Comparative Study of Fake News Detection Using Machine Learning Algorithms and Deep Learning with Embeddings

Ramacharan Reddy Kasireddy

Vellore Institute of Technology, Chennai

Abstract

Fake news has become a major concern in recent times due to its potential to mislead and deceive the general public. To detect fake news using text-based approaches, machine learning models were tested. Four datasets were gathered and processed using NLP techniques with different linear learners and ensemble methods. Deep learning models and embedding techniques were also employed to compare all the models before selecting the best two for each dataset. The top-performing deep learning models were Hybrid CNN + BiLSTM model with Word2Vec embedding and Hybrid CNN + BiLSTM model with FastText embedding, achieving 0.9106 and 0.9087 accuracy on our combined dataset, respectively. The best machine learning model was the SVM model with TF-IDF features, which scored an accuracy of 0.9012. Our findings demonstrate that combining deep learning models with embedding techniques is an effective approach for fake news detection, but it is crucial to choose appropriate datasets for training and testing as the performance of the models can vary depending on the data used. This study contributes towards developing efficient techniques to address the spread of fake news.

Introduction

As the age of information evolves, so does the problem of fake news. The propagation of false information has become a growing concern as it can spread like wildfire through social media with ease. This phenomenon is alarming since it misleads and manipulates the masses on a massive scale, rendering people unable to differentiate what's true from what's not. To remedy this pressing issue, we must develop effective measures for detecting fake news reliably.

Various methods for identifying fake news have been suggested, including text-based, image-based, and user-based techniques. Text-based approaches involve examining the language and content of news articles to detect indications of fake news. The project emphasizes using text-based methods that involve scrutinizing the language and substance of news articles to establish their truthfulness since online access to article texts is readily available.

Machine learning algorithms such as Support Vector Machines (SVM), Random Forests, k-Nearest Neighbors (kNN), Decision Trees, Multilayer Perceptron (MLP), Logistic Regression, Voting Classifiers, AdaBoost, and XGBoost are used to compare the performance of individual learners and ensemble learners.

Deep learning algorithms like Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Convolutional Neural Network (CNN), and moreover we'll integrate newer techniques such as word embeddings using GloVe Word2Vec FastText compare their results against more traditional approaches to ensure maximum accuracy.

Objective

Creating a reliable system that utilizes machine learning and deep learning algorithms to identify fake news is the primary aim of this project. The model's precision and adaptability will be enhanced by utilizing numerous datasets, including LIAR, ISOT, Fake News from Kaggle, and Fake News Detection from Kaggle. Additionally, to enhance the comprehension of text-word associations and optimize the model's effectiveness, embeddings such as GloVe, Fasttext, and Word2Vec will be evaluated. The ultimate objective is to produce a dependable fake news detection system that can promptly and precisely detect phony news.

Methodology

In our proposed framework, as illustrated, current literature was expanded by introducing ensemble techniques with various linguistic feature sets to classify news articles from multiple domains as true or fake. The ensemble techniques and word embeddings used in this research for the combined dataset are the novelty of our proposed approach.

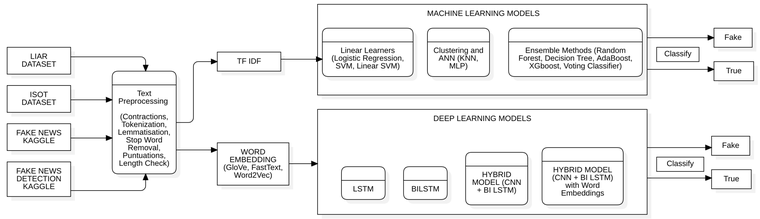

Figure 1. Proposed Methodology.

Results

Figure 2 lists the accuracy of learning algorithms and ensemble methods for five datasets. SVM, MLP, Random Forest, Voting Classifier, and Linear SVM performed best on the ISOT Fake News dataset with a maximum accuracy of 98. Logistic regression scored 97. The average accuracy of ensemble learners is 96.2 compared to individual learners at 95.75. KNN performed best on the Liar Dataset with an accuracy of 76, while MLP was least accurate at 64. Individual learners had an accuracy of 73 while ensemble learners reported a similar accuracy rate. For the Fake News Dataset, individual learners had an average accuracy of 89.75, whereas ensemble learners received an accuracy rate of 88.8. The Combined Dataset showed that SVM had the highest performance at 90, followed by random forest at 89, and KNN with the lowest score at 62.

Classification Model	ISOT`	LIAR	Fake News	Fake News Detection	Combined	Average
LOGISTIC REGRESSION	0.97	0.76	0.92	0.96	0.87	89.6
SVM	0.98	0.76	0.93	0.98	0.9	91
MLP	0.98	0.64	0.92	0.98	0.87	87.8
KNN	0.9	0.76	0.82	0.93	0.62	80.6
RANDOM FOREST	0.98	0.74	0.91	0.98	0.89	90
VOTING CLASSIFIER	0.98	0.75	0.93	0.98	0.88	90.4
XGBOOST	0.96	0.75	0.9	0.97	0.86	88.8
ADABOOST	0.96	0.75	0.89	0.97	0.84	88.2
DECISION TREE	0.93	0.67	0.81	0.93	0.81	83
Linear SVM	0.98	0.71	0.93	0.98	0.87	89.4
LSTM	0.97	0.68	0.86	0.96	0.87	86.8
BI-LSTM	0.97	0.65	0.87	0.97	0.86	86.4
CNN-BI LSTM (HYBRID)	0.98	0.69	0.91	0.94	0.89	88.2
GLOVE (HYBRID) 6B	0.99	0.67	0.92	0.96	0.91	89
GLOVE (HYBRID) 42B	0.99	0.71	0.93	0.96	0.91	90
WORD2VEC (HYBRID)	0.98	0.68	0.92	0.97	0.91	89.2
FASTTEXT (HYBRID)	0.99	0.69	0.93	0.93	0.91	89

Figure 2. Accuracy for ML Models.

Coming to DL algorithms Figure 2 shows the accuracy of Deep Learning algorithms for five datasets. The hybrid model with GloVe and FastText embeddings achieved Word2Vec embeddings also had an accuracy of 98, and the average accuracy for all hybrid models with embeddings was 98.75. Other models scored an accuracy of 97.33, with a negligible difference of 1.42. On the Liar Dataset, Hybrid Model with GloVe 42b embeddings performed best with 71, while Bi LSTM scored lowest (65). The hybrid model with embeddings achieved an accuracy of 68.75, and other models reported a similar score of 67.2. For Fake News Dataset, the hybrid model with embeddings had an average accuracy of 92.5 compared to others with a lower score (88). This trend continued in Fake News Detection too. In Combined Dataset, all Hybrid Models with embeddings performed well with 90, followed by Hybrid Models with 89, but Bi LSTM performed poorly at just 86.

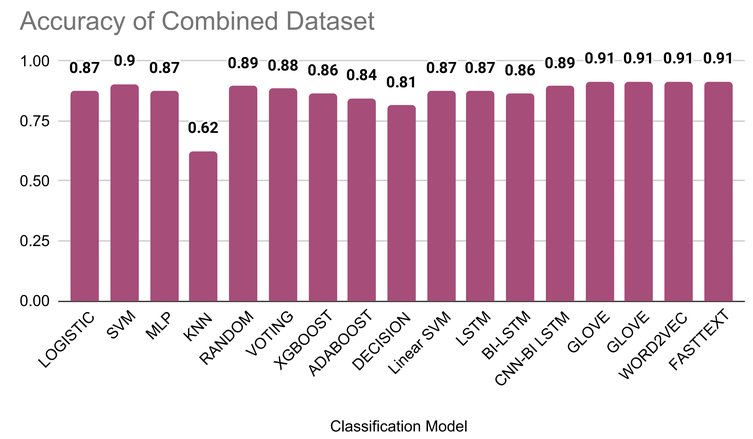

Figure 3. Accuracy for Combined Data.

The summary of average accuracy for all ML models across 5 datasets can be obtained from figure 2. SVM, Voting Classifier, Random Forest and Linear SVM are the best performers, where as KNN and Decision are tree poor performers. In DL models Hybrid Model with Word2Vec and GloVe42b embeddings are best performers. LSTM and Bi LSTM have the lease accuracy.

Future Research

For future research, we can embed datasets that will be trained using deep learning models and select the best model for future work. Using transfer learning, we can input data into the best deep learning models and get the new feature set using CNN. This new feature set can be used to train over ML algorithms, potentially improving the classification task's accuracy. Also, other embeddings, such as BERT or ELMo, can be used to compare their performance with the embeddings used in this project. Furthermore, other techniques for combining datasets can be explored and used in ensemble methods to improve the performance of the models. Finally, the model's behavior can be analyzed to interpret their decisions to gain insights into their internal mechanisms.

Phone No: 9110784930

krcreddy200226@gmail.com