USING LARGE LANGUGAGE MODELS TO IDENTIFY FAKE NEWS

Pranshu Savani[1]**, Raj Paynik**[2]
[1]Stevens Institute of Technology
psavani@stevens.edu

ABSTRACT

Automated Fake news stratification is the task of assessing the legitimacy of stated facts in news. Due to the increase in consumption of online media through social networking platforms like twitter, Facebook, Instagram etc there has been a major increase in accessible information. The said data may not be validated or credible, subsequently resulting in the dispersion of fake news. Due to this increasing volume of news content, there is a need to single out fake news with the help of computational methods. Automated detection of fake news is tedious to achieve as it requires the model to understand the ambiguity of sentences and how real information is incorporated within them. We systematically perform experiments and compare the task formulations, data sets and solutions that have been developed for this problem. Based on our insights we highlight the differences between the models and their evaluations. Our results show data classification through 3 major approaches: classification algorithm based, Deep Learning based(from scratch) and hybrid/pre-trained DL model based. This study provides academics and professionals, a review of advances in stratification of Fake News for the English language, most reviews focus on comparing 2 or more methods but not on the timeline of the advancements throughout the years.

[1]

1 Introduction

Fake news detection is a problem that has been taken on by large social-networking companies such as Facebook and Twitter to inhibit the propagation of misinformation across their online platforms. Some fake news articles have targeted major political events such as the 2016 US Presidential Election and Brexit [6]. Individuals falsely reported that a golden asteroid on target to hit Earth contains 10 quadrillion worth of precious metals in an attempt to increase the value of Bitcoin [6]. Widespread subjection to fake news can seed attitudes of cynicism, alienation, and inefficacy towards certain political candidates. Fake news even relates to real-world violent events that threaten public safety. Recent events such as the The news disseminated on social media platforms may be of low quality carrying misleading information intentionally. This sacrifices the credibility of the information. Millions of news articles are being circulated every day on the Internet how one can trust the credibility of these articles? Thus identifying such news is one of the biggest challenges in our digitally connected world. Fake news detection on social media has recently become an emerging research domain [4]. Research communities expressed concern about this flood of misinformation and introduced automated fake news identification solutions based on Natural Language Processing (NLP) techniques. Machine learning algorithms are used for classification of unreliable content and analyzing the accounts that share such content. We have used various existential models and classifiers for identification of fake news and predict the accuracy of different models and classifiers. In this project, we compare and evaluate every model in terms of its accuracy of predicting the news as fake and real. We combined different datasets from sources like Kaggle which contains labelled mixed records of real and fake news, using this we performed various exploratory analyses to identify linguistic properties that are present in deceptive content. We discuss the pros and cons, as well as the potential pitfalls and drawbacks of each task. More specifically systematic comparison of their task definitions, data sets, model construction, and performances. We also discuss a guideline for future research in this direction.

2 Related work and dataset

A news article is a sequence of words. Hence in past, many authors propose the use of text mining techniques and machine learning techniques to analyze news textual data to predict the news credibility. Authors [7] proposed a simple approach for fake news detection using naive Bayes classifier and tested it against a data set of Facebook news posts. Performance evaluation of multiple classification algorithms namely Support Vector Machines, Gradient Boosting, Bounded Decision Trees and Random Forests on a corpus of about 11,000 news articles are presented in [5]. Ahmed et al [2] have utilised TF-IDF (Term Frequency-Inverse Document Frequency) as a feature extraction method with different machine learning models. Extensive experiments have been performed with LR (Linear-regression model) and obtained an accuracy of 89 percent. Subsequently, they have shown an accuracy of 92 percent using their LSVM (Linear Support Vector Machine) and an accuracy of 93.5 percent using the black-box method. In one of the studies, Singh et al [16] have investigated with LIWC (Linguistic Analysis and Word Count) features using traditional machine learning methods for classifying fake news.They have explored the problem of fake news with SVM (support vector machine) as a classifier and obtained an accuracy of 87 percent.

In one of the studies, Jwa et al [8] have explored a deep learning approach towards automatic fake news detection using Bidirectional Encoder Representations from Transformers model (BERT) model to detect fake news by analyzing the relationship between the headline and the body text of the news story. Their results improve the 0.14 F-score over existing state-of-the-art models. The word embeddings enable the model to measure word correlation by calculating the distance between two embedding vectors. Neural networks reveal their best performance in many NLP tasks with the pre-trained word embedding [12]. A Neural network architecture based on term frequency-inverse document frequency (TF-IDF) and bag-of-words (BOW) representations as input to a multi-layer perception (MLP) are used for fake news detection in news articles[6].

2.1 Dataset Collection

We combined multiple real-or-fake news data sets from kaggle.com in our experiments to evaluate semantic features. It contains 43759 real news articles and 33497 fake news articles. The individual data sets span through multiple subject domains including politics, sports, world-events etc. The final counts of records were 77256. Different data sets had different feature columns but we chose to keep only the news-title and news-article-text as the input features as they had the highest correlation with our target variable and are most relevant to our analysis. a single news article may contain up-to 2500-3000 words which makes the data set huge considering the fact that we have around 77K records.

3 Our approach

Our study was conducted by using a modified methodology by Kitchenham (2004) [11], following three phases: (1) Planning the review, (2) Execution of the review and (3) Reporting the results, next described.

The first phase boils down to 3 main questions :

Q1. which are the main processes for assessing the credibility of a news article?

Q2. what methodologies are popular or widely known for FND?

Q3. Which methods are relevant and proved to be effective?

The next step is to perform an extensive search for relevant studies and filtering the research papers according to the above criterion. After compiling the searches and the applying inclusion, exclusion criteria(filtering like above) the candidate studies were selected.

Table 1. shows the selected studies. The first column(ID) is the identification number for the paper, the second column(Paper Title) has the title for the paper and the third column(Author) is the reference of the author.

ID	Paper Title	Author
[] [17]	Fake News Detection with Different Models	S. Vijayaraghavan et al.
[4]	Fake News Detection using Bi-directional LSTM-Recurrent Neural Network	P. Bahada et al.
[14]	Fake news detection: A hybrid CNN-RNN based deep learning approach	J.A Nasir et al.
[9]	Fake news detection in social media with a BERT-based deep learning approach	RK Kaliyar et al.
[3]	Fake Detect: A Deep Learning Ensemble Model for Fake News Detection	N. Aslam et al.
[10]	A Hybrid Model for Effective Fake News Detection with a Novel COVID-19 Dataset	RK Kaliyar et al.

Table 1: Candidate studies.

Some papers performed several tests comparing the performance of different approaches and algorithms, testing multiple sets of configuration features, or even different domains. For example, [17] presented 15 different implementations comparing five different ML classification algorithms with several features: corpus size, n-gram size, number of classes, balanced or unbalanced corpus and same or different domain along with different word embeddings. The implementations with the best result per domain were obtained from each of the 6 papers. However, it is necessary to clarify that in [17] three ML implementations with the same performance were identified as with the best results. Therefore, a total of 10 implementations will be analyzed, representing the basic information to answer the research questions.

3.1 Which are the main processes for assessing the credibility of a news article?

3.1.1 Information Extraction

The combined dataset includes reviews, essays, forums, blogs or published news articles extracted from various sources of the internet.

3.1.2 Preprocessing

Within the context of a news article title or text, numbers simply quantify claims and do not change the meaning of the text.Moreover, Em dashes are used in various linguistic contexts like joining independent clauses. They do not add to the meaning of the text, however they are surrounded by two words of different clauses. To mitigate the noise effects and reduce data size by keeping only relevant information the raw data needs to be processed by certain refinements namely lower casing, stop-word removal, punctuation and non-word filtering, stemming, tokenization etc.

- Stop word removal : Stop words are insignificant words in a language that will create noise when used as features in text classification. These are words frequently used in sentences to help connect thought or to assist in the sentence structure. Articles, prepositions, and conjunctions and some pronouns are considered stop words.
- Stemming : Stemming, put simply, is changing the words into their original form and decreasing the number of word types or classes in the data.[cite] For example, the words “Running,” “Ran,” and “Runner” will be reduced to the word “run.”We use stemming to make classification faster and efficient.

3.1.3 Features extraction

One of the challenges of text categorization is learning from high-dimensional data. There is a large number of terms, words, and phrases in documents that lead to high computational burden for the learning process. Furthermore, irrelevant and redundant features can hurt the accuracy and performance of the classifiers. Thus, it is best to perform feature reduction to reduce the text feature size and avoid large feature space dimension.

TF-IDF

TF-IDF stands for term frequency - inverse document frequency which measures how important a given word is for sentence or document in relation to the entire corpus. The term consequence increases with the number of times a word appears in the document, consequently weighing down the high frequency terms while mounting up the occasional occurrences.

In this study TF-IDF is used for vectorizing the input for ML classification algorithms.

$$tf(t,d) = \frac{n(t,d)}{V(d)}$$

$$n(t,d) = \text{ occurrence of the word } t \text{ in the document } d \qquad idf(t,D) = \log\left(\frac{N}{df(t,D)}\right)$$

Glove

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.https://nlp.stanford.edu/pubs/glove.pdf

3.1.4 Fake News Classification

It is the assessment of news orientation in two classes, fake or real. There are 3 main classification levels News text level, News title Level, combined. the classification model accepts news articles as inputs and outputs a binary output whether the article is fake or real. some models may even output the probability of the article being real or fake. many implementations exits which we will discuss in later sections.

3.1.5 Evaluation

The most frequently performed metrics in the papers to contrast results are accuracy, precision, recall and F-measure, with the units: true positive (tp), false positive (fp),true negative (tn) and false negative (fn). These metrics are calculated following the next equations (Kharde and Sonawane, 2016): Accuracy is calculated by dividing the addition of the number of tp, and tn correctly assigned, between the number of texts analyzed (N),

$$\text{Accuracy} = \frac{tp + tn}{N}$$

Precision is the number of tp divided between all positively assigned documents, the addition of tp plus fp,

$$\text{Precision} = \frac{tp}{(tp + fp)}$$

Recall is the number of tp out of the actual positive documents,

$$\text{Recall} = \frac{tp}{(tp + fn)}$$

3.2 What methodologies are popular or widely known for FND?

most papers classify news based on news text alone, as news title may polute the actual content due to the presence of out of context idioms, sarcasm etc. While the most common approaches before were ML algorithms, since 2017 the use of the Deep Learning (DL) approaches seems to be increasing in the research community. There are numerous implementations of different models for this task which take in numerous types of features as input, some are traditional ML classification algorithms like Logistic regression or linear SVM primararily taking tf-idf vectorized inputs, and other are deep learning based models like bidirectional LSTMS, Hybrid CNN's + LSTMs, GRU which work well with word embeddings. Novel proposals have recently emerged which propose classification through Genetic Programming (GP), an evolutionary algorithm. Another interesting proposal is, where Transfer Learning (TL) approach is used. TL aims to find ML methods that retain and reuse previously learned knowledge, capable of learning different multiple tasks simultaneously like BERT.

3.3 Which methods are relevant and proved to be effective?

The following 9 experiments are the closest to answer the above question filtered from the 6 candidate papers listed in Table 2. Note: traditional ml classification has 3 experiments namely, Logistic Regression, SVM with linear kernel and Decision Tree.

ID	Model	Embedding
1-3	Traditional ML Classification	TF-IDF
4	LSTM	trainable WE
5	GRU	trainable WE
6	LSTM	GloVe
7	GRU	GloVe
8	CNN+LSTM	Trainable WE
9	BERT	Pre-trained WE

Table 2: Candidate studies.

4 Experimental Design

All the implementations were carried out in python and run on Google Colab. Deep learning models were created using TensorFlow Keras and HuggingFace. Pandas, Numpy, NLTK and Porter Stemmer are used for reading and preprocessing the data. Pretrained GloVe embeddings were downloaded from Stanford [15]

Each record in the dataset has features id, title, text and target labelled 0 and 1, after preprocessing the models are trained on the data with a 0.2 test split along with cross-validation split of 0.1 on train.

4.1 ML classification algorithms

Logistic Regression with sigmoid activation

Logistic regression is a statistical machine learning algorithm that classifies the data by considering outcome variables on extreme ends providing a discriminatory line between classes.

Support Vector Machine with linear kernel

SVM is a supervised machine learning algorithm in which a hyperplane is created in order to separate and categorize features. The optimal hyperplane is usually calculated by creating support vectors on both sides of the hyperplane in which each vector must maximize the distance between each other.

Decision Tree

Decision Trees are a non-parametric supervised learning methods used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

We fit the above models with TF-IDF vectorized inputs.

Parameter	Logistic Regression	SVM	Decision Tree
penalty	L2	-	-
tol	1E-4	1E-3	-
C	1	1	-
solver/kernel	lbfgs	linear	gini

Table 3: parameters used.

4.2 Bi-directional Recurrent Neural Networks (RNN)

Recurrent Neural Network is a feed-forward artificial neural network. RNNs handle a variable-length sequence input by comprising a recurrent hidden layer whose activation at each time is dependent on the previous time. Hence RNNs are better choice for long-distance contextual information [17].Bi-directional processing is an evident approach for a large text sequence prediction and text classification. A Bi-Directional RNN network steps through the input sequence in both directions at the same time.

Long Short-Term Memory networks

Long Short-Term Memory networks (LSTM) are a special type of RNN competent in learning long-term dependencies [18]. LSTM is a very effective solution for addressing the vanishing gradient problem. In LSTM-RNN the hidden layer of basic RNN is replaced by an LSTM cell. LSTMs help to preserve the error that can be back-propagated through time and in lower layers of a deep network [19].

Gated Recurrent Units

The GRU is the newer generation of Recurrent Neural networks and is pretty similar to an LSTM. GRU's got rid of the cell state and used the hidden state to transfer information. It also only has two gates, a reset gate and update gate. The update gate acts similar to the forget and input gate of an LSTM. It decides what information to throw away and what new information to add. The reset gate is another gate is used to decide how much past information to forget. When comparing GRU with LSTM, it performs good but may have a slight dip in the accuracy. But still we have less number of trainable parameters which makes it worth comparing the trade-off.

Hybrid CNN + LSTM

This model makes use of the ability of the CNN to extract local features and of the LSTM to learn long-term dependencies. First, a CNN layer of Conv1D is used for processing the input vectors and extracting the local features that reside at the text-level.The output of the CNN layer (i.e. the feature maps) are the input for the RNN layer of LSTM units/cells that follows. The RNN layer uses the local features extracted by the CNN and learns the long-term dependencies of the local features of news articles that classify them as fake or real[13]

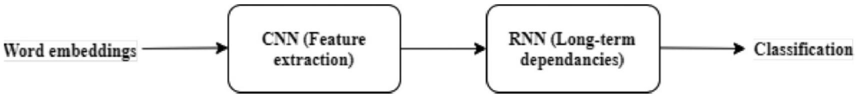


Figure 1: CNN-LSTM hybrid model

we create, 2 Bi-directional LSTM networks and 2 Gated Recurrent Units, where 1 of each is fit on keras trainable embeddings from scratch and the other uses pre-trained GloVe embeddings for vectorizing the input. We can observe the layered architecture for the RNN in Table 4. the proposed implementations share a common architecture to give way for unbiased comparison.

Layer	Output Shape	params LSTM	params GRU
Embedding	(None,None,300)	3000000	3000000
Bi-Directional	(None,None,128)	186880	140544
Bi-Directional	(None,32)	18560	14016
Dense	(None,64)	2112	2112
Dropout	(None,64)	0	0
Dense	(None,1)	65	65

Table 4: Hybrid Layered Architecture.

another model implemented is the hybrid CNN + LSTM stated above, the layered architecture is given in table 5.

Layer	Output Shape	params
Embedding	(None,None,300)	3000000
Dropout	(None,None,300)	0
Conv1D	(None,None,64)	96064
MaxPooling	(None,None,64)	0
Bi-Directional	(None,None,128)	186880
Bi-Directional	(None,32)	18560
Dense	(None,64)	2112
Dropout	(None,64)	0
Dense	(None,1)	65

Table 5: Hybrid Layered Architecture.

Automatic Hyper-parameter tuning was computationally too expensive, so the following hyper parameters were used.

Hyper Parameter	Value
Embedding Layers	1
Hidden Layers	2
Dense Layers	2
Dropout rate	0.3
Optimizer	Adam
Activation	relu
Loss	binary Cross-Entropy
Epochs	10
Batch size	30
Validation split	0.1
Early stop patience	2

Table 6: Hyper Parameters.

4.3 BERT transformer

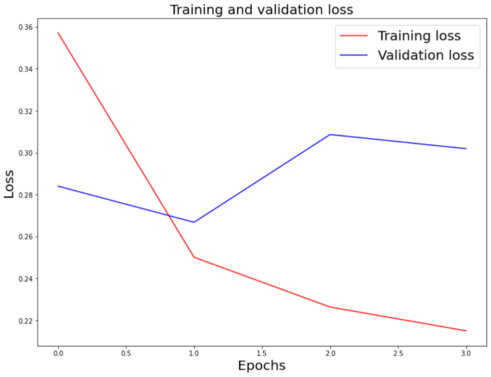
Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weights between them are dynamically calculated based upon their connection. BERT is pre-trained on two different, but related, NLP tasks: Masked Language Modeling and sequence Prediction.

The reason for using BERT in fake news detection is self-attention which is made possible by the bidirectional Transformers at the center of BERT's design. This is significant because a word may change its meaning as sentence develops. The more words that are present in total in each sentence or phrase, the more ambiguous the word in focus becomes.

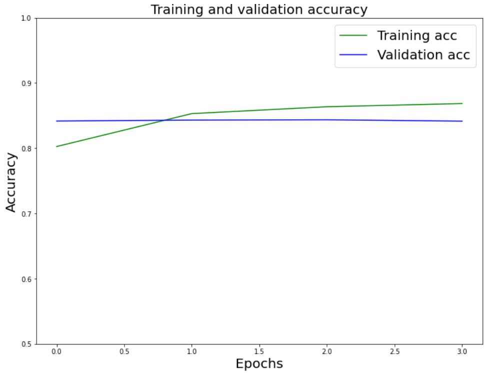
5 Experimental Results

We have investigated and analyzed the results with these models having different types of learning paradigms (different optimal hyper-parameters and architectures). Classification results demonstrate that the capability of automatic feature extraction with deep learning models plays an essential role in the accurate detection of fake news. The classification models give an accuracy of around 81 percent, Decision Tree performs poorly with an accuracy of only 73 percent. Deep learning Models provide a good improvement over the traditional methods with an increase of 2-5 percent in accuracy, in which the Bi-directional LSTM with pretrained GloVe embeddings performs best. The BERT produced more accurate results as compared to any of the above benchmarks with an accuracy of 98.90 percent. The Results and Plots are shown below .

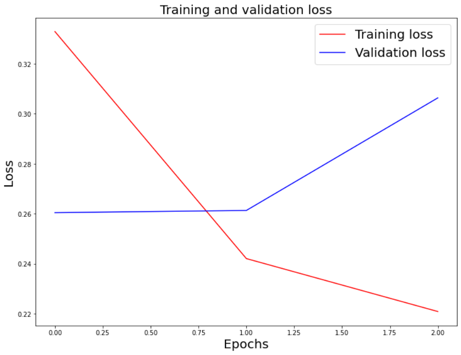
Classifier	Accuracy	Precision	Recall
Logistic Regression	0.81	0.73	0.85
Support Vector Machine	0.81	0.61	0.75
Decision Tree	0.73	0.65	0.78
Bi-Directional LSTM	0.83	0.66	0.95
GRU	0.84	0.68	0.93
Bi-Directional LSTM(GloVe)	0.85	0.70	0.97
GRU(Glove)	0.85	0.75	0.95
CNN+LSTM	0.82	0.65	0.91
BERT	0.99	-	-



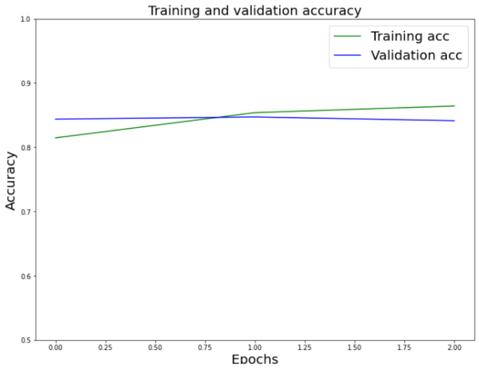
(a) training loss for LSTM



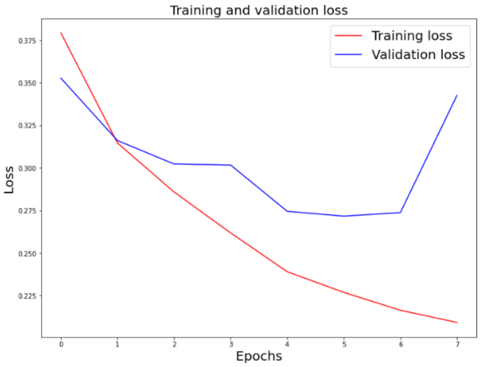
(b) training accuracy for LSTM



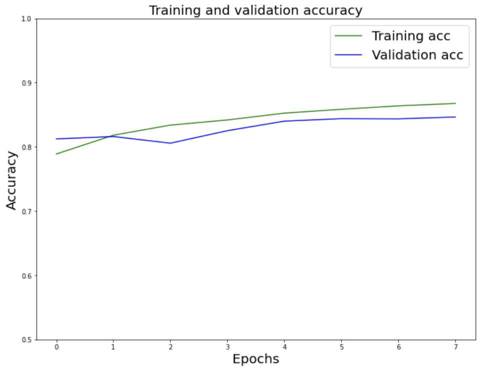
(a) training loss for GRU



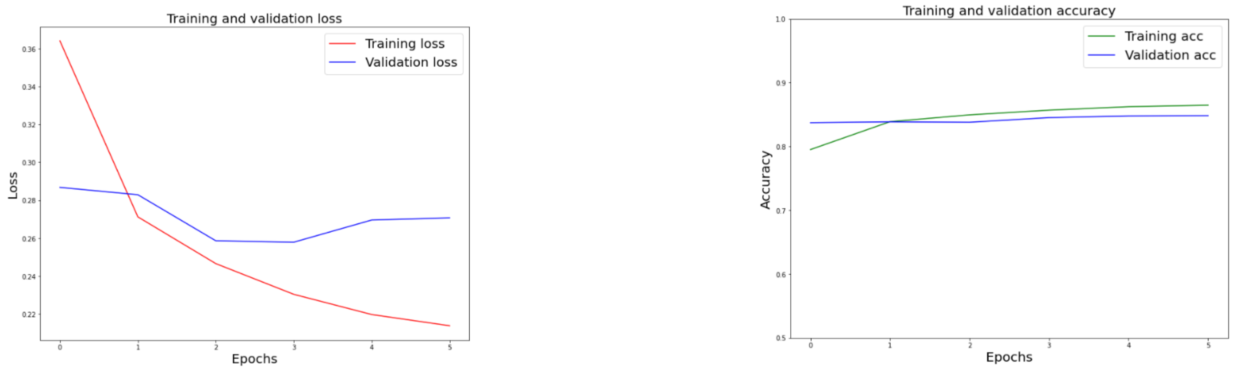
(b) training accuracy for GRU



(a) training loss for LSTM with GloVe



(b) training accuracy for LSTM with GloVe



(a) training loss for GRU with GloVe

(b) training accuracy for GRU with GloVe

6 Conclusion and future work

The aim of this work is to systematically compare candidate experiments undertaken for fake news detection. The accuracy of Logistic regression, SVM, DT, Bi-directional LSTM,GRU standalone, LSTM with CNN and BERT are evaluated and compared. CNN performs better for extracting local and position-invariant features while LSTM-RNN is well suited for a long-range semantic dependency based classification. The results show deep learning models are more effective than traditional models, in which Bi-directional LSTM with GloVe embeddings performs best among the RNN, BERT is best over-all. The experiments also show that the choice of the adaptive learning rate algorithm plays a major role in the output to handle vanishing gradient problem of RNN. The proposed model works well for the balanced and imbalanced high dimensional news data set. More thorough experiments will be required in the future to further understand how deep learning model with attention can help to evaluate the automatic credibility analysis of News.

There is plenty of space for experimentation in observing the nature of fake news to get more insights which will also lead to more efficient and accurate models. We observed that models tend to perform well on specific type of dataset. The dataset is limited to english language and there is need to experiment on other languages as well. Apart from considering CNN and RNN models there is scope to experiment on more complex neural network architectures for future analysis. Old traditional models are also beneficial if they are combined with task specific feature engineering techniques.

References

[1]

[2] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer, 2017.

[3] Nida Aslam, Irfan Ullah Khan, Farah Salem Alotaibi, Lama Abdulaziz Aldaej, and Asma Khaled Aldubaikil. Fake detect: A deep learning ensemble model for fake news detection. *complexity*, 2021, 2021.

[4] Pritika Bahad, Preeti Saxena, and Raj Kamal. Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165:74–82, 2019.

[5] Arvinder Pal Singh Bali, Mexson Fernandes, Sourabh Choubey, and Mahima Goel. Comparative performance of machine learning algorithms for fake news detection. In *International conference on advances in computing and data sciences*, pages 420–430. Springer, 2019.

[6] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism*, pages 84–89, 2017.

[7] Mykhailo Granik and Volodymyr Mesyura. Fake news detection using naive bayes classifier. In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, pages 900–903. IEEE, 2017.

[8] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, 9(19):4062, 2019.

[9] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.

[10] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. A hybrid model for effective fake news detection with a novel covid-19 dataset. In *ICAART (2)*, pages 1066–1072, 2021.

[11] Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.

[12] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[13] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007, 2021.

[14] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007, 2021.

[15] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[16] Vivek Singh, Rupanjal Dasgupta, Darshan Sonagra, Karthik Raman, and Isha Ghosh. Automated fake news detection using linguistic analysis and machine learning. In *International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation (SBP-BRiMS)*, pages 1–3, 2017.

[17] Sairamvinay Vijayaraghavan, Ye Wang, Zhiyuan Guo, John Voong, Wenda Xu, Armand Nasseri, Jiaru Cai, Linda Li, Kevin Vuong, and Eshan Wadhwa. Fake news detection with different models. *arXiv preprint arXiv:2003.04978*, 2020.