional results, but among them also LS-SVM and LS-T-SVM showed state-of-the-art performance. The LS-T-SVM was not only the most accurate model among the three but also the fastest.

Liliana [5] tries to perform news classification in the Indonesian Language. They have used SVM to train their model. The dataset corpus consists of 180 document articles, out of which 70% (126) is

© STM Journals 2021. All Rights Reserved

used for training and the remaining 30% (54) is used for testing. The model is trained for different values of c and gamma, to get the best set of hyper-parameters. The researchers have followed a proper pre-processing pipeline: Tokenization, stop words removal, stemming, and TF-IDF vectorization. The model achieves the highest accuracy of 85% with c = 110 and gamma = 1. The researchers observed the effect of c and gamma on the model's accuracy. Accuracy increases with the rising value of c, while the trend is opposite in the case of gamma. Even though the model seems to have good accuracy (85%), it is more vulnerable to underfitting due to its low training dataset (126 samples). The researchers have not specified the confusion matrix to confirm the same.

Yuliani [6] have used 6 different classifiers: Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree, Logistic Regression, Stochastic Gradient Descent (SGD), and Neural Network-MLP (Multi-Layer Perceptron), to classify the news into "Hoax" and "Non-Hoax". The dataset corpus consists of 1486 news samples. The researchers have followed a proper pipeline with tokenization, stop word removal, stemming, and one-hot-encoding (bag of words) for creating word embeddings. The accuracy and F1 score for Naive Bayes is 88% and 0.88, for Decision Tree it's 92% and 0.92. For Logistic Regression, it's 91% and 0.91, for SVM it's 93% and 0.93, for SGD it's 89% and 0.88 and for NN-MLP it's 93% and 0.93. Evidently, most models performed well for this binary classification. But among them also there were some exceptional performers, SVM and NN-MLP. The SVM performed toe to toe with NN-MLP and delivered great accuracy.

Mangal [7] uses Naive Bayes for news classification in the Punjabi Language. The researchers have not provided any information about the training corpus size and pre-processing step. They have used a Multinomial Naive Bayes classifier to train the model. The model is tested on a separate dataset collected from various classifiers. The model got an average accuracy of 72% and an F1 score of 0.74. The model achieved significant accuracy considering the scarcity of information available in the Punjabi language. The model performed well taking these factors into account.

Jehad [8] uses Random Forest and Decision Tree for Fake news detection. Both the models are trained on a dataset with 20,761 records. The model classifies the news into one of "fake" and "real". The researchers have followed a proper pipeline with tokenization, stop word removal, stemming, and vectorization. They have used TF-IDF for creating the word embeddings. Accuracy and F1 score for Decision Tree are 89.11% and 0.88 while that of Random Forest is 84.97% and 0.85. Surprisingly Decision Tree performed better than Random Forest but due to the randomized feature selection in Random Forest, it can generalize over the data in a better way, thus making it suitable for large datasets. The researchers also emphasized the importance of the data pre-processing step by training the same data on the unprocessed text which gave less accuracy in both the models, with 78% for Decision Tree and 73% for Random Forest.

Fanny [9] compared k-Nearest Neighbor (k-NN), Naive Bayes (NB), and Support Vector Machine (SVM) for news classification. The dataset with a sample size of 30 documents, is collected from various sources and divided into 6 categories. The researchers have followed a proper pipeline with tokenization, stopword removal, stemming, and used TF-IDF for creating the word embeddings. The researchers trained the 3 models on the same dataset. The Accuracy achieved for the k-NN is 86.11%, for Naive Bayes it's 72% and for SVM it's 86.11. The Naive Bayes gave a stable performance, while surprisingly k-NN's performance remained toe to toe with SVM. The researchers also emphasized "stemming" as they trained all three models on 3 different types of stemmers (Porter, Lovin, WordNet). Porter and Wordnet stemmer produced better accuracies with all the 3 models when compared to Lovin.

Analysis of Deep Learning Models

Various natural language and deep learning models have been created till now for classifying news. Not all of them are perfect, but we have come very far from where we started. Nowadays, Deep

© STM Journals 2021. All Rights Reserved

learning classification models are the new trend as they provide more accuracy. LSTM models are proving to be efficient in storing and analyzing the data present in long sentences [10]. Neural network models were explored because they perform huge calculations very easily on a large number of inputs [11]. In [12, 13], a medical dataset having health information classes was trained and tested using a neural network, and the resultant accuracy was 15% higher than achieved by the rest of the models at that time.

The first neural model was proposed by Bengio et al. [14] in 2001, based on the feed-forward neural network trained on 14 million words. But this very first neural network model was not the best and was outperformed by many other classification models of that time [15]. Since then many deep learning models have emerged which has allowed researchers to increase the efficiency of their work.

Ghosh and Shah [10] use a simple feed-forward neural network to classify news articles into three categories: "Fake", "Legit" and "Suspicious". The dataset was made by combining 5 datasets: Kaggle Fake News (2016), Fake News Challenge Dataset (2015), LIAR (2017), and University of Washington Fake News dataset. For the veracity-based model, they retrieved articles using the TF-IDF model. The hyperparameter k was chosen to be 10. The model was evaluated using various values of k, to observe its effect on the model. When the value of k was increased, the recall rate improved but precision suffered. The accuracy of ternary classification is 67.1% and of binary classification is 72.12%. For the style-based model, bidirectional LSTM architecture was used. The accuracy for the style-based model is 81.83%. The combined accuracy of the model came out to be 82.4%. The researchers emphasized their dataset collection as they categorized it into two types based on the length and structure of the sentences.

Ramdhani [16] performed news classification in the Indonesian language using CNN. The data was collected from several online news websites such as CNN Indonesia, Merdeka.com, Aneka news, DialySocial.com, etc. The dataset was labeled into four categories: "hiburan", "olahraga", "tajuk utama" and "teknologi". Researchers followed a proper pipeline with exploratory data analysis, tokenization, stopword removal, stemming. The EDA analysis showed that the maximum length of news data found in the dataset is 1929. They used Porter stemmer for stemming and a collection of Indonesian stopwords for stopword removal. There were a total of 472 news texts, 377 were used for training and 95 for testing. 5 models were produced by the CNN algorithm and they each had 10 epochs. The number of hidden layers in the model are 100, 200, and 300. The experiment resulted in 90.74% accuracy. The researchers observed how the number of hidden layers affects the accuracy of the model. It was detected that increasing the number of hidden layers led to a slight decrease in the training process accuracy but the loss value of the training process was increasing. More hidden layers gave a more accurate model. The result is visualized through GloVe embedding. They have visualized their result through a confusion matrix.

Monti [17] uses geometric deep learning to recognize propagation patterns in fake news. They used the combined set of 1084 fake and true claims spread on Twitter in the years 2013 to 2018. They grouped the features into 4 categories: "User Profile", "User Activity", "Network and Spreading", and "Content". With a 92.7% ROC-AUC score, the model has very high accuracy. They emphasized a lot on their data collection because in fake news classification it's difficult to collect a sufficiently large and reliably labelled dataset. Researchers detected fake news in two ways: URL-wise and Cascade-wise. The network was trained for 25×10^3 and 50×10^3 iterations, respectively. The result was visualized using a ROC curve between false positive rate and true positive rate. The performance achieved by model for both settings is $92.70\pm1.80\%$ and $88.30\pm2.74\%$, respectively.

Abdullah et al. [18] used a bimodal CNN and LSTM to classify news based on its source and previous history. The dataset used is Fake News from Kaggle which has 12 categories. They used three features: "Domain", "Authors", and "Title". The model consists of 4 layers, 2 CNN layers, 1

© STM Journals 2021. All Rights Reserved

LSTM layer, and 1 Fully-Connected layer. As for their preprocessing process, they converted capital letters to small, tokenized their data, and then applied padding on the data, lastly followed by one-hot encoding. They could have preprocessed their data using stemming and stopword removal to acquire a more accurate model. The model achieved 99.7% and 97.5% accuracy for training and testing, respectively. They visualized the performance of the model through a confusion matrix. Though they achieved high accuracy, they should have used a much larger dataset to find out the genuine accuracy of the model.

Nugroho [19] classifies large amounts of news topics using BERT architecture. They used pre-trained models for classification. They used the AG dataset for the corpus of news articles. The dataset has 120,000 training and 7,600 testing samples. They used a learning rate of 1e-4 and set the warm proportion as 0.1. They tested two models. First, BERT without Spark NLP gave an accuracy of 91.87% and took an average of 25 minutes for training. In the second model, they used the pipeline from Spark NLP and added it to the BERT pre-trained model. It achieved an accuracy of 86.65% with the duration of training of 9 minutes. So it was observed that BERT without Spark gave a higher accuracy but was less efficient. Since the authors are using the pre-trained models they haven't described the data preprocessing in detail.

González-Carvajal [20] compares the BERT model with the traditional ML NLP approach for text classification. They conducted four experiments to compare how BERT performs against other models. In the first experiment, BERT was compared with Linear SVC, Logistic Regression, Ridge Classifier, Passive Aggressive Classified of, and more. The accuracy for the BERT model is 93.87% which was the highest of all. In the second they compared BERT and H20 AutoML and Bert outperformed with an accuracy of 83.61%. In the third experiment, they changed the language to Portuguese and compared BERT with Predictor (auto_ml). And lastly, they used Chinese text for comparison. It is observed that the language of text does not affect the accuracy of the BERT model a lot. They haven't discussed the pipeline they used and have not visualized any of their results.

Aggarwal [21] uses a fine-tuned BERT model for the classification of news articles. They compare the results of BERT, LSTM and Gradient Boosted tree. The dataset used here is the NewsFN dataset consisting of 6335 items labelled as "Fake" or "Real". Researchers remove outliers and noise from the data to increase the efficiency of the model and then divide it into training and testing sets. For best results, the learning rate is kept as 2e-5, and the accuracy achieved by the BERT model is 97.02% with a ROC-AUC score of 0.99. A second experiment was conducted using the LSTM model which achieved an accuracy of 86.23% with a ROC-AUC score of 0.92. Confusion matrices visualize the result of all the models. This showed that the BERT model gave a higher accuracy compared to other models. They have used a well-balanced dataset and have visualized word clouds for both Fake and Real labels.

Umer [22] uses CNN-LSTM for automated fake news detection. The experiment is done with two different dimensionality reduction approaches, Chi-Square and Principal Component Analysis (PCA). The dataset used is the Fake News Challenge dataset which is divided into 49,972 and 25,413 instances for training as testing respectively. They followed a proper pipeline with tokenization, stemming, stopword removal, and vectorization. The CNN-LSTM with PCA model having an accuracy of 97.8% and F-score of 97.8%, outperforming all the other models.

COMPARISON AND RESULTS

We analyzed that Support Vector Machine (SVM) is the most popular Machine Learning Technique among the researchers. Not only the most popular, but SVM seems to be the most reliable one also. Table 1, clearly shows the accuracy of different Machine Learning models that we analyzed. Some of the variations of SVM (LS-T-SVM and LS-SVM) [4] produced state-of-the-art accuracy. Naive Bayes, Logistic Regression, and Random Forest remained consistent performers and were able

© STM Journals 2021. All Rights Reserved

to achieve decent accuracy. Decision Tree and Scholastic Gradient Descent (SGD) surprised us with their accuracy and low training time. For Generalization for a big dataset, Random Forest and SVM seems to be labelled, great choices. While most of the news classifications were done in the English Language, we also encountered some classification performed in "Punjabi" [7] and "Indonesian" [5] languages. This allowed us to further improve our study by analyzing the effect of languages on news classification. For the "Indonesian" language model performance was very similar to the model trained on the English dataset. But for the "Punjabi" Language, the model performance suffers due to insufficient data to train.

In deep learning models, we observed that the accuracy of each model depends on many factors other than the algorithm being used for classification. Fine-tuning and hyperparameters play an important role in deep learning classification models. As evident from Table 2, the BERT model is the most accurate Deep Learning transformer model which can be used for classification giving us state-of-the-art accuracy. And though LSTM is the most sought-after deep learning classification model for researchers it has its limitations.

Table 1. Comparative Study on Various Machine Learning Algorithms

Technique	Article	Dataset Size	Language	Accuracy
SVM	[3]	149 news articles	English	83.0%
	[4]	Reuters-21578 (21578 samples)	English	**98.2%**
	[4]	20 Newsgroups (20000 samples)	English	92.9%
	[5]	180 articles	Indonesian	85.0%
	[6]	1486 news samples	English	93.0%
	[9]	30 news Articles	English	86.1%
Naive Bayes (NB)	[6]	1486 news samples	English	88.0%
	[7]	------	Punjabi	72.0%
	[9]	30 news Articles	English	72.0%
Decision Tree	[6]	1486 news samples	English	92.0%
	[8]	20,761 records	English	89.1%
Random Forest	[8]	20,761 records	English	84.9%
k-NN	[9]	30 news Articles	English	86.1%
Logistic Regression	[6]	1486 news samples	English	91.0%
Scholastic Gradient Descent (SGD)	[6]	1486 news samples	English	89.0%

Table 2. Comparative Results of Various Deep Learning Classification Models

Technique	Article	Dataset Size	Language	Accuracy
BERT	[20]	5000 reviews	English	93.87%
	[20]	------	English	83.61%
	[21]	6335 news articles	English	**97.02%**
	[19]	127600 articles	English	91.87%
	[20]	------	Portuguese	91.96%
	[20]	------	Chinese	93.81%
Feed-Forward NN	[10]	------	English	72.2%
LSTM	[10]	------	English	81.83%
	[21]	6335 news articles	English	86.23%
CNN	[16]	472 news texts	Indonesian	90.7%
CNN-LSTM	[18]	20,761 records	English	97.5%
	[22]	75,385 records	English	97.8%
Geometric DL	[17]	1084 Twitter claims	English	92.7%

© STM Journals 2021. All Rights Reserved

Due to its sequential nature, it's difficult to perform parallel training in LSTM models. It can also be observed that the BERT model provides a decent accuracy for text languages other than English too, but these discrepancies can be linked to the small dataset size. We can see from the results that LSTM is a consistent performer and is toe-to-toe with CNN classification models. A feed-forward neural network is the simplest and basic classification model and isn't able to achieve higher accuracies as the other models. We were surprised by the accuracy of the models which used the bimodal CNN-LSTM approach for news classification. It was seen that adding an LSTM layer on top of CNN layers helps achieve high accuracy and one model even achieved an accuracy of 97.8% [22]. The geometric deep learning architecture can be seen performing better than the LSTM models.

It's crystal-clear from the above results that the highest accuracy achieved by an SVM [4] model gave us an accuracy of 98.2% whereas, the highest accuracy achieved by the CNN-LSTM [22] model is 97.8%. At first glance, it can be perceived that machine learning models performed better than deep learning models but that's not the case. When observed closely, it can be seen that there is a huge difference in the dataset sizes used by both models: machine learning models mostly are trained on small datasets whereas deep learning is trained on huge corpora. Thus, deep learning models being trained on a lot of data makes itself a better-generalized model. But everything is not as perfect as it seems. In order to achieve great results, deep learning models pay the price with their huge training time. So with a small dataset SVM is the clear choice because of its high accuracy and low training time. Therefore, for a large dataset, the Deep Learning techniques are favourable, otherwise, Machine Learning techniques are a clear choice.

CONCLUSION

In this paper, we surveyed many machine learning and deep learning models, which have been developed over the past decade. These models have significantly improved the results of news classification over the years. We gave you insight into each model which will help you choose classifications models and techniques more easily for future purposes.

REFERENCES

1. S. Kaur and N.K. Khiva, "Online news classification using Deep Learning Technique," International Research Journal of Engineering and Technology, vol. 03, no. 10, pp. 558–563, 2016.
2. A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D.C.L. Ngo, "Text mining for market prediction: A systematic review," Expert Systems with Applications, vol. 41, no. 16, pp. 7653–7670, Nov. 2014, doi: 10.1016/j.eswa.2014.06.009.
3. R.B. Kumar, B.S. Kumar, and C.S.S. Prasad, "Financial News Classification using SVM," International Journal of Scientific and Research Publications, vol. 2, no. 3, pp. 1–6, 2012.
4. P. Saigal and V. Khanna, "Multi-category news classification using Support Vector Machine based classifiers," SN Applied Sciences, vol. 2, no. 3, p. 458, Mar. 2020, doi: 10.1007/s42452-020-2266-6.
5. D.Y. Liliana, A. Hardianto, and M. Ridok, "Indonesian News Classification using Support Vector Machine," World Academy of Science, Engineering and Technology, vol. 5, no. 9, pp. 767–770, 2011.
6. Yuliani and S. Sahib, "Hoax News Classification using Machine Learning Algorithms," International Journal of Engineering and Advanced Technology, vol. 9, no. 2, pp. 3938–3944, Dec. 2019, doi: 10.35940/ijeat.B3753.129219.
7. S.B. Mangal and D.V. Goyal, "Text News Classification System using Naïve Bayes Classifier," An International Journal of Engineering Sciences, vol. 3, pp. 209–213, 2014.
8. R. Jehad and S.A. Yousif, "Fake News Classification Using Random Forest and Decision Tree (J48)," Al-Nahrain Journal of Science, vol. 23, no. 4, pp. 49–55, Dec. 2020, doi: 10.22401/ANJS.23.4.09.
9. Y. Muliono and F. Tanzil, "A Comparison of Text Classification Methods k-NN, Naïve Bayes, and Support Vector Machine for News Classification," Jurnal Informatika: Jurnal Pengembangan IT, vol. 3, no. 2, pp. 157–160, May 2018, doi: 10.30591/jpit.v3i2.828.

© STM Journals 2021. All Rights Reserved

10. A.M. Somsen, D. Langbroek, H. Borgman, C. Amrit, and T.X. Bui, Rerouting Digital Transformations: Six Cases in the Airline Industry. HICSS, 2019.
11. G. Kaur and K. Bajaj, "News Classification using Neural Networks," Communications on Applied Electronics, vol. 5, no. 1, pp. 42–45, May 2016, doi: 10.5120/cae2016652224.
12. M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical Text Classification using Convolutional Neural Networks," Informatics for Health: Connected Citizen-Led Wellness and Population Health, pp. 246–250, 2017.
13. T.B. Shahi and A.K. Pant, "Nepali news classification using Naïve Bayes, Support Vector Machines and Neural Networks," in 2018 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, 2018, pp. 1–5. doi: 10.1109/ICCICT.2018.8325883.
14. Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," Journal of Machine Learning Research, vol. 3, pp. 1137–1155, Mar. 2003.
15. S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," ACM Computing Surveys (CSUR), vol. 1, no. 1, Jan. 2021, [Online]. Available: http://arxiv.org/abs/2004.03705
16. M. Ali Ramdhani, D.S. Maylawati, and T. Mantoro, "Indonesian news classification using convolutional neural network," Indonesian Journal of Electrical Engineering and Computer Science, vol. 19, no. 2, pp. 1000–1009, Aug. 2020, doi: 10.11591/ijeecs.v19.i2.pp1000-1009.
17. F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake News Detection on Social Media using Geometric Deep Learning," arXiv:1902.06673 [cs, stat], Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.06673
18. A. Yasin, M. J. Awan, M. F. Shehzad, and M. Ashraf, "Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory," International Journal on Emerging Technologies, vol. 11, no. 5, pp. 209–212, 2020.
19. K.S. Nugroho, A.Y. Sukmadewa, and N. Yudistira, "Large-Scale News Classification using BERT Language Model: Spark NLP Approach," in 6th International Conference on Sustainable Information Engineering and Technology 2021, Malang Indonesia, Sep. 2021, pp. 240–246. doi: 10.1145/3479645.3479658.
20. S. González-Carvajal and E.C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," arXiv:2005.13012 [cs, stat], pp. 1–10, Jan. 2021.
21. A. Aggarwal, A. Chauhan, D. Kumar, M. Mittal, and S. Verma, "Classification of Fake News by Fine-tuning Deep Bidirectional Transformers based Language Model-eudl," EAI Endorsed Transactions on Scalable Information Systems, vol. 7, no. 27, pp. 1–12, 2020, doi: 10.4108/eai.13-7-2018.163973.
22. M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G.S. Choi, and B.-W. On, "Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)," IEEE Access, vol. 8, pp. 156695–156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

© STM Journals 2021. All Rights Reserved