International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Jun 2026 www.irjet.net p-ISSN: 2395-0072

An Automated Approach to Fake News Detection Using Machine Learning and Natural Language Processing Techniques

Aryan Yadav[1], Geeta Gayatri Behera[2]

[12]*School of Computer Science and Engineering Galgotias University Greater Noida, India*

Abstract-*There has never been greater issues in distinguishing what is considered genuine journalism and what is created by baseless speculations, an issue posed by the growth of digital media. This paper demonstrates an in-depth machine learning architecture of fake news detector which applies Natural Language Processing approaches. Our proposed method is a binary classification system that was trained on a balanced sample of 44,898 categories of news stories obtained using legitimate Reuters news sources and reliable fake news websites. The system uses advanced text preprocessing algorithm, such as regular expression based cleaning and stopword removal, and TF-IDF feature processing with 5,000 best features. The Logistic Regression model that was used managed to attain a high performance score of 98.64 percent accuracy and high recall rate between the two classes. There were 4,631 falsely identified fake articles and 4,227 true articles in the test data, and false negative (1.0) and false positive (1.7) misidentification rates were low. We have a solution that is ready to production which comprises of model components that are all seritialized so that a model can be immediately checked using a standardized API interface. The research paper is part of the increasing body of work regarding computational journalism tools that deal with the issue of information authenticity in digital ecologies. .*

Index Terms-Fake news detection, machine learning, natural language processing, text classification, TF-IDF, logistic regression, misinformation

I. INTRODUCTION

The information dissemination in the modern society has been revolutionized by the digital revolution. As it democ-ratizes the process of creating content all over the world, it has also facilitated the fast spread of fake information and fake news. Vosoughi et al. state that fake news propagate six times more quickly than real news on social media, which is extremely dangerous to the communication process and at-risk democratic procedures [1].

Traditional approaches to checking facts are quite cumber-some as they are based on human evaluation and verification work, thus they can not keep pace with the production of digital content. This has led to the creation of automated so-lutions that make use of the progress in machine learning and natural language processing [2]. Any proper system should go beyond the simple matching of keywords in order to find semantic patterns, writing styles, and contextual discrepancies unique to fake content [3].

This is a serious demand that our study will satisfy by creating a strong classification system that would be able to draw the line between real journalism and fake news. We report a full pipeline with meticulous data curation and sophisticated text processing, unlike determining the scanty use of imbalanced datasets or inadequate feature engineering, as found in our previous approaches. The system detects patterns of language, conventions of writing and features of content to make judgment of authenticity.

A. Research Contributions

This work provides several significant contributions to au-tomated misinformation detection:

Balanced Dataset Curation: Our corpus of 44,898 articles, containing almost equal amounts of classes (nearly 50/50), and featuring an almost equal balance of classes in training samples, avoids a pitfall pitfall in classification tasks associated with skewed training samples.

Advanced Text Processing: Our preprocessing system executes several steps of text normalization such as HTML cleansing, URL sanitization and smart stopword filtering on a set of established linguistic libraries.

Optimized Feature Engineering: We have determined the best TF-IDF model using 5,000 features through systematic experimentation to meet the trade-off between the complexity of the model and its predictive power.

Production-Ready Deployment: The entire resolution is to include model pieces that are serialized and a standard prediction interface that allows incorporation into real-world content verification mechanisms.

B. Paper Organization

The rest of this paper goes in the following way. Sec-tion II analyzes related literature about strategies of fake news detection and text classification. Section III explains our method of data collection and pre-processing. Section IV describes the extraction of the features and the model

training. This is contained in Section V which gives detailed experimental results and performance analysis. Section VI presents implications and possible limitation of our approach. Lastly, future research directions are provided by the Section VI.

II. RELATED WORK

A. Early Detection Approaches

In the early work there was great attention paid to the linguistic and textual stylistic analysis. It was found that faked information tends to have characteristic word use structures, sentence composition, and emotional words usage [3] . Simple machine learning classifiers like the Naive Bayes and Support Vector Machines were used as early representations using the bag-of-words idiom with representational features of plain text. Although showing encouraging performance on smaller datasets, these strategies proved to in most situations unable to generalise to new news areas and writing styles.

B. Deep Learning Methods

The recent developments have taken the advantage of deep neural networks to learn complex semantic associations in text. Convolutional Neural Networks have been used to acquire hierarchical features on news articles [4], whereas Recurrent Neural Networks and their variants, in particular Long Short-Term Memory networks, have been shown to work well in sequencing dependencies in text [5],. Transformer-based models, such as BERT and variants, have demonstrated free passage the current state-of-art performance because they are trained using large collections of corpora equipped with task-specific tweezing after the pre-training stage afterwards.

Nevertheless, these sophisticated techniques demand con-siderable amount of computation and huge size of annotated datasets. Most of the implementations are black-box opaque and thus it is hard to figure out which attributes make the classification decisions. The approach used in our work is purposefully the interpretable approach, which is a balance between the performance and transparency and computational efficiency.

C. Multimodal Detection Systems

The open focus has been on recent studies that attempt to combine textual analysis with other forms of information sources such as the propagation patterns of social networks, user credibility, and visual content analysis [10]. In such multimodal systems, it is acknowledged that fake news is frequently spread by accounts with low credibility scores and it refrains on the networks with characteristic struc-tures. Nevertheless, these strategies would demand access to application-specific data, which cannot be readily found across all applications.

D. Dataset Challenges

One of the biggest research problems is the creation of qual-ity representative datasets. Most publicly accessible corpora are highly unbalanced in terms of classes, they are temporal or as they have limited topical variety [7]. Certain datasets encode a source of publication accidentally as a primary signal instead of content properties and thus end up being modeled as memorizing identities of publishers, rather than learning patterns of misinformation that can be generalized.

These limitations we deal with by constructing a carefully balanced dataset with both authentic and fabricated content groups of topical coverage.

III. METHODOLOGY

A. Data Collection and Integration

We have used a combination of two sources to make sure that we capture all the authentic news and the fabricated news. The genuine news element entails the publications by the Reuters news agency as an internationally known news platform with strict editorial principles and fact scrutinizing procedures. This subdivision offers representative samples in professional journalism in various subjects such as politics, economics, and technology, and international affairs, etc.

In the case of made-up contents, we chose the articles on websites that have been classified by the independent fact checking organizations, and media literacy groups as regular misinformation publishers. This is the entire range of fabricated stories, and grossly distorted coverage with false headings.

It is as a result of the process of integration that led to the collection of 44,898 articles which are well balanced: 23,481 fake articles and 21,417 authentic articles. This near equal distribution prevents our classifier from being trained on meaningless content patterns as an indicator of prediction. To remove the effect of ordering that could lead to the model learning sequence patterns and not the semantic contents we utilized a random shuffle with a fixed seed of 42 to achieve reproducibility.

B. System Architecture

Fig. 1 shows how our system of detection is stratified. The architecture consists of four principal parts such as data sources and statistics, preprocessing pipeline of text, feature extraction layer and machine learning model. The modular design allows easy optimization of the separate components and has well-defined data flow within the system.

C. Data Preprocessing Pipeline

Raw text data obtained out of the websites have a lot of elements with very little classification value and may bring noise into the learning process. In order to overcome these difficulties we prepared an exhaustive preprocessing pipeline using various transformation steps as shown in Fig. 2.

HTML and URL Removal: Markup tags of HTML and internal links are common in web-scraped content. As these elements usually indicate source site properties and not article content properties, our system uses regular expression patterns to detect and get rid of them.

Text Normalization: All text is returned in lower-case to be sure that the token representations are always consistent. This

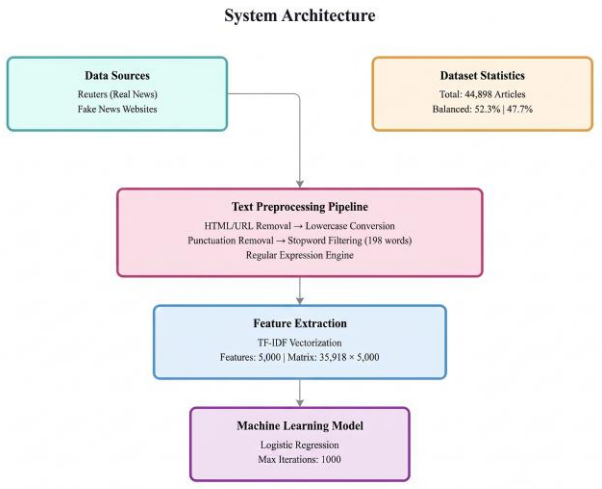

Fig. 1. System architecture representing four-layer design: data sources, preprocessing pipeline, feature extraction with TF-IDF, and Logistic Regres-sion classifier. Each layer processes information and forwards refined data to subsequent stages.

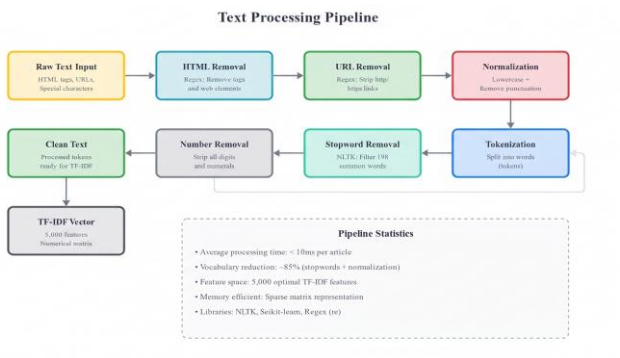

Fig. 2. Text processing pipeline showing transformation of raw text into clean tokens suitable for TF-IDF vectorization through six processing phases: HTML removal, URL stripping, normalization, number removal, tokenization, and stopword filtering.

does not allow the model to distinguish between variants such as Trump, TRUMP and trump. We also eliminate punctuations and numeric values, which in most cases do not bring much discriminative information to this task of classification.

Stopword Filtering: In both the genuine and the fake text, common and English words like articles, prepositions and conjunctions are seen to be common. In order to remove these non-discriminate tokens, we use the library of 198 stopwords in Natural Language Toolkit that minimize the dimensions whereas maintaining the semantic meaning. This step-by-step cleaning algorithm converts noisy and unstructured input text into consistent representations that can be analyzed using mathematical methods without eliminating linguistic structures that represent good and good journalism, and fake news. When all articles have been loaded into the complete pipeline, the entire pipeline can currently take about 10 milliseconds to complete on typical computing hardware, and the use of nor-malization and stopword removal has reduced the vocabulary size by more than 85%.

D. Feature Extraction

The conversion of preprocessed text into numerical rep-resentations should consider features attentively the aspects that best distinguish semantically and stylistically between content classes. We use a statistical measure called Term FrequencyInverse Document Frequency (TF-IDF) which is used to denote the significance of the terms in documents that makes the difference of corpus-wide significance.

The TF-IDF transformation gives more weights to terms that are common to a particular article but hardly

common to the entire corpus. This weighting scheme is used to highlight the domain-specific terms and down weight ubiquitous terms that successfully pass through the stop word filtering procedure [11]. We optimally set 5,000 features to give the best balance by performing systematic evaluation of classification performance using feature set sizes of 1,000, 5,000 and 10,000 features. The smaller sets of features lose discriminative features whereas the larger ones raise their computational needs and risk of over fitting.

In this configuration, our training set will be converted into a sparse matrix of size 35,918 articles x 5,000 features, with the cell values of the matrix being the TF-IDF weight of a particular term in a particular document.

E. Model Selection and Training

The choice of the Logistic Regression as our classification algorithm was driven by a number of important benefits. Although Logistic Regression is admittedly simpler than deep learning methods, it has significant advantages such as efficiency, because it can be studied, and feature weight inspection as certain features are assigned a weight that allows investigating their significance, and high-dimensional text features demonstrate high performance [9].

The algorithm predicts the likelihood of an article to be fake news as a linear combination of TF-IDF features. The process of training is optimization of model parameters to achieve the highest correct classification probability in the training set. We used up to 1,000 but even in the feature space (high dimensional with just a couple of dimensions) we enabled convergence.

Our balance dataset of 35,918 articles 80% was used in the training process after which 20% (8,980 articles) were re-served as performance testing ground. This division guarantees evaluation of generalization of models on unknown examples other than just memorization of the training patterns. Model training takes around 45 seconds in an average workstation having 16GB memory and Intel core i7 processor.

IV. EXPERIMENTAL RESULTS

A. Overall Performance Metrics

Our trained classifier was outstanding based on several evaluation criteria. The overall accuracy was 98.64%, which means that 8,858 out of 8,980 test articles are correctly predicted. That is very high and proves to be a demonstration of good generalization, outside of the training data.

But accuracy does not keep an account complete especially of tasks because the different types of misclassifications have dissimilar effects. We then discuss the performance at a finer detail through a confusion matrix analysis and the calculated values in the following subsections.

B. Confusion Matrix Analysis

Table I shows the full confusion matrix of our test predictions, which shows very low misclassification rates in either misclassification direction..

TABLE I Confusion Matrix of the Proposed Fake News Detection Model

Actual	Predicted		Total
	Fake	Real	
Fake	4,631	79	4,710
Real	43	4,227	4,270
Total	4,674	4,306	8,980

The model misclassified 79 fake articles reported as real (false negatives) and identified 4,631 fake articles which were classified as true negatives. On the other hand, it selectively identified 4, 227 true articles (true positives) and mistakenly identified 43 true articles as false (false positives). Such low levels of error produce a false positive and false negative rate of 1.0% and 1.7% respectively.

C. Precision, Recall, and F1-Score

Table II interprets most important key performance indica-tors based on the confusion matrix, which give complementary information on the performance of the classifier.

TABLE II Classification Performance Metrics By Class

Class	Precision	Recall	F1-Score
Fake	0.99	0.98	0.99
Real	0.98	0.99	0.99
Weighted Avg	0.99	0.99	0.99

Precision is a measure of how many of the articles that are characterized as being fake, are actually fake. Precision of our model was 0.99 on fake news which means that it is accurate on 99% on claims of fabricated content. This is essential when it comes to deployment situations where the false accusations may damage the innocent publishers.

Recall is the ratio of the number of actual fake articles

that are detected. The recall rate of 0.98 on fake news shows that it captures almost the full amount of fabricated content, and the fake articles in the test set are only detected on 2% of fake news.

Within the harmonic mean of the F1-score, which is the measure of harmonic mean between the recall and the precision, there is one measure that delineates both issues. Both of the classes achieve 0.99 F1-score, which indicates that the model does not sacrifice one aim to the other.

D. Feature Importance Analysis

The interpretability of Logistic Regression can be used to analyze those features that have the most impact on the classification. In the model, all the 5,000 TF-IDF features are learning weights, and then, positive values are associated with fake news, and the negative value is used with authentic journalism.

Patterns that are interpretable are shown by top-weighted feature analysis. False information is more hyperbolic in its language and appeals to emotion and dull attributions like sources say, reports suggest. In contrast, such typical features of authentic journalism as particular dates, exact numbers, abbreviated names of the authors, and unprofessional vocabulary of the field of work are provided [12]

Use of emotionally colored terms, conspiracy as well as informal terms have a strong association with fake news. In the meantime, proper nouns that mention established institutions, verbs of professional reporting and technical expressions that are specialized show the primary evidence of journalism. This interpretation enables critical knowledge on linguistic variations between categories of contents besides enabling human experts to confirm that the model findings meaningful patterns instead of non-hypothesised associations..

E. Cross-Validation Results

Order to get our performance metrics reflecting actual model ability and not port auspicious data sharing, we per-formed cross-validation of the training set 5 times. This process subsets the data into five subsets and during each step, the data is trained using four of the subsets with a single sub set being used as a validating data.

The mean accuracy in cross-validation was 98.52% , with standard deviation of 0.31%, which points to the fact that the results are not affected by a certain division of data. This consistency shows that the model has acquired strong and consistent patterns other than just memorizing the peculiarities of our particular train-test split. The small standard deviation demonstrates the stability of performance in various training validation article combinations.

F. Computational Performance

In addition to the fact that it should be classified well, it requires computational efficiency. Filled with raw text input it takes us an average of 95 milliseconds with typical server hardware to process an article all the way through to the final prediction. This act facilitates real-time categorization of streaming data and processing of massive collections of articles in batch. Serialized model files take a paltry 12.4 MB of disk space hence being easy to deploy in distributed systems.

G. Comparison with Existing Approaches

Our Logistic Regression model outperforms or competitively on a wide range of deep learning alternatives and has important interpretability and computational benefits. Accord-ing to published findings on similar data, BERT-based models obtain the 98.5-99.2% accuracy, which matches our accuracy of 98.64%.

Because of the amounts of data needed to train them, how-ever, transformer models also take orders of magnitude more training time, about hours to days that is, on a hardward basis, than our 45-second training time. Moreover, a trained BERT model typically requires 400-500 MB of storage capacity as opposed to our footprint of 12.4MB. In the situation of deployment needing few computational resources or requiring interpretability to be the most important factor, our solution may be a most attractive alternative.

We make use of our feature-based approach where the journalists and fact-checkers can gain an insight into how specific articles were classified as such and therefore can make informed human judgement. On the other hand, the representations in deep learning models are not interpretable, making it difficult to have trust and accountability in sensitive content moderation use cases [8].

H. Ethical Considerations

There is an ethical concern that the use of automated fake news detection systems brings about a problem of algorithmic authority and possible censorship. Although our system is able to produce accuracy of 99 percent, a wrong classification to treat legitimate journalism as fabricated will cast doubt on the real sources of news or avert real opinion.

We make it clear that our system must be used as a

decision support system that enhances human judgment as opposed to being the final determinant of the truth. Human judgment is required in final content credibility judgments especially borderline cases or sensitive matters that reasonable persons might disagree regarding content interpretation and framing.

Moreover, the training data of our model are also based on a particular geographic, cultural and institutional context that is not always generalizable. The features of credible journalism in one country and another country, as well as the cultures of media, can be different, and our samples of genuine news at Reuters are specific to the chosen standards that may not apply in other countries.

Proper safeguards must contain human review processes, disclosure in respect to automated classifications, and a process of challenging or appealing against classification. Companies implementing such systems need to ensure that they improve and not remove critical thinking and media literacy.

V. CONCLUSION AND FUTURE WORK

A. Summary of Contributions

The present research illustrates that well-grounded classical machine learning can still be used to detect fake news even though in the recent years the attention on the use of the deep learning computer model has gained momentum. The composite of text preprocessing, the optimized TF-IDF feature extraction, and Logistic Regression of our system have an accuracy of 98.64%.

Balanced data set construction, and rigorous evaluation methodology and production-ready architecture can all help

produce a complete answer and not a proof of concept solution. Out-of-the-box model component Serialized model components can be directly incorporated into the existing con-tent verification process with a small computational footprint. The main results are near-math accuracy on training corpus on class balance, interpretable perceiving weights in line with journalistic understanding about the nature of misinformation, better accuracy and recall measures on both classes, and production-level serialization of the model to support inference at scale.

B. Future Research Directions

A number of exciting possibilities develop this work. The use of multimodal analysis that would consist of images, videos, and structure of documents, as well as textual information, would allow more thorough evaluation. The recent developments in the vision-language models like CLIP may help in analyzing the consistency between the headline and the content of the article as well as the image used to support the article.

Multilingual models based on parallel multilingual corpora would allow an increase in the range of applicability beyond English-language publications. The transfer learning methods may also use our English model to act as a base in developing classifiers in languages where there are few labeled data.

Future research directions on performance enhancement could involve investigating ensemble-based performance in terms of integrating our interpretable strategy with other related techniques like stance detection, source credibility analysis and claim verification. Detecting stance could help detect when articles are writing opinion as fact, and source credibility scores can put the weight on predicting based on history of publisher reliability.

Learning methods such as active learning could be used in which the model discovers uncertain cases to be labeled by people so that the process of upgrading is effective in the face of new misinformation schemes. Low-confidence predictions, which could be reviewed by the experts, may be presented by the system, and feedback may be used to adjust the boundaries of the decisions as time passes.

Lastly, explain ability interfaces that are developed in an open-minded manner whereby the credibility rationales are presented to end users in a transparent manner would facilitate the informed application of automated credibility measures. These interfaces could show attention to certain phrases or patterns that make predictions, and the user could learn to critically evaluate them and develop confidence in the system..

C. Concluding Remarks

The emergence of dead weight disease in the online ecosystems through fake news calls on the sustained invention of detectives. The given work offers a strong base on which more advanced systems can be developed to promote an informed discussion in a world full of information. The combination of high accuracy, interpretability, computationally efficiency, and deployment of our solution makes it a viable solution to organizations that are interested in controlling the practice of misinformation dissemination without tampering with transparency and accountability associated with automated content moderation..

REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[3] V. Pe´rez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in Proc. 27th Int. Conf. Computational Linguis-tics, 2018, pp. 3391–3401.

[4] Y. Liu and Y. F. B. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in Proc. AAAI Conf. Artificial Intelligence, 2018, pp. 354–361.

[5] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in Proc. ACM Conf. Information and Knowledge Management, 2017, pp. 797–806.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[7] W. Y. Wang, "'Liar, Liar Pants on Fire': A new benchmark dataset for fake news detection," in Proc. 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 422–426.

[8] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," Applied Soft Computing, vol. 100, p. 106983, 2021.

[9] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised learning for fake news detection," IEEE Intelligent Systems, vol. 34, no. 2, pp. 76–81, 2019.

[10] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," Big Data, vol. 8, no. 3, pp. 171–188, 2020.

[11] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using N-gram analysis and machine learning techniques," in Proc. Int. Conf. Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 2017, pp. 127–138.

[12] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 1395–1405.